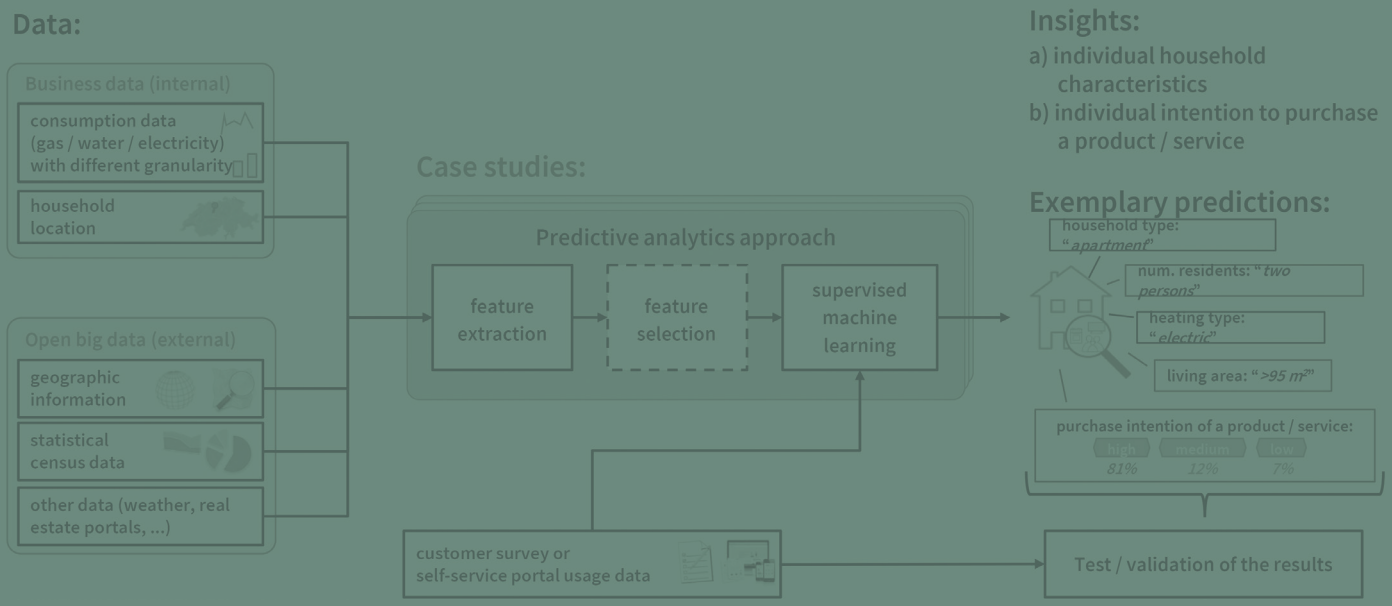
36

Schriften aus der Fakultät Wirtschaftsinformatik und
Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Predictive Analytics for Energy Efficiency and Energy Retailing

Konstantin Hopf



University of Bamberg Press

36 Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Contributions of the Faculty Information Systems and Applied Computer Sciences of the Otto-Friedrich-University Bamberg

Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Contributions of the Faculty Information Systems and Applied Computer Sciences of the Otto-Friedrich-University Bamberg

Band 36

University of Bamberg Press
2019

Predictive Analytics for Energy Efficiency and Energy Retailing

Konstantin Hopf

University of Bamberg Press
2019

Bibliographische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.d-nb.de/ abrufbar.

Diese Arbeit hat der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.
1. Gutachter: Prof. Dr. Thorsten Staake
2. Gutachterin: Prof. Dr. Ute Schmid
Tag der mündlichen Prüfung: 30.04.2019

Herstellung und Druck: docupoint, Magdeburg
Umschlaggestaltung: University of Bamberg Press

ISSN: 1867-7401
ISBN: 978-3-86309-668-7 (Druckausgabe)
eISBN: 978-3-86309-669-4 (Online-Ausgabe)
URN: urn:nbn:de:bvb:473-opus4-548335
DOI: http://dx.doi.org/10.20378/irbo-54833

Contents

List of Figures

List of Tables

List of Abbreviations

AIS Association for Information Systems

ANN Artificial Neural Network

AUC Area Under ROC Curve

BRG Bias Random Guess

CBR Customer Based Reputation

CPD Consumption Per Day

CRM Customer Relationship Management

CTA Call-to-Action

DiD Difference-in-Differences

FTTH Fiber-to-the-Home

FSM Feature Selection Method

GPS Global Positioning System

HT High Tariff ("Hochtarif")

IoT Internet of Things

IS Information Systems

IT Information Technology

kNN k Nearest Neighbors

LDA Linear Discriminant Analysis

ML Machine Learning

MCC Matthews Correlation Coefficient

NILM Non-Intrusive Load Monitoring

NPS Net Promoter Score

NT Low Tariff ("Niedertaif")

OLS Ordinary Least Squares

OSM OpenStreetMap

PI Purchase Intention

RF Random Forest

RG Random Guess

ROC Receiver Operating Characteristic

RQ Research Question

SMD Smart Meter Data

SVM Support Vector Machine

VGI Volunteered Geographic Information

XGB Extreme Gradient Boosting

Abstract

In the course of digitalization, large amounts of data are created. These are present in every single firm or they are freely available as public data online. They can be used, for example, to implement new products and services, or to make existing processes more efficient. The data are of diverse nature and range from transaction data as a result of business processes (such as purchases in online shops, usage information from video or music-on-demand providers, app usage data), communication data (e.g., correspondence, chat protocols), to sensor data from industrial plants or smart home devices. Additionally, a large number of online data sources have been created, which can be used freely. Examples are geographic information (e.g., from the freely editable map OpenStreetMap), weather data, and public statistics.

From a practical perspective, firms are searching for new business models and ways to commercially leverage the increasing amount of data. From a research perspective, a better understanding of the data value creation process is necessary, and success factors as well as obstacles to this process must be identified. Better knowledge of this process will enable the creation of economic, environmental, and social value from the growing amount of data.

Current information systems literature points to the need for empirical research in this area. Additionally, a comprehensive explanation of how to successfully create value from available data is missing, thus far. Value does not arise automatically from the use of data. There is rather a complex process necessary in which insights are first gained from data, leading to better decisions that ultimately can create value. Machine learning (ML) techniques, a class of artificial intelligence methods that derive patterns from data, have the particular potential to prepare the large amount of data in a way that knowledge can be generated from it. The insights can then create value through better decisions.

In my dissertation, I focused on answering five central research questions regarding predictive analytics along with the data value creation process. The resulting core contributions of my work, outlined below, help to achieve a better theoretical understanding of the value generation from data using ML methods. First, based on a systematic literature analysis of information system research journals, an overview of firm internal and external data sources for possible data

analyses was created. This overview has been validated and expanded with the knowledge gathered in seven case studies (two of which are described in detail in this dissertation), and gives firms the opportunity to inventory their databases and identify data sources for analysis. Second, throughout eight examples of empirical feature extraction, the work demonstrates how human cognition, theory, and expert knowledge can help in the effective preparation of data for ML applications, despite the many automated ways to preprocess data for further analysis. The findings point out that experts are needed to set up and effectively combine automated data analysis techniques. These professionals must be trained and promoted. Third, current ML algorithms for classification and variable selection methods were empirically benchmarked using real data sets from energy retailing. In this dissertation, recommendations for the use of data analyses in utility companies are derived. Concretely, I tested how strong unfavorable factors of modeling (i.e., a lower degree of detail in the raw data, and different geographical locations of the training and test data) affect the predictive quality. My investigation revealed, in particular, that despite these unfavorable factors negatively effect the predictive quality, the influence of the lower performance is ultimately not too strong as to impair the success of predictive systems. Finally, the successful use of the information gained through ML applications is shown through two main case studies in this dissertation. The step from insight to value is demonstrated using the examples of automated energy feedback and relationship marketing. The two case studies considered are based on real data from energy providers in Germany and Switzerland and cover all steps of the data value creation process.

The energy sector is an ideal field for the research carried out, as the demand for data-driven innovations is particularly high in that industry. Energy utilities must become more competitive in increasingly liberalized markets, but are mandated to motivate their customers to consciously conserve energy. Additionally, they must increase the acceptance of sustainable—often more expensive—energy products and pioneer new fields of business in order to implement the transformation of energy systems. On the contrary, energy utility companies have a large customer base and possess an increasing amount of data on their customers (e.g., from smart grids) that contain valuable insights.

In the first case study, the prediction of household characteristics related to residential energy efficiency (e.g., type of heating, age of house, number of occupants, children in the household) from energy consumption data together with freely available data using ML methods was investigated. Knowledge of such characteristics can be used to personalize energy efficiency campaigns and thus

make them more effective (e.g., through household specific savings recommendations, load estimation and shifting).

The second case study dealt with the recognition of customer attitudes and behavior (e.g., attitudes towards energy efficiency, willingness to buy photovoltaic systems or new products from energy providers). The successful prediction of such information enables the development of new products and services in the energy sector as well as their targeted promotion to relevant customer segments.

The use of ML-based data analysis in energy retailing can thus improve energy efficiency in the residential sector, increase customer value, and improve the service quality. Therefore, the dissertation shows how economic, ecological, and social value can be created from data and is a blueprint for other industries.

Kurzzusammenfassung

Im Zuge der Digitalisierung entstehen große Datenmengen. Diese sind in jedem einzelnen Unternehmen präsent oder als öffentliche Daten frei zugänglich im Internet verfügbar. Sie können genutzt werden, um beispielsweise neue Produkte und Dienstleistungen zu realisieren oder vorhandene Prozesse effizienter zu gestalten. Die Daten sind von unterschiedlichster Natur und reichen von Transaktionsdaten aus digitalen Geschäftsprozessen (wie Einkäufe in Online-Shops, Nutzungsinformationen von Video- oder Musik-On-Demand-Anbietern, App-Nutzungsdaten), Kommunikationsdaten (z.B. Schriftverkehr, Chatprotokolle) bis hin zu Sensordaten aus Industrieanlagen oder Smart-Home-Geräten. Im Internet ist zudem eine große Anzahl von Datenquellen entstanden, welche frei nutzbar sind. Beispiele hierfür sind geographische Informationen (z.B. aus der frei editierbaren Landkarte OpenStreetMap), Wetterdaten und öffentliche Statistikdaten.

Aus praktischer Sicht suchen Unternehmen nach neuen Geschäftsmodellen und Wegen, wie sie die steigende Menge an Daten kommerziell nutzen können. Aus dem Blickwinkel der Forschung ist es nötig, den Daten-Wertschöpfungsprozess besser zu verstehen und Erfolgsfaktoren sowie Hindernisse für diesen Prozess zu identifizieren. Mit der besseren Kenntnis dieses Prozesses kann ökonomischer, ökologischer und sozialer Wert aus der wachsenden Menge an Daten geschaffen werden.

Der Bedarf an empirischer Forschung in diesem Bereich wird aus der jüngsten Informationssystemliteratur deutlich. Bisher fehlt nämlich eine umfassende Erklärung, wie man aus den verfügbaren Daten erfolgreich Wert schaffen kann. Fest steht, dass Wert nicht automatisch aus der bloßen Nutzung von Daten entsteht, sondern ein komplexer Prozess zugrunde liegt, bei dem zunächst Erkenntnisse aus Daten gewonnen werden und infolgedessen sachkundigere Entscheidungen möglich sind, die dann schließlich Wert schaffen können. Verfahren des maschinellen Lernens (ML), eine Klasse von Methoden aus dem Bereich der künstlichen Intelligenz die Muster aus Daten ableitet, haben besonderes Potential, die große Menge an Daten so aufzubereiten, dass daraus Erkenntnisse entstehen, durch welche bessere Entscheidungen Wert schaffen können.

In meiner Dissertation habe ich mich mit der Beantwortung von fünf zentralen Forschungsfragen zu Predictive Analytics im Rahmen des Daten-Wertschöpfungsprozesses beschäftigt. Die aus der Arbeit resultierenden Kernbeiträge sind nachfolgend genannt und tragen zu einem besseren theoretischen Verständnis über die Wertgenerierung aus Daten mit Hilfe von ML-Verfahren bei. Erstens wurde, basierend auf einer systematischen Literaturrecherche in Zeitschriften der Informationssystemforschung, eine Übersicht über firmeninterne und externe Datenquellen für mögliche Datenanalysen erstellt. Diese Übersicht wurde mit den Erfahrungen von sieben Fallstudien (zwei davon sind in dieser Dissertation im Detail ausgeführt) validiert sowie erweitert und gibt Unternehmen die Möglichkeit, ihre Datenbestände zu inventarisieren oder Datenquellen für die Analyse zu identifizieren. Zweitens stellt die Arbeit anhand von acht Beispielen dar, wie die kognitiven Fähigkeiten des Menschen, Theorie und Expertenwissen helfen können, Daten effektiv für ML-Anwendungen aufzubereiten, auch wenn zahlreiche automatische Verfahren existieren, um Daten für die weitere Analyse vorzubereiten. Die Erkenntnisse unterstreichen, dass Experten benötigt werden, um automatische Datenanalyseverfahren aufzusetzen und effektiv zu kombinieren. Diese Fachkräfte müssen ausgebildet und gefördert werden. Drittens wurden aktuelle ML-Algorithmen zur Klassifikation und Variablenselektionsmethoden mithilfe von realen Datensätzen aus dem Energievertrieb empirisch verglichen. In der Arbeit werden daraus Empfehlungen für den Einsatz von Datenanalysen in Energieunternehmen abgeleitet. Hierbei wird insbesondere deutlich, dass sich eine geringere Detailtiefe in Rohdaten oder eine unterschiedliche geographische Lokation der Trainings- und Testdaten zwar negativ auf die Vorhersagegüte auswirkt, der Einfluss jedoch nicht so stark ist, dass dieser am Ende den Erfolg von prädiktiven Systemen beeinträchtigt. Schließlich wird der erfolgreiche Einsatz der gewonnenen Informationen durch ML-Anwendungen anhand von zwei Fallstudien, welche in der Dissertation beschrieben werden, aufgezeigt und der Insight-to-Value Schritt an den Beispielen des automatisierten Energiefeedbacks und des Beziehungsmarketings aufgezeigt. Die beiden Fallstudien im Fokus dieser Dissertation basieren auf realen Daten von Energieanbietern aus Deutschland sowie der Schweiz und decken alle Prozessschritte des Daten-Wertschöpfungsprozesses ab.

Die Energiebranche bietet sich für die durchgeführte Forschung an, da der Bedarf an datengetriebenen Innovationen dort besonders hoch ist: Energieanbieter müssen einerseits in zunehmend liberalisierten Märkten wettbewerbsfähiger werden, sind jedoch andererseits durch den Gesetzgeber angehalten, ihre Kunden für einem bewusst sparsamen Umgang mit Energie zu motivieren. Darüber hinaus müssen sie die Akzeptanz von nachhaltigen—oft teureren—Energieprodukten

steigern und neue Geschäftsfelder erschließen, um die Transformation der Energiewirtschaft, welche sich aus der Energiewende ergibt, umsetzen zu können. Andererseits verfügen Energieanbieter über eine große Kundenbasis und haben Zugriff auf eine zunehmende Menge an Daten (z.B. aus intelligenten Stromnetzen).

In der ersten Fallstudie, die in dieser Arbeit dargestellt wird, wurde die Erkennung von Haushaltseigenschaften in Bezug auf die Energieeffizienz (z.B. Heizungstyp, Alter des Hauses, Anzahl der Bewohner, Kinder im Haushalt) aus Energieverbrauchsdaten zusammen mit frei verfügbaren Daten mit Hilfe von ML-Verfahren untersucht. Die Kenntnisse über solche Merkmale lassen sich nutzen, um Energieeffizienzkampagnen zu personalisieren und damit wirkungsvoller zu gestalten (z.B. durch haushaltsspezifische Sparempfehlungen, sowie Lastabschätzung und -verschiebung).

Die zweite Fallstudie behandelte die Erkennung von Kundeneinstellungen und -verhalten (beispielsweise die Einstellung zu Energieeffizienz, die Kaufbereitschaft von Photovoltaikanlagen oder neuen Produkten von Energieanbietern). Mit Hilfe der ermöglichten Vorhersagen können neue Produkte und Dienstleistungen im Energiebereich entwickelt sowie zielgerichteter vermarktet werden.

Durch den Einsatz von ML-basierten Datenanalysen im Energievertrieb kann somit die Energieeffizienz im Privatsektor verbessert, der Kundenwert gesteigert und die Servicequalität verbessert werden. Die Dissertation zeigt damit schlussendlich auf, wie ökonomischer, ökologischer und sozialer Wert aus Daten generiert werden kann und ist damit eine Blaupause für weitere Branchen.

Acknowledgements

This dissertation is a result of my work as a researcher at the Chair of Information Systems and Energy Efficient Systems at the University of Bamberg (Germany) in the period from 2015 to 2019 and my visiting stay at the Copenhagen Business School (Denmark) in 2018. My work was enriched by projects and studies in the context of the Bits-to-Energy Lab, a joint research initiative of the ETH Zurich, the University of St. Gallen, and the University of Bamberg that also involved industry partners.

Foremost, I want to express my sincere gratitude to my doctoral advisor Thorsten Staake. He has provided me guidance, support, and honest feedback, ever since I started working as a research assistant in 2014 and wrote my Master thesis at his group. Thorsten inspired me to create new ideas and think out the box through his open mind which overarches many disciplines. Next, I want to thank Ioanna Constantiou. Far more than enabling my research stay at the Copenhagen Business School, she helped me to better understand the Information Systems research discipline and to more clearly identify the theoretical contribution of my, so far, rather applied research. I also thank Mariya Sodenkamp, for being a mentor between 2014 and 2016.

Research is a collaborative effort. Therefore, I want to deeply thank my colleagues and peers at the Bits-to-Energy-Lab. First of all, Ilya Kozlovekiy, Liliane Ableitner, and Andreas Weigert. It was a pleasure to work with you on several studies and research projects that covered significant parts of my work. Without our teamwork, the entire work would not have been half as exciting. Next, I would like to thank my colleagues and the cohort of PhD students at the Chair of Information Systems and Energy Efficient Systems at the University of Bamberg for their friendly support and the nice work environment: Anna Kupfer, Sarah Appeldorn, Jürgen Wenig, Samuel Schöb, Sebastian Günther, Viktoria Hirschfeld, and Carlo Stingl. Beyond the boundaries of our group in Bamberg, I would like to thank Verena Tiefenbeck for many interesting discussions and the exchange. It was also great opportunity to be in contact with the research groups around Elgar Fleisch at the Chair of Information Management at ETH Zurich and at the Chair of Operations Management at the University of St. Gallen. All doctoral consortia in which I participated were advantageous.

Furthermore, I was fortunate to collaborate with excellent students on their Master thesis topics that we managed to continue. These research ideas have always been a source of inspiration, and we managed to translate some curious ideas into scientific contributions. So, many thanks to Michael Kormann, Sascha Riechel, Carlo Stingl and Felix Jungmann.

My work was only possible through the financial support of public funding bodies and industrial partners who shared much resources and provided practical experience. My work was financially supported by the Swiss Federal Office of Energy (grant numbers SI/501053-01 and SI/501202-01), the Swiss Commission for Technology and Innovation (grant number 16702.2 PFEN-ES), and the European Commission and member countries (grant number E!9859). I thank the industry partners, especially the whole BEN Energy team for the amazing and fruitful cooperation, in particular, Tobias Graml, Jan Marckhoff, Claire-Michelle Sévin, Matthias Dhum, Lutz Gegner, Sarvesh Dwivedi, and Simon Tietze. In regards to the utility companies, I particularly appreciate the cooperation with Arbon Energie AG, Werkbetriebe Frauenfeld and Centralschweizerische Kraftwerke AG.

Finally, I want to express my heartfelt thanks to my family and friends. Most of all, I would like to thank my dear wife, Julia, who has taken on much to support me in realizing my dreams, this dissertation in particular. She supports me always and is my best advisor. I further thank my parents and brothers for all the confidence and backing. I also thank my friends on whom I can always rely, especially Martin Radenz, Christian Lang and Haley Culpepper for feedback on this dissertation from an outsider's perspective. Without all your support, I would have never started and finished this endeavor. Thank you so much for everything!

1 Introduction and motivation

Digitalization is advancing in commercial as well as private environments. Electronic devices in firms and at home are increasingly equipped with network connections and produce large amounts of data that contain valuable information. Information Technology (IT) and the Internet of Things will enable smart cities and make life in urban environments more efficient, sustainable, and resilient (Brandt et al. 2018). The energy system will be soon connected through a smart grid to meet the requirements for a stable and sustainable energy supply (Ketter et al. 2018). At home, sensors and automatic controls for lighting, household appliances, heating, and air conditioning will take over routine tasks. Digitalization thus generates data that are not only huge in numbers, but also diverse in nature (Hashem et al. 2016). It is a great challenge to process this data in a meaningful and value-adding way. To illustrate the magnitudes of data that are already being processed by ▷Information Systems (IS) today, one can take a look at large enterprises: Walmart, a retailing firm, is said to process 2.5 petabyte[1] of data every hour (Marr 2017).

Massive amounts of diverse data are created through digitalization

The above mentioned examples give an idea on how IT is becoming more embedded and omnipresent in our life. Sensors will be nearly everywhere and the acceptance of connected devices in broad areas of life will increase further, because interaction between humans and IT devices continuously improves and will run even more smoothly in the future. This development leads to the fact that detailed data about peoples' living conditions, habits, and behavioral patterns are recorded. Therefore, individuals voluntarily disclose—wittingly or unwittingly—information

The data contain details on living conditions and behavior

[1] One petabyte stands for 10^{15} bytes.

that was previously difficult to collect while using smartphones, smart watches, voice controls, etc.

Data have high strategic relevance for firms, but realizing value from raw data is difficult

Newspapers and consultancies recon that data are the "new oil", or the "new gold", and describe thereby the prospects ascribed to them. Though these analogies have weaknesses, oil was the business enabler for a century, and data have the potential to be it for the current one (The Economist 2017; Bergers and Meijerink 2017). Data are advantageous sources for business model innovations and the development of data-driven services. According to a recent investigation by Hartmann et al. (2016), a majority of startups rely on business models that use data as resources. The relevance of big data for corporate strategy is well recognized from research (Constantiou and Kallinikos 2015; Yoo 2015; Kallinikos and Constantiou 2015). Moreover, the World Economic Forum (2011) describes personal data as a "new asset class" due to its economic value for firms and the global political agendas concern about the use of personal data[2].

ML can bridge the data-value gap

The value of data—an intangible asset—is hard to estimate, considering the fact that a piece of data can be useless for one application, but highly relevant for another. However, methods of Machine Learning (ML) and predictive analytics are powerful tools for realizing value from data. This was shown in a number of insightful studies. In the *banking industry*, for example, credit card fraud can be detected from payment transaction data (Bhattacharyya et al. 2011). Martens et al. (2016) analyze similar data to obtain the customer interest in financial products, and machine learning is successfully used for credit scoring (Kruppa et al. 2013). Shi et al. (2017) identify the source of incidents in the *aviation* business. In *retail*, G. Cui, Wong, and Wan (2012) identify high-value customers in marketing campaigns, Shrivastava and Jank (2015) predict customer spending during promotional events. In the *telecommunication* business, interaction data of cell phone users are utilized to predict the customer's insolvency risk (Daskalaki et al. 2003) or contract cancellation (Backiel et al. 2014; Braun and Schweidel 2011).

Successful ML applications in the banking, aviation, telecommunication and retail industry

[2]The European Commission (2017) works on rules for a "European Data Economy" and the G20 countries works on a world-wide digital political agenda (OECD 2017).

This work examines the reasonable use of ML on data available to organizations

Despite its success, initial euphoria about rapid gains from applying ML to data has faded. Companies as well as researchers have realized that the simple application of algorithms to internal business data does not necessarily bring sustainable value or competitive advantage. Reasons for this disenchantment are, among others, the awareness that ML is not a universal remedy for all problems, that the quality of predictions is sometimes not sufficient enough for real applications, that high quality data are needed for successful predictions, that detailed domain knowledge is required to understand the data, and that expert knowledge is needed in order to apply algorithms correctly and evaluate the results. To foster the meaningful use of ML on data that are available to organizations, which I refer to as *ambient data*, this dissertation answers five research questions along the data value creation process. Two case studies from the energy retailing industry serve as empirical base for the investigation.

Following sections introduce the research topic and outline the content of this dissertation

The remainder of this chapter first introduces ambient data as a new source for analytics. Second, the research topic of this dissertation—the value creation from ambient data with the help of machine learning—is outlined and the research questions are presented at a glance. Thereafter, the two case studies from the energy retailing industry are presented which are adduced to answer the research questions. Finally, the structure of the dissertation is shown and the contents of the individual chapters briefly summarized.

1.1 Ambient data as a new source for analytics

Showcases of data analytics, so far, often rely only on structured and firm-internal data

Successful examples, in which the value-realization from data through ML is already showcased, are often limited to the use of firm-internal business data. These data are typically available in structured formats (relational databases, spreadsheet files, etc.), well embedded into business processes, and organized to specifically support the business context. Characteristics of firm-internal business data are listed in Figure 1.1 (right box). Business processes that stem from the pre-digital era have most likely

been supported by IS at some point. New business models were directly set up with respective IS support, for example, as an e-commerce platform. The data resulting from both types of business processes are well-aligned with the business goals and stored in databases with structured formats. This makes the inclusion of internal business data in analytics relatively easy. Besides the well-structured internal business data, we are currently witnessing the emergence of several ▷*ambient data* sources. These datasets are by-products of pursued business activities and often not essential for the fulfillment of contracts. Nevertheless, they have high relevance for data-driven innovations when new insights are to be generated from data. Examples for ambient data are high-frequent transaction data (e.g., ▷smart meter data in the utility industry, payment transaction data, music streams listened, or purchase data from Internet of Things devices like the Amazon dash button[3]), internal data on business processes (e.g., modification logs of files or database records), or communication data (e.g., call-center notes, e-mails).

Ambient data are often by-products business activities

I describe three reasons why organizations collect such ambient data below, together with examples. This list of reasons is most likely not exhaustive:

Reasons why firms collect ambient data

- Firms *start to collect data* in the context of their existing product portfolio *to develop additional services or new business models*. Home appliances, for instance, are increasingly equipped with an internet connection and come with smartphone apps that allow controlling the devices remotely, or connect them into the smart home environment. Data-logging in industrial machinery or vehicles helps to collect data for predictive maintenance and services around physical products. With additional data, heating manufacturers

[3] The Amazon dash button is a small WiFi connected device to quickly order products. It can be, for instance, placed next to the washing machine to re-order washing powder. With such a device, orders are recorded not only when the customer remembers to buy a consumer good, but also when the good is consumed. This timestamp may contain much information on the customers' behaviors and living situations, considering the ordered good.

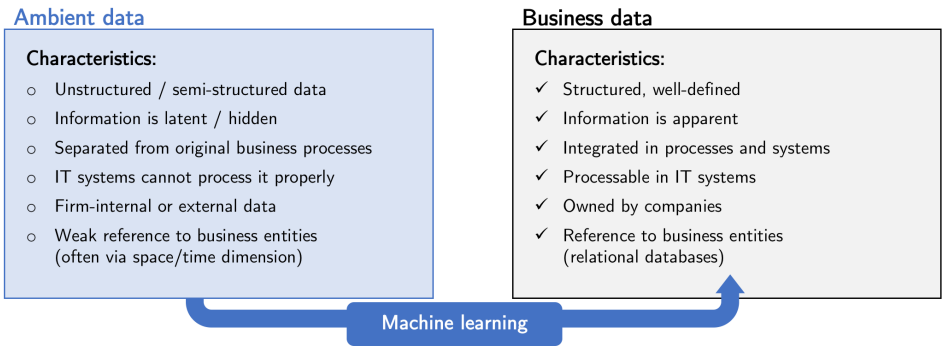


Figure 1.1: Characteristics of ambient data and classical business data; the goal of this work is to investigate how machine learning can be used to harness ambient data in business applications

are not only able to sell a heating installation, but the service of having a warm home. Truck manufacturers can not only sell lorries, but the guarantee of goods being transported from one location to another.

- Data have already been incurring through business processes and can be used for *different purposes*. Communication data are stored in mail servers, communication logs (e.g., in messengers, e-mails, call-center notes), and transaction data exist in several databases (e.g., time stamps of order placements or payment transactions).

- Organizations are mandated to collect data because of *legal requirements*. This is, for instance, the case in the utility industry where smart meters for electricity must be installed. Another example is the requirement for workers to "clock in" that became necessary because of minimum wage law.

The use of data that firms are obligated to collect, or have collected for various reasons in the past is often dedicated to a purpose. In the case of personal data, the consent of the data subject must be collected to meet data protection regulations. This consent should be feasible to obtain when the data analysis also benefits the data subject.

Ambient data also include public or open data sources

In addition to ambient data accumulated *within* firms, there are multiple origins of such publicly available data. Governments, for example, publish public sector information (e.g., weather data, public statistics, satellite images). In addition, many websites exist that contain publicly accessible content and users create web portals with large amounts of crowd-sourced data—often having geographic locations associated. These data are known as Volunteered Geographic Information (VGI) and offer information on various subjects, even for some on which data was never collected before (Goodchild 2007; Sester et al. 2014). Geographic data contain information on the living conditions, gentrification, etc. Examples for such VGI initiatives are OpenStreetMap[4], Geocaching[5], and Runtastic[6].

Open and public data have often a geographic reference

Ambient data come with opportunities and challenges

Ambient data offer great opportunities for firms to create new data-driven innovations. Nevertheless, they have several characteristics that make them distinct from (classical) business data. These characteristics are illustrated in Figure 1.1. First, the information is represented in unstructured or weakly structured formats (e.g., text, log messages, time series data). When it is stored in a structured format, the representation is often insufficient for the use in contexts that are different from those for which the data were initially collected. Second, the data contain *latent information*, for instance, on the behavior, living conditions and socio-demographic characteristics of customers. This makes the data highly interesting. Third, it contains "noise", irrelevant or missing data points, requiring advanced data processing techniques to prepare the data for further analyses. Fourth, the data often have only a weak reference to business entities. This means that the geographic location must be used to connect public data to customer entries in corporate databases, or date and time must be used to connect calendar data or environmental observations. In contrast, data concerning customers can be attributed via a unique customer number.

The data contain latent information on behavior, living conditions, etc.

[4] Free world map that is editable by everybody, available at `http://www.opentreetmap.org`.

[5] Outdoor game where players seek and hide containers at different geographic locations using Global Positioning System (GPS) devices.

[6] GPS fitness tracking app that allows to upload and ahare running tracks.

1.2 Research goal: Value creation from ambient data through machine learning

This work supports the so-far rather conceptual research on big-data value-creation with empirical findings

Several research efforts on how the value from data can be realized have already been undertaken in IS research, as Günther et al. (2017) observe in a comprehensive literature review. The authors come—in accordance with Markus (2017)—to the conclusion that research on this topic has been mostly conceptual, so far. They call for more empirical research regarding the value-creation from data. I follow this call in my dissertation and explore, *how machine learning can be used to harness ambient data in business applications and thus, create new insights from data*. These insights can be used to realize value from datasets.

Data-driven decision making process as outline of this dissertation

The complex process to create value from data was conceptualized and divided into several stages. Sharma et al. (2014) consider the *data to insight*, the *insight to decision*, and the *decision to value* stage. In other works, this resulting process is also called "information value chain" (Abbasi et al. 2016; Koutsoukis and Mitra 2003). Thiess and Müller (2018) argue that the "data-driven decison making" process should start with a question and expand the information value chain by the stage *question to data*. Using this process model (illustrated in the top part of Figure 1.2), the complex issue to investigate the value creation from data can be broken down in a research agenda.

Five Research Question (RQ)

I formulate five Research Questions (RQs), adhering to this process, and use the stages as a structure for my dissertation. The RQs are listed in Table 1.1 to give an overview to the work. A detailed motivation of each RQ, the underlying challenge and their relation to literature are described in the introductory part of the respective chapters as described below.

Chapter 3 focuses on available data sources for business analytics and answers RQ 1 in its first part. In the second part, preparation of raw data to usable data points (features) is described and advantages as well as pitfalls of automatic and theory-driven feature extraction is compared (RQ 2). Chapter 5 describes available datasets for both case studies and answers RQ 3. The last

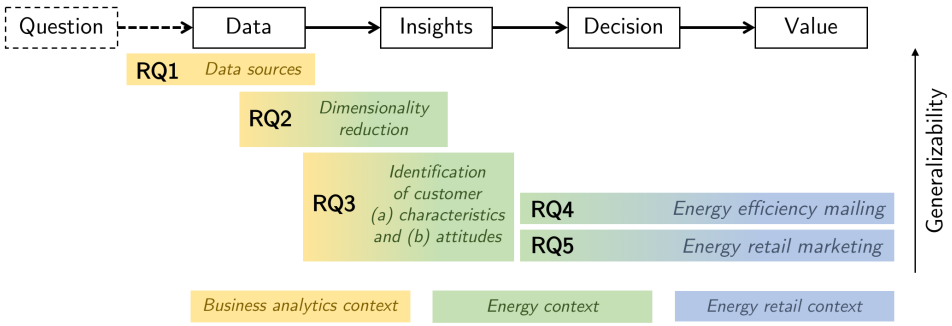


Figure 1.2: Research questions (RQs) covered in this dissertation and their positioning in the data-driven decision making process

two application-related RQs are answered in chapter 6 and 7 respectively. The contents of these and the remaining chapters are summarized in section 1.4.

Figure 1.2 illustrates to which step of the data-driven decision making process the RQs belong, and to which context the findings mainly contribute. The core results of my research for each step of the data-driven decision making process are summarized in section 8.1 on page 193. The following section introduces the empirical context of energy efficiency and energy retailing, and describes two case studies that are the base of the empirical investigation pursued in this thesis.

Table 1.1: Research questions and chapters at a glance

Research question (RQ)		Chapter
RQ1	*Which data sources are considered in predictive analytics IS research studies, which typically exist in firms or are publicly available, and what are characteristics of the data sources?*	3
RQ2	*Does theory, expert knowledge and human cognition notably help to reduce data dimensionality, although several computational methods exist for this task?*	3
RQ3	*How well can (a) customer characteristics and (b) intentions be revealed from ambient data, in the context of energy retail, using state-of-the-art ML methods?*	5
RQ4	*Which added value can be realized from predicted customer characteristics on the example of personalized energy feedback?*	6
RQ5	*Which added value can be realized from predicted customer intentions on the example of relationship marketing?*	7

1.3 Empirical research within the the context of energy efficiency and energy retailing

Case studies from energy retailing in Germany and Switzerland

The research conducted in this dissertation uses energy retailers in Switzerland and Germany as empirical context. This established industry is as relevant as ever, given that it provides power to other companies and individuals, without which many activities of our modern society would not be viable. Energy retailers posses manifold ambient data sources (like electricity smart meter data), and can use public data (like weather data or geographic information) together with their owned data.

Green IS and energy informatics research

Through the energy context of this dissertation, I position my research in the tradition of *Green IS*, following Melville (2010) and R. T. Watson, Boudreau, et al. (2010), as well as in the field of *energy informatics* (Goebel et al. 2014; Gholami et al. 2016; Ketter et al. 2018). Thereby, I investigate how IS can help to obtain new insights on energy consumers that can be

used to promote energy conservation in the residential sector, and to disseminate sustainable energy technology among private customers (through the identification of market potentials).

In this section, I first give an overview to the electricity market, outline current challenges in electricity retail and describe how data-driven innovation can help firms to cope with the challenges. Second, I introduce the two case studies that are used to answer RQ 4 in chapter 6 and RQ 5 in chapter 7.

1.3.1 Electricity retail market: Current challenges and opportunities

Electricity is a commodity of our everyday life, but it is a good that can not be stored, like food or other tangible goods. In fact, the generation of electricity and its consumption must be synchronized. Multiple players ensure that electricity—meaning both: power (measured in Watt) and amount of energy (power over time, usually measured in kWh)—is delivered across the production and supply chain to private customers.

Depending on the sales region, utilities act in monopoly and liberalized markets

In *monopoly* market settings (e.g., Switzerland currently), one regional player controls the whole centralized generation of electricity (e.g., in coal-fired, nuclear or water power plants), the distribution of electricity via the grid network, and the delivery to industry or private customers with all electricity procurement and billing. In *liberalized* markets (such as the EU, US, Australia, Japan and Singapore) grid operation is separated from electricity trading. This enables competition between utility companies. The process of deregulating the electricity market has taken place in the mentioned countries from 1990 on. It led to the fact that many utility companies that possessed a monopoly market position before needed to split up into grid operation and energy trading firms. Private customers were finally able to switch their electricity supplier without changing the physical connection to the grid, as it is for example the case in some telecommunication industries worldwide.

Severe challenges exist in both markets

Both market settings are affected these days by groundbreaking market transitions, changing the way energy will be produced and used in the future. These market transitions in the electric

utility industry become visible considering the energy production, grid operation and energy retailers perspective, whereas I focus here on the energy retailers' perspective. Energy retailers in Europe operate in a market that is not growing, but competitive. The share of wallet for housing, water and energy remains nearly constant during the last 20 years at 20-25% (Eurostat 2017). This market cannot be expected to rise strongly in the future, as energy efficiency will be increased and thus less energy is demanded. Moreover, high churn rates in the energy retailing industry exist, for example in Germany 6.4% in 2015 (BNetzA 2016, p. 184), or in Norway 12.7% in 2014 (NordREG 2017). In countries with a market that is still characterized by regional monopolies, upcoming market liberalizations is a specter for utility companies (e.g., in Switzerland the market liberalization was postponed multiple times) and may obviously cause radical changes in their business (Markard and Truffer 2006). In addition, utilities are mandated to implement energy efficiency programs and regulations in the residential sector. The policies range from incentives for efficient behavior and subsidies for energy-efficient construction (BMUB 2014) to stronger regulatory instruments, like decoupling.[7] Along with this development, the European Energy Efficiency Directive forces energy retailers to achieve 1.5% energy savings per year through the implementation of efficiency measures in the residential sector (EU 2012).

High competition on EU's energy retailing market

Energy efficiency regulations reinforces burdens on energy suppliers

Besides the mentioned challenges that utility companies are still aware of, there are some new players potentially entering the electricity retail market, that have the capability to fundamentally change the game for many utility companies and call the existence of classical energy retailers in question. For example, Drift[8], a community-based start-up providing sustainable energy

New players in the energy market emerge

[7] Decoupling is an instrument of utility regulation (in place in some U.S. states), where a utility's profits are disassociated from energy commodity sales. Instead, the returns are aligned with meeting previously defined revenue targets and are adjusted at the end of a previously defined period (Lazar et al. 2016; Eto et al. 1997). With this approach, policy makers try to enforce, for example, energy efficiency among energy consumers, or distributed energy generation.

[8] `https://www.joindrift.com/`, last accessed 10.09.2018

to end-customers in New York City. It uses algorithmic trading of electric energy between producers and end-consumers. In 2017, this platform traded 517,140kWh and saved their customers, by their own account, approximately $26,000 (which equals to 5 ct per kWh). Such data and community driven organization can make traditional energy retailers obsolete. Additionally, a couple of smart home energy management systems have been developed. The most prominent example is the smart thermostat "Nest", a start-up recently acquired by Google. The smart thermostat controls the heating by learning from the behavior of its user. Substantially integrated in a smart energy grid, this may have significant influence on the way energy is consumed in private homes. Owed to the demand of residential customers for autarky, some experiments are made with micro-grids combining local energy producers and consumers to self-sufficient units that may have only a connection to the national grid as a fallback-option. Finally, carmakers, governments, and firms are searching for business models to implement electric vehicle charging infrastructure. This may also have a significant impact on the electricity market, since joint-ventures of these three parties are likely to become substantial players. A few energy providers are experimenting with new business models. Some offer energy consulting, thermography services, services around electric mobility, products in the smart home market, or fibre-optic internet access. Others enter the market for photovoltaics or heat pumps. These examples are initiatives of some energy suppliers. No superior strategy has yet emerged.

Online platforms ans smart home vendors enter the energy market

Established utilities experiment with new business models

For all these new offers, firms rely on detailed knowledge about customers, as known household details for residential customers, for example, help to realize new services (e.g., energy consulting) or to advertise new products to the relevant target audience. I further illustrate the necessity for such customer insight in the two case study descriptions below.

Detailed customer knowledge is essential for new energy products and services

This dissertation shows how ambient data in utility companies can be combined with ML to obtain such detailed customer knowledge. To investigate the nexus of ML and ambient data in a real-world setting, two case studies from electricity retailing are investigated in detail. Both are motivated and described below.

1.3.2 Case 1: Scalable energy efficiency campaigns

Consumption feedback as a successful mass-market energy efficiency measure

Energy efficiency campaigns are, as argued above, relevant activities of energy retailers to fulfill legal mandates or enter new business areas. Energy feedback was demonstrated to be an effective measure for reducing the residential energy demand (Allcott 2011; Tiefenbeck, Wörner, et al. 2018). Feedback is thereby an instrument that is more cost-effective than price-incentives and leads to better public acceptance than prohibitive regulations (Allcott and Mullainathan 2010). Thereby, more specific feedback leads to higher savings (Vassileva et al. 2012). A major limitation for scalable energy feedback campaigns is, however, the sparsity of information available on residential customers (Tiefenbeck 2017). Household characteristics of energy consumers (e.g., household type and size, type of heating, age of appliances) are therefore demanded to realize energy feedback campaigns on scale.

ML can reveal the information for targeted feedback from ambient data

In the field of energy efficiency, Beckel and colleagues (2014; 2013) showed that it is possible to predict household characteristics (e.g., number of residents, living space area, employment of the residents) based on 30-minute electricity consumption data using ML. This finding was the starting point of my dissertation research, and I investigated how applicable this approach is for several real-world datasets of different time resolution. The detailed analysis and results are presented in chapter 5, the use of the revealed household data in a personalized energy feedback intervention is demonstrated in chapter 6.

Chapters 5 and 6 continue this case

The case study demonstrates, how well characteristics of residential households can be predicted based on ambient data (i.e., smart meter electricity consumption data, weather observations, geographic information) using various ML methods. Furthermore, predicted characteristics are used in an exemplary field study to personalize home energy reports and display highly targeted energy feedback to private energy consumers.

1.3.3 Case 2: Relationship marketing in energy retailing

Deep customer knowledge is necessary for relationship marketing

Firms try to establish better relationships with their customers in order to maximize the customer lifetime value (Brassington and Pettitt 2006). This demands adequate knowledge about their customers, which goes beyond the typical data stored for order fulfillment and invoicing. Such information can be *general characteristics*, like the living situation (single vs. cohabiting, size and type of housing, etc.) the employment status, or general household characteristics (e.g., type of heating, age of the house, ownership of the house). These socioeconomic variables help to better communicate with the customer and increase the service-level, but also help to better value the customer, for example with the customer lifetime value approach (Kumar 2018). Besides, *attitudes of customers* are highly relevant for firms, as they express the interest of them to purchase products and services.

ML can help to overcome expensive customer data collection or data purchase

Data that companies store internally are often unreliable, incomplete, outdated or redundant (Alshawi et al. 2011; Reid and Catterall 2005). Several strategies have been tried by firms to obtain the desired information, but all of them have limitations. On the one hand, the purchase of personal data from data brokers can be expensive, because data of customers on an address-level can account for $5 - 30\%$ of the budget of a typical mailing campaign (Hopf, Riechel, et al. 2017). On the other hand, collection of such data through customer-loyalty programs, lotteries, or customer surveys suffer from low response rates and can thus cause bias in the data (Groves 2006). Obtaining the desired information from ambient data through the use of ML is therefore a promising alternative that I elaborate in chapter 5. Moreover, I demonstrate the benefits of the ambient data processing through ML in the case of supporting a cross-selling marketing campaign in chapter 7 and thereby answer RQ 5.

Chapters 5 and 7 continue this case

The case study shows how well customer intentions can be revealed from ambient data (i.e., smart meter electricity consumption data, weather observations, geographic information) using ML. The application of resulting predictions in the form of cross-selling scores is illustrated among the marketing case

of identifying customers intending to purchase a Fiber-to-the-Home (FTTH).

1.4 Structure of this work and earlier publications

During my dissertation research, I have investigated how available data can be transformed into insights and finally be used in measures that increase economic or societal value. The dissertation is therefore structured along the *data-driven decision making process* (Abbasi et al. 2016; Sharma et al. 2014; Koutsoukis and Mitra 2003; Thiess and Müller 2018). I illustrate this structure in Figure 1.3.

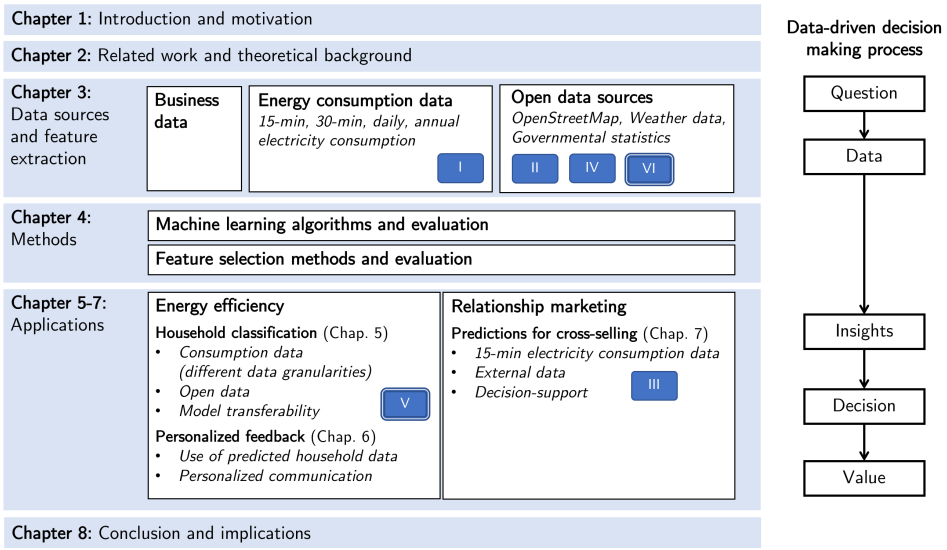


Figure 1.3: Structure of this dissertation and related publications along the data-driven decision making process

This dissertation consists of research results that I obtained through my investigation and that have not been published in their present form yet. The illustration shows how ancillary publications, that I authored or co-authored, relate to this dissertation and the research topic. The referenced papers in Figure 1.3 (numbers I to VI) are as follows:

I. Hopf, K., Sodenkamp, M., Kozlovskiy, I., Staake, T. (2014). *Feature extraction and filtering for household classification based on smart electricity meter data.* 3. D-A-CH+ Energieinformatik Konferenz 2014, Zurich, Switzerland, November 13–14. In: Computer Science–Research and Development 31 (3), pp. 141–148, DOI: 10.1007/s00450-014-0294-4

II. Hopf, K., Sodenkamp, M., Kozlovskiy, I. (2016). *Energy Data Analytics for Improved Residential Service Quality and Energy Efficiency.* 24. European Conference on Information Systems (ECIS), Research-in-Progress, Istanbul: Turkey, June 12–15

III. Kozlovskiy, I., Sodenkamp, M., Hopf, K., Staake, T. (2016). *Energy Informatics for Environmental, Economic and Societal Sustainability: A Case of the Large-Scale Detection of Households with Old Heating Systems.* 24. European Conference on Information Systems (ECIS), Istanbul: Turkey, June 12–15

IV. Hopf, K., Riechel, S., Sodenkamp, M., Staake, T. (2017). *Predictive Customer Data Analytics—The Value of Public Statistical Data and the Geographic Model Transferability.* 38. International Conference on Information Systems (ICIS), Seoul: South Korea, December 10–13

V. Hopf, K., Sodenkamp, M., Staake, T. (2018). *Smart Meter Data Analytics for Enhanced Energy Efficiency in the Residential Sector.* Electronic Markets, 28 (4), DOI: 10.1007/s12525-018-0290-9

VI. Hopf, K. (2018). *Mining Volunteered Geographic Information for Predictive Energy Data Analytics.* Energy Informatics, 1:4, DOI: 10.1186/s42162-018-0009-3

The aim of this work is not to substantially improve algorithms or to develop new algorithms. Rather, the aim is to investigate how existing data analysis methods can be applied and combined with ambient data from the energy retail industry in order to create value-added services (i.e., energy efficiency measures) or support of energy retail marketing.

A main focus of the work lies on the demonstration that the developed methods can be translated into practice. This is achieved by implementing all analyses in the statistical standard environment GNU R[9]. This platform is freely available for everyone. R

[9] R is a free software environment for statistical computing and data science (see `https://www.r-project.org/`, last accessed 18.11.2018)

also has a close integration into major software products, like Microsoft's enterprise products, or SAP's in-memory database technology. The dissertation also has the aim to be a useful guide for data analysts in organizations, data scientists and students, who want to have a case-related insight into business intelligence and analytics, or are looking for exemplary solutions in similar applications.

Subsequently, I briefly summarize the contents of the different chapters of this dissertation. The practical problem, the research gap, as well as the core contribution of each chapter is outlined.

Chapter 2: Related work and theoretical background I resume related works in the field of IS, relationship marketing, and energy informatics that are related to my dissertation. I identify research gaps in the three areas and position this work in IS literature on value creation from data. This lays the theoretical background for the deduction of RQs—which is done in the following chapters.

Chapter 3: Data sources and feature extraction In the first part of the chapter, I give an overview to data sources typically available in organizations and thereby answer RQ 1. A systematic literature analysis in the eight major IS research journals is conducted that serves, together with case studies conducted during my dissertation research, as the base to develop a taxonomy of data sources in organizations.

In the second part of the chapter, I introduce two fundamental approaches to prepare data for further analyses and thereby reduce its dimensionality: Empirical ▷feature extraction and automatic ▷feature selection. The question, which of these two approaches is better, conveys the debate about an advantage of human knowledge in the context of ML to the task of data preparation. I first present examples of empirically defined features (from electricity consumption data, weather data, geographic data from OpenStreetMap, and governmental statistical data) and describe 43 automatic methods for selecting features (available in R). Finally, RQ2 is answered argumentatively.

Chapter 4: Machine learning methods and model evaluation An introduction to the main approaches of supervised ML is given. I describe six algorithms and the evaluation of them in detail. Additionally, I explain the foundations of machine learning model evaluation and discuss several classification performance metrics. The contents of this chapter serve as an introduction to the topic and as a base for the following chapters.

Chapter 5: Household classification approach The main contribution of this chapter is to thoroughly evaluate how predictive analytics can predict individual characteristics and intentions of residential energy customers. Thus, RQ 3 is answered. Furthermore, I present additional analyses to finally answer RQ 2. In this chapter, the foundations are laid to further investigate the two case studies in the following chapters.

The chapter starts with a detailed description of the datasets that are the base for the research presented in this dissertation. Dependent variables are defined and descriptive statistics are shown.

In the following, several ML models are evaluated. First, I present the results achievable with 15-minute data and different classification algorithms. Second, I compare 43 feature selection methods using the systematic benchmark methodology proposed before. Third, I evaluate how well households characteristics can be predicted based on daily and annual electricity consumption data together with external data sources. Finally, the transferability of the prediction models between Ireland and Switzerland, as well as between Switzerland and Germany is positively tested. This analysis bolsters the reliability of investigated models and the research conducted.

I move beyond the state-of-the-art in the area of household classification based on energy consumption data by adding additional data, thereby strongly increasing the classification performance and enabling the recognition of household details with data of low data resolution (daily and annual consumption intervals). Furthermore, I show that it is not only possible to recognize

household details from the consumption data, but also interests and attitudes.

Chapter 6: Personalized home energy reports for user engagement and residential energy efficiency I demonstrate how the predicted knowledge about households can be used to realize highly personalized energy reports in the first case study. Such digital services increase customer satisfaction and energy efficiency in the residential sector.

The benefit of the predicted data was examined in a two-phase experiment with 414 energy customers from Switzerland. In the first phase, an energy report without predicted data was sent and the reaction of the recipients was analyzed with regard to the use of digital services, customer satisfaction and energy demand. The second phase involved adding a more personalized element to the e-mail and testing the benefits of the predicted data. With this latter experiment, the value of predicted household data in the field of energy efficiency is demonstrated and thus RQ 4 is answered.

The contribution of this case study to the theory is that the effectiveness of personalized electricity consumption feedback via e-mail was demonstrated in a field study. Such energy saving campaigns also have a positive effect on customer satisfaction, as the survey of customers who received the feedback showed. The case study shows the potential of such measures and motivates future research to confirm these effects in larger experiments. Finally, the example application of predictive analytics results exemplifies how environmental and societal value can be realized from data sources available to firms.

Chapter 7: Supporting cross-selling marketing The application of the predictive analytics approach in relationship marketing is showcased in the second case study. Hereby, the revealed attitudes of customers through the ML based predictive analytics approach are used in the concrete application of identifying customers intending to purchase a cross-selling product (i.e., Fiber-

to-the-Home internet access). Thereby, RQ 5 is answered and the value of the approach for relationship marketing is demonstrated.

This application shows how the predictive analytics approach presented in this thesis can solve two important business problems related to relationship marketing. First, existing customers can be selected based on the potential of purchasing a cross-selling product or service. A score can thus be calculated for each customer, which can specifically support sales. Second, the result of the prediction supports managerial decision making. Scoring can be used to determine the cost-benefit optimal number of customers to be addressed in an advertising campaign.

This case study contributes to theory by showing that customer intentions can be extracted from ambient data and how these data-driven insights can be used to support managerial decision making on an operational and strategic level. This showcase is a contribution to the insight to decision stage of the data-driven decision making process.

Chapter 8: Conclusion Finally, I summarize the findings of my research in chapter 8, discuss the limitations and name implications for research as well as practice.

Appendices and glossary There are four appendices. Appendix A describes the pursued systematic literature review in detail. Appendix B gives an overview to the seven case studies conducted through my research so far. The cases are the foundation for this dissertation as well as the related publications. Appendix C contains survey instruments to measure personal attitudes and intentions that were used in the data collection. Finally, I compiled a glossary with important terms and their definition used in this dissertation. Terms in the glossary are highlighted with a triangle (▷) at their first occurrences.

2 Related work and theoretical background

This dissertation deals with the problem of how companies can make use of ambient data sources (e.g., energy consumption data, communication data or public data) in business analysis to realize value. The research follows the IS research tradition and, more specifically, the recent calls to investigate how value can be created from big data (Sharma et al. 2014; Abbasi et al. 2016; Günther et al. 2017; Markus 2017). The first section of this chapter summarizes therefore current research on value creation from big data. The second section gives an overview to previous research on the characterization of residential households based on energy consumption data, as my research is carried out in the field of energy retailing. The idea of extracting household characteristics from electricity customers' consumption time series with the help of ML—which was the starting point of my research—stems from that field, more specifically from the seminal work of Beckel, Sadamori, and Santini (2013). Energy informatics research is connected to IS, and several scholars have outlined that more research should be conducted regarding topics of energy and sustainability (R. T. Watson, Boudreau, et al. 2010; Melville 2010; Ketter et al. 2018). I respond to these calls and support the relevant energy industry to benefit from digitization.

Customer relationship management and the promotion of products is the core topic of the marketing and the business administration field. I give an overview to the research on relationship marketing in the final section of this chapter, and describe the research gaps that are related to my work.

2.1 Value creation from big data

Through ongoing digitalization in commercial and private contexts, more and more data are recorded. In the energy domain, for example, electricity consumption is measured by smart meters in 15- or 30-minute intervals and communi-

cated to grid operators or energy suppliers. Smart thermostats[1] record activity at home to detect presence of residents, can combine it with weather forecasts and adjust the heating. With smart phones, data on movements, daily activity, and the personal health status is captured, stored, and maybe never deleted.

The existence of big data within organizations is well recognized and conceptualized from IS research (Constantiou and Kallinikos 2015; Yoo 2015; Kallinikos and Constantiou 2015; Günther et al. 2017; Abbasi et al. 2016). It is acknowledged that advanced data processing techniques, skilled personnel, and adequate organizational culture is required to enable data-driven decision making (McAfee and Brynjolfsson 2012). Regarding the nature of big data, previous works have identified its sources, possible applications, and associate big data with the following characteristics:

- *Volume*: The most obvious characteristic of big data is the large amount of entries in databases and data sources
- *Variety*: Structured, semi-structured, and unstructured data from various data sources including transactional data, user-generated data, images, sensor data, web content, spatial-temporal data
- *Velocity*: Much data change over time, are time-dependent or must be processed in real-time (e.g., sensor-data, user-generated content)
- *Veracity*: Not all data are trustworthy or they contain noise, especially web content may be affected by wrong information or fraudulent data

The research on big data has a strong focus on the opportunities that big data provides for organizations. As it is mostly conceptual, Günther et al. (2017) and Markus (2017) call for more empirical research on the realization of value from big data. It also focuses mostly on data that are available internally in organizations.

A number of additional characteristics have been suggested to describe the concept of big data. "Value" is often cited in ths context. Value is, however, not a unique characteristic of big data, because other data have also value, and the actual gained value from big data strongly depends on the context and the analysis pursued. In addition, it is the final goal of firms to increase economic value, or the goal of other organizations or governments to increase societal value by exploiting data (Günther et al. 2017). The definition of big data with

[1] Examples are tado (`https://www.tado.com/de/`, last accessed 14.11.2018) or Nest (`https://nest.com/de/`, last accessed 14.11.2018)

the help of the four or more V's has recently reached an exaggerated extent, as an article explaining 10 V's of big data (Khan et al. 2018) illustrates. The addition of further keywords starting with V does not bring any more clarity in how to use the data. Furthermore, considering the mentioned characteristics, a dataset that has a low volume can even be considered as big data, given that the other characteristics apply. For example, a few hundred customer reviews need little storage space, but still require advanced analysis methods to be evaluated automatically.

Ihe term "big data" is therefore misleading, because some of the named characteristics apply to various datasets that have not a large volume. I introduced the concept of *ambient data* in the previous section describing those kinds of information within and outside of organizations that firms can make sense of.

2.1.1 Data-driven decision making process

Value does not automatically evolve from data (Sharma et al. 2014). There is rather a complex process necessary in which many actors are involved, in order to gain insight from raw data and to derive knowledge. This knowledge can consequently be used to make better decisions, which can ultimately lead to added value in organizations (or in a private context). The information value chain or *data-driven decision making process* was already introduced in the previous chapter, since this dissertation is structured along this process.

The process is described by a number of recent authors in IS research (Abbasi et al. 2016; Sharma et al. 2014; Thiess and Müller 2018). The process model was also subject of investigation in earlier research on the value creation through knowledge-based or business intelligence systems (Koutsoukis and Mitra 2003; Chiang et al. 2012; Negash 2004).

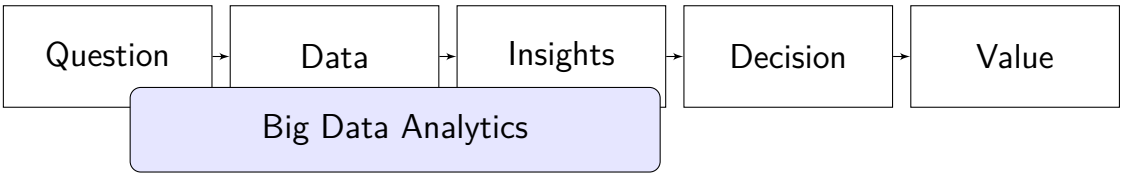


Figure 2.1: Big data analytics and the data-driven decision making process

In contrast to the studies on big data that are rather conceptual than empirical (Günther et al. 2017; Markus 2017), there is a stream of research investigating the methods to realize value from big data explicitly. Müller et al. (2016) describe *big data analytics* as "statistical modeling of large, diverse, and dynamic data sets of user-generated content and digital traces." The methods enable manifold

ways to process and make sense of data. They cover a variety of approaches (including algorithms and best practices) that allow firms to detect new patterns of available data and gain insights to improve decision making. The big data analytics methods are usually categorized into:

- *Descriptive analytics*: analyzing the past to explain what has happened,
- *Diagnostic analytics*: analyzing why something happened in the past,
- *Predictive analytics*: building models from historic data to predict future outcomes,
- *Prescriptive analytics*: development of recommendations for future actions.

LaValle et al. (2011) and H. J. Watson (2014) name only descriptive, predictive, and prescriptive analytics as areas of big data analytics. There is also a stream of methods using models primarily developed to predict, but are used to explain causalities (Shmueli 2010). The difference between descriptive and diagnostic analysis is that the first reveals correlations as well as potential pattern, and the latter one investigates causality as well as relations between variables. In that sense, descriptive analysis is used to develop hypotheses (e.g., using descriptive statistics, clustering, or frequent pattern analyses), and diagnostic analysis is used to verify them (e.g., using inferential statistics). The method tool set of big data analytics is very rich and supports the definition and refinement of the initial question, and the discovery of relevant data sources (from question to data) as well as the insight generation process (from data to insight).

A couple of challenges are, notwithstanding, associated with the analysis of big data with ML algorithms. The massive sample size and high dimensionality of big data introduce computational and statistical challenges that require special care. Fan et al. (2014) name algorithm scalability, noise in the data, spurious correlation, incidental endogeneity and measurement errors that can lead to biased models. L'Heureux et al. (2017) give an overview to such issues and resume the current state-of-the-art on dealing with the challenges. Nevertheless, Müller et al. (2016) point out that when researchers consider big data analytics, they "might have to grow comfortable with the idea that research can start with data or data-driven discoveries, rather than with theory". This work considers selected issues big data analytics challenges and investigates aspects empirically: feature extraction, feature selection, and algorithm transferability.

Having derived insights from data, it does not mean that value is created automatically from. On the contrary, the gap between insights to value is significant, given that the insights must be used, for example, to make more informed

decisions, and the decisions must be implemented right (Sharma et al. 2014). This part of the data value creation process is much harder to examine, as managerial decision making is a complex field in which, for example, cognitive biases and simplistic heuristics are involved (Tversky and Kahneman 1974). Shollo and Galliers (2016) have shown that value is created from business intelligence systems through organizational knowing. It can be assumed that big data analytics methods, have similar or even more contribution to the daily business of managers, but also in private contexts. This dissertation uses two case studies to examine the insight to value gap in the data-driven decison making process and provides empirical results to the discourse on the value creation from data.

2.1.2 Predictive analytics in information systems research

▷Predictive analytics is the activity of applying "statistical models and other empirical methods that are aimed at creating empirical predictions (as opposed to predictions that follow from theory only), as well as methods for assessing the quality of those predictions in practice (i.e., predictive power)" (Shmueli and Koppius 2011). Some years ago, Shmueli and Koppius (2011) found that the number of predictive studies in the IS field was low. Since then, several predictive analytics studies have been published in the IS discipline so far.

To obtain an overview to studies on predictive analytics in IS research, I conducted a literature search in the Association for Information Systems (AIS) basket of top journals[2] in September 2018, querying databases that contain a full index of publications in the respective journals. I used the following list of search terms: "predictive (analytics OR analysis OR modeling OR modelling OR power OR value)", "forecasting", "detection", "statistical learning", "machine learning", "neural network". The number of the resulting hits to the search over time is shown in Figure 2.2. The figure contains several duplicated hits, resulting from papers that were found with multiple search terms. In the last ten years $(2009-2018)$, studies related to the topic increased strongly. However, the total number is still low compared, considering the 55 hits for the years 2014–2018 and the fact that 2,094 articles were published in the journals in that time[3].

[2]The list contains: European Journal of Information Systems (EJIS), Information Systems Journal (ISJ), Information Systems Research (ISR), Journal of the Association for Information Systems (JAIS), Journal of Information Technology (JIT), Journal of Management Information Systems (JMIS), Journal of Strategic Information Systems (JSIS), Management Information Systems Quarterly (MISQ)

[3]The total number of articles is the number entries in the respective databases in the years 2014–2018, queried on 14.02.2019.

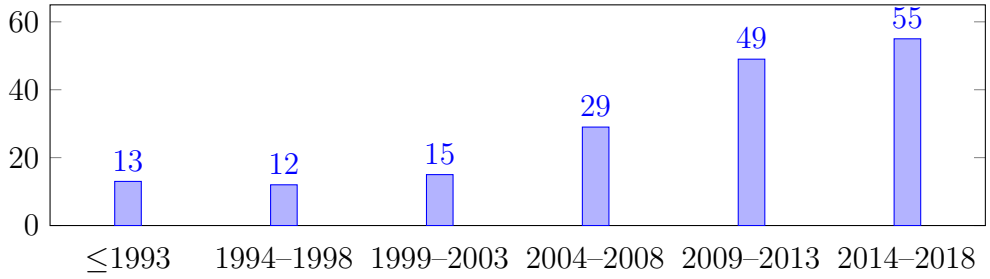


Figure 2.2: Total number of hits for the search terms related to predictive analytics (duplicate mentions possible, as papers were found through multiple search terms)

After having filtered the unique publications, I pursued a content analysis of the resulting 155 unique articles (considering publication title, abstract, and keywords), which I describe in detail Appendix A, and categorized the papers into one of the following categories:

1. Empirical predictive analytics studies, including predictive IS prototypes when the creation and evaluation was the main focus of the work ($n = 56$)

2. Conceptual studies, reviews that discuss predictive analytics, or big data analytics ($n = 7$)

3. Other studies, including literature reviews and editorials, explorative or conformative data analysis investigating phenomena or theories, forecasting of future market or business developments, the use or the value of intelligent systems, research method papers, without focus on predictive analytics ($n = 92$)

From this systematic literature review, I summarize that predictive analytics in the energy domain was not yet in the focus of publications in the leading IS journals, even after the call for more research in this field by R. T. Watson, Boudreau, et al. (2010) and Melville (2010). Besides, a variety of algorithms and methods are used in the works, but the selection of algorithms is often ambiguous, evaluation does not comply with standards from machine learning, and evaluation metrics are not coherently used.

Research on predictive analytics, as well as research on big data analytics in companies, is still in its beginning stage and the number of empirical research studies is low. In my work, I investigate those methods, that provide data-driven decision support, in the context of energy retail, which has not yet been done so far.

2.1.3 Frameworks and process models for big data analytics

Several frameworks and process models for big data analytics and data mining have been proposed so far. As a first attempt to identify key activities for data mining, Fayyad et al. (1996) describe the knowledge discovery in databases process "for extracting useful knowledge from volumes of data" as a sequence of data selection, preprocessing, transformation, data mining, interpretation, and evaluation. Activities in this process are understood to be "repeated in multiple iterations" to "identifying valid, novel, potentially useful, and ultimately understandable patterns in data." The goal of this process is knowledge creation, reuse of models is not explicitly envisaged.

From a joint industry and research effort, the CRISP-DM ("Cross Industry Standard Process for Data Mining) framework was developed (Chapman et al. 2000; Shearer 2000). The framework has emphasis on embedding data analytics in the business context and developing models including the main phases "business understanding", "data understanding", "data preparation", "modeling", "evaluation", and "deployment". The main goal of this process is the development of models that can be further used and deployed in productive environments.

In their work on predictive analytics in information systems research, Shmueli and Koppius (2011) present a list of necessary activities to develop predictive models from a research perspective. Their understanding of the process on model building and evaluation is linear, which does not represent the reality in data analytics, even though the iterative nature of data modeling generation was already postulated earlier. Through the variety of problems, in an actual instance of data analytics, the transition between phases or the activities are somewhat overlapping. For example, the phase of data preparation and modeling is intermingled when human knowledge is implemented in algorithms to better preprocess data for improving models, or when additional variables are considered, because of the experience of a data scientist or manager.

Figure 2.3 shows the three frameworks (or process models) in comparison. All of them have a large overlap in activities and cover the key steps. The KDD process and CRISP-DM foresee that activities can be repeated, which Shmueli and Koppius (2011) do not explicitly indicate.

These frameworks can provide aid (e.g., to data scientists, or managers) making first steps in new data analytics problem classes. For concrete data analysis problems (like segmentation in marketing, or new predictive analytics cases), they are too abstract and provide only marginally guidance to data scientists. In the same vein, Venter et al. (2015) point out that so far "academic researchers

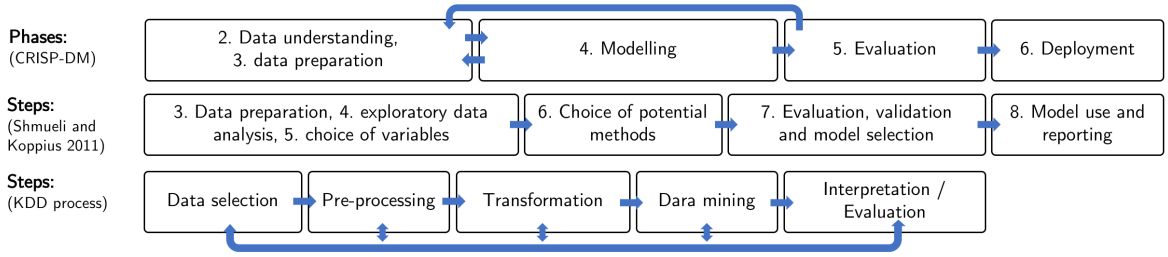


Figure 2.3: Frameworks and process models for data analytics

have focused on theoretical and technical issues [...], practitioners must tackle practical and pragmatic implementation problems."

2.1.4 Research gaps in information systems research

Several IS scholars have recently highlighted the need for empirical research on the value-creation from big data (Abbasi et al. 2016; Sharma et al. 2014; Günther et al. 2017; Markus 2017). Although a number of insightful studies exist that show how predictive analytics can be used to solve relevant business problems, the predictive analytics literature mainly considers firm-internal data that are stored in convenient formats, easy to analyze. Ambient data sources, such as customer communication data, geographic information, and data from smart devices, were hardly used in any IS works, so far. My work fills this gap and presents empirical research on predictive analytics in the energy retailing industry.

I operationalize the data-driven decision making process and answer five tangible RQs by means of datasets stemming from real companies. In doing so, I obtained new theoretical insights and practical knowledge on the effective use of ML in business analytics.

In detail, my work regarding RQ 1 helps to narrow the research gap on value creation from big data, as there is currently no systematic overview of which data silos exist inside firms and outside of organizations. I suggest a taxonomy of data sources that are available to organizations. RQ 2 addresses the problem of transforming raw data into a form that can be further processed by algorithms. I illustrate two fundamental approaches to deal with the problem in chapter 3 where the first is based on human knowledge end theory and the second relies on automatic methods. Besides that, my research helps to obtain an understanding of how predictive analytics can be used. The proposed approach, described in detail and evaluated in chapter 5 as my answer to the RQ 3, is a method tool-set

for analytics in energy retail. It may also serve a blueprint for other industries with strong customer focus and shows how value can be generated from ambient data through advanced business analytics.

With the practical and theoretical findings, the research is more specific than general process models for data analysis, like the knowledge discovery in databases process (Fayyad et al. 1996), the CRISP-DM process model (Chapman et al. 2000; Shearer 2000), or the predictive analytics process of Shmueli and Koppius (2011).

Finally, this work is—to the best of my knowledge—the fist comprehensive investigation of predictive analytics in energy retailing, following several successful approaches of predictive analytics in IS in other industries. By means of case studies from the energy retailing industry, my work also helps the energy industry getting more attention from an IS research perspective.

2.2 Household characterization based on electricity consumption data

The application domain of my research is energy retailing with the focus on residential customers. This section therefore resumes the related research streams on the use of electricity consumption data to characterize private utility customers. I also identify limitations of existing works and point to the contributions of this thesis to the field of energy informatics.

Existing research that analyzes electricity consumption with the aim to find out details on electricity customers is comprehensive. I structure the related studies into four streams of research: 1) the identification of influence factors to residential electricity demand, 2) the disaggregation of high-resolution electric consumption data, known as NILM, 3) clustering (unsupervised ML) of electricity customers, and 4) classification (supervised ML) of households among predefined household classes based on data that are available to energy utility companies.

The research streams 1–3 brought advancements for domain experts. Although this research is comprehensive, concrete aid for scalable energy efficiency campaigns that can be conducted in mass-market settings and marketing efforts is limited. For this objective, only the fourth approach, household classification, provides such aid.

This dissertation continues therefore the research on household classification. Beyond the further development of predictive models, and the test, if new household characteristics and intentions of customers can be revealed from the data

(as it is the focus of RQ 3), I apply the results in two case studies. Thus, in RQ 4, I investigate how predicted household characteristics can be used in the form of tailored energy consumption feedback. Finally, the analysis of RQ 5 shows how household classification can add value to relationship marketing.

2.2.1 Factors influencing residential electricity consumption

The characteristics of households and residents' behavior have a strong impact on the household electricity demand. McLoughlin et al. (2012) show this correlation for households in Ireland, and Kavousian et al. (2013) for US (Californian) households. In addition, geographic information (Heiple and Sailor 2008), weather, and climate are shown to have a strong impact on the energy use in residential buildings (Apadula et al. 2012; Druckman and Jackson 2008; Hernández et al. 2012). These works show that causalities between environmental conditions and the electricity consumption of households can be found in data. This helps to explain, why predictive models for household are feasible.

2.2.2 Non-Intrusive Load Monitoring (NILM)

NILM refers to the analysis of high-frequency meter data (often with more than one measurement per second) that tries to disaggregate the electric load curves to identify single appliances in households. Research goes back to Hart (1992) and applies various ML algorithms (e.g. neural networks, decision-tree learners) to detect refrigerators, coffee machines, microwave ovens, and other household appliances. Even the difference between ordinary light bulbs and energy efficient lamps can be recognized (Birt et al. 2012; Zeifman and Roth 2011). While the high sampling rates necessary for NILM are common in industrial or lab settings without transmission of data outside local networks, the smart metering infrastructure currently being rolled out does not provide such fine-grained data (the current standard is 15- or 30-minute). In addition, the data processing of such detailed data will only be relevant in local settings because of privacy reasons (Quinn 2009; Enev et al. 2011; Greveler et al. 2012; Armel et al. 2013; H. Kim et al. 2011).

2.2.3 Clustering of energy customers

Based on data from off-the-shelf smart meters data with 15-minute or less data resolution, unsupervised ML (clustering) algorithms are applied to identify groups of customers with similar electricity consumption pattern (A. Albert and Rajagopal 2014; Al-Otaibi et al. 2016; Gianfranco Chicco 2012; Silva et al.

2011; Flath et al. 2012; Kwac et al. 2013; McLoughlin et al. 2012; Figueiredo et al. 2005; Sánchez et al. 2009). The resulting customer segments are hard to interpret, and the interpretation relies on expert as well as domain knowledge. Therefore, the contribution of these analysis results to energy efficiency or marketing campaigns is limited.

2.2.4 Household classification

Several works investigate supervised ML (classification) algorithms to predict the characteristics of household from electricity consumption data. The works use consumption measurements where information on household characteristics is known, for example, from surveys conducted. Fei et al. (2013) detect heat pumps from daily Smart Meter Data (SMD) and weather data in a study with 6,385 US customers. Verma et al. (2015) detect the existence of electric vehicles among 1,250 US households with hourly SMD.

A. Albert and Rajagopal (2013) apply Hidden Markov models to raw 30-minute SMD and predict the existence of 10 household properties in a study with 1,100 US households, however, with a high prediction error. Beckel, Sadamori, and Santini (2013), Beckel, Sadamori, Staake, et al. (2014), and Beckel (2015) suggest reducing the raw load curve of 30-minute data to a set of 22–35 features and use five classification algorithms[4] to predict 18 household characteristics by means of a smart meter dataset with 4,232 Irish households. The same dataset is used by Wang et al. (2018) who achieved better results using convolutional neural networks.

I describe the limitations and open research directions in the context of the household classification works below and describe the contribution of my work towards these research gaps.

2.2.5 Research gaps in household classification

In this work, especially in chapter 5 (RQ 3), I replicate and significantly extend the work of Beckel and colleagues (2013; 2014; 2015), raise data in Switzerland and Germany with several household characteristics that have not been collected in this context so far. I extend the data processing within the approach, test different consumption data resolutions (15-minute, daily, and annual electricity consumption data), additional data sources, and the geographic transferability

[4]Beckel (2015) uses linear discriminant snalysis, Mahalanobis distance classifier, Support Vector Machine, k Nearest Neighbors, and AdaBoost; the latter three are also used in this work for replication and are described in section 4.1.

between Switzerland, Germany, and Ireland. I list the extensions to the approach made and published in earlier studies, as well those advanced documented in this thesis below.

Additional algorithms and predictor variables The prediction accuracy can be improved through empirical feature definition and extraction based on the 30-minute data. We extended the feature set from 22 to 88 and made first experiments with feature filtering in Hopf, Sodenkamp, Kozlovskiy, and Staake (2014). The feature extraction was transferred to 15-minute SMD, and we added features on the correlation of five weather variables (temperature, sky cover, precipitation, wind speed, air pressure) with the electricity consumption to the classification in Hopf, Sodenkamp, and Staake (2018). This advanced the existing works that had only considered the outside temperature. In chapter 5, I include geographic data into the classification models that further improves the prediction performance.

Geographic scope of the studies The existing studies have only been carried out in North America and Ireland. The empirical validation of the approach in Central Europe is beneficial, as it underlined the stability of the predictions. In addition, it worth knowing whether models trained in one geographical region would also produce meaningful results in another. In this way, the stability and reliability of the models can be demonstrated, as models that are transferable to other regions are unlikely to be over-fitted to datasets.

Transferability tests of predictive models have been conducted in several other application fields. For example in geography (Vanreusel et al. 2007; Wenger and Olden 2012), recreation demand (Loomis et al. 1995), forecasting of travel demand for transportation planning (Everett 2009; Sikder et al. 2013), or accident prediction (Sawalha and Sayed 2006).

Limited number of household characteristics So far, only general household characteristics that are directly reflected in electricity consumption (e.g. electric car, number of persons in the household, size of dwelling, employment) were investigated. On the other hand, it would also be helpful for energy-saving campaigns to have characteristics related to heating. So far, only heat pumps have been identified here. For marketing purposes, it would be very interesting for firms to extract not only household characteristics but also the attitudes or interests of residents from the electricity consumption sensor data.

Selected time resolutions The previous works test 30- and 60-minute, as well as daily time resolutions with a different set of household characteristics. From the existing results, no proposition can be made which household characteristic can be predicted on which time resolution. The question is relevant, considering that SMD is not available a majority of regions. On the one hand, The market in central Europe, for example, lacks a sufficient amount of ˜smart meters[5] In addition, SMD may not be usable for direct marketing purpose because of privacy or data governance reasons. In the European legal environment, for example, energy consultants, or retailers need to have the consent of the data subject (i.e., a customer) to process its personal data as long as it is not necessary to process the data to fulfill a contract. The feasibility of prediction should therefore also be investigated for lower data granularities, for example, with annual electricity consumption data that is available to all energy retailers for billing purposes.

The prediction of household characteristics using *annual electricity consumption data* is therefore a relevent question. Together with geographic data, we tested this successfully in Hopf, Sodenkamp, and Kozlovskiy (2016) and Hopf (2018) and found that it is feasible to predict household characteristics with strong impact on the energy demand (e.g., living space area, number of residents, household type, type of space, and water heating). We suggested also a method to test the transferability of such predictive models and investigated the added value from governmental statistical data (Hopf, Riechel, et al. 2017).

Missing field studies on the contribution to energy efficiency Previous studies on household classification aim to support energy efficiency campaigns, automated energy consulting and marketing. To the best of my knowledge, no study has yet tested the predicted household details in energy efficiency campaigns and only Fei et al. (2013) test their prediction in a marketing campaign. In an experiment with Swiss residential energy customers, I tested this successfully and describe the results in chapter 6 (RQ 4). In chapter 7 (RQ 5), I demonstrate the usage of predicted purchase intention scores in a cross-selling case.

[5] Einhellig et al. (09.07.14) analyzed the market in Germany and came to the conclusion that the amount of smart meters in 2030 will be only 27.1 %, if current legislation status does not change.

2.3 Relationship marketing

A core aspect of energy retailing is the advertisement of products that utility companies offer, and the the management of customer relations. Consequently, my research has a considerable intersection with the research area of relationship marketing.

Traditional understanding of marketing focuses on the view that seller and buyer transactions (exchange of goods or services) are totally discrete and lacking any of the personal and emotional overtones established in long-term relationships between actors in markets (Brassington and Pettitt 2006). Following this understanding, firms use market segmentation to describe groups of similar customers with adjectives or personas (e.g., "social responsible", "innovative", "ecological", "political") and specialize products or services for each segment. This is the first step in the segmentation-targeting-positioning process of strategic marketing (Varadarajan 2010). For each segment, offers are developed and advertised. Thereby, firms try to attract specific segments with special offers varying product features, pricing, and quality to realize profits from hyperdifferentiation (Clemons, Gu, et al. 2003; Clemons, Gao, et al. 2006).

The offers based on the segmentation results are promoted with different marketing instruments (Alaimo and Kallinikos 2017; Weinstein 2013). It is assumed that customers in the segments will act in the desired way an advertisement is designed and placed. The response rates of typical promotional campaigns are problematic to measure and have rather a long-term than a short-term effect (Dekimpe and Hanssens 1995).

A more contemporary understanding of marketing includes the focus on *buyer-seller relationships* (Dwyer et al. 1987) in *relationship marketing*. Based on the understanding how parties regard each other (Thibaut and Kelley 1959), Jap and Ganesan (2000) describe four phases of a relationship: exploration, buildup, maturity, and decline. The traditional marketing view mainly focuses on the exploration and buildup phase, uses advertisement, or other promotional activities. All customers (including existing ones) are treated as new ones. Relationship marketing focuses on intensifying the relationship and tries to keep customers loyal to the company.

In this section, I inform the reader about the fundamentals of relationship marketing, customer segmentation, and the gap between buying intention and buying behavior. The two case studies which will be considered later in chapters 6 and 7 are both related to customer interaction and therefore have to do with relationship marketing. RQ 5 in particular addresses the question of how the

predictive analytics methods considered in this thesis can be used for cross-selling marketing campaigns.

2.3.1 Information systems for customer relationship management

The systematical storage and management of information on customers has a long tradition in IS research and practice. Synnott (1978), for example, documented the development and use of a "Total Customer Relationship system" in the banking sector in an early IS publication. S. H. Kim and Mukhopadhyay (2010) differentiate two types of Customer Relationship Management (CRM) systems:

- *Support-related CRM*: Act as direct support for front-line employees, store and manage data for providing customized service; the systems are also known as "front-office" or "operational" CRM.
- *Targeting-related CRM*: Support relationship marketing by analyzing customers' preferences and purchasing behaviors; the systems are known as "analytical", "strategic", or "back-office" CRM).

CRM systems have the primary scope to store known data on customers and persist them together with transaction and interaction data. The stored data may serve for customer prioritization or the support of marketing campaigns, but Padmanabhan and Tuzhilin (2003) advice to be careful with data from CRM systems, because the stored data might be wrong or unreliable. Other studies have documented further data quality issues in CRM systems (Reid and Catterall 2005; Alshawi et al. 2011).

Studies in IS have investigated different aspects of CRM systems in an organizational context. One stream of research investigates effective CRM systems design and implementation (Gefen and Ridings 2002; S. H. Kim and Mukhopadhyay 2010; Ward et al. 2005; Xu and Walton 2005). This research is supported by studies on the user satisfaction of front-line employees with CRM that leads to a better perceived service quality by customers (Hsieh et al. 2012). Another stream of research focuses on influence of CRM systems on firm performance (Coltman 2007; Coltman et al. 2011; Karimi et al. 2001; Zablah et al. 2012). A third stream of research focuses on how customer centric websites must be built to create satisfied customers and improve customer relationships (T. Albert and Goes 2004; Lee et al. 2003). Moreover, personalized customer communication based on known customer information is a successful measure to enhance customer loyalty (Otim and Grover 2006; Zhang et al. 2011).

Some empirical studies motivate the use of predictive analytics in CRM systems, for example by obtaining the likelihood a customer purchases an offer (G. Cui, Wong, and Wan 2012) or by modeling the responses for marketing campaigns (G. Cui, Wong, and Lui 2006). Advanced models, for example by incorporating sensor data (e.g., electricity smart meter data) and open big data to predict socioeconomic variables on customers, as I do in my dissertation research was, to the best of my knowledge, not yet considered in the CRM literature.

2.3.2 Predictive segmentation and customer scoring

From a methodological point of view, segmentation in marketing can be divided into descriptive and predictive segmentation approaches. *Descriptive segmentation* techniques are used to identify within-group similarities or between-groups dissimilarities (Banasiewicz 2013, p. 188ff). This approach has three major limitations. First, identified pattern or customer segments need interpretation from experts and are therefore highly subjective and arbitrary, given that different analysts would interpret segments differently and may come to other solutions. Second, the data used in segmentation studies usually not available in common databases at scale (Turner et al. 2013). Relevant demographic and socioeconomic data on customers must be collected in an earlier customer contact (and stored in CRM systems). Third, individual segment memberships of customer can only be given for customers with sufficient available data and not for all customers. This holds especially for the case of anonymously conducted experiments, or survey panels, where no relation between the segmentation result and customer records is possible.

Predictive segmentation aims at obtaining future events (e.g., mailing response, purchase event, churn event, bankruptcy of a borrower) that are most likely to occur. The approach is also known as scoring and results in "an ordered lists of probabilities that customers act in an assumed way that contributes revenues" (Wachtel and Otter 2013). For that means, a predictive segmentation model is built to estimate individual-record-level probabilities of a certain outcome, such as product repurchase or promotional response. A variety of modeling techniques are used for customer scoring, ranging from regression to more advanced approaches, such as Bayesian Networks with evolutionary programming (G. Cui, Wong, and Lui 2006), artificial neural networks, genetic algorithms (Y. Kim et al. 2005), and general machine learning algorithms (D. Cui and Curry 2005). In the field of IS research, Martens et al. (2016) describe a data analytics method for predictive segmentation, obtaining the purchase likelihood score for financial products using payment transaction data.

Scoring overcomes some above-mentioned limitations of descriptive segmentation, but the approach has other drawbacks. First, there is hardly any method that can convert purchase intention scores into a purchase *probability*. Buying behavior deviates from purchase intention, because of several biases involved (Sun and Morwitz 2010). Factors related to the way of offering (Venkatesh and Agarwal 2006) also influence purchase decisions. Second, it is also commonly agreed that scores for customer alone do not matter. It is often unclear which concrete guidance the scores contain. The likelihood that a customer will switch the energy supplier, for example, can have several reasons (relocation, unsatisfied with the service, too many advertisements, campaign of a competitor, ...). In this case it is possible that a customer terminates his contract if the company contacts him, because he is annoyed by too much advertising. To overcome the problem of hard to interpret scores, one can try to infer preferences of customers using post-purchase data (Jagabathula and Vulcano 2017) or usage data of an online product configurator (Huang and Luo 2016). The mentioned studies describe case-specific methods that are hardly transferable to other industries. Further investigation is therefore necessary to generalize findings for a broader community. Finally, the customer groups formed by a predictive model cannot be described easily. The variables that are used to describe a segment often differ from those used in predictive segmentation (Y. Liu et al. 2010, p. 881).

2.3.3 Purchase intention and behavior

Consumer buying behavior has attracted much attention among the IS researchers, for example, in the context of e-commerce (Jiang et al. 2010; van Der Maaten et al. 2009) and mobile commerce (Lu and Su 2009). This research has typically an empirical nature, related to the technology acceptance models, theory of planned behavior, and flow theory (Gefen, Karahanna, et al. 2003; George 2004; Koufaris 2002). The typical objective of these studies is to design user interfaces for intensifying buying behavior. Reported intentions—asked for instance in a survey—do not necessarily respond to actual behavior of humans (Fishbein and Ajzen 1975; Brown and Venkatesh 2005; Venkatesh, Thong, et al. 2012). Consequently, purchase intention scores, estimated by a statistical model, cannot be easily be converted to purchase probabilities for individual customers.

Stated purchase intentions give an indication towards the actual buying behavior, but this relation is influenced by several determinants (Morwitz et al. 2007). Sun and Morwitz (2010) propose a statistical model to convert stated intentions to purchase probabilities and validate it in several industries. In com-

bination with ML and predictive analytics techniques can be used to predict purchase intention scores.

2.3.4 Research gaps in relationship marketing research

Relationship marketing is a state-of-the-art approach that prioritizes customer relationships over individual transactions and attempts to increase customer value. To build enduring relationships, detailed customer knowledge is necessary. Internal information systems and databases often contain only incomplete information. For the relationship-building measures, details about the customer are required that have not previously been collected. The need for detailed customer information is therefore considerable.

In chapter 5 (RQ 3), I show how information about customers can be extracted from a large amount of internal and external data that, if used correctly, can increase service quality and customer value. Among the examples of energy efficiency mailings, I demonstrate in chapter 6 (RQ 4) how relationship building can lead to increased customer satisfaction and increased usage of a customer-engagement online portal.

Segmentation approaches are important for marketing in order to divide customers into groups and then address them according to their specific needs. Predictive segmentation approaches use customer scoring to rank existing customers according to importance (e.g, purchase likelihood, churn risk). The scoring approaches are usually based on data stored in firm internal information systems and databases. External data is, as far as I know, only considered occasionally in scoring models. The extensive integration of such data is constrained by the data formats in which external data or sensor data are available and the fact that necessary information is hidden in the data.

In chapter 7 (RQ 5), I showcase how smart meter electricity consumption data and geographic information can be used to obtain purchase intention of individual customers. The case study illustrates how the household classification approach can support relationship marketing on different levels of managerial decision making.

3 Data sources in organizations and extraction of predictor variables

Highlights

- ▷ Firm internal and external data sources are summarized from predictive analytics studies in eight leading IS journals. The developed *taxonomy of data sources* for business analytics was extended through seven industry case studies.
- ▷ *Empirical feature extraction* is introduced as an approach to prepare data using the nexus of human cognition, theory, and expert knowledge.
- ▷ The approach is demonstrated among eight examples using data sources that are typically present at energy retailing firms: Transaction data, environmental data, geographic information, and governmental statistics.
- ▷ An overview to *automatic feature selection* techniques is given.

Data stocks in organizations are rapidly growing[1] driven by the ongoing digitalization in private and industrial environments. Besides that, massive amount of data are publicly available, for example, open government data (also referred to as ▷open data). It stems from public sector activities that collect, process, and store a broad range of data from numerous fields like taxation, social security, geography, weather observations, patent administration, and education. Governments have begun to publish such data in order to meet political demands for openness and transparency,[2] and firms can use this data in their business an-

[1] According to the EMC (2014) Digital Universe Study, the data that are created and copied each year will reach 44 zettabytes and is growing by 40% every year.

[2] The EU (2003) decided to launch open data initiatives like the "INSPIRE" project (EU 2007), a geographical and environmental data infrastructure, or the "Copernicus" project (EU 2013) which publishes data from the European Space Agency for Earth Observation.

alytics. A second example of public data are user-generated, or crowd-sourced, data that have geographic references associated. Goodchild (2007) calls this kind of data Volunteered Geographic Information (VGI). The data resulting from the VGI phenomenon are freely available and cover a wide variety of fields, even data on subjects that have never been collected before (Sester et al. 2014). In Hopf (2018), I have investigated VGI and demonstrated that this data are promising sources for insights in energy data analytics.

To describe these and other heaps of data that are emerging, practitioners and researchers coined the term "big data". Meanwhile, it is acknowledged that *volume* is not the only characteristic of big data. Yoo (2015) underlines that "a simplistic quantitative approach to big data in fact diminishes its strategic importance." Research ascribes three other characteristics to the phenomenon (LaValle et al. 2011; Constantiou and Kallinikos 2015; Abbasi et al. 2016). First, *variety* specifies that different data types are available (particularly unstructured or weakly structured data types, such as textual descriptions or images). Second, *velocity* describes the high frequency in which data may change (e.g., newly created social media content, Web pages can be modified at every time and are not stable over time). Third, *veracity* characterizes the unclear trustworthiness of data (e.g., fake news or fake reviews on social network sites).

Consequently, even small datasets with numerous attributes together with additional data from various sources can be described as big data. The term "big data" is therefore misleading and I introduced the concept of *ambient data* in section 1.1, which characterizes data that are available to firms (stemming from internal IS or from external sources) and are at the same time by-products of core business activities. The examples of ambient data that stem from firm's internal sources (e.g., customer communication data, energy consumption data) as well as the examples of external data sources (e.g., open government data, VGI data) give an idea of what promising data are available to organizations.

Data—regardless which term is used for their description—have significant business value: The World Economic Forum (2011) praises personal data as a "new asset class" and the global political agendas concern about the use of personal data (European Commission 2017; OECD 2017). To realize the value from data, a complex process of insight generation from data is necessary. This process leads to knowledge that can be turned into decisions that finally can create economic, ecologic, or societal value. The *data-driven decision making process*, introduced in subsection 2.1.1, starts with the phases *question to data* and *data to insights*.

The US and UK governments have also launched open data initiatives (Immonen et al. 2014).

This chapter deals with these initial phases and answers two important RQs in this context, both are subsequently motivated by resuming existing literature and highlighting the research gap. The second section answers the first RQ and gives an overview to data sources in organizations, that were mentioned in IS research studies or used during my investigation. I present this overview in the form of a taxonomy of data sources. This result supports the *question to data* stage, as it helps to categorize existing datasets and identify new data sources for analytics. In the third and fourth section, two fundamentally different approaches to prepare the data sources for further analysis (empirical feature extraction and automatic feature selection) are presented. Based on the exposition of both approaches, I give an argumentative answer to the second RQ in the last section from a theoretical point of view (further empirical investigation regarding RQ 2 is presented in section 5.4). This concluding section also gives a summary of this chapter and names implications for research as well as practice.

3.1 Theoretical background and research questions

Entering the data-driven decision making process and defining a specific problem to be solved with analytics (*question to data* phase) soon raises the question which data sources are available to the respective organization or team of analysts. In the phase *data to insight*, the first task is to prepare data and define suitable predictor variables. Here, a debate is ongoing about whether algorithms alone can solve this problem, or whether human knowledge is needed. These mentioned issues in each phase of the data-driven decision making process are further explained below and specific RQs are derived from the literature review.

3.1.1 Need for a systematic overview to available data sources

Research has conceptualized the general characteristics of big data (LaValle et al. 2011; Constantiou and Kallinikos 2015; Abbasi et al. 2016) with the “four V’s” (volume, variety, velocity, veracity). This characterization is helpful when it comes to data processing, because suitable methods can be chosen to respond to the special properties of the data. The dictum of describing big data with terms beginning with the letter ”V” has now taken an odd turn when looking at publications that continue to add words to this definition, like Khan et al. (2018) proposing “The 10 Vs [...] of Big Data”. For the identification of possible data sources in the beginning of the insight-generation process, those descriptions only

have a limited contribution. Therefore, a systematic overview to data sources that are available to firms for analytics is necessary.

Related works are not providing a systematic overview so far, although it would support the work of data analysts. In addition, empirical research on the value-creation from big data is demanded (Günther et al. 2017; Markus 2017).

To the best of my knowledge, only the following works provide enumerations of available data sources for organizations. First, a quite obvious differentiation between internal and external data is made by Hartmann et al. (2016) who investigate data-driven business models of start-ups. Second, Kitchens et al. (2018) give an overview to available "data lakes" within organizations for "advanced customer analytics" and name *transaction data* (customers core data records, orders, product descriptions, etc.), *customer interaction data* (campaigns, marketing messages, message interaction logs, etc.), and *voice of the customers* (product recommendations, surveys, notes from call-center, etc.). This categorization provides only limited aid to analysts, as it focuses on company-internal information. Besides, it includes also external data, as product recommendations from users or social media discussions, for example, do often not occur on a company's homepage, so that the data generated here is not necessarily in the possession of the respective organization. Third, a McKinsey report (Henke et al. 2016, p. 81) lists a number of additional data sources (e.g., inputs from sales agents, call center customer notes, telecommunication customer patterns, data from government agencies, regular surveys / satisfaction data), but this collection was not systematically created. The lack of an overview to available data sources in companies most likely results from the fact that this information is difficult to collect. I therefore limit my research to predictive analytics studies published in IS literature as well as case studies that I have worked on myself. Thus, I formulate the following first research question:

RQ 1 *Which data sources are considered in predictive analytics IS research studies, which typically exist in firms or are publicly available, and what are characteristics of the data sources?*

The question is examined by developing a taxonomy of data sources in organizations available for analytics in section 3.2 that are mentioned in IS studies on predictive analytics or used during my research. To establish the taxonomy, the results of the literature survey that was mentioned in the previous chapter (section 2.1.2) are synthesized and combined with findings from industry case studies from energy retailing.

3.1.2 Algorithmic versus theory-based extraction of predictor variables (features)

Having identified available data sources for analytics, the challenge is to harness the value residing in them, as "big data without algorithm may end up being just a heap of digital dust" (Yoo 2015). Competitive advantage, which firms want to gain from data (LaValle et al. 2011), will only be realized by firms that are capable of processing the data in a meaningful way. In turn, when all firms possess huge amounts of data, only the capabilities to make use of that data matter.

The resulting practical consideration is whether firms should invest into more skilled personnel or into better algorithms. The problem becomes particularly apparent, when looking at the very beginning of data analysis: Data scientists spend up to 80% of their time on data collection, preparation, and cleaning (Bowne-Anderson 2018; Press 2016). This includes understanding the datasets, bringing the data into a suitable format, selecting relevant variables, and integrating them to one data matrix that algorithms can process. In this challenge, little has changed since the emergence of the first business intelligence applications (Negash 2004; Power 2002), although good software exists to support the ETL (Extraction, Transformation, and Load) process from operational IS into data warehouses.

This unsatisfactory situation raises the question whether automatic procedures can be used to prepare data. A recent development in the artificial intelligence community has brought forth the idea of *AutoML*, which means to automatize ML procedures almost completely. Besides the demand to reduce heavy workload of analysts for data preparation, arguments for a seamless algorithmic approach are that experienced users of algorithms waste time by re-implementing already existing procedures and non-experienced users make mistakes in that endeavor (Guyon, Bennett, et al. 2015). Moreover, selecting the best suitable ML models "should be treated with rigor as an optimization and statistics problem, not by applying haphazard heuristics." (Guyon, Bennett, et al. 2015). Furthermore, Madsen (2015, p. 11) points out that, "Despite being 'blind' in their processing of data, the [ML] approach rests on the assumption that algorithms can guide analysts to innovative analytic concepts and categorizations." Consequently, AutoML may not only reduce the workload of analysis, selects parameters more rigorously, but may also provide new insights.

On the opposite, IS as well as ML researchers underline the relevance of preparing data before advanced analytics algorithms are being applied. Guyon and Elisseeff (2006), for example, state that the "art of machine learning starts with

the design of appropriate data representations". Similarly, Kitchens et al. (2018) note that "Data management has long been considered a cornerstone of the IT function (Goodhue et al. 1992) and 'big data's rise has further amplified the importance of IT in this role' [...] no single data integration strategy is sufficient." Humans posses *knowledge* as well as *experience*, and can combine both with a variety of *cognitive skills* (e.g., perception, evaluation, comprehension, reasoning, judgment, and problem solving). These capabilities play an important role in setting the algorithms along the right course and defining the goals of analytics. This happens in particular with the definition of input variables, the to be predicted outcomes, or by writing code to execute procedures. In other words, "the analyst train[s] the algorithm" (Madsen 2015, p. 12).

The definition of predictor variables (features) using human expert knowledge, or theory, can realize also *human-in-the-loop* concepts (Schirner et al. 2013) in the context of business analytics. The ML algorithms are thus provided with theory and human knowledge in the form of algorithms or calculations. In that way, not only the dimensionality of the input data becomes smaller, but also ML models can be better explained, because single variables can be interpreted by humans.

Sharma et al. (2014) conclude, following Lycett (2013), that the question of "How can human sense making and machine learning *work together* to improve the generation of insights from the use of business analytics?" (emphasis added) is necessary to be examined in future research to support the insight generation phase of the data-driven decision making process. The question on whether "algorithmic and human-based intelligence" is the superior way of doing big data analytics is one of the six debates around the question how organizations can realize value from big data, that Günther et al. (2017) identify in their comprehensive literature review.

The discourse is fundamental and may constitute the beginning of a more general discussion on the relevance of human expert knowledge or theory that was developed so far. Jankel (2017), for example claims that "Management Theory Is Dead", because in the current "Volatile, Uncertain, Complex, Ambiguous, Networked and Stressed (VUCANS) world, management theory is no longer entirely fit for purpose. [...] The theory was designed to fit a very specific historical context" (18th to 20th Century). He further explains that there is rather the "need for constant, proactive and dynamic transformation and innovation within organizations that were designed for control, certainty and predictability is the greatest challenge that your organization faces today." Inductive, hence, algorithm-driven insight-generation in a bottom-up direction may be one solution to cope with the challenges, but the implications of the fact that algorithmic

decision making becomes more prevalent are still unclear (Newell and Marabelli 2015).

The fundamental discussion about the added value of theory, human knowledge, and cognition in the context of machine learning is wide and complex. Therefore, I limit my research to one key step of data preparation, namely the preselection of variables, also called feature selection or feature extraction. Thus, the second RQ in this thesis is formulated as follows:

RQ 2 *Does theory, expert knowledge and human cognition notably help to reduce data dimensionality, although several computational methods exist for this task?*

I examine this question by first illustrating the empirical definition of exemplary features from various ambient data sources in the context of energy retailing (utility transaction data in the form of energy and water consumption data of different measurement frequencies, geographic information, and governmental statistical data) in section 3.3. For the presented examples of empirically defined features, I explain how theory, expert knowledge, and cognition is necessary to prepare the data sufficiently. Second, I give an overview to automatic feature selection and review existing methods that are available in the statistical programming environment R in section 3.4. A first resume to answer this RQ is given in the final section 3.5 of this chapter. Further results from the application of both approaches to real-world datasets are presented in chapter 5, where I test the empirically defined features as well as the automatic feature selection approach quantitatively.

3.2 Taxonomy of data sources available for predictive business analytics

A systematic overview to data sources available for predictive analytics is, as delineated above, currently missing but would benefit the *question to data* stage of the data-driven decision making process. To close this research gap and to answer the first RQ, I developed a taxonomy of data sources mentioned in IS research studies and case studies conducted during my dissertation research.

Concretely, I started from the mentioned collection of data sources (Kitchens et al. 2018; Henke et al. 2016) and separated company-internal business data as well as external data sources, following Hartmann et al. (2016). Subsequently, I developed a first draft for a taxonomy of data sources that are used for predictive business analytics. To expand this taxonomy, a content analysis was

conducted (Drisko and Maschi 2015; Weber 1990), considering 56 predictive analytics studies in major IS journals with regard to the data sources used. The considered studies for this analysis have been selected by means of a systematic literature review that is described in Appendix A. To validate the taxonomy, it was discussed and developed further together with researchers and industry partners during the completion of predictive analytics case studies. A list of conducted case studies can be found in Appendix B.

The resulting taxonomy of data sources consists of 13 categories of possible data sources that can be considered for analytics. An overview to it is shown in Table 3.1. Internal company data can usually be linked with each other, because a unique matching criterion exists (e.g., customer number, order number). External data sources, on the other hand, usually cannot be simply linked to internal data using an unique identifier. Instead, one can use information on time or space to connect the data records. In order to highlight the respective properties of the data sources in Table 3.1, I marked all categories that usually have a time stamp with $\times$ in the column *time dimension*, and those usually contain geographic location information with $\times$ in the column *spatial dimension*. I placed the markers in brackets when information about time or space is only rarely available. A star ($\star$) in the column *empirical features* denotes that empirical features are presented for this category of data in the latter part of this chapter (section 3.3). In the following sections, each category of data sources is briefly described and limitations of this taxonomy are discussed.

3.2.1 Internal business data

Business data from IT systems within organizations, or such data that are in possession of a respective company, but may be stored by another firm or service provider belong to this category. The internal business data can be typically related to business partners of the focal firm. I categorize the internal business data into one of the following types of data sources:

Customer core data (I1) Basic information on customers that is necessary for order-processing and fulfillment. Examples are name, address, contact details like e-mail address or phone number, as well as bank account number.

Transaction and accounting data (I2) All kind of business transactions that are recorded in a more or less structured way over time. The type of data depends strongly on the industry. It may consist of data on loans and bank transactions in the financial industry, energy consumption data or such of energy installations

Table 3.1: Taxonomy of internal and external data to categorize or identify data sources for analytics

Category		Temporal dimension	Spatial dimension	Empirical features
Data sources within organizations (internal business data)				
I1	Customer core data		×	
I2	Transaction and accounting data	×		⋆
I3	Interaction data	×		
I4	Internal socio-demographic data			
External data sources				
E1	Socio-demographic data from address traders		×	
E2	Environmental data	×	×	⋆
E3	Public statistical data	(×)	×	⋆
E4	Geographic data		×	⋆
E5	Calendar events	×	×	
E6	Official publications			
E7	Website content			
E8	Electronic business platforms	(×)	(×)	
E9	Social media data	(×)	(×)	

in the energy industry, purchase data in retail, or usage data for digital services. In addition, firms have accounting data with details on transactions, billings, duns, and payment preferences of customers that I also count to this category.

Interaction data (I3) All data resulting from the information exchange between a company and its business partners and not recorded by transactional IS (e.g., e-mail communication, call-center notes, unstructured CRM system data). The data are typically unstructured. Communication data are considered as internal, as long as the firm has a copy of it (e.g., e-mails, business documents), or it is non-public communication (e.g., messenger communication, messages on Facebook to the company page), but not when the content is public and the company does not posses the data (e.g., social media content).

Internal socio-demographic data (I4) Variables that are not necessarily needed to pursue business, but are helpful in CRM or sales. Data in this category is owned by the company and was raised during the customer contact or in

personalized surveys (e.g., age, interests, earlier suppliers). Purchased marketing data are considered as external data (see below).

3.2.2 External data sources

This data stem neither from company IT systems, nor is the company owner of the data. External data are published online or purchased from data providers. The data typically cannot be directly matched to respective business partners, as a unique identifier (like a customer or supplier ID) is missing, but it can connected either using the geo-location or the time-dimension.

Socio-demographic data from address traders (E1) in comparison with the internally stored socio-demographic variables, this data stem from address traders, such as Acxiom[3], Panaddress[4], and MB Research[5]. These data are often predicted based on statistical models rather than raised in surveys. The data are available on the level of single households or persons and are different to public statistical data which are only available on a geographically aggregated level and are often marketing-related (e.g., customer is interested in soccer, or the buying power of a neighborhood).

Environmental data (E2) Weather measurements (like temperature, wind, sunshine hours, precipitation, ...), traffic data or other environmental observations (like air pollution[6], bird counts[7]). This data are often available as time series data and can therefore help to analyze internal data with a time dimension.

Public statistical data (E3) National offices of statistics and census bureaus provide public available figures about different topics of interest. With these figures, the change of states (e.g., population movements, historical economic data) can be used in data analytics tasks. The data are typically aggregated on a certain statistical geographic region (e.g., municipality, or state) and thus not directly related to single households or customers. Given that public statistical is update once a year or less often, the analysis of the time dimension is only meaningful over longer time periods.

[3] `https://www.acxiom.com/`, last access 20.09.2018
[4] `https://www.panadress.de/`, last access 20.09.2018
[5] `http://www.mb-research.de/`, last access 20.09.2018
[6] see for example the fine dust pollution measurement project luftdaten.info (`https://luftdaten.info/en/home-en/`, last accessed 01.10.2018)
[7] see for example the Christmas Bird Count project (`https://www.audubon.org/conservation/science/christmas-bird-count`, last accessed 01.10.2018

Geographic data (E4) Spatial data contain important information about environment and living conditions in one area. For example, the type of land use and near geographical objects can be used to detect the house type, and the environment in which the house or apartment is built can influence the energy consumption. There are mainly two sources of geographic information: First, official cadaster data from land surveying offices and proprietary data from companies that capture the landscape that are both available for purchase. Second, recently many VGI sources have emerged (Goodchild 2007) that can be used freely.

Calendar events (E5) Special events and holidays have a high influence on consumption pattern. Such events and public holidays (e.g., Christmas, music festivals, football games) can be included in predictive analytics for data cleaning and the improvement of models. Such data can be accessed online via event calendars and news sites.

Official publications (E6) Publications from institutions or organizations, financial statements, scientific publications, or patent data. The data are often digitally available in specialized platforms and, in contrast to website content, the contents are static over time and will not be changed.

Website content (E7) Almost every organization has a website today. This data can provide information on business partners. In contrast to electronic business platforms, this content changes less frequent, but is not static, like the content of official publications.

Electronic business platforms (E8) are related to social media data, as they contain user-generated content, but users fulfill rather business transactions than establish social relations to each other. Examples are business rating portals, trading platforms, real estate advertisement sites.

Social media data (E9) the data stem from social network platforms (like Facebook, Twitter, Youtube) and are different from electronic business platform data, as it has the social interaction as a focus rather than the business aspect. Kitchens et al. (2018) describe it as “voice of the customer”.

3.2.3 Contribution of the taxonomy

The taxonomy of data sources presented in this section provides an initial overview of available data sources for business analytics. It is based on IS studies and the experience of seven case studies from the energy retail industry, and answers RQ 1. As the literature so far lacks such a comprehensive overview to data sources, it offers an empirical contribution to a better understanding of how value can be created from data. The overview also assists future research, for example with classifying studies regarding to used data sources, or to get an overview of common data types. For each data source, typical characteristics or difficulties with the data were briefly mentioned.

From a practical point of view, the overview helps with the first step of the data-driven decision making process (*question to data* phase), because potential data sources for possible analyses can be identified. The taxonomy also supports analysts and managers, for example, in inventorying corporate data assets, or in the planning of possible analyses, as typical data properties have been described and effort of data analyses can be better estimated.

3.2.4 Limitations and future research

The presented overview to data sources has limitations and needs further investigation. First, it is most likely that data sources within firms or additional external data sources exist that have neither been mentioned in the considered IS studies nor appeared in the investigated case studies. Different firms in various industries may collect data that does not fit clearly in one of the proposed categories. It is, for example, questionable in which category data from personal health-apps or self-tracking-devices, log-files from manufacturing-machines, data from Internet of Things devices, or driving data from cars belong to. With the chosen method and data foundation to build the taxonomy, I could not identify these data sources in the literature and decided to not include them. Future research may therefore validate and extend this overview, for example, by conducting interviews with industry experts or data analysts.

Second, the categories of data sources were identified using the content analysis method (Drisko and Maschi 2015; Weber 1990) on the basis of 55 predictive analytics articles in major IS journals that were systematically selected (see subsection 2.1.2 for further details). Due to the qualitative nature of this research method, the coding of data sources, mentioned in the analyzed texts, into four internal and nine external data sources may be not conclusive. Therefore, this coding should be validated in the future, for example, by obtaining intercoder-reliability metrics.

3.3 Dimensionality reduction through empirical feature extraction

The insight generation step of the data-driven decision making process (*data to insight*) begins with the transformation of raw data into representations that are ready for further processing in computing procedures and ML algorithms. As the raw data sources stem from different origins (internal or external data, various departments, and IT systems, etc.), they are available in heterogeneous formats. Besides, the data can be inconsistent or contain wrong values. For example, various number formats are present, measurements can include erroneous values, and data inserted by humans can contain typing mistakes. Data integration and harmonization are therefore critical activities that are necessary before any data analysis can be made. In addition to the technical challenge of data preparation and the connection of multiple datasets, the high volume of raw data must be reduced to overcome the ▷curse of dimensionality which refers to the fact that performance of ML algorithms usually decrease with larger input dimensions (Guyon, André, et al. 2003; Keogh and Mueen 2011).

Feature extraction can help to overcome the technical challenge to integrate and harmonize different data sources. Moreover, it is a tool of dimensionality reduction. I conceive feature extraction as a major activity of model building, because theory and human expert knowledge can be encoded into variables and thus be made usable to algorithms. In this sense, the empirical definition of features that are calculated from raw data is a realization of the human-in-the-loop concept of ML (Schirner et al. 2013) in the context of business analytics. ML algorithms are thus provided with theory and human knowledge in the form of feature extraction procedures or calculations.

Furthermore, feature extraction can increase the ability to explain statistical models. When humans define features and label them meaningfully instead of using algorithms that produce data points—sometimes with cryptic names, statistical models are more explainable because estimated coefficients in models can be interpreted with relation to a natural dimension.

The extraction of features is a qualitative activity relying on the cognitive skills of humans together with expert knowledge and theory. The definition of features through an analyst combines, among others, the following resources:

- Human sense-making (e.g., by recognizing pattern in data from repeated observations)

- Personal experience and process knowledge (e.g., sales managers know characteristics of their customers, analysts search for explanations of extreme values)
- Domain expert knowledge (e.g., an energy consultant can interpret a "load curve" of energy consumption over time)
- Semantic relations between data (e.g., the number of dunning letters and the amount overdue have a relation)
- Theory (e.g., using landscape metrics to describe a geographic map section)

I illustrate the empirical feature extraction by means of eight examples from four data sources available to energy retailing companies in the remainder of this section. The first category is *transaction data* of utility companies (i.e., energy and water consumption data). The second is half-hourly weather observations as a representative of environmental data. The third category is *geographic data* (i.e., VGI data), and the fourth category is *government statistic* data. I explain the specialty of each exemplary data source and illustrate how feature extraction can help to obtain distinct information from the respective data using theory, expert knowledge, and human cognition. The detailed description of features also serves as a reference for following chapters.

All features and the findings described in this section stem from the case studies conducted through my dissertation research. The features have been developed together with research and industry partners while pursuing several predictive analytics case studies (see Appendix B for an overview). Additional sources of features are interviews with domain experts, as it was done by Beckel (2015), and reviews of research literature.

3.3.1 Features from utility transaction data

Transaction data in the utility industry mainly consist of energy and water consumption information for each customer. In energy retailing for residential customers, which is the context of my dissertation, utility companies usually posses consumption data with at least a yearly time resolution. Through the continuous roll-out of smart meters, energy and water consumption data are often available with time resolutions of up to 15-min, but daily measurements are also common.

Hereafter, I present features that can be extracted from different data granularities (annual, daily, and smart meter data). Additionally, I show how clustering algorithms can be used to extract features from time series data and how

the spatial location of consumers can be used to derive neighborhood-related features.

Empirical features from annual electricity consumption data

Each utility has at least annual consumption data (e.g., electricity, gas, water) from its customers. Even this small piece of information contains valuable details about customers, but must be further processed to reveal insights.

Annual consumption information usually stems from manual read-out of meters. Because of this, the time period of measurements can differ. The transactional IS therefore usually store not only the consumption information for a certain time span but also the number of days in which the consumption was created. The first data preparation objective is therefore to combine this information. The second data processing objective is to remove redundancy in the data. Namely, the fact that consecutive years of consumption data are highly correlated, as household characteristics and consumption behavior do not change much over time. The observation that consumption values of several years correlate is supported by theory, because household characteristics and peoples' living situations—the main causes of energy or water consumption—do not strongly change over time (Wells and Gubar 1966). The third objective of feature extraction from annual data is to incorporate information on other customers into variables for analysis. To realize that, it is reasonable to set the consumption of each household in relation to that of neighboring households. To pursue the three data preparation objectives, the following features from the annual consumption data can be used:

1. The normalized Consumption Per Day (CPD) for each year i:

$$CPD_i = \frac{TotalConsumption_i}{NumberDays_i} \tag{3.1}$$

   The log transformation is used to achieve a symmetric distribution of the variable, given that the distribution of energy consumption on a household level has typically a positive skew. The arithmetic mean of the normalized consumption over a number of years is used, because the consumption of different years are usually highly correlated.

2. The consumption trend as the relative change β_1 between the consumption of all n years $y_i,\ i \in 1, ..., n$, obtained with a linear regression model of the years of consumption (Hess et al. 2001):

$$CPD_i = \beta_0 + \beta_1 * y_i + \epsilon_i \tag{3.2}$$

3. The deviation of the mean (logarithm) consumption $log(CPD)$ to the mean consumption of all neighbors $log(\overline{CPD})$, expressed as a multiple of the standard deviations σ, to quantify the household's consumption deviation from it's neighborhood (the number equals to the Z-score):

$$Z = \frac{log(CPD) - log(\overline{CPD})}{\sigma} \tag{3.3}$$

   The neighborhood can either be defined using predefined geographic borders (e.g., all household within the same postal code region), or using the distances between household that can be obtained from geo-coordinates. In the latter case, one can consider to use a fixed radius (e.g., all neighbors within a radius of 1,000m) or with a fixed number of neighbors (i.e., a k nearest neighbor approach).

The features have been applied in several case studies with electricity consumption data (Hopf, Sodenkamp, and Kozlovskiy 2016; Hopf, Riechel, et al. 2017; Hopf 2018) and gas consumption data (Kozlovskiy et al. 2016). Nevertheless, the features can be also applied to water or other consumption data.

From the case of annual consumption data, the relevance of feature extraction is already visible: An automatic procedure is certainly able to identify the relationship between the annual consumption, the number of days the consumption was generated, the consumption of previous years, and the consumption of neighbors. Nevertheless, feature extraction allows to reduce these four dimensions (even more, if earlier years of consumption data are considered) into three highly expressive features from this data source.

Empirical features from electricity consumption smart meter data

A growing number of households are being equipped with smart meters due to political mandates and the renewal of electricity grids. These meters automatically read, for example, electricity consumption in 15- or 30-minute intervals and communicate the readings to the energy provider. Consumption data that are recorded in such frequent time intervals contains an extensive number of latent variables about living conditions and the behavior of the energy consumer. An exemplary electric load curve with a measurement interval of 15-minute from one household in Switzerland for one week (Monday–Sunday) is illustrated in Figure 3.1.

The high number of observations per week (672 measurements for each smart meter with 15-minute readings per week) already reveals insights about the

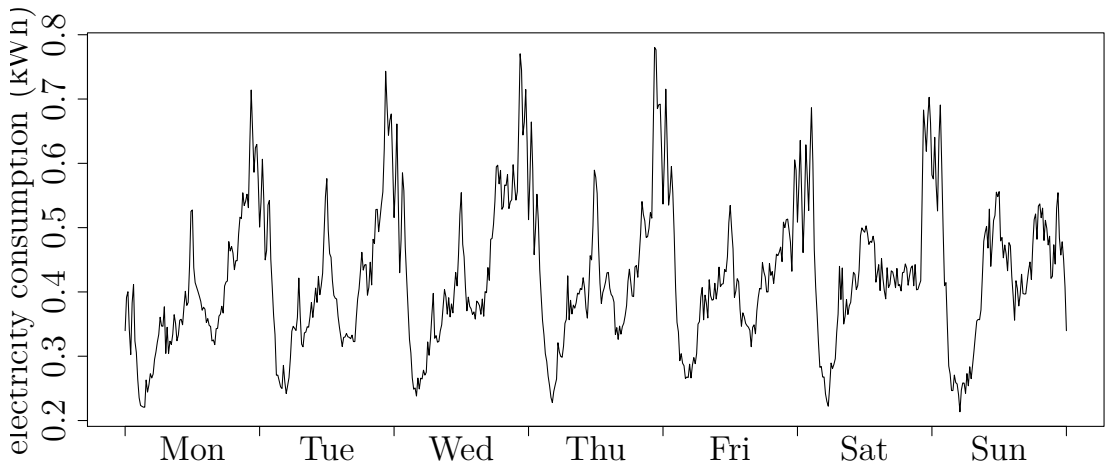


Figure 3.1: Exemplary load curve of one week (June 02–08, 2014) with 15-min smart meter data from Switzerland

energy consumer that we—as humans—can interpret from the illustration. For example: The electricity consumption in this load curve is not evenly distributed throughout the day, rather times of presence and absence of residents can be seen. The data contain redundant information, because the load profiles of weekdays are relatively similar. Peaks in the morning and evening as well as the lower consumption during the day indicate that the residents are not at home during the day. This can also be seen from the different load profiles at weekends and during the weekdays.

This primitive interpretation of a household's load curve already exposes that it is beneficial to derive empirical features from time series data and thereby reduce the high dimension of 672 measurements per week. Several works have investigated smart meter electricity consumption time series data to predict characteristics of residential customers so far and defined features for one week of electricity consumption. Based on the feature collection of Beckel, Sadamori, and Santini (2012) and incorporating other studies that suggest features for SMD (G. Chicco et al. 2001; Beckel, Sadamori, and Santini 2013; Hopf, Sodenkamp, Kozlovskiy, and Staake 2014; Beckel, Sadamori, Staake, et al. 2014; Beckel 2015; Al-Otaibi et al. 2016), a summary of electricity smart meter features is presented in Hopf, Sodenkamp, and Staake (2018).

In total, 93 features for 15-min SMD for one week have been suggested, so far. The proposed features belong to four categories:

- consumption (e.g., in the morning, noon, evening)

- ratios of consumption figures (e.g., consumption in the morning vs. noon, daytime vs. night, weekend vs. weekdays)
- statistics (e.g., variance, quantiles, auto-correlation of the time series)
- time series related figures (e.g., number and average heights of consumption peaks, time-slots with more than 0.5kWh consumption)

The number of features proposed for SMD to date, and the variety of aspects they cover, shows that there is no prescribed way for empirical feature definition. It is, in fact, a creative task that implies to reflect which characteristics of a load curve could help a computer model make better predictions and formulate the resulting thoughts into program code that feature values can be computed. With some consideration, one can certainly find additional indicators that could be calculated based on the load curve which could be also used for model building. Fear of defining too many empirical features is quite unreasonable, as in most cases it is likely to be below the dimension of the original data (which is, for the example of 15-minute SMD, 672 measurements per week and household). Conversely, there are no limits to creativity and it is unclear which feature ultimately provides a surplus in the models. Therefore, empirical feature extraction is an opportunistic approach (i.e., trial-and-error) and all features that come to one's mind through creativity and experience should be tested.

Empirical features from multiple consumption traces with daily granularity

Daily readings of energy or water consumption in households are a compromise between the deployment of smart meters, in which utility companies receive meter readings in intervals of up to 15 minutes while customers disclose detailed insights into their consumption habits, and classical annual readings, usually recorded manually. Daily consumption data, that are automatically communicated to the utility company, help to better plan resources, automate the billing process and enable comprehensive energy feedback campaigns (see chapter 6 for an example). Daily consumption data offer added value in digitalizing the commodity business of utility companies, especially when gas and water meters communicate their measurements automatically, not only electricity smart meters.

Commonly two types of daily consumption data are distinguished: one reading per day (i.e., gas and water consumption) and two readings per day (i.e., electricity consumption). Specifically, electricity consumption can be measured in single-tariff meters that record one reading per day and double-tariff meter that record the consumption during High Tariff ("Hochtarif") (HT) and Low Tariff

("Niedertaif") (NT) times. The HT time is usually during the day, NT usually during the night, but also on public holidays and during the weekend. A daily time resolution is already sufficient to detect several energy efficiency related household characteristics from electricity consumption data (Hopf, Sodenkamp, and Staake 2018).

To prepare this kind of consumption data, I followed the features defined for SMD and defined features on a daily basis covering each commodity good (e.g., electricity, gas, or water) separately. The features represent the four categories: consumption values, statistical indicators, relations, and temporal properties. For the daily data, I consider not only a weekly but also a 12 weeks time windows (quarter of a year). Table 3.2 lists all features that I defined for the available daily time series data (electricity, gas, and water).

In addition to features that cover one single consumption trace, I also defined features from the linear dependency of multiple consumption traces. The features `cdc_lmCoef_gas` and `cdc_lmCoef_wa` are calculated using multiple linear regression with the electricity consumption c_i^{el}, the gas consumption c_i^{gas} and the water consumption c_i^{water} for each available day i with measurements for all consumption meters:

$$c_i^{el} = \beta_0 + \beta_1 * c_i^{gas} + \beta_2 * c_i^{water} + \epsilon \tag{3.4}$$

Table 3.2: Features based on daily energy consumption time series data

Feature	Scope	Description
Electricity consumption (HT/NT)		
`el_cons`	week, weekday, weekend	Overall el. consumption on average during the week / weekdays / weekend (Saturday and Sunday)
`el_mean_HT`	week, weekday, weekend	Mean el. HT consumption during the week / weekdays / weekend
`el_mean_NT`	week, weekday, weekend	Mean el. NT consumption during the week / weekdays / weekend
`el_var_HT`	week, weekdays	Variance in el. HT consumption during the week / weekdays
`el_var_NT`	week, weekdays	Variance in el. NT consumption during the week / weekdays
`el_var_day`	week	Variance overall el. consumption (HT and NT)
`el_max`	week total, HT, NT	Maximum of overall / HT / NT el. consumption
`el_min`	week total, HT, NT	Minimum of overall / HT NT el. consumption
`el_r_we_wd`	week total, HT, NT	Relation mean overall el. consumption weekend / weekday (overall / HT / NT consumption)

Feature *(continued)*	Scope	Description
el_mean_max	week total, HT, NT	Relation between mean and max (overall / HT / NT consumption)
el_mean_min	week total, HT, NT	Relation between mean and min (overall / HT / NT consumption)
el_r_HT_NT	week total, HT, NT	Relation between HT and NT during the week / weekend / weekday
Multiple weeks		
el_max_avgWeek	multiple weeks	Average maximum consumption in all given weeks (if multiple weeks are given)
el_min_avgWeek	multiple weeks	Average maximum consumption in all given weeks (if multiple weeks are given)
el_ts_acf_week	multiple week	Mean auto correlation of the weeks
Gas and water consumption (one reading per day)		
gas/wa_mean	week, weekday, weekend	Mean gas / water consumption
gas/wa_var	week	Variance in gas / water consumption
gas/wa_max	week	Maximum in gas / water consumption
gas/wa_min	week	Minimum in gas / water consumption
gas/wa_r_we_wd	week	Relation mean gas / water cons weekend / weekday
gas/wa_r_mean_max	week	Relation between mean and max in gas / water consumption
gas/wa_r_mean_min	week	Relation between mean and min in gas / water consumption
cor_el_wa	week, weekdays	Correlation between overall electric and water consumption during the week / weekdays
cor_el_gas	week, weekdays	Correlation between overall electric and gas consumption during the week / weekdays
cor_wa_gas	week, weekdays	Correlation between water and gas consumption during the week / weekdays
cdc_lmCoef_gas	week	Estimated coefficient for the gas consumption in a multiple regression model (Equation 3.4) explaining the electricity consumption
cdc_lmCoef_wa	week	Estimated coefficient for the water consumption in a multiple regression model (Equation 3.4) explaining the electricity consumption.

The combination of multiple consumption traces takes account for the fact that human activity is represented in the use of all available considered commodity goods. The consumption of one commodity may cause the consumption of another (e.g., consumption of water may cause the use of electricity because people are at home and use electric water boilers). Conversely, one good can substitute another (e.g., water can be heated with an installed gas boiler or using a pot on the stove, residents can temporarily use an electric fan heater

because of a defect gas boiler). This leads, on the one hand, to partly correlated time series data and, on the other hand, to time series data that reveals new insights through combination. Redundancies in recorded data, which was visible in the one-dimensional time series data through recurring load profiles, are also present in this case, but further redundancies are added because the multiple synchronized time series that can contain duplicate observations.

The measurements of different commodity goods, however, cannot be treated completely similarly, because each consumption trace has different confounding factors. For example, water can be subject to leakage, electricity is perturbed with stand-by consumption, and gas as well as electricity can be consumed when residents are not at home (heating or appliances relies on constant energy supply). An additional practical difficulty in processing multiple consumption data time series lies in the fact that usually only a few households exist with consumption data on all commodity goods offered by an utility. The reason for this can be that households either have different suppliers for the energy sources, do not have a connection to the grid or the supplier has not yet installed a communicative meter.

All the mentioned characteristics of the multi-dimensional consumption data are hard to handle with automated methods. I tried to find features that account for the different cases and suggested features for isolated daily electricity, gas and water consumption, as well as defined some for the combination of all three consumption traces.

Feature extraction using unsupervised learning

Besides the empirically defined features where the calculation is hard-coded in program code and results from a manual interpretation of the energy consumption curve, additional features for consumption pattern can be calculated using clustering algorithms. Additional insights can thereby be generated using unsupervised ML methods.

The idea is to identify times in which the household is unoccupied in a certain time period (e.g., more than four days) and calculate features based on the duration of absence. With the proposed algorithm below, I follow the idea by Becker and Kleiminger (2017) who use unsupervised ML algorithms to detect residents' occupancy based on SMD. The functional principle is illustrated using daily electricity consumption data (similar to the data described previously). The approach can be also transferred to other data granularities by adjustment of parameters.

Due to the lack of true knowledge on occupancy times of residents that could be used to train supervised ML models, I considered visualizations of load profiles of 12 weeks and developed a k-Means clustering based algorithm being able to identify time spans that belong to low and high consumption clusters. The algorithm for identifying low and high consumption clusters is described in Figure 3.2.

Data: $Cons_t$: energy consumption of each point in time $t \in T$,
δ: margin between high and lox consumption cluster,
$nDaysCheck$: number of days of minimum absence time
Result: $Clus_t$: cluster-membership of each point in time $t \in T$
identify two clusters of consumption measurements with the centers M_1 and M_2 using k-Means;
if $|M_1 - M_2| > \delta$ **then**
  if $M_1 > M_2$ **then**
    set M_1 as the high consumption cluster, M_2 as the low consumption cluster;
  else
    set M_2 as the high consumption cluster, M_1 as the low consumption cluster;
  end
  for $t \leftarrow 0$ **to** T **do**
    change all time-windows of $[t; t + nDaysCheck]$ from low to high consumption clusters, when both clusters exist in the time-window;
  end
else
  set all days to the high consumption cluster;
end

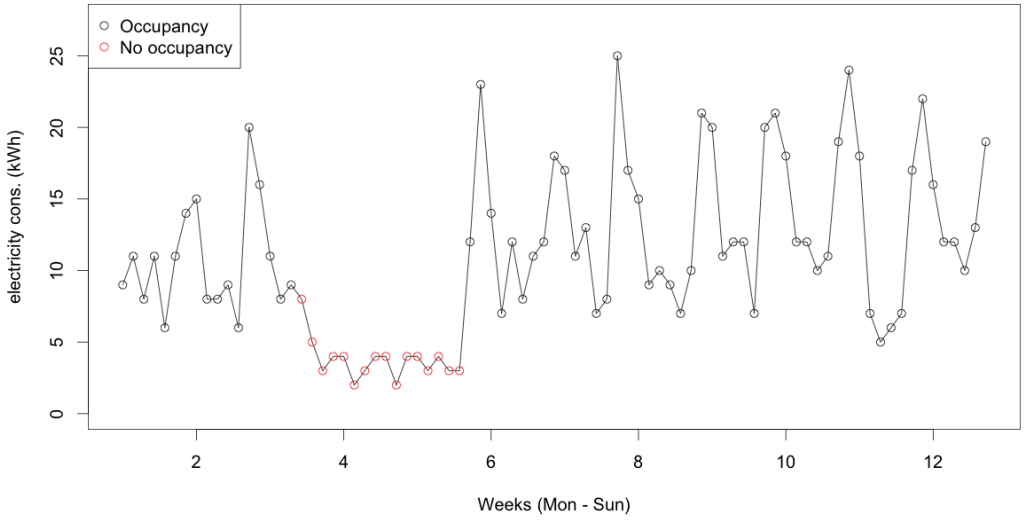
Figure 3.2: Algorithm for identifying low and high consumption clusters

I consider two consistency checks in the algorithm to increase the reliability of the method. First, I ensure that both identified clusters are distinguishable. This is done by testing that the absolute distance between the cluster centers is larger than $\delta = 1.25 * \sigma$, where σ is the average standard deviation within both clusters. Second, the $nDaysCheck = 4$ parameter ensures that the times of low consumption (i.e., absence of the residents) must consist of at least four consecutive days (e.g., an extended weekend) in the case of daily consumption measurements.

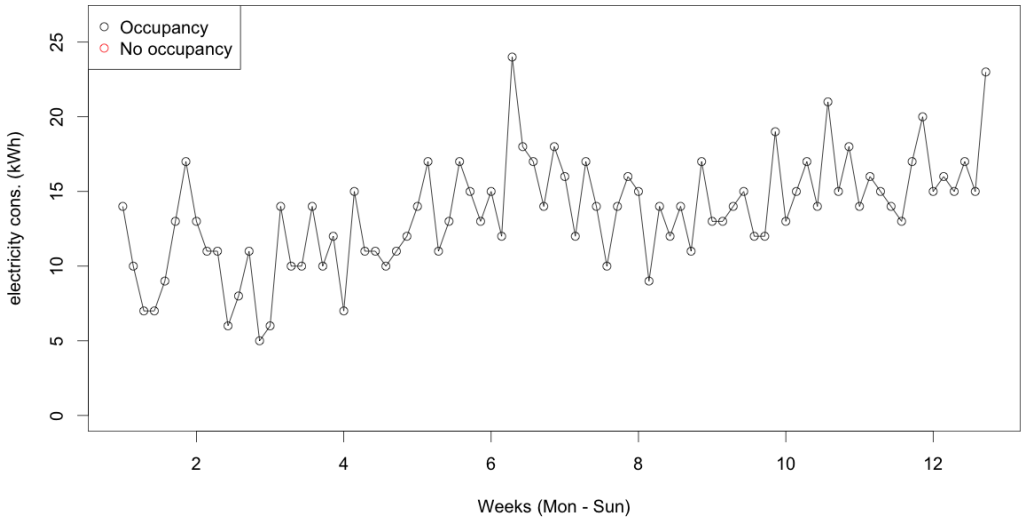
To illustrate the functional principle, two example electricity consumption curves with daily measurements for twelve weeks and the results of the algorithm are shown in Figure 3.3. In Figure 3.3a, a time of low consumption can be seen during week 3–5 that was identified by the algorithm (red circles) that we can interpret as absence of the residents. In Figure 3.3b, presence or absence times cannot be distinguished. With the algorithm, it is not necessarily possible to detect presence or absence times of residents in a household, but clusters with times of low and high consumption which is sufficient for the purpose of feature extraction. These resulting high and low consumption time spans are used in feature extraction to calculate the following statistics:

1. Average consumption in low-consumption times
2. Average consumption in high-consumption times
3. Relation of the average consumption in high / low consumption clusters
4. Average length of the low consumption cluster time span
5. Average length of the high consumption cluster time span

This exemplary application of unsupervised ML methods for feature extraction illustrates three aspects. First, an additional data analysis step can lead to new insights. The application of clustering helps to extract latent information from time series data (i.e., presence and absence information from daily electricity consumption data). Second, even if an unsupervised learning approach is used for feature extraction in this example, the complete feature extraction step cannot be automatized, as the implementation of the algorithm (as shown in Figure 3.2) must be done by an analyst. The clustering algorithm could not be executed on the raw data without identification of the correct parameters and follow-up processing to ensure that the resulting high and low consumption clusters are meaningful (i.e., using the *nDaysCheck* criterion). Third, theory and expert knowledge was combined to create an algorithm being able to identify two states of a household: high consumption and low consumption times which roughly matches to residents' presence or absence. The theoretical consideration of having a binary state (presence or absence) led to the determination of $k = 2$ clusters. The idea of using unsupervised ML for occupancy detection stems from earlier research SMD (Becker and Kleiminger 2017).



(a) Example for detectable occupancy time



(b) Example for non-detectable occupancy time

Figure 3.3: Electricity consumption load trace for daily consumption values in a 12-week period with occupancy any non-occupancy times

Features comparing the electricity consumption with neighboring households

Consumption data for commodities are spatio-temporal data, since each consumption event can be assigned a time stamp (i.e., time and date) and a location (e.g. address of the meter). In the above presented examples for characteristic extraction, the main focus was on reducing the time dimension. As the goal of feature extraction is the support of recognizing latent variables from the consumption data—which are difficult to predict from the data—the geographical dimension is also taken into account.

The spatial dimension contains plenty of information. Apparently, households often have similarities with their neighbors. People live, for example, in housing complexes where dwellings have equal sizes or buildings in a district were built in similar times. Such neighborhood-effects are long known and proven to be existing in social science (Dietz 2002), for example in research on social exclusion (Bauder 2002) or on child and adolescents development (Brooks-Gunn et al. 1993). The geographic location of customers should be present in all cases, as firms need a billing address and utility companies know where they deliver electricty.

Having defined a rudimentary neighborhood feature for annual electricity consumption data (see Equation 3.3 on p. 54), I added features that compare the electricity consumption to the nearest neighbors and expressed the number of neighbors in a radius of 50m, 250m, 500m, and 1,000m around the household location. I consider not only the average consumption difference from one household to its neighbors, but also other metrics that take the time series data into account. The proposed neighborhood features are listed in Table 3.3.

The presented examples of features on neighborhood-related information of households shows that theory (here: homophily[8]) can help to identify new variables by modeling known relations into data. This reduces the overhead of ML to recognize pattern that are already known from previous research.

Human knowledge is incorporated in the data preparation, as the radius in which neighbors are identified is set manually with reasonable distances (i.e., from 50m to 1,000m) and neighbors are identified using their distance to the household which is in the focus and the k nearest neighbors are used. The creativity of feature definition is, however, limited by reality. For instance, it was not possible to quantify the electricity consumption in relation to the neighbors in a radius around a household—which is the most natural understanding of neighborhood comparison—in the form of features, as there were only low

[8]Homophily describes the tendency of individuals to associate with similar ones.

Table 3.3: Neighborhood features based on smart meter electricity consumption data

Feature	Description
nnk_avgDist	Average distance to the ten nearest neighbors
nn_numNBs50m	Number of neighbors in a 50m radius
nn_numNBs250m	Number of neighbors in a 250m radius
nn_numNBs500m	Number of neighbors in a 500m radius
nn_numNBs1000m	Number of neighbors in a 1,000m radius
nnk_cons_relDiff	Total consumption, relative to k nearest neighbors
nnk_max_relDiff	Maximum consumption in the week, relative to k nearest neighbors
nnk_corDays_absDiff	Average correlation of days in the week, relative to the k nearest neighbors
nnk_numPeaks_relDiff	Mean number of peaks, relative to the k nearest neighbors
nnk_numAboveMean_relDiff	Number of observations (in the respective time resolution) above the mean, relative to the k nearest neighbors
nnk_consNoon_wd_wd_absDiff	Consumption during noon weekdays vs. weekend, relative to the k nearest neighbors
nnk_meanCor	Mean correlation of the load curve and the load curve of the k nearest neighbors
nnk_meanCor_wd	Mean correlation of the load curve and the load curve of the k nearest neighbors on weekdays
nnk_meanCor_we	Mean correlation of the load curve and the load curve of the k nearest neighbors on the weekend

numbers of households in many cases so that this approach would create many missing values. I decided therefore to calculate the statistics based on the k nearest neighbors[9].

3.3.2 Features from environmental data

The spatio-temporal dimensions of consumption data allow the addition of further data sources to reveal latent information from ambient data. Environmental information, such as weather observations, has strong impact on energy use and can help to eliminate environmental-related variations from the data (e.g., energy consumption suddenly increases with a temperature drop). As the

[9] I used $k = 10$ in my implementation, but this number should be subject to further investigation.

environmental data recorded also have time stamps and geographic locations, they can be connected to consumption data traces. Other environmental data, besides weather information, are observations on air quality[10] or pollen count. Both affects peoples' health and might therefore influence the presence time at home or human behavior on ventilation, heating, and energy use. Nevertheless, the addition of environmental data further increase the data dimension. In the following, I show, how spatio-temporal data that are highly correlated with the consumption data can be reduced to an overseeable number of features.

When the ordinary weather variables—I consider the five variables temperature, wind speed, wind direction, precipitation, sky cover in this work—are directly added to predictive analytics models, the dimensionality increases strongly, given that each weather variable is represented as time series data. In the case of five weather observations that are recorded in 30-minute intervals, this would lead to 1,680 measurements per week.

Previous research has identified a positive correlation between the electricity consumption and weather data (A. Albert and Rajagopal 2013; Hernández et al. 2012; Apadula et al. 2012). Taking this knowledge as given, it is not necessary to add the raw weather observations to predictive models. Instead, the correlation of energy consumption data with each weather variables can be used to define expressive features. In Hopf, Sodenkamp, and Staake (2018), we defined the following features for each weather variable:

1. `cor_overall` (correlation over the complete time series)
2. `cor_daily` (average correlation between weather and electricity consumption on each day)
3. `cor_night` (correlation during the night, 0:00–5:59)
4. `cor_daytime` (correlation during daytime, 6:00–17:59 Mon–Fri)
5. `cor_evening` (correlation during the evening, 18:00–23:59 Mon–Fri)
6. `cor_minima` (correlation of minima)
7. `cor_maxmin` (correlation of weather minima with consumption peaks)
8. `cor_weekday_weekend` (relation of the correlation on weekend / weekdays)

[10] A VGI project that collects data on air pollution is `https://luftdaten.info`, last accessed 06.02.2019

The features allow adding comprehensive weather information to models by using 40 features representing five weather variables (eight features per weather variable) instead of using the raw data of 1,680 measurements per week. Certainly, a good ML algorithm can also recognize the relation between weather variables and energy consumption from a dataset, wherefore one might argue that feature extraction is not necessary in this case. With feature extraction, however, this relationship in the data does not need to be learned, as it is known from research. Computational effort is thereby saved. Moreover, errors can occur by learning such true relationships from data, for example, when time spans without any change in the weather variable exist (e.g., no snowfall in summer, no sky cover over more than one week). An algorithm might assume that no relationship exists when the variable does not change in a considered time window that is too short to recognize a true relation (there are, for example, often weeks without change in sky cover). Important information is then erroneously omitted. Feature extraction helps in this case to explicitly model knowledge from theory in data representations and avoids to recognize wrong pattern.

3.3.3 Features from geographic information

Geographic conditions influence the characteristics of houses, determine the size of households, and affect many other aspects of human life. The geographic data can be connected to other data using the spatial dimension and, like environmental data considered before, helps to reduce variance and noise in ambient data, through providing additional data. Therefore, geographic data can provide detailed insights on customer living conditions and behavior patterns. The relationship between geographic conditions and living conditions has been long observed in social science research (Dietz 2002; Bauder 2002). Making use of this relation, geographic features can render aid in customer data analytics.

Because of its nature, geographic data do not only consist of numeric observations (e.g., spatial coordinates), but also of categorical data or textual descriptions that have semantic relations to objects in the surrounding. Information in geographical databases is also not always explicitly stored and must be sometimes retrieved by reasoning. A train station, for example, is a conceptual region that is maybe not mapped explicitly, given that only isolated parts of the train station like the entrance hall, tracks, and platforms are present in the geographic data base. In order to interpret the data, the semantic relations of geographical objects to others must be considered (Hopf, Dageförde, et al. 2015). Data structures that allow the storage of weakly structured information and the fact that even two-dimensional map data has a high complexity because of seman-

tic information makes the usage of geographic data without any preprocessing virtually impossible.

Geographic data have long been collected and maintained exclusively by central authorities. Since the emergence of the participatory web, however, users have begun to collect data with location information and have been making it publicly available. Goodchild (2007) describes this phenomenon as Volunteered Geographic Information (VGI) because users voluntarily provide the collected data.

The amount of currently available VGI Data is immense. The largest VGI initiative, OpenStreetMap (OSM), for example, has almost five million users and over 6 billion geo-coordinates stored[11]. In addition, the data has become more and more complete and accurate in recent years [12]. So, the data in OSM are available as a promising data source for business analytics. In Hopf (2018), I describe VGI data sources in detail and show how 60 features from OSM can be extracted. The features belong to the four categories:

1. *Topology*: describing the structure of and relations between one household and spatial neighbors (number of GPS coordinate points, distance to objects in the surrounding, without considering their context)

2. *Landmarks and points of interests*: distances and frequencies of objects by considering their meaning within the spatial context (frequency, distance, and other measures to sights, shops, cafes, etc.)

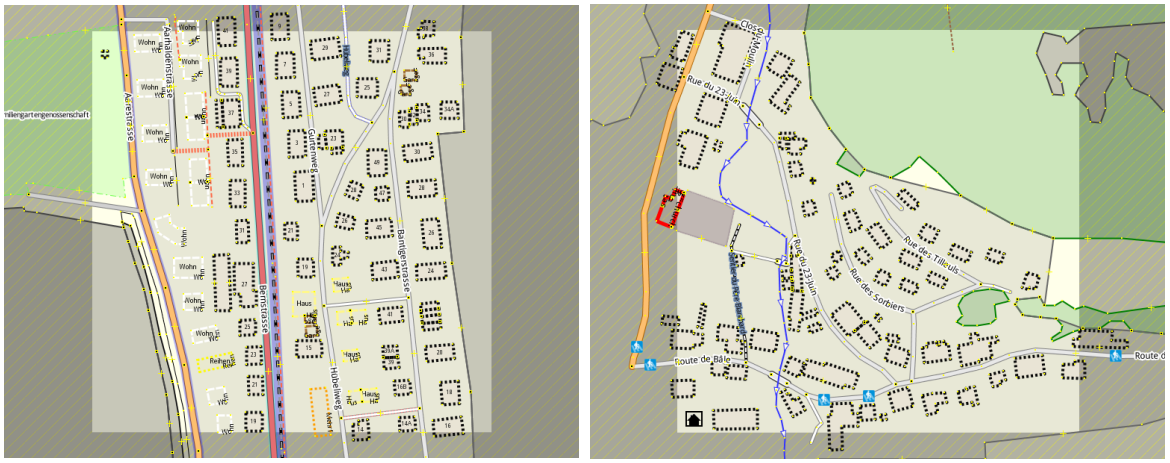
3. *Buildings*: mean and variance of the building basal area, the distance to buildings, and the type of buildings in the surrounding, etc.

4. *Land use*: land use type embracing the geo-location, area distribution in different land use types, etc.

The calculation of the feature values is based on a data selection using a bounding box around one household location, as illustrated in Figure 3.4. Typically, a bounding box of less than 1,000m $\times$ 1,000m is reasonable in the context of customer data analytics, but the choice of an appropriate geographical scope is case-specific. I recommend to not use the exact geo-coordinates of one household as feature in predictive models, because this would limit the resulting models

[11] On October, 02 2018, 4,914,714 users were registered on OSM and have created 6,388,813,141 GPS points (`https://www.openstreetmap.org/stats/data_stats.html`, accessed 02.10.2018)

[12] See for example Ciepłuch et al. (2010) for a study on the OSM Accuracy in Ireland and Zielstra and Zipf (2010) for Germany.



(a) Example address 1 (b) Example address 2

Figure 3.4: Map visualizations of OSM-data in 300m × 300m bounding-boxes around two customer address locations (the bounding-boxes are highlighted with brighter colors)

to the geographic region of the available training data. Predictions outside this area are hardly meaningful.

By looking at the geographic features, it becomes evident that theory or expert knowledge is necessary to extract variables from geographic data sources, as the automatic computation of features from such data sources is hardly possible. With the example of OSM, the geographic data are represented as 2D map data (i.e., points and polygon-lines) with semantic information (e.g., points of interests, geographic areas) can not be simply converted to variables related to a customer's address.

Moreover, theory from geography and relationships known from geographic information research can be included in landscape metrics to express spatial relations. These metrics have been used in the analysis of a variety of fields, such as the investigation of change in urban land uses, gentrification, urban sprawl and biodiversity (Herold et al. 2002; Irwin and Bockstael 2007; McGarigal et al. 2009). I adopted some of the metrics in my definition of features from OSM data (Hopf 2018).

3.3.4 Features from governmental statistical data

Open data from governments (also referenced as "public sector information") is on the rise. The US and the UK administrations, for example, have started

open data initiatives (Immonen et al. 2014), also in Germany[13], Switzerland[14], and other European countries, numerous open data portals are being created currently. Open data is considered to have a high economic value[15]. This is one reason why I consider the data source in this work. The other reason is that the data source is an example for geographic information that is represented as aggregated values for an entire region (each country defines their own statistical regions).

We investigated open government data from the EU, Germany, and Switzerland in Hopf, Riechel, et al. (2017) and reviewed the available official government statistics for their usability in predictive analytics. For that, we identified all household-related statistical data that were published by the three institutions and defined respective features from the data. I list the obtained features in Table 3.4.

Table 3.4: Statistics that have been identified in Hopf, Riechel, et al. (2017) as meaningful for predictive customer analytics (• = data available, ∗ = no or incomplete data for Germany, ⋆ = no or incomplete data for Switzerland)

		Data provider		
Feature	Description	EU	DE	CH
Building statistics				
rf.singleFamilyHome	Frequency of single family homes	∗		•
rf.multipleFamilyHome	Frequency of multiple family homes	∗		•
rf.residentialHomeAncillaryUse	Frequency of residential homes with ancillary use			•
rf.housePartlyResidential	Frequency of houses with partial residential use			•
mean.HouseAge	Average house age		•	•
rf.NewHouses	Share of new houses		•	•
rf.homeOwners	Number of residents which hold a share or are owner of the building they live in in the region			•
mean.NumRooms	Average number of rooms per apartment		•	•
rf.oneDwellingBuildings	Amount of buildings with one apartment	∗		
rf.twoDwellingBuildings	Amount of buildings with two apartments	∗		

[13] `https://www.govdata.de/`, last accessed 24.08.2018

[14] `https://opendata.swiss/en/`, last accessed 24.08.2018

[15] "The total economic value of published open data in Europe is estimated to be between €27 billion (Dekkers et al. 2006) and €140 billion (Vickery 2011)" (Hopf, Riechel, et al. 2017)

Feature *(continued)*	Description	Data provider EU	DE	CH
rf.threeOrMoreDwellingBuildings	Amount of buildings with three or more apartments	∗		
rf.nonResidentialBuildings	Amount of buildings that are not used for residential purposes	∗		
Population statistics				
rf.male	Amount of male population in percent	•	•	•
rf.permResidentials	Amount of permanent residents			•
mean.ResidentialAge	Average age of permanent residents	•	•	•
num.permResidents	Quantity of permanent residents			•
num.migration	Difference of immigration and emigration	•	•	•
num.residents	Quantity of total residents (permanent and non-permanent)	•	•	•
Economic statistics				
rf.zeroEmployeeBusinesses	Amount of businesses that have 0 employees	⋆	•	
rf.smallBusinesses	Amount of businesses with 0-9 employees	⋆	•	•
rf.mediumBusinesses	Amount of businesses with 10-250 employees		•	•
rf.bigBusinesses	Amount of businesses with 250 and more employees		•	•
num.gdpEuroPerCitizen	Quantity of GDP in Euro per citizen	⋆	•	
rf.publicInvest	Relative frequency of investments in buildings by the public sector			•
rf.newInvest	Relative frequency of investments in new buildings instead of renovation			•

The statistical data published by governments in the form of one figure for a geographical region (e.g., absolute or relative frequency of a variable for a municipality) cannot be used directly, as whole tables would have to be connected to the customer data record. Besides, the data contains many redundancies, like frequencies or percentages (e.g., of male and female population). The transformation of published statistics to usable features that can be used in analytics must be done manually.

Surprisingly, the majority of published governmental statistical data are hardly usable for business data analytics, because the data are only available on a national or regional level and the number of available datasets is low. The handful of 23 features that are usable in our case brought only notable performance improvements in Switzerland, not in Germany.

3.3.5 Contribution of empirical feature extraction to model building

This section introduced empirical feature extraction as an approach to integrate and harmonize different data sources. The approach also helps to reduce the input data dimension to avoid congestion of ML algorithms. Moreover, feature extraction is a major activity of model building, because human cognition is used to encode theory and expert knowledge into variables and thus make it available to algorithms.

To illustrate the approach, I presented eight examples of feature extraction from four typical data sources (transaction data, environmental observations, geographic information, and government statistic data) in this section and delineated how feature extraction can help to obtain distinct information from the respective data source. Additionally, I described the specialty of each data source and named typical characteristics of ambient data sources which represent the challenge in the computational processing of each data source. Hereafter, I underline the importance of conquering specialties of available data with feature extraction and summarize the advantages that emerge from the nexus of human cognition, theory, and expert knowledge to obtain features.

Methods needed to conquer issues in raw data

The large diversity of data sources that are available to firms for analytics and their varying quality make it hard to pursue a proper strategy for data preparation. Kitchens et al. (2018) point out that "no single data integration strategy is sufficient." I exemplified several issues in real-world datasets from the energy retailing industry that firms have to solve before they can apply advanced analytics methods in this chapter. These main issues are summarized below.

Foremost, the sheer amount of data is a problem for advanced analytics—including ML methods—to make sense of. Whereas more *observations* of real world objects or events are welcomed, because they allow creating more accurate models of the reality, a high number of *variables* are a challenge for algorithms. This became obvious with the presented example of consumption information regarding electricity, gas, and water that can be combines with free available data (weather data, environmental observations, geographic information and government statistics). It is nearly impossible to detect pattern from that large amount of data without dimensionality reduction. It is generally unclear, which variable helps to reveal a certain information of interest. Feature extraction is therefore an opportunistic approach in which all features that come to ones

mind (through creativity, experience, etc.) are defined and tested to obtain their value.

Second, the data contain redundancies: Temporal redundancies, for example, are caused by daily living cycles (similar consumption pattern on weekdays), or can result from the fact that activities are recorded in multiple variables (for example, electricity, gas, and water consumption together with weather data). Geographic redundancies can result from considering neighboring households that are co-located and are likely to have similar household characteristics. Through feature extraction, it is possible to explicitly give extra weight to important information, and less weight to unimportant characteristics (e.g., eliminate redundancies where beneficial, normalize values to bring certain variables in the same range).

Third, the diversity of information stored and the respective high number of data types available makes it hard to implement algorithms that can work on all data types. Spatio-temporal information—often present in customer data analytics—consist of a mixture of variable types and information with recorded change over time. Geographic data are difficult to process, as they contain not only numeric measurements but also textual description and semantic information.

Fourth, data sparsity is a problem. For example, data streams contain can missing values, multiple time-series data may not be present for all households (e.g., a household has only an electricity contract with the utility company, but none for gas delivery), and information can be not recorded in a certain geographic area.

All four specialties of raw data sources must be handled in data analysis. Empirical feature extraction can help to conquer these mentioned issues.

Contribution of human cognition to obtain features

Beyond the difficulties in the data, all variables contain latent information on consumer behavior pattern that firms aim to extract. To achieve this goal, it is necessary to perform several data processing steps, whereby algorithmic methods need assistance of humans. The examples of empirical feature extraction emphasize the need of human cognition and the ability to translate the observations into program code.

Human sense-making allows implementing algorithms that identify distinct characteristics of real-world objects. On the base of SMD, for example, humans can implement algorithms that identify peaks and quantify their magnitude, search for low and high consumption times, or define reasonable thresholds for

data (e.g., electricity consumption over 0.5kWh indicates activity). Moreover, humans have the creativity to create new variables by transforming or combining available data, interpret existing data in a new way, can detect new information, and include additional data sources into the analysis. For example, human reasoning leads to the explicit consideration of phenomena recorded in multiple time series variables (the increase of electricity use when water is consumed is an indicator for electric water heating, when gas heating is defect, residents can temporarily use electric fan heater).

Having different features available, humans can conclude a meaningful codomain of variables, recognize programming errors, or interpret which calculation can lead to infinite or not available values. For such cases, a special treatment of variables, or a value that can be used as replacement in the case of missing values can be defined. This can hardly be done autonomously by algorithms that are currently known.

Contribution of theory or expert knowledge to obtain features

In all considered examples of empirical feature extraction from the four data sources, theory, and expert knowledge played an important role in defining features, as the knowledge is explicitly or implicitly modeled by the data analyst.

Theory can provide guidance to the cognitive process of feature extraction, as relationships between variables and actual facts known, for example from research, do not need to be learned from data through an algorithm. I gave several examples of how theory guided the feature extraction:

- The information that peoples' living situations do only rarely change much over time (Wells and Gubar 1966) helped to remove redundancy in annual consumption data of consecutive years.
- The fact that neighborhoods often show similarities (Dietz 2002; Bauder 2002; Brooks-Gunn et al. 1993) brought the idea to implement neighborhood-related features.
- Knowledge on the correlation of energy use with environmental conditions (A. Albert and Rajagopal 2013; Hernández et al. 2012; Apadula et al. 2012) helped to express the weather information in eight features per weather variable (which is a decent reduction of the data dimension).
- The use of landscape metrics to quantify geographical regions (Herold et al. 2002; Irwin and Bockstael 2007; McGarigal et al. 2009) was a blueprint to define geographic features.

These examples demonstrate that theory and expert knowledge provides value for data analytics. Moreover, computational effort can be reduced by considering known facts instead of learning such relationships from data.

3.4 Dimensionality reduction through automatic feature selection

Selecting the relevant predictors from available datasets is a serious challenge in data-driven decision making process (*data to insight* phase). This is true in both cases: When raw data is directly used and when it has already been reduced through empirical feature definition. In fact, taking many data sources into account, the number of available variables is still large, even when the initial dimension was reduced by empirical feature definition.

In many cases, only some variables of a given dataset carry relevant information, whereas others contain noise or irrelevant values.[16] This curse of dimensionality lowers the quality of ML applications and the success of predictive systems. Automatic approaches can help to overcome the problem of many input variables.

Automatic Feature Selection Methods (FSMs) can help to reduce the model complexity, increased model generalization performance, lower training times, and lead to models that need less storage space (Guyon, André, et al. 2003; Kudo and Sklansky 2000). An extensive number of FSMs is available and multiple software libraries exist that provide these methods to data analysts (Chandrashekar and Sahin 2014; Guyon, André, et al. 2003; Saeys, Inza, et al. 2007). Moreover, several studies have been published that describe new FSMs or present literature reviews (Bolón-Canedo et al. 2015; Chandrashekar and Sahin 2014; Guyon, André, et al. 2003; Huan Liu and Motoda 2008; Saeys, Inza, et al. 2007). For practitioners as well as researchers that start working with automatic FSM, an overview to methods in the widely used statistical programming environment R is missing. Besides, a comprehensive benchmark of current methods in open source program libraries is—to the best of my knowledge—not existent.

The remaining section first gives a brief overview to the three fundamental approaches of feature selection. Thereafter, I present a collection of FSMs that are available in the statistical programming environment R.

[16] The problem is serious: Whereas in the last decades, a number of 50-100 features was called a "large" feature set (Kudo and Sklansky 1998), today we are confronted with hundreds or even thousands (Hua et al. 2009) of features.

3.4.1 Types of automatic feature selection approaches

Existing FSM are typically classified into the three categories (Guyon and Elisseeff 2006; Lal et al. 2006) filter, wrapper, and embedded methods, which are briefly introduced below.

Wrapper methods Approaches in this category use the actual classification performance (see section 4.2 for an introduction) to assess the quality of a complete feature set in combination with a learning algorithm and try to optimize this overall classification performance by adding or removing features from the model. Most wrapper selectors use greedy-algorithms to expand or reduce the feature set iteratively. Two basic approaches are, for example, the *forward selection* approach, in which the procedure starts with an empty set and progressively adds features yielding to the improvement of a performance metric; the *backward elimination* approach, in which the procedure starts with all the features and progressively eliminates the least useful ones. Further approaches exist that overcome certain limitations of basic approaches. General advantages of wrapper methods are that feature combinations are evaluated and that the approaches are suitable for any classification problem. Disadvantages are the high computational complexity (model training and test must be done for each modification in the feature set) and that not all features might be tested, as a local optimum in classifier performance is selected.

Filter methods The class of FSMs use statistical measures (e.g., correlation coefficients, statistical tests and entropy measures) that try to quantify the expressiveness of features for the classification task. In that way, filter methods do not incorporate the learning step, as they are based on heuristics regarding good features. These measures can be used to rank features according to their importance and select the most appropriate number of features based on this rank. The advantages are a low computational complexity (as the calculation of statistical measures is done once) and the reasons for selection a certain feature are reasonable. Disadvantages are, that interaction between features can only hardly be measured, the measures quantify only certain characteristics of features (e.g., correlation between features and the dependent variable, or correlations within the feature set) and not all measures are appropriate for all variable types problems (e.g., multi-class classification problems, or categorical variables). Previous studies indicate that wrapper methods cannot outperform filter methods (Haury et al. 2011). Therefore, this work focuses mainly on filter methods.

Embedded methods This type of feature selection does not separate between the learning part and the feature selection part, rather the interaction of both makes the set of methods distinct from the other two categories. A conceptualization of this set of methods is given by Lal et al. (2006). An example of such methods is the implementation feature selection into the optimization problem of Support Vector Machine algorithm (an explanation of this algorithm can be found in subsection 4.3.3), as done by Carrizosa et al. (2016).

3.4.2 Collection of Feature Selection Method (FSM) in R

In order to create an overview to implemented FSM, I conducted a survey of software libraries for automatic future selection in GNU-R[17] between April and June 2017. This review led to 43 implemented methods available for use.

I list the identified methods in Table 3.5 together with short descriptions that summarize the main idea of each algorithm (Romanski and Kotthoff 2014; Kursa and Rudnicki 2010; Robnik-Sikonja and Alao 2016). For further details, the interested reader can follow the referenced sources.

Table 3.5: Identified filter methods for feature selection together with the software library and literature reference (if applicable)

Method	Description	Reference
R Package 'Boruta'		
Boruta	Boruta iteratively compares the importance of features (based on the Random Forest algorithm) with the importance of shadow features, created by shuffling the original ones. Features that have a worse importance than shadow ones are consecutively *dropped*. Features that are better than shadows are *confirmed*. Shadows are re-created in each iteration. Algorithm stops when only confirmed attributes are left, or when the algorithm reaches a maximum number of iterations. When the maximum number of iterations is reached, all features without a decision are considered as *tentative*. The sets of dropped, confirmed and tentative feature can be analyzed separately.	(Kursa and Rudnicki 2010)
R Package 'CORElearn'		
DistAngle	Cosine of angular distance between splits.	
DistAUC	AUC distance between splits	
DistEuclid	Euclidean distance between splits	

[17]The survey was conducted in the Comprehensive R Archive Network (`https://cran.r-project.org/`)

Method *(cont'd)*	Description	Reference
DistHellinger	Hellinger distance between class distributions in branches	
DKM	A measure following Dietterich et al. (1996) that is suitable for two class problems	(Dietterich et al. 1996)
DKMcost	Cost-sensitive variant of DKM	(Robnik-Šikonja 2003)
EqualDKM	DKM with equal weights for splits	
EqualGini	Gini index with equal weights for splits	
EqualHellinger	Two equally weighted splits based Hellinger distance	
EqualInf	Information gain with equal weights for splits.	(Hunt et al. 1966)
GainRatio	Gain ratio, which is normalized information gain to prevent bias to multi-valued attributes	(Quinlan 1986)
GainRatioCost	Cost-sensitive variant of GainRatio	(Robnik-Šikonja 2003)
Gini	Gini-index	(Breiman 1984)
ImpurityEuclid	Euclidean distance as impurity function on within node class distributions	
Impurity Hellinger	Hellinger distance as impurity function on within node class distributions.	
InfGain	Information Gain as used in the original decision tree.	(Quinlan 1986)
MDL	Minimum Description Length, a method introduced by Kononenko (1995) with favorable bias for multi-valued and multi-class problems.	(Kononenko 1995)
MDLsmp	Cost-sensitive variant of MDL where costs are introduced through sampling	(Robnik-Šikonja 2003)
MyopicReliefF	Myopic version of the ReliefF algorithm resulting from assumption of no local dependencies and attribute dependencies upon class.	(Kononenko 1995)
Relief	The original algorithm of Kira and Rendell (1992) working on two class problems. The algorithm calculates scores for each feature, based on the Euclidean distance to nearest neighbor training instance pairs.	Kira and Rendell (1992)
RReliefF	An updated version to the Relief algorithm using the Manhattan distance and is applicable to work with multi-class problems.	(Kira 1992)
ReliefFavgC	Cost-sensitive 'ReliefF' version with average costs.	(Robnik-Šikonja 2003)
ReliefFbestK	'ReliefF' variant testing all possible k nearest instances for each feature and returns the highest score.	(Robnik-Šikonja 2003)

Method *(cont'd)*	Description	Reference
ReliefFdistance	'ReliefF' variant where k nearest instances are weighed directly with its inverse distance from the selected instance.	(Robnik-Šikonja 2003)
ReliefFequalK	'ReliefF' algorithm where k nearest instances have equal weight.	(Robnik-Šikonja 2003)
ReliefFexpC	Cost-sensitive 'ReliefF' algorithm with expected costs.	(Robnik-Šikonja 2003)
ReliefFmerit	'ReliefF' algorithm where for each random instance the merit of each feature is normalized by the sum of differences in all attributes.	
ReliefFpa	Cost-sensitive 'ReliefF' algorithm with average probability.	(Robnik-Šikonja 2003)
ReliefFpe	Cost-sensitive 'ReliefF' algorithm with expected probability	(Robnik-Šikonja 2003)
ReliefFsmp	Cost-sensitive 'ReliefF' algorithm with cost sensitive sampling.	(Robnik-Šikonja 2003)
ReliefFsqr Distance	'ReliefF' variant where k nearest instances are weighed with its inverse square distance from the selected instance.	(Robnik-Šikonja 2003)
ReliefKukar	Cost-sensitive 'Relief' variant	(Kukar et al. 1999)
UniformDKM	DKM measure with uniform priors	
UniformGini	Gini index with uniform priors	
UniformInf	Information gain with uniform priors	
R Package 'FSelector'		
cfs	The algorithm finds attribute subset using correlation and entropy measures for continuous and discrete data	(Hall 1999)
chi.squared	The algorithm finds weights of discrete attributes basing on a chi-squared test	(Huan Liu and Setiono 1995)
consistency	The algorithm finds attribute subset using consistency measure for continuous and discrete data	(Dash, Huan Liu, and Motoda 2000)
gain.ratio	The algorithms find weights of discrete attributes basing on their correlation with continuous class attribute	(Cover and Thomas 2006; Hunt et al. 1966)

Method *(cont'd)*	Description	Reference
oneR	Find weights of discrete attributes basing on very simple association rules involving only one attribute in condition part	(Holte 1993)
random.forest. importance	Finds weights of attributes using Random Forest algorithm (see subsection 4.3.5); for each tree, the prediction error on the out-of-bag portion of the data is recorded (error rate for classification, mean squared error for regression)	(Breiman 2001)
symmetrical. uncertainty	Symmetrical uncertainty measure	(Yu and Huan Liu 2003)

I reviewed the technical documentation of the identified methods together with student research assistants and categorized the methods into one or multiple of the three following categories:

1. 16 methods consider *interdependencies between features*. Reversely, 27 methods that evaluate features separately do not take the context of other features into account. The latter ones are called "impurity based methods" (Huan Liu and Motoda 2008, p. 170).

2. Nine methods have the ability to assign *class importance* (or cost) factors to single classes in a classification problem. This enables fine-tuning of ML algorithms and FSM in order to better recognize less frequent classes.

3. 29 methods have *explicit support for multi-class problems*, whereas others are (per definition) designed for binary classification problems. In practice, all tested FSMs work for both binary and multi-class classification problems due to internal class binarization (*implicit multi-class support*) but it can be expected that methods with explicit support for multi-class problems perform better for those than others.

Obviously, it is unclear which method should be chosen to select a suitable subset of features for a prediction problem. Several methods follow similar working principles and deviate from each other only in small details (e.g., the variants of the *Relief* algorithm). Experience of an analyst or a rigorous benchmark is necessary to select the right method for feature selection. To the best of my knowledge, a comprehensive benchmark of FSM with the available methods, like those listed in Table 3.5, is not available yet. I will therefore provide a systematic benchmark of the identified methods using a dataset from energy retailing in section 5.4 to provide aid to analysts that want to use methods that are available in the statistical programming environment R.

3.5 Discussion and implications

Data is the feedstock of business analytics. This chapter presented an overview of available data sources for business data analytics and gave examples of empirically extracted features for energy consumption data, environmental data, geographic information, and public statistical data, as well as introduced into the topic of automatic feature selection.

Answer to RQ 1 Research is currently lacking a systematic overview of which data sources exist within and outside of organizations. As the answer of the first RQ, I developed a taxonomy of data sources available for analytics (see Table 3.1 on p. 47), based on data sources mentioned in IS research studies and the experiences from seven case studies from energy retailing. I consider *internal data* that are created by firms and stored within their databases. Such business data comprise information on customer details (e.g., name, address, contact details), transactions, interactions, and basic demographic variables. In addition, *external data* are considered that can be used by firms to generate additional insights. As external data, I identified nine categories including public statistical data, geographic information, weather data. This taxonomy of data sources available for analytics fosters the research on value creation from big data, as it gives a systematic overview to them. The taxonomy is also a starting point for future research. Created with the base of research publications and case studies from this dissertation, the systematic overview needs further validation. This may include interviews with data scientists and case studies in other industries.

Answer to RQ 2 In the second part of this chapter, I presented a synopsis for two fundamental approaches of data preparation for ML algorithms: First, *empirical feature extraction* is introduced in section 3.3 and exemplified among four ambient data sources (energy consumption data, geographic information, and government statistical data). Second, the three approaches of *automatic feature selection* (wrapper, filter, embedded methods) are explained and an overview to 43 FSMs is given in section 3.4. The second RQ adds to the discussion on whether human cognition, theory, and expert knowledge can support data analytics in the activity of data preparation, even when several automatic approaches exist. The main arguments and findings presented in this chapter are summarized below.

The reasons for an increased application of automated methods in the preparation of data are plausible. First, data preparation is an expensive task, as data scientists spend up to 80% of their time on this activity. Second, auto-

matic ML approaches select data points and parameters more rigorously. Third, algorithmic approaches may also provide new insights to previously unknown pattern. To conclude from the benefits of automated approaches, that "Theory is Dead" (Jankel 2017), however, is short-sighted. The claim may be valid for business models that quickly change in current times, but scientific theory—and the knowledge on associated concepts as well as their relations—is still of inestimable value, as it makes principles of the world explicit and provides in that way guidance for data analysts in the preparation of data.

My research results support the argument of Sharma et al. (2014) that "insights do not emerge automatically out of mechanically applying analytical tools to data." The huge amount of data that are available to firms must be integrated, prepared and its dimension reduced. As there is no single strategy to this complex task, I demonstrated the benefits of the nexus of human cognition, theory and expert knowledge to the complex data preparation task. Additionally, it is evident from earlier research that trials in which empirical features were derived from data (Beckel, Sadamori, Staake, et al. 2014; Beckel, Sadamori, and Santini 2013) lead to better prediction results than in trials without a decent feature extraction data processing (A. Albert and Rajagopal 2013). In Hopf, Sodenkamp, Kozlovskiy, and Staake (2014) we also substantiated that the definition of features from SMD can lead to a significant improvement in prediction performance.

Figure 3.5 further illustrates how empirical feature extraction can be applied in the field of energy data analytics (this case is continued in the later chapters) and backs the argument for using human cognition, theory, and expert knowledge in data preparation: Considering one week of electricity consumption data and respective weather data from five variables, geographic map data with semantic information together with statistical data, the raw data have a huge complexity. With empirical feature extraction, this can be at least reduced to a manageable number of features. The definition of features is an empirical engineering task and can be theory-driven considering literature (e.g., landscape indices from geography to describe map data), qualitative description (e.g., inspection of load curves or map data), statistical analysis (correlation, time series, etc.). Human expert knowledge can also be a source for features. Beckel (2015), for example, conducted interviews with energy consultants and used the insights for feature definition.

The RQ 2 can be answered positively, as human cognition, theory, and expert knowledge provides value to data analysis even when several automated methods exist. The most successful approach to data preparation for ML is certainly to *combine both approaches* and *first define empirical features and then*

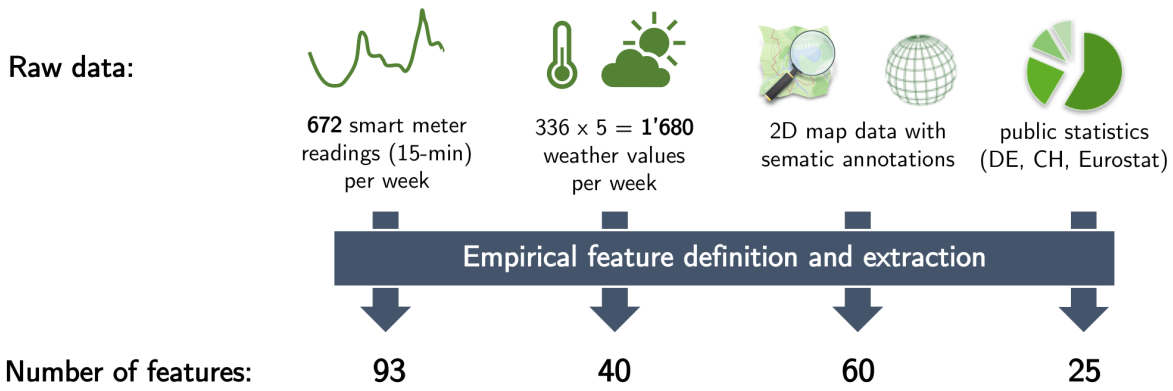


Figure 3.5: Theory and human expert knowledge based dimensionality reduction through feature extraction

apply automatic feature selection algorithms to further select variables based on computational procedures.

So far, I presented an argumentative answer to RQ 2. I am inspecting automatic dimensionality reduction techniques in chapter 5 and will continue the answer to this RQ there.

4 Machine learning methods for predictive analytics

Highlights

- ▷ A summary of the main supervised ML algorithm categories is given and six algorithms are explained.
- ▷ Classification performance metrics are described and compared with each other.
- ▷ The algorithms and methods presented are the basis for analyses in the following chapters; the overview can serve as an introduction to ML for interested readers.

Firms and their data scientists are likely to be faced with millions of data points related to their customers that are available for predictive analytics. This becomes explicit, considering the different ambient data sources in organizations that contain a variety of information which is represented in different data types, including numeric and non-numeric data. The data contain missing values, noise, and outlier. Outlier refer thereby to values that are obviously extreme (e.g., a private customer ordering a pallet of beer, or a household that consumes over 50,000kWh electricity oer year) and noise describes means that a variable is confounded by unwanted influences (e.g., imprecise measurements, environmental influences). Moreover, events that are aimed to be revealed from the data have often infrequent occurrence (e.g., only one in a thousand customers has the willingness to pay for a premium electricity tariff). In short, ML algorithms have to deal with challenging pattern recognition tasks and must be effective in their sense-making of the data to deliver compelling results.

Not all algorithms can adequately process a high number of input variables, noisy data, or data that contain differnt data types. The challenge for algorithms is to identify the relevant features and encounter the curse of dimensionality. Whereas in the last decade, a number of 50–100 features was called a “large”

feature set (Kudo and Sklansky 1998), today we are confronted with hundreds or even thousands of features (Hua et al. 2009).

Additionally, a requirement for business analytics models is the ability to explain the obtained models, at least to some extent. This requirement is necessary, because the reliability of a model (and the respective variables that are the base for this model) can sometimes only be judged, for example by an analyst or decision maker, when interpreting the underlying data base. Besides, managerial decison makers might want to understand how predictions are made and how reliable they are. It is also important to make sure that models are not too closely aligned with the data (overfitting), as good models deliver meaningful results even when the data are noisy or only a few data points are available for model training.

This chapter introduces the methods needed to obtain and evaluate predictive models, when features are already prepared (as illustrated in Figure 4.1). After the overview to data sources, feature extraction, and feature selection in the previous chapter, an introduction to ML algorithms and the evaluation of classification models is given here.

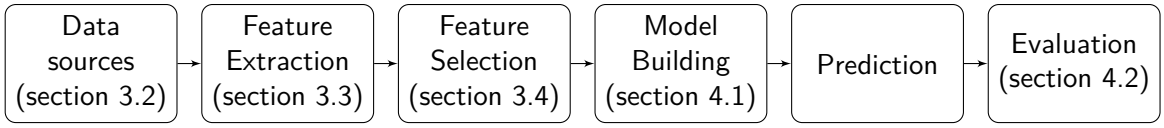


Figure 4.1: The predictive modeling and evaluation process

In the first section of this chapter, the six most important classes of learning algorithms are briefly reviewed. The second section introduces evaluation metrics for the classification performance. Finally, six acknowledged ML algorithms are explained, together with the most important parameters of each algorithm that can be modified and "tuned" during the modeling process. The focus of this dissertation is not to further develop ML algorithms. It is rather to investigate the effective combination of existing algorithms together with other data analytics techniques. The review of supervised ML algorithms in this chapter serves as an introduction to the topic and as a background for the following analyses.

4.1 Overview to supervised machine learning algorithms

Machine learning describes a class of algorithms that detect pattern in data to generalize a model from data. In the case of *supervised machine learning*, the training instances are labeled with a specific outcome (e.g., customer bought a product or not) so that an algorithm can "learn" to separate a number of classes based on input variables (called *features*, *explainable variables*, or *predictors*). This is also called "classification". The primary goal of supervised machine learning is therefore to use the feature values to predict an outcome. The outcome variable is called *dependent variable* or *responses* (Hastie, Tibshirani, and J. Friedman 2009). There are also algorithms for *unsupervised learning* (mainly *clustering* or *segmentation*) that detect pattern from unlabeled data, but these algorithms are not in the focus of this work. Besides the primary goal of prediction, the trained models can also be used to explain phenomena, as it is done classically to develop or verify theories or hypotheses (Shmueli and Koppius 2011). Well-known textbooks and the state of the art machine learning literature (Han et al. 2012; Hastie, Tibshirani, and J. Friedman 2009; Kotsiantis et al. 2007; Mitchell 1997; Russell and Norvig 1995; Zaki and Meira Jr. 2014) provide different categorization of the existing machine learning algorithms. The most common categories are:

- Logic based learners (including the learning rule sets and decision trees)
- Instance based learners (including nearest neighbor classifier)
- Statistical learning (e.g., Naïve Bayes, Linear Discriminant Analysis)
- Support vector machines
- Ensembles (e.g., Bagging, Boosting, Random forest)
- Artificial neuronal networks (single and multi-layer perceptron, neural networks)

All types of algorithms have different capabilities to cope with challenges of machine learning, for example the curse of dimensionality, complex decision boundaries (i.e., not linearly separable classes), imbalanced classes. For each category, I explain the functional principle of the machine learning algorithms.Finally, I briefly discuss the suitability of the algorithm class for predictive analytics in energy distribution.

Logic based learners Logic based learners are one of the first concepts in machine learning. They are built upon set of rules or decision trees and have a low generalization ability, are sensitive to noise in the data and lack in handling continuous features. These methods have not found much attention in business analytics, except decision tree learners. One reason might be, that logic based learner are more suitable for semantic data that must be specially encoded than for numeric data. Mitchell (1997) points out, that several more advanced algorithms exist that are more appropriate for learning tasks than algorithms in this category.

Instance-based learners Instance based learners delay the processing of learning examples until new examples are classified (they are therefore also known as "lazy learners"). Therefore, the classification step can be computationally expensive. As the most prominent representative of this category, the k Nearest Neighbors (kNN) was frequently used. Instance-based learners have low ability to handle high-dimensional data and are therefore less appropriate for multi-dimensional classification.

Statistical learning algorithms They use concepts of statistical analysis for classification. Example algorithms are Logistic regression, Naive Bayes, and Linear Discriminant Analysis (LDA). Algorithms ususally have a low computational complexity, but their generalization performance is limited. They can handle complex and non-linear decision boundaries only in some way. Besides that, they are susceptible to noise and multiple input vectors.

Support Vector Machines (SVMs) To address the classification problems with complex decision boundaries with non-linearly separable classes, SVM performs the so-called "kernel trick" to transform the input vector in the higher dimensional space and used a soft-margin to separate classes. SVMs are among the best currently known classifiers (Fernández-Delgado et al. 2014).

Ensembles Ensemble learning methods combine multiple machine learning models with the goal to create an improved composite classification model. Ensemble classifier predictions are based on the votes of the base classifiers (Han et al. 2012) that are mostly simple learning algorithms, such as decision trees. Two types of ensemble methods are known: 1) *Bagging* (bootstrap aggregation): algorithms train various "weak" classifier using different subsamples of the training set that all classify new examples. With this approach, the variance of prediction is reduced and the accuracy is increased. 2) *Boosting*: trains

multiple classification models, and aggregates the final prediction including a weight for each base classifier, resulting from its assessed accuracy.

The first boosting algorithm was AdaBoost. Random Forest (RF) who was rated as the best performing algorithm in a comprehensive study by Fernández-Delgado et al. (2014) and Extreme Gradient Boosting (XGB) who won a number of machine learning online competitions[1]. Therefore, and because they have a higher ability to be explained than other ML algorithms, Ensemble learner are good choices for business analytics.

Artificial Neural Networks (ANNs) This class of machine learning algorithms estimate models from training data with network structures inspired by biological neural networks. Each node of the network (so-called "neurons") has multiple weighted inputs. An activation function converts the input for each neuron to one single output. During the learning process, the weights of the inputs are learned for each neuron. According to the structure of the network, the learning algorithm and the activation function, many types of ANN exist. Through recent advances in computing power and cluster-computing, impressive application of large neural networks (so-called *deep neural networks* or *deep learning*) have been made. Deep learning algorithms need, however, large datasets with many thousand examples or more, which makes the approach less useful in business analytics.

4.2 Classification performance evaluation

Quantifying the performance of ML predictions is key to evaluate and improve models. Classification performance metrics must—similarly to ML algorithms—account for several difficulties of learning tasks. For example, the measures must be reliable (e.g., rate a lower performing models worse than a good one), good to interpret (e.g., by having a natural benchmark that also laypeople can understand), it should be possible to obtain the metric also for multi-class prediction problems, and extreme cases (e.g., a model does not recognize one class of many). In fact, no metric fulfills all requirements, wherefore several metrics must be taken into account when evaluating models. This section gives an overview to the most important metrics, names their strengths and weaknesses.

All metric have in common that they quantify the amount of correctly classified examples in a test sample. The number of correctly recognized examples (true positives, TP), the number of correctly recognized examples that do not

[1] see `https://www.kaggle.com/dansbecker/xgboost`, last accessed 03.10.2018

belong to the class (true negatives, TN), and examples that either were incorrectly assigned to the class (false positives, FP) or that were not recognized as class examples (false negatives, FN) are counted. These four cases constitute the so-called confusion matrix shown in Table 4.1 for the case of the binary classification (Sokolova and Lapalme 2009).

Table 4.1: Confusion matrix for binary classification

		Predictions		
		Positive	Negative	Total
Ground Truth	Positive	TP	FN	P
	Negative	FP	TN	N
	Total	P*	N*	

The remaining section is structured as follows. First, metrics in the scope of one dependent variable (binary and multi-class) prediction problems are described. Second, metrics for binary classification problems are presented that can also be calculated for single classes in a multiclass classification problem. Third, the sound calculation of the metrics using a holdout or ▷cross-validation is explained. In the concluding section, the metrics are compared.

4.2.1 Metrics for dependent variables with multiple classes

The metrics for dependent variables express an average classification performance for the trained model and do not distinguish between different classes. Though, they help to get a first impression of the prediction and enable the comparison of classifier between different classification problems.

Accuracy is defined as the portion of correctly classified instances from the number of total classification instances. The formula for the binary classification case can be found in Equation 4.1. In the case of more than two classes, the accuracy is calculated by counting all correct classified examples divided by the sample size. For example, when one class of the dependent variable occurs only in 1% of the cases, a classifier that does not detect the class at all can reach an accuracy of 99%.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

The measure can take values between 0 and 1, where 1 corresponding to the perfect prediction and 0 total misclassification. Accuracy is easy to interpret, but in the situation the classes are imbalanced (i.e., one class occurs much more often than the others) a classifier that always predicts a majority class can achieve high accuracy. Therefore, this measure can be slightly misleading if applied to such unbalanced properties. Accuracy is, thus, influenced by the class distribution of the dependent variable.

Matthews Correlation Coefficient (MCC) is an alternative measure for multi-class classification problems, that is more suitable for dependent variables with ▷imbalanced classes (meaning that one class is more frequent than others). It is a correlation coefficient between the observed and predicted classifications. In the case of binary classification problem, it is equal with the ϕ statistic (Cramer 1946). In this work, the definition of MCC for multi-class classification problems as given by Jurman et al. (2012) and Gorodkin (2004) is used and shown in Equation 4.2.

$$MCC = \begin{cases} \sqrt{\phi^2}, & \text{for two class problems} \\ \frac{cov(X,Y)}{\sqrt{cov(X,X) * cov(Y,Y)}}, & \text{for n class problems} \end{cases} \tag{4.2}$$

MCC can take values between -1 and 1, where 1 corresponds to the perfect classification, -1 to the total disagreement between the predictions and real observations and 0 for the classification that is no better than random prediction. Following the above-mentioned literature, all $MCC \leq 0$ performance results shall be treated as random classification and therefore unreliable predictions. MCC lacks the easy interpretability of the accuracy measure (there is no natural benchmark, which makes it hard to judge which MCC value is a "good" one), but it is more robust and more suitable for the comparison between the classifiers.

Another restriction of the measure is that it is not defined for instances where the confusion matrix has zero observations in rows or columns. This is the case when a model does not predict a class label that is present in the test data or predicts a class label that was not available in the test data. It happens in the case of multi-class problems, imbalanced data, or when the model is overfitted to the data. In such cases and when a two-class problem is evaluated, an approximation can be used that was suggested by Burset and Guigó (1996) and Anderberg (1973)

$$\widehat{MCC} = \frac{1}{4}\left[\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right] \tag{4.3}$$

The approximation slightly overestimates the performance, but the deviation from MCC is supposed to be manageable (Burset and Guigó 1996).

4.2.2 Metrics for two-class problems

The in-depth analysis of the prediction performance on the level of single classes is often necessary. I present four performance metrics for classes in this section.

Precision expresses the amount of correct classified examples from all positively predicted ones. The measure is biased by the relative class size and it is therefore not recommended to compare the precision values of one class with another. The measure is also known as Positive Predictive Value (PPV). The number of FN and TN is not considered in this metric.

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

Recall expresses the amount of correct identified examples from all examples belonging to this class. It is also known as Sensitivity, or True Positive Rate (TPR). The number of FP and TN is disregarded.

$$Recall = \frac{TP}{TP + FN} \tag{4.5}$$

F-score Precision and recall are—taken for itself—only meaningful for model benchmarking, as the increase of one of this measure without consideration of the other one would lead to biased results. Consider the following extreme cases: 1) a dummy classifier that assigns one class to all examples would have $recall = 1$ for that class and a low precision, 2) a classifier that assigns only one example to the class of interest and this example is correct, would achieve a $precision = 1$, but a low recall, since the number of FN is large. The F-score combines both using a weighted harmonic mean and is robust against extreme small or large classes. The distribution of F_1 is illustrated in Figure 4.2 with the corresponding precision and recall values.

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{4.6}$$

$$F_\beta = \frac{(1 + \beta^2) * precision * recall}{(\beta^2 * precision) + recall}()/(),\ \beta \in \mathbb{R}^+ \tag{4.7}$$

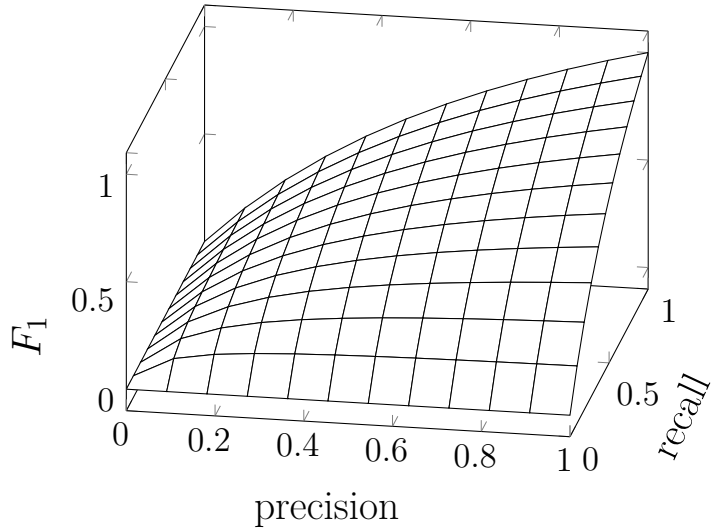


Figure 4.2: Illustration of the F_1 classification quality measure in relation to precision and recall

In its commonly used form F_1, precision and recall have equal weight. F_2 gives higher weight to recall than precision and $F_{0.5}$ vice versa. The measure is, however, class specific and its strength lies in the application of comparing different classification settings. The interpretation of single values is difficult, as a natural benchmark cannot be given.

Specificity This measure quantifies the number of true negatives from all actually negative examples and quantifies the avoiding of false negatives. The measure is also called True Negative Rate (TNR).

$$Specificity = \frac{TN}{FP + TN} \tag{4.8}$$

Area Under ROC Curve (AUC) This measure is based on the Receiver Operating Characteristic (ROC). ROC is a graphical illustration of the classifier performance for one single class and is created by plotting the TPR (Recall) against the fall-out (calculated by 1-specificity). The AUC is therefore an unbiased measure for the classification performance of one single class. AUC is a portion of the area of the unit square, and its value varies between 0 and 1. Because random guessing produces the diagonal line between (0, 0) and (1, 1), which has an AUC of 0.5, usable classifiers are expected to achieve values above 0.5 (Fawcett 2006). AUC is a proper metric for practical purposes because firms

are ultimately interested in recognizing specific customer groups for targeted measures. Some studies criticize the AUC measure (Hanczar et al. 2010; Hand 2009; Lobo et al. 2008). The main purport is that the AUC measure might be misleading for small sample sizes. Therefore, multiple performance measures should be reported in scientific reports.

4.2.3 Reference statistics for interpretation of performance metrics

The interpretation of classification performance results is challenging. As some performance metrics have no fixed reference value that indicate the performance of a random classification, two metrics are considered in this work are explained below.

Random Guess (RG) Without any knowledge on the class distribution within one property (for instance, how much percent of the customers are single-households), an equal class distributions can be assumed. For n classes within one property, the random guess metric is therefore:

$$RG = 1/n \tag{4.9}$$

Bias Random Guess (BRG) This metric was introduced by Beckel, Sadamori, and Santini (2013) and Beckel, Sadamori, Staake, et al. (2014) and is defined as the sum of the squared relative class sizes within one property. When h_k denotes the relative class size of the class k, the metric is defined as:

$$BRG = \sum_{k=1}^{K} h_k^2 \tag{4.10}$$

This metric equals to the Herfindahl Index that is used to measure the concentration in monopoly markets (Fahrmeir et al. 2007, p. 87). The properties of this index can be transferred to the BRG measure. So the extreme values for BRG are $BRG_{max} = 1$ and $BRG_{min} = 1/K$.

4.2.4 Calculation of performance measures

Two main approaches exist to calculate the performance measures: the holdout method and cross-validation. I briefly explain the approaches and refer the interested reader to Hastie, Tibshirani, and J. H. Friedman (2013, chapter 7) where a detailed discussion of the approaches is given.

Holdout The available training examples (features and class labels) are separated into *training* and *test* set. The ML algorithm is trained using the data in the train set and performance metrics are calculated based on the predicted data for the test set. Separation of training and test data is preferably done using a stratified split. This means that the random selection (without replacement) considers the class distribution.

Cross-validation The available data are separated into k disjoint subsets of equal size (so-called "folds"), preferably using stratified random sampling. With this approach, the sampling procedure takes the distribution of a target variable (e.g., the chosen dependent variable) into account and obtains random samples that have approximately the same ratio of observations per class in the drawn samples as in the target variable. The training and test is then repeated k times, each with a different fold as test and the remaining data as training set. Performance metrics are then calculated using the arithmetic mean and can be reported with a confidence interval. The procedure is illustrated in Figure 4.3.

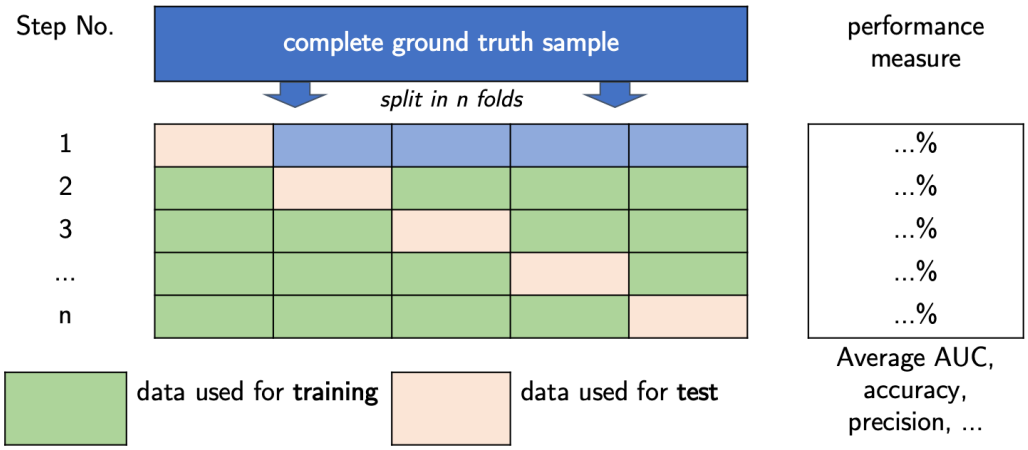


Figure 4.3: Illustration of the k-fold cross-validation

4.2.5 Comparison of performance metrics

The performance measures presented above have different strengths and weaknesses. There is no measure that satisfies all requirements in business analytics applications. Likewise, different performance measures are used in the literature. Table 4.2 gives an overview of the quality metrics presented so far, as well as their strengths and weaknesses.

Table 4.2: Comparison of classification performance metrics

Dimension	Accuracy	Precision	Recall	F_β	AUC	MCC
Multi-class support	yes	no	no	no	no	yes
Biased by class-distribution	yes	yes	no	yes	no	no
Interpretability	high	high	high	low	medium	low
Undefined cases	no	no	no	no	no	yes

This section gave an overview to classifier performance evaluation and raises no claim to be complete. The interested reader is referred to the comprehensive introduction to model evaluation by Hastie, Tibshirani, and J. H. Friedman (2013) and the comparison of classifier performance metrics of Sokolova and Lapalme (2009).

4.3 Description of selected supervised machine learning algorithms

Based on the review of existing classification methods, I selected six supervised ML algorithms from different algorithm classes and describe their functional principle briefly below. The sample of algorithms contains basic procedures that have often been used in previous works, but also state-of-the-art algorithms. I also summarize the most important parameters of each algorithm together with typical parameter values that can be used to "tune" the algorithms.

4.3.1 k Nearest Neighbors (kNN)

This lazy-learner infers the class-memberships by considering the k training instances with the lowest distance (e.g. Euclidean distance) to the example that has to be classified. Due to its sensitivity to outliers (Han et al. 2012), all features should be normalized to a range of $[0; 1]$. In my analysis, the implementation of Wing et al. (2015) is used and I tested various values for k that are listed in Table 4.3.

4.3.2 Naïve Bayes

Naïve Bayes is a Bayesian classifier that predicts the class membership based on a probability that a given data point belongs to the class. The probabilities needed for this prediction are calculated by means of the Bayes' theorem. In R,

I can recommend to use the implementation of Meyer et al. (2014) and identified no parameters that are reasonable to be included in a parameter tuning.

4.3.3 Support Vector Machine (SVM)

SVM was proposed by V. N. Vapnik and V. Vapnik (1998). The algorithm searches for a hyper plane in the vector space that separates all training examples with a maximal margin. In the case of not separable training data, a kernel-function is used that transforms the training vector into a higher dimension. Three kernels (polynomial, radial basis function, sigmoid) have been tested in 317 configurations (overview see Table 4.4), and we found that radial basis kernel having a coefficient of 50 and a cost of misclassification parameter of 50 leads to the best results (Sodenkamp, Hopf, Kozlovskiy, et al. 2016).

4.3.4 AdaBoost

The AdaBoost algorithm of Freund and Schapire (1997) was the first practical boosting algorithm that combines multiple weak learners (i.e., decision trees) to build a strong learner. This combination is done by weighting of the learned models (by the weak learners) in multiple iterations. In R, the implementation of Alfaro et al. (2013) can be recommended, since it is able to deal with multi-class problems. The parameters available for tuning are listed in Table 4.5.

4.3.5 Random Forest (RF)

This algorithm generates multiple low correlated decision trees that are learned and evaluated with ensemble methods (Breiman 2001). In R, the implementation of Hothorn et al. (2006) and Strobl et al. (2008) can be used which provides the parameters listed in Table 4.6 for tuning. Biau (2012) gives an introduction to Random Forest model tuning.

Table 4.3: Parameters of the kNN algorithm

Parameter	**Description**	**Possible range**	**Considered range**
k	Number of considered neighbors	All positive integers	1, 5, 15, 20, 50
distance metric	The distance metric to use	Euclidean, Manhattan, ...	Euclidean

Table 4.4: Parameters of the SVM algorithm

Parameter	Description	Possible range	Considered range
kernel	The kernel function	linear, radial, polynomial, sigmoid	
cost	The cost of misclassification parameter	Decimal	$2E^2(-5,\ldots,10)$
deg	The degree of the polynomial kernel	Decimal	2, 3, 4, 5
coef	A coefficient used in polynomial and sigmoid kernels	Decimal	0, 1, 5, 10, 100
gamma	A coefficient for polynomial, radial base and sigmoid kernel	Decimal	$\frac{1}{Num.features}$
eps	Paramer ϵ of the insensitive loss function	Decimal	0.1

Table 4.5: Parameters of the AdaBoost algorithm

Parameter	Description	Possible range	Considered range
coeflearn	The boosting algorithm	'Breiman','Freund','Zhu'	
mfinal	Number of iterations for which boosting is run or the number of trees to use	Integer	50, 100, 200

The RF classifier provides internal feature importance measures, that can be used to assess the predictive power of single features. The feature importance is measured as *mean decrease in accuracy* or *mean decrease in gini*[2]. A high feature importance score indicates a high contribution of the respective feature for the prediction performance of the model. It is important to mention that a score of zero does not mean a feature has no influence to the classification. A negative score does consequently not mean that the feature has a negative impact on the classification, because these values are just internal weights of the RF classifier and neither quantify the magnitude, nor the direction.

4.3.6 Extreme Gradient Boosting (XGB)

As an extension to the gradient tree boosting algorithm of J. H. Friedman (2001), Chen and Guestrin (2016) describe a scalable machine learning algorithm that builds multiple decision trees, by iteratively splitting the training data in smaller parts and aggregating the predictions of all base classification trees. The algorithm implements three techniques that avoid overfitting: a regularized learning

[2]The gini coefficient is a statistical measure of dispersion (i.e., inequality).

Table 4.6: Parameters of the RF algorithm

Parameter	**Description**	**Possible range**	**Considered range**
ntree	Number of trees to grow. This should not be set to too small a number, to ensure that every input row gets predicted at least a few times.	Integer	300*, 500, 1000, 2000
nodesize	Minimum size of terminal nodes. Setting this number larger causes smaller trees to be grown (and thus take less time).	Integer	1*, 10, 30
mtry	Number of features randomly sampled as candidates at each split.	Integer	Square root of number of features*, 20%, 50%, 70% of all features

objective that penalizes model complexity, tree shrinkage that limits the influence of each single tree, and feature subsampling, which means that only subsets of features are used to grow trees. Basically, there are three booster algorithms: Tree booster ("gbtree"), Dart booster ("dart"), Linear booster ("gblinear"). A further advantage of this algorithm is, that it explicitly can handle missing values. The number of XGB parameters is large (DMLC 2016b) and they are listed in Table 4.7. Some remarks from the developer to tune the XGB algorithm are available online (DMLC 2016a). For the interpretation of the model, XGB provides "Gain" as a feature importance score for each attribute, similarly to the feature importance in the RF model.

Table 4.7: Parameters of the XGB algorithm

Parameter	Description	Possible range	Considered range
booster	Which booster to use	gbtree, gblinear, dart	
nrounds	The max number of iterations	Integer	5, 10, 20
Parameters for tree booster			
eta	After each boosting step, one can directly get the weights of new features, and eta actually shrinks the feature weights to make the boosting process more conservative.	$[0, 1]$	0.001, 0.01, 0.05, 0.1, 0.3*
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger, the more conservative the algorithm will be.	$[0, \infty[$	0*

Parameter	Description	Possible range	Considered range
max_depth	Maximum depth of a tree, increase this value will make model more complex / likely to be overfitting.	Integer	3, 7, 10, 20
min_child_weight	Minimum sum of instance weight needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning. In linear regression mode, this simply corresponds to minimum number of instances needed to be in each node. The larger, the more conservative the algorithm will be.	$[0, \infty[$	1*
max_delta_step	Maximum delta step that is allowed each tree's weight estimation to be. If the value is set to 0, it means there is no constraint. If it is set to a positive value, it can help making the update step more conservative. Usually this parameter is not needed, but it might help in logistic regression when class is extremely imbalanced. Set it to value of 1-10 might help control the update	$[0, \infty[$	0*
subsample	Subsample ratio of the training instance. Setting it to 0.5 means that XGB randomly collected half of the data instances to grow trees and this will prevent overfitting.	$[0, 1]$	1*
colsample_bytree	Relative number of features that are used when constructing each tree.	$]0, 1]$	1*, 0.8, 0.6, 0.4
colsample_bylevel	Relative number of features that are used for each split, in each level.	$]0, 1]$	1*, 0.8, 0.6, 0.4
lambda	L2 regularization term on weights, increase this value will make model more conservative. When the parameter is set to 0, the optimization objective equals to the gradient tree bosting (Friedman, 2001)	$[0, \infty[$	1*
alpha	L1 regularization term on weights, increase this value will make model more conservative.	$[0, \infty[$	0*
tree_method	The tree construction algorithm used in XGB; 'auto' uses a heuristic to choose the faster one of 'exact' and 'approx'	'auto', 'exact', 'approx'	
sketch_eps	This is only used for tree_method='approx'	$]0, 1[$	0.03*, 0.05
scale_pos_weight	Control the balance of positive and negative weights, useful for unbalanced classes.	$[0, 1]$	0*
Additional parameters for dart booster			
sample_type	Type of sampling algorithm	'uniform', 'weighted'	

Parameter	Description	Possible range	Considered range
normalize_type	Type of normalization algorithm	'tree', 'forest'	
rate_drop	Dropout rate.	$[0, 1]$	0*
skip_drop	Probability of skip dropout; if a dropout is skipped, new trees are added in the same manner as 'gbtree' booster.	$[0, 1]$	0*
Parameters for linear booster			
alpha, lambda	See "Parameters for tree booster".	–	–
lambda_bias	L2 regularization term on bias, default 0 (no L1 reg on bias because it is not important)	$[0, \infty[$	0

4.4 Discussion and conclusion

This chapter gave an overview to the main approaches of supervised ML and performance evaluation of ML algorithms. Six classification algorithms were described (kNN, Naïve Bayes, SVM, AdaBoost, RF, and XGB) together with their main parameters. I identified ranges in which the parameters can be meaningfully varied, and described briefly what each parameter controls in the respective algorithm. The parameters with each identified range are a base for the systematic parameter search for best parameters. The ranges also help to perform a random selection of parameters, which was shown by Bergstra and Bengio (2012) to be equally effective.

Classifier evaluation is a difficult task and multiple performance measures exist. The choice of the best metric is dependent on the evaluation goal, the predicted variable and its distribution. When the classification results should, for example, be presented in front of persons that have no or less knowledge on ML, a metric should be chosen that is easy to interpret (e.g., accuracy or precision) and benchmark measures of random classification (e.g., random guess, biased random guess) should be shown. When models are to be presented in front of experts, rather unbiased metrics like MCC or AUC should be used. Following Demšar (2006), I call for the usage of cross-validation to calculate the metrics and reporting confidence intervals or test statistics when classifiers are compared.

5 Household classification for energy efficiency and personalized customer communication

Highlights

- ▷ It is possible to predict 18 out of 22 investigated characteristics of residential energy customers with data that are available to energy retailers.
- ▷ From the six selected machine learning algorithms, *Random Forest showed the best overall classification performance* and could be used successfully for a wide range of energy data analysis problems.
- ▷ In a benchmark, *five feature selection methods out of 43 provide reasonable results* in the investigated problem space and outperform the other 38 Feature Selection Method (FSM) in terms of classification performance and stability of the selected features sets, when the dataset for training is changed.
- ▷ Household classification is possible with 15-minute and daily smart meter data, as well as annual electricity consumption data; the geographic transferability of trained models was possible between Germany and Switzerland, as well as Switzerland and Ireland with manageable performance loss.
- ▷ The results enable personalized customer communication and render automatic energy consulting services possible.

Energy conservation is, despite its presence in the last decades of political discussion, still an important societal duty and will still be relevant in the future, considering the ongoing climate change and the limitation of fossil energy resources which are supposed to be exhausted in this century. Implemented mea-

sures to increase energy efficiency, like regulations towards efficient appliances, the prohibition of incandescent light bulbs, or incentives to conserve energy have not decreased the final energy consumption of the EU-28 countries (Eurostat 2015). Additionally, especially European states rely on fossil energy sources that are typically produced in foreign states. A decrease of energy consumption and the substitution of the remaining energy demand with renewables, that are generated within a country, has also a geostrategic motivation and increases independence in energy supply from other countries.

Private households account for a large share of the final energy demand[1], and residential homes hold a high energy saving potential considering the existence of old appliances, low insulation, or the customer behavior being a major cause of energy waste. Even without notable loss of comfort, some extent of this energy saving potential can be lifted (Balzer et al. 2015).

Concretely, the residential energy consumption can be lowered by targeted consumer feedback. Comparing consumers to similar households in the neighborhood (Allcott 2011), providing suitable energy saving goals (Loock et al. 2013), and focusing on particular target behaviors (Tiefenbeck, Goette, et al. 2016) are successful examples of energy efficiency campaigns. By looking at the costs per kWh energy saved, behavioral interventions are cheaper and achieve higher public acceptance than prohibitive regulations (Allcott and Mullainathan 2010). Likewise, studies from the area of marketing show that personalized messages are more likely to be recognized by the recipient (Wattal et al. 2011).

Specific behavioral interventions or personalized communication require detailed information about the receiver, for example, data on household characteristics or consumption behavior (Tiefenbeck 2017). Such data are typically not available to metering operators and utility companies. Obtaining household characteristics is therefore a major obstacle to implement feedback campaigns. Classical customer surveys suffer from high costs and low response rates (Groves 2006). The purchase of data from data brokers is expensive and the data quality is unclear. I argue that the relevant information can be extracted from data that are already present to utility companies. Several studies (Beckel, Sadamori, and Santini 2013; Beckel, Sadamori, Staake, et al. 2014; Wang et al. 2018) show that a decent number of household characteristics can be revealed from half-hourly electricity SMD using a large dataset from Ireland. In other studies, special household details are investigated. Fei et al. (2013) detect heat pumps from daily electricity consumption data and Verma et al. (2015) predict the existence of electric vehicles in households based on hourly electricity consumption data.

[1] In the EU-28 countries, private households account for 26.8 % (Eurostat 2015) and in Switzerland, even 30 % (Kemmler et al. 2015) of the final energy consumption.

Table 5.8 on page 138 compared the mentioned studies with this work. All studies investigate datasets outside central Europe (North America and Ireland) and no study has yet included geographic data in the prediction of household data.

The studies show that the recognition of selected household characteristics based on electricity consumption data with 30-min, hourly, or daily measurement intervals are feasible. In my dissertation, I move beyond this state-of-the-art and advanced the topic especially in the following areas: (1) Investigation of further household properties, (2) the application of previously untested machine learning algorithms for the problem, (3) test of additional data sources, (4) the selection and application of methods for automatic feature selection, and (5) test the model transferability. In this chapter, I compiled unpublished results out of my research to answer the following question. To obtain a complete picture on the topic, I partly cite results from papers that were published before, but mark this accordingly in the text.

RQ 3 *How well can (a) customer characteristics and (b) intentions be revealed from ambient data, in the context of energy retail, using state-of-the-art ML methods?*

In detail, I extended the approach of Beckel and colleagues to achieve a higher prediction accuracy with the same dataset by defining additional features from the load traces (Hopf, Sodenkamp, Kozlovskiy, and Staake 2014), adapted the algorithms to 15-min SMD added features on weather data as well as geographic information, and validated the methods with data raised in Switzerland (Hopf, Sodenkamp, and Staake 2018). In addition, I investigated the recognition of households characteristics with strong impact on the overall electricity consumption (household type, number of residents, or presence of electric heating systems) based on annual consumption data, governmental statistical data, and geographic information from OSM (Hopf, Sodenkamp, and Kozlovskiy 2016; Hopf, Riechel, et al. 2017; Hopf 2018).

The household classification approach proposed in this work is illustrated in Figure 5.1, The figure shows how all analytical steps that have been described in the previous chapters interrelate. The feature extraction was described in section 3.3 and serves as a data preparation, data integration and dimensionality reduction step. Feature selection is a further step to reduce the data dimension and select relevant variables for the prediction and was introduced in section 3.4. Supervised ML (classification) algorithms, as introduced in the previous chapter, detect pattern in the data based on ground truth information (in my work obtained through customer surveys) and create models that can be used predict the household classes for new instances.

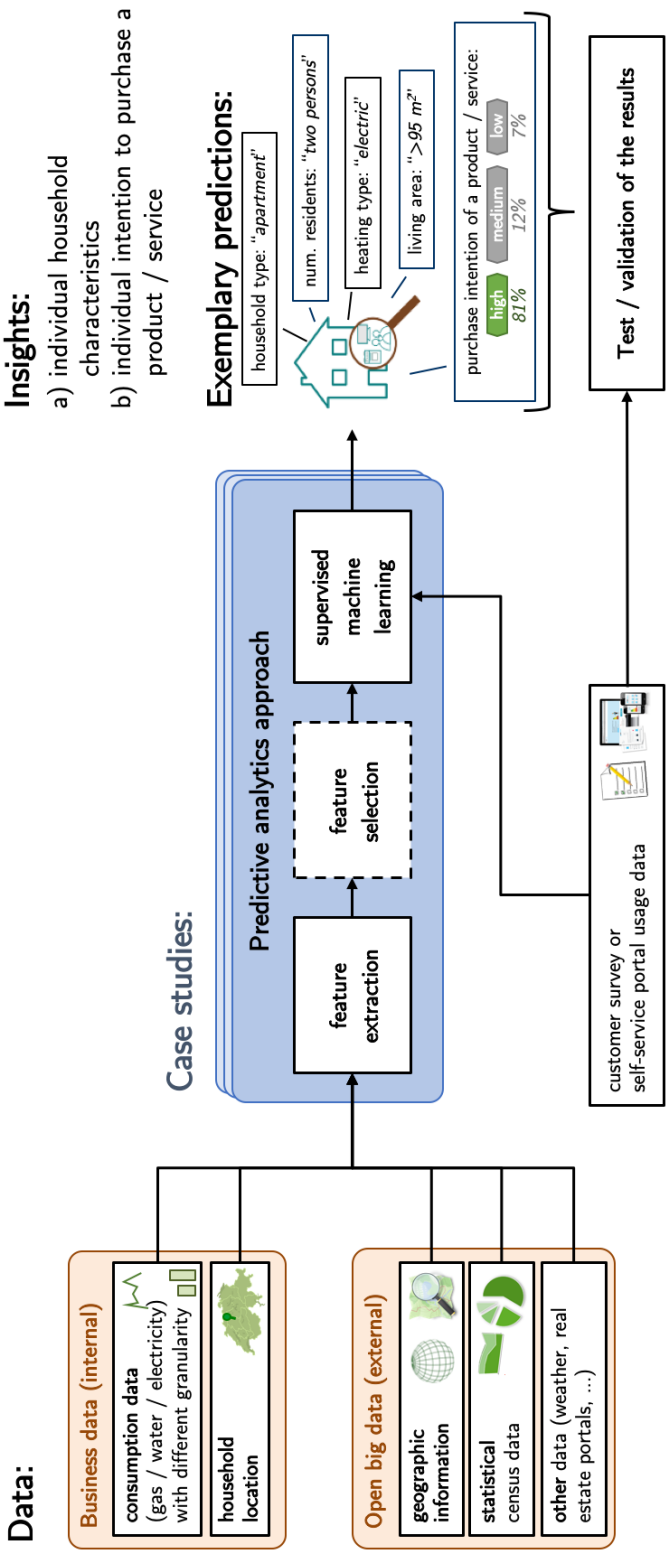


Figure 5.1: Illustration of the household classification approach used in this work

The classification quality of the resulting models is calculated using well-known performance metrics described in section 4.2. An essential part of the predictive analytics approach is the selection of relevant features from a large set of available ones. After having introduced 43 such methods in section 3.4, I present a benchmark of FSMs and provide an additional contribution to answer RQ 2 on whether human cognition, theory, and human expert knowledge become irrelevant when many automatic methods for ML exist to reveal pattern automatically from data. I will therefore discuss the trade-off between algorithmic vs. theory-driven dimensionality reduction with the empirical results from the benchmark of FSMs in the end of this chapter.

The remainder of this chapter is organized as follows: First, I describe the datasets available for the investigation in this chapter. The datasets will also be used in the studies described in the later chapters. Thereafter, results of the household classification with different consumption data granularity (annual, daily, and smart meter data) together with external data are described. Finally, the geographic transferability of trained models is tested. With the empirical results, I give answers to RQ 2 as well as RQ 3, and discuss the implications of the results for research and practice in the concluding section.

5.1 Datasets with residential energy consumption, household location and survey data

Four comprehensive datasets on residential energy customers were acquired together with utility companies and used in this dissertation research project. The datasets contain energy consumption data together with the address of the location where the energy was consumed. Through web portals that engage private customers towards energy efficiency and customer surveys, additional information regarding the households were obtained. An overview of the key information of the four datasets is presented in Table 5.1. Raising new datasets was a necessary effort, because public available datasets that have been used in earlier studies contain only a limited number of household details, cover data on households outside central Europe[2] and have no location information available due to privacy reasons.

The first three datasets (A, B, C) have been raised together with the industry partner BEN Energy AG (Switzerland). As part of its product portfolio, the company offers energy efficiency online platforms to utility companies in Europe.

[2]Beckel, Sadamori, Staake, et al. (2014) and Wang et al. (2018) use a dataset from Ireland and A. Albert and Rajagopal (2013) use a dataset from North America.

Table 5.1: Energy consumption datasets used in this dissertation

	Dataset			
Characteristics	**A**	**B**	**C**	**D**
Empirical context				
Location	Switzerland	Germany	Switzerland	Switzerland
City size (approx.)	(multiple)	89,000	25,000	14,000
Utility companies	6	1	1	1
Energy consumption data				
Energy source	electricity	electricity	electricity, gas, water	electricity
Data resolution	annual	annual	daily	15-min SMD
Time span	2009–2014	2009–2016	2013–2018	2014–2015
Ground truth data on households				
Data source	portal data	portal data	portal data	survey
Survey responses	5,446	2,058	567	451

Technically, the platform-as-a-service offering is instantiated and branded for each utility company client. The platforms use story-telling around a mascot (e.g., an energy hero) and aim to engage residential customers for energy literacy and energy efficiency by means of energy usage feedback, household efficiency check, saving tips and gamification elements. According to BEN Energy, $5-15\%$ of the residential customers use these portals when the utility is offering the platform. From earlier research on the energy saving effect of such customer engagement platforms, it is known that especially consumers with an energy demand above the average register and use them (Lossin 2016).

During registration or through an "efficiency check" (a functionality where users can compare themselves to similar households), customers are asked to enter several household details to enable the energy consumption benchmark. An example questionnaire to obtain household characteristics is depicted in Figure 5.2. The five basic properties (household type, living area, number of residents, water and space heating type) are required information necessary to use the portal. Depending on the portal instance, additional information are obtained from users through single-question surveys (framed as "introductory round"). A full list of survey variables raised in the portals is listed in Table 5.2.

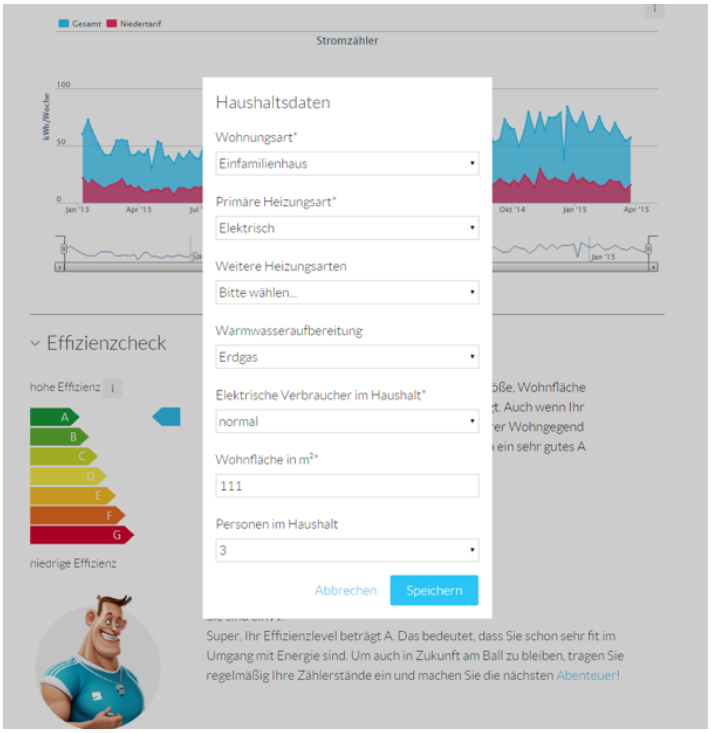


Figure 5.2: Efficiency check in the energy efficiency web portals to acquire household characteristics for classification (Source: BEN Energy)

5.1.1 Dataset A and B: Annual electricity consumption data

The datasets A and B have a similar structure, their only distinction is that data in A stem from six utility companies in Switzerland and data in B stem from a German utility. For each household, the total annual electricity consumption for several years together with the number of days in which the electricity was consumed is given. In addition, the customer address (street and street number, postal code and city name) is known.

All customers in the sample are users of an energy efficiency portal and gave their consent for data analysis for research purposes. During registration or by completing the "efficiency check", they provided information on five household properties: household type, living area, number of residents, water and space heating type (see Table 5.2, questions 1, 2, 3, 4, and 6). No additional data on the portal usage are available in both datasets. The users of the efficiency portals have either been attracted by an advertisement from the utility company, have been invited by e-mail, or by a physical mailing that contained an eye-catching motivation related to energy efficiency (e.g., households have been benchmarked with their neighborhood in terms of electricity consumption). Lossin (2016) de-

scribes and analyzes the effectiveness of different motivational measures on such portal activation mailing campaigns with one of BEN Energy's portal instances.

Table 5.2: Survey variables with respective scale type (N = numeric, C = categorical) asked in BEN Energy's energy efficiency portals

	Questions	Answer possibilities	Type
Basic household properties from "introductory round" (all energy-efficiency portals)			
1	Household type	Apartment, house	C
2	Living area	Integer $1, ..., 1000$ (in square meter)	N
3	Household members	Integer $1, ..., 9$	N
4	Primary space heating type	Electric, natural gas, heating oil, electric storage heating, central heating (multiple family house), district heating, air heat pump, ground heat pump, solar thermal, pellets, piece of wood	C
5	Additional space heating type	(see question 5)	C
6	Water heating type	(see question 5)	C
7	Number of devices	1, 5, 10, 15, 20	C
Introductory round (the questions are embedded in a narrative)			
8	How well do you know your meter? (How often do you look at your meter?)	1) I've never seen it before, 2) We're casual acquaintances, 3) We see each other regularly, 4) I even keep a diary about it.	C
9	Does your meter have children? (Have you bought or borrowed energy cost meters or smart meters?)	Yes, no	C
10	Does your meter have a productive (girl-)friend? (Do you have solar cells on the roof?)	Yes, no	C
11	Why do you use the energy efficiency portal?	Because 1) I want to collect efficiency points, 2) I want to know how much energy I use, 3) I want to know how well I compare to others, 4) I want to do something good for the environment, 5) I enjoy such actions, 6) I can save money that way, 7) my family / friends also take part	C
12	How did you hear about the energy efficiency portal?	1) Through poster advertising, 2) Through the Internet, 3) Through friends and acquaintances, 4) Television/Newspaper/Radio, 5) By mail	C
13	Is your meter on a diet? (Have you already bought energy saving light bulbs?)	Yes, no	C

Questions *(continued)*		Answer possibilities	Type
14	Does your meter have professional support? (Have you ever taken advantage of an energy consultation?)	Yes, no	C
15	How strongly are you interested in sustainability?	1) Not at all, 2) A little, 3) Medium, 4) Strong, 5) Very strong	C
16	Are you male or female?	male, female	C
17	How old are you?	younger than 18, 18 - 20, 21 - 30, 31 - 40, 41 - 50, 51 - 60, 61 - 70, older than 70 years	C
18	[Mascot name] has an energy saving diploma. What is your education?	1) Compulsory schooling, 2) vocational training, 3) higher vocational training, 4) university of applied sciences or university	C
19	[Mascot name] works all day long in the energy-saving world. What about you?	1) I work full time, 2) I work part-time, 3) I'm unemployed	C
20	Since [mascot name] is in the energy efficiency portal all day long, he does not spend much time in the household. How much time do you spend in your household?	1) Very little, 2) Little, 3) Medium, 4) Much, 5) A lot	C

5.1.2 Dataset C: Daily smart meter data

Dataset C is a comprehensive dataset containing energy consumption data (electricity, gas, and water), household address data, household property information (see Table 5.2, question $1-7$), detailed usage data on the energy efficiency portal (times of logins, page visits, used functionalities, earned bonus points), and additionally asked survey-responses from questions in the efficiency portal (see Table 5.2, question $8-20$).

All customers in the dataset are likewise users of the energy efficiency portal and gave their consent for data analysis for research purposes. In addition to the portal data, a survey was conducted to which all customers that have been registered on the portal were invited. The survey stood in relation to an experiment that is described in chapter 6.

5.1.3 Dataset D: 15-min smart meter and survey dataset

The fourth dataset stems from a project together with a utility in a German-speaking municipality in Switzerland with about 9,000 customers and contains household electricity smart meter readings at 15-min granularity in the time span June 1, 2014 to May 31, 2015. Details on the dataset and the empirical context

are described in Sodenkamp, Hopf, Kozlovskiy, et al. (2016). We conducted a web-based customer survey on household characteristics, heating, photovoltaic installations, completed energy-efficiency measures, satisfaction with the utility company, and purchase intention for a Fiber-to-the-Home (FTTH) offer. Invitations to the survey were sent together with the bimonthly bill to all household customers between June and September 2015. As the utility serves as the only energy supplier in the municipality, one can assume that all households received the survey invitation. 541 households participated in the survey which corresponds to a response rate of 6%.

The survey questions are listed in Table 5.3. Some questions have been taken from the "residential pre-trial survey" of the Irish Commission for Energy Regulation (2011) smart meter study to provide comparability with earlier works, as this dataset has been used in earlier studies on household classification. For these questions, I list the question number beginning with *CER* in the reference column.

Personal attitudes of the respondent were asked through constructs known from literature. I list the respective measurement instruments together with the considered German translation in Appendix C and summarize them briefly below. The purchase intention towards a photovoltaic system was asked in question 21 through a three item measure *PI1–PI3* (see Table C.3 on p.221) following H.-W. Kim et al. (2007). Attitudes towards energy efficiency (questions 26) follow the factor "energy conservation" of the behavior-based attitude scale for environmental attitude of Kaiser et al. (2007) together with a question of a Swiss environmental study (Diekmann and Bruderer Enzler 2012; Diekmann and Franzen 1999) and constitute a seven item scale of *EA1–EA7* (see Table C.1 on p.219). The customer satisfaction was raised in question 27 through four items *REP5, REP6, REP7, REP3* (see Table C.2 on p.220) of the customer-based reputation of a service firm scale (Walsh, Beatty, and Shiu 2009; Walsh and Beatty 2007). Finally, the purchase intention towards FTTH was raised with the scale of the purchase intention and purchase probability following Juster (1966) with some adaptions in the German translation (see Table C.4 on p.222). The survey was developed and tested together with colleagues and feedback from the partnering utility company was considered.

The survey questions in this datasets, and those presented before, will be used in the remaining work as the dependent variables for predictive analytics.

Table 5.3: Questions and answer possibilities to the customer survey related to dataset D with answer types (N = numeric, C = categorical, T = free text, L = logical) and references to survey items of measurement instruments

Questions		Answer possibilities	Type	Reference
Page 1: General				
1	Acceptance of the privacy declaration	Yes, no (survey finishes with answer "No")	L	
2-7	Name and address fields		T	
Page 2: Household members				
8	Number of persons in the household (without children of 16 years or younger)	1, 2, 3, 4, 5 or more	C	CER420
9	Number of children (16 years or younger)	0, 1, 2, 3, 4, 5 or more	C	CER43111
10	How many persons are typically at home during the day (e.g., 5-6 hours per day)	0, 1, 2, 3, 4, 5 or more	C	CER430, CER4312
Page 3: Household characteristics				
11	How do you live?	Apartment in a multifamily house, semi-detached house, single-family house (detached), terraced house	C	CER450
12	Living area in m^2 (estimate, if unknown)		N	CER6103
13	Ownership	Rent, own	C	CER542
14	Do you know the construction year of your house?	Yes, no	L	CER453
14a	Construction year of the building (if question 14 is "No")	Newer than 5, 10, 30, 75, older than 75 years	C	
14b	Construction year of the building (exact or estimated) (if question 14 is "Yes")		N	
Page 4: Heating and appliances in the house				
15	Type of heating	(separated by main heating, secondary heating, water heating) natural gas, heating oil, electric storage heating, central heating (multiple family house), district heating, heat pump, solar thermal, pellets, piece of wood, other / none	C	CER470, CER471
16	In which year was your heating installed	(separated by main heating, secondary heating, water heating)	N	
17a	How many of the following devices are present in your household: (electric cooker, fridge, separate freezer, washing machine, tumbler, dishwasher, TV)	0, 1, 2, 3 or more	C	CER4704, CER490, CER49001, CER49002, CER4901
17b	Estimate the age of the oldest appliance for each category	2, 5, 7, 10, 15, older than 15 years	C	
Page 5: Solar and geo-thermal potential				
18a	Please indicate, if you have one of the following installations: photovoltaic system, solar thermal system, air heat pump, geothermal heat pump	Yes, no	C	

Questions *(continued)*		Answer possibilities	Type	Reference
18b	If yes, when was the system installed?	Year of installation	N	
19	How strongly is your roof inclined? (With illustrations)	$0-40°, 40-70°, 70-90°$	C	
20	Does one side of your roof face the midday sun? (With illustrations)	North/south, south-west/south-east, east/west, unknown	C	
21	I could imagine buying a solar system in the next 1-2 years	7-point Likert scale	C	PI1
	I intend to purchase a solar system in the next 1-2 years	7-point Likert scale	C	PI2
	I plan to purchase a solar system in the next 1-2 years	7-point Likert scale	C	PI3
Page 6: Energy efficiency				
22	Share of energy saving light bulbs (estimated)	100%, 75%, 50%, 25%, *none*	C	CER4905
23	Percentage of double or triple glazed windows (estimated)	100%, 75%, 50%, 25%, *none*	C	CER4906
24	Which of the following energy-saving measures have been realized in the past 15 years? (Insulation of roof or upper floor, building insulation, cellar insulation, window replacement, none, do not know)	Yes, no	C	
Page 7: Energy efficiency				
25	After one day of use, my sweaters or trousers go into the laundry*	7-point Likert scale	C	EA1
	As the last person to leave a room, I switch off the lights	7-point Likert scale	C	EA2
	I leave electrically powered appliances (TV, stereo, printer) on standby*	7-point Likert scale	C	EA3
	In the winter, I turn down the heat when I leave my room for more than 4 hours	7-point Likert scale	C	EA4
	In the winter, it is warm enough in my room to only wear a T-shirt*	7-point Likert scale	C	EA5
	In hotels, I have the towels changed daily*	7-point Likert scale	C	EA6
	I do what is right for the environment, even when it costs more money or takes more time.	7-point Likert scale	C	EA7
26	My household saves energy to help the environment	7-point Likert scale	C	CER4331
	My household saves energy to save money	7-point Likert scale	C	CER4331
	My household saves energy to set a good example for children.	7-point Likert scale	C	
	My household saves energy because we generally live economically.	7-point Likert scale	C	
	My household saves energy because it is so common in the family or among friends.	7-point Likert scale	C	
	My household is interested in new technologies.	7-point Likert scale	C	
Page 8: Satisfaction with the utility company and interest in Fiber-to-the-Home (FTTH)				
27	The utility company offers high quality products and services	7-point Likert scale	C	REP5
	The utility company is a strong, reliable company	7-point Likert scale	C	REP6

Questions *(continued)*		Answer possibilities	Type	Reference
	The utility company develops innovative services	7-point Likert scale	C	REP7
	The utility company is concerned about its customers	7-point Likert scale	C	REP3
28	How do you estimate the prospects that you will buy FTTH within the next 12 months?	10-point Juster (1966) scale	C	
29	Remarks		T	

5.2 Household properties (dependent variables)

This work investigates the predictability of several household characteristics (also called "household properties"), based on ambient data available to energy retailers. Hereinafter, I define the set of dependent variables that are used consistently in the remainder of this work. This definition, as it is listed in Table 5.4, follows earlier household classification studies based on smart meter data (Beckel, Sadamori, Staake, et al. 2014), and annual electricity consumption data (Hopf, Sodenkamp, and Kozlovskiy 2016). The table shows the definition of dependent variables and the respective descriptive statistics in the four datasets, if the variable is available in the respective dataset.

Numeric variables have been converted to categorical variables because of multiple reasons. First, a majority of machine learning algorithms are designed to do classification (the prediction of a limited number of classes) rather than regression (the prediction of a numeric value). Second, Beckel, Sadamori, Staake, et al. (2014) investigated how well regression can be used to predict the number of residents, appliances, bedrooms, the age of residents, and the floor area of households based on SMD and conclude that "utilities should rely on the estimated class rather than striving for exact, continuous values". The benefit of concrete predicted values is limited because of a high uncertainty and in most application cases, the identification of one specific class is relevant (e.g., large or small dwellings, single-households). Third, survey questions where respondents estimate numeric values one can assume that guesses of unknown exact values (e.g., living space area, age of something), tend to be round numbers which leads to a distribution of the variables that is not realistic. Thus, binning continuous values helps also to overcome this survey data issue. In Hopf, Sodenkamp, and Kozlovskiy (2016), we discuss the binning of continuous values into categorical dependent variables in detail considering three examples:

> *pLivingArea* The variable living area takes integer values in the range of 10 to 5,443. Therefore, any definition of this property is ambiguous. We defined the class borders at 95m^2 and 145m^2 based on the following motivation:

First, the class borders are empirically defined and based on quantiles. The 33% quantile is 100m^2, the 66% quantile is 150m^2, and the 99% quantile is 400m^2. Since we assume that people estimate their living area in a survey to the next upper bound, we define the categories 5m^2 below this round number. Second, we find further evidence in our class definition in European statistics (Statistical Office of the European Communities 2014, p. 54): the average dwelling size in the EU-28 countries is 95.9m^2, in Switzerland it is according to the statistics 117.1m^2.

pNumResidents The number of residents in a household takes fewer values than the living area, but the variable has nevertheless a range of 1 to 10 household and the class borders can be defined ambiguously. We tested a set of definitions in the classification: a) 1 / 2 / $>$ 2, b) 1 / 2 / 3–5 / $>$ 5, c) 1 / 2 / 3 / 4 / $>$ 4, d) 1 / $>$ 1. Our results show that the definition (b) has the best trade-off between gained information, number of classes and classification performance. Therefore, we include only this definition in this paper.

Table 5.4: Household properties as dependent variables in predictive analytics: definition of the classes and descriptive statistics for the four datasets used in this work; when no statistics are listed, the variable is not existent in the dataset; the asterisk (*) marks that the property has been investigated in earlier studies

Property	Classes	Definition	Statistics A		Statistics B		Statistics C		Statistics D	
			n	%	n	%	n	%	n	%
Age of appliances	New	Avg. appliance age $< q_{0.25}$							153	33.41
	Average	Avg. appliance age between$q_{0.25}$ and $q_{0.75}$							152	33.19
	Old	Avg. appliance age $> q_{0.75}$							153	33.41
Num. of appliances*	Few	Number appliances $< q_{0.25}$ for dataset D, < 15 for C					458	87.60	149	28.27
	Average	Number appliances between $q_{0.25}$ and $q_{0.75}$							280	53.13
	Many	Number appliances $> q_{0.75}$ for dataset D, ≥ 15 for C					65	12.40	98	18.60
Efficiency measures	No	Number completed energy efficiency							304	57.69
	Few	measures during the last 15 years							109	20,68
	Multiple	(insulation of basement/roof/building							114	21.63
		envelop, or window replacement)								
Age of residency	< 10								71	13.65
	$10 - 29$	Age (in years) of the building the							147	28.27
	$30 - 74$	household is living in							219	42.12
	≥ 75								83	15.96
Cooking type*	Electric	Number electric stoves > 0							484	91.84
	Not electric	Number electric stoves $= 0$							43	8.16
Heat pump*	No	No heat pump existing							453	85.96
	Yes	Heat pump existing							74	14.04
Solar installation	Yes	Photovoltaic or solar heating existent					17	9.39	29	5.50
	No	Neither photovoltaic nor solar heating existent					164	90.61	498	94.50
Interest in solar	Low	Purchase intention coefficient $< q_{0.5}$							387	73.43
	Average	Purchase intention coefficient between $q_{0.5}$ and $q_{0.75}$							49	9.30
	High	Purchase intention coefficient $> q_{0.75}$							91	17.27

Property	Classes	Definition	Statistics A		Statistics B		Statistics C		Statistics D	
(continued)			n	%	n	%	n	%	n	%
Space heating type	Electric	Space heating = "Electric heating"	781	15.65	39	2.12	32	11.52	21	3.98
	Heat pump	Space heating = "Heat pump"							66	12.52
	Gas	Space heating = "Gas"					155	55.14		
	Other	Other space heating	4,210	84.35	1,801	97.88	94	33.34	440	83.49
Water heating type	Electric	Water heating = "Electric heating"	2,431	50.03	187	10.49			81	15.37
	Heat pump	Water heating = "Heat pump"							63	11.95
	Other	Other water heating	2,428	49.97	1,596	89.51			383	72.68
Age of heating	New	Space heating age $< q_{\frac{1}{3}}$							128	33.51
	Average	Space heating age between $q_{\frac{1}{3}}$ and $q_{\frac{2}{3}}$							135	35.34
	Old	Space heating age $> q_{\frac{1}{3}}$							119	31.15
Number of residents*	1	Number of persons (adults and children)	612	12.26	354	19.23	24	8.66	121	23.05
	2		2,004	40.14	653	35.47	123	43.76	238	46.33
	$3-5$		2,251	45.09	808	43.89	124	43.97	150	28.57
	> 5		125	2.50	26	1.41	10	3.61	16	3.05
Single*	Yes	Adults = 1 and children = 0	612	12.26	354	19.23	24	8.66	121	23.05
	No	Otherwise	4,380	87.74	1,487	80.77	257	91.34	404	76.95
Children*	Yes	Number of children > 0							69	13.09
	No	Number of children = 0							458	86.91
Family*	Yes	> 1 adults and > 0 children							111	21.06
	No	Otherwise							416	78.94
Household type	Apartment	Apartment in a multi-family home	2,228	44.38	863	46.88	119	42.20	323	61.88
	House	Otherwise	2,798	55.62	978	53.12	162	57.80	199	38.12
Home ownership	Own	The property is owned							280	53.95
	Rent	The property is rented							239	46.05
Living space area**	$\leq 95\text{m}^2$	Living area $\leq 95\text{m}^2$	1,059	21.10	623	33.80	46	16.62	166	32.49
	$\leq 145\text{m}^2$	Living area $> 95\text{m}^2$ and $\leq 145\text{m}^2$	1,852	36.91	748	40.59	87	31.44	180	35.23
	$> 145\text{m}^2$	Living area $> 145\text{m}^2$	2,107	41.99	472	25.61	144	51.94	165	32.29
Satisfaction with the utility	Low	Reputation coefficient $< q_{0.25}$							127	25.45
	Medium	Reputation coefficient between $q_{0.25}$ and $q_{0.75}$							261	52.30
	High	Reputation coefficient $> q_{0.75}$							111	22.24

Property	Classes	Definition	Statistics A		Statistics B		Statistics C		Statistics D	
(continued)			n	%	n	%	n	%	n	%
Interest in sustainability	Low	Coefficient $< q_{0.5}$					97	59.64		
	High	Coefficient $\geq q_{0.5}$					66	40.36		
Energy saving light bulbs	Yes	Energy saving light bulbs used					145	86.28		
	No	Energy saving light bulbs not used					23	13.72		
Employment	Yes	Employed or part-time employed					125	78.34		
	No	No employment					35	21.66		

5.3 Performance of machine learning algorithms in smart meter household classification

As the first analysis, I compare the prediction accuracy of seven machine learning algorithms on the 19 household properties defined in Table 5.4 for dataset D. For this analysis, no automatic feature selection is used, as it adds additional complexity and makes it hard to determine the quality of the different machine learning algorithms. As predictor variables, 312 features are available. These encompass the features on SMD (see section 3.3.1), on the correlation between electricity consumption and weather data (see subsection 3.3.2), and on geographic information from OSM (see subsection 3.3.3).

As classifiers, I considered for this analysis: logistic regression (multinomial logistic regression[3] for variables with more than two classes), kNN[4] with $k = 20$, SVM[5] with the standard parameters $\epsilon = 0.1$, $cost = 1.0$and a radial basis kernel with $\gamma = \frac{1}{308}$, Random Forest (RF)[6], AdaBoost[7] and XGB[8], each with 500 trees.

For the SMD and weather features, one exemplary week (January 12–18, 2015) without school holidays or special events was selected. A detailed analysis on the seasonal impact of electricity and weather data on the household classification performance in Hopf, Sodenkamp, and Staake (2018), we found that the time of year has an influence on the classification performance, but the results do not change dramatically. All numeric features were scaled to have a mean of 0 and a standard deviation of 1 in the training sample, the examples in the test sample were scaled with the same factors. For the SVM, kNN, and XGB algorithms (which have no direct support for categorical data in the used implementation), four features (specifying the land use type and the type of building next to the household location) have been excluded.

Classification performance is measured in accuracy, as this metric is good to interpret and can be easily compared to random guess and biased random guess metrics. The performance is calculated using 4-fold cross-validation and the folds were created applying stratified random sampling. The reason for choosing four folds was that several variables have infrequent classes (e.g., only 16 households with more than five people, or only 43 households with an electric cooking place) and I wanted to ensure that there were enough training examples

[3] using the implementation of the R "nnet" package (version 7.3-12)
[4] using the R package "class" (version 7.3-14)
[5] using the R package "e1071" (version 1.6-8)
[6] using the R package "randomForest" (version 4.6-12)
[7] using the R package "adabag" (version 4.1)
[8] using the R package "xgboost" (version 0.6-4)

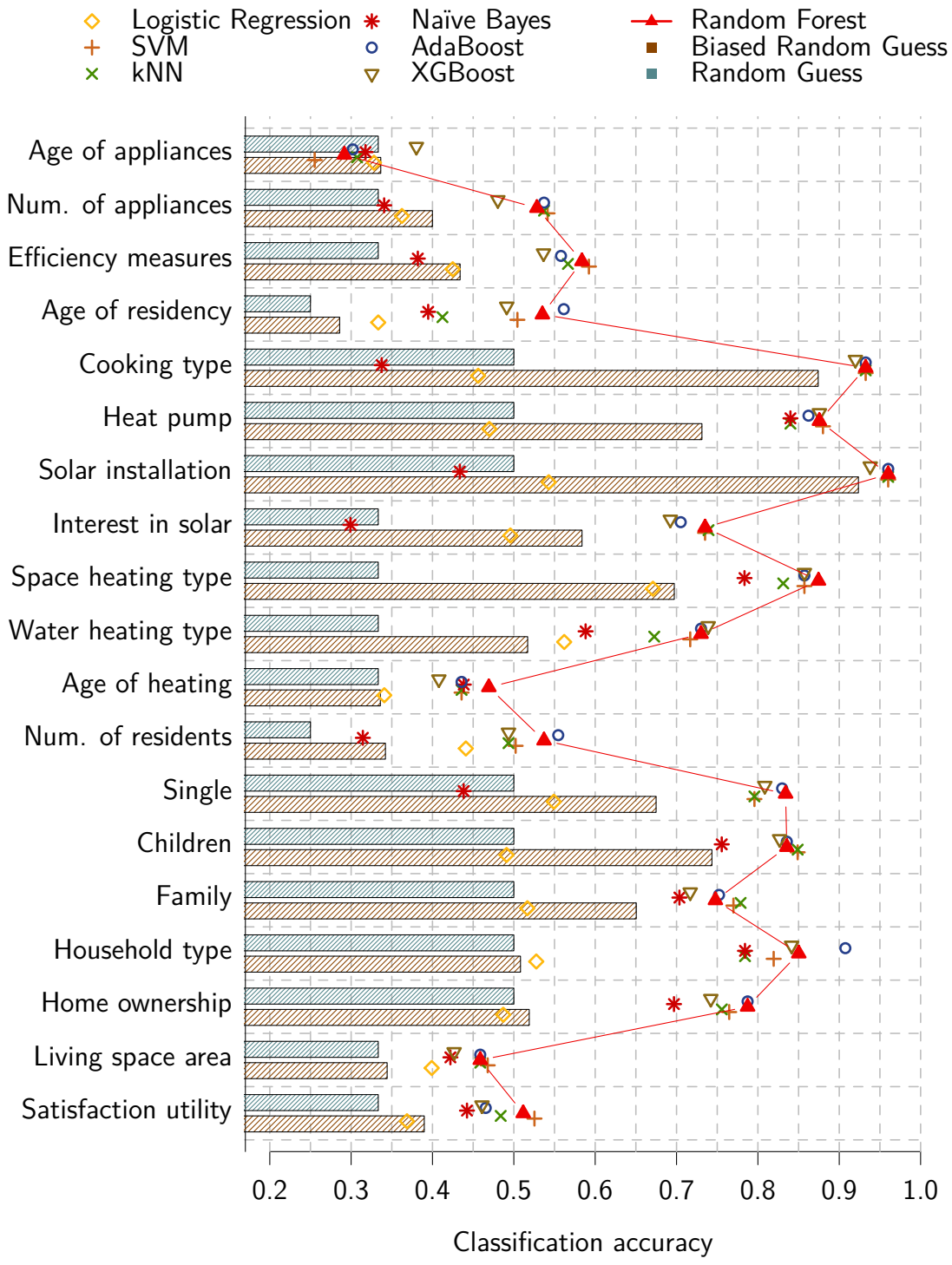


Figure 5.3: Classification accuracy for 19 household properties in dataset D based on seven classifiers with standard parameters; Random Forest achieved the best overall results

from all classes in each fold. Figure 5.3 shows the classification results of the six considered classification algorithms for the 19 considered household properties. In the experiment, RF outperformed all algorithms except AdaBoost in terms of the average classification accuracy of $M = 0.6846$ with $SD = 0.1891$ (paired $t(18) > 1.7903, p < 0.05$). The difference in classification accuracy is, however, only notable for logistic regression and Naïve Bayes (Cohen's $d > 1$). The effect size between the performance of RF and the other classification algorithms is small ($d < 0.2$).

As the SVM classifier has many parameters to vary, I computed a sensitivity analysis and varied the parameters of this algorithm in meaningful bandwidths: For the *cost* parameter $10^a, a \in \{-4, \ldots, 15\}$, was considered, "linear", "radial basis", "polynomial" and "sigmoid" kernels were tested, and depending on the used kernel, the parameters $degree \in \{2, 3, 4, 5\}$, $coef \in \{0, 1, 5, 10, 100\}$ and $\gamma \in 2^a, a \in \{-15, \ldots, 3\}$, were used. All accuracy values are calculated using 4-fold cross-validation.

Figure 5.4 shows the result of this sensitivity analysis. The standard configuration of SVM (see description above) is indicated with red marks in the graphic. From the analysis, I conclude that the SVM classification performance can be improved through parameter tuning, but the standard parameters already provide a good starting point for the investigated problems.

To conclude, I can support Fernández-Delgado et al. (2014) and conclude that RF works well in the investigated problem class of predicting household properties based on energy consumption data. The algorithm can deal with a large feature set of 312 potentially expressive features without automatic feature selection and produce good results on average. Consequently, I use the RF classification algorithm in all analysis where the predictability of household properties is investigated.

5.4 Benchmark of FSMs for smart meter household classification

Feature selection is an advantageous intermediate step between feature extraction and machine learning. In section 3.4 on page 74, I introduced feature selection as an algorithmic step to reduce the complexity of the prediction task. I presented the three general approaches to feature selection (wrapper, filter, and embedded methods) and gave an overview of 43 FSMs that are available in the statistical programming environment R.

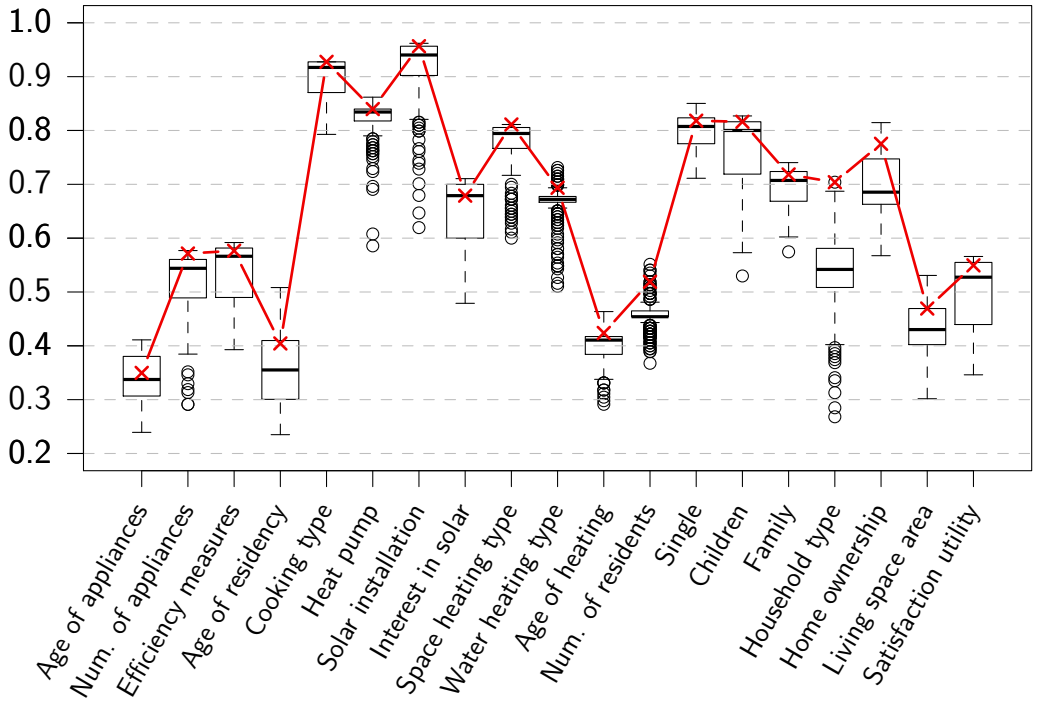


Figure 5.4: Boxplots of classification accuracy of 209 different SVM parameter configurations for the 19 household properties in dataset D (the standard configuration is highlighted in red)

Despite several studies comparing selected algorithms for feature selection, there is to the best of my knowledge no comprehensive benchmark of such methods in the problem area of energy data analysis. In this section, I therefore perform such a benchmark of the identified methods and first review related studies, then describe quality criteria of FSMs, and present the results accordingly.

5.4.1 Earlier studies comparing FSMs

Several works compare the effectiveness of FSMs. Early works (Kudo and Sklansky 2000; Reunanen 2003) compare wrapper methods on the basis of different datasets. The comparisons in terms of classification accuracy show that no single method is superior to others. Two recent studies affirmed this finding by investigating ten (Hua et al. 2009) four (Chandrashekar and Sahin 2014) wrapper as well as filter methods using multiple datasets from the field of Bioinformatics. A major drawback of these studies is that the FSM benchmarking only considers the classification performance as a quality criterion.

More carefully designed studies (Haury et al. 2011; Saeys, Abeel, et al. 2008) include also the stability of FSMs, which means that the selected feature sets are similar when slightly varied training and test data are used, and found that filter methods can outperform more complex wrapper or embedded methods. Furthermore, Lo et al. (2016) describe a method to estimate the predictive power of feature sets independently of the classifier. For practical use, this approach is limited, as features are required to be categorical with up to three categories.

Existing studies consider only a low number of FSMs in limited problem spaces. Mostly problems from the field of Bioinformatics (microarray, mass spectrometry data to predict diseases) are used that lack to enclose features of different scale types (binary, nominal, ordinal, integer, real valued). Besides that, only binary classification problems (e.g., low or high risk for breast cancer) are considered. A systematic comparison of existing FSM is therefore needed.

I present such a benchmark of FSMs in this section. First, I describe four criteria for FSM comparison that overcome limitations of earlier studies. Thereafter, I present results of the empirical analysis which was done together with a colleague[9]. The considered feature selection algorithms for this benchmark are described in Table 3.5 on 76.

As an empirical basis, I have chosen dataset D because it has ideal characteristics to determine the quality of FSMs for a wide range of classification problems. On the one hand, the dataset contains a wide range of classification problems (see Table 5.4): binary problems (e.g., *Single*, or *Household type*), multi-class problems (e.g., *Age of residency*, *Number of residents*), balanced class sizes, and skewed class distributions (e.g., *Space heating type*, *Cooking type*). On the other hand, many heterogeneous predictors are available: The data set contains, in combination with all considered data sources, 308 features from the 15-minute electricity meter readings[10], address data, geographical data from OSM and weather data. The features vary also widely in their nature: the spectrum ranges from several features with a value ranging (without scaling) between -1 and $+1$, features with large negative as well as positive numbers, binary features that indicate for example the existence of a geographical property, and finally categorical features (such as the land use type).

[9] I acknowledge the contribution of Andreas Weigert, who has implemented interfaces to several FSMs and prepared parts of the comprehensive data analysis.

[10] For this analysis, one representative week (February 09—15, 2015) without special days or school holidays is used.

5.4.2 Quality criteria for FSM performance

To fill the research gap outlined above and overcome the limitations in earlier studies, I suggest to compare FSMs among the criteria 1) classification accuracy improvement, 2) stability, 3) average size of the resulting feature set, and 4) algorithm runtime. Each criterion is explained and defined below.

Classification accuracy improvement The classification quality is the most critical criterion for assessing FSMs and was used in most previous works. The classification accuracy is, even when it has flaws in expressing the classification quality with imbalanced data, a good metric for an initial analysis. It expresses the average classification performance, can be calculated for all classification problems, and is good to interpret. Being able to compare the accuracy for different classification problems, the accuracy improvement respective to classification performance without feature selection is considered. MCC can be used to obtain a more detailed analysis, but some cases exist where the metric cannot be calculated, as discussed in section 4.2.

Stability The ability of FSMs to find similar feature sets by taking different subsets of training data into account is an indicator for the reliability of the method. An appropriate measure for the similarity of two feature sets K_1 and K_2 is the Jaccard index (Saeys, Abeel, et al. 2008):

$$Jaccard(K_1, K_2) = \frac{|K_1 \cap K_2|}{|K_2 \cup K_2|} \tag{5.1}$$

Complete similarity of sets is expressed by 1, disjoint feature sets 0. For each combination (of classification problem, FSM and classifier) c, the stability is estimated as mean Jaccard index on all permutations of feature sets K in the k-fold cross-validation (Kalousis et al. 2007):

$$Stability_c = \frac{2}{k^2 - k} \sum_{i=k}^{k-1} \sum_{j=i+1}^{k} Jaccard(K_i, K_j) \tag{5.2}$$

Number of selected features Simple models having a smaller number of predictor variables are preferred to complex ones (Hastie, Tibshirani, and J. Friedman 2009). The number of features is also be a meaningful criterion to assess the quality of a FSMs.

Runtime The computational complexity is finally a criterion that is relevant for practical considerations. As the exact determination of the computational complexity is sophisticated, I propose to use the algorithm runtime as a fourth criterion to evaluate FSMs. Of course, the hardware that is used for the computation should be equal when multiple FSM experiments are conducted.

5.4.3 Accuracy improvement of FSMs in a minimal viable setup

The major goal of feature selection is the improvement of classification quality. I quantify this improvement by subtracting the classification accuracy achieved without applying any FSM from the accuracy using a FSM. As machine learning algorithms have functionalities to weight features or internally select features, I use logistic regression with ordinary least squares estimation, given that this method has no internal feature transformation. For result comparison, I consider SVM and RF as two state-of-the-art classification algorithms to better interpret the results in the context of machine learning.

Figure 5.5 shows the accuracy improvement compared to no feature selection taking logistic regression classifier. The methods with names beginning with "F:" stem from the R "FSelector" package (Romanski and Kotthoff 2014), and those with "C:" from the "CORElearn" package (Robnik-Sikonja and Alao 2016). "Boruta" comes from a separate package of the same name (Kursa and Rudnicki 2010). Except six, all FSMs improve classification accuracy by more than 10%. The results of a previous study by Haury et al. (2011), namely that feature selection improves the overall classification performance, can be therefore confirmed.

The results were calculated using 5-fold cross-validation and the procedure is repeated ten times. Initially, the cross-folds were selected randomly and independently for each run. During the analysis, a high variance in performance was visible, even for the same classification problem with the same combination of FSM and machine learning algorithm. This results from the variation in training and test data (through random assignment in the five folds), or due to random components in the FSM. Therefore, we decided to use a minimum viable setup to benchmark FSMs: logistic regression with the same random allocation of data points in the five cross-folds (yellow bars in Figure 5.5). In a second setup, the training data was varied by randomly creating new cross-folds allocations (orange bars in Figure 5.5). The mean accuracy together with a 95% confidence interval was estimated considering the $t(9)$ distribution based on the ten repetitions for each setup.

Surprisingly, logistic regression (as a classifier with low generalization performance) can—together with the best FSMs—achieve accuracy results close to performance results of advanced classifiers (SVM, RF) without a feature selection before the algorithm training.

Most methods show no variance in the minimum viable setup, in which no variation through the separation of training examples was introduced. Six methods exhibit some deviations in the results, even when the same training and test data are used. I attribute this to random components utilized in the methods. In the second setup (where training and test data are varied in each of the 10 iterations), the variance of accuracy are larger, as expected, and I can conclude that 5-10% deviation in classification accuracy is attributed to the variation in separating the cross-folds.

5.4.4 Stability of feature selection

Through solely considering the classification accuracy improvement, no clear decision on the best performing FSMs can be made, when the high variance in the results is taken into account. We therefore also included the stability of the feature selection and the number of selected features. We normalized the stability by the logarithm of the number of selected features, as there is a significant relationship between stability and the number of selected features (Pearson's $\rho = 0.51, t = 57.601, p < 0.0001$). This correlation is plausible, given that the likelihood to select similar feature sets increases by the number of selected features,

The results are illustrated in Figure 5.6. Five methods (F:consistency, F:cfs, Boruta, C:Gini, C:InfGain) have a high classification performance improvement (20.4%–21.8%) and relatively stable results (the 95% confidence interval of accuracy improvement is 1.2%). Comparing the methods' characteristics, no clear pattern can be seen (e.g., only two methods consider interdependencies between features). I conclude therefore that these five methods are good candidates to be the best FSMs for the investigated case.

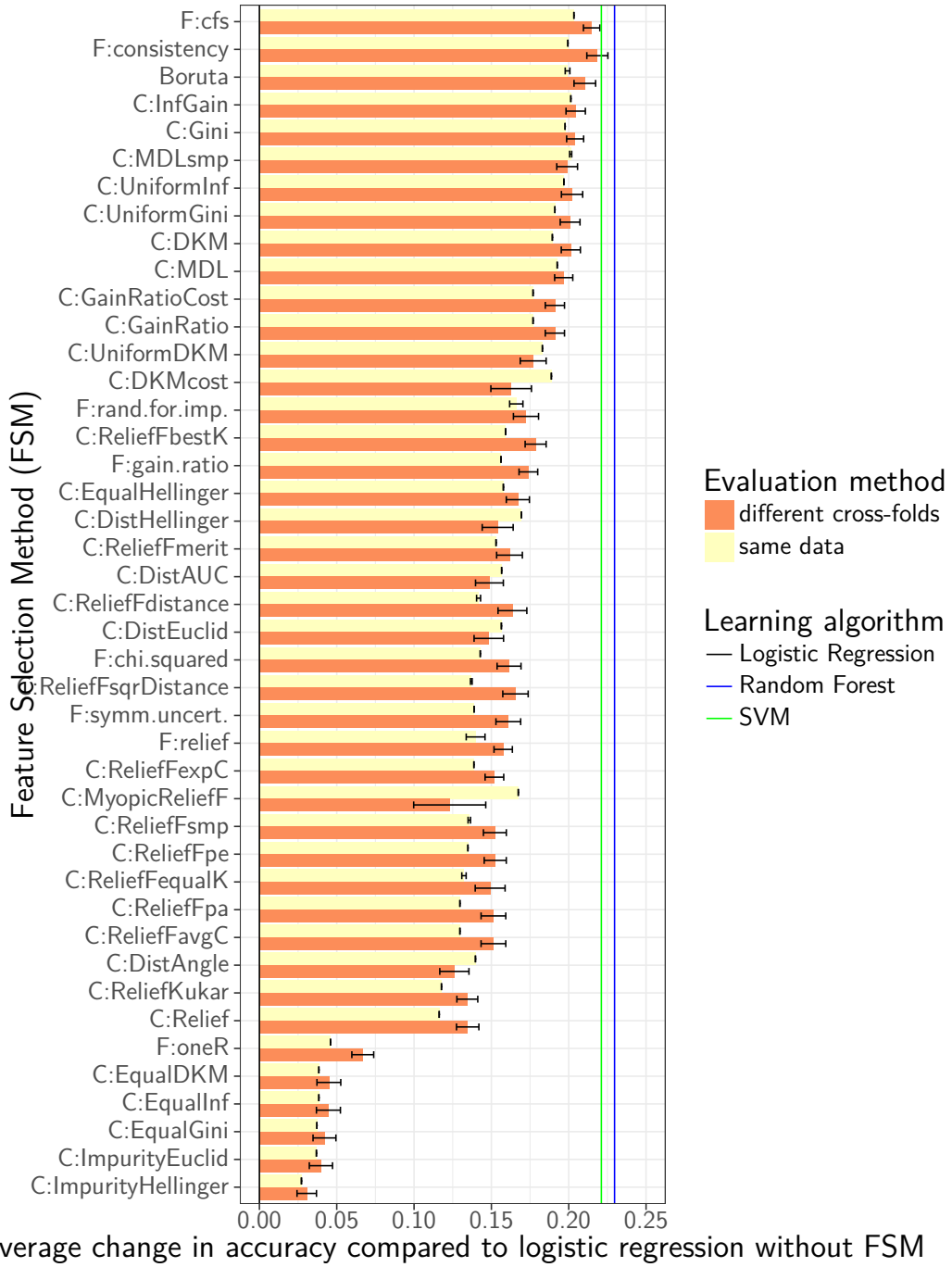


Figure 5.5: Performance results of different FSMs in average change in accuracy compared to no feature selection (using logistic regression as classifier)

5.4.5 Correlation of algorithm runtime and accuracy improvement

Finally, we investigated whether the algorithm runtime of a FSM has a correlation with the quality of feature selection. The algorithm runtime on a computer is a proxy for the complexity of the pursued computation (assuming that the algorithm is implemented efficiently). If this naive relationship between the algorithm runtime and the quality of the selection exists, a clear trade-off between runtime and quality could be weighed up.

In the described experiments, the runtime of FSMs was less than one second in most cases and there were also methods that exceeded a runtime of 250 seconds. I cannot find a relation between runtime and classification improvement, because the Pearson's correlation coefficient is very low $\rho = 0.01 (p < 0.01)$. Consequently, good feature selection does not necessarily take long. I cannot deduct from the findings that computational expensive FSMs deliver bad results, given that the "consistency" method, for example, produces good results even when it has a high computation time.

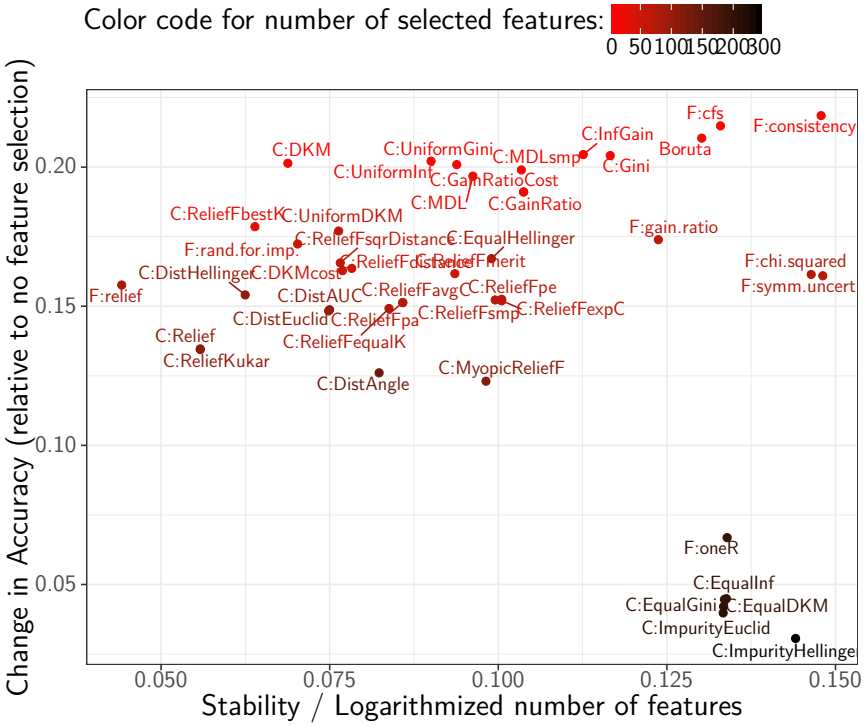


Figure 5.6: FSMs with classification performance improvement in comparison with the stability (normalized by the logarithm of the number of selected features)

5.5 Predictability of household characteristics based on different data granularities

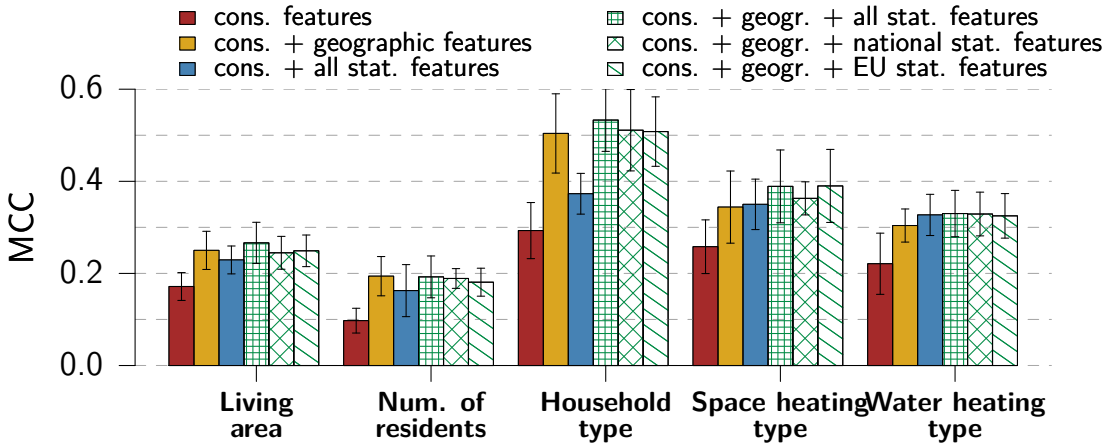
After having analyzed the performance of classification algorithms and FSMs, the predictability of different household characteristics based on the four considered datasets is evaluated. The datasets cover different granularity of the energy consumption data (annual, daily, and smart meter data). I further examine how the prediction performance can be improved through combining dependent variables and analyze how well predictive models trained in one geographic region (i.e., country) can be transferred to another region. The results give an overview to what extent energy utilities can make sense of data they possess.

So far, I found that RF works well in the investigated problem class of predicting household properties based on energy consumption data. Consequently, I use the RF classification algorithm in the all analysis where the predictability of household properties is investigated.

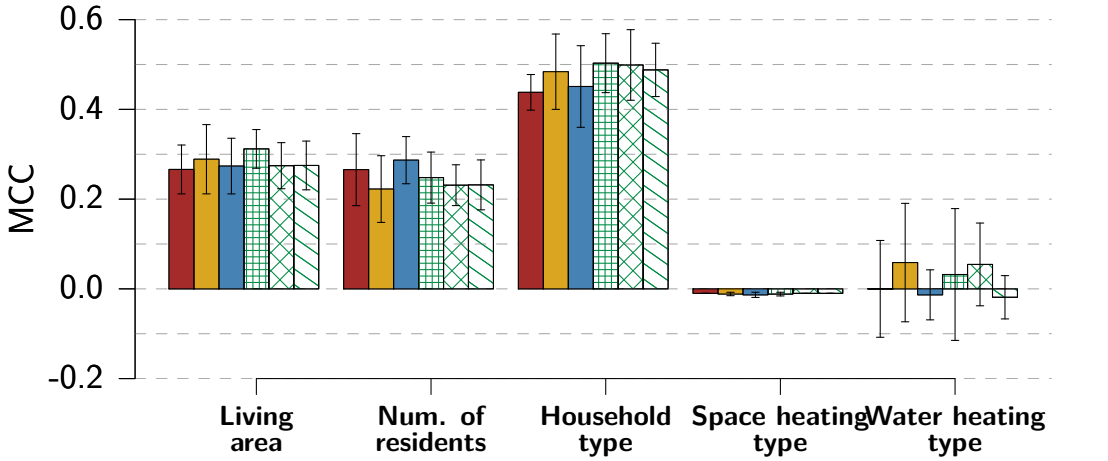
5.5.1 Annual electricity consumption data with geographic and statistical data

I tested the classification performance of a RF classifier with annual electricity consumption data and examined to what extent governmental statistical data, and geographic data from OpenStreetMap can improve the prediction performance. For this analysis, dataset A and B were used. The results are taken from Hopf, Riechel, et al. (2017). Figure 5.7 shows the performance of the classification considering different feature sets: The first setup includes solely consumption features as the base case (red bars). Then, consumption together with geographic features (yellow bars), consumption together with all available statistical features (dark blue bars), consumption together with geographic features, and different kind of statistical features (different shades of green) is taken. The MCC results are obtained in 10-fold cross-validation, the standard deviation is depicted as error-bars in the graphs. In the investigated case, all variables—except the heating types in Germany—can be predicted better than random (MCC values greater than zero).

Statistical data improve the classification only in Switzerland (red vs. blue bars in Figure 5.7a) to some extent, but significantly (paired $t(4) = -9.1478$, $p < 0.001$). For Germany, no improvement of statistical data can be seen. I attribute the lower contribution of this kind of data in Germany to the fact that the geographic areas for which the statistical data is published are larger than



(a) Results for Switzerland (dataset A)



(b) Results for Germany (dataset B)

Figure 5.7: Classification performance (in MCC) for five household properties in dataset A and B using consumption, governmental statistical, and geographic data in various combinations with the Random Forest classifier

in Switzerland[11]. Using all statistical features (from national and EU agencies) achieved the highest classification performance, but considering only one source of such features—either national or from EU—does only lead to a small, but significant loss in classification performance (paired $t(4) > 2.3, p < 0.05$).

The geographic features from OSM together with electricity consumption and statistical data improves the model performance strongly, but including geographic information in the models, the added value from open statistical data is lowered for some properties and disappeared completely for others (orange vs. green bars). I assume that the geographic data describes regional differences between the households quite well and the level of detail in the statistical data are too low to add further information to the prediction model.

I expect that the low performance for the heating types in Germany results from the uneven distribution in the class sizes (2% "electric" vs. 98% "not electric" for *Space heating type*, and 10% "electric" vs. 90% "not electric" for *Water heating type*).

5.5.2 Daily electricity consumption data

For the prediction of nine household properties (dataset C) considering features from daily, HT and NT electricity consumption data, I tested six classification algorithms and considered features from automatic feature extraction (see section 3.3.1), geographic, and weather data. The standard configurations of each classifier was applied as described above.

Figure 5.8 shows the classification accuracy for each algorithm. All considered household properties can be predicted better than a biased random guest and thus, the prediction of household properties based on daily electricity consumption data is possible. The RF classifier produces again good results on average, whereas some classifiers outperform RF in some cases. The Naïve Bayes classifier is worse than the BRG metric in four cases.

The classification of household characteristics on the basis of daily consumption data is possible, but naturally less accurate than the availability of more detailed data. In many use cases it is not necessary to accurately predict all household characteristics because one is interested in a combination of characteristics. A marketer for photovoltaics, for example, would like to identify all households that live in a house and also own it; an energy consultant is probably more interested in the type of household and the number of people living in the

[11] The size of statistical geographical unit in Germany is $M = 33.32k\text{m}^2$ on average 95% larger than in Switzerland, where the average area of statistical regions is $M = 17.12k\text{m}^2$ (paired $t(13,670) = 11.405, p < 0.0001$)

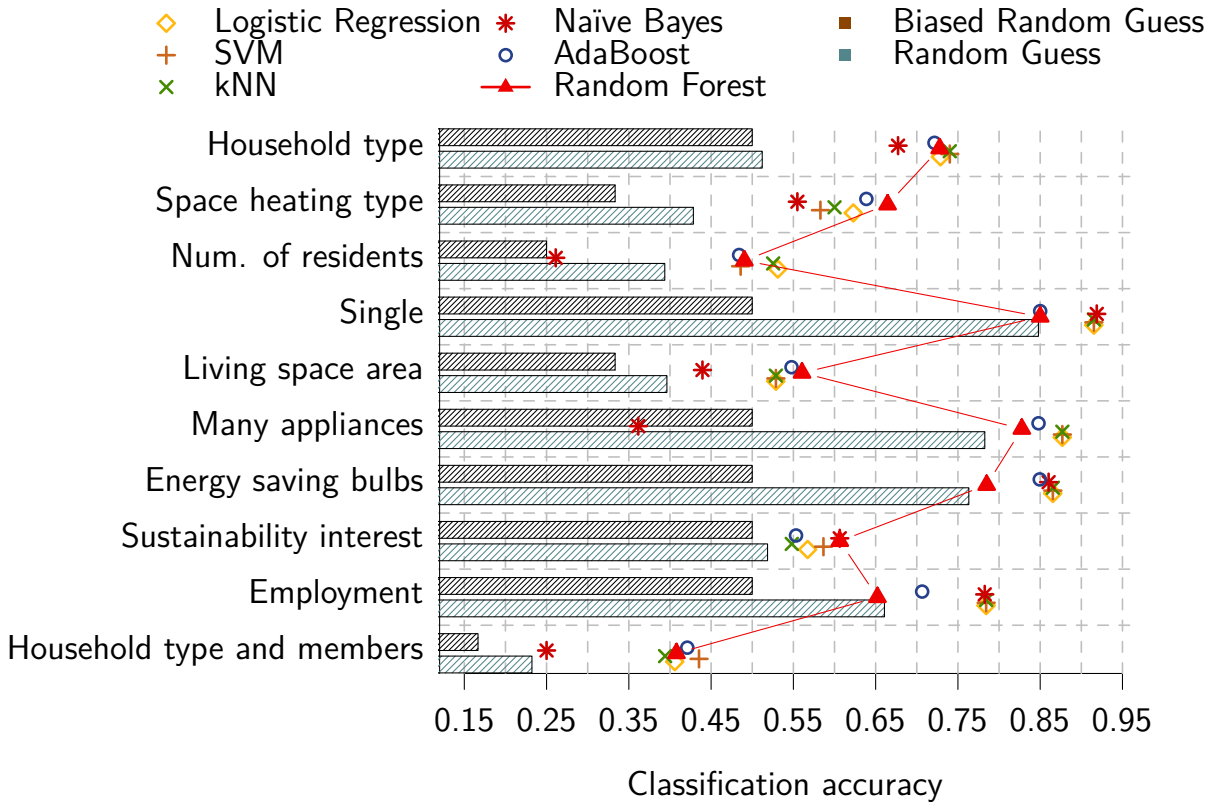


Figure 5.8: Classification accuracy for 9 household properties in dataset C (plus one property-combination) with six classifiers

household. Such information is also often combined in the data, as household properties are represented together as latent variables in the features. Electricity consumption (which accounts for a large proportion of the features considered), for instance, is influenced by the number of residents, the living space area, and the type of heating system. The geographical data contain the base area of a house, indicating either a large single-family house or a multi-family house with several small apartments.

I have examined the prediction of property combinations with dataset C using the example of *Number of residents* and *Household type*. For this, I tried different classification algorithms and FSM, and identified SVM together with "Random Forest importance" and "Boruta" FSMs as the best combination. The classification performance for SVM together with "Random Forest importance" feature selection led to $AUC = 0.7312$ when predicting the property combination jointly. A separate prediction of the variables and later combining them led to $AUC = 0.3355$, which is obviously worse. SVM together with "Boruta"

feature selection resulted in $AUC = 0.6178$ for the combined and $AUC = 0.3445$ for the separate prediction. I therefore conclude that the prediction of property combinations rather than individual predictions should be considered for use cases where a combination of household properties is relevant.

5.5.3 Smart meter data

In the previous sections, I have investigated the performance of seven classification and 43 feature selection algorithms. To test the predictability of household properties based on SMD, weather data, and geographic information from OSM, I combine RF and the "consistency" feature selection (both showed best results in the previous analyses) and trained prediction models for the 19 household properties available in dataset D. For the evaluation of the models, 4-fold cross-validation is applied and the procedure (including the random sampling to create the cross-folds) is repeated 50 times when each time another week of smart meter and weather data are used.

Table 5.5 shows the average classification performance for each household property in terms of accuracy and MCC, and the performance on the level or individual classes in terms of AUC. I calculated a statistical $t(49)$-test to examine whether $MCC > 0.05$, which means that the classification is significantly better than random. The column "MCC Sign." indicates all cases, where this criterion is met with at least one asterisk. From this analysis, I conclude that all household properties are predictable based on the data and applied algorithms, except the variables *Efficiency measures*, *Cooking type*, *Solar installation*, and *Satisfaction utility*.

The analysis completes the study results documented in Hopf, Sodenkamp, and Staake (2018) where 11 properties were investigated. In the results presented herein, the geographic features from OSM are used and a FSM is applied. However, the new results of this updated study confirm the figures presented in the earlier publication.

Table 5.5: Classification performance (Random Forest algorithm) for all properties and classes measured in Accuracy, MCC, and AUC on average based on all 50 weeks of electricity consumption and weather data individually

Property	Class	Accuracy		MCC			AUC	
		Mean	SD	Mean	SD	Sign.[a]	Mean	SD
Age of appliances	New	0.3805	0.04	0.0714	0.06	**	0.6175	0.03
	Moderate						0.5165	0.04
	Old						0.5423	0.04
Num. of appliances	Few	0.5241	0.03	0.1178	0.07	***	0.6846	0.05
	Moderate						0.5503	0.05
	Many						0.6446	0.05
Efficiency measures	No	0.4301	0.04	−0.0042	0.05		0.5028	0.04
	One						0.5107	0.05
	Multiple						0.4825	0.06
Age of residency	< 10	0.5710	0.03	0.3979	0.05	***	0.8225	0.04
	$10 - 29$						0.7726	0.03
	$30 - 74$						0.7057	0.03
	$>= 75$						0.8382	0.02
Cooking type	Electric	0.8732	0.02	0.0004	0.09		0.5046	0.09
	Not electric						0.5046	0.09
Heat pump	No	0.8467	0.03	0.3468	0.14	***	0.7461	0.08
	Yes						0.7461	0.08
Solar installation	No	0.9265	0.02	0.0498	0.11		0.5547	0.09
	Yes						0.5547	0.09
Interest in solar	Low	0.6563	0.05	0.0663	0.07	*	0.5988	0.05
	Moderate						0.5306	0.06
	High						0.5840	0.06
Space heating type	Electric storage	0.8098	0.04	0.2667	0.18	***	0.5930	0.12
	Heat pump						0.6898	0.10
	Other						0.6868	0.10
Water heating type	Electric storage	0.7124	0.02	0.3460	0.06	***	0.8052	0.03
	Heat pump						0.6932	0.06
	Other						0.7932	0.03

Property	Class	Accuracy		MCC			AUC	
(continued)		Mean	SD	Mean	SD	Sign.[a]	Mean	SD
Age of heating	New	0.4772	0.04	0.2028	0.06	***	0.6444	0.05
	Medium						0.6628	0.05
	Old						0.6012	0.05
Num. of residents	1 person	0.5066	0.04	0.2197	0.06	***	0.7824	0.05
	2 persons						0.6111	0.04
	$3-5$ persons						0.6759	0.05
	>5 persons						0.7521	0.07
Single	No single	0.8099	0.03	0.3266	0.12	***	0.7548	0.06
	Single						0.7548	0.06
Children	Children	0.7842	0.03	0.0768	0.09	*	0.6065	0.06
	No Children						0.6065	0.06
Family	Family	0.7076	0.03	0.0952	0.08	***	0.6207	0.06
	No Family						0.6207	0.06
Household type	Apartment	0.8325	0.03	0.6618	0.06	***	0.8942	0.02
	House						0.8942	0.02
Home ownership	Rent	0.7827	0.02	0.5420	0.05	***	0.8475	0.02
	Own						0.8475	0.02
Living space area	≤ 95	0.5134	0.04	0.2534	0.06	***	0.7901	0.03
	≤ 145						0.5699	0.04
	> 145						0.7477	0.03
Satisfaction utility	Low	0.4112	0.05	−0.0006	0.06		0.5162	0.06
	Moderate						0.5096	0.06
	High						0.4686	0.06

a) Significance code: "***" $p < 0.001$, "**" $p < 0.01$, "*" $p < 0.05$

5.5.4 Geographic transferability of models

Finally, I evaluate the transferability of the household classification models among two case studies. This aspect is relevant, because it shows how strong the prediction models are fitted to the data, or in other words: how generalizable the models are. This aspect is on the one hand relevant from a research perspective, because models should avoid dataset specific findings. On the other hand, it is relevant for practice, as transferable models are not needed to be extended with new data, when they should be applied to a new region. Multi-national

firms, for example, do not need to create ML models for each regional division separately, and vendors could offer pre-trained models as a service.

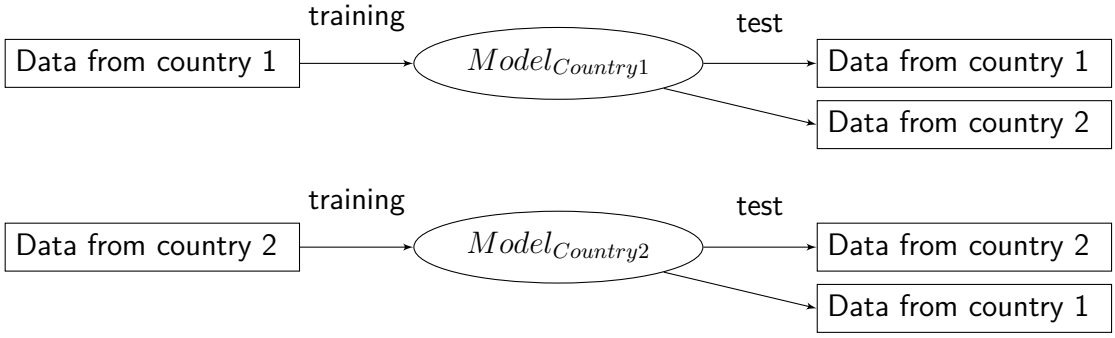


Figure 5.9: Illustration of the four considered cases to test the geographic transferability of machine learning models

The transferability of models based is investigated with two cases. In the first case, annual electricity consumption data (dataset A and B) together with statistical data from Germany to Switzerland and vice-versa are considered. As a second case, I consider the transferability of 30-min smart meter classification models from Switzerland (based on dataset D) to Ireland and vice-versa. The detailed analysis of this case is documented in the project report Sodenkamp, Hopf, Kozlovskiy, et al. (2016). For this analysis, a 30-min smart meter dataset was from the Irish Commission for Energy Regulation (2011) was used, which was also subject to previous studies (Beckel, Sadamori, Staake, et al. 2014; Wang et al. 2018). In both cases, the experiment setup depicted in Figure 5.9 was applied.

In the first case, five household properties and the predictor variables from annual electricity consumption and geographic data were considered (see the MCC classification performance in Figure 5.7 on page 129, orange bars). The RF classifier was applied. The detailed accuracy results and the percentage change in accuracy for the transferabilitiy experiment is listed in Table 5.6.

Classification accuracy of models trained in Germany and applied to Switzerland data had a loss in prediction accuracy of 7.58%. This difference can be interpreted as a small effect (Cohen's $d = 0.3013$), but statistically weakly significant (paired $t(4) = 1.6182, p = 0.0905$). For models trained in Switzerland and applied to Germany, the loss is only 1.55% ($t(4) = 0.5025, p = 0.3208, d = 0.0648$). Further details on this analysis are documented in Hopf, Riechel, et al. (2017).

In the second case, the transferability of classification models based on 30-min SMD together with weather data. The outside temperature was used to represent the weather information, as this data were available for both regions

Table 5.6: Accuracy of models trained with annual electricity consumption data from households in Switzerland and Germany together with geographic data and are applied to households in both countries

Property	CH⇒CH	DE⇒CH	Diff. (%)	DE⇒DE	CH⇒DE	Diff. (%)
Living space area	0.5298	0.4647	−12.29	0.5429	0.5108	−5.91
Num. of residents	0.5274	0.5138	−2.58	0.5160	0.5235	1.44
Household type	0.7480	0.7156	−4.33	0.7320	0.7870	7.52
Space heating type	0.8216	0.8472	3.12	0.9754	0.9814	0.61
Water heating type	0.6396	0.5000	−21.82	0.8965	0.7945	−11.36

in the time span of the electricity consumption data. Four household characteristics were available in both data sets and were used to test the geographical transferability between Switzerland and Ireland: *Children* existent in the household, *Cooking type*, *Family*, and *Single* (all binary variables). In both datasets, a week in early summer without special days or school holidays was selected, however in different years (June 02–08, 2014 in datset D, and May 31–June 06, 2010 in the Irish data). For classification, the RF classifier was applied. Results of this transferability experiment are listed in Table 5.7. I can conclude that models trained in Ireland can be applied in Switzerland with an accuracy loss of $M = 2.74\%$ and models trained in Switzerland can be used in Ireland with a loss in accuracy of $M = 3.38\%$.

Table 5.7: Accuracy of models trained with 30-min electricity SMD from households in Switzerland and Ireland and are applied to households in both countries

Property	$CH \Rightarrow CH$	$IR \Rightarrow CH$	Diff. (%)	$IR \Rightarrow IR$	$CH \Rightarrow IR$	Diff. (%)
Cooking type	0.9280	0.9383	1.11	0.9624	0.9624	0.00
Children	0.8406	0.8123	−3.36	0.7067	0.6787	−3.95
Single	0.8278	0.7584	−8.39	0.8703	0.8093	−7.01
Family	0.7609	0.7584	−0.34	0.7243	0.7058	−2.56

To summarize the findings, I conclude that ML models trained with ground truth data from a geographically limited region can be applied to new data points that are located in another area. The performance loss that must be taken into account is manageable and lied significantly below 10% in the conducted experiments. Consequently, the models developed in this work can be assumed to be not overfitted to the specific datsets raised.

5.6 Discussion and implications

Detailed customer knowledge is most helpful to improve energy efficiency and service quality in the residential energy sector. On the one hand, targeted behavioral interventions (like feedback campaigns to trigger energy savings or the shift of electric load to other times of the day) can be realized when characteristics of the receiving household are known. Such campaigns are more effective when the communication is tailored to the feedback receiver (Allcott 2011; Loock et al. 2013; Tiefenbeck, Goette, et al. 2016). On the other hand, knowledge on household characteristics and intentions of customers enable utility companies to develop new products and advertise them to relevant customer groups. This leads to improved conversion rates but also decreases the advertisement effort that are not interested in a certain offering.

In this chapter, I investigated ML models that reveal residential households characteristics and customer intentions from energy consumption data together with external data (i.e., weather data, geographic information, governmental statistical data). The results help firms to develop and advertise energy-related products. The findings also help analysts with the second stage of the data-driven decision making process *data to insight*, as I found out which algorithms work well with the investigated problem class and tested the stability of predictive models under changing conditions (different data granularity and different geographic location).

Below, I summarize the results of this chapter and give the answer to RQ 3 as well as finalize my answer to RQ 2. Thereafter, I name limitations of my study and outline future research direction. Finally, I compiled some lessons learned regarding model development and tuning as practical implications for predictive modeling.

5.6.1 Recognition of household characteristics based on electricity consumption data

In total, 22 dependent variables from four datasets were investigated. Thereby, I evaluated six machine learning and 43 feature selection algorithms. With this comprehensive analysis, I went significantly beyond existing studies, considering the combination of the four topics: Features, algorithms, dependent variables, and model stability.

The present work is, to the best of my knowledge, the first comprehensive study on household classification with data from central Europe (Germany and Switzerland). In addition to the earlier studies, I have used external data to-

gether with electricity consumption data for the prediction of household characteristics. To illustrate this contribution in comparison with earlier studies, Table 5.8 lists the related works which investigate the recognition of residential household characteristics from electricity consumption data together with the used dataset, the consumption data resolution and considered external data.

Table 5.8: Studies predicting household characteristics with different electricity consumption datasets and external data (the symbols are used in Table 5.9)

Study	Reference	Dataset	Consumption data	External data
•	Beckel (2015) and Beckel, Sadamori, Staake, et al. (2014)	Commission for Energy Regulation (2011) smart meter trial with 4,232 Irish households, 18 variables	30-minute electricity SMD	–
○	Hopf, Sodenkamp, Kozlovskiy, and Staake (2014)			–
△	Wang et al. (2018)			–
⋆	A. Albert and Rajagopal (2013)	Energy feedback study with 1,100 US households, 28 variables	5 – 15 minute SMD	Weather data
×	Fei et al. (2013)	4,564 utility customers with a heat pump and 1,821 without	Daily SMD	Weather data
∗	Verma et al. (2015)	1,250 utility customers in Michigan (US) with 50% plug-in electric vehicles	Hourly SMD	–
⧫	This work, including earlier published studies (see section 1.4)	See Table 5.1	Multiple	Geographic, weather, governmental statistical data

Whereas earlier works successfully tested weather information, I added geographic data from OSM and governmental statistical data and tested their contribution for the predictive modeling. Additionally, I tested the prediction of household data with different data granularity. Household details with strong

impact on the energy consumption (e.g., size and type of the household) can be predicted from annual consumption data. Furthermore, I tested various machine learning algorithms in the field of energy data analytics and support findings of Fernández-Delgado et al. (2014), and extended the comparison of FSMs following Haury et al. (2011).

In addition to the technical improvements of the classification approach, I have also investigated the predictability of household characteristics that were not tested so far. Table 5.9 lists the characteristics on which earlier studies and my works focused on.

From Table 5.9, it can be clearly seen that earlier works investigated general household characteristics, the existence of single appliances, and behavior of the residents (categories A-C). Beyond that, I examined variables related to heating as well as energy production, and found that they are predictable with the proposed approach. I showed that even interests and the house ownership can be predicted better than random with data that are available to energy utility companies.

Answer to RQ 3 The ambient data sources investigated in this work (consumption data of electricity, gas and water, geographic information, geographic information, and governmental statistical data) contain many latent variables that can be revealed through ML methods. In response to RQ 3 of this work (*How well can (a) customer characteristics and (b) intentions be revealed from ambient data, in the context of energy retail, using state-of-the-art ML methods?*), I can give a positive answer: Household characteristics and customer intentions can be predicted significantly better than a random predictor. In my analysis, 18 household characteristics and intentions of residents could be predicted better than random from the considered data sources by using state-of-the art ML algorithms using datasets from Switzerland and Germany. Especially the prediction of variables related to heating and energy production in residencies, as well as intentions (e.g., *Interest in sustainability*, *Interest in solar installation*), that have not been mentioned in earlier works, were positively tested. The prediction of general household characteristics was tested with smaller data resolutions than in earlier studies (yearly and daily electricity consumption data in addition to 15-/30-minute smart meter data).

Table 5.9: Household characteristics that could be predicted based on electricity consumption data of different time series resolution in earlier studies (symbols are listed in Table 5.8) and this work (marked with ⧫)

		Consumption data resolution		
Property	Example classes	Annual	Daily	$15-60$ minute
A) Characteristics with strong impact on electricity consumption				
Living space area	$\leq 95, \leq 145, > 145\text{m}^2$	⧫	⧫	•, △, ⧫
Num. residents	$1, 2, 3-5, > 5$	⧫	⧫	•, ○, ⧫
Children	Existent, not existent	⧫	⧫	•, ○, △, ⋆, ⧫
Household type	Apartment, house	⧫	⧫	•, △, ⧫
C) Appliances, installations and other household details				
Cooking type	Electric, not electric			•, △, ⧫
Num. bedrooms	Few, average, many			•, ○, △
Energy light bulbs	Yes, no		⧫	•, △, ⧫
Age house	$< 10, 10-29, > 75$ years			•, ○, △, ⧫
Num. appliances	Few, moderate, many		⧫	•, ○, ⧫
Age of appliances	New, moderate, old			⧫
Electric dryer	Existent, not existent			⋆
Washing machine	Existent, not existent			⋆
Fridge	Existent, not existent			⋆
Plasma TV	Existent, not existent			⋆
Electric vehicles	Existent, not existent			∗
D) Characteristics and behavior of the residents				
Employment	Employed, unemployed		⧫	•, ○, ⋆, ⧫
Age person	Young, medium, high			•, △, ⋆
Retirement	Retired, not retired			•, △
Income	Low, high			•
Social class	A/B, C1/C2, D/E			•, △
Pets	Existent, not existent			⋆
Unoccupied during the day	Yes, no			•
B) Heating, cooling and energy production				
Air conditioning	Existent, not existent			⋆
Space heating	Electric, not electric	⧫	⧫	⧫
Water heating	Electric, not electric	⧫	⧫	⧫
Age heating	Old, medium, new			⧫
Heat pump	Existent, not existent		×, ⧫	⧫
Solar installation existent	Yes, no			⧫
E) Interests of the residents and property ownership				
Interest in sustainability	High, medium, low		⧫	
Interest in solar installation	Yes, no			⧫
House ownership	Yes, no			⧫

In accordance with related works, I found no superior ML algorithm always leading to best results. Nevertheless, the RF algorithm achieved, on average, the most accurate classification results and outperformed five other well-known supervised machine learning algorithms. This supports the findings obtained by Fernández-Delgado et al. (2014), who came to this conclusion in an investigation with several simulated datasets and datasets from other application domains. The RF algorithm is able to recognize patterns from many features for a variety of problems, even if the amount of training examples is small.

Models predicting household characteristics that were trained in one geographic region can be applied to other regions with acceptable performance loss of $1-8\%$ on average in the investigated regions Switzerland, Germany and Ireland. This indicates the generalizability and stability of the models. The reliable models enable manifold innovations in energy retailing that can improve residential energy efficiency, for example, personalized feedback or automated energy consulting.

Empirical comparison of FSMs (continued answer to RQ 2) As a second theoretical contribution of this chapter, I provided further analysis to RQ 2 (*Does theory, expert knowledge and human cognition notably help to reduce data dimensionality, although several computational methods exist for this task?*) in which the value of human cognition, theory, and human expert knowledge in advanced data analytics is questioned. In particular, I tested 43 feature selection algorithms with the classification problems of my study. The usage of FSMs in all considered classification problems brought on average 15.31% improvement in accuracy, compared with the results achieved by considering all features. Thereby the improvements range between 3.06% and 21.85%, using logistic regression as a classifier with no internal feature transformation or selection. Interestingly, logistic regression (an explainable model which has not the generalization capability as other ML algorithms have) together with feature selection can achieve results that are close to advanced algorithms (RF, SVM) for the considered problems in my analysis.

From this study, I conclude that feature selection is a valuable step in model development. The simple application of such methods to raw data is, however, not meaningful, considering two important considerations. First, the benchmark of methods conducted in this research disclosed that there is no superior technique to select feature sets. Thus, one cannot assume that the algorithms perform the very important task of data preparation properly. Second, the feature selection uses heuristics to judge whether a feature is relevant for a prediction or not. Thereby, the dependent variable is mostly considered to calculate

the feature importance measure. This means that the computation can identify variables with a high correlation with the dependent variable *in the specific dataset*. This does, conversely, not necessarily mean that a causality relation exist (the feature is a real predictor) and it does not mean that this correlation exists outside the selected training sample.

Therefore, I conclude that *the nexus of theory, expert knowledge, and human cognition is necessary* to efficiently select and combine the methods for effective modeling.

5.6.2 Limitations and future research

The study presented in this chapter has limitations and can be extended in the future. The performance of classification algorithms in combination with FSMs can be further evaluated considering, for example, artificial datasets that are problem-independent. Besides that, some FSMs have special support for different feature types and are appropriate for either continuous features, categorical features with low or high number of categories, or the ordinal relation of categorical features (Kononenko 1995). Other methods recognize interdependencies between features and should be evaluated in appropriate classification problems (Bolón-Canedo et al. 2015). Looking at the dependent variable, many machine learning algorithms and some FSMs support the assignment of weights to specific classes, so that it is possible, for example, to identifying customers with "high" interest in a product rather than customers that are not interested (Robnik-Šikonja 2003). I therefore encourage further studies with specific datasets for the mentioned special characteristics of machine learning algorithms and FSMs. Finally, further FSMs could be implemented and tested, also together with different classification algorithms to find adequate FSMs for different types of algorithms. Future research in business analytics should also agree on guidelines needed on how to report classification performances and model configurations.

5.6.3 Practical implications: Model improvement beyond algorithms tuning

The empirical analyses in this chapter comparing learning algorithms and FSMs have disclosed one fundamental point: The complexity of predictive data modeling. Thereby, I have not even considered the optimization of parameters (Bergstra and Bengio 2012) of individual algorithms, except for the example of the SVM algorithm. The search space consisting of algorithms, their parameters, variable selection methods, possible additional data sources and the

modification of the prediction variable is infinite. I argue therefore, that the creation of "good" models is not limited to "tuning" or "tweaking" methods, it is rather intelligently combining existing techniques to meet the requirements from practice. This development of models must always be specific to the question investigated and I support Thiess and Müller (2018) in their argument that the data-driven decision making process must start with a question to clarify the objective of analytics.

If the requirement is, for example, to achieve the best possible recognition accuracy and if labeled data are available, countless learning algorithms or variants of them can be implemented. A model selection can be performed and the best model selected. In the end, it is not necessarily possible to explain why this model was chosen and why the prediction was calculated this way.

Conversely, if there is the requirement that one should be able to explain the trained model at least to some extent, the search for a prediction algorithm has to be restricted accordingly. Additional insights from detected pattern can then be derived from the model, often at the cost of a lower prediction accuracy. Feature selection methods, for instance, help to find out which features are included in a model. Ensemble models can calculate how relevant a feature is for the model (e.g. RF importance). A linear or logistic regression model can even calculate how strong the influence of a variable is on the prediction and how the prediction changes when the feature value is changed.

The prediction performance of models is strongly dependent on the used data sample. In an investigation with logistic regression (a deterministic statistical learning technique) variations in the classification accuracy of $1 - 4\%$ (aggregated over all tested classification instances) occurred when the training samples were mutated. Therefore, I recommend reporting classification results together with a standard error or confidence interval, especially when different models or algorithms are compared.

6 Personalized home energy reports for user engagement and residential energy efficiency

Highlights

- ▷ A personalized energy report was tested in an experiment with 414 residential customers; it led to *increased usage* of an energy efficiency web portal and resulted in *customers providing more data* about themselves on this portal.
- ▷ Household customers receiving the energy report *reduced their electricity consumption by 6%* and receivers rated the energy report positively in a post-trial customer survey.
- ▷ Participants that received an energy report with predicted household characteristics in a second experiment ($n = 400$) showed *similar open and click rates* than recipients that have self-reported the household data.
- ▷ Results demonstrate that predicted data on customers enable scalable services with a high level of personalization resulting in increased energy efficiency and customer satisfaction.

The value creation from data implies to gain insights from data—this was subject to the previous chapters—and to realize benefits from the newly created insights. The latter part of the data value creation process is difficult to investigate, as "there is no one-to-one correspondence between an insight and a specific course of action to exploit that insight." (Sharma et al. 2014). Moreover, the transformation of insights to value can hardly be observed in controlled laboratory environments. It rather needs investigated in the field. This chapter therefore describes a comprehensive field study in which these aspects are examined using the example of a highly personalized customer communication in

the form of quarterly electricity consumption reports, which was introduced as a new service for customers of an energy supplier in Switzerland. The case study demonstrates how insights from predictive analytics can be used to improve energy efficiency in the residential sector and enhance the service quality of the energy provider, thereby creating *ecological* and *social* value. Such measures also increase customer satisfaction and therefore help to increase customer value (Kumar 2018), an *economic* target.

Given that it is difficult to measure the value of predictive analytics in a robust study, the field investigation was accordingly extensive. It included two experiments with 414 and 400 customers each, and a survey for which 919 customers were invited (96 participated). The results portray how predictive analytics can enable mass personalized services.

This chapter is structured as follows: First, the relevance of personalized energy feedback is motivated from a theoretical as well as a practical point of view, and the research question of this chapter is derived. Second, the study design including the the energy report, the experiment groups, and the timeline of the study is explained. Third, the data gathered in the first experiment and results from the customers survey are analyzed to investigate how the personalized energy report led to energy savings, increased energy efficiency portal usage and customer satisfaction. This baseline information on the performance of the energy report—it was created with known household data in this first experiment—is necessary to obtain the value of predicted data. Afterwards, the second experiment is described and evaluated, in which the added value of predictive analytics was investigated. The final section summarizes the findings, gives an answer to RQ 4, and names limitations of this study.

6.1 Customer engagement through tailored energy feedback

Feedback on energy use was shown to be a proper intervention to decrease residential energy consumption. In contrast to regulatory tools, like price incentives, or prohibitions, energy feedback is a measure that comes at low cost and yields high public acceptance (Allcott 2011). Feedback is more effective when it is tailored to the receiver. Successful formats compare households to similar ones in the neighborhood (Allcott 2011), focus on a particular target behavior (Tiefenbeck, Goette, et al. 2016), or use energy saving goals (Loock et al. 2013).

From an energy supplier's perspective, energy efficiency among customers has mostly a subordinate relevance and is mainly triggered by regulatory pressure.[1] It is far more important for utility companies to increase customer satisfaction, brand image, and sales. Using energy efficiency feedback as an anchor, utility companies can conduct effective marketing campaigns that focus on customer relationship building. This supports relationship marketing, where firms aim to build long term seller-buyer relationships and maximize customer value (Brassington and Pettitt 2006; Kumar 2018). Value-added services, like an energy report, offer the possibility to make specific advertisements for higher-value tariffs, or ▷cross-selling products and thereby increase the customer satisfaction.

For all forms of tailored feedback, precise information about the feedback recipient is needed, but this is usually not available at scale (Tiefenbeck 2017; Hopf, Riechel, et al. 2017). In the previous chapters, I have demonstrated that detailed information on household characteristics (e.g., household type, number of residents, type of heating), which are necessary to carry out such feedback campaigns, can be extracted from ambient data sources using ML methods. The RQ examined in this chapter investigates the further use of such insights and refers to the *insight to value gap* in the data-driven decision making process. It is formulated as follows:

RQ 4 *Which added value can be realized from predicted customer characteristics on the example of personalized energy feedback?*

The question is answered utilizing a field study in the context of energy efficiency web portals, developed and operated by BEN Energy as white-label solutions for utility companies. Several utility companies in central Europe have followed the above mentioned reasons to promote energy conservation in the residential sector and offer BEN Energy's efficiency platforms to customers under their brand. An exemplary screenshot of such a portal is shown in Figure 6.1. Customers can, for example, track and compare the personal energy consumption, as well as receive energy saving advises. The effectiveness of that energy feedback tool is described by Graml et al. (2011), Loock et al. (2013), and Lossin (2016).

In spring 2017, BEN Energy developed an energy report in the form of an e-mail as an additional functionality of the energy efficiency portals (the design of the e-mail is explained in the next section). The new service was evaluated in a pilot study between April and October 2017 which I describe in the remainder of this chapter.

[1] Regulatory mandates may force energy retailers to promote energy conservation among private customers, for instance in Europe (EU 2012).

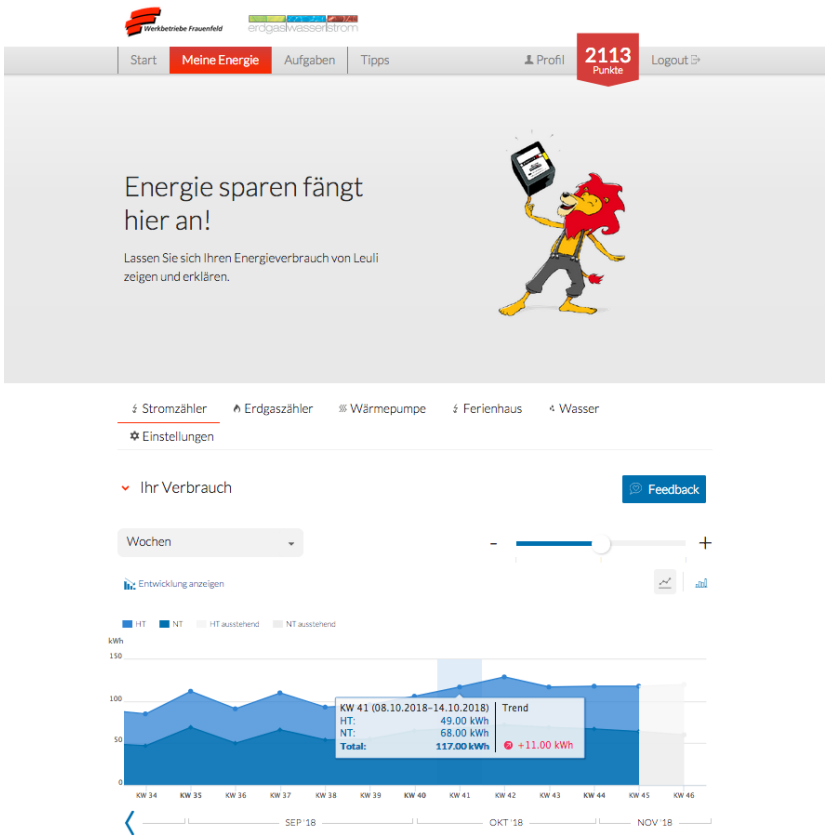


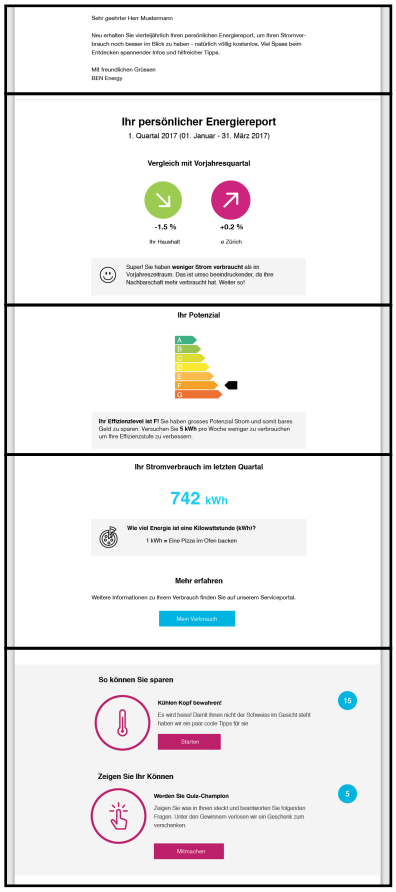
Figure 6.1: Screenshot of the energy efficiency web portal (Source: BEN Energy)

Besides evaluating the energy report as a whole, my study objective was particularly to test, how well the results of household classification can be used for energy feedback measures. This completes the research results of the earlier research questions, where conditions for good ML models were evaluated, in general, and in the context of energy retail.

The study described in this chapter was conducted together with a colleague, Liliane Ableitner, who investigated the user experience of the energy reports. In cooperation with her, the survey and the experiment were developed and carried out.

6.2 Development of a personalized e-mail energy report

The energy report was mainly designed by the Head of User Interface Design at BEN Energy, who was also responsible for the design of the energy efficiency



a) Introduction

b) Consumption trend and comparison with neighbors

c) Energy efficiency comparison with similar households

d) Total consumption and analogy

e) Energy saving tips

Figure 6.2: Energy report variant sent in April 2017 (Design: BEN Energy)

portals. Several feedback rounds involved colleagues from BEN Energy, my research colleague, and myself. The design of the energy report was created considering the following requirements: the e-mail is sent out once a quarter, it should give the customer a vivid overview of the consumption in the past period, display the consumption trend, and compare the consumption with the neighborhood. The report should also contain a rating of the household's energy efficiency, as well as energy saving tips.

An example of the first energy report for a selected group of customers is shown in Figure 6.2. The following three reports were of the same structure and the order of the elements was similar. Variations of the components in future reports were foreseen.

The first element (a) is an introductory text to explain the meaning of this e-mail and the reason why users receive it. The second element (b) compares the electricity consumption in the last quarter with the same period one year before. The trend figure is additionally compared to the trend of all available households

in the city. When electricity consumption data of less than 15 months (one year and one quarter) of consumption data are available, customers receive a comparison of their consumption with regard to that of all available household in the city. Element (c) the household efficiency rating, calculated based on households with similar household characteristics, as long as the customer inserted details on its household (household type and size, space and water heating types) on the portal before. If this information was not present, a Call-to-Action (CTA) button was displayed that asked the receivers to insert their household characteristics online. The total consumption in the last quarter and an analogy (d) what amount of energy is used, for instance to prepare a pizza, is shown that users have a relation to actions of their everyday life. As the last element (e) of the report, seasonal energy saving tips are displayed and finally, the e-mail footer contains a CTA to give feedback to the report and necessary legal details about the company. Below of all content, CTA links to the corresponding pages functionalities of the online portal is placed to lower the barrier for further analysis of the electricity consumption online.

6.3 Experimental and survey-based evaluation of the energy report

The energy efficiency portal and the customers that are subject to this study belong to an utility company in Switzerland operating in a city with approximately 25,000 inhabitants. The data collection, the field experiment and a customer survey took place during the year 2017. An overview to the study timeline, the send-out dates of the energy reports and the customer survey is illustrated in Figure 6.3. This section describes the timeline of the experiment, the sample of study participants, and the conducted customer survey.

6.3.1 Timeline of the study

The study included an experiment and a customer survey. In the experiment the effect of the energy efficiency report on the user behaviour of the customers, the energy consumption was examined. The experiment run in three phases that are also depicted in Figure 6.3:

1. Baseline / No-Intervention-Phase (January–April 2017): During this period, energy consumption and portal usage was measured without the influence of energy reports. The data serves as baseline reference points for expected behavioral changes through the experiments.

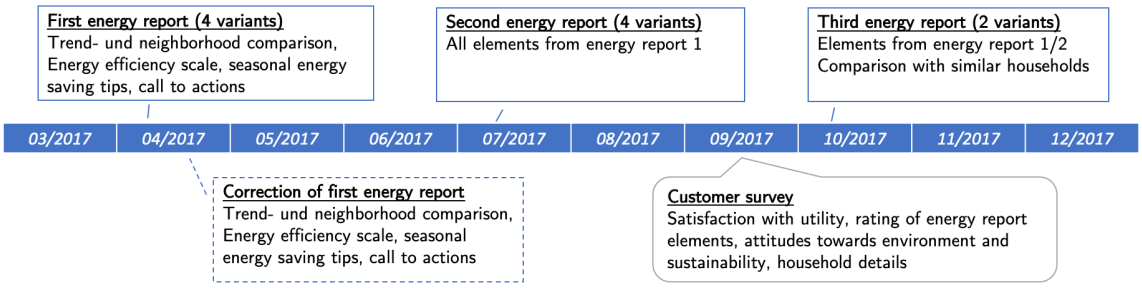


Figure 6.3: Timeline of the experiment

2. First experiment with intervention based on *known data* (April–September 2017). A selected group (approximately 55% of the portal users with daily electricity meters) received the first version of the energy report with general consumption feedback (comparison of electricity consumption with prior year quarter and neighborhood, energy efficiency scale, and seasonal energy saving tips).

3. Second experiment with intervention based on *predicted data* (October–December 2017). All customers with daily electricity meters receive an extended version of the energy report with an energy feedback element that compares the household with similar neighbors. For customers that have not provided household detail, data was predicted with the household classification approach described in chapter 5.

In September 2017, a customer survey was conducted to gather data on customer satisfaction, attitudes towards the environment and sustainability, and the experience gained with the energy report.

6.3.2 Sample description

In the beginning of the accompanying research, $n = 969$ customers were registered in the energy efficiency portal. 562 of them have registered within one month after an "activation mailing" that contained a motivation to use the portal. This activation mailing was sent on December 3, 2013 to 7,043 customers of the utility company. It contained feedback on the energy consumption of the household and a comparison to the average consumption of household in the city. Since then, further customers registered that got notice of the portal via other

channels (website of the utility company, information mail after newly installed meter, etc.) as shown in Figure 6.4.

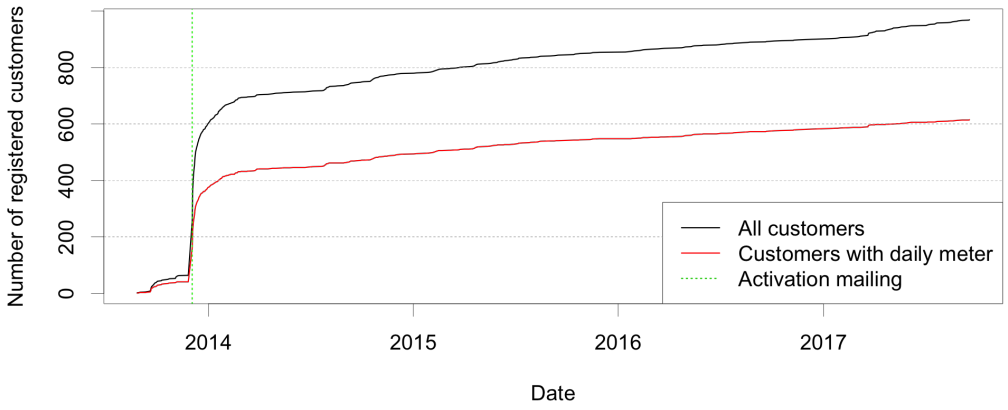


Figure 6.4: Number of registered customers on energy efficiency portal

In preparation of the experiment, all households that had been registered on the energy efficiency portal were separated into groups before the experiment started. The group distribution is shown in Table 6.1. In the experiment, only households that are equipped with a daily electricity meter ($n = 414$) received any consumption feedback during the experiment, because the energy report contains feedback based on daily energy consumption data. Households with yearly meter readings ($n = 555$) are available as a baseline group for several analyses, but not for the analysis of energy consumption. A randomly selected group of customers received the energy consumption feedback from the beginning of our experiment in April ($n = 263$). I call it *early treatment group* ($\mathcal{A}$) The remaining households with daily meters installed received their first energy report in October ($n = 151$). I call this group *late treatment group*($\mathcal{B}$). The latter group serves as a control group for effects on dependent variables in the analysis.

For the goal of this study, it was relevant to separate portal users that inserted data on their household characteristics within the portal and those who did not. Table 6.1 therefore lists the distribution of customers in columns “given” or “not given” respectively.

Randomization checks on several variables (e.g., electricity consumption 2016, known households characteristics) have been made, but customers are added

Table 6.1: Customer groups and sample sizes at the beginning of the study

Customer group		Group	Date of first report	Household details given	Household details not given	Household details total
portal user ($n = 969$)	daily meter ($n = 414$)	$\mathcal{A}$	April	151	112	263
		$\mathcal{B}$	October	75	76	151
	yearly meter ($n = 555$)	$\mathcal{C}$	(no)	259	296	555

later to the experiment groups due to new registrations on the web portal and new rolled-out daily meters. Considering the 672 customers that already registered in 2012 where the annual electricity consumption is available, a t-test was calculated to ensure that there is no statistical difference in the electricity consumption between the groups that receive a mailing and those who not ($t(629) = 1.1464, p = 0.2521$). Between the group of customers that received the first mailing in April and those received it in October is also no significant difference in the electricity consumption with $t(205) = -0.090084, p = 0.9283$. Nevertheless, households registered in such investigated energy efficiency portals typically have a higher energy consumption than customers that do not respond to this service (Lossin 2016, chapter 4).

6.3.3 Customer survey

A customer survey was conducted to complete the data that was collected throughout the experiments. There, we raised the user opinions regarding the energy report, their attitudes, and other variables through a survey that we conducted in September 2017 among all customers that received the energy report and those who did not receive any report, but were registered in the energy efficiency portal. The survey covered the following topics:

- User acceptance and feedback regarding the energy report (only customer that received the report)
- Customer based reputation of the utility company
- Attitudes towards energy efficiency
- Household details and socio-demographics

The invitations to the survey were sent on September 14 via e-mail to all customers that are registered on the web portal ($n = 969$). Customers that

received an energy report have got a different survey with questions regarding the energy report. The mapping of the survey responses to customer data was done with personalized invitation links to the survey. The survey was technically implemented using the SurveyGizmo platform[2]. In total, 96 customers answered the survey. 73% of the respondents were male and the average age was $M = 55.99, SD = 15.01$. 26 of the respondents received at least the second energy report during the experiment and answered questions regarding the report. The findings are described below.

6.4 Analysis of the first experiment and results for the base energy report

In this section, results from the mailing experiment as well as the customer survey are presented and answers to the RQ are given. In detail, the impact of the mailings in terms of customer reactions to the mailings, the portal usage, and the energy consumption is investigated and the survey is evaluated with descriptive statistics.

6.4.1 Customer reactions to the energy report mailing

As reactions to the mailing, all opening events of the the e-mail and clicks on hyperlinks in each e-mail are considered. The links provided in the e-mail have an explicit CTA (i.e., buttons highlighted text). The customer reactions on the different mailings are visualized in Figure 6.5.

Each link has an unique identifier, so that clicks can be associated with the e-mail a customer received. A detailed overview to number of emails sent, open and click rates, such as the time until first user interactions are shown in Table 6.2. The first energy report was sent in four tranches, later emails were sent at one time. Since the first e-mail contained a mistake in the calculation of consumption trends, a corrected version was sent on April 26 to all recipients of the first energy report.

6.4.2 Portal usage

One major goal of the energy reports is to attract customers using the energy efficiency web portal. To examine this goal achievement, I analyze the number of user sessions in the portal and the amount of changes in user data.

[2] `http://www.surveygizmo.com/`, last visited 30.12.2017

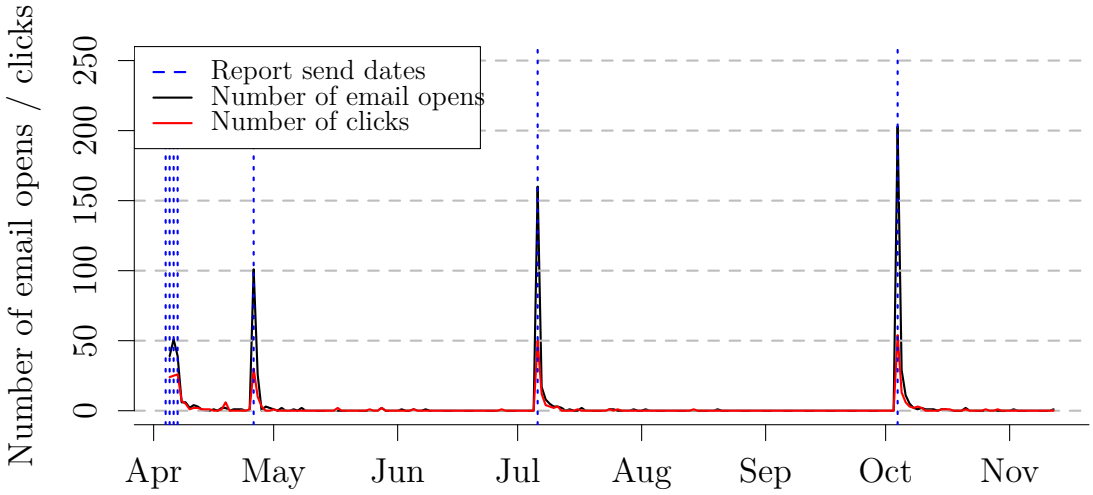


Figure 6.5: Customer reactions to the energy report (e-mail open and click events)

User sessions The first variable of interest is the number of *user sessions* on the web portal in a period of three weeks beginning with the date when a report was sent. A session always starts with an user login. The number of sessions per user was extracted from portal log data and a session identifier was used to matched the activity with user accounts. Figure 6.6 shows the number of user sessions in time intervals after each mailing for four separate customer groups: received the energy report (yes/no) and provided household details on the portal (yes/no). The latter can be a proxy on how much the customer got motivated to use the web portal. For comparison, two example periods from January and February 2017, before the intervention started, and two periods between the reports in May and September are included in the figure.

In addition to the average number of sessions per user in each group, the figure also shows the average number of sessions without the sessions that have been triggered by a click on a CTA element in the e-mail. The comparison of these sessions with the total number of sessions gives an impression, to what extent the energy report has an impact on the portal usage.

The results indicate that the energy report has a positive influence on the portal usage, as the average number of sessions per user has increased in group $\mathcal{A}$ (early treatment group) after each report. The average number of sessions per customer in $\mathcal{A}$ increased from $M = 0.1068$ ($SD = 0.6749$) in the 12 weeks before the experiment to $M = 0.2079$ ($SD = 0.9081$) in the three weeks after the first report. This effect is significant with $t(725.76) = -2.2926, p < 0.05$. The effect is

Table 6.2: Customer reactions to three energy reports

Date	Energy report	Amount mails	Open rate	Click rate	Median hours to open	first click
2017-04-04	First, 1. tranche	66	66.7%	33.3%	3.31	6.79
2017-04-05	First, 2. tranche	66	75.8%	37.9%	4.93	5.91
2017-04-06	First, 3. tranche	65	78.5%	32.3%	2.35	8.84
2017-04-07	First, 4. tranche	68	64.7%	32.4%	3.90	5.53
2017-04-26	First, correction	195	72.3%	19.5%	2.10	2.61
2017-07-06	Second	289	70.9%	23.5%	3.19	4.38
2017-10-04	Third	417	62.8%	18.0%	3.66	7.34

also visible comparing the number of sessions of group $\mathcal{A}$ with sessions of group $\mathcal{B}$ that received their first report in October Figure 6.6. This effect is not statistical significant, because customers in group $\mathcal{A}$ visited the web portal also slightly more often in the three weeks after the first report $M = 0.1369$ ($SD = 2.5256$) compared with the time before any reports $M = 0.0650$ ($SD = 0.6179$). I attribute this increase in sessions to external factors (e.g., electricity bill or cold winter), because theses customers did not receive an energy report.

Interestingly, customers that provided household details on the portal increased their presence on the portal even when the sessions that follow a click on a CTA in the e-mail are subtracted (see red dashed line in Figure 6.6). Besides, the report also attracted customers that did not provide any data in the portal (which I assume is an indicator for less interest in the portal), even if this effect is small.

The activation effect of the energy report to use the portal more frequently seems to be stable over time, as households that received the energy report have more sessions on the portal (even adjusted with the sessions triggered by mails) than households without receiving the report in the later reports. The effect is only significant in the group of households that are anyway active on the portal. With a large enough sample, future research may investigate this effect in more detail.

Number of changes in portal data As a second aspect, the number of changes in household details provided by users on the portal is investigated. Figure 6.7 shows the number of changes made in 2017 and in the years 2014–2016 on average

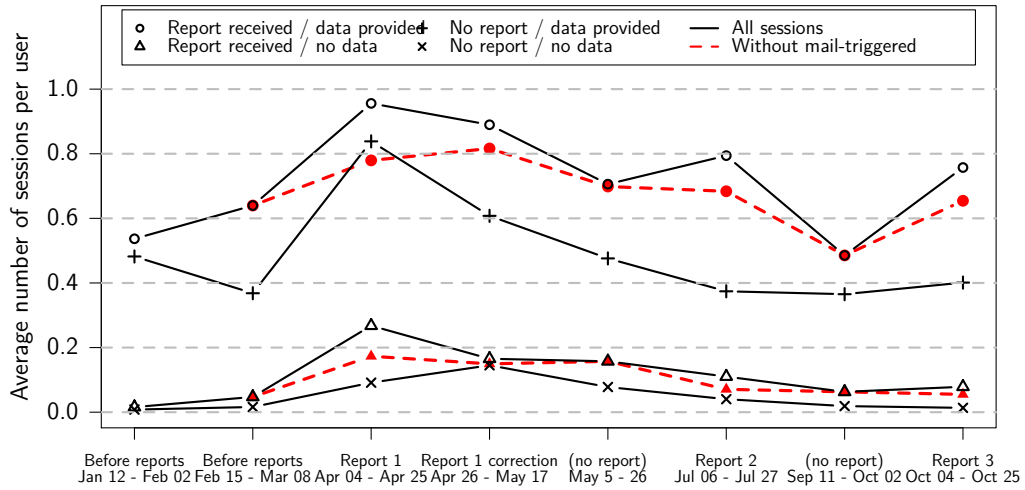


Figure 6.6: Portal user sessions in three week timespans after each energy report mailing was sent and two timespans without a report for comparison

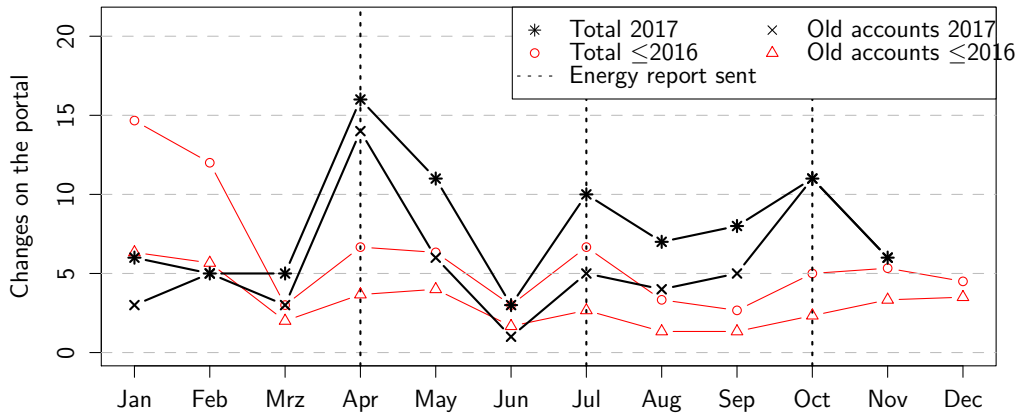


Figure 6.7: Changes in household details by users per month

in each month. The detailed recording mechanism of changes in household portal data was implemented in March 2017. Before this functionality existed, only the timestamp of the last change was saved in the database. The number of change-events before March 2017 in this analysis is therefore lower than it actually was. In total, users made $M = 1.14$ ($SD = 0.39$) changes after March 2017, including the first insert of household data when users register. Thus, we can interpret at the total number of changes as an indicator for user involvement. Furthermore, the metric is critical, since users must be willing to provide household data which are personal information and are not provided to everyone.

It is obvious that changes are frequently made in the month of an energy report. The number of changes in 2017 is clearly higher than the average of earlier years, even when the figures of earlier years are biased. Two aspects are in particular interesting: First, most changes are made after a household receives its first energy report. This is visible in April, when $\mathcal{A}$ received the first e-mail. In July, when the same group received a second e-mail, the number of changes is as low as in the years before. After $\mathcal{B}$ received the e-mail in October, an increase in changes is also visible for this study participants. Second, it seems that the report activates especially users who have longer been inactive. The number of changes made to old[3] accounts represent the majority of changes in 2017.

Summary The energy report clearly increased the portal usage. This effect lasts even beyond the first click on a CTA in the e-mail. Besides, customers updated their user data and provided more details in the portal. Particularly users with old accounts inserted new data. The energy report is therefore a good tool to reactivate inactive customers and collect data on those households.

6.4.3 Electricity consumption

The energy report aims to get customers motivated to conserve electricity. Therefore, the impact of the energy feedback intervention on the electricity consumption of all households in the study is investigated. For this analysis, the daily electricity consumption[4] for each customer is available. I consider only customers that had a daily electricity meter installed at the beginning of the

[3] Accounts are considered as 'old' when a change is made to an account that existed for more than 16 days.

[4] The total consumed electricity was measured by smart meters separately for HT and NT times. If the reading for one or more days was missing, the reading count of each meter for these days was linearly interpolated based on the known reading before and after the missing interval. This case of missing readings occurs far less than once in 1,000 cases. As multiple meters per customer exist, all meter readings are summarized to one daily total

experiment. 16 consumption values less than zero and 239 outliers from a total number of 180,538 daily consumption measurements are excluded. The outliers were identified following the *two-sigma-rule*, so values higher than the mean ($M = 12.77$kWh) plus double standard deviation ($SD = 121.50$kWh) have been excluded.

Weather data from the Global Hourly Integrated Surface dataset provided by US National Centers for Environmental Information[5] is used in the analysis. The hourly temperature readings are converted from °F to °C and the average temperature for each day was calculated.

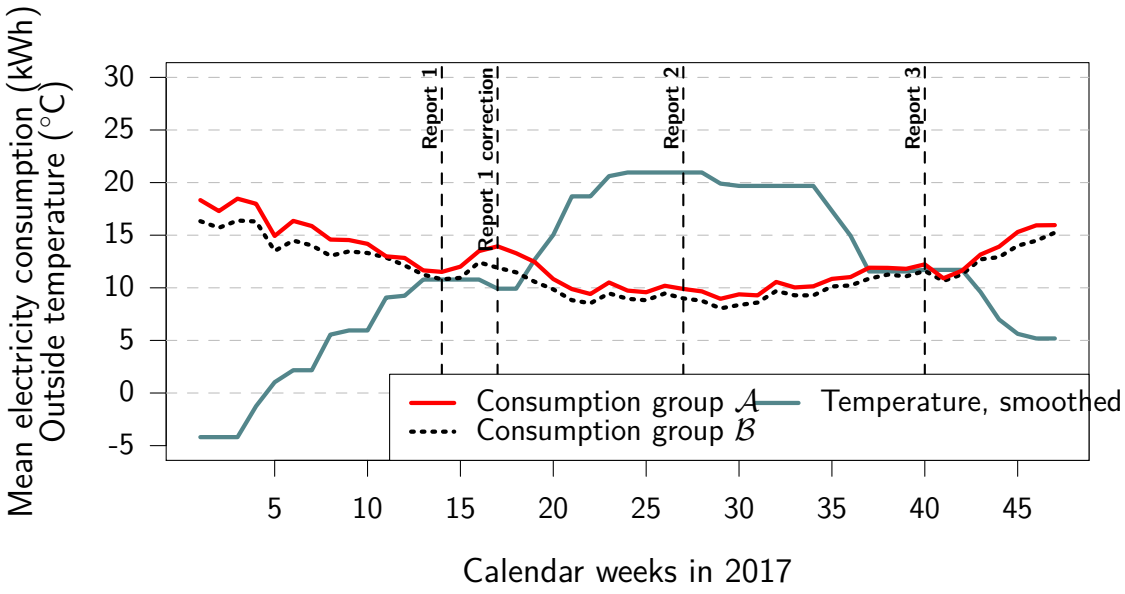


Figure 6.8: Electricity consumption of both experiment groups with temperature and dates of the energy report interventions

Descriptive interpretation Figure 6.8 shows the average electricity consumption of customers in group $\mathcal{A}$ and $\mathcal{B}$, together with the average outside temperature in the study region (a single town in Switzerland) and the dates when the energy reports have been sent. The experimental group $\mathcal{A}$ shows a slightly higher electricity consumption ($M = 15.01, SD = 15.57$) than the group $\mathcal{B}$ ($M = 14.00, SD = 11.36$), but this difference is not statistically significant

meter count for this analysis. Finally, the daily electricity consumption was calculated considering the meter reading difference between one day and the day before.

[5] `https://www.ncei.noaa.gov/data/global-hourly/archive/`, last accessed 14.01.2018

($t(358) = 0.74068, p = 0.4594$). From the figure, only a marginal energy saving effect after the three treatments in April and July can be recognized. This is aggravated by the change in temperature over the seasons that affects to some extend the electricity consumption.

Difference-in-Differences (DiD) analysis To investigate the feedback intervention in more detail, DiD models are estimated using Ordinary Least Squares (OLS) linear regression with the daily electricity consumption of each household and the effect of the energy feedback intervention. Table 6.3 shows the estimated coefficients and the goodness of fit of three different models that are explained below.

Table 6.3: Difference-in-Differences models explaining the daily electricity consumption of households in both experiment groups

	Model 1	Model 2	Model 3
(Intercept)	13.68***	5.86***	6.69***
	(0.11)	(0.09)	(0.09)
TEMP_Celsius			−0.25***
			(0.01)
baseline_cons_el		0.56***	0.56***
		(0.00)	(0.00)
d_phaseT1	−3.81***	−3.40***	−0.49***
	(0.15)	(0.12)	(0.13)
d_phaseT2	−4.21***	−3.75***	−0.25
	(0.15)	(0.12)	(0.14)
d_groupA	1.50***	0.87***	0.83***
	(0.14)	(0.11)	(0.10)
d_phaseT1:d_groupA	−0.12	−0.54***	−0.51***
	(0.19)	(0.15)	(0.15)
d_phaseT2:d_groupA	−0.49*	−0.89***	−0.85***
	(0.19)	(0.15)	(0.15)
R^2	0.03	0.41	0.42
Adj. R^2	0.03	0.41	0.42
F-statistic	***	***	***
Num. obs.	112244	112244	112244
RMSE	12.52	9.73	9.63

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The first model includes only dummy variables that assign the observations with the time interval and the experiment group. In detail, *d_phaseT1* describes observations in the first intervention phase after the first energy report on April 4, *d_phaseT2* describes observations in the second intervention phase after the second energy report on July 6 and *d_groupA* indicates observations that belong to experiment group $\mathcal{A}$ that receives the energy reports. Interactions between the dummy variables are considered between all phases and the experiment groups, and are denoted with a colon. The first model indicates no significant effect after the first energy report, but a small effect after the second report. Admittedly, the model is weak and explains only 3% of the variance (see R^2 in Table 6.3). Therefore, the average daily electricity consumption before the experiment, between Jan 2 and Apr 3, (Variable *baseline_cons_el*) is included in the second Model, and the average daily outside temperature in Celsius (Variable *TEMP_Celsius*) additionally in the third Model.

The three models show a significant reduction of the energy consumption of 6%-13% compared to the baseline consumption of households in group $\mathcal{A}$. [6] This becomes visible in Figure 6.8. Group $\mathcal{A}$ decreased its energy consumption more than group $\mathcal{A}$ over time until November (week 40). The result of this analysis are interesting, as it shows that personalized electricity consumption feedback can lead to energy savings with a low effort by utility companies.

One must be careful by interpreting the results, because the number of households considered in this study is limited (263 received the energy report, 151 served as control group), the models used for explanation indicate significant effects, but they are associated with the high number of observations that result from the time-series data.

Summary Even if one cannot speak of a robust effect, the data indicate that households received an energy report consumed 6% less electricity than the control group in the same time span. Nevertheless, saving effects that are documented in earlier electricity consumption feedback studies using energy reports, in form of letters (Allcott 2011), lead to savings of 2% with a sample of 600,000 households. I conclude from the results, that the digital version of such energy reports can at least replicate these energy saving effects.

[6] According to Model 3, the electricity consumption of group $\mathcal{A}$ in the baseline phase is $6.69 + 0.83 = 7.52$kWh and it is reduced by 0.51kWh (6.78%) after the first report. Considering the electricity consumption of this group in the phase *T1*, which is $6.69+0.83-0.49-0.51 = 6.52$kWh, it is additionally reduced by 0.85kWh (13.04%) after the second report.

6.4.4 Usability and user perception of the report

The energy report was well accepted by all customers that answered the survey. In a question regarding the preferred frequency in which the participants like to receive the report, eight (29%) said they like to receive the report once a quarter (this is the choosen interval), more than half would appreciate to receive the report more often (14 choose "monthly" and one choose "weekly"). Three customers prefer the report half or once a year and only two customers said they do not want to receive the report at all.

To investigate the user acceptance in more detail, the *user engagement* and *ease of use* were measured using the Usability Perception Scale (UPscale), a recently published and tested measurement instrument for perceived usability of eco-feedback (Karlin and Ford 2013) that I list in Table 6.4.

Table 6.4: UPScale: Items and descriptive statistics obtained in the customer survey

Survey items		*M*	*SD*
UP1	I am able to get the information I need easily	4.11	0.74
UP2	I think the image (energy report) is difficult to understand*	1.70	0.72
UP3	I feel very confident interpreting the information in this image (energy report)	4.19	0.79
UP4	A person would need to learn a lot in order to understand this image (energy report)*	1.70	0.61
UP5	I gained information from this image (energy report) that will benefit my life	2.93	1.00
UP6	I do not find this image (energy report) useful*	1.59	0.93
UP7	I think that I would like to use this image (energy report) frequently	3.22	1.25
UP8	I would not want to use this image (energy report)	4.11	1.01

* negative question

In contrast to the well-known usability scale (SUS) by Brooke et al. (1996), the UPScale is a specific scale for energy feedback context and fits better to the research pursued. A limitation to the scale is, however, that Karlin and Ford (2013) do not provide concrete estimates for single items or both measured factors. This makes it impossible to compare user ratings of the energy report with their research. The original scale was published in English. For our survey, the questions were translated to German. The translation procedure and the German version of the scale is shown in section C.4 on page 223. As the original scale was developed for one single visualization. As we evaluated multiple images

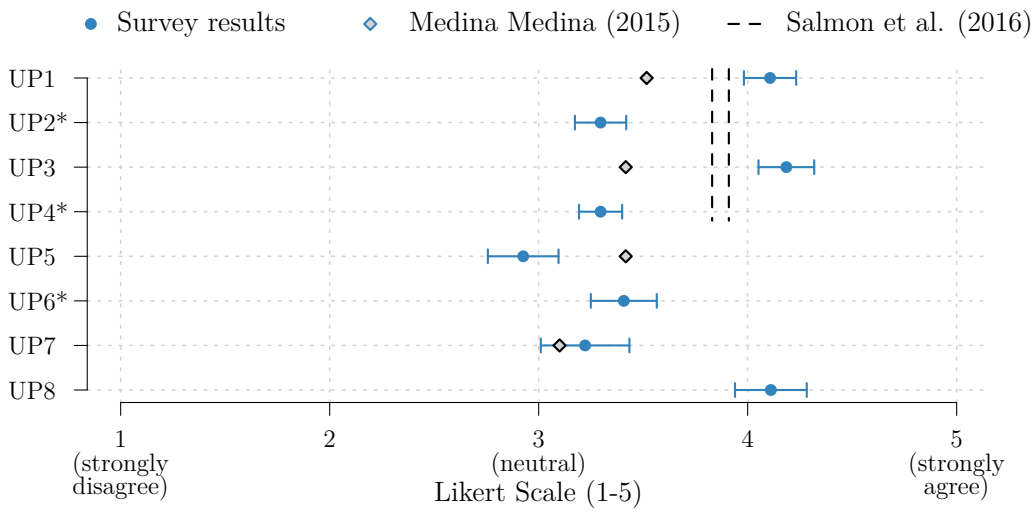


Figure 6.9: Survey results for UPScale with results from previous studies, (* indicates transformed negative questions)

combined in one energy report, we replaced the word "image" by "energy report" in all questions.

The detailed survey results are depicted in Figure 6.9. The factor *ease of use* is $M = 3.72$ ($SD = 0.49$) and *engagement* is $M = 2.61$ ($SD = 0.49$). Remarkably, the user ratings for all negative questions (UP2, UP4 and UP6) received less agreement, which might be a specialty of this instrument. The figure contains also the results of two earlier studies that I could identify. First, Medina Medina (2016) uses the items UP1, UP3, UP5, and UP7 to evaluate feedback in learning environments with $n = 24$ students. Neither a factor of the UPScale nor the variance of the single items are reported. Second, Salmon et al. (2016) use items UP1–UP4 for the factor *ease of use* to evaluate a campus energy education dashboard with $n = 277$ Amazon Mechanical Turk participants, but document only estimates of the whole factor of $M = 3.83$ ($SD = 0.75$) for a bar chart and of $M = 3.93$ ($SD = 0.74$) for a map visualization. Statistics on single items are not reported. The ease of use rating in this study is slightly higher than in the investigated case with $t(26) = -1.6638, d = 0.2120, p = 0.05408$, but I assume that the Mechanical Turk participants rate an energy feedback element other than utility customers who receive an illustration on their own energy consumption.

I conclude that the energy report was rated positively, even if the factor *ease of use* is slightly lower as an earlier study. The majority or survey participants want the energy report once a quarter or more often.

6.4.5 Customer satisfaction with the utility company

Another important dimension in this study is the satisfaction of customers with the utility company. For this, two separate instruments have been used: the Customer Based Reputation (CBR) of a service firm and the Net Promoter Score (NPS).

Walsh and Beatty (2007) define the corporate reputation as a multi-dimensional attitude and measure it among five dimensions: 1) Customer Orientation, 2) Good Employer, 3) Reliable and Financially Strong Company, 4) Product and Service Quality, 5) Social and Environmental Responsibility. The measurement instrument was initially developed in the US market. In a later study, Walsh, Beatty, and Shiu (2009) validated it in UK and Germany, and reduced the number of items to 15 (three for each factor). In our study, items from the factors Customer Orientation (1), Reliable and Financially Strong Company (3), and Product and Service Quality (4) were chosen. German translations of the questions have been provided by Gianfranco Walch[7]. Besides that, three questions from the provided translations were included that have not been reported in the studies of Walch and colleagues, but were reasonable in our context. As the electricity market in Switzerland is not liberalized and consists of regional monopolies, the questions regarding competitors or assuming a free market were skipped. Original questions and German translations are listed in section C.2 (page 220). The used items in this study together with descriptive statistics Table 6.5.

Figure 6.10 shows the results obtained in the customers survey among all responses and separated into the experiment groups $\mathcal{A}$ (received energy report before the survey) and $\mathcal{B}$ (did not receive an energy report before the survey). Results of the CBR scale from earlier studies in other industries (Walsh, Beatty, and Shiu 2009) are depicted for better interpretation of the results. The overall customer satisfaction is slightly higher among the customers in group $\mathcal{A}$ that received the mailing ($M = 4.27, SD = 0.59$) than in group $\mathcal{B}$ ($M = 4.07, SD = 0.72$), but this difference is not significant with $\alpha < 0.1$. This low difference can be ascribed to the fact that the CBR scale has generally a low variance when only one company is subject to an investigation[7].

[7]E-mail communication with G. Walsh on July 28, 2017

Table 6.5: Selected items from CBR-Short Scale with descriptive statistics obtained in the customer survey

Survey item		M	SD
Factor 1: Customer Orientation			
REP1	Has employees who are concerned about customer needs	4.23	0.86
REP2	Has employees who treat customers courteously	4.36	0.79
REP3	Is concerned about its customers	4.09	0.89
Factor 3: Reliable and Financially Strong Company			
REP4	Seems to recognize and take advantage of market opportunities	3.84	0.92
Factor 4: Product and Service Quality			
REP5	Offers high quality products and services	4.14	0.76
REP7	Develops innovative services	3.81	1.05
Other items, not included in original scale (Walsh and Beatty 2007)			
REP8	I am satisfied with the services that the company offers	4.18	0.95
REP9	You can trust this company	4.12	0.82
REP10	I would probably report good things about the company to others	4.26	0.80

As a second metric for the customer satisfaction, the NPS is often used as a predictor for the success of a company. The metric goes back to research from Reichheld (2003) and raises customer loyalty with one single item "How likely is it that you would recommend our company/product/service to a friend or colleague?" with a 11-point likert scale (from 0 to 10).

To calculate the NPS of a company, the relative number of responses in the interval of $[0, \ldots, 6]$ (called "detractors") is subtracted from the relative number of responses with 9 or 10 (called "promoters"). The remaining responses are called "passively satisfied" and not further counted in the metric. "Companies that garner world-class loyalty receive net-promoter scores of 75% to more than 80%" (Reichheld 2003). The scores vary widely through different industries and I could not identify a reliable source for typical NPS scores in the utility industry (especially in non-liberalized markets). According to Delighted Inc. (2017), the scores for twelve US utilities are on average $NPS = 27$ and range between 5 and 41. Nonetheless, I doubt that these numbers are representative, as their collection is not documented and the sample is unclear.

With 18 promoters and 13 detractors from 48 respondents that answered the question for the NPS, the utility achieved $NPS = 10.4$ with is a comparable low

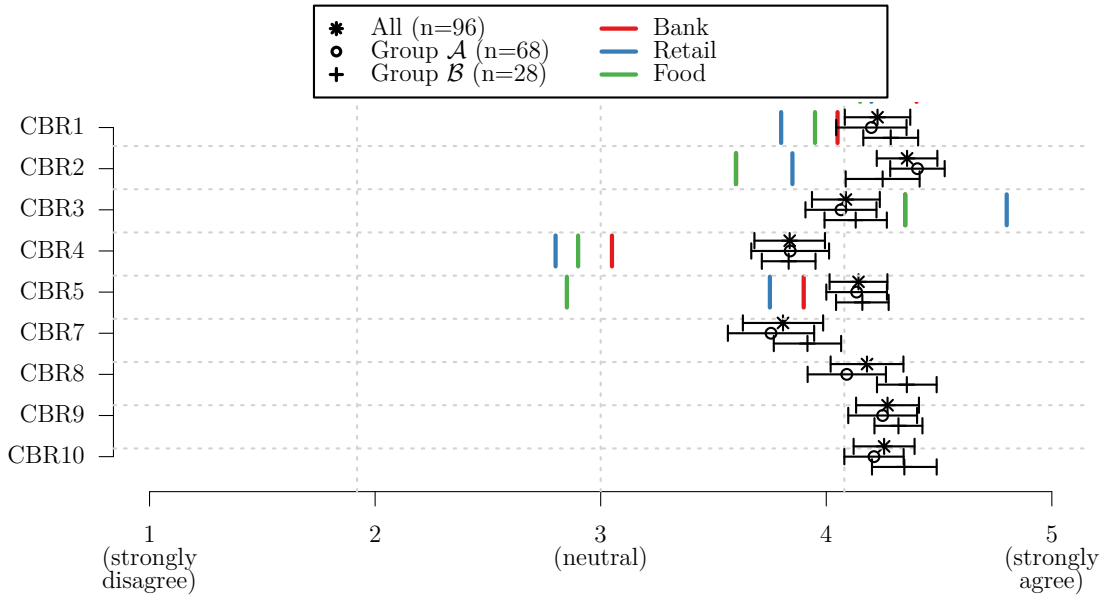


Figure 6.10: Survey results for CBR-Short scale with benchmark values from other industry branches

value. It remains questionable, how valid this metric is in a monopoly market like Switzerland, when customers cannot choose their utility company.

6.4.6 Attitudes towards energy conservation

Finally, customers were asked about their attitudes towards the environment and energy conservation. I compare the answers in both experimental groups. For this, an instrument developed by Kaiser et al. (2007) and one item about environmental awareness from Diekmann and Franzen (1999) was used.

The detailed questions are listed in Table 6.6. Questions BEC2–BEC6 are taken from the factor *energy conservation* in Kaiser et al. (2007). The utility company desired to skip statement BEC1 ("After one day of use, my sweaters or trousers go into the laundry"), because of the apprehension that this could be an extraordinarily personal aspect of hygiene. Translations for the remaining five items (BEC2–BEC7) are taken from the earlier survey that belongs to dataset C (see subsection 5.1.3). The translation for the sixth item was obtained from a Swiss environmental survey (Diekmann and Bruderer Enzler 2012, p. 57). German translation can be found in Table C.1 on 219.

There are small differences in the answers between group $\mathcal{A}$ and $\mathcal{B}$, but considering the average of all items, the difference is not statistically significant

Table 6.6: Items for attitudes towards energy conservation with descriptive statistics obtained in the customer survey

Survey items		M	SD
BEC2	As the last person to leave a room, I switch off the lights	4.62	0.55
BEC3	I leave electrically powered appliances (TV, stereo, printer) on standby*	2.79	1.41
BEC4	In the winter, I turn down the heat when I leave my room for more than 4 hours	1.80	1.10
BEC5	In the winter, it is warm enough in my room to only wear a T-shirt*	2.27	1.27
BEC6	In hotels, I have the towels changed daily*	1.46	0.81
BEC7	I do what is right for the environment, even when it costs more money or takes more time.	3.09	1.09

* negative question

calculating a t-test. Nevertheless, the last item shows small effect: group $\mathcal{A}$ shows higher agreement with BEC7 ($M = 3.21, SD = 0.96$ instead of $M = 2.58, SD = 1.12$ in group $\mathcal{B}$ $t(34.55) = -2.0206, p < 0.05$).

6.5 Analysis of the second experiment and contribution of household classification

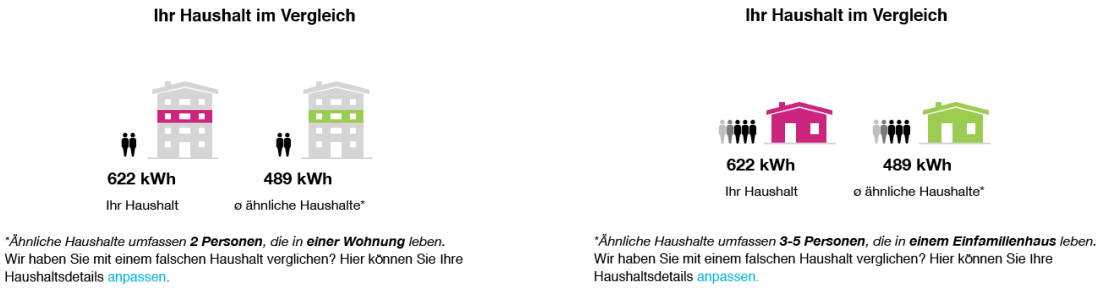
In the third phase of this ancillary research (beginning in October 2017), the energy report was developed further. A new feedback element was designed, that contains a comparison the electricity consumption of the receiving household with similar ones. The new element was tested in the second experiment among two groups of customers: those who have inserted household details (e.g., household type and size, number of residents, heating types), named $\mathcal{D}$ in the following, and those who have not $\overline{\mathcal{D}}$.

With this experiment, I tested the capabilities of household classification for targeted energy efficiency campaigns. The new feedback element, developed for the third energy report, showed an electricity consumption benchmark of household with a relevant comparison group (social-normative feedback). All households that inserted the necessary data in the efficiency portal received the comparison to a reference group based on known data. For all other customers, the household classification approach was used to predict the household details. I describe the element below and argue why this combination of available household variables (household type together with number of residents) was chosen.

The details on the prediction of household characteristics are described in subsection 5.5.2.

6.5.1 Personalized feedback element for comparison with similar households

Before the additional feedback element that compares the receiving household with similar ones could be designed, we had to define the comparison group of "similar household". Thereby, the following considerations were made: The feedback should be easy to interpret by customers, the comparison should be feasible to implement with a given technology stack and it must be possible to find such a comparison group for all households that is large enough to estimate the average electricity consumption. Besides, the variables used to define the comparison groups need a limited set of values and the considered household details should have significant impact on the electricity consumption, so that the receiver is able to realize the feedback. We also preferred household characteristics that ML algorithms can predicted with a low error rate.



(a) Apartment with two residents (b) House with three to five residents

Figure 6.11: Variants of the social-normative feedback element with comparison of the energy consumption with similar households (Source: BEN Energy)

In the study team, and together with the research partner, we selected the variables *household type* (either "house" or "apartment") and *number of residents* (with the classes "1 person", "2 persons", "$3-5$ persons", and "≥ 5 persons") in combination. To confirm our decision, we estimated an OLS linear regression model (which is not relevant for the study and therefore not further explicated here) explaining the electricity consumption of all households in the sample. In that model, we found that more than a fifth of the variance in average

electricity can be explained with this combination of household characteristics. In addition, the electricity consumption has a significant correlation with the number of residents ($r = .14, p = .043$) and the classification of the household detail combination achieves a comparably high prediction performance (see subsection 5.5.2).

The resulting element is shown in Figure 6.11. It has the headline "Your household in comparison", shows the number of residents and the household type visualized with pictograms. Below of the image, the comparison group is described with text and a CTA with a link to the energy efficiency portal is placed, where users can change their household details in the case of any wrong information.

6.5.2 Experiment setup

The third energy report was sent to 417 customers in total. Due to data changes in the portal, new registration in the portal, unsubscription of the energy report, or contract changes, only 400 are included in this experiment. Table 6.7 lists the distribution of customers that did or did not provide data on their household characteristics. For 174 customers without given data, the household classification procedure, described in subsection 5.5.2, was applied and the necessary data are predicted based on a machine learning model that has been trained with labeled customer data. No prediction could be made for 13 customers because of missing data.

Table 6.7: Distribution of customers into groups for the second experiment

	No prediction	Prediction	Sum
Household data given ($\mathcal{D}$)	213	0	213
Household data not given ($\overline{\mathcal{D}}$)	13	174	187
Sum	226	177	400

6.5.3 Experiment results

To analyze the experiment and the difference in portal usage and energy consumption, I compare the behavior of study participants in the group $\mathcal{D}_{noPrediction}$ with $\overline{\mathcal{D}}_{prediction}$, shown in Figure 6.12. The open rates of the emails are therefore similar in both experiment groups. In group $\mathcal{D}_{noPrediction}$, 63.38% of the emails were opened and in group $\overline{\mathcal{D}}_{prediction}$, 62.64% ($\chi^2(1) = 0.0019, p = 0.9653$).

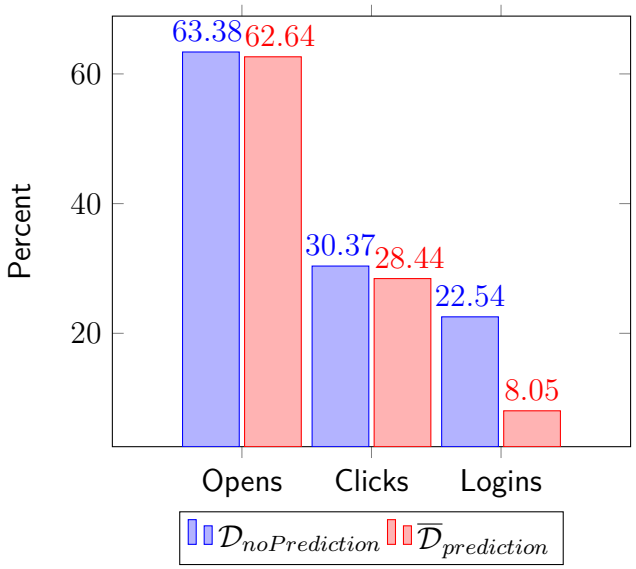


Figure 6.12: Reactions of both experiment groups to the third energy report e-mail

Considering all emails that were reported to be opened, the recipients in $\mathcal{D}_{noPrediction}$ clicked on a link in the e-mail with a ratio of 30.37% and those in group $\overline{\mathcal{D}}_{prediction}$ with 28.44%. The difference is not significant with $\chi^2(1) = 0.03514, p = 0.8513$. This shows that the source of information (i.e., self-reported vs. predicted data) in such an energy report has, if indeed, only a limited influence on the user's activation to use the efficiency portal of the website of the utility company.

The click events on CTAs yet let not to a strong increase in the use of the energy efficiency portal. While 22.54% of recipients in group $\overline{\mathcal{D}}_{prediction}$ that opened the e-mail started a session, only 8.05% of customers in group $\overline{\mathcal{D}}_{prediction}$ logged in (the numbers are calculated by considering three weeks after the report has been sent). The average number of sessions is lower, too: 0.0919 instead of 0.7652 in the group that provided data ($t(216.32) = 2.7264, p = 0.0069$). Interestingly, an equal number of five customers changed or inserted household data in the energy efficiency portal.

6.6 Discussion and implications

This chapter described the results of a research project conducted together with BEN Energy in which a personalized home energy report was developed and evaluated. The two-phase experiment and one customer survey was conduced

to answer RQ 4 (*Which added value can be realized from predicted customer characteristics on the example of personalized energy feedback?*) regarding the effectiveness a personalized energy feedback instrument and the benefit of ML algorithms to obtain customer characteristics to realized targeted feedback.

In detail, the results of this study demonstrate that an energy report e-mail which is regularly sent to customers is beneficial: In the present case, the open rates were always greater than 60%, and between 18% and 38% of receivers clicked on CTA links in the e-mail. These results were rated positively by the research partner who has experience in other emailing campaigns. Users that received the energy report increased their presence on the energy efficiency portal. This positive effect is even recognizable when the first click on an energy report e-mail is subtracted. Furthermore, the energy report attracted users to insert more data into the web portal. This holds especially for users with accounts older than 16 days. The energy report is therefore a powerfull tool to reactivate users. The results of my DiD analysis indicates that the customer group that received the report consumed 6% less electricity between April 4 and July 5, and 13% less electricity between July 6 and October 3, considering a control group of customers that did not receive such energy report.

In the customer survey that was conducted in September 2017 (between the second and the third energy report), the energy report was rated positively and customers mostly liked to receive the energy feedback in the current frequency every quarter or even more often. Customers that received the energy report documented a stronger satisfaction with the utility company and concern for the environment, but the differences to customers that did not receive the energy report are small.

Answer to RQ 4 In the second experiment, it was finally tested if predicted household characteristics lead to similar usage pattern than an energy report that is based on data that was self-reported by customers. In terms of open and click rates to the e-mail, no difference between the groups can be found. This supports the argument that the reports based on predicted data have the potential to trigger similar behavior than reports that are created on the base of data that was inserted by recipients. Further research, however, must investigate, why the number of user sessions of customers that have not inserted data was significantly lower than the activity of users that provided such data, because this observation can have different reasons.

The results of this investigation demonstrates that personalized energy efficiency campaigns are feasible on scale using residential household characteristics

that were predicted using ML models. The similar reactions on personalized mails based on compared to true information back this conclusion.

Limitations Despite the positive results presented, this study has limitations. First, the experiments were conducted in only one city in Switzerland and only with a special customer group (users of an energy efficiency portal that were equipped with a smart meter). Second, the number of study participants was low. This leads to the fact that some analyses could not be conducted. Third, the separation of study groups was—despite a number of randomization checks in preparation of the experiment—not a random split because of technical reasons of data matching and difficulties in the initial sending of the energy reports.

Future research I motivate further research on the topic, as several aspects could be investigated in more carefully designed experiments that consider findings and limitations of this study. This holds especially for experiments and data analysis to better investigate overlapping effects. For example, if customers that provided data are more frequently interacting with the web portal because of the more specific energy report possible, or because they were willing to enter their data into the portal. This could be one reason for the observation that users that inserted data in the portal had a higher likelihood to actually log in after having opened the energy report e-mail. Another question is, which of the feedback elements in the report was most successful. Moreover, the effects found in the time series data should be verified with panel regression methods as they are more suitable to estimate effects in time-series data. Finally, future work should investigate the long-term effects of such energy reports on the energy usage behavior of recipients.

7 Supporting cross-selling marketing campaigns with predictive analytics

Highlights

- In a Fiber-to-the-Home (FTTH) cross-selling case from the utility industry, I demonstrate how firms can use their existing customer data in combination with public available online data to predict the purchasing intention in relationship marketing campaigns.
- Location-related information (geographic and weather data) can significantly improve the prediction performance of the purchase intention model.
- Predictive analytics can address two important business problems in the context of relationship marketing: First, customers with a high likelihood to purchase a cross-selling product or service from the existing customer base can be identified; Second, the cost-benefit optimal number of households to be addressed in a campaign can be determined.

Cross-selling products and services are often offered in markets that are not within the companies' core business. The planning of cross-selling marketing campaigns in foreign markets is challenging, because it is unclear which customers are likely to respond to the offerings. The company may also not be familiar with the market and the purchasing behavior related to the offered product. Utility companies, for example, begin to offer Fiber-to-the-Home (FTTH) internet access to customers and enter a liberalized market that is still dominated by traditional telecommunication companies (Hongju Liu et al. 2010).

Despite promising benefits of cross-selling, including increased revenue, customer retention, and enhanced brand image, there is still room for improvement in performance of cross-selling campaigns in many companies (Li et al. 2011; Schmitz et al. 2014). Besides that, cross-selling to all customers is not a uni-

versal remedy, because addressing the wrong customers may lead to negative effects. There are, for example, four customer groups that lower the profit in companies and should be better left out in cross-selling campaigns (Shah and Kumar 2012; Shah, Kumar, et al. 2012): *Service demanders* (characterized by an overuse of service, this customer group might cause more service-demand per cross-buy), *revenue reversers* (try to retract gained revenue from companies, for example through early termination of contracts), *promotion maximizers* (mainly interested in discounts and not in the cross-selling offer), and *spending limiters* (avoid to increase their total expenses for one company and reduce the spending on initial products per each cross-buy).

The identification and targeting of early adopters within the existing customer base, on the contrary, is a good way to start cross-selling activities, because the contact to those customers already exists. Early adopters help to faster realize return on the investments through purchasing the product or service itself, but also influence peers that are motivated to likewise buy the product, thus speeding up the adoption process and maximizing turnover (Kamakura 2008). Therefore, one major challenge in cross-selling campaigns is to address the right customers (e.g., early adopters), which implies to identify customers with a high likelihood to purchase the product or service to be promoted, and to find the cost-benefit optimal number of customers to be addressed. For both, the *purchasing probability* of individual customers is a desired information. Estimating the purchasing probability of a new product is challenging, because no data on past purchases exists. As an alternative, customer surveys can be used to obtain purchase intentions of survey participants. This purchase intention can only give a initial estimate and cannot be directly considered as a purchasing probability. Besides, it is not cost-effective to conduct surveys with all customers.

The prediction of individual purchase intention scores is beneficial for marketing campaigns because a ranking of the customers can be created in order to answer the "whom address first?" question. In addition, it is relevant for decision making in marketing, which cost-benefit optimal number of customers should be included in a marketing campaign.

From the work on household classification (Beckel, Sadamori, Staake, et al. 2014; Hopf, Sodenkamp, and Staake 2018), it is known that characteristics of households (e.g., household type, dwelling size) can be revealed from electricity smart meter consumption data. The household classification studies so far are almost limited to physical household characteristics, but did not focus on intangible aspects, such as interests of customers. Moreover, plenty of ambient data can be analyzed, for example, geographic data and weather data can be used to reveal household characteristics and intentions. Having shown in chapter 5

(RQ 3) that customer intentions (i.e., interest in sustainability and solar installations) are possible to be predicted from ambient data that is available to energy retailing companies, it can be assumed that customers with a high purchase intention towards a product or service can be identified likewise. In addition, it is known that predictive segmentation in marketing (Banasiewicz 2013) can obtain scores for individual customers that express likelihoods for customer actions, based on marketing databases. Scores alone, however, do not help to improve marketing. I therefore investigate the following RQ in this chapter:

RQ 5 *Which added value can be realized from predicted customer intentions on the example of relationship marketing?*

This study showcases a general framework for decision support in cross-selling marketing campaigns through predictive analytics for the utility industry and instantiates this by means of a case study of promoting a FTTH cross-selling product by a utility company. This method can be—with some adjustments regarding the data and the feature definition—also applied in domains outside energy retail for the prediction of purchase probabilities based on data that a company already owns or which is available.

The remainder of this chapter is structured as follows: First, the relevance of the FTTH technology for utility companies and customers is outlined. The experimental data and some descriptive insights on the interest of customers towards the purchase of a FTTH internet access from the utility company are presented. The predictive analytics method is then applied to answer the stated research questions. Finally, the statistical model of Sun and Morwitz (2010) is instantiated for the purchase intention scores obtained through the predictive analytics method.

7.1 Fiber-to-the-Home (FTTH) as a relevant product for utility companies

FTTH describes the final expansion stage of the fiber optic network, the so-called "last mile", laid to private dwellings. So far, broadband internet access technologies for the residential homes and small businesses has mainly relied on twisted pair copper wires or coaxial cables, where fiber optic networks connected only distribution points for internet connections of endpoints (Green 2004; Frigo et al. 2004).

Larger penetration of the FTTH technology in the telecommunication market has economic, social, and environmental effects thus can support a corporate

strategy of the "Triple Bottom Line" (Elkington 1994). So, the advantage for private customers to have faster internet connections is a *social* benefit. This becomes visible in the broadly conceived political initiatives in Europe, for example in Germany (TÜV Rheinland Consulting GmbH 2016) and the US, for example in Wisconsin (2018), where governments are encouraging the broadband expansion with subsidies. The rashly developing Internet of Things (IoT) applications (including connected home with its automated energy management, automated domestic chores and security, as well as telemedicine, and distance learning) will generate significant upstream bandwidth demands already in the nearest future, on the order of a few hundred megabytes per second (Deichmann et al. 2015). FTTH has also broader *economic* effects. Atasoy (2013) and Kolko (2012), for example, find a positive relationship between the broadband expansion and employment growth in the US, based on the data collected between 1999 and 2006. Finally, ecologic benefits are associated. FTTH belongs to technologies of green telecommunication networks, because FTTH uses significantly less energy when it comes to high levels of data transferred compared to copper–fiber telecommunication networks (Lange et al. 2015) and by hybrid fiber coax networks provided by cable network operators (Gladisch et al. 2008).

For the utility companies, the FTTH connections laid to their energy customers have an opportunity to resolve the problem of being dependent on telecommunication providers who deliver the data from smart grid endpoints to the internal grid infrastructure and lower the grid operation costs. In spite of the advantages of owing FTTH for energy suppliers, the telecommunication business is competitive. Therefore, utility companies need to plan the roll-out of such services carefully and conduct their marketing campaigns in a cost-effective manner.

7.2 Descriptive insights on the purchase intention of residential customers

The analysis in this chapter uses dataset D, as described in subsection 5.1.3 in detail. The survey that was conducted in 2015, the following question was asked about the purchase intention towards FTTH: "How do you estimate the prospects that you will buy FTTH within the next 12 months?" (In German: "Wie hoch schätzen Sie die Chance ein, dass Sie innerhalb der nächsten 12 Monate FTTH beziehen werden?"). 436 households answered this question. The survey participants could choose one of 12 following values: 11 textual levels of willingness to buy (Juster 1966) and the option "I don't know what FTTH is".

The complete scale on "Consumer Buying Intentions and Purchase Probability" is listed in Table C.4 on p. 222 together with the used German translations. Several survey participants did not answer the question. The distribution of answers and the number of missing responses is illustrated in Figure 7.1.

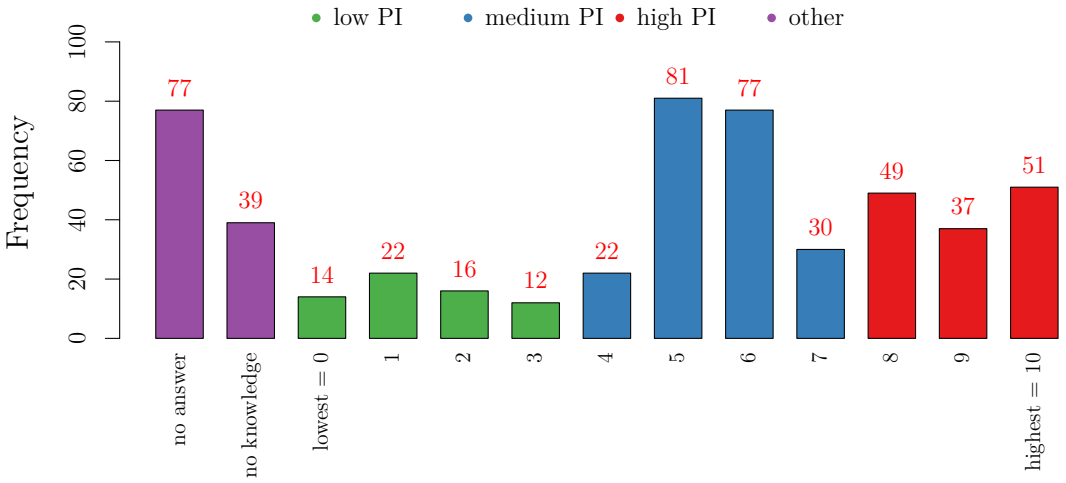


Figure 7.1: Distribution of answers to purchase intention towards FTTH

I aggregated the answers of the 11-piece Purchase Intention (PI) scale into three classes: "high PI" (8-10), "medium PI" (4-7), and "low PI" (0-3). The class borders have been chosen with the motivation to identify particularly those households with an extremely high or low purchase intentions, but also to allow for sufficient number of samples in each class. Table 7.1 shows the distribution in the aggregated customer groups.

Besides the purchase intention question, households were asked further variables regarding household characteristics and attitudes towards sustainability.

Table 7.1: The groups for purchase intention towards FTTH

Group	**Intention level**	**Absolute**	**Relative**
Low PI	0-3	64	14.22%
Medium PI	4-7	210	46.67%
High PI	8-10	137	30.44%
No knowledge on FTTH	-	39	8.67%
Overall		450	100.00%

The detailed survey and accompanying study is described in Sodenkamp, Hopf, Kozlovskiy, et al. (2016). In the following, some case-related descriptive insights are presented.

The majority of customers have a positive attitude towards FTTH (30.4% said that they have a high PI and 46.7% said that they have a medium PI towards FTTH). Besides the question about PI towards FTTH, the customer survey contained questions on the household characteristics of the survey participants, their attitudes, and completed measures towards energy-efficiency and about renewable energy sources. I present the descriptive insights about the customers with high PI and no knowledge on FTTH below. All presented findings have been verified with Yates (1934) χ^2-test or Welch (1947) t-test. Other combination of variables in the survey towards the purchase intention or the knowledge on FTTH were not statistically significant.
Findings on customers with a high PI towards FTTH:

- Families have high PI towards FTTH ($\chi^2(1) = 4.28, p = .039, \omega = 0.162$)
- Households that are likely to adopt a solar energy system (survey items are based on H.-W. Kim et al. (2007) and Davis (1985)) said also that they have a high PI towards FTTH ($t(150.59) = 2.40, p = .009, d = 0.33$)
- Households with interest in new technologies also show a high PI towards FTTH ($\chi^2(1) = 4.38, p = .036, \omega = 0.176$)
- The customer-based reputation of the utility company (using selected items of Walsh and Beatty (2007) and Walsh, Beatty, and Shiu (2009)) has a positive correlation of $\rho = .126$ with the PI towards FTTH (Spearman's rank correlation, $S = 4{,}684{,}500, p = 0.025$)

Findings on customers without knowledge on FTTH:

- Households with a low amount of multiple-glazed windows have less knowledge on FTTH than other households ($\chi^2(1) = 5.89, p = .015, \omega = 0.180$)
- People in small homes (less than 100m^2) have less knowledge about FTTH ($\chi^2(1) = 4.12, p = .042, \omega = 0.160$)
- People without knowledge about FTTH live more frequently in rented homes than in own homes ($\chi^2(1) = 3.43, p = .064, \omega = 0.152$) and more often have old houses ($\chi^2(1) = 4.52, p = .034, \omega = 0.165$)

7.3 Predictive analytics to identify customers with high interest in FTTH

Following the household classification approach, I evaluate how well households with a high PI towards FTTH can be recognized from SMD, weather data, and geographic information. The analysis described in this chapter is illustrated in Figure 7.2. It comprises three main phases: feature extraction, feature selection, and the training of a machine learning model that is able to predict the PI of individual households. The predictive analytics procedure is described below.

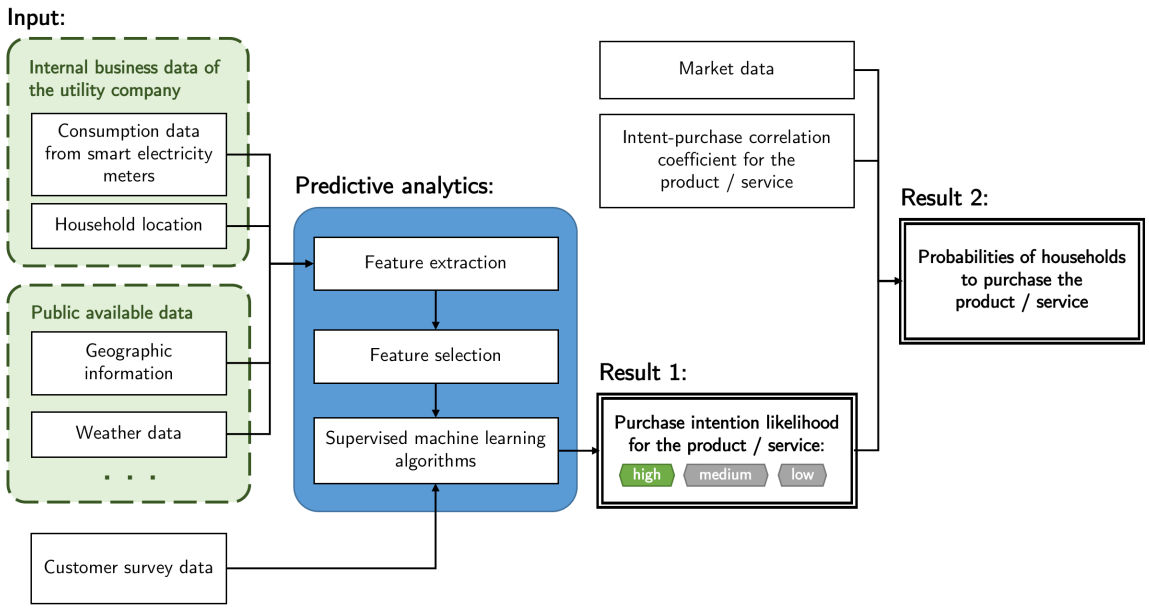


Figure 7.2: Method to predict purchasing probability for cross-selling products or services using the data available to energy utility companies

The result of the household classification approach is the likelihood that a respective household has a high purchase intention level towards the cross-selling product (*Result one*). This score can be used for decision support in marketing. Besides, the prediction can be used together with market data and intent-purchase correlation coefficients from marketing research to obtain a purchase likelihood for each household (*Result two*). This helps to estimate the market for a cross-selling product and further supports decision making. The both mentioned applications of this predictive analytics result are further described in the discussion section of this chapter.

7.3.1 Data and features

In analogy to the taxonomy of data sources (see section 3.2), I consider available company internal business data (i.e., SMD, the customer address), and external data (i.e., geographic data, weather data). Features from SMD, the electricity consumption in comparison with the neighborhood, geographic information, and weather data are used as described in section 3.3.

In total, 207 features (93 from smart meter data, 40 from weather information, 60 from OSM, 14 from neighborhood) are usable for the predictive analytics in this case. To reduce the number of features, consistency based feature selection (Dash and Huan Liu 2003; Romanski and Kotthoff 2014) was used, because it showed the best trade-off between the number of selected features, classification quality, runtime, and reproducibility of the feature selection and classification results. The feature selection is performed for each week of the data separately, because the time-series features are defined for one week of data. With that procedure, also seasonal effects are avoided, because each feature is considered that was selected at least in one week. Table 7.2 shows the selected features together with the frequency of selection in the 52 weeks. 20 of the 50 features have been selected only in one week.

Table 7.2: Selected features by the consistency feature selection method (Dash and Huan Liu 2003) in any of the weeks with Random Forest feature importance scores, obtained with the predictive model used in the case study

Category	Feature	Frequency of selection	Mean	SD
geographic	buildingTypeMode	39	0.55	1.11
consumption	c30_we_afternoon	1	1.73	1.57
neighborhood	nextbuildingType	50	0.33	1.23
neighborhood	nextlanduseType	39	2.19	1.58
neighborhood	nn10_consNoon_wd_wd_absDiff	4	0.87	1.78
neighborhood	nn10_numAboveMean_relDiff	2	1.57	1.94
neighborhood	nn10_corDays_absDiff	1	1.74	1.80
neighborhood	nn10_meanCor	1	0.91	1.39
neighborhood	nn10_meanCor_we	1	0.72	1.55
relations	r30_night_wd_we	11	1.25	1.84
relations	r15_mean_max_no_min	9	1.87	1.41
relations	r30_we_night_day	6	0.80	1.45
relations	r30_noon_wd_we	3	1.25	1.30
relations	r30_we_evening_noon	3	0.94	1.37
relations	r30_evening_wd_we	2	0.81	1.65
relations	r30_wd_night_day	2	1.11	1.60
relations	r15_var_wd_we	2	1.64	1.63

Category	Feature	Frequency of selection	Mean	SD
relations	r30_morning_wd_we	1	0.92	1.35
relations	r30_evening_noon	1	0.63	1.55
relations	r15_max_wd_we	1	1.76	1.70
relations	r15_day_night_no_min	1	0.01	1.75
statistical	s15_num_peaks	5	0.83	1.68
statistical	s15_cor	1	2.73	1.78
statistical	s15_cor_wd	1	2.81	1.68
temporal	thisbuildingType	39	0.40	1.27
temporal	t15_above_2kw	38	0.85	1.24
temporal	thislanduseType	36	0.06	0.97
temporal	t15_above_1kw	29	1.31	1.49
temporal	t15_above_0.5kw	18	2.02	1.45
temporal	ts15_acf_mean3h	18	0.44	1.34
temporal	t15_above_base	5	0.53	1.47
temporal	t15_above_mean	4	2.59	2.21
temporal	t15_percent_above_base	3	1.78	1.70
temporal	ts15_stl_varRem	2	0.71	1.67
temporal	t15_daily_max	1	0.98	1.61
temporal	ts15_acf_mean3h_weekday	1	1.34	1.50
temporal	t15_value_min_guess	1	1.63	2.07
weather	w_temp_cor_overall	6	2.39	1.69
weather	w_temp_cor_daytime	5	2.10	1.83
weather	w_windSp_cor_overall	2	1.59	1.65
weather	w_windSp_cor_daytime	2	1.50	2.18
weather	w_windSp_cor_minima	2	1.26	1.99
weather	w_prec_cor_evening	2	0.73	1.76
weather	w_temp_cor_night	1	1.09	1.52
weather	w_windSp_cor_maxmin	1	1.35	2.06
weather	w_prec_cor_overall	1	1.62	1.53
weather	w_prec_cor_night	1	0.73	1.66
weather	w_prec_cor_daytime	1	1.50	1.88
weather	w_skyc_cor_overall	1	1.95	2.07
weather	w_skyc_cor_daytime	1	2.12	2.17

7.3.2 Supervised machine learning and performance evaluation

In this analysis, Random Forest (Breiman 2001) is primarily used. For benchmark reasons, kNN, SVM, Naïve Bayes, AdaBoost, and neuronal networks are considered. Some of the algorithms have the functionality to consider weights for the to-be-predicted classes, enabling to adjust the sensitivity of the classification. As the interest of this study lies in households with high PI towards FTTH, I tested class weights that favoring the class "PI high": 0.5, 0.75, 0.875,

0.9, 0.9475, 0.96875. The best classification results could be achieved using a class weight of 0.75. For comparison reasons, the default class weight of 0.5 is also used.

To measure the performance of the classification algorithms, I use AUC of the class "high PI" and use 5-fold cross-validation and report arithmetic means of the AUC cross-validation results. Given that consumption data on 52 calendar weeks is available, cross-validation is performed for each week which leads to 52 AUC values for the performance.

In a first step, different feature sets are used to perform classification (consumption, weather, and geographic features). The average AUC of different combinations of the feature sets is shown in Table 7.3. For each combination of features, a t-test is computed to estimate whether the AUC value is different from a random classification with $AUC = 0.5$.

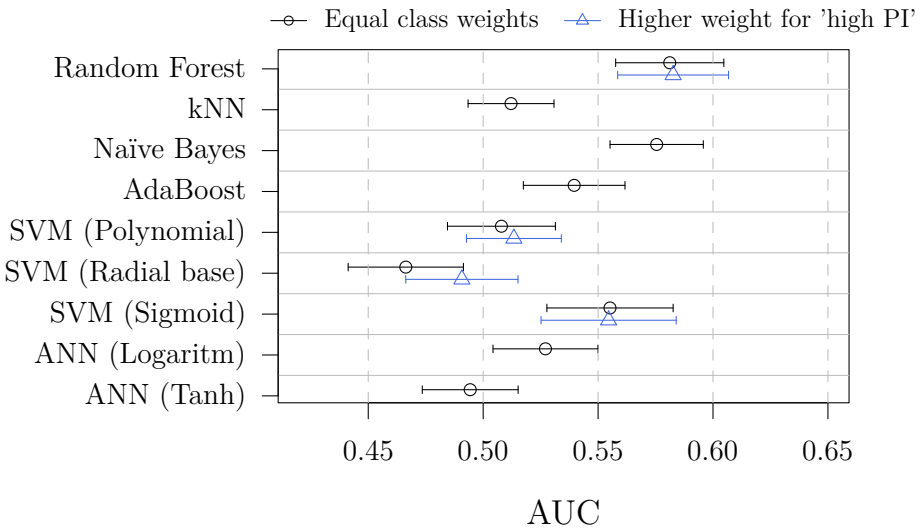


Figure 7.3: AUC results for different classifiers for the class "high PI" with 95% confidence intervals

The prediction of the PI based on SMD alone is not significantly better than random. Adding features from weather data (instance 2) improve the predictive power of the machine learning model, so that it is slightly better than a random selection, but the results remains not satisfactory. Using consumption and geographic features (instance 3) improves the results, and using all features (instance 4) provides performance results that show significant better prediction results than a random selection of customers.

Table 7.3: Classification results with data from different data sources

Instance	Consumption	Weather	Geographic	Average AUC*	
1	X			0.5116161	
2	X	X		0.5161658	*
3	X		X	0.5451374	***
4	X	X	X	0.5554294	***
5			X	0.5560899	

Interestingly, the classification only based on geographic features is better than random (only one AUC value can be calculated here, as the time dimension is missing) and seems quite as good as the complete model. It is necessary to evaluate this model carefully, because only one feature ("nextbuildingType") was included.

Based on the present dataset, I can conclude that the prediction of high PI towards FTTH solely based on geographic information is possible in the present dataset, but I cannot deduce a general finding for other datasets because of the regional limitation of the data. Further data would be needed to confirm the hypothesis, that this finding holds in general.

7.4 Contribution to the planning and execution of cross-selling marketing campaign

Considering the results of the above analysis, planning, and execution of cross-selling campaigns can be improved. Predictive analytics has three main contributions to operational, tactical, and strategical decision making that I describe below.

7.4.1 Customer scoring (operational decision support)

On the operational level, the result of predictive analytics benefit marketing managers by ordering the customers according to their likelihood that they have a high purchase intention towards a cross-selling product.

When a marketing campaign for FTTH in the present case is considered, an energy retailer with 100,000 customers has 14,220 customers that are interested in the product (14.22% indicated in the survey that they have a high PI). When

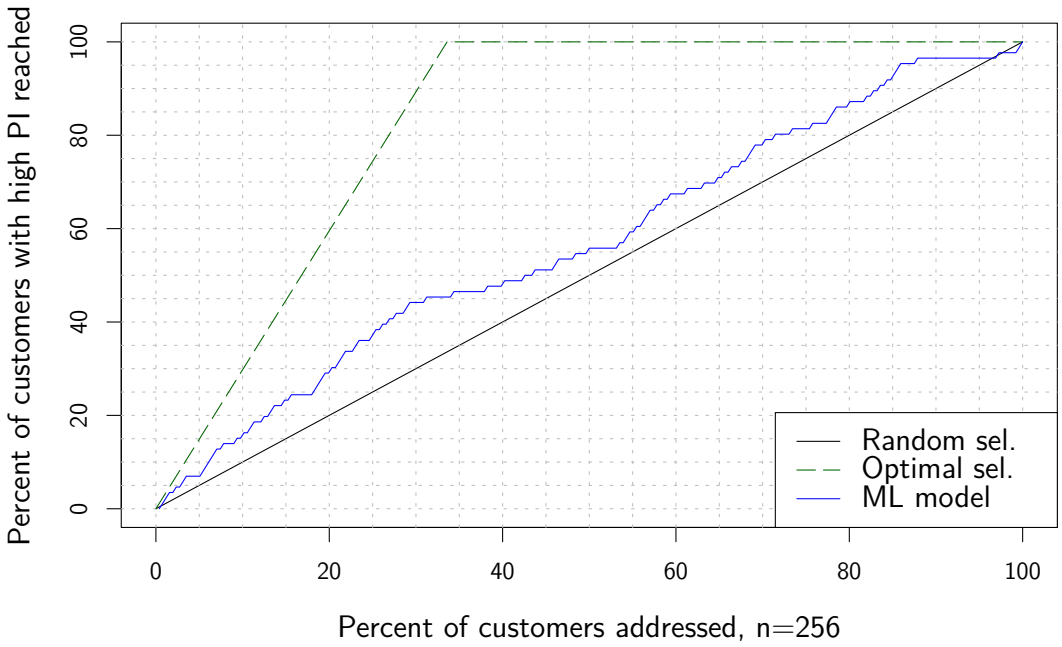


Figure 7.4: Algorithm selection in comparison with a random selection

40% of them should be targeted in a marketing campaign without having knowledge on the interest in FTTH, one would need to send letters to 40% of all the customers. Assuming that each letter to a customer costs €1 as variable cost, this campaign would cost €40,000. When the predictive analytics approach is used (including electricity consumption, weather data, and geographic information with the RF algorithm) to select top scored customers, only 27% of the customers must be addressed in order to reach the 40% target segment (see Figure 7.4). In the example of €1-per-customer mailing, this would cost only €27,000 and would save 32.5% of the costs. This example illustrates, how the marketing budget can be more efficiently allocated to reach high value customers. Thus, the proposed approach helps to support the operational decision making in marketing.

7.4.2 Cost-benefit analysis (tactical decision support)

The planning of a marketing campaign—which I consider to be a decision on the tactical management level—additional information is needed to allocate the

right budget and set the right dimension of the campaign. The predicted scores for all customers can be used to calculate a cost-optimal number of customers that should be addressed in a campaign, to maximize the expected total benefit.

For that, I consider fixed costs of a marketing campaign C_{fix} (e.g., setting up the mailing template, or the concept for call-center agents), variable costs per customer C_{var} (e.g., printing and postage), the benefit B for each acquired customer, as well as the probability that one customer is interested in the product p_i, as obtained with the ML analysis. The resulting equation for the total benefit B^* from the mailing campaign with n customers is

$$B^* = -C_{fix} + \sum_{i=1}^{n} *B * p_i - C_{var} \tag{7.1}$$

Figure 7.5 shows the cost-benefit visualization of a mailing campaign with $n = 100{,}000$ potential customers in the considered FTTH cross-selling case. As benefit for each customer $B = €55$ is assumed, variable costs per customer $C_{var} = €1.50$, fixed costs $C_{fix} = €5{,}000$. Taking the predicted purchase intention scores p_i for each customer i, the optimal number of customers that should be addressed can be calculated. Results for the present case are illustrated in Figure 7.5 and indicate an optimal number of 37,891 customers.

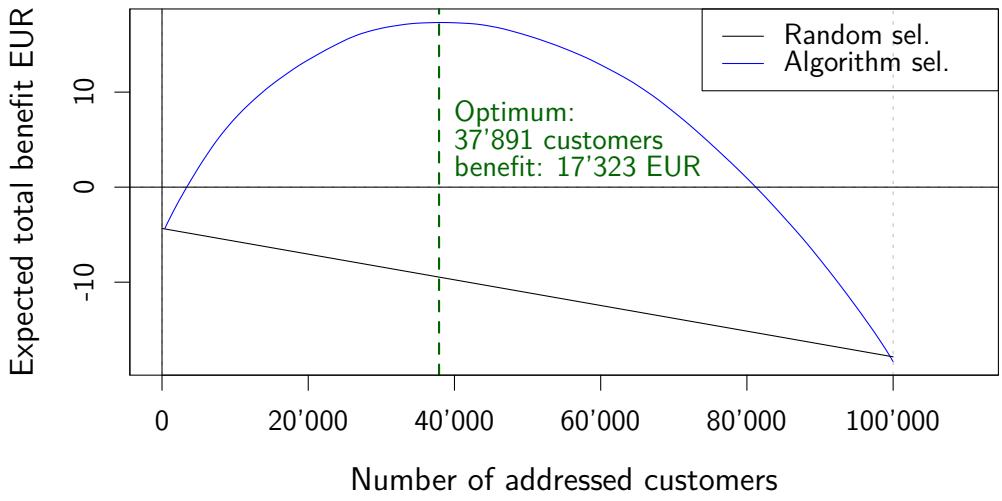


Figure 7.5: Cost-benefit visualization for a marketing campaign

This information helps to better plan the marketing campaign. By changing the border conditions, several scenarios can be tested. For example, a call-center

can be engaged to do the sales. In this case, the variable costs would be higher, but the likelihood to win the customer would be improved.

7.4.3 Converting predicted purchase intentions into purchase probabilities to estimate market size (strategic decision support)

So far, the predicted customer intention to buy a product was used to prepare information that is relevant for operational and tactical decision making in marketing. Research in marketing shows that it is possible to convert the purchase intention—stated in a survey—into purchase probabilities using empirically obtained intent-purchase correlation coefficients. With this information, a third application of the predictive analytics analysis renders possible: The estimation of the market size. This insight helps, for example, to make the strategic decision to enter a certain market with an offer or not.

Below, I outline the calculation steps necessary to pursue a conversion of the predicted purchase intention scores into a purchase likelihood for each customer. Nevertheless, this approach needs further validation in the field, for instance, with a field experiment.

Forecasting purchase probabilities from purchase intentions Where early marketing studies state that intentions might be good indicators of consumers' purchase behavior (Morrison 1979; Bemmaor 1995; Armstrong et al. 2000), Sun and Morwitz (2010) emphasize that this assumption does not hold, and show that several factors lower the probability of an actual purchase compared to stated intentions by customers. Morwitz et al. (2007) show that purchase intention ratings of customers can be converted to buying probabilities. The authors determine the coefficients based on their meta-analysis of 40 commercial and academic studies conducted between the 1957 and 2006 with more than 65,000 consumers on more than 200 different products.

Unified model to convert stated purchase intentions into purchase probabilities Sun and Morwitz (2010) proposed a unified model to convert stated purchase intentions to the likelihood that a customer buys a product. In this model, it is assumed that $n = 1, 2, \ldots, N$ respondent expressed their purchase intention among $j = 0, \ldots, M$ ordered intention levels. Considering the probability $P(y_{jn} = 1)$ as those that the jth intention level is true and the probability $P(z_{j^P n} = 1)$ as that the purchase at j^Pth level is true, $F^{*P}_{jj^P n}$ is the joint probability for each customer n that the reported intention j_n is equal or lower to

the threshold for purchasing at j_n^P. When π_{jn} is the probability that the stated intention is suspected to a bias, Sun and Morwitz (2010) define their unified model as

$$P(y_{jn} = 1, z_{j^P n} = 1) = \pi_{jn} \left[\sum_{i=0}^{j} F_{jj^P n}^{*P} \right] + \left[1 - \sum_{i=j}^{M} \pi_{jn} \right] F_{jj^P n}^{*P} \tag{7.2}$$

That stated intentions do not reflect the true intention levels because of several biases involved. It is assumed that the stated intentions are in general over-reported. These biases are described in several studies (Sun and Morwitz 2010; Morwitz et al. 2007) and can be corrected in the calculation. Finally, field tests help to obtain intent-purchase correlation coefficients for different product categories.

Market data and intend-purchase correlation coefficients In addition to a general purchase intention of a customer, it must be considered that multiple competitors offer similar products. Therefore, the relative market size should be incorporated into the purchasing probability calculation. The market data is expressed as an empirically found coefficient of the *relative market size* in the outlet the product or service is placed.

In the present case, data on the market share of internet service providers can be considered. In Switzerland, for example, the Swiss Federal Communications Commission (2015) publish market shares for broadband internet service providers. The market in 2014 is distributed to three large telecommunications companies holding a share of 84.6% and other companies holding a share of 15,4%. For a utility company that aims to enter the market with a FTTH product, a market share coefficient of $p_m = 0.154$ would be a conservative estimate.

Calculating individual purchase probabilities The individual purchase probabilities for each household customer can finally be obtained using a three-factor model of market data, intent-purchase correlation coefficients, and the results from the ML prediction:

$$p_i = p_{PI=high} * p_{PI=high \wedge buys} * p_m \tag{7.3}$$

The first factor $p_{PI=high}$is the probability that the customer belongs to the class of customers with a high PI. This number is the score obtained using the ML prediction described earlier in this chapter and that gives probabilities for each customer belonging to the class "high PI". The second factor $p_{PI=high \wedge buys}$ is

the intent-purchase correlation coefficient, a conditional probability that households with a high reported PI also purchase the good. The third factor p_m is the probability that a customer who actually buys the product or service would buy it from the utility company and not from one of its competitors.

Strategic marketing planning can use the resulting purchase probabilities to estimate the market size using data from a customer survey, the intent-purchase coefficients and market share.

7.5 Conclusion and limitations

Cross-selling is a relevant task in relationship marketing which has the goal to "nurture" existing customers and increase the customer lifetime value (Kumar 2018). In this chapter, I demonstrated the use of the household classification approach to support a FTTH cross-selling campaign.

Answer to RQ 5 The presented application of ML algorithms shows how scores for individual customers, that express the likelihood that a customer belongs to the group of customers with high purchasing intention towards a cross-selling product, can be obtained from ambient data available to utility companies. These insights help marketing managers in operational, tactical, and strategic decision making.

In response to RQ 5 (*Which added value can be realized from predicted customer intentions on the example of relationship marketing?*), the case study of predictive analytics in energy retailing revealed two contributions. First, the obtained cross-selling scores can be used to select high-value customers for targeted campaigns. Second, the prediction helps to obtain the cost-optimal size of a marketing campaign and finally helps to estimate the reachable market size for a product that was not offered by the firm before. By selecting the customers with high likelihood to respond to an offering, the customer value can be increased, and the marketing budget allocated more efficiently.

Additionally, I proposed an approach to convert the predicted scores into purchase likelihood following the unified model of Sun and Morwitz (2010). This conversion of purchase intention into purchase likelihoods is necessary, as intentions stated in a survey, and likewise intentions predicted using ML models that rely on the ground truth data obtained in such surveys, do not reflect the true buying behavior. Nevertheless, the proposed theoretical construct needs empirical validation in the future. Thus, the presented approach provides therefore decision support on three managerial levels: operational, tactical and strategic marketing management.

Limitations and future research I consider the following limitations to the presented study. The customer survey used to develop and test of the predictive analytics method includes 436 utility customers located in one town in Switzerland. The approach should be validated with a larger sample with customer located in other countries.

The measurement of purchase intention using Juster's (1966) single-item scale was criticized in the literature. Future research raise the purchasing intention as acknowledged in IS research (H.-W. Kim et al. 2007; Davis 1985).

The presented approach is a general framework for the prediction of individual purchase intention and behavior of private households for cross-selling products or services, offered by energy utility companies. This is a starting point for further research that may validate the approach with additional data and possibly in other industries.

8 Summary and implications

Ambient data emerge as by-products of business activities and in publicly available sources

Digitization creates large amounts of data in every organization, but also several publicly available data sources emerge. Considerable parts of the data are by-products of business activities (like a customer's payment history, app usage data, or sensor data in smart homes), that are often not necessarily a requirement to fulfill contracts. Additionally, a remarkable number of data sources are freely available online, like government statistics, user-generated web content, and weather observations. I describe data sources that are available to firms for analytics as *ambient data* and recognize various characteristics: On the one hand, the data are difficult to process, as they are available in unstructured or weakly structured formats (e.g., text, log messages, time series data, geographic information), contain "noise", irrelevant or missing data points, and require advanced data processing techniques to prepare the data for further analyses. Moreover, data often have only a weak reference to business entities and can therefore generally not be easily connected to traditional business data. On the other hand, ambient data contain plenty of latent information (e.g., household characteristics, socio-demographic information, and data on buying behavior) that can be made explicit with state-of-the-art computing.

Plenty of latent variables are hidden in ambient data sources

Ambient data therefore have enormous value for firms, but the actual process to create value from that data is complex and not yet fully explored. An attempt to guide the investigation of this process, the data-driven decision making process model was proposed (Koutsoukis and Mitra 2003; Sharma et al. 2014; Thiess and Müller 2018). It helps to investigate the opaque transformation of raw data to, for example, value-added products, services, and process improvements. The process model splits the value-creation into four stages: question to data, data to insight, insight to decision, and decision to value.

Data-driven decision making process as a research agenda

Exploring ML to extract value from ambient data

In order to explore the stages of the process more closely, my research aimed to answer the leading question *how ML can be used to harness ambient data in business applications and thus gain new insights from data*. Thus, I examined five research questions in the context of my dissertation, which are described in detail in the following section along with the results. Furthermore, I have demonstrated how the resulting insights can be leveraged to generate value, by means of two case studies from energy retailing.

Case 1: Scalable energy efficiency campaigns

The first case study focused on the identification of household characteristics related to residential energy efficiency (e.g., type of heating, age of house, number of occupants, children in the household). I showed, how energy consumption data together can be processed together with freely available data using ML methods to obtain knowledge on household characteristics. The insights can be used to personalize energy efficiency campaigns and thus make them more effective (e.g., through household specific savings recommendations, load estimation and shifting).

Case 2: Relationship marketing in energy retailing

The second case study showcased the recognition of customer attitudes and behavior (e.g., attitudes towards energy efficiency, willingness to buy photovoltaic systems or new products from energy providers). The successful prediction of such information enables the development of new products and services in the energy sector as well as their targeted promotion to relevant customer segments.

Structure of this chapter

This chapter resumes the results of my work in the form of answers to the RQs. This is done along with the four stages of the data-driven decision making process in the following section. Thereafter, I outline future work directions in the field of IS, energy informatics, relationship marketing, and machine learning. In the third section, I name limitations regarding the research pursued. Finally, I compile the practical implications resulting from my work for an industry audience.

8.1 Summary of the results

This work's contributions are structured along the data-driven decison making process

The research conducted in this dissertation is structured along the data-driven decision making process. With regard to the four stages of this process, I formulated five RQs for my dissertation research and answered them by means of case studies from the energy retail sector—a key industry that faces many challenges. This process model has four stages, for which I formulated research questions that I recapitulate below. I name the core results respectively and briefly discuss my contribution to the field of IS.

Stage 1: Question to data Like in research, data-driven investigations should focus on answering important questions. As firms have huge amounts of ambient data available that they can analyze, it is relevant to clearly formulate the goal of investigation before any analysis starts. This also means clearly defining the scope and the objective of the analysis, including the dependent variables to be predicted, and the criteria under which an analysis is considered successful.

Overview to available data sources needed

An important consideration in this scoping phase is to identify the relevant data sources that should be included into the analysis. To formulate a good question as a starting point for analytics, it is relevant to know which data is available and can be potentially analyzed. Predictive analytics studies in IS so far focus mainly on firm-internal data available in easy-to-use formats. This could be because a systematic overview of data sources for business analytics is currently missing. I wanted to overcome this gap and answered the first research question (RQ 1: *Which data sources are considered in predictive analytics IS research studies, which typically exist in firms or are publicly available, and what are characteristics of the data sources?*).

Taxonomy of internal and external data

As the result of a systematic literature review in IS research journals, I developed a taxonomy of data sources that are available for analytics. The taxonomy differentiates between internal data, that is created by firms and stored within their databases, and external data, that can be accessed by firms to generate additional insights. Most of the data sources identified are ambient data, as they are not necessary to fulfill a contract, or they can be used in a differ-

ent context than they were collected. The internal data contains information on customer details (e.g., name, address, contact details), transactions, interactions and basic demographic variables. External data includes, for example, public statistics, geographic information, or weather data that are published as open data by governments or created by users.

Insight generation through ML based predictive analytics

Stage 2: Data to insight The insight generation process is realized by combining several types of analyses (descriptive, diagnostic, predictive and prescriptive) that are selected depending on the question focused on. This dissertation proposes predictive analyses using ML as a tool to generate insights from ambient data. Two research questions are examined regarding this stage and the respective contributions outlined below.

The first contribution to this stage is a debate on the trade-off between human cognitive skills, theory as well as expert knowledge and algorithmic power in building predictive models (RQ 2: *Does theory, expert knowledge and human cognition notably help to reduce data dimensionality, although several computational methods exist for this task?*). I investigated both aspects in detail. On the one hand, feature extraction was introduced as an approach to prepare data using the nexus of human cognition, theory and expert knowledge. The approach is demonstrated among eight examples using data sources that are typically present at energy retailing firms: Transaction data, environmental data, geographic information, and governmental statistics. On the other hand, feature selection methods are tested as automatic procedures. The results of my analysis conclude that automatic feature selection methods (FSMs) provide beneficial aid for the insight-generation process. Therefore, they should be considered as an *additional step* in this process. Nevertheless, the simple application of such methods to raw data cannot be recommended because of three reasons. First, in the comprehensive analysis of 43 FSMs using real-world data, I could not find any algorithm that is superior to others. I can rather recommend testing several methods, including a shortlist of five methods. Second, theory-based preparation of the data by empirical feature extraction can be particularly recommended because there is no need to

Human input is essential to set up ML methods

Automatic approaches can support insight-generation

waste computational power to learn relationships from data that have already been proven by studies or are known from human experience. Third, the models learned from many input dimensions, are difficult to interpret and it is not possible to judge on their generalizability. For example, a model in which an algorithm considers the combination of five arbitrary measurements from a long time series as predictive features because of its ability to accurately predict the dependent variable is not particularly reliable because the predictive power of these selected points in time may be smaller for another data set. In general, it is hardly possible to explain why exactly these points in time are relevant for the prediction. Thus, the stability of the model cannot be evaluated. In other words, algorithms can at most learn correlations, but rarely causalities.

Combining human input and automatic methods is key for effective modeling

I conclude, in response to RQ 2, that the derivation of features from the raw data solves several problems: The dimension of the data becomes smaller, resulting in more efficient data processing (resource efficiency). Simultaneously, predictive accuracy can be improved because known knowledge (theory and expert knowledge) is made available to the algorithm without the need to acquire it. Furthermore, models become more explainable, as the features are defined by humans. Admittedly, automatic methods can also help to significantly improve model quality. In the experiments I carried out, a logistic regression and feature selection methods were used to achieve classification accuracy similar to that of SVM or random forest without feature selection.

New insights: customer characteristics and intentions

The second contribution to this stage is the evaluation of RQ 3 (*How well can (a) customer characteristics and (b) intentions be revealed from ambient data, in the context of energy retail, using state-of-the-art ML methods?*). With this research question, I have gone through the insight generation process in the field of energy retail. Electricity consumption and geographic data contain many latent variables that can be revealed through ML. From the 22 household characteristics and intentions of residents that were investigated, 18 can be predicted from electricity consumption data (of different granularity) along with available data on residential consumers (i.e., governmental statistics, open geographic information from OSM) by using state-of-the art machine learning algorithms which are significantly better than random. With this in-

Models can be applied to data with various degree of detail and work outside the training region

vestigation, I went beyond existing research by testing additional household characteristics (especially heating related variables and house ownership) and customer intentions (e.g., purchase intention and interest in sustainability). Additionally, I added new predictor variables from ambient data sources that bolstered the classification performance, tested different data resolution and the model transferability between countries.

Insights lead to more informed decisons

Stage 3: Insight to decision Single pieces of information created by predictive ML models cannot be simply put together. They must rather be put into context, interpreted, and correctly used to make better—especially more informed—decisions that finally lead to value. Among two cases from the energy retail industry, this step from insights to decision was investigated. Respective to the cases, I formulated RQ 4 (*Which added value can be realized from predicted customer characteristics on the example of personalized energy feedback?*) and RQ 5 (*Which added value can be realized from predicted customer intentions on the example of relationship marketing?*) which addresses this stage from single insights to decisions which finally can create value. The cases evaluate the predictions for household characteristics and customer intentions obtained in practical settings.

Fostering personal and corporate decisions

In the first case, which is described in chapter 6, the predicted household characteristics are tested in a monthly home energy report that fosters energy conservation in residential households. The availability of detailed data on individual customers make such personalized energy feedback campaigns possible. Without the predicted data, it would be unclear how such a personalized energy feedback campaign could be realized. The tested home energy report triggers energy conservation and improves customer satisfaction. This case demonstrates how decisions in commercial and personal environment are enabled through data-driven innovations.

Operational and strategic decision-support

In the second case, described in chapter 7, predicted purchase intentions towards a cross-selling product of an energy retailer are worked up to foster management decisions. Concretely, cross-selling scores for each customer are obtained from ambient data with ML. This supports managerial decision making in two ways. First,

operational decisions in marketing are supported by identifying customers with a high score to purchase a cross-selling product. Second, strategic marketing is supported with obtaining the cost-benefit optimal number of customers to be addressed in a campaign.

Stage 4: Decision to value The final stage of the data-driven decision making process is hard to measure. The ultimate consequence of better decisions, meaning the added value of new insights for more informed decisions, is difficult to quantify. We generally cannot judge how the world would have developed if we had made a different decision. Consequently, I only give a careful appraisal of the value created through the additional insights that can be generated through my research in the two investigated case studies on energy efficiency and energy retail marketing. Nevertheless, my cautious credit to the results are the following.

Personalized energy reports increase service quality, customer satisfaction and energy efficiency

In a vivid experiment, described in chapter 6, I showed how the predicted household data could be converted into targeted energy feedback interventions. The developed home energy reports, which was send once a quarter, led not only to energy conservation—which was the original goal—but also to an increased usage of the online portal, more data inserted by customers and a slightly increased customer satisfaction. Thus, the additional insights have not only created environmental value through energy conservation, but also societal value in the form of higher customer satisfaction and finally, economic value as the improved customer communication that improves brand image of the energy supplier and offers possibilities of efficient advertisements to loyal customers. This outlet can be used further to increase sales in the future.

Increased customer value through targeted advertising

The detailed investigation of a cross-selling campaign in chapter 7 uncovered the possible support through ML based insight-generation from ambient data. It showed that marketing budgets could be allocated more efficiently. Furthermore, customers that have a low interest in certain products receive less advertisements when the only customers addressed are those identified as high-value targets. Aside from the promotional benefit, this approach therefore also benefits the customer.

8.2 Implications for research and future work

The results of my work have implications for future research in the fields of energy informatics, relationship marketing, big data analytics, and machine learning. In addition, some counter-intuitive results appeared during my investigation that need further inquiry. My research is embedded in broader contexts, wherefore some concepts proposed in this dissertation need further validation. I summarize the resulting consequences and open research topics in this section and order them according to the related research fields summarized in chapter 2.

8.2.1 Value creation through predictive analytics

Overview to data sources needs further validation

Data have a high strategic relevance for organizations (Constantiou and Kallinikos 2015; Yoo 2015; Kallinikos and Constantiou 2015). A broad research agenda was presented to understand how the information value chain can realize value from data (Abbasi et al. 2016; Sharma et al. 2014) and several scholars call for more empirical research (Günther et al. 2017; Markus 2017). The taxonomy of firm-internal and external data sources, developed as result regarding the first RQ, is a building block in future research on how value can be created out of different data sources. The sources through which the taxonomy was developed were research studies in IS and seven case studies from the energy retail industry. As data sources in firms might differ from those reported in research studies, I call for future research validating the taxonomy, for example, by interviewing domain experts, data scientists or managers.

Proposed features should be tested in other fields of application

I propose empirical feature extraction as an insightful data preparation and integration step. Pursuing this engineering task, a variety of ambient data such as high-frequency transaction data (e.g., electricity smart meter data), weak-structured data (e.g., geographic information from OSM, weather data, governmental statistical data) are usable for such analyses. I tested the new data sources and the application of predictive analytics in seven case studies in the central European energy retail industry. Future work should, on

the one hand, expand the set of empirical features to utilize new data sources, and on the other hand, validate the contribution of the proposed features in other application domains or industries. A very promising data source for future research are VGI web portals. OpenStreetMap, the source of geographic features for my research, is only one example of that class of open data.

Support of data scientists and managers

The proposed predictive analytics approach to analyze ambient data feature extraction, feature selection and ML to gain new insights on residential customers is a tool-set for business analytics. Developed for the energy retail industry, it serves as a blueprint for other industries with strong customer focus and shows how value can be generated through advanced business analytics. Thereby, my work supports the work of data analysts or managers that aim to set up analytics processes in organizations. The support of data analysts or managers is much more concrete than existing general process models for data analysis describe (e.g., the knowledge discovery in databases process, the CRISP-DM process model, or the predictive analytics process of Shmueli and Koppius (2011)). This is, for example, done by demonstrating how sense-making from big data can be supported through empirical feature extraction (see section 3.3). With this technique, one can overcome several challenges in sense-making from big data (L'Heureux et al. 2017), for example: Handling noise in data, measurement errors, biased models, small datasets, imbalanced classes.

New research field: Decison-making based on predicted data

Finally, the question of how predicted information (that is associated with uncertainty) can be properly used by managers and in business applications should be investigated in the future. Decades of IS and computer science research tried to improve the data available in firms' databases. Predictive analytics makes latent variables tangible, thereby creating data of low quality because it is unknown if the prediction is correct in reality. Efficiently handling such data must be investigated in theory and practice.

8.2.2 Energy informatics to support energy efficiency

Energy data analytics is a relevant field at the intersection of IS and related fields. Despite the relevance of this industrial sector,

it has not gained much attraction from a business analytics and IS point of view. I followed the call for research by Melville (2010), R. T. Watson, Howells, et al. (2012), and Ketter et al. (2018) in the field of energy informatics and conducted empirical research in that area. In particular, I investigated how characteristics and intentions of residential energy customers can be extracted from ambient data. This helps to realize targeted energy efficiency and marketing measures. Hereby, I adopted the household classification approach of Beckel and colleagues (2013; 2014; 2015) and identified several limitations that I addressed in my research. Together with research partners, I conducted seven case studies (two are described in this dissertation in deatail) involving energy data from more than 20,000 residential customers data in Switzerland and Germany. Several household characteristics were tested as being predictable from the ambient data. The most important contribution to the household classification approach from my work was the newly defined features from SMD and the inclusion of additional data sources using empirical feature extraction. The data processing within the approach was largely expanded through the definition of new features, the testing of ML algorithms and FSMs. In addition, different consumption data resolutions were tested and external data sources such as geographical data from OSM, weather data and statistical data were tested. The transferability of models trained in one country can be applied in another country with little loss of classification accuracy.

Research encompassed seven case studies from energy retailing with up to 20,000 residential customers in central Europe

Earlier research mainly excluded testing predicted household data in field studies. I pursued one such field study in which predicted household characteristics were tested in an energy feedback mailing campaign. A personalized home energy report led to energy conservation, user engagement and customer satisfaction. Future research should extend this approach to other applications in the field of energy production or supply. I myself will continue this research, for example on predicting characteristics of small and medium enterprises. In a recent study (Stingl et al. 2018), we found indications that this seems possible. The prediction of general household characteristics should be extended to investigate single household properties in more detail. In Hopf, Kormann, et al. (2017), for instance,

The approach should be transferred to other fields of energy data analytics

we investigated the detection of households with high potential for photovoltaic installations based on open data.

8.2.3 Relationship marketing

Maximizing customer value through up- and cross-selling

Contemporary marketing aims to develop long-term seller-buyer relationships (Brassington and Pettitt 2006). Thereby, firms try to better manage customers and maximize customer lifetime value (Kumar 2018). Their ultimate goal is to realize up-selling (i.e., selling products of the same category with higher margins) and cross-selling (e.g, selling products of different categories), or try to reduce customer churn. For all of these activities, detailed knowledge on customers is necessary, but often not available.

Insights on indivisual customers support targeted marketing

The proposed approach supports relationship marketing by providing new insights into the customer base. Forecasts of purchasing intentions help operational marketing to better select customers for marketing campaigns and strategic marketing to plan cost-benefit optimal campaign sizes. I demonstrated this contribution in chapter 7 with a cross-selling campaign from energy retail. The prediction of existing photovoltaic installations, home ownership, and purchasing intention of solar installations are also feasible and create manifold ways to develop and implement new business models in energy retail. Besides the single cross-selling score obtained through ML—which is often not meaningful to sales agents—the approach also reveals other household characteristics that help to explain high or low scores or to better select customers. Finally, the socioeconomic variables available for customers help to better value each customer. From previous research it is known that firms which systematically evaluate their customers outperform competitors. I call for future research that investigates how the predicted household characteristics change the valuation of customers, as additional opportunities for firms exist to make sense of data through ML.

8.2.4 Machine learning

The case studies from energy retail conducted in this dissertation helped to show the application of state-of-the-art ML models in

the energy retail industry, which has not been done extensively so far. The results of my dissertation indicate that the Random Forest algorithm is a good choice for business data analytics. This supports findings from a comprehensive benchmark of ML algorithm by Fernández-Delgado et al. (2014). Automatic feature selection is a powerful tool to cope with data dimensionality, but has received less attention from business analytics researchers so far. Even when multiple approaches exist, a rigorous benchmark of many available methods is—to the best of my knowledge—missing. I proposed a systematic benchmark approach that considers classification accuracy improvement, stability of the selected feature set, number of selected features, and computational complexity that I used to compare 43 feature selection methods available in the programming environment R using a dataset from my case studies. The results reveal that there is no superior approach, but some feature selection methods perform well in many cases. The benchmark should be validated with other datasets that are not domain-specific in the future.

8.3 Assumptions and limitations

In addition to limitations named in the separate chapters, I consider some overarching limitations of my present work that are summarized below.

Legal and Ethical issues The processing of personal data must always be carried out with the consent of the customer and on a legal basis. Nonetheless, such a requirement should be no obstacle for data-driven innovations. Obtaining consent from individuals is feasible when a clear value proposition is associated with the data use. Customers willingly disclose their personal data in cases where they receive an additional value. This can be seen with operators like Facebook (users want to be part of the social network and want to use these functions, so they share their data) and Google (people want to use the services offered for free and are willing to provide app usage data or personal location information).

Clear value propositions help to gain the consent of individuals for data use

A major limitation of public available data is the copyright under which the data is released. For open data that is released with the explicit right to use and share it, this is not a problem for the users of data. Websites and online portals in general, admittedly, can only be used or redistributed in those cases where the copyright holder grants a right of use.

Bias in the data The customer surveys conducted have in some cases only a limited number of responses (especially dataset C and D) and customers were located in one town in Switzerland. The data may contain a selection bias. The consideration of a longer timespan for the smart meter data could further improve the prediction.

8.4 Practical implications

My work has several practical implications. On the one hand, results have consequences for the business of utility companies that have residential customers. Several challenges that these established companies face can be turned into opportunities considering the power of ML and predictive analytics. On the other hand, my work revealed several issues that are relevant for data scientists and managers for all companies which have ambient data present. I describe the key take-aways for energy utility companies first and conclude then with those relevant for all firms.

8.4.1 Utilities can turn challenges into opportunities through data-driven innovations

Opportunities through data-driven innovations

The utility industry faces groundbreaking changes in their business. The most radical one is probably the impending replacement of the current fossil-nuclear energy supply with an alternative, sustainable energy supply. This step becomes necessary due to the scarcity of fossil resources, the risks of nuclear energy production, and the associated environmental pollution of old energy sources (Dangerman and Schellnhuber 2013). The expansion of renewable energy is often

decentral and residential customers are more often becoming producers of photovoltaic energy or become load-balancers when they install a battery at home to supply their own energy. The question for energy suppliers is therefore, whether they can integrate the customers that begin to produce energy (technical question), and how utility companies can enter new businesses by offering photovoltaic and storage solutions (market positioning question). The approach developed and evaluated in this dissertation provides aid for utility companies in this process. First, suitable customers for renewable energy installations can be found and second, the intention of residents towards renewable energy production and buying power can help to tailor personalized offerings and prioritize sales. Third, the knowledge about customers can be used to develop new products that help energy utilities to survive in their competitive market.

Business models for expensive smart-meter infrastructure

The energy transition will force utilities to highly invest in their fixed assets. It is necessary to invest not only in new power plants, but also in the energy transmission infrastructure. In particular, in a "smart grid" that is able to balance between volatile energy production from solar and wind and the actual demand of energy on the consumer-side. This work demonstrated the possibilities of smart meter data processing. It is now in the hands of firms to create new business models based on the predicted customer knowledge.

Possible innovations in energy retail

These challenges in the utility industry also entail opportunities and innovative companies can gain competitive advantages through service innovation. Three examples of such innovations are (Gebauer et al. 2014): 1) to improve the customer service, 2) foster basic and advanced innovations to increase energy efficiency, and 3) enter new business fields (e.g., electric mobility, home-automation, telecommunication, sales of and service for renewable energy installations). All three possible innovations can be supported by the ongoing digitalization and datication, because utility companies hold millions of data points on their customers. The available amount of data will further increase in the future due to the roll-out of smart meter and IoT infrastructures. Techniques of predictive analytics will help utilities to make sense of that data and can enable them to create innovative products and services. Possible applications of known household characteristics may include:

- Up-selling: Identify customers with willingness to pay for sustainable energy (e.g., locally produced energy, green electricity)
- Development of new products meeting the special needs of several customer groups, considering ideas of dynamic pricing, carbon-offsetting to gain competitive advantage through hyperdifferentiation of electricity products, or offering of energy consulting services
- Cross-selling: customers that are suitable for photovoltaic and battery self-supply and storage solutions, that are interested in FTTH internet access and other products that utilities might offer in addition to electricity
- Identify customers that may change their consumption behavior and relive the grid through load-shifting, adding flexible loads (with buffers such as batteries and electric cars)
- Increase customer satisfaction through better customer communication and new services like (semi-)automated energy consultancy by informing households about extreme high or low energy consumption compared with similar households or earlier periods

Hyper-differentiation

One possible strategy to use the obtained customer knowledge is *hyperdifferentiation*. This marketing strategy tries to gain value from offering more and more variants of the same product and thus broaden the product lines. This meets the customers' habits of seeking variety (Feinberg et al. 1992; Kahn 1995) and the fact that customers choose among all available product offerings instead of only selecting products from the same brand (Fader and Hardie 1996). Considering both phenomena, companies can profit from increased product differentiation because of the fact that nearly all products and their variants are producible. Clemons and colleagues (2003; 2004) coined the term hyperdifferentiation for the ability of firms to produce almost anything that any potential customer might want and the fact that these products generate profitability through variations that attract customers belonging to targeted micro-segments

(by using flexible pricing strategies and tailored offerings for individual customers based on detailed data on customers available). Furthermore, the "information availability and the use of this information by consumers has so profoundly affected consumer purchasing behavior that all of the underlying premises of corporate strategy require careful reexamination" (Clemons 2008, p. 14). Many firms already use hyperdifferentiation strategies. Examples are cosmetics, craft-beer (Clemons, Gao, et al. 2006), air travel (Granados et al. 2012) or gambling (Nair et al. 2017). The energy industry is still less innovative than other industries (Defeuilley 2009) and may profit from differentiating energy products—which are, from a customer viewpoint, just necessary infrastructure—and sell more added-value-services (e.g., energy-audits for homeowners, carbon offset possibilities for eco-oriented customers, flat-rate-charging for electric vehicle owners) as product bundles and make use of hyperdifferentiation strategies.

8.4.2 Recommendations for introducing predictive analytics in firms

Companies are at the beginning of the productive adoption of ML based predictive analytics. Consequently, they need to learn how to use existing tools in a meaningful way. At the moment there are many calls to hire data scientists, but the existing staff in organizations can already use many of the methods and techniques because they are freely available to everyone in open software environments.

The core resource needed to make sense out of data is, in my opinion, already often present in most companies: This is expert knowledge in the minds of employees. The results of my investigation backed this statement, as there is no superior algorithm (in the areas ML and feature selection) to solve the variety of business problems. Besides, data preparation must involve human cognition, theory and expert knowledge to transform data into highly expressive, and thus predictive, variables.

Data analytics is a source of innovation. Not only the final result of an analysis—be it a report created once or the prediction of customer details embedded into a business process—creates value for firms. This is because companies that begin not only to inventorize

their data stocks (which has already extensively happened in many companies due to the introduction of business intelligence systems), but also to analyze them, find massive problems in their databases. Such problems can be missing values, redundancies (e.g., when several departments maintain address data of customers), problematic data formats (e.g., if a current contract has saved the year "9999" as the end of the contract) or weakly structured data (plain text notes on the customer). Furthermore, the insights generated on the way to the final analytics result provide many insights into the business and demonstrate potential for improvements. I call therefore to not only evaluate the final result of an analysis, but also to honor the positive side effects of data-driven decision making processes. These positive side effects are a reason why firms should be careful with outsourcing of analytics.

For companies or departments that are taking their first steps in analytics, I can make the following recommendations:

- Random Forest is a powerful algorithm to kick-start predictive analytics projects. It yielded good classification results in this and earlier investigations and is—at least to some extent—explainable, because the influence of each feature can be quantified in terms of the feature importance. The feature importance scores help to find good predictor variables and can also help to identify good predictors.

- Insights do not result from simply applying algorithms to data. It is rather a process of testing algorithms, refining the data, varying parameters, always keeping the analytics goal in mind. Hopes for "quick wins" gaining value from data and ML may not be fulfilled.

- It is important to define the dependent variable in each specific case. Each problem must be analyzed in detail, with different data sources and features. This means that there is no off-the-shelf solution for predictive analytics.

- Evaluation of prediction models should be done rigorously with different performance metrics to avoid wrong assumptions regarding the classification quality. Cross-validation

should be used to estimate the performance with a confidence interval.

- The prediction of events that occur less frequent (resulting in skewed distribution of dependent variables) is tough. Models can be trained using under- or oversampling where observations of the more frequent class are excluded (undersampling), or observations of the minor class are created (oversampling). Alternatively, some algorithms support class-weights that help to better recognize infrequent events. Nevertheless, such approaches can easily lead to over-fitted models and should be tested on independent training data.

- When predictive analytics is used in organizations, the managers must learn to make decisions based on such predicted (uncertain) information. Business processes that rely on such uncertain data might need to be adapted. Predicted information should be used only in suitable cases where uncertain outcomes are acceptable.

To summarize, firms are better off beginning to use analytics today and training their experienced staff in the basics of statistics and data analytics. Even if data analytics tasks are outsourced to consulting agencies, it is necessary to be able to evaluate the quality of predictions based on ML from an internal perspective. In many cases, dedicated data analysis vendors specializing in certain industries are better suited than all-encompassing consulting agencies with little context knowledge or industry experience.

A Systematic literature analysis on predictive analytics

A literature analysis was conducted, intended to collect studies related to the topic of *predictive analytics* in Information Systems (IS) research. This literature survey focused thereby on empirical studies that develop or investigate a predictive model (in an empirical/real-world context). In particular, the focus of this literature analysis was to gain an overview to (a) empirical contexts for which predictive analytics was used so far, (b) predictive analytics (machine learning) algorithms that are typically applied in IS research studies, (c) data that is used in predictive analytics applications, as well as (d) discussion and review articles on the topic of predictive analytics, In a first step, articles were retrieved using a keyword search in literature databases. This was done between April and September 2018. Thereafter, I conducted a content analysis (Weber 1990) of the articles' title, abstract and keywords to extract the mentioned information from the article metadata.

A.1 Data collection

The search terms used for the literature analysis are listed below, together with a brief explanation of their relevance for the topic and the context where these search terms appear in the found articles (terms consisting of multiple words have been used for the database search with quotation marks to identify their co-occurence):

predictive X where X stands for analytics, analysis, modeling, modelling, power, or value; all word combinations indicate empirical predictive analytics studies

forecasting aims to predict future developments; this term is used to describe empirical forecasts as well as the description of future developments (e.g., in markets or a research field)

detection this word is often used to describe cases where a circumstance shall be identified (e.g., fraud or or a suspicious activity); methods of predictive anaytics are typically applied in such cases

machine learning describes methods and algorithms that are typically used for predictive analytics

neural network a class of machine learning algorithms that has attracted much research interest and is often used as a surrogate for machine learning

statistical learning another term for machine learning algorithms with a focus on statistical techniques

The keywords probably do not fully cover the variety of possible applications of predictive analytics in IS research, as it is described by Shmueli and Koppius (2011). Nevertheless, the set of search terms allows to obtain a sample of articles to gain an overview to the use of predictive analytics in the variety of fields and help to identify data sources available for analytics.

Initially, the search terms **data mining** and **regression analysis / analyses** were also considered as related but did not lead to many additional relevant papers, because most of the studies that were found with these terms were also found with the focus terms. The term "data mining" is, like the term "data analytics", a broader term without the focus on predive modleling and the studies related have often an explorative nature than a prediction.

In total, 155 articles were found in the AIS basket of top journals. An overview to the journals, used databases and fields are listed in Table A.1. The table contains also the not considered search terms "data mining" and "regression analysis" for illustrative purpose. Articles that were identified only based on these search terms are excluded in the remaining analysis. The frequency of search terms in the journal articles are listed in Table A.2 and the articles for each journal in five-year time intervals in Table A.3 and in Figure A.1. It is noticeable that most articles on predictive analytics have been published in JMIS, ISR and MISQ. The number of articles published in other journals has, however, increased in the recent years.

A.2 Content analysis

The identified articles were categorized using a content analysis (Weber 1990). Based on the title, abstract and keywords, all articles were categorized according to the paper type, machine learning algorithms, industry applications and used

Table A.1: Journals and database fields that have been searched, together with the corresponding databases

Journal	ISSN	Database	Fields
EJIS	0960-085X	Ingenta Connect	Title, Keywords, Abstract
ISJ	1350-1917	Business Source Ultimate (EBSCO Host)	Title, Subject Terms, Keywords, Abstract
ISR	1047-7047	Business Source Ultimate (EBSCO Host)	Title, Subject Terms, Keywords, Abstract
JAIS	1536-9323	AIS electronic library	Title, Subject, Abstract
JIT	0268-3962	Business Source Ultimate (EBSCO Host)	Title, Subject Terms, Keywords, Abstract
JMIS	0742-1222	Business Source Ultimate (EBSCO Host)	Title, Subject Terms, Keywords, Abstract
JSIS	0963-8687	Science Direct	Title, Keywords, Abstract
MISQ	2162-9730	AIS electronic library	Title, Subject, Abstract

Table A.2: Articles in the AIS basket of top journals for the search terms (one article may belong to multiple search terms)

Search term	EJIS	ISJ	ISR	JAIS	JIT	JMIS	JSIS	MISQ	Sum
forecasting	1	1	9	3	2	19	1	6	42
machine learning	0	0	7	0	0	15	0	2	24
neural network	0	1	1	0	2	14	1	0	19
predictive X	1	2	10	1	1	13	3	12	43
detection	2	1	14	4	5	25	0	5	56
statistical learning	0	0	0	0	0	0	0	4	4
data mining	1	3	14	2	3	32	0	8	63
regression analysis	5	5	8	1	1	23	2	3	48
Sum	10	13	63	11	14	141	7	40	299

Table A.3: Identified articles for the considered search terms over time

Years	EJIS	ISJ	ISR	JAIS	JIT	JMIS	JSIS	MISQ	Sum
<=1993	0	0	3	0	4	8	0	0	15
1994–1998	0	0	6	0	1	8	1	0	16
1999–2003	0	0	4	0	0	10	1	0	15
2004–2008	0	2	2	3	1	11	0	5	24
2009–2013	3	2	10	0	1	8	1	6	31
2014–2018	1	0	12	4	2	22	2	12	55
Sum	4	4	37	7	9	67	5	23	156

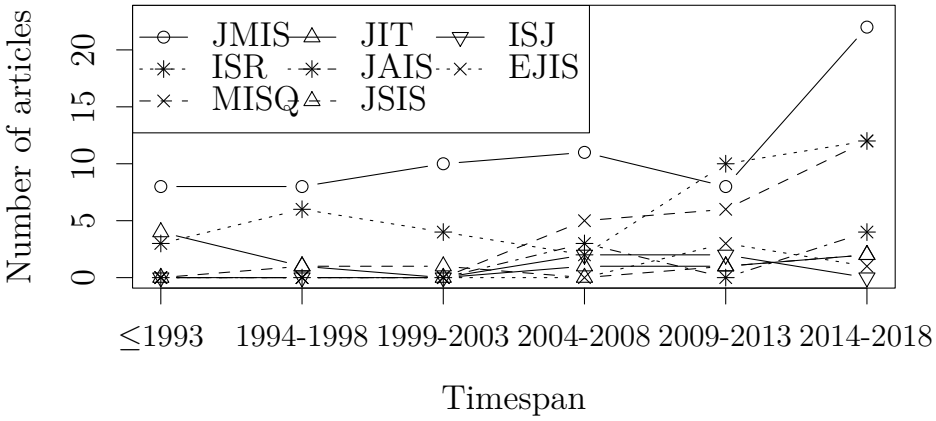


Figure A.1: Identified articles for the considered search terms in journals over time

data. This categorization was done using a coding scheme that was developed during the analysis in several iterations. In the first iteration, all articles were coded into one of three paper types (see Figure A.2). In the second iteration, all 58 focal articles were coded according to their industrial context, used machine learning algorithms and data sources that were used in the study.

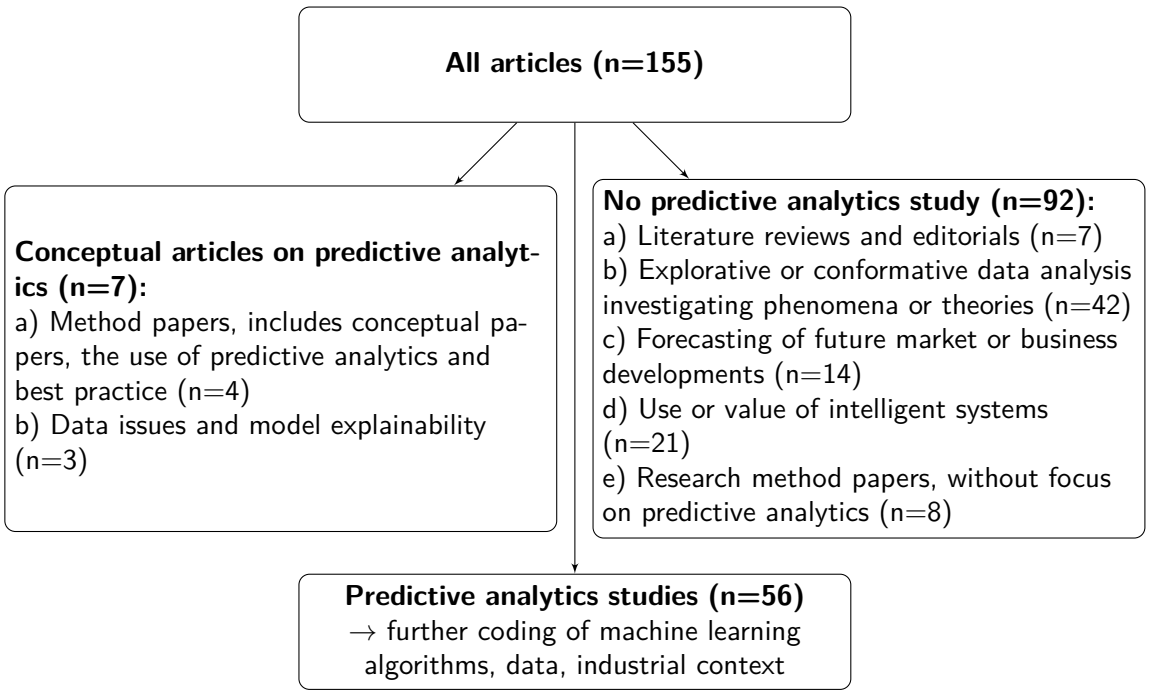


Figure A.2: Illustration of the codings into three paper types: predictive analytics studies (focus of this work), conceptual articles on predictive analytics and studies not investigating the application of predictive analytics

Table A.4: Industrials contexts in predictive analytics studies with the frequency of studies in AIS basket of top journals (multiple assignments possible)

Industrial context	Number of studies
financial / banking	19
sales and marketing	10
crime detection	7
e-business	6
business documents / communication	6
health care	4
deception (communication, security)	3
web search	2
mergers and acquisitions	2
demand forecasting	2
security screening	2
innovation management	2
knowledge acquisition (expert systems)	2
web usage workplace	1
other (single examples for patent administration, process mining, plagiarism, automotive, movie industry, hotel industry)	6

Industrial contexts Table A.4 shows the frequencies of industrial contexts in the studies in the sample. Studies can be assigned to more than one empirical context. The industrial contexts are diverse, however, the most frequent empirical context investigated so far seems to be the finance and banking industry, as well as security, crime and abuse.

Data sources In most of the papers, the used data was given in the article meta-data, but for nine articles, the full-text must be considered. Table A.6 shows the frequencies of data sources considered as predictor variables in the articles, multiple data sources have been used in several studies. Internal business data (including customer core data, transaction data, business process data, documents, ...), and available online data (especially public websites, social media and search engines usage data) was used most frequently. From 11 studies, the used data could not be read out of the article meta-data.

Methods Finally, the articles were coded regarding the applied predictive analytics method. Table A.8 shows the frequencies of studies. ANNs were the most used class of algorithms, followed by regression analysis. For 13 articles, the concrete algorithm could not be read out of the title, abstract or keywords and for further 10 articles document only that "machine learning" was used, but the concrete algorithm was not mentioned.

Table A.6: Data sources in predictive analytics studies with the frequency of studies in AIS basket of top journals (multiple assignments possible)

Data source for prediction	Number of studies
Business data (internal)	
transactions: purchase history, data on loans	4
sensor data (specially collected)	4
payment transactions	2
medical record database	4
customer data / demographics	2
communication and business documents	6
business process data	1
accounting data	1
External data	
online website content	12
social network / media data	10
business indicators / financial statements	7
search trend data	3
e-business platform data	2
public statistical data	1

Table A.8: Predictive modelling algorithms in reviewed studies in AIS basket of top journals (multiple assignments possible)

Data source for prediction	Number of studies
ANN	14
machine learning (no algorithm specified)	10
text analysis	9
log / linear regression	7
genetic / evolutionary algorithms	6
decision tree learner	5
feature extraction / engineering	3
kernel-based learners and SVM	3
LDA	3
time-series analysis	2
kNN	2
bayesian learning	2
grammatical inference	1
social network analysis	1
recursive partitioning	1

B Conducted case studies in energy retail

During the dissertation research project, several case studies have been conducted together with energy retailers in Germany and Switzerland, the data analytics vendor, BEN Energy AG (Zurich, Munich) and by using data that is publicly available for research purposes. This chapter gives an overview to the case studies with their empirical context, data used, challenge pursued, and publications.

1. Household classification with smart meter electricity consumption data
2. Household classification (characteristics and intentions) with smart meter electricity consumption data and external data
3. Predicting purchasing intention for cross-selling (Fiber-to-the-Home internet access)
4. Household classification (socioeconomic variables) with annual electricity consumption data and external data
5. Value of statistical and geographical data for household classification
6. Cross-selling sustainable product (detecting old heating systems)
7. Household classification with daily electricity, gas and water consumption data, as well as external data for energy efficiency feedback (with field test)

The research in case studies 1–3 was partly funded by Swiss Federal Office of Energy, grant numbers SI/501053-01 and SI/501202-01. The research in case studies 1 was partly funded by Commission for Technology and Innovation in Switzerland (CTI Grant number 16702.2 PFEN-ES). The research in case studies 4–7 was funded by European Commission, grant number E!9859.

Table B.1: Overview to conducted case studies in this dissertation research project

No.	Years	Data on households	Country	Sample	Impact / Application	References
1	2014–2015	Smart meter data and surveys	IR	4,232	Automated energy consulting	Hopf, Sodenkamp, Kozlovskiy, and Staake (2014), Sodenkamp, Hopf, and Staake (2015)
2	2015–2016	Smart meter data and surveys	CH	451	Automated energy consulting	Sodenkamp, Kozlovskiy, et al. (2017), Hopf, Sodenkamp, and Staake (2018), Sodenkamp, Hopf, Kozlovskiy, et al. (2016), chapter 5
3	2015	Smart meter data and surveys	CH	451	Purchasing probability, estimating market size	Chapter 7
4	2015–2017	BEN Energy customer engagement portals	CH	5,446	Household classification with annual data, support for targeted marketing, features from VGI data	Hopf, Sodenkamp, and Kozlovskiy (2016), Hopf (2018)
5	2016–2018	BEN Energy customer engagement portals, statistical data, VGI data	DE	2,058 (plus data from case 4)	Algorithm transferability, features from statistical and VGI data, Relationship marketing	Hopf, Riechel, et al. (2017)
6	2015–2016	Utility company data	NL	7,582	Economic and ecologic benefits	Kozlovskiy et al. (2016)
7	2017	BEN Energy customer engagement portals and surveys	CH	969, (414 in experiment)	Energy efficiency mailing and customer satisfaction	Chapter 6

C Survey instruments

This appendix summarizes survey questions and measurement instruments used in this work. As all instruments were used in surveys with German-speaking participants, the used translation is given.

C.1 Environmental attitude

The scale for environmental attitude is taken from the factor *Energy conservation* in the behavior-based environmental attitude scale of Kaiser et al. (2007) and has six items (EA1–EA6) that was translated to German Table C.1. In addition, a question (EA7) from the Swiss environmental survey of Diekmann and Bruderer Enzler (2012) and Diekmann and Franzen (1999).

Table C.1: Items for the behavior based measurement instrument for attitudes towards energy conservation with German translations

Name	Item (Original English question and *German translation*)
EA1	After one day of use, my sweaters or trousers go into the laundry* *Wenn ich Pullover oder Hosen einen Tag lang getragen habe, kommen sie immer in die Wäsche**
EA2	As the last person to leave a room, I switch off the lights *Wenn ich als letzte Person den Raum verlasse, mache ich immer das Licht aus*
EA3	I leave electrically powered appliances (TV, stereo, printer) on standby* *Ich lasse strombetriebene Geräte (TV, Stereoanlage, Drucker, ...) immer auf Standby laufen**
EA4	In the winter, I turn down the heat when I leave my room for more than 4 hours *Im Winter drehe ich das Thermostat runter, wenn ich die Wohnung für mehr als vier Stunden verlasse*
EA5	In the winter, it is warm enough in my room to only wear a T-shirt* *Im Winter ist es in meiner Wohnung immer warm genug, um nur ein T-Shirt zu tragen**
EA6	In hotels, I have the towels changed daily* *In Hotels lasse ich die Handtücher immer täglich austauschen**

Name	Item (Original English question and *German translation*)
EA7	I do what is right for the environment, even when it costs more money or takes more time. *Ich verhalte mich auch dann umweltbewusst, wenn es erheblich höhere Kosten und Mühen verursacht.*

* negative question

C.2 Customer-based reputation of a firm

To raise the customer satisfaction, the short version of the customer-based reputation (CBR) scale of firmWalsh, Beatty, and Shiu (2009) and Walsh and Beatty (2007) is used. The factor "Good employer" and "Social and Environmental Responsibility" is left out. Due to the use in the non-liberalized energy market in Switzerland, the statements that the firm "... tends to outperform competitors" and "... looks like it has strong prospects for future growth" was not used.

Table C.2: Selected items from CBR-Short Scale: Item and German translation

Name	Item (Original English question and *German translation*)
Factor 1: Customer Orientation	
REP1	Has employees who are concerned about customer needs *Die Mitarbeiter des Unternehmens kümmern sich um die Bedürfnisse der Kunden*
REP2	Has employees who treat customers courteously *Die Mitarbeiter des Unternehmens behandeln die Kunden höflich*
REP3	Is concerned about its customers *Das Unternehmen kümmert sich um seine Kunden*
Factor 3: Reliable and Financially Strong Company	
REP4	Seems to recognize and take advantage of market opportunities *Das Unternehmen scheint neue Marktchancen zu erkennen und zu nutzen*
Factor 4: Product and Service Quality	
REP5	Offers high quality products and services *Die Produkte und Dienstleistungen des Unternehmens sind von hoher Qualität*
REP6	Is a strong, reliable company *Die Firma ist ein starkes, verlässliches Unternehmen*
REP7	Develops innovative services *Das Unternehmen entwickelt innovative Dienstleistungen*
Other items, not included in original scale (Walsh and Beatty 2007)	

Name	Item (Original English question and *German translation*)
REP8	I am satisfied with the services that the company offers *Ich bin mit den Leistungen, die das Unternehmen anbietet, zufrieden*
REP9	You can trust this company *Diesem Unternehmen kann man vertrauen*
REP10	I would probably report good things about the company to others *Ich würde wahrscheinlich gute Dinge über das Unternehmen Anderen gegenb̈er berichten*

C.3 Purchase intention

The purchase intention scale with three items as listed in Table C.3 was used by H.-W. Kim et al. (2007) and was translated into German.

Table C.3: Purchase intention scale used by H.-W. Kim et al. (2007) with German translations

Name	Item (Original English question and *German translation*)
PI1	I could imagine buying [product] in the next n years. *Ich könnte mir vorstellen, in den nächsten n Jahren [Produkt] anzuschaffen.*
PI2	I intend to purchase [product] in the next 1-2 years. *Ich beabsichtige die Anschaffung von [Produkt] in den nächsten n Jahren.*
PI3	I plan to purchase a solar system in the next 1-2 years. *Ich plane, in den nächsten n Jahren [Produkt] anzuschaffen.*

As the number of questions may be a criterion in survey development, an alternative single-item scale for the purchase intention is relevant, An acknowledged scale is that of Juster (1966) which consists of the question "How do you estimate the prospects that you will buy [product] within the next n months?" (in German: "Wie hoch schätzen Sie die Chance ein, dass Sie innerhalb der nächsten n Monate [Produkt] beziehen werden?"). I refer to it as *JUS* in the text. The responses to the question is given by selecting a number from an eleven-point probability scale shown in Table C.4.

The scale was, in agreement with the research partner and after discussions with uninvolved persons, freely translated to German and the reference to "even chance (50-50)" removed, as this was regarded as difficult to assess for the survey participants. The fine gradations in the language still express an increasing support of the statement from 1 to 10 after the translation.

Table C.4: Responses to the purchase intention scale of Juster (1966) with German translation

Alternative	Statement (Original English question and *German translation*)
10	Absolutely certain to buy *Sicher, auf jeden Fall*
9	Almost certain to buy *Fast sicher*
8	Much better than even chance *Sehr wahrscheinlich*
7	Somewhat better than even chance *Wahrscheinlich*
6	Slightly better than even chance *Gut möglich*
5	About even chance (50-50) *Eventuell möglich*
4	Slightly less than even chance *Mit geringer Wahrscheinlichkeit*
3	Somewhat less than even chance *Mit sehr geringer Wahrscheinlichkeit*
2	Much less than even chance *Unwahrscheinlich*
1	Almost no chance *Sehr unwahrscheinlich*
0	Absolutely no chance *Nein, überhaupt nicht*
–	I don't know what [product] is. *Ich weiß nicht, was [Produkt] ist*

C.4 Usability perception scale for energy feedback

To investigate the user acceptance of an energy report, the Usability Perception Scale (UPscale), a recently published and tested scale for perceived usability of eco-feedback (Karlin and Ford 2013) was used. The original scale was published in English. We therefore translated the items to German (see Table C.5). To obtain a proper translation, we followed a common three step procedure. First, the questions were translated by me and two team members independently and almost literally. Second, we merged the German questions to one common translation and discussed them with one additional team member and our research partner. In this step, we changed some formulations into a more free translation. Third, the German questions have been translated back to English to ensure validity to the original scale. As the original scale was developed for one single visualization, we replaced the word "image" by "energy report" in all questions before translation.

Table C.5: UPScale: Items and German translation

Name	Item (Original English question and *German translation*)
UP1	I am able to get the information I need easily *Der Energiereport enthält nützliche Informationen*
UP2	I think the image (energy report) is difficult to understand* *Ich denke, der Energiereport ist schwer verständlich**
UP3	I feel very confident interpreting the information in this image (energy report) *Ich bin mir sicher, die Informationen aus dem Energiereport richtig zu interpretieren*
UP4	A person would need to learn a lot in order to understand this image (energy report)* *Eine Person müsste viel lernen um den Energiereport zu verstehen**
UP5	I gained information from this image (energy report) that will benefit my life *Ich erhalte Informationen aus dem Energiereport, die mein Leben bereichern*
UP6	I do not find this image (energy report) useful* *Ich finde diesen Energiereport nutzlos**
UP7	I think that I would like to use this image (energy report) frequently *Ich nutze die Informationen aus dem Energiereport häufig*
UP8	I would not want to use this image (energy report) *Ich möchte diesen Energiereport gerne weiter nutzen*

* negative question

Bibliography

Abbasi, Ahmed, Suprateek Sarker, and Roger Chiang (2016). "Big Data Research in Information Systems: Toward an Inclusive Research Agenda." In: *Journal of the Association for Information Systems* 17.2.

Alaimo, Cristina and Jannis Kallinikos (Aug. 8, 2017). "Computing the Everyday: Social Media as Data Platforms." In: *The Information Society* 33.4, pp. 175–191. DOI: 10.1080/01972243.2017.1318327.

Albert, A. and R. Rajagopal (2013). "Smart Meter Driven Segmentation: What Your Consumption Says About You." In: *IEEE Transactions on Power Systems* 28.4, pp. 4019–4030.

— (Nov. 2014). "Cost-of-Service Segmentation of Energy Consumers." In: *IEEE Transactions on Power Systems* 29.6, pp. 2795–2803. DOI: 10.1109/TPWRS.2014.2312721.

Albert, Terri and Paulo Goes (June 1, 2004). "GIST: A Model for Design and Management of Content and Interactivity of Customer-Centric Web Sites." In: *MIS Quarterly* 28.2.

Alfaro, Esteban, Matıas Gámez, and Noelia Garcia (2013). "Adabag: An R Package for Classification with Boosting and Bagging." In: *Journal of Statistical Software* 54.2, pp. 1–35.

Allcott, Hunt (Oct. 2011). "Social Norms and Energy Conservation." In: *Journal of Public Economics*. Special Issue: The Role of Firms in Tax Systems 95.9–10, pp. 1082–1095. DOI: 10.1016/j.jpubeco.2011.03.003.

Allcott, Hunt and Sendhil Mullainathan (2010). "Behavior and Energy Policy." In: *Science* 327.5970, pp. 1204–1205.

Alshawi, Sarmad, Farouk Missi, and Zahir Irani (Apr. 1, 2011). "Organisational, Technical and Data Quality Factors in CRM Adoption — SMEs Perspective." In: *Industrial Marketing Management*. Special Issue on Industrial Marketing Strategy and B2B Management by SMEs 40.3, pp. 376–383. DOI: 10.1016/j.indmarman.2010.08.006.

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Apadula, Francesco et al. (2012). "Relationships between Meteorological Variables and Monthly Electricity Demand." In: *Applied Energy* 98.0, pp. 346–356. DOI: 10.1016/j.apenergy.2012.03.053.

Armel, K. Carrie et al. (Jan. 1, 2013). "Is Disaggregation the Holy Grail of Energy Efficiency? The Case of Electricity." In: *Energy Policy*. Special Section: Transition

Pathways to a Low Carbon Economy 52 (Supplement C), pp. 213–234. DOI: 10.1016/j.enpol.2012.08.062.

Armstrong, J. Scott, Vicki G. Morwitz, and V. Kumar (July 2000). "Sales Forecasts for Existing Consumer Products and Services: Do Purchase Intentions Contribute to Accuracy?" In: *International Journal of Forecasting* 16.3, pp. 383–397. DOI: 10.1016/S0169-2070(00)00058-3.

Atasoy, Hilal (2013). "The Effects of Broadband Internet Expansion on Labor Market Outcomes." In: *Industrial & Labor Relations Review* 66.2, pp. 315–345.

Backiel, Aimée, Bart Baesens, and Gerda Claeskens (June 1, 2014). "Mining Telecommunication Networks to Enhance Customer Lifetime Predictions." In: *Artificial Intelligence and Soft Computing*. International Conference on Artificial Intelligence and Soft Computing. Lecture Notes in Computer Science. Cham: Springer, pp. 15–26. DOI: 10.1007/978-3-319-07176-3_2.

Balzer, Frederike et al. (June 2015). *Daten zur Umwelt 2015*. Dessau-Roßlau, Germany: Umweltbundesamt.

Banasiewicz, Andrew D. (2013). *Marketing Database Analytics: Transforming Data for Competitive Advantage*. London, UNITED KINGDOM: Taylor and Francis.

Bauder, Harald (Jan. 1, 2002). "Neighbourhood Effects and Cultural Exclusion." In: *Urban Studies* 39.1, pp. 85–93. DOI: 10.1080/00420980220099087.

Beckel, Christian (2015). "Scalable and Personalized Energy Efficiency Services with Smart Meter Data." Doctoral Thesis. ETH Zurich. DOI: 10.3929/ethz-a-010578740.

Beckel, Christian, Leyna Sadamori, and Silvia Santini (2012). "Towards Automatic Classification of Private Households Using Electricity Consumption Data." In: *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. Ed. by George J. Pappas. Toronto: ACM, pp. 169–176.

— (2013). "Automatic Socio-Economic Classification of Households Using Electricity Consumption Data." In: Proceedings of the the Fourth International Conference on Future Energy Systems - e-Energy '13. Ed. by David Culler et al. Berkeley, California, USA: ACM Press, p. 75. DOI: 10.1145/2487166.2487175.

Beckel, Christian, Leyna Sadamori, Thorsten Staake, et al. (2014). "Revealing Household Characteristics from Smart Meter Data." In: *Energy*. Vol. 78, pp. 397–410.

Becker, Vincent and Wilhelm Kleiminger (Aug. 30, 2017). "Exploring Zero-Training Algorithms for Occupancy Detection Based on Smart Meter Measurements." In: *Computer Science - Research and Development*, pp. 1–12. DOI: 10.1007/s00450-017-0344-9.

Bemmaor, Albert C. (May 1995). "Predicting Behavior from Intention-to-Buy Measures: The Parametric Case." In: *Journal of Marketing Research (JMR)* 32.2, pp. 176–191.

Bergers, Ron and Jasper Meijerink (Nov. 16, 2017). *The New Gold: Crafting the Business Case for an Insight Driven Organisation*. URL: https://www2.deloitte.com/nl/nl/pages/data-analytics/articles/the-new-gold.html (visited on 11/28/2018).

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization." In: *Journal of Machine Learning Research* 13 (Feb), pp. 281–305.

Bhattacharyya, Siddhartha et al. (Feb. 1, 2011). "Data Mining for Credit Card Fraud: A Comparative Study." In: *Decision Support Systems*. On Quantitative Methods for Detection of Financial Fraud 50.3, pp. 602–613. DOI: 10.1016/j.dss.2010.08.008.

Biau, Gérard (2012). "Analysis of a Random Forests Model." In: *Journal of Machine Learning Research* 13 (Apr), pp. 1063–1095.

Birt, Benjamin J. et al. (2012). "Disaggregating Categories of Electrical Energy End-Use from Whole-House Hourly Data." In: *Energy and Buildings* 50, pp. 93–102.

BMUB (Dec. 3, 2014). *Aktionsprogramm Klimaschutz 2020*. Kabinettsbeschluss. Berlin, Germany: Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit.

BNetzA (Nov. 30, 2016). *Monitoring Report 2016*. Bonn, Germany: Federal Network Agency and Federal Cartel Office.

Bolón-Canedo, Verónica, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos (2015). *Feature Selection for High-Dimensional Data*. Artificial Intelligence: Foundations, Theory, and Algorithms. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-21858-8.

Bowne-Anderson, Hugo (Aug. 15, 2018). "What Data Scientists Really Do, According to 35 Data Scientists." In: *Harvard Business Review*.

Brandt, Tobias et al. (June 1, 2018). "Smart Cities and Digitized Urban Management." In: *Business & Information Systems Engineering* 60.3, pp. 193–195. DOI: 10.1007/s12599-018-0537-1.

Brassington, Frances and Stephen Pettitt (2006). *Principles of Marketing*. 4th ed. New York: Prentice Hall. 1264 pp.

Braun, Michael and David A. Schweidel (Sept. 2011). "Modeling Customer Lifetimes with Multiple Causes of Churn." In: *Marketing Science* 30.5, pp. 881–902. DOI: 10.1287/mksc.1110.0665.

Breiman, Leo (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

— (2001). "Random Forests." In: *Machine Learning* 45.1, pp. 5–32.

Brooke, John et al. (1996). "SUS-A Quick and Dirty Usability Scale." In: *Usability evaluation in industry* 189.194, pp. 4–7.

Brooks-Gunn, Jeanne et al. (Sept. 1, 1993). "Do Neighborhoods Influence Child and Adolescent Development?" In: *American Journal of Sociology* 99.2, pp. 353–395. DOI: 10.1086/230268.

Brown, Susan A. and Viswanath Venkatesh (2005). "Model of Adoption of Technology in Households: A Baseline Model Test and Extension Incorporating Household Life Cycle." In: *MIS Quarterly* 29.3, pp. 399–426. DOI: 10.2307/25148690.

Burset, Moisès and Roderic Guigó (June 15, 1996). "Evaluation of Gene Structure Prediction Programs." In: *Genomics* 34.3, pp. 353–367. DOI: 10.1006/geno.1996.0298.

Carrizosa, Emilio, Amaya Nogales-Gómez, and Dolores Romero Morales (Feb. 1, 2016). "Strongly Agree or Strongly Disagree?: Rating Features in Support Vector Machines." In: *Information Sciences*. Special Issue on Discovery Science 329, pp. 256–273. DOI: 10.1016/j.ins.2015.09.031.

Chandrashekar, Girish and Ferat Sahin (Jan. 2014). "A Survey on Feature Selection Methods." In: *Computers & Electrical Engineering*. 40th-Year Commemorative Issue 40.1, pp. 16–28. DOI: 10.1016/j.compeleceng.2013.11.024.

Chapman, Pete et al. (2000). *CRISP-DM 1.0*. SPSS.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System." In: KDD '16, August 13-17, 2016. San Francisco, USA, pp. 785–794. DOI: 10.1145/2939672.2939785.

Chiang, Roger H. L., Paulo Goes, and Edward A. Stohr (Oct. 2012). "Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline." In: *ACM Transactions in MIS* 3.3, 12:1–12:13. DOI: 10.1145/2361256.2361257.

Chicco, G. et al. (2001). "Electric Energy Customer Characterisation for Developing Dedicated Market Strategies." In: *Power Tech Proceedings, 2001 IEEE Porto*. Vol. 1.

Chicco, Gianfranco (2012). "Overview and Performance Assessment of the Clustering Methods for Electrical Load Pattern Grouping." In: *Energy* 42.1, pp. 68–80. DOI: 10.1016/j.energy.2011.12.031.

Ciepłuch, Błażej et al. (July 20, 2010). "Comparison of the Accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps." In: *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*, p. 337.

Clemons, Eric K. (Sept. 1, 2008). "How Information Changes Consumer Behavior and How Consumer Behavior Determines Corporate Strategy." In: *Journal of Management Information Systems* 25.2, pp. 13–40. DOI: 10.2753/MIS0742-1222250202.

Clemons, Eric K., Guodong Gao, and Lorin Hitt (Oct. 1, 2006). "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry." In: *Journal of Management Information Systems* 23.2, pp. 149–171. DOI: 10.2753/MIS0742-1222230207.

Clemons, Eric K., B. Gu, and R. Spitler (2003). "Hyper-Differentiation Strategies: Delivering Value, Retaining Profits." In: *Proceedings of the Thirty-Sixth Hawaii International Conference on System Sciences*. IEEE, 7 pp. DOI: 10.1109/HICSS.2003.1174592.

Clemons, Eric K. and R. Spitler (Oct. 7, 2004). "The New Language of Consumer Behavior." In: *Financial Times*, pp. 4–5.

Coltman, Tim (Sept. 1, 2007). "Why Build a Customer Relationship Management Capability?" In: *The Journal of Strategic Information Systems* 16.3, pp. 301–320. DOI: 10.1016/j.jsis.2007.05.001.

Coltman, Tim, Timothy M. Devinney, and David F. Midgley (Sept. 1, 2011). "Customer Relationship Management and Firm Performance." In: *Journal of Information Technology* 26.3, pp. 205–219. DOI: 10.1057/jit.2010.39.

ComCom (Mar. 2015). *Annual Report 2014*. Federal Communications Commission ComCom.

Commission for Energy Regulation (Mar. 16, 2011). *Electricity Smart Metering Customer Behaviour Trials (CBT) Findings Report*. Information Paper CER11080a, p. 146.

Constantiou, Ioanna D. and Jannis Kallinikos (Mar. 2015). "New Games, New Rules: Big Data and the Changing Context of Strategy." In: *Journal of Information Technology* 30.1, pp. 44–57. DOI: 10.1057/jit.2014.17.

Cover, T. M. and Joy A. Thomas (2006). *Elements of Information Theory*. 2nd ed. Hoboken, N.J: Wiley-Interscience. 748 pp.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Cui, Dapeng and David Curry (2005). "Prediction in Marketing Using the Support Vector Machine." In: *Marketing Science* 24.4, pp. 595–615.

Cui, Geng, Man Leung Wong, and Hon-Kwong Lui (Apr. 2006). "Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming." In: *Management Science* 52.4, pp. 597–612. DOI: 10.1287/mnsc.1060.0514.

Cui, Geng, Man Leung Wong, and Xiang Wan (Sum. 2012). "Cost-Sensitive Learning via Priority Sampling to Improve the Return on Marketing and CRM Investment." In: *Journal of Management Information Systems* 29.1, pp. 341–374.

Dangerman, A. T. C. Jérôme and Hans Joachim Schellnhuber (Dec. 2, 2013). "Energy Systems Transformation." In: *Proceedings of the National Academy of Sciences* 110.7, E549–E558. DOI: 10.1073/pnas.1219791110.

Dash, Manoranjan and Huan Liu (Dec. 2003). "Consistency-Based Search in Feature Selection." In: *Artificial Intelligence* 151.1-2, pp. 155–176. DOI: 10.1016/S0004-3702(03)00079-1.

Dash, Manoranjan, Huan Liu, and Hiroshi Motoda (2000). "Consistency Based Feature Selection." In: *Knowledge Discovery and Data Mining. Current Issues and New Applications*. Springer, pp. 98–109.

Daskalaki, S. et al. (Mar. 1, 2003). "Data Mining for Decision Support on Customer Insolvency in Telecommunications Business." In: *European Journal of Operational Research* 145.2, pp. 239–255. DOI: 10.1016/S0377-2217(02)00532-5.

Davis, Fred (Jan. 1, 1985). "A Technology Acceptance Model for Empirically Testing New End-User Information Systems." Doctoral Thesis. Massachusetts Institute of Technology.

Defeuilley, Christophe (Feb. 1, 2009). "Retail Competition in Electricity Markets—Expectations, Outcomes and Economics: A Reply." In: *Energy Policy* 37.2, pp. 764–765. DOI: 10.1016/j.enpol.2008.09.091.

Deichmann, Johannes, Matthias Roggendorf, and Dominik Wee (Nov. 2015). *Preparing IT Systems and Organizations for the Internet of Things — McKinsey & Company*. URL: https://www.mckinsey.com/industries/high-tech/our-insights/preparing-it-systems-and-organizations-for-the-internet-of-things (visited on 02/25/2018).

Dekimpe, Marnik G. and Dominique M. Hanssens (1995). "The Persistence of Marketing Effects on Sales." In: *Marketing Science* 14.1, pp. 1–21.

Dekkers, Makx et al. (June 2006). *Measuring European Public Sector Information Resources: Final Report of Study on Exploitation of Public Sector Information – Benchmarking of EU Framework Conditions*. European Commission.

Demšar, J. (2006). "Statistical Comparisons of Classifiers over Multiple Data Sets." In: *Journal of Machine Learning Research* 7, pp. 1–30.

Diekmann, Andreas and Heidi Bruderer Enzler (Feb. 21, 2012). *Projekt "Zeitpräferenzen Und Energiesparen": Codebook*. Zürich: ETH Zürich.

Diekmann, Andreas and Axel Franzen (July 1999). "The Wealth of Nations and Environmental Concern." In: *Environment and Behavior* 31.4, pp. 540–549. DOI: 10.1177/00139169921972227.

Dietterich, Tom, Michael Kearns, and Yishay Mansour (1996). "Applying the Weak Learning Framework to Understand and Improve C4.5." In: *Proceedings of the 13. International Conference on Machine Learning*. Morgan Kaufmann, pp. 96–104.

Dietz, Robert D. (Dec. 2002). "The Estimation of Neighborhood Effects in the Social Sciences: An Interdisciplinary Approach." In: *Social Science Research* 31.4, pp. 539–575. DOI: 10.1016/S0049-089X(02)00005-4.

DMLC (2016a). *Notes on Parameter Tuning — Xgboost 0.6 Documentation*. URL: http://xgboost.readthedocs.io/en/latest//how_to/param_tuning.html (visited on 10/14/2016).

— (2016b). *XGBoost Parameters — Xgboost 0.6 Documentation*. URL: http://xgboost.readthedocs.io/en/latest//parameter.html (visited on 10/13/2016).

Drisko, James and Tina Maschi (Dec. 1, 2015). *Content Analysis*. Oxford University Press. DOI: 10.1093/acprof:oso/9780190215491.001.0001.

Druckman, A. and T. Jackson (2008). "Household Energy Consumption in the UK: A Highly Geographically and Socio-Economically Disaggregated Model." In: *Energy Policy* 36.8, pp. 3177–3192. DOI: 10.1016/j.enpol.2008.03.021.

Dwyer, F. Robert, Paul H. Schurr, and Sejo Oh (1987). "Developing Buyer-Seller Relationships." In: *Journal of Marketing* 51.2, pp. 11–27. DOI: 10.2307/1251126.

Einhellig, Ludwig, Kamila Behrens, and Laetitia v. Preysing (9.07.14). *Einführung von Smart Meter in Deutschland*. Berlin: Deutsche Energie - Agentur GmbH.

Elkington, John (Win. 1994). "Towards the Sustainable Corporation: Win-Win-Win Business Strategies for Sustainable Development." In: *California Management Review* 36.2, pp. 90–100.

EMC (2014). *Digital Universe Study*. URL: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm (visited on 09/15/2018).

Enev, Miro et al. (2011). "Televisions, Video Privacy, and Powerline Electromagnetic Interference." In: *Proceedings of the 18th ACM Conference on Computer and Communications Security*. ACM, pp. 537–550.

Eto, Joseph, Steven Stoft, and Timothy Belden (1997). "The Theory and Practice of Decoupling Utility Revenues from Sales." In: *Utilities Policy* 6.1, pp. 43–55.

EU (Nov. 17, 2003). *Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the Re-Use of Public Sector Information*.

— (May 15, 2007). *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*.
— (2012). *Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on Energy Efficiency, Amending Directives 2009/125/EC and 2010/30/EU and Repealing Directives 2004/8/EC and 2006/32/EC Text with EEA Relevance*.
— (Dec. 9, 2013). *Commission Delegated Regulation (EU) No 1159/2013 of 12 July 2013 Supplementing Regulation (EU) No 911/2010 of the European Parliament and of the Council on the European Earth Monitoring Programme (GMES) by Establishing Registration and Licensing Conditions for GMES Users and Defining Criteria for Restricting Access to GMES Dedicated Data and GMES Service Information Text with EEA Relevance*.
European Commission (Jan. 10, 2017). *Building a European Data Economy*. COM(2017) 9 final. Brussels: European Commission.
Eurostat (Aug. 5, 2015). *Consumption of Energy - Statistics Explained*. Eurostat.
— (Jan. 25, 2017). *Final Consumption Expenditure of Households, by Consumption Purpose - Eurostat (Code: Tsdpc520, Last Update: 25/01/17)*. URL: http://ec.europa.eu/eurostat/web/products-datasets/-/tsdpc520 (visited on 01/25/2017).
Everett, Jerry Don (2009). "An Investigation of the Transferability of Trip Generation Models and the Utilization of a Spatial Context Variable." Dissertation. Knoxville: University of Tennessee.
Fader, Peter S. and Bruce G.S. Hardie (Nov. 1996). "Modeling Consumer Choice Among SKUs." In: *Journal of Marketing Research (JMR)* 33.4, pp. 442–452.
Fahrmeir, Ludwig et al. (2007). *Statistik*. Springer.
Fan, Jianqing, Fang Han, and Han Liu (June 1, 2014). "Challenges of Big Data Analysis." In: *National Science Review* 1.2, pp. 293–314. DOI: 10.1093/nsr/nwt032.
Fawcett, T. (2006). "An Introduction to ROC Analysis." In: *Pattern Recognition Letters* 27.8, pp. 861–874.
Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth (Nov. 1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data." In: *Commun. ACM* 39.11, pp. 27–34. DOI: 10.1145/240455.240464.
Fei, Hongliang et al. (2013). "Heat Pump Detection from Coarse Grained Smart Meter Data with Positive and Unlabeled Learning." In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. New York: ACM, pp. 1330–1338. DOI: 10.1145/2487575.2488203.
Feinberg, Fred M., Barbara E. Kahn, and Leigh McAlister (1992). "Market Share Response When Consumers Seek Variety." In: *Journal of Marketing Research* 29.2, pp. 227–237. DOI: 10.2307/3172572.
Fernández-Delgado, Manuel et al. (2014). "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" In: *The Journal of Machine Learning Research* 15.1, pp. 3133–3181.

Figueiredo, V. et al. (2005). "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques." In: *IEEE Transactions on Power Systems* 20.2, pp. 596–602.

Fishbein, Martin and Icek Ajzen (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley Series in Social Psychology. Reading, Mass: Addison-Wesley Pub. Co. 578 pp.

Flath, Christoph et al. (Feb. 1, 2012). "Cluster Analysis of Smart Metering Data." In: *Business & Information Systems Engineering* 4.1, pp. 31–39. DOI: 10.1007/s12599-011-0201-5.

Forum, World Economic (2011). *Personal Data: The Emergence of a New Asset Class*.

Freund, Yoav and Robert E Schapire (Aug. 1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." In: *Journal of Computer and System Sciences* 55.1, pp. 119–139. DOI: 10.1006/jcss.1997.1504.

Friedman, Jerome H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." In: *The Annals of Statistics* 29.5, pp. 1189–1232.

Frigo, N. J., P. P. Iannone, and K. C. Reichmann (Aug. 2004). "A View of Fiber to the Home Economics." In: *IEEE Communications Magazine* 42.8, S16–S23. DOI: 10.1109/MCOM.2004.1321382.

Gebauer, Heiko, Hagen Worch, and Bernhard Truffer (2014). "Value Innovations in Electricity Utilities." In: *Framing Innovation in Public Service Sectors*. Ed. by Rolf Rønning, Bo Enquist, and Lars Fuglsang. Vol. 30. Routledge Studies in Innovation, Organization and Technology. Routledge, 85ff.

Gefen, David, Elena Karahanna, and Detmar W. Straub (2003). "Trust and TAM in Online Shopping: An Integrated Model." In: *MIS Quarterly* 27.1, pp. 51–90.

Gefen, David and Catherine M. Ridings (Sum. 2002). "Implementation Team Responsiveness and User Evaluation of Customer Relationship Management: A Quasi-Experimental Design Study of Social Exchange Theory." In: *Journal of Management Information Systems* 19.1, pp. 47–69.

George, Joey F (2004). "The Theory of Planned Behavior and Internet Purchasing." In: *Internet research* 14.3, pp. 198–212.

Gholami, Roya et al. (Aug. 26, 2016). "Information Systems Solutions for Environmental Sustainability: How Can We Do More?" In: *Journal of the Association for Information Systems* 17.8.

Gladisch, Andreas, Christoph Lange, and Ralph Leppla (2008). "Power Efficiency of Optical versus Electronic Access Networks." In: *Proceedings of the European Conference on Optical Communication*. European Conference on Optical Communication. Brussels, Belgium.

Goebel, Christoph et al. (Feb. 1, 2014). "Energy Informatics." In: *Business & Information Systems Engineering* 6.1, pp. 25–31. DOI: 10.1007/s12599-013-0304-2.

Goodchild, Michael F (2007). "Citizens as Sensors: The World of Volunteered Geography." In: *GeoJournal* 69.4, pp. 211–221.

Goodhue, D.L. et al. (1992). "Strategic Data Planning: Lessons From the Field." In: *MIS Quarterly* 16.1, pp. 11–34.

Gorodkin, J. (2004). "Comparing Two K-Category Assignments by a K-Category Correlation Coefficient." In: *Computational biology and chemistry* 28.5, pp. 367–374.

Graml, Tobias et al. (2011). "Improving Residential Energy Consumption at Large Using Persuasive Systems." In: *ECIS 2011 Proceedings*. 19. European Conference on Information Systems (ECIS). Helsinki, Finland: AIS electronic library.

Granados, Nelson et al. (May 1, 2012). "À La Carte Pricing and Price Elasticity of Demand in Air Travel." In: *Decision Support Systems*. Information Issues in Supply Chain and in Service System Design 53.2, pp. 381–394. DOI: 10.1016/j.dss.2012.01.009.

Green, P. E. (Sept. 2004). "Fiber to the Home: The next Big Broadband Thing." In: *IEEE Communications Magazine* 42.9, pp. 100–106. DOI: 10.1109/MCOM.2004.1336726.

Greveler, Ulrich, Benjamin Justus, and Dennis Loehr (2012). "Multimedia Content Identification through Smart Meter Power Usage Profiles." In: *Computers, Privacy and Data Protection* 1, p. 10.

Groves, Robert M. (2006). "Nonresponse Rates and Nonresponse Bias in Household Surveys." In: *The Public Opinion Quarterly* 70.5, pp. 646–675.

Günther, Wendy Arianne et al. (Sept. 1, 2017). "Debating Big Data: A Literature Review on Realizing Value from Big Data." In: *The Journal of Strategic Information Systems* 26.3, pp. 191–209. DOI: 10.1016/j.jsis.2017.07.003.

Guyon, Isabelle, André, and Elisseeff (2003). "An Introduction to Variable and Feature Selection." In: *Journal of Machine Learning Research* 3, pp. 1157–1182.

Guyon, Isabelle, K. Bennett, et al. (July 2015). "Design of the 2015 ChaLearn AutoML Challenge." In: *2015 International Joint Conference on Neural Networks (IJCNN)*. 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. DOI: 10.1109/IJCNN.2015.7280767.

Guyon, Isabelle and André Elisseeff (Jan. 1, 2006). "An Introduction to Feature Extraction." In: *Feature Extraction*. Ed. by Isabelle Guyon et al. Vol. 207. Studies in Fuzziness and Soft Computing. Berlin, Heidelberg: Springer, pp. 1–25.

Hall, Mark A. (Apr. 1999). "Correlation-Based Feature Selection for Machine Learning." Hamilton, New Zealand: The University of Waikato.

Han, Jiawei, Micheline Kamber, and Jian Pei (2012). *Data Mining: Concepts and Techniques*. 3. The Morgan Kaufmann Series in Data Management Systems. Amsterdam: Elsevier.

Hanczar, Blaise et al. (Mar. 15, 2010). "Small-Sample Precision of ROC-Related Estimates." In: *Bioinformatics* 26.6, pp. 822–830. DOI: 10.1093/bioinformatics/btq037.

Hand, David J. (June 16, 2009). "Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve." In: *Machine Learning* 77.1, pp. 103–123. DOI: 10.1007/s10994-009-5119-5.

Hart, George Wiliam (1992). "Nonintrusive Appliance Load Monitoring." In: *Proceedings of the IEEE* 80.12, pp. 1870–1891. DOI: 10.1109/5.192069.

Hartmann, Philipp Max et al. (Oct. 3, 2016). "Capturing Value from Big Data – a Taxonomy of Data-Driven Business Models Used by Start-up Firms." In: *International Journal of Operations & Production Management* 36.10, pp. 1382–1406. DOI: 10.1108/IJOPM-02-2014-0098.

Hashem, Ibrahim Abaker Targio et al. (Oct. 1, 2016). "The Role of Big Data in Smart City." In: *International Journal of Information Management* 36.5, pp. 748–758. DOI: 10.1016/j.ijinfomgt.2016.05.002.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman (2013). *The Elements of Statistical Learning*. 2. ed., corrected at 7. print. Springer Series in Statistics. New York: Springer. 745 pp.

Haury, Anne-Claire, Pierre Gestraud, and Jean-Philippe Vert (Dec. 21, 2011). "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures." In: *PLOS ONE* 6.12, e28210. DOI: 10.1371/journal.pone.0028210.

Heiple, Shem and David J. Sailor (Jan. 1, 2008). "Using Building Energy Simulation and Geospatial Modeling Techniques to Determine High Resolution Building Sector Energy Consumption Profiles." In: *Energy and Buildings* 40.8, pp. 1426–1436. DOI: 10.1016/j.enbuild.2008.01.005.

Henke, Nicolaus et al. (Dec. 2016). *The Age of Analytics: Competing in a Data-Driven World*. McKinsey Global Institute, p. 136.

Hernández, Luis et al. (2012). "A Study of the Relationship between Weather Variables and Electric Power Demand inside a Smart Grid/Smart World Framework." In: *Sensors* 12.12, pp. 11571–11591. DOI: 10.3390/s120911571.

Herold, Martin, Joseph Scepan, and Keith C. Clarke (2002). "The Use of Remote Sensing and Landscape Metrics to Describe Structures and Changes in Urban Land Uses." In: *Environment and Planning A* 34.8, pp. 1443–1458. DOI: 10.1068/a3496.

Hess, Ann, Hari Iyer, and William Malm (Oct. 1, 2001). "Linear Trend Analysis: A Comparison of Methods." In: *Atmospheric Environment*. Visibility, Aerosol and Atmospheric Optics 35.30, pp. 5211–5222. DOI: 10.1016/S1352-2310(01)00342-9.

Holte, Robert C. (1993). "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets." In: *Machine Learning* 11.1, pp. 63–90.

Hopf, Konstantin (2018). "Mining Volunteered Geographic Information for Predictive Energy Data Analytics." In: *Energy Informatics* (1:4). DOI: 10.1186/s42162-018-0009-3.

Hopf, Konstantin, Florian Dageförde, and Diedrich Wolter (2015). "Identifying the Geographical Scope of Prohibition Signs." In: *COSIT 2015*. COSIT. Vol. 9368. Lecture Notes in Computer Science. Santa Fe, NM: USA: Springer, pp. 247–267. DOI: 10.1007/978-3-319-23374-112.

Hopf, Konstantin, Michael Kormann, et al. (2017). "A Decision Support System for Photovoltaic Potential Estimation." In: *Proceedings of the 1st International Confer-*

ence on Internet of Things and Machine Learning. IML '17. New York, NY, USA: ACM, 3:1–3:10. DOI: 10.1145/3109761.3109764.

Hopf, Konstantin, Sascha Riechel, et al. (2017). "Predictive Customer Data Analytics – The Value of Public Statistical Data and the Geographic Model Transferability." In: *ICIS 2017 Proceedings*. 38. International Conference on Information Systems (ICIS). Seoul, South Korea: AIS electronic library.

Hopf, Konstantin, Mariya Sodenkamp, and Ilya Kozlovskiy (2016). "Energy Data Analytics for Improved Residential Service Quality and Energy Efficiency." In: *ECIS 2016 Research in Progress Proceedings*. 24. European Conference on Information Systems (ECIS). Istanbul, Turkey: AIS electronic library.

Hopf, Konstantin, Mariya Sodenkamp, Ilya Kozlovskiy, and Thorsten Staake (2014). "Feature Extraction and Filtering for Household Classification Based on Smart Electricity Meter Data." In: *Computer Science-Research and Development*. D-A-CH Energieinformatik, Zürich. Vol. (31) 3. Zürich: Springer, pp. 141–148. DOI: 10.1007/s00450-014-0294-4.

Hopf, Konstantin, Mariya Sodenkamp, and Thorsten Staake (2018). "Enhancing Energy Efficiency in the Residential Sector with Smart Meter Data Analytics." In: *Electronic Markets* 28.4. DOI: 10.1007/s12525-018-0290-9.

Hothorn, Torsten et al. (Jan. 7, 2006). "Survival Ensembles." In: *Biostatistics* 7.3, pp. 355–373. DOI: 10.1093/biostatistics/kxj011.

Hsieh, J. J. Po-An et al. (Dec. 1, 2012). "Impact of User Satisfaction with Mandated CRM Use on Employee Service Quality." In: *MIS Quarterly* 36.4, pp. 1065–1080.

Hua, Jianping, Waibhav D. Tembe, and Edward R. Dougherty (Mar. 2009). "Performance of Feature-Selection Methods in the Classification of High-Dimension Data." In: *Pattern Recognition* 42.3, pp. 409–424. DOI: 10.1016/j.patcog.2008.08.001.

Huang, Dongling and Lan Luo (May 2016). "Consumer Preference Elicitation of Complex Products Using Fuzzy Support Vector Machine Active Learning." In: *Marketing Science* 35.3, pp. 445–464. DOI: 10.1287/mksc.2015.0946.

Hunt, Earl B, Janet Marin, and Philip J Stone (1966). *Experiments in Induction.* Oxford: Academic Press.

Immonen, Anne, Marko Palviainen, and Eila Ovaska (2014). "Towards Open Data Based Business: Survey on Usage of Open Data in Digital Services." In: *International Journal of Research in Business and Technology* 4.1, pp. 286–295.

Inc., Delighted (2017). *NPS Benchmarks: Compare Your NPS by Industry*. URL: https://delighted.com/nps-benchmarks (visited on 01/02/2018).

International, Open Knowledge (2018). *The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge*. URL: http://opendefinition.org/ (visited on 12/02/2018).

Irwin, Elena G. and Nancy E. Bockstael (Dec. 26, 2007). "The Evolution of Urban Sprawl: Evidence of Spatial Heterogeneity and Increasing Land Fragmentation." In: *Proceedings of the National Academy of Sciences* 104.52, pp. 20672–20677. DOI: 10.1073/pnas.0705527105.

Jagabathula, Srikanth and Gustavo Vulcano (Apr. 24, 2017). "A Partial-Order-Based Model to Estimate Individual Preferences Using Panel Data." In: *Management Science*. DOI: 10.1287/mnsc.2016.2683.

Jankel, Nick (Mar. 31, 2017). *Management Theory Is Dead. Here's Why.* URL: https://www.huffingtonpost.com/entry/management-theory-is-dead-heres-why_us_58de21a4e4b0efcf4c66a7bc (visited on 01/10/2019).

Jap, Sandy D. and Shankar Ganesan (2000). "Control Mechanisms and the Relationship Life Cycle: Implications for Safeguarding Specific Investments and Developing Commitment." In: *Journal of Marketing Research* 37.2, pp. 227–245.

Jiang, Zhenhui et al. (Jan. 2010). "Effects of Interactivity on Website Involvement and Purchase Intention." In: *Journal of the Association for Information Systems* 11.1, pp. 34–59.

Jurman, Giuseppe, Samantha Riccadonna, and Cesare Furlanello (Aug. 8, 2012). "A Comparison of MCC and CEN Error Measures in Multi-Class Prediction." In: *PLoS ONE* 7.8. DOI: 10.1371/journal.pone.0041882.

Juster, F. Thomas (1966). "Consumer Buying Intentions and Purchase Probability: An Experiment in Survey Design." In: *Journal of the American Statistical Association* 61.315, pp. 658–696. DOI: 10.2307/2282779.

Kahn, Barbara E. (July 1, 1995). "Consumer Variety-Seeking among Goods and Services: An Integrative Review." In: *Journal of Retailing and Consumer Services* 2.3, pp. 139–148. DOI: 10.1016/0969-6989(95)00038-0.

Kaiser, Florian G., Britta Oerke, and Franz X. Bogner (Sept. 2007). "Behavior-Based Environmental Attitude: Development of an Instrument for Adolescents." In: *Journal of Environmental Psychology* 27.3, pp. 242–251. DOI: 10.1016/j.jenvp.2007.06.004.

Kallinikos, Jannis and Ioanna D. Constantiou (Mar. 2015). "Big Data Revisited: A Rejoinder." In: *Journal of Information Technology (Palgrave Macmillan)* 30.1, pp. 70–74. DOI: 10.1057/jit.2014.36.

Kalousis, Alexandros, Julien Prados, and Melanie Hilario (May 2007). "Stability of Feature Selection Algorithms: A Study on High-Dimensional Spaces." In: *Knowledge and Information Systems* 12.1, pp. 95–116. DOI: 10.1007/s10115-006-0040-8.

Kamakura, Wagner A. (2008). "Cross-Selling: Offering the Right Product to the Right Customer at the Right Time." In: *Journal of Relationship Marketing* 6.3-4, pp. 41–58.

Karimi, Jahangir, Toni M. Somers, and Yash P. Gupta (Spr. 2001). "Impact of Information Technology Management Practices on Customer Service." In: *Journal of Management Information Systems* 17.4, pp. 125–158.

Karlin, Beth and Rebecca Ford (July 21, 2013). "The Usability Perception Scale (UPscale): A Measure for Evaluating Feedback Displays." In: *Design, User Experience, and Usability. Design Philosophy, Methods, and Tools*. International Conference of Design, User Experience, and Usability. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 312–321. DOI: 10.1007/978-3-642-39229-0_34.

Kavousian, Amir, Ram Rajagopal, and Martin Fischer (2013). "Determinants of Residential Electricity Consumption: Using Smart Meter Data to Examine the Effect of Climate, Building Characteristics, Appliance Stock, and Occupants' Behavior." In: *Energy* 55, pp. 184–194.

Kemmler, Andreas et al. (Sept. 30, 2015). *Analyse des schweizerischen Energieverbrauchs 2000 - 2013 nach Verwendungszwecken*. Bern, Switzerland: Bundesamt für Energie, p. 77.

Keogh, Eamonn and Abdullah Mueen (2011). "Curse of Dimensionality." In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer, pp. 257–258. DOI: 10.1007/978-0-387-30164-8_192.

Ketter, Wolfgang et al. (Sept. 2018). "Information Systems for a Smart Electricity Grid: Emerging Challenges and Opportunities." In: *ACM Trans. Manage. Inf. Syst.* 9.3, 10:1–10:22. DOI: 10.1145/3230712.

Khan, Nawsher et al. (2018). "The 10 Vs, Issues and Challenges of Big Data." In: *Proceedings of the 2018 International Conference on Big Data and Education - ICBDE '18*. The 2018 International Conference. Honolulu, HI, USA: ACM Press, pp. 52–56. DOI: 10.1145/3206157.3206166.

Kim, H. et al. (Apr. 28, 2011). "Unsupervised Disaggregation of Low Frequency Power Measurements." In: *Proceedings of the 2011 SIAM International Conference on Data Mining*. 0 vols. Proceedings. Society for Industrial and Applied Mathematics, pp. 747–758.

Kim, Hee-Woong, Hock Chuan Chan, and Sumeet Gupta (Feb. 2007). "Value-Based Adoption of Mobile Internet: An Empirical Investigation." In: *Decision Support Systems*. Mobile Commerce: Strategies, Technologies, and ApplicationsDSS on M-Commerce 43.1, pp. 111–126. DOI: 10.1016/j.dss.2005.05.009.

Kim, Seung Hyun and Tridas Mukhopadhyay (Nov. 18, 2010). "Determining Optimal CRM Implementation Strategies." In: *Information Systems Research* 22.3, pp. 624–639. DOI: 10.1287/isre.1100.0309.

Kim, YongSeog et al. (Feb. 2005). "Customer Targeting: A Neural Network Approach Guided by Genetic Algorithms." In: *Management Science* 51.2, pp. 264–276. DOI: 10.1287/mnsc.1040.0296.

Kira, Kenji (1992). *New Approaches to Feature Selection, Instance-Based Learning and Constructive Induction*. v, 69 leaves, bound.

Kira, Kenji and Larry A. Rendell (1992). "A Practical Approach to Feature Selection." In: *Proceedings of the Ninth International Workshop on Machine Learning*. Aberdeen, Scotland, United Kingdom: Morgan Kaufmann, pp. 249–256.

Kitchens, Brent et al. (Apr. 3, 2018). "Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data." In: *Journal of Management Information Systems* 35.2, pp. 540–574. DOI: 10.1080/07421222.2018.1451957.

Kolko, Jed (2012). "Broadband and Local Growth." In: *Journal of Urban Economics* 71.1, pp. 100–113.

Kononenko, Igor (1995). "On Biases in Estimating Multi-Valued Attributes." In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Vol. 2. Montreal, Quebec, Canada, pp. 1034–1040.

Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas (2007). "Supervised Machine Learning: A Review of Classification Techniques." In: *Informatica* 31, pp. 249–268.

Koufaris, Marios (June 1, 2002). "Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior." In: *Information Systems Research* 13.2, pp. 205–223. DOI: 10.1287/isre.13.2.205.83.

Koutsoukis, Nikitas-Spiros and Gautam Mitra (2003). *Decision Modelling and Information Systems: The Information Value Chain*. Operations Research/Computer Science Interfaces Series. Boston: Kluwer Academic Publishers. 366 pp.

Kozlovskiy, Ilya et al. (2016). "Energy Informatics for Environmental, Economic and Social Sustainability: A Case of the Large-Scale Detection of Households with Old Heating Systems." In: *ECIS 2016 Proceedings*. 24. European Conference on Information Systems (ECIS). Istanbul, Turkey: AIS electronic library.

Kruppa, Jochen et al. (Oct. 1, 2013). "Consumer Credit Risk: Individual Probability Estimates Using Machine Learning." In: *Expert Systems with Applications* 40.13, pp. 5125–5131. DOI: 10.1016/j.eswa.2013.03.019.

Kudo, Mineichi and Jack Sklansky (1998). "Classifier-Independent Feature Selection for Two-Stage Feature Selection." In: *Advances in Pattern Recognition*. Ed. by Adnan Amin et al. Red. by G. Goos, J. Hartmanis, and J. van Leeuwen. Vol. 1451. Berlin, Heidelberg: Springer, pp. 548–554.

— (2000). "Comparison of Algorithms That Select Features for Pattern Classifiers." In: *Pattern Recognition* 33.1, pp. 25–41.

Kukar, M. et al. (May 1999). "Analysing and Improving the Diagnosis of Ischaemic Heart Disease with Machine Learning." In: *Artificial Intelligence in Medicine* 16.1, pp. 25–50.

Kumar, V. (Jan. 2018). "A Theory of Customer Valuation: Concepts, Metrics, Strategy, and Implementation." In: *Journal of Marketing* 82.1, pp. 1–19. DOI: 10.1509/jm.17.0208.

Kursa, Miron B. and Witold R. Rudnicki (2010). "Feature Selection with the Boruta Package." In: *Journal of Statistical Software* 36.11. DOI: 10.18637/jss.v036.i11.

Kwac, Jungsuk et al. (Oct. 2013). "Utility Customer Segmentation Based on Smart Meter Data: Empirical Study." In: *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference On*. Vancouver, BC, Canada, pp. 720–725. DOI: 10.1109/SmartGridComm.2013.6688044.

L'Heureux, Alexandra et al. (2017). "Machine Learning With Big Data: Challenges and Approaches." In: *IEEE Access* 5, pp. 7776–7797. DOI: 10.1109/ACCESS.2017.2696365.

Lal, ThomasNavin et al. (Jan. 1, 2006). "Embedded Methods." In: *Feature Extraction*. Ed. by Isabelle Guyon et al. Vol. 207. Studies in Fuzziness and Soft Computing. Springer, pp. 137–165.

Lange, Christoph et al. (Aug. 2015). "Analysis of the Energy Consumption in Telecom Operator Networks." In: *Photonic Network Communications* 30.1, pp. 17–28. DOI: 10.1007/s11107-015-0492-4.

Laudon, Kenneth C. and Jane Price Laudon (2010). *Management Information Systems: Managing the Digital Firm*. 11th ed., global ed. Upper Saddle River, N.J: Pearson. 653 pp.

LaValle, Steve et al. (Dec. 21, 2011). "Big Data, Analytics and the Path From Insights to Value." In: *MIT Sloan Management Review* 52.2.

Lazar, Jim et al. (Nov. 2016). *Revenue Regulation and Decoupling: A Guide to Theory and Application (Incl. Case Studies)*. Montpelier, Vermont, U.S.: The Regulatory Assistance Project (RAP).

Lee, Jae-Nam et al. (Mar. 1, 2003). "The Contribution of Commitment Value in Internet Commerce: An Empirical Investigation." In: *Journal of the Association for Information Systems* 4.1.

Li, Shibo, Baohong Sun, and Alan L Montgomery (Aug. 2011). "Cross-Selling the Right Product to the Right Customer at the Right Time." In: *Journal of Marketing Research (JMR)* 48.4, pp. 683–700. DOI: 10.1509/jmkr.48.4.683.

Liu, Hongju, Pradeep K. Chintagunta, and Ting Zhu (July 2010). "Complementarities and the Demand for Home Broadband Internet Services." In: *Marketing Science* 29.4, pp. 701–720. DOI: 10.1287/mksc.1090.0551.

Liu, Huan and Hiroshi Motoda, eds. (2008). *Computational Methods of Feature Selection*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Boca Raton: Chapman & Hall/CRC. 419 pp.

Liu, Huan and Rudy Setiono (1995). "Chi2: Feature Selection and Discretization of Numeric Attributes." In: *TAI '95 Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*. IEEE, p. 88.

Liu, Ying et al. (2010). "Multicriterion Market Segmentation: A New Model, Implementation, and Evaluation." In: *Marketing Science* 29.5, pp. 880–894.

Lo, Adeline et al. (Nov. 29, 2016). "Framework for Making Better Predictions by Directly Estimating Variables' Predictivity." In: *Proceedings of the National Academy of Sciences*, p. 201616647. DOI: 10.1073/pnas.1616647113.

Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real (Mar. 1, 2008). "AUC: A Misleading Measure of the Performance of Predictive Distribution Models." In: *Global Ecology and Biogeography* 17.2, pp. 145–151. DOI: 10.1111/j.1466-8238.2007.00358.x.

Loock, Claire-Michelle, Thorsten Staake, and Frédéric Thiesse (2013). "Motivating Energy-Efficient Behavior with Green IS: An Investigation of Goal Setting and the Role of Defaults." In: *MIS Quarterly* 37.4, pp. 1313–1332.

Loomis, John et al. (Mar. 1, 1995). "Testing Transferability of Recreation Demand Models Across Regions: A Study of Corps of Engineer Reservoirs." In: *Water Resources Research* 31.3, pp. 721–730. DOI: 10.1029/94WR02895.

Lossin, Felix (2016). "Customer Engagement for Utilities: Information Systems to Curb Residential Energy Consumption." Doctoral Thesis. ETH Zurich. DOI: 10.3929/ethz-a-010782581.

Lu, Hsi-Peng and Philip Yu-Jen Su (Aug. 14, 2009). "Factors Affecting Purchase Intention on Mobile Shopping Web Sites." In: *Internet Research* 19.4, pp. 442–458. DOI: 10.1108/10662240910981399.

Lycett, Mark (July 1, 2013). "'Datafication': Making Sense of (Big) Data in a Complex World." In: *European Journal of Information Systems* 22.4, pp. 381–386. DOI: 10.1057/ejis.2013.10.

Madsen, Anders Koed (May 1, 2015). "Between Technical Features and Analytic Capabilities: Charting a Relational Affordance Space for Digital Social Analytics." In: *Big Data & Society* 2.1, p. 2053951714568727. DOI: 10.1177/2053951714568727.

Markard, Jochen and Bernhard Truffer (June 2006). "Innovation Processes in Large Technical Systems: Market Liberalization as a Driver for Radical Change?" In: *Research Policy* 35.5, pp. 609–625. DOI: 10.1016/j.respol.2006.02.008.

Markus, M. Lynne (Sept. 1, 2017). "Datification, Organizational Strategy, and IS Research: What's the Score?" In: *The Journal of Strategic Information Systems* 26.3, pp. 233–241. DOI: 10.1016/j.jsis.2017.08.003.

Marr, Bernard (Jan. 23, 2017). *Really Big Data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud*. URL: https://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/ (visited on 12/01/2018).

Martens, David et al. (Dec. 2016). "Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics." In: *MIS Quarterly* 40.4, pp. 869–888.

McAfee, Andrew and Erik Brynjolfsson (Oct. 2012). "Big Data: The Management Revolution." In: *Harvard Business Review* 90.10, pp. 60–68.

McGarigal, Kevin, Sermin Tagil, and SamuelA. Cushman (Mar. 1, 2009). "Surface Metrics: An Alternative to Patch Metrics for the Quantification of Landscape Structure." In: *Landscape Ecology* 24.3, pp. 433–450. DOI: 10.1007/s10980-009-9327-y.

McLoughlin, Fintan, Aidan Duffy, and Michael Conlon (2012). "Characterising Domestic Electricity Consumption Patterns by Dwelling and Occupant Socio-Economic Variables." In: *Energy and Buildings* 48, pp. 240–248. DOI: 10.1016/j.enbuild.2012.01.037.

Medina Medina, Esunly (Feb. 9, 2016). "An Approach to Pervasive Monitoring in Dynamic Learning Contexts : Data Sensing, Communication Support and Awareness Provision." Dissertation. Barcelona, Spain: Universitat Politècnica de Catalunya.

Melville, Nigel P. (2010). "Information Systems Innovation for Environmental Sustainability." In: *MIS Quarterly* 34.1, pp. 1–21.

Meyer, David et al. (2014). *E1071: Misc Functions of the Department of Statistics (E1071)*. TU Wien. URL: http://CRAN.R-project.org/package=e1071.

Mitchell, Tom M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science. New York: McGraw-Hill. 414 pp.

Morrison, Donald G. (1979). "Purchase Intentions and Purchase Behavior." In: *Journal of Marketing* 43.2, pp. 65–74. DOI: 10.2307/1250742.

Morwitz, Vicki G., Joel H. Steckel, and Alok Gupta (July 2007). "When Do Purchase Intentions Predict Sales?" In: *International Journal of Forecasting* 23.3, pp. 347–364. DOI: 10.1016/j.ijforecast.2007.05.015.

Müller, Oliver et al. (July 1, 2016). "Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines." In: *European Journal of Information Systems* 25.4, pp. 289–302. DOI: 10.1057/ejis.2016.2.

Nair, Harikesh S. et al. (July 31, 2017). "Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation." In: *Marketing Science* 36.5, pp. 699–725. DOI: 10.1287/mksc.2017.1039.

Negash, Solomon (Feb. 15, 2004). "Business Intelligence." In: *Communications of the Association for Information Systems* 13.1. DOI: 10.17705/1CAIS.01315.

Newell, Sue and Marco Marabelli (Mar. 1, 2015). "Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification'." In: *The Journal of Strategic Information Systems* 24.1, pp. 3–14. DOI: 10.1016/j.jsis.2015.02.001.

NordREG (Jan. 13, 2017). *Nordic Market Report 2015 - NMR Dataset*. Helsinki, Finland: Energy Market Authority.

OECD (Jan. 12, 2017). *Key Issues for Digital Transformation in the G20*. Berlin: OECD.

Al-Otaibi, R. et al. (Apr. 2016). "Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data." In: *IEEE Transactions on Industrial Informatics* 12.2, pp. 645–654. DOI: 10.1109/TII.2016.2528819.

Otim, Samuel and Varun Grover (Dec. 1, 2006). "An Empirical Study on Web-Based Services and Customer Loyalty." In: *European Journal of Information Systems* 15.6, pp. 527–541. DOI: 10.1057/palgrave.ejis.3000652.

Padmanabhan, Balaji and Alexander Tuzhilin (Oct. 1, 2003). "On the Use of Optimization for Data Mining: Theoretical Interactions and eCRM Opportunities." In: *Management Science* 49.10, pp. 1327–1343. DOI: 10.1287/mnsc.49.10.1327.17310.

Power, Daniel J. (2002). *Decision Support Systems: Concepts and Resources for Managers*. Greenwood Publishing Group. 284 pp.

Press, Gil (Mar. 23, 2016). *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. URL: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/ (visited on 01/16/2019).

Quinlan, J. R. (Mar. 1, 1986). "Induction of Decision Trees." In: *Machine Learning* 1.1, pp. 81–106. DOI: 10.1007/BF00116251.

Quinn, Elias Leake (2009). "Privacy and the New Energy Infrastructure." In: *SSRN Electronic Journal* 2009.02. DOI: 10.2139/ssrn.1370731.

Reichheld, Frederick F. (2003). "The One Number You Need to Grow." In: *Harvard Business Review* 81.12, pp. 46–55.

Reid, Andrea and Miriam Catterall (July 2005). "Invisible Data Quality Issues in a CRM Implementation." In: *Journal of Database Marketing & Customer Strategy Management* 12.4, pp. 305–314. DOI: 10.1057/palgrave.dbm.3240267.

Reunanen, Juha (Mar. 2003). "Overfitting in Making Comparisons Between Variable Selection Methods." In: *Journal of Machine Learning Research* 3, pp. 1371–1382.

Robnik-Šikonja, Marko (Sept. 22, 2003). "Experiments with Cost-Sensitive Feature Evaluation." In: *Machine Learning: ECML 2003*. European Conference on Machine Learning. Berlin, Heidelberg: Springer, pp. 325–336. DOI: 10.1007/978-3-540-39857-8_30.

Robnik-Sikonja, Marko and Petr Savicky with contributions from John Adeyanju Alao (2016). *CORElearn: Classification, Regression and Feature Evaluation*. URL: https://CRAN.R-project.org/package=CORElearn.

Romanski, Piotr and Lars Kotthoff (2014). *FSelector: Selecting Attributes*. URL: http://CRAN.R-project.org/package=FSelector.

Russell, Stuart J. and Peter Norvig (1995). *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice-Hall. 1132 pp.

Saeys, Yvan, Thomas Abeel, and Yves van de Peer (Sept. 15, 2008). "Robust Feature Selection Using Ensemble Feature Selection Techniques." In: *Machine Learning and Knowledge Discovery in Databases*. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer, pp. 313–325. DOI: 10.1007/978-3-540-87481-2_21.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga (Jan. 10, 2007). "A Review of Feature Selection Techniques in Bioinformatics." In: *Bioinformatics* 23.19, pp. 2507–2517. DOI: 10.1093/bioinformatics/btm344.

Salmon, Kiernan et al. (Aug. 20, 2016). "The Iterative Design of a University Energy Dashboard." In: ACEEE Summer Study on Energy Efficiency in Buildings. Asilomar, CA.

Sánchez, I. B. et al. (2009). "Clients Segmentation According to Their Domestic Energy Consumption by the Use of Self-Organizing Maps." In: *Energy Market, 2009. EEM 2009. 6th International Conference on the European*, pp. 1–6.

Sawalha, Ziad and Tarek Sayed (Mar. 2006). "Transferability of Accident Prediction Models." In: *Safety Science* 44.3, pp. 209–219. DOI: 10.1016/j.ssci.2005.09.001.

Schirner, G. et al. (Jan. 2013). "The Future of Human-in-the-Loop Cyber-Physical Systems." In: *Computer* 46.1, pp. 36–45. DOI: 10.1109/MC.2013.31.

Schmitz, Christian, You-Cheong Lee, and Gary L. Lilien (May 2014). "Cross-Selling Performance in Complex Selling Contexts: An Examination of Supervisory- and Compensation-Based Controls." In: *Journal of Marketing* 78.3, pp. 1–19.

Sester, Monika et al. (2014). "Integrating and Generalising Volunteered Geographic Information." In: *Abstracting Geographic Information in a Data Rich World*. Ed. by Dirk Burghardt, Cécile Duchêne, and William Mackaness. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, pp. 119–155.

Shah, Denish and V. Kumar (Dec. 2012). "The Dark Side of Cross-Selling." In: *Harvard Business Review*.

Shah, Denish, V. Kumar, et al. (May 2012). "Unprofitable Cross-Buying: Evidence from Consumer and Business Markets." In: *Journal of Marketing* 76.3, pp. 78–95. DOI: 10.1509/jm.10.0445.

Sharma, Rajeev, Sunil Mithas, and Atreyi Kankanhalli (July 1, 2014). "Transforming Decision-Making Processes: A Research Agenda for Understanding the Impact of Business Analytics on Organisations." In: *European Journal of Information Systems* 23.4, pp. 433–441. DOI: 10.1057/ejis.2014.17.

Shearer, Colin (2000). "The CRISP-DM Model: The New Blueprint for Data Mining." In: *Journal of data warehousing* 5.4, pp. 13–22.

Shi, Donghui et al. (Oct. 2, 2017). "A Data-Mining Approach to Identification of Risk Factors in Safety Management Systems." In: *Journal of Management Information Systems* 34.4, pp. 1054–1081. DOI: 10.1080/07421222.2017.1394056.

Shmueli, Galit (Aug. 2010). "To Explain or to Predict?" In: *Statistical Science* 25.3, pp. 289–310. DOI: 10.1214/10-STS330.

Shmueli, Galit and Otto R. Koppius (Sept. 2011). "Predictive Analytics in Information Systems Research." In: *MIS Quarterly* 35.3, pp. 553–572.

Shollo, Arisa and Robert D. Galliers (July 2016). "Towards an Understanding of the Role of Business Intelligence Systems in Organisational Knowing." In: *Information Systems Journal* 26.4, pp. 339–367. DOI: 10.1111/isj.12071.

Shrivastava, Utkarsh and Wolfgang Jank (June 26, 2015). "A Data Driven Framework for Early Prediction of Customer Response to Promotions." In: *AMCIS 2015 Proceedings*. Americas Conference on Information Systems (AMCIS). Puerto Rico: AIS electronic library.

Sikder, Sujan et al. (Sept. 1, 2013). "Spatial Transferability of Travel Forecasting Models: A Review and Synthesis." In: *International Journal of Advances in Engineering Sciences and Applied Mathematics* 5.2-3, pp. 104–128. DOI: 10.1007/s12572-013-0090-6.

Silva, Daswin et al. (2011). "A Data Mining Framework for Electricity Consumption Analysis From Meter Data." In: *IEEE Transactions on Industrial Informatics* 7.3, pp. 399–407. DOI: 10.1109/TII.2011.2158844.

Sodenkamp, Mariya, Konstantin Hopf, Ilya Kozlovskiy, et al. (June 2, 2016). *Smart-Meter-Datenanalyse Für Automatisierte Energieberatungen ("Smart Grid Data Analytics")*. Final Report 291131. Bern, Switzerland: Bundesamt für Energie.

Sodenkamp, Mariya, Konstantin Hopf, and Thorsten Staake (2015). "Using Supervised Machine Learning to Explore Energy Consumption Data in Private Sector Housing." In: *Handbook of Research on Organizational Transformations through Big Data Analytics*. Ed. by Madjid Tavana and Kartikeya Puranam, p. 320.

Sodenkamp, Mariya, Ilya Kozlovskiy, et al. (2017). "Smart Meter Data Analytics for Enhanced Energy Efficiency in the Residential Sector." In: *Wirtschaftsinformatik 2017 Proceedings*. 13. International Conference on Wirtschaftsinformatik (WI2017). St. Gallen, Switzerland: AIS electronic library.

Sokolova, Marina and Guy Lapalme (2009). "A Systematic Analysis of Performance Measures for Classification Tasks." In: *Information Processing & Management* 45.4, pp. 427–437.

Statistical Office of the European Communities (2014). *Living Conditions in Europe: 2014 Edition*. Luxembourg: Publications Office of the European Union.

Stingl, Carlo, Konstantin Hopf, and Thorsten Staake (2018). "Explaining and Predicting Annual Electricity Demand of Enterprises – a Case Study from Switzerland." In: *Energy Informatics* (1(Suppl 1):50). DOI: 10.1186/s42162-018-0028-0.

Strobl, Carolin et al. (2008). "Conditional Variable Importance for Random Forests." In: *BMC Bioinformatics* 9.1, p. 307. DOI: 10.1186/1471-2105-9-307.

Sun, Baohong and Vicki G. Morwitz (2010). "Predicting Purchase Behavior from Stated Intentions: A Unified Model." In: *International Journal of Research in Marketing* 27.4, pp. 356–366.

Synnott, W. R. (Sept. 1978). "Total Customer Relationship." In: *MIS Quarterly* 2.3, pp. 15–24.

The Economist (May 6, 2017). "Data Is Giving Rise to a New Economy." In: *The Economist*.

Thibaut, John W. and Harold H. Kelley (1959). *The Social Psychology of Groups*. In collab. with University of California Libraries. New York : Wiley. 346 pp.

Thiess, Tiemo and Oliver Müller (2018). "Towards Design Principles for Data-Driven Decision Making – an Action Design Research Project in the Maritime Industry." In: *ECIS 2018 Proceedings*. 26. European Conference on Information Systems. Portsmouth, UK: AIS electronic library.

Tiefenbeck, Verena (May 22, 2017). "Bring Behaviour into the Digital Transformation." In: *Nature Energy* 2.6, p. 17085. DOI: 10.1038/nenergy.2017.85.

Tiefenbeck, Verena, Lorenz Goette, et al. (Nov. 28, 2016). "Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation." In: *Management Science*. DOI: 10.1287/mnsc.2016.2646.

Tiefenbeck, Verena, Anselma Wörner, et al. (Nov. 19, 2018). "Real-Time Feedback Promotes Energy Conservation in the Absence of Volunteer Selection Bias and Monetary Incentives." In: *Nature Energy*, p. 1. DOI: 10.1038/s41560-018-0282-1.

Turner, David, Michael Schroeck, and Rebecca Shockley (May 2013). *Analytics: The Real-World Use of Big Data in Financial Services*. GBE03555-USEN-01. IBM Institute for Business Value.

TÜV Rheinland Consulting GmbH (2016). *Mid-2016 Report on the Broadband Atlas Commissioned by the Federal Ministry of Transport and Digital Infrastructure*. Berlin: German Federal Ministry of Transport and Digital Infrastructure.

Tversky, A. and D. Kahneman (Sept. 27, 1974). "Judgment under Uncertainty: Heuristics and Biases." In: *Science* 185.4157, pp. 1124–1131. DOI: 10.1126/science.185.4157.1124.

Van Der Maaten, Laurens, Eric Postma, and Jaap van den Herik (2009). "Dimensionality Reduction: A Comparative Review." In: *Journal of Machine Learning Research* 10, pp. 66–71.

Vanreusel, Wouter, Dirk Maes, and Hans van Dyck (Feb. 1, 2007). "Transferability of Species Distribution Models: A Functional Habitat Approach for Two Regionally Threatened Butterflies." In: *Conservation Biology* 21.1, pp. 201–212. DOI: 10.1111/j.1523-1739.2006.00577.x.

Vapnik, Vladimir Naumovich and Vlamimir Vapnik (1998). *Statistical Learning Theory*. Vol. 1. Wiley New York.

Varadarajan, Rajan (Apr. 1, 2010). "Strategic Marketing and Marketing Strategy: Domain, Definition, Fundamental Issues and Foundational Premises." In: *Journal of the Academy of Marketing Science* 38.2, pp. 119–140. DOI: 10.1007/s11747-009-0176-7.

Vassileva, Iana et al. (2012). "The Impact of Consumers' Feedback Preferences on Domestic Electricity Consumption." In: *Applied Energy* 93, pp. 575–582.

Venkatesh, Viswanath and Ritu Agarwal (Mar. 1, 2006). "Turning Visitors into Customers: A Usability-Centric Perspective on Purchase Behavior in Electronic Channels." In: *Management Science* 52.3, pp. 367–382. DOI: 10.1287/mnsc.1050.0442.

Venkatesh, Viswanath, James Y.L. Thong, and Xin Xu (2012). "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology." In: *MIS Quarterly* 36.1, pp. 157–178.

Venter, Peet, Alex Wright, and Sally Dibb (Feb. 2015). "Performing Market Segmentation: A Performative Perspective." In: *Journal of Marketing Management* 31.1-2, pp. 62–83. DOI: 10.1080/0267257X.2014.980437.

Verma, Anoop et al. (Dec. 2015). "A Data-Driven Approach to Identify Households with Plug-in Electrical Vehicles (PEVs)." In: *Applied Energy* 160, pp. 71–79. DOI: 10.1016/j.apenergy.2015.09.013.

Vickery, Graham (2011). "Review of Recent Studies on PSI Re-Use and Related Market Developments." In: *Information Economics, Paris*.

Wachtel, Stephan and Thomas Otter (2013). "Successive Sample Selection and Its Relevance for Management Decisions." In: *Marketing Science* 32.1, pp. 170–185. DOI: 10.1287/mksc.1120.0754.

Walsh, Gianfranco and Sharon E. Beatty (Mar. 8, 2007). "Customer-Based Corporate Reputation of a Service Firm: Scale Development and Validation." In: *Journal of the Academy of Marketing Science* 35.1, pp. 127–143. DOI: 10.1007/s11747-007-0015-7.

Walsh, Gianfranco, Sharon E. Beatty, and Edward M.K. Shiu (Oct. 1, 2009). "The Customer-Based Corporate Reputation Scale: Replication and Short Form." In: *Journal of Business Research* 62.10, pp. 924–930. DOI: 10.1016/j.jbusres.2007.11.018.

Wang, Y. et al. (2018). "Deep Learning-Based Socio-Demographic Information Identification from Smart Meter Data." In: *IEEE Transactions on Smart Grid* PP.99, pp. 1–1. DOI: 10.1109/TSG.2018.2805723.

Ward, John, Christopher Hemingway, and Elizabeth Daniel (June 1, 2005). "A Framework for Addressing the Organisational Issues of Enterprise Systems Implementation." In: *The Journal of Strategic Information Systems*. Understanding the Contex-

tual Influences on Enterprise Systems (Part II) 14.2, pp. 97–119. DOI: 10.1016/j.jsis.2005.04.005.

Watson, Hugh J. (2014). "Tutorial: Big Data Analytics: Concepts, Technologies, and Applications." In: *Communications of the Association for Information Systems* 34.65.

Watson, Richard T., Marie-Claude Boudreau, and Adela J. Chen (2010). "Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community.(Essay)." In: *MIS Quarterly* 34.1, p. 23.

Watson, Richard T., Jeffrey Howells, and Marie-Claude Boudreau (2012). "Energy Informatics: Initial Thoughts on Data and Process Management." In: *Green Business Process Management*. Ed. by Jan vom Brocke, Stefan Seidel, and Jan Recker. Berlin, Heidelberg: Springer, pp. 147–159. DOI: 10.1007/978-3-642-27488-6_9.

Wattal, Sunil et al. (Nov. 3, 2011). "What's in a "Name"? Impact of Use of Customer Information in E-Mail Advertisements." In: *Information Systems Research* 23 (3-part-1), pp. 679–697. DOI: 10.1287/isre.1110.0384.

Weber, Robert Philip (1990). *Basic Content Analysis*. 2. ed. Quantitative Applications in the Social Sciences. Newbury Park u.a.: Sage. 96 S.

Weinstein, Art (2013). *Handbook of Market Segmentation: Strategic Targeting for Business and Technology Firms, Third Edition*. 3rd ed. Binghamton, UNITED STATES: Routledge.

Welch, B. L. (Jan. 1, 1947). "The Generalization of 'Student's' Problem When Several Different Population Varlances Are Involved." In: *Biometrika* 34.1-2, pp. 28–35. DOI: 10.1093/biomet/34.1-2.28.

Wells, William D. and George Gubar (1966). "Life Cycle Concept in Marketing Research." In: *Journal of Marketing Research* 3.4, pp. 355–363. DOI: 10.2307/3149851.

Wenger, Seth J. and Julian D. Olden (Apr. 1, 2012). "Assessing Transferability of Ecological Models: An Underappreciated Aspect of Statistical Validation." In: *Methods in Ecology and Evolution* 3.2, pp. 260–267. DOI: 10.1111/j.2041-210X.2011.00170.x.

Wing, Max Kuhn Contributions from Jed et al. (2015). *Caret: Classification and Regression Training*. URL: http://CRAN.R-project.org/package=caret.

Wisconsin, Public Service Commission of (2018). *Broadband Expansion Grants for Fiscal Year 2018 (Round 2)*. URL: https://psc.wi.gov/Pages/Programs/BroadbandGrants.aspx (visited on 02/24/2018).

Xu, Mark and John Walton (Sept. 1, 2005). "Gaining Customer Knowledge through Analytical CRM." In: *Industrial Management & Data Systems* 105.7, pp. 955–971. DOI: 10.1108/02635570510616139.

Yates, F. (1934). "Contingency Tables Involving Small Numbers and the $X2$ Test." In: *Supplement to the Journal of the Royal Statistical Society* 1.2, pp. 217–235. DOI: 10.2307/2983604.

Yoo, Youngjin (Mar. 1, 2015). "It Is Not about Size: A Further Thought on Big Data." In: *Journal of Information Technology* 30.1, pp. 63–65. DOI: 10.1057/jit.2014.30.

Yu, Lei and Huan Liu (2003). "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution." In: *ICML*. Vol. 3, pp. 856–863.

Zablah, Alex R. et al. (Mar. 7, 2012). "Performance Implications of CRM Technology Use: A Multilevel Field Study of Business Customers and Their Providers in the Telecommunications Industry." In: *Information Systems Research* 23.2, pp. 418–435. DOI: 10.1287/isre.1120.0419.

Zaki, Mohammed J. and Wagner Meira Jr. (May 2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Zeifman, Michael and Kurt Roth (2011). "Nonintrusive Appliance Load Monitoring: Review and Outlook." In: *IEEE Transactions on Consumer Electronics*, pp. 76–84. DOI: 10.1109/TCE.2011.5735484.

Zhang, Tongxiao Catherine, Ritu Agarwal, and Jr. Lucas Henry C. (Dec. 2011). "The Value of It-Enabled Retailer Learning: Personalized Product Recommendations and Customer Store Loyalty in Electronic Markets." In: *MIS Quarterly* 35.4, 859–A7.

Zielstra, Dennis and Alexander Zipf (2010). "A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany." In: *13th AGILE International Conference on Geographic Information Science*. Vol. 2010.

Glossary

Ambient data

Data that is incurred though business activities, but is not essential for the fulfillment of contracts. The data can be firm-internal (e.g., high-frequent transaction data, data on business processes, communication data) or public data (e.g., governmental statistics, user-generated online content). 2, 4, 6, 12

Cross-selling

The activity of offering additional products or services to existing customers is considered as *cross-selling*, when the items to be sold differ from those a customer has already purchased or has expressed an interest in buying previously (Schmitz et al. 2014). 98, 120

Cross-validation

Method to calculate classification performance metrics. The approach is explained in subsection 4.2.4. 58, 64, 78, 81, 89, 126

Curse of dimensionality

Refers to the fact that the performance of machine learning algorithms typically decreases with a large input dimension and. This makes it hard to find relevant information in high dimensional data; for details, see Keogh and Mueen (2011). 32, 45, 47, 58

Euclidean distance

A distance measure between two points quantifying the "straight line" distance. 49, 60, 61, 176

Feature extraction

The activity of (empirically) defining and the process of calculating features from raw input data in data analysis. The activity involves human

knowledge and helps to decrease the dimensionality of the input data, encodes human knowledge into models, reduces model complexity and helps to make models more explainable. See Guyon and Elisseeff (2006) for a general introduction and section 3.3 for an introduction in the context of this work. 9, 28, 32

Feature selection

Automatic selection of input variables for data analysis tasks, using algorithms or statistical criterion. See Guyon and Elisseeff (2006) for a general introduction and section 3.4 for an introduction in the context of this work. 10, 46

Hellinger distance

A distance measure to quantify the similarity between two probability distributions. 60

Imbalanced classes

One class in a dependent variable is larger than others. When the class imbalance is strong, many classification algorithms tend to omit the infrequent class. 55

Information systems

An ambiguous term, which on the one hand describes a research discipline investigating the practice and study of information systems, on the other describes technical systems. The latter one are defined by (K. C. Laudon and J. P. Laudon 2010) as "Interrelated components working together to collect, process, store, and disseminate information to support decision making, coordination, control, analysis, and visualization in organizations." Considering the emerging of Internet of things and smartphones, the definition should also also include personal information systems. 1, 12

Machine learning

Describes a collection of algorithms and methods that enable machines (i.e., computers) to acquire knowledge automatically. In *supervised* machine learning, data with labels is used to learn a model, in *unsupervised* machine learning, pattern from data is detected without having ground truth data. Different approaches for supervised machine learning algorithms are summarized in section 4.1. 46

Manhattan distance

A distance metric that is also known as taxicap or city-block distance, as it is calculated with the sum of the absolute differences of their cartesian coordinates, like a taxi driver who cannot drive the Euclidean distance between start and destination, but must navigate in city-blocks. 61

Open data

Data or knowledge that "can be freely used, modified, and shared by anyone for any purpose." (International 2018). 27, 42

Predictive analytics

Shmueli and Koppius (2011) define predictive analytics as "statistical models and other empirical methods that are aimed at creating empirical predictions (as opposed to predictions that follow from theory only), as well as methods for assessing the quality of those predictions in practice (i.e., predictive power)". 16, 142

Smart meter

An electronic device that records consumption of energy and communicates the information to the electricity supplier for monitoring and billing. 2, 4, 12, 20, 21, 32, 71, 107, 135

University
of Bamberg
Press

Digitization causes large amounts of data in organizations (e.g., transaction data from business processes, communication data, sensor data). Besides, a large number of data sources are emerging online and can be freely used. Firms are looking for ways to commercialize this increasing amount of data and research aims to better understand the data value creation process. The present dissertation answers five central research questions in this context and examines how machine learning (ML) can be used to create value from data, using case studies from energy retailing and energy efficiency. First, a systematic literature review gives an overview of firm internal and external data sources for potential analyses. Second, the importance of human cognition, theory, and expert knowledge in effective data preparation for ML is demonstrated. Third, current ML algorithms and variable selection methods are empirically compared using industry data sets. Implications for theory and practice are identified. Finally, the successful use of the information gained through ML is exemplified through case studies where increased energy efficiency, customer value, and service quality can demonstrate economic, environmental, and social value. Thus, this empirical work contributes to the so far rather conceptual discussion on value creation from big data in information systems research.

eISBN: 978-3-86309-669-4

9 783863 096694

www.uni-bamberg.de/ubp