Value creation from analytics with limited data: A case study on the retailing of durable consumer goods

Konstantin Hopf[1*], Andreas Weigert[1], Thorsten Staake[12]

1) University of Bamberg, Chair of Information Systems and Energy Efficient Systems, 96047 Bamberg, Germany

2) ETH Zurich, Department of Management, Technology and Economics, Weinbergstrasse 56/58, 8092 Zurich, Switzerland

**) correspondence: konstantin.hopf@uni-bamberg.de*

April 06, 2022

The Version of Record of this manuscript has been published and is available in the Journal of Decision Systems, April 7, 2022, https://www.tandfonline.com/10.1080/12460125.2022.2059172

Abstract

Companies are pinning high hopes on competitive advantages through data analytics. So far, value gains through analytics have been demonstrated for IT-heavy and data-rich business areas. Yet, research has paid little attention to value creation through data analytics in the plethora of companies with limited data (i.e., having transactions in the hundreds and attributes in the tens). Building on the literature of big data value creation and the resource-based view, we carried out an in-depth analytics case study with a retailer of renewable energy systems. Firms in this business area operate with expensive but few sales, so their available data are notoriously limited. Our findings demonstrate that data analytics capabilities and value creation mechanisms (democratize, contextualize, experiment with data, and execute data insights) are also effective in situations with limited data. Practice and research should therefore put not only emphasis on the volume and the variety of data but also on contextual factors related to managers (e.g., clear strategy, vision, leadership) and all employees (e.g., openness for agile working mode, data awareness).

Keywords: Data analytics, information systems (IS) value creation, resource theory, value creation mechanisms, machine learning (ML), retail, renewable energy systems (RES)

1 Introduction

Firms compete on analytics—what Davenport and Harris (2007) had forecasted in the last decade became reality for many companies. *Analytics* describes the combination of organizational processes and computational tools—like methods from statistics, machine learning (ML), and data mining—for the analysis of available business data with the goal of making better decisions through data-driven insights (Chen et al. 2015; Ghasemaghaei et al. 2018). The euphoria about the possibilities of analytics (Chiang et al. 2018) and their strategic relevance for companies (Constantiou and Kallinikos 2015) is contrasted by the fact that productivity gains from investments are visible in only a few business sectors. Firms that have large amounts of data available (Tambe 2014), who rely on massive information processing (Wu et al. 2019), and firms in the IT industry (Müller et al. 2018) demonstrated productivity gains from data analytics investments. For companies in other sectors, studies could not substantiate economic contributions from such investments (Tambe 2014; Müller et al. 2018; Wu et al. 2019).

Information systems (IS) research has already devoted substantial attention to value creation through data analytics (Günther et al. 2017). A common perspective to explain successful value creation through analytics follows the resource-based view (e.g., Mikalef et al. 2017; Ghasemaghaei et al. 2018; Grover et al. 2018). In this logic, firms build analytics resources (like data, human talent, and infrastructure) and capabilities (e.g., competency in data analytics) to achieve their value targets including revenue growth, cost savings, and service innovation (Elia et al. 2020). Firms that have built resources and capabilities must in a next step activate and examine value creation mechanisms (e.g., executing insights from descriptive or predictive analytics) to gain value from data. Value creation mechanisms are virtually the gear needed to create competitive advantages from resources and capabilities because business value does not result from the sole possession of resources and capabilities, rather these must be actively used and managed (Sirmon et al. 2007).

The investigation of analytics in organizations, however, is difficult, given that it is a complex and iterative rather than a straightforward process (Shearer 2000; Asamoah and Sharda 2019) and because "insights do not emerge automatically out of mechanically applying analytical tools to data. Rather, insights emerge out of an active process of engagement between analysts and business managers using the data and analytic tools to uncover new knowledge." (Sharma et al. 2014, p. 435) The process of successful value creation from data remains not fully understood, even if recent studies shed more light on the effective interplay between human resources and technical analytics tools in the data-driven decision-making process (Elgendy et al. 2021) and underline the cyclical nature of analytics value creation (Grønsund and Aanestad 2020; van den Broek et al. 2022).

Moreover, research on data value creation (capabilities and mechanisms), so far, dominantly focuses on *big data*. Thereby, the predominant field assumption in business and IS research is—unlike other pronouncements (Constantiou and Kallinikos 2015; Yoo 2015)—that large volumes of data are available in firms. In practice, high volumes of data are often not (yet) available: One reason is that relevant events in business processes occur infrequently (e.g., rare defects in quality management). Another reason is that many firms are just starting with the collection of ground truth data (through manual labeling, asking customers, etc.) and have not much data available in their early days of using analytics. Limited data *(i.e., entries in databases in the order of hundreds, with attributes in the order of tens)* are often present, for example, when companies enter a new market, introduce a new product, and operate in a limited geographic region or market niche. Limited data are also often typical for analyses that focus on several subgroups or data segments. Thus, limited data occur in organizations of all sizes. Such situations are often challenging and have led to the empirical finding that the limited amount of data is a significant barrier to value creation (Brodley et al. 2012; Baier et al. 2019; Someh et al. 2020; Wilson and Daugherty 2020). Therefore, we see a need for research on analytics capabilities and data value creation mechanisms in cases where only limited data are available. Knowledge about analytics capabilities and value creation mechanisms would particularly help companies in non-IT and data-scarce business areas with their investments into analytics resources and capabilities.

To study this area, we selected a company for an in-depth case study that operates in a business field that is less data-rich and less IT-supported. Our study site is a vendor of renewable energy systems (RES). The company sells "big-ticket" durable consumer goods, like heating systems and photovoltaic installations. The situation that this business area has only limited data available mainly results from the fact that customers often make investments in consumer durables only once or twice in their life, given that the products are expensive and have a long lifespan. In addition, the market for such systems is fragmented and consists of many small vendors. Both lead to only a few sales transactions being available per company, which limits the data availability. Another reason is that the sales of such goods is a highly personalized process, which is mediated by sales agents who underwent long training. This makes the sales process—or parts of it—hard to automate. Less digitized business processes generate fewer data points for potential analytics. Moreover, the low number of individual sales transactions stands in contrast to a high number of product variants, so that even large companies with a relatively high number of customers have only a few sales transactions per product variant. Finally, vendors and installers of such goods commonly resemble handicraft enterprises, at least for the case of RES (Seel et al. 2014). All in all, these vendors find it difficult to use analytics on a large scale, as for example many online retailers (e.g., Amazon, Walmart, Alibaba) can do.

Vendors of durable goods have started to introduce online configurators on their websites that serve as easy-to-access sales assistants to attract new customers or to support the sales process (Abbasi et al. 2013; Sandrin 2017). These tools cannot replace professional consulting but can help to attract new potential customers and collect more information about them than before. In addition, amplified through the COVID-19 pandemic, the demand for remote sales consulting increased (Smith et al. 2020), and such sales tools help to meet this demand. Next to the objective to introduce such tools for attracting new customers, decision makers place such investments with the hope to gain value from data analytics.

Yet, research and practice lack knowledge on how firms that are not blessed with large amounts of data could also benefit from analytics. From the three broad areas of data analytics—descriptive, predictive, and prescriptive analytics (LaValle et al. 2011; Sivarajah et al. 2017), predictive analytics (with ML as the core technique) today seems to be the most powerful for firms in "successfully competing with business analytics" (Kraus et al. 2020, p. 628). Predictive analytics is the main principle of advanced data analytics (Donoho 2017; Kitchens et al. 2018). It makes use of "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, p. 554) As predictive analytics uses empirical data to create models, it particularly suffers from limited data. Therefore, we focus on this area of analytics and examine the following research question:

Which organizational capabilities and which mechanisms for creating business value can support the application of predictive ML methods in business areas with limited amounts of data?

Some existing studies investigate the application of data analytics methods to provide guidance in the sales process (Olson and Chae 2012; Martens et al. 2016), demand estimation (e.g., Shrivastava and Jank 2015; Loureiro et al. 2018), and lead generation as well as qualification (Cui et al. 2012). However, these studies focus on the sale of services or products that are sold with much higher quantity than durable consumer goods and the databases in their specific areas are reasonably large. Beyond that, the focus of these works does not lie on the capabilities and value creation mechanisms in organization, as our study does.

In the next section, our paper reviews the theoretical background on value creation from data (analytics) and formulates our research propositions. Section 3 describes our case study research approach. In Section 4, we present the empirical findings of applying ML in the selected case. Section 5 puts these findings in the theoretical context. We formulate the theoretical as well as practical implications of our study in Section 6 and close with a summary in the final section.

2 Theoretical background

Analytics—often referred to as big data analytics—has a strong link to the *"big data"* phenomenon[1], which emerged from the ongoing digitization of business processes and the proliferation of data. This section begins with a cursory review of how IS literature conceives the concept of big data and contrasts this with the fact that not all valuable data sources need to be large in volume. The following subsections review the literature on business value of IS investments and big data value creation. At the end of this section, we formulate two propositions that we test during our empirical study.

2.1 (Big) data in organizations

Much of research on data value creation in organizations focuses on big data. The concrete definition of the big data concept is, however, controversial. A common understanding of big data centers around the high volumes of data available in organizations (LaValle et al. 2011; Mikalef et al. 2017). Some definitions of big data refer to datasets with sizes "beyond the ability of common software tools to capture, curate, manage, and process the data within a specified elapsed time" (Bharadwaj et al. 2013, p. 476f). Other studies similarly focus on such upper-bound definitions regarding data volume (Kamioka and Tapanainen 2014; Rouhani et al. 2017) and also consider hardware limits (Kaisler et al. 2013; Fan et al. 2015), but such limits are constantly extended through technological improvements. These perceptions underline the relevance of appropriate infrastructure (i.e., computational techniques and computing power) to process the data, but they do help to not specify any measurable limit of storage and processing capacity. Looking at the lower bound of data that is necessary to perform analytics methods associated with big data analytics (e.g., ML, statistical analyses), there are some technical studies that explore the minimum number of observations for certain applications. They come to different conclusions, depending on the analysis technique chosen: Less than ten positive observations seem to be enough for error analyses in special industrial processes (Indira et al. 2010), less than a hundred observations for investigating the effect of certain medical applications (Motrenko et al. 2014), a few hundred observations for identification of customer characteristics from sensor data (Hopf et al. 2018), or a few thousand observations for the use of neural networks in discrete choice analysis (Alwosheel et al. 2018). Given this huge possible bandwidth of dataset sizes that can be subsumed under the concept of big data, several studies criticize the sole focus on the volume of data as it does not properly account for the multifacetedness of the big data phenomenon. The phenomenon, in fact, particularly includes internal, and external data (open and crowdsourced data), structured and unstructured, but also digital traces from sensors, which are data sources that were previously unknown to business analytics (Gillon et al. 2014; George et al. 2014). These works put more emphasis on other properties of big data—aside from volume—namely variety, velocity, and veracity (Lycett 2013; Constantiou and Kallinikos 2015; Yoo 2015). These characteristics suggest that even datasets with few observations that are barely suitable for applying data analytics methods might be a valuable resource for business value creation. Such datasets are abundant not only in companies from the IT sector, but also in non-IT businesses or are publicly available.

Albeit these accentuations, research on the value creation from data in organization still often focuses on high volume (e.g., Ghasemaghaei et al. 2018; Mikalef and Krogstie 2020; Elia et al. 2020). Motivated by this tension, our study relieves the field assumption that large volumes of data are necessary to realize value from data that is available to firms. In doing so, our study focuses on *limited data, by which we*

[1] On the one hand, the scientific discourse on *big data* is closely related to *big data analytics* (e.g., George et al. 2014; Woerner and Wixom 2015; Abbasi et al. 2016), sometimes the two concepts are even used interchangeably (McAfee and Brynjolfsson 2012; Constantiou and Kallinikos 2015). In any case, *analytics* is "a part of processing the data and one of the potential first steps in trying to realize value from big data" (Günther et al. 2017, p. 191). On the other hand, the discourse on the *use of analytics in business* needs to talk about big data as a core resource for which new, more efficient analytical methods are needed (Chen et al. 2012; Gillon et al. 2014).

mean data collections in organizations that are small in volume—i.e., entries (rows) in databases in the order of hundreds—and limited in variety—i.e., attributes (columns) in the order of tens. Yet, limited data are still sufficient to apply analytics methods and obtain meaningful results[2]. The other characteristics related to the concept of big data, like velocity and veracity (Lycett 2013; Constantiou and Kallinikos 2015; Yoo 2015) apply similarly to limited data. In addition, when firms start their analytics efforts, they usually need to collect ground truth data to train and validate statistical or ML models (Walczak 2001; Roh et al. 2021). As the collection of ground truth data is expensive and can take a long time, decision makers need to begin their analysis based on limited data to provide a basis for making an informed decision for or against further data collection.

2.2 Business value creation through IS investments

The relation between IS investments and business value is one of the major areas of investigation in IS research. Therefore, it has attracted substantial research attention in the past (Melville et al. 2004; Kohli and Grover 2008; Schryen 2013). Schryen (2013) defines IS business value as "the impact of investments in particular IS assets on the multidimensional performance and capabilities of economic entities at various levels, complemented by the ultimate meaning of performance in the economic environment." (p. 141) Following this definition, IS investments are characterized by making tangible (i.e., monetary contributions) as well as intangible contributions (e.g., new capabilities, knowledge) that can be observed from an internal and external perspective.

Many works (Melville et al. 2004; Schryen 2013) have used the resource-based view of the firm (Barney 1991; Bharadwaj 2000; Sirmon et al. 2007) as a theoretical frame to investigate IT value creation. In addition, studies have found dynamic capabilities (Teece et al. 1997), which build and alter capabilities in turbulent environments as important to sustain value creation from IT investments (Sambamurthy et al. 2003; Roberts et al. 2012). In the next section, we summarize this theoretical lens because recent research used it as foundation to conceptualize the process of big data value creation.

2.3 Resources and capabilities related to data analytics

A core concept of the resource-based view are capabilities, which describe high level routines (or collection of routines) in organizations. Capabilities can transform available resources into managerial decision options and the ability to produce significant outputs with a specific value proposition (Winter 2003). Bharadwaj (2000) found that the IT capability relies essentially on three types of resources for successful value creation, namely IT infrastructure, human IT resources, and IT-enabled intangibles. Studies on the data analytics capability find this triad of resources, as well (e.g., Gupta and George 2016; Fosso Wamba et al. 2017; Mikalef et al. 2017; Ghasemaghaei et al. 2018; Ghasemaghaei 2019; Mikalef and Krogstie 2020). But they underline that the resources necessary for analytics differ from classical IT resources (e.g., necessity of data, big data infrastructure, data-driven culture). There is evidence that the analytics capability—which Ghasemaghaei et al. (2018) describe as "analytics competency"—itself is a dynamic capability that is able to alter other capabilities (Ghasemaghaei et al. 2018; Ghasemaghaei 2019; Kristoffersen et al. 2021), or is capable to at least positively influence other dynamic capabilities like a firm's ambidexterity, agility, and innovative power (Mikalef et al. 2019; Božič and Dimovski 2019a; Rialti et al. 2019; Yasmin et al. 2020).

[2] Two conditions explain that limited data is sufficient for analytics in practice: (i) data should allow to be applied in appropriate algorithms (technical requirement) and that (ii) reasonable results can be expected from this (business requirement).

Even if the literature on data analytics resources and capabilities is fairly extensive, it seems to neglect the fact that limited data can be a valuable resource that companies can use to create competitive advantages. We therefore postulate the following proposition that we investigate in our fieldwork:

P1: Firms can form capabilities related to data analytics (i.e., analytics competency) already with limited data available.

2.4 Mechanisms for value creation through data analytics

Research that identifies resources and capabilities that are necessary for data value creation clarifies the antecedents of the value creation process (Galetsi et al. 2020). Yet, decisions on investments in this area are difficult to make because the value of resources is realized indirectly, not before they are selected and used purposefully (Makadok 2001; Sirmon et al. 2007). Therefore, further examination of the whole data value creation is necessary. Recent works do so by focus on identifying the mechanisms of value creation in organizations. These value creation mechanisms contribute to the attainment of value targets, which are usually defined strategically, in domain-specific business processes (Brinch 2018). Grover et al. (2018) develop a conceptual framework on data analytics value creation. They name twelve value creation mechanisms (transparency, access, discovery, experimentation, prediction, optimization, customization, targeting, learning, crowdsourcing, monitoring, and proactive adaption), but do not explain them in detail. Zeng and Glaister (2018) identify four mechanisms of data analytics value creation from internal data and two mechanisms of data analytics value creation from external sources, based on their empirical inquiry. We believe that value creation mechanisms offer research and practice a more palpable instrument to examine investment decision on data analytics projects than a sole focus on analytics resources and capabilities. Thus, we adopt the conceptualization of Zeng and Glaister (2018) on value creation mechanisms from firm-internal data and relat the mechanisms of Grover et al. (2018) to these mechanisms, as we describe below. The four internal value creation mechanisms are:

1. *Democratize data* describes "the capability to integrate data across the firm and enable a wider range of employees to access and understand data where it is needed at any given time." (Zeng and Glaister 2018, p. 120). Grover et al. (2018) describe this as the mechanism of "transparency and access" (p. 401). The ability to generate descriptive insights and disseminate them widely across an organization allows consistency in viewing the available data and facilitates a more comprehensive view on business processes and outcomes.
2. *Contextualize data* is the ability "to assign meaning as a way of interpreting the data within which an action is executed." (Zeng and Glaister 2018, p. 124) "[D]igging into data for both deep and pragmatic insights can yield important outcomes for various data analytics targets." (Grover et al. 2018, p. 401)
3. *Experiment with data* refers to the "capability to promote 'trial and error', cultivate an inquisitive attitude towards data, encourage continuous experimenting with the data and monitor the changes." (Zeng and Glaister 2018, p. 126) Grover et al. (2018) explain that "in an increasingly digital world, big data can involve many small experiments."
4. *Execute data insights* describes the "ability to transform data insights into actions that lead to … creating value." (Zeng and Glaister 2018, p. 127) Grover et al. (2018) name "prediction," "optimization," "customization," "targeting," (machine) "learning," and "proactive adaption" as separate value creation mechanisms, but we consider them as different alternative activities in the *execute data insights* mechanism.

Several studies provide evidence that the mechanisms provide business value from *big* data (see the special issue of Chiang et al. (2018) for a selection of studies), but research dominantly assumes that big data (especially in terms of high volumes of data) are available. Value creation mechanisms in firms in business sectors with limited data have not gained much attention. Thus, based on the four internal value creation mechanisms, we seek to investigate the following proposition with our empirical fieldwork:

P2: (a) Democratizing data, (b) Contextualizing data, (c) Experimenting with data, and (d) Executing data insights are internal value creation mechanism that can, each individually, help to achieve value targets even when only limited data is available in the company.

3 Research method

Starting from the theoretical basis, we examine data analytics value creation mechanisms in an organization that just had limited data available to do so. To investigate this revelatory case, we selected an in-depth case study approach (Yin 2018). This approach allowed us (i) to conduct exploratory research on a complex contemporary phenomenon—the process of value creation through data analytics (Sharma et al. 2014)—in its breadth and depth under real-world conditions (Yin 2018, p. 15) and draw "attention of researchers to interesting theoretical issues" (Narasimhan 2014). The approach also allowed us (ii) to identify boundaries of the big data concept and, to do so, (iii) to integrate information from multiple sources of evidence. Concretely, we examined how ML methods (a core technology of data analytics) can be used by a company that sells small quantities of durable consumer goods to private customers and, hence, has only limited amounts of data available. Our fieldwork was conducted between August 2018 and December 2020 and covered the complete process of designing, implementing and evaluating several ML models in a naturalistic setting (Bailey and Barley 2020). Earlier studies have demonstrated that the case study approach is useful for investigating, on the one hand, the resource theory in organizations (e.g., Lockett and Thompson 2001; Coates and McDermott 2002; Tarafdar and Gordon 2007) and, on the other hand, the understanding of the use of data analytics methods in firms (e.g., Lehrer et al. 2018; Grønsund and Aanestad 2020). Following a deductive approach (Barratt et al. 2011), the first two authors acted as quantitative analysts and carried out the data analytics and ML model development, given their backgrounds in ML and statistics. We involved a data analytics consulting firm to include industry expertise. Our deep involvement allowed us to gain detailed insights into the case. We ensured the necessary distance of a critical researcher by opting for a longitudinal research design and by analyzing the qualitative data after March 2020, when the data analytics project was completed. In addition, the third author was less involved in the fieldwork, ensuring the necessary distance from the field material. We describe our case selection strategy, our study design, the collection of qualitative data, and its analysis below.

3.1 Case selection and study design

For our project, we found a vendor of durable consumer goods in the area of RES, such as photovoltaic, battery storage, and heat pump installations who had started to invest in digital sales tools, but had not build significant data analytics competencies (i.e., no data scientists or analyst positions were created by that time). The company operated in Switzerland and was willing to evaluate the possibilities of ML-based data analytics to support their sales process. This project was set up as a research project from the very beginning. We also asked a consultancy company that was specialized on data analytics in the energy domain, with which the authors have been cooperating for several years, to support the effort and to ensure that our data analyses followed the state of practice.

We consider the selected case as a revelatory study (Yin 2018, p. 50) because we had the opportunity to observe how data analytics can be applied in an environment of limited data. This situation was—to the best of our knowledge—previously not accessible to IS inquiry. In addition, the early stage of a company experimenting with data analytics was a good research area to explore the range of potential capabilities and value creation mechanisms. The context of our investigation is a single data analytics project within one organization. It covered the computational models, but also related tools and organizational processes. Within this analytics project, context-specific new knowledge was generated by researchers closely collaborating with practitioners utilizing ML approaches. This knowledge qualifies as theory in

flux[3] (Tremblay et al. 2021), which is a novel type of practice-oriented theory that is built with ML and data analytics. The study we are reporting here focused on the *internal* value creation (Schryen 2013; Zeng and Glaister 2018) from this knowledge, rather than the development of this knowledge. The development of this theory-in-flux knowledge is therefore an embedded unit of analysis in our case study method, following Yin's (2018, p. 51f) suggestion of conducting embedded single-case studies in such a case.

3.2 Collection of case study research data

Our qualitative data covers multiple sources of evidence to allow data triangulation (Yin 2018, p. 126). To avoid researcher bias (Johnston et al. 1999; Barratt et al. 2011), we employed a longitudinal study design and collected data over a period of 17 months (08/2018 until 12/2020). In detail, we held seven semi-structured interviews, eleven focus group discussions during our fieldwork, and one focus group meeting ten months after project completion (see Appendix A for a detailed list of the data collection events). In addition, we considered all material that we gathered throughout the project (six email conversations with 38 emails in total, six flipchart notes from meetings, eight slide decks from presentations, six meeting notes, the dataset documentation, data analytics source code and models, and the dashboard web application).

3.3 Qualitative data analysis

We pursued a content analysis (Krippendorff 2018) to all qualitative data. With the support of the qualitative data analysis software MAXQDA, we coded empirical evidence for the existence of resources and capabilities related to analytics (see Section 2.3), the four value creation mechanisms (see Section 2.4), any barriers that hindered their functioning, and all business value contributions that were reported. We translated all quotes included in this paper from German to English.

4 Data analytics case study

In this section, we present our empirical case study material using a chronological structure. We describe the data analytics efforts that we carried out together with the consultancy company specialized in data analytics. The case study illustrates the various stakeholder expectations that became apparent during the project and which we addressed with the conducted data analyses. It also illustrates the activities necessary to prepare data—even of small volumes and variety—and to obtain results from ML analytics based on online configurator data.

First, we describe the work practice of the vendor's sales unit in the first phases of a typical personal selling process of renewable energy systems. Then, we present the steps of data acquisition (from the IS as well as the collection of ground truth data labels), data preparation, analysis, and the setup of ML models.

4.1 Initial situation and starting objectives of the project

The RES vendor applied a typical personal selling process (Brassington and Pettitt 2006), which starts with the prospecting and qualifying phase (see Figure 1). This process started with the generation of a list of *leads*[4], followed by an assessment of each lead's need and its motivation to purchase. Sources for

[3] Tremblay et al. (2021) define theories in flux "as evidence-based inferences that emerge from analyzing large amounts of data or big data, often gathered from business processes and in partnership with practitioners. … A [theory in flux] generally takes shape when a pattern of a phenomenon emerges from the analysis of data."

[4] A sales lead refers to a potential customer for whom address or contact details are known.

leads of the RES vendor were "classic" ones, like store visits, direct contacts via phone or email, recommendations by other clients, sales events, consumer fairs, but also "digital" leads that originated from online configurators (OCs) for their products. These OCs, one for photovoltaic and one for heat pump installations, are digital sales assistants that RES vendors can purchase as standard software products[5]. OCs can be branded to every RES vendor. They enable end-user services like automatically creating cost estimates, customized to the lead's local conditions. These tools are effective marketing measures to generate leads because they raise awareness of consumers to the brand, and they record the general interest in a product and contact information. Such OCs, however, still require professional support later in the consultation process. The vendor's value target was, thus, to prevent resources from being wasted by offers without a chance of win. To achieve this target, the initial project goal was to use the data from OCs to separate promising from unpromising leads. With the project, the vendor wanted to examine how data analytics can help to make this selection process more efficient.

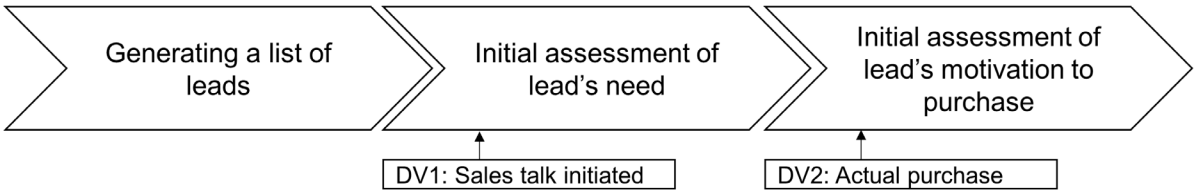

Figure 1: Prospecting and qualifying phase of the personal selling process (Brassington and Pettitt 2006) together with dependent variables that ML predicts and visualizes for sales agents

The processing of leads at the vendor started with a sales agent who received several leads ("classic" and "digital" ones), prioritized them (according to her/his individual judgement), and appointed an on-site sales visit. This visit—where an agent collects information on the potential installation and additional data on the building (e.g., roof, technical circumstances) to create an offer—and the preparation of an offer were the most resource-intensive steps in the sales process, according to multiple informants in our study. Informants estimated the effort of creating an individual offer and project plan with roughly one working day of experienced staff. Sales agents often don't have enough resources to work through all potential leads, which is aggravated by fluctuating seasonal demand, as one of the sales agents described: "Our first priority is to process the [classic] inquiries that are in the pipeline anyway … at the very end, we process [the digital] inquiries from the online calculator … we are not processing those at all at the moment. Yet, I am convinced that if we can channel those somehow, it will be attractive for us to approach these really high-potential customers." (#7) The caseload of the sales agents seemed to be seasonal, as one salesperson said in the beginning of the project that they were calling off all inquiries from the calculator (#5) but said at the end: "At the moment [early spring], we don't have the capacity to actively address these leads. There is simply not enough time." (#17) At the moment of our inquiry, thus, the revenue potential was, according to the informants, likely not fully exhausted.

Together with the RES vendor and the data analytics consultancy, we identified two companion pieces of information for this prospecting and qualification process (see Figure 1) that can support the sales agent's work. These two dependent variables (DV1: sales call initiated, DV2: actual purchase) should then be predicted from the data in the next step so that the sales rep can better assess the importance of the lead before launching costly and time-consuming activities.

4.2 Data collection and preparation

After having concretized the data analysis goal, we started to collect and prepare the datasets. On the one hand, data had to be exported from two OCs (one for photovoltaic, the other for heating systems)

[5] For example, EnergySage (https://www.energysage.com/) and NREL (https://pvwatts.nrel.gov/) for the U.S., or Eturnity (https://eturnity.ch/en/home-eng/) for the European market.

and other ISs. On the other hand, ground truth data had to be collected to apply ML methods. While describing the conducted ML analyses, we follow guidelines for documenting supervised ML studies in IS research (Kühl et al. 2021) and report adequate details on our data analysis steps. Finally, we describe the uses of project results at the vendor in different business sectors.

4.2.1 Preparation of OC data and feature extraction

At the time when we conducted the analyses, the OCs contained 5,038 entries with user inquiries that were made between 2016 and 2019 (roughly 1,260 entries per year, with a rising trend). Because of seasonal fluctuations in the sales agents' caseload, the number of entries was almost too large for serious manual investigation but at the same time small, when ML approaches should be applied.

The entered data contained general *building data* (e.g., location, living space) and *RES-specific data* such as the planned size of a photovoltaic system or dependencies to the existing heating infrastructure (e.g., the radiator type). Based on this input, the configurator suggests a pre-configured RES, calculates project metrics such as financial figures. The users and sales agents receive the results of this calculation via email. We derived 53 features[6] from the OC dataset, of which 24 were for heating and 29 for the photovoltaic application.

The features belong to seven categories (see Table 1): *Building characteristics* (e.g., building type, living space), *sociodemographic* (e.g., number of residents), *RES-specific data* (e.g., previous heating system, roof characteristics), *monetary figures* (e.g., the investment cost for a planned PV system and the cost changes when a battery system is added, whether subsidies can be received), *autarky figures* (e.g., share of self-consumption that would result from the planned project), *environmental figures* (e.g., estimated CO_2 savings of the planned project), and *consulting* (e.g., whether the lead desires a callback). The features result mainly from variables in the available datasets, but we also consulted research works on economic, sociodemographic, and psychographic factors (e.g., attitudes, literacy) that influence the adoption of RES (e.g., Heiskanen and Matschoss 2017). For example, we included the estimated CO_2 savings of the planned system as a feature in the model, because the perceived environmental benefit influences the adoption of PV systems (Korcaj et al. 2015; Wolske et al. 2017). A minority of entries (6.7%) in the heating system OC had missing values for some features. We imputed numerical features with the median. We centered and scaled all features. Categorical features are imputed with the mode and dummy encoded. There were no missing values in the photovoltaic OC data.

Table 1: Features[6] derived from heating and photovoltaic system online configurator

Feature category	Heating system OC	Photovoltaic OC
Building characteristics	• Region (first two digits of postal code) • Building type • Living space • Building construction year • Heights above sea level • Size of photovoltaic system (when existent)	• Region (first two digits of postal code) • Building type • Building orientation • Heating • Heating system type • Type of hot water production
Socio-demographic	• Number of residents	-
RES specific data	• Type of previous heating system • Age of previous heating system • Installed radiator type • Previous hot water production	• Existence of an old PV system • Annual electricity consumption in kWh • Roof orientation

[6] Features are predictor variables for the ML models.

	• Energy consumption old heating system in kWh • Thermal heat demand in kWh • Existence of a thermal solar system • Heat pump suggested as new heating system • Energy consumption for suggested heating system in kWh	• Roof type • Roof slope • Planned panel area • Planned number of panels • Planned size of PV system in kWp • Suggested size of a battery in kWh
Monetary figures	• Investment cost in CHF • Energy costs to date in CHF • Eligibility for subsidy program A-D (multiple programs possible)	• Energy produced in the first year in kWh • Investment cost in CHF* • Internal rate of return (IRR)* • Production cost per CHF/kWh* • Payback period in years*
Autarky figures	-	• Autonomy PV system* • Self-consumption*
Environmental figures	• CO_2 savings in kg / year	• CO_2 savings in kg / year*
Consulting	• Return call requested • Weekday when the online configurator was used	• Opt-in for being contacted

** Number each with and without battery storage*

4.2.2 Ground truth data collection and definition of dependent variables

Ground truth data, which contains labels for all entries in the dataset, is necessary for training and evaluation of ML models. We used data from the vendor's customer relationship management (CRM) system, which stored records from historical sales talks, to define *DV1 (sales talk initiated)*. We found 165 leads with a sales talk on a photovoltaic system and 208 sales talks for heat pumps (see Table 2). For *DV2 (actual purchase)*, the sales records from the vendor's CRM system contained only a few data points that we could match to leads from the OC. This is because the vendor introduced the OC only a few years ago and a significant share of customers still come from traditional marketing channels. To collect ground truth data and thereby remedy this situation, we conducted online surveys for the two OCs in June 2019. For these, we invited all users of the OCs that gave their consent to be contacted (4,253). In the surveys, we asked participants for their demographics, building characteristics (more detailed than available in the OC), details on the sales process they underwent, attitudes, and preferred characteristics regarding the respective RES (photovoltaic or heating system). We provide questionnaires in Appendix B. To collect data on the dependent variable, we asked if they had purchased a RES at some vendor (not necessarily our partnering company) after they have filled out the OC. In total, we could connect 560 survey answers to leads of the photovoltaic OC, where 289 have already purchased a system, and 238 records from the heating OC, where 66 have purchased heat pumps. We used the buyers as positive and the non-buyers as negative example in model training (see Table 2).

4.3 Predictive modelling and evaluation

Having collected and prepared the data, we developed predictive models for the two DVs, and the two products (heat pumps and photovoltaic) and evaluated those predictions. Knowledge about the quality of predictions is necessary to decide for their further use in business processes. The following efforts describe the analytics capabilities we have provided to the company as a means of examining data value creation mechanisms.

We tested logistic regression with a maximum likelihood estimator, Breiman's (2001) Random Forest (RF), as well as Vapnik and Vapnik's (1998) Support Vector Machine (SVM) as exemplary ML algorithms. The latter two algorithms were shown to provide good results across a variety of real world

prediction problems (Fernández-Delgado et al. 2014). In three cases, the dependent variables exhibited class imbalance: Both instances of DV1 (only 6-10% of examples are in the positive class) and the heat pump case of DV2. Class imbalance lowers the predictive power for rare events because typical cost parameters of ML models have a bias towards predicting the majority class (to increase overall model performance). We applied down-sampling (Kuhn and Johnson 2013) to reach an almost equal class distribution. This technique is preferable to other methods for overcoming the class imbalance problem (Kaur et al. 2019). Thereby, we randomly left out observations for the majority class during model training (for evaluation, we used the original distribution). For DV2 (photovoltaic), we used the original distribution.

Our main ML analysis, which we present below, used the standard hyperparameters of each algorithm[7] to get a (rather unbiased) first estimate of the model performances. We adhered to standard procedures to train and evaluate ML models with a tenfold cross-validation and applied the following well-known classifier performance metrics (Hastie et al. 2009):

$$Sensitivity = \frac{true\ positives}{positives}$$

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$F_\beta = \frac{(1+\beta^2) * precision * sensitivity}{(\beta^2 * precision) + sensitivity}$$

In addition, we used the Receiver Operating Characteristic (ROC) curve to evaluate the results. This curve is a two-dimensional figure with true positive and false positive rates on vertical and horizontal axes. The area under the ROC curve (AUC) is a performance metric derived from the ROC portion of the area of the unit square, and its value varies between 0 and 1. Random guessing produces a diagonal line between (0, 0) and (1, 1), which has an AUC of 0.5. Consequently, good prediction models are expected to achieve values clearly above 0.5 (Fawcett 2006). Table 2 shows the results of the classifier evaluation as mean values of a tenfold cross validation prediction, together with the standard deviation in brackets.

Table 2: Prediction performance results for lead scoring using three dependent variables

	Dependent variable: Sales talk initiated (DV1)					
	photovoltaic			heat pump		
Positives	165 (6%)			208 (10%)		
Negatives	2,780 (94%)			1,885 (90%)		
Sampling	down			down		
Method	RF	LR	SVM	RF	LR	SVM
AUC	0.68 (0.07)	0.64 (0.06)	0.72 (0.07)	0.56 (0.05)	0.56 (0.05)	0.55 (0.10)
Sensitivity	0.67 (0.10)	0.65 (0.09)	0.79 (0.09)	0.53 (0.09)	0.50 (0.08)	0.50 (0.09)
Precision	0.10 (0.02)	0.08 (0.01)	0.10 (0.01)	0.11 (0.02)	0.12 (0.02)	0.12 (0.03)
F2	0.31 (0.05)	0.27 (0.03)	0.32 (0.04)	0.31 (0.05)	0.30 (0.04)	0.30 (0.06)
	Dependent variable: Actual purchase (DV2)					
	photovoltaic			heat pump		
Positives	289 (52%)			66 (28%)		
Negatives	271 (48%)			172 (72%)		
Sampling	none			down		

[7] RF was used with *mtry* of 5 and 500 *trees*. SVM was used with a gaussian radial base *kernel*, *sigma* of 0.009 and *cost* of 1.

Method	RF	LR	SVM	RF	LR	SVM
AUC	0.77 (0.08)	0.75 (0.06)	0.78 (0.08)	0.67 (0.13)	0.63 (0.13)	0.62 (0.09)
Sensitivity	0.72 (0.09)	0.67 (0.10)	0.65 (0.10)	0.68 (0.18)	0.67 (0.22)	0.63 (0.16)
Precision	0.73 (0.08)	0.75 (0.06)	0.82 (0.07)	0.35 (0.08)	0.34 (0.09)	0.34 (0.07)
F2	0.72 (0.08)	0.69 (0.09)	0.67 (0.10)	0.57 (0.13)	0.55 (0.16)	0.53 (0.11)

For the RES vendor, it was considerably important to identify a large share of leads from the OC that convert to a sales talk or purchase of RES (true positives). For this prediction problem, we selected a model with a higher *Sensitivity* to not lose sales opportunities. As sensitive models can be prone to predict many false positives (i.e., leads that will not have a sales talk or buy but the model suggests that they will), an overhead effort for sales agents to prepare for sales meetings could be generated. To cope with this issue, we computed the F_2 measure, which gives *Sensitivity* a higher importance than *Precision*.

Considering all performance metrics for all dependent variables, RF and SVM lead to better predictions that logistic regression in most cases. In terms of the F_2 metric, RF shows slightly better results compared to LR or SVM. In in terms of AUC, SVM achieves slightly higher values in the case of photovoltaics and RF in the case of heat pumps. Thus, there seems to be no clearly superior approach for all variables. Among the photovoltaic OC users that also participated in the survey, the prediction models of DV2 can clearly separate buyers from non-buyers. All in all, the prediction performance was rated as sufficient for the use case.

In addition to this main ML analysis, we conducted two sensitivity analyses to corroborate the predictive performance (see Appendix C for detailed results). In the first analysis, we tested twelve variations of hyperparameters for the RF and SVM algorithm in a grid search. These experiments led only to changes in the prediction performance of few percentage points, the predictions of only some variables could be improved. The incremental improvements of parameter tuning, however, did not change the core findings of our study. We report this as a side-analysis because experimenting with hyperparameters raises the risk of model overfit, especially in the case of limited data. In the second sensitivity analysis, we explored if the predictive models could also have been built with fewer data than present. For that, we varied the amount of data for training and examined how the model performance behaves. In this analysis, we found that with 80% or less of the data available for this study, the model quality (in terms of AUC) was significantly lower, and the variance of the performance was significantly higher. Thus, the models became less reliable. Therefore, we cannot recommend building the presented models with less training data.

4.4 Use of data analytics results in the organization

Throughout the project, several use cases have emerged in which the results of the analytics project could potentially create value at the RES vendor. We present these below and describe the unique value created by each of them.

4.4.1 Improved prioritization of sales efforts through predictions

The ML predictions allowed to prioritize incoming leads according to their relevance regarding a potential sales talk (DV1) and the purchase likelihood (DV2). Both predictions could help to allocate scarce and expensive human resources to the leads with the highest potential. To illustrate this, we compiled a numerical example based on the predictions for DV1 and DV2 in the photovoltaics' case in which we compare the sales performance with and without lead prioritization using the developed prediction model[8]. We depict the example in in Figure 2.

[8] The numbers for the heat pump case with a selection of the top 20% are as follows: *Sensitivity* DV1: 25.5%, *Precision* DV2: 12.7%, *Sensitivity* DV2: 34.8%, *Precision* DV2: 48.9%.

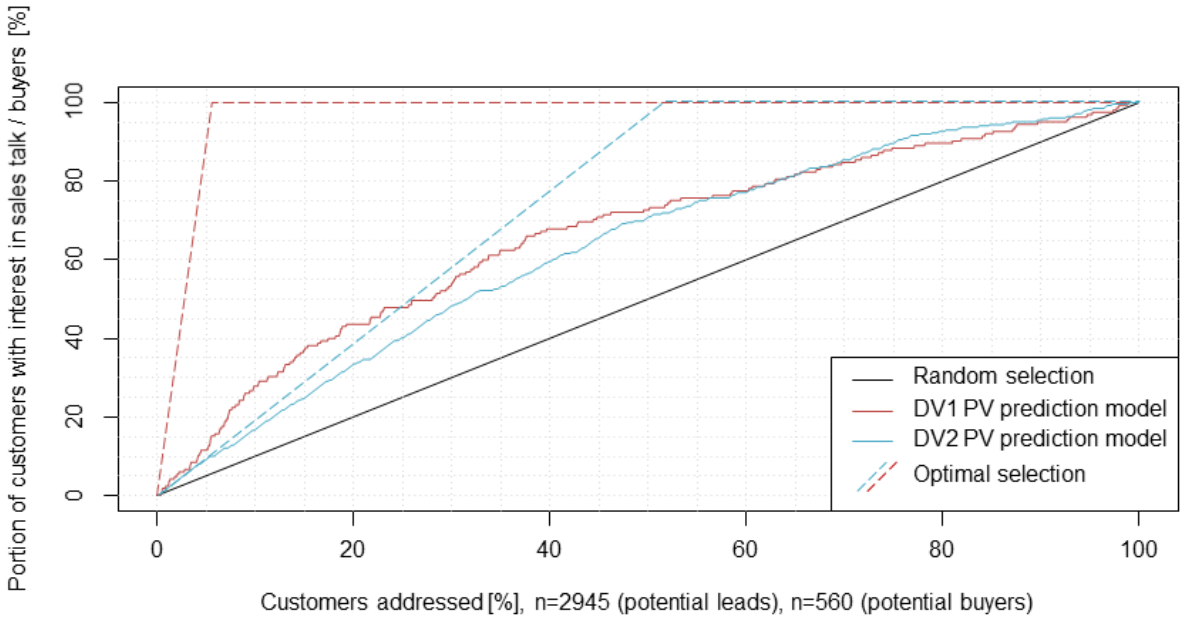

Figure 2: Impact of prediction models in the photovoltaics' case

When the resources of sales agents allow handling only 20% of the incoming leads in an initial sales talk (DV1), they reach on average 20% of the truly interested leads when incoming leads are processed randomly, which is a reasonable assumption for service queues (Rogiest et al. 2015). Based on our model, the prioritization could detect 43.6% of truly interested leads (*Sensitivity* of the model for the top 20%), which equals to a 118% higher chance to select promising leads. The rate of sales talks that would actually take place with the top-20%-selection is 12.2% (*Precision*), which is an acceptable number in a marketing setting. Going one step further, we consider DV2 (actual purchase of a photovoltaic system) and assume that the top 20% of the interested leads should be further processed. In this case, the prediction model identifies 33.6% of the actual buyers with a *Precision* of 86.6%. The *Sensitivity* of the model is 68% higher than in the absence of a selection. Moreover, it is close to a hypothetical optimal selection, where only 38.5% could be achieved with 20% of the customers processed (see dashed and solid lines in Figure 2). Both predictions allow sales agents to invest their time more effectively. This is not only beneficial for the vendor on an organizational level but also for the personal advantage of sales agents, as their task performance is often evaluated based on achieved orders.

4.4.2 Visualizing the drivers of customer behavior

During the presentation of the prediction models, we observed a great desire of the RES vendor to understand what is going on within the ML tools. The head of RES sales put it: “We need rules, because we are interested in the score and the relations between electric mobility and photovoltaics, between installed heating type and the probability of investing in a heat pump etc. … These are the most exciting conclusions for us, because there are also qualitative things that we try to pick up certain users in the next marketing campaign.” (#10) Likewise, a sales agent explained “It would be interesting to know how much [the leads] have already studied the topic, how much information they got, how much consulting they received. Do they already have other offers? … If someone has ten offers on the table, he is perhaps not very interesting for us.” (#17) To satisfy this demand, we used facilities of the applied ML models to identify influential variables for the predictions (e.g., estimates of logistic regression and feature importance scores from RF). We added another view in the dashboard that allowed to browse the most influential variables in each prediction model. The sales agents rated this additional view as helpful, because it was a novel source of information in addition to reports from already existing systems. However, this view has even a much higher relevance for the business development department that defines targeted customer profiles and marketing campaigns.

In addition, we observed that just showing the patterns in the data, i.e., influences of individual variables on the target variable, triggered discussions, and sales agents increased their understanding of customer decisions. There was, for example, an effect in the data that leads in more densely populated areas had a higher propensity to purchase heat pumps. We (in this case as the data analysts) could not explain this effect and one manager doubted the model. But sales and business development team members were able to explain such an effect and convinced the manager about this likely true causality (#13).

4.4.3 Integration of prediction in operational processes

After a longer discourse on how the scores could be integrated most effectively into the sales agents' work environment, it turned out that the most effective form would be a colored flag in the email system of the sales agents in which the computed report for every new OC lead arrives. The head of RES sales argued that a new user interface for sales agents is not necessary: "We actually need to make a triage as quickly as possible, as soon as a lead comes in, we need the ability … based on a lead scoring … where we actually say: 'Wow, this is a good, hot, dark red lead or this is more of a blue, pale gray guy that we leave aside.'" (#9) Likewise, a sales agent reported: "I see ten new leads and then I look through … the cover mail. Sometimes I look in and sometimes I just look at the mail address, that is all I see. If I have to open [the attachment with detailed information on the configuration] and close it again then it would take me too much time, … only if I am interested, I go into the depth, then I open the PDF" (#16). The scores particularly help to sort out leads without any purchase intention, as a sales agent explains: "there are still a lot of those who click to [check out RES installations of their neighbors]." (#17)

4.4.4 Presentation of predictions in a dashboard

Another aid for the business of sales agents was the visualization of scores on a map (see Figure 3). An agent said: "The geographic views of course help to plan routes more effectively. We do not always need that, but especially with an electric car it is important to drive the right routes. Also, in general … because driving is actually a waste of time, right? … This morning I was out [for an on-site visit], 45 minutes to drive, and it would have been interesting to see what I have in the region … for the way back … I would have liked to pick the most relevant one." (#16)

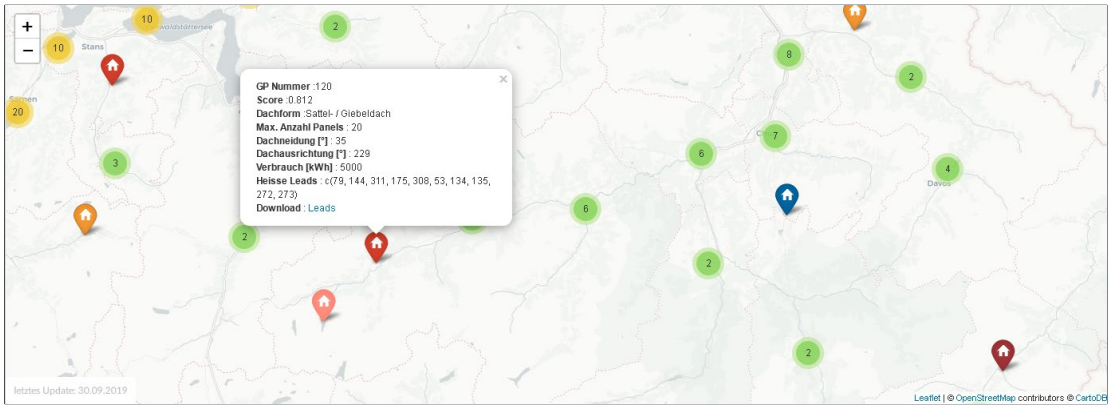

Figure 3: Interactive map that visualizes the lead scoring with a color scale (red is highest and blue is the lowest relevance) and provides a summary of the lead

5 Analysis of case findings

The presented case study documented a data analytics project, which we carried out in a company with limited available data that was at the beginning of their journey to build a data infrastructure and data analytics capabilities. The starting point for the initiative was that the company began to digitally gather data on their potential customers by using a sales tool to generate leads. The introduction of two OCs—one for heat pumps and another one for photovoltaic systems—were motivated by the desire to gain a

competitive advantage in lead generation. But the company struggled to achieve their value targets with their recent data infrastructure investments. Concretely, employees had problems to manage the incoming workload of leads trough this new channel with their current sales processes and to correctly prioritizing their sales efforts. Applying data analytics to solve this problem is, in general, not a new approach because many companies are currently trying to use data analysis to generate value from available data. Yet, given that ML applications are usually data hungry, companies often only start with data analysis projects when a significant amount of data is available. The dataset size of our case seems, at a first glance, not small (OCs contained 5,038 entries with user inquiries). When looking, however, at the positive examples, which range between 66 and 289 depending on the to-be-predicted variable, the dataset size was just sufficient to apply data analytics methods (Motrenko et al. 2014; Hopf et al. 2018; Alwosheel et al. 2018). Our detailed sensitivity analysis (see Appendix C) also demonstrated that 80%-90% of the data that was available was truly necessary to create stable models.

Another problem is that many companies are aware of the possibilities but do not have the necessary capabilities and resources to tackle such projects. In this study, we supplied know-how in data analytics and human resources in a research project and had the opportunity to explore in detail how a company can use data analytics at this early stage and with limited data. The project, for example, raised the data awareness among a broad group of employees, increased the understanding of possibilities and limits of data analysis, and finally helped employees with understanding and interpreting analytics results. In the end of the project, the company hired two data scientists to continue the capability building and realization process after our involvement. Thus, we *find support for our first proposition*.

The second focus of our inquiry was the function of value creation mechanisms in the field even in the situation of limited data. Figure 4 summarizes the data analytics efforts, which we carried out throughout the project together with the outcomes and value created which we mapped to the value creation mechanisms of Zeng and Glaister (2018) and the value targets articulated by the informants. The project started with available data from OCs and CRM systems. We collected ground truth data via two surveys, computed features from the input data, and carried out several data analyses with ML algorithms. We identified four outcomes (right part of Figure 4), which we consider as instances of the four value creation mechanisms from internal data. The first outcome (a scored list of leads), which was also the starting objective of the project, can improve the sales lead prioritization through a computed score for each incoming inquiry. We have demonstrated how this information can be used to better allocate personal resources and thus create value through more efficient operations. The second outcome (a visualization of the drivers of the customer behavior) appeared through setting up the analytics models, presenting interim results, and clarifying anomalies in the data. The discussion around the issues increased the data literacy in the organization and led to a sense of potential data problems and their prevention. Third, through presentation of results and discussing model details, the demand for explanations of the predictions became apparent (outcome: patterns and correlations). Employees wanted to “understand the rules” by which models ranked the importance of an incoming lead. When we presented more details for the models, employees started to relate the findings to their experience. They imagined possible implications of their work or questioned the models in the case of this finding was counterintuitive. These correlations also triggered intense discussions during the meetings between different perceptions of the participants, which led to an increased customer understanding.

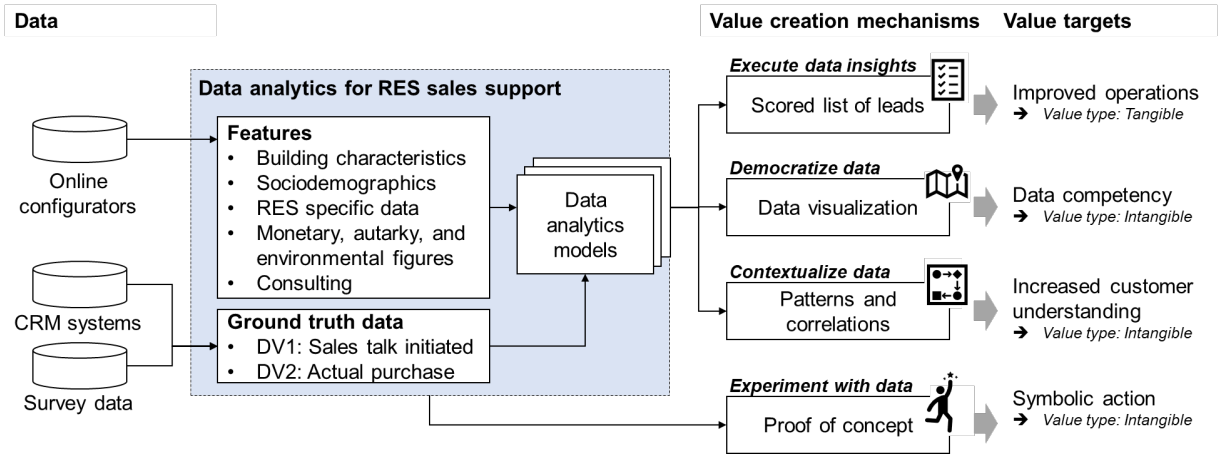

Figure 4: Outcomes of data analytics project and value contributions

Finally, we saw evidence that the project had a symbolic impact both internally and externally (outcome: proof of concept). On the one hand, the examination of datasets that existed in the company and the problems that arose during this process led to a greater understanding of data. This stimulated discussions on how to collect, process and store data in a tidier manner. In this sense, the project had a kind of internal beacon effect. On the other hand, we noticed that towards the end of the project the available resources for data analytics increased at the vendor. The company hired two data scientists and started to invest in a data management platform. Even though this observation is probably confounded with other developments and business needs, it does indicate an indirect symbolic effect of the project. The two data scientists were also convinced that "precisely with such projects, the [company] attracts young talents from university who are now engaged with these topics. So, it reaches a new audience for its [potential] employees." (#19) Hiring skilled personnel encounters the shortage of skills identified in the literature (Brynjolfsson et al. 2017).

The four outcomes that we identified from the project can be mapped to the four value creation mechanisms of Zeng and Glaister (2018), which we have introduced earlier even with limited data. *Thus, we find support for our second proposition.*

To sum up, Table 3 shows the four value creation mechanisms, the concrete outcome of the project, the type of value created (Schryen 2013; Fosso Wamba et al. 2015). We corroborated this conceptualization by our case study in which we tested to what extent value creation can also be achieved with limited data available. We ordered the value creation mechanisms according to their appearance in the literature, not according to their relevance in the case study. Informants agreed that the largest, yet intangible, value contribution of this project was to contextualize data (e.g., achieve a better understanding of customer behavior, knowledge on data that should be automatically collected in the future). In addition, we summarize barriers to each value creation mechanisms, which we identified during our fieldwork and which we derive from the collected qualitative data.

Table 3: Case findings on value creation mechanisms and identified barriers from our case study

Value creation mechanism	**Realizable business value through data analytics**	**Type of value**	**Barriers to value creation** *(as became apparent in our case)*
Democratize data	- Analytics uncover flaws in data and drive data literacy - Dashboard makes information available to different stakeholders (e.g., sales agents, marketeers, business developers)	Intangible	- Existing systems provide already access to much data - Data access may be restricted to selected roles (e.g., only some departments have access to systems)
Contextualize data	- Increase customer understanding in market niche (e.g., drivers of customers' purchase decisions) - Understand what data should be collected in the future	Intangible	- Model output not necessarily actionable for business (insights need to be transformed to prescriptions) - Contradictory findings and old convictions - Right balance between model performance and interpretability
Experiment with data	- Software (dashboard) guided decision making on data analytics investments - Continuous update of prediction models with new data	Intangible	- Definition of appropriate success criteria - Sufficient data and users available
Execute data insights	- Individual estimation of purchase probabilities (*prediction*) - Increase effectiveness of marketing efforts and allocate resources (*optimization*) - Integrate results in business processes (*use*): Present predictions with simple visualization (e.g., traffic light symbols), using them for targeting (e.g., marketing automation), and use detected pattern to define target groups	Tangible	- Integration of prediction in existing IT systems - Changes in business processes to act upon predictions - Real-time predictions (e.g., when new data comes in)

6 Discussion

Our finding that creation and execution of data analytics capabilities works even with limited data has several implications for theory and practice. Beyond this, our use case practically demonstrated how ML can be employed in the sales context to gain operational decision support and deeper business knowledge on customers.

6.1 Less emphasis on high volumes of data

The extensive literature on data value creation so far predominantly assumes the existence of large volumes and large variety of data (big data). This, at least implicit, field assumption allows the conclusion that companies without big data are unlikely to be successful in creating value from data. Questioning and investigating such a field assumption is not trivial (Alvesson and Sandberg 2011) but our revelatory case allowed unique insights into a company that was starting to invest in data analytics. Our study demonstrated that value creation can even happen with limited data, which means that firms can start with datasets on transactions in the hundreds and attributes in the tens, but still sufficient to apply analytics methods and obtain meaningful results. Although research acknowledges that data value creation is a complex, cyclical and interactive (Grønsund and Aanestad 2020; van den Broek et al. 2022) process that is not understood very well (Sharma et al. 2014), our findings indicate that it is feasible for firms to master this process with only limited available data.

Our literature review pointed out that there is no coherent definition of big data. While some papers define big data with a focus on the limits of current software and hardware capabilities (e.g., Bharadwaj et al. 2013; Kamioka and Tapanainen 2014; Fan et al. 2015), our study is in line with others (e.g., Lycett 2013; George et al. 2014; Constantiou and Kallinikos 2015; Yoo 2015) suggest putting less weight on the volume of data rather than on their other characteristics, like the variety, velocity, and veracity of these data.

Findings of earlier studies indicate that merely firms with large amounts of data available (Tambe 2014), who rely on massive information processing (Wu et al. 2019), and firms in the IT industry (Müller et al. 2018) benefit from data analytics. This might be confounded by the fact that these firms are already used to work with data and act on data-driven insights. This means that they create and execute these (dynamic) capabilities and create competitive advantage. However, our study suggests that it is the capabilities rather than the resource of high volumes and large variety of data that matter. Similarly, Ghasemaghei et al. (2018) find in their empirical study that "bigness" of data does only have a small effect on the decision quality and no effect on decision effectiveness.

6.2 When volume does not matter, contextual conditions affect value creation

The literature that employs the resource-based view to investigate data value creation (e.g., Ghasemaghaei et al. 2018; Grover et al. 2018) served as a theoretical lens for our study. We found evidence that data analytics capabilities and value creation mechanisms also work in the case of limited data. From this, we conclude that the most critical factor affecting value creation with limited data are the presence of *contextual factors* (Grover et al. (2018) call them "moderating factors"), like strategy, leadership, data-driven culture in organizations. Our case study provided anecdotal evidence for several of these factors as we describe in the following.

6.2.1 Clear strategy and vision

We found that the goals of the analytics efforts changed during the project (from predictive to explorative and then again to predictive after the possibility of analytics became clear). We assume that this was mainly because of an unclear strategy how to exploit the data resources in the beginning which was clarified over time. Thus, we contend that awareness of available datasets in an organization and

the existence of a convinced vision on how to exploit them are crucial to achieve analytics value targets. If organizations lack such clear strategy and vision, they can—like in our case—cooperate with researchers or specialized analytics vendors to learn about possibilities and practices and derive novel knowledge and working theories for their business (Tremblay et al. 2021).

6.2.2 Leadership

Connected to vision and strategy is the necessity that managers need pursue their role as leaders for enabling the improvement of existing processes and practices. Researchers see analytics competency as a form of a dynamic capability (Ghasemaghaei et al. 2018; Božič and Dimovski 2019b; Kristoffersen et al. 2021), which means that it can alter other capabilities in an organization to create competitive advantage (Winter 2003) to implement data-driven findings. In our case, employees were also sometimes reluctant to appreciate new findings until their manager made a positive comment. This observation stays in line with findings from an earlier study (Frick et al. 2021) that found that too rigid participative management approaches (i.e., empowering leadership) do currently not work well in the area of data-driven innovation, because employees seek stability and have fear to be rationalized by algorithms.

6.2.3 Openness for agile working mode

The analytics performed in this case study was characterized by many interactions with the analytics and business stakeholders (e.g., inquiries about data, necessary explanations, follow-up questions, and goal-refinement). This characteristic of analytics projects was already demonstrated by several earlier studies (Shearer 2000; Asamoah and Sharda 2019; Grønsund and Aanestad 2020; van den Broek et al. 2022), Yet, we found the use of minimum viable products (MVPs) helpful to start small, then iterate and finally achieve the project goal. This approach is often applied in start-ups and becomes also accepted in established organizations (Dennehy et al. 2019). MVPs work thereby as boundary objects (Marabelli et al. 2017) and help in communicating between analytics experts and businesspeople. They help, in particular, in cases when the goal of analytics needs to be clarified over time.

6.2.4 Data awareness among all employees

Data quality is an often-cited and severe challenge in data analytics (Brodley et al. 2012; Baier et al. 2019; Someh et al. 2020). It limits the acceptance of such applications (Passlick et al. 2020), and might cause harmful misinformation. Several of our interviewees resume that the project has increased awareness of data at different levels and areas of the organization. Managers, for example, became aware of datasets and of data issues. Employees have understood, through the intensive discussion and inquiries about data, what consequences incorrect entries in systems can cause. With limited data, wrong information may get aggravated because there are not many database entries to alleviate this. A practice of adhering to frameworks for effective documentation of analytics results (e.g., Kühl et al. 2021) or data quality management (Bai et al. 2018) can help to establish an effective data governance (Alhassan et al. 2019).

Data awareness also helps to identify potential uses of datasets to exploit or help to identify further datasets, like open data (Hopf et al. 2017; Hopf 2019). Data should cover multiple facets of the business to enable modeling real-world phenomena, make high quality predictions and also enable the creation of new knowledge (Fredriksson 2018).

6.3 Potential limits of limited data

Nevertheless, some expectations to data value creation most likely will not hold for the case of limited data or need the existence of further moderating factors. First, the value promise of "democratizing data," which implies "transparency and access" to data (Grover et al. 2018, p. 401) by a large number of employees might be restricted by legal and competence reasons, but also require the right corporate

culture. Second, as amounts of data are limited, advanced analytics might be of limited help to carry out descriptive analytics, given that already existing BI solutions of spreadsheets are often sufficient to achieve similar results. Third, the mechanism "experiment with data" implies to some extent to carry out field experiments (e.g., A/B-Tests). This is, however, difficult with only few transactions available. In this case, measurable effects might require experiments with long run times.

6.4 Practical implications

Our empirical data analytics case demonstrated that two relevant variables in the sales process (DV1: *sales talk initiated* and DV2: *actual purchase*) can be predicted from the limited amount of available data with suitable performance, using out-of-the box ML algorithms. Even with data on less than 300 positive examples (initiated sales talks and purchased installations), meaningful estimations for process-guiding variables (scores) are feasible and exceed a common state of practice, where no sales prioritization is made. We found that this information can help to effectively allocate scarce human resources in the sales process and can assist sales agents in their daily work. In addition to the predictions of hot leads, ML can also derive additional knowledge about the clientele from the data and thus improve the general sales expertise in an organization.

6.5 Limitations and future research

We note the importance of four aspects that influence the success of data analytics but could not be examined in our in-depth qualitative study. First, data analytics in organizations emerges over time. This implies, on the one hand, that all data is associated with time. Therefore, characteristics and attitudes of customers change over time and models must be adapted in the future. On the other hand, we examined a firm that was just beginning its investments in data analytics. Value contributions certainly change in nature and scope as the maturity of the data analytics environment increases. This extended longitudinal perspective should be part of future research. Second, with an intelligent design of user interfaces or systematic recording of relevant data, the value created by data analyses can be further improved. In our case, adding new fields to the OC could, for example, further improve the quality of the predictions. Third, existing knowledge in companies and heuristics can already be very effective in making predictions about customer behavior, given that sales agents themselves have gained experience and developed own practices in sales conversations. Our study design did not allow for a comparison of the ML predictions with the predictive power of such heuristics (as baseline estimates). We motivate future studies, though, to carry out such a comparison in dedicated experiments, which could thereby quantitatively measure the business value creation from analytics solutions in comparison with a baseline without an analytics solution. Fourth, the consent of sales leads and customers that their data can be processed must be collected. Even if firms do not plan to introduce data analytics soon, they should start early to collect the consent of business partners for data processing, because the collection of ground truth data is a long process and customer data exchanges require established customer relations (Krafft et al. 2021).

7 Conclusion

Data together with analytics capabilities can provide strategic competitive advantages for firms. Yet, research so far has documented value contributions through data analytics for data-rich and IT-intensive business sectors but has paid little attention to situations with limited data. Our case study demonstrated that analytics capabilities and value creation mechanisms, originally identified for big data analytics, can also work with limited data. With this insight, all companies are encouraged to invest in resources and capabilities in this area early on, so that potentials can be leveraged. In addition, the results of our study underline the relevance of contextual factors, like a clear strategy, vision, and leadership of managers to steer data analytics value creation, the openness for an agile working environment, and the data awareness among all employees.

Acknowledgements

The work presented in this paper was financially supported within the framework of the ERA-Net SES initiative (project “SmartLoad”). We gratefully acknowledge this joint funding by the European Union, the Swiss Federal Office of Energy (grant number SI/501521-01), and the German Federal Ministry for Economic Affairs and Energy (grant number: 03050010).

Bibliography

Abbasi A, Sarker S, Chiang R (2016) Big Data Research in Information Systems: Toward an Inclusive Research Agenda. J Assoc Inf Syst 17:

Abbasi EK, Hubaux A, Acher M, et al (2013) The Anatomy of a Sales Configurator: An Empirical Study of 111 Cases. In: Salinesi C, Norrie MC, Pastor Ó (eds) Advanced Information Systems Engineering. Springer, Berlin, Heidelberg, pp 162–177

Alhassan I, Sammon D, Daly M (2019) Critical success factors for data governance: a telecommunications case study. J Decis Syst 28:41–61. https://doi.org/10.1080/12460125.2019.1633226

Alvesson M, Sandberg J (2011) Generating Research Questions Through Problematization. Acad Manage Rev 36:247–271

Alwosheel A, van Cranenburgh S, Chorus CG (2018) Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. J Choice Model 28:167–182. https://doi.org/10.1016/j.jocm.2018.07.002

Asamoah DA, Sharda R (2019) CRISP-eSNeP: Towards a data-driven knowledge discovery process for electronic social networks. J Decis Syst. https://doi.org/10.1080/12460125.2019.1696614

Bai L, Meredith R, Burstein F (2018) A data quality framework, method and tools for managing data quality in a health care setting: an action case study. J Decis Syst

Baier L, Jöhren F, Seebacher S (2019) Challenges in the Deployment and Operation of Machine Learning in Practice. In: ECIS 2019 Proceedings. AIS electronic library

Bailey DE, Barley SR (2020) Beyond design and use: How scholars should study intelligent technologies. Inf Organ 30:100286. https://doi.org/10.1016/j.infoandorg.2019.100286

Barney J (1991) Firm resources and sustained competitive advantage. J Manag 17:99–120

Barratt M, Choi TY, Li M (2011) Qualitative case studies in operations management: Trends, research outcomes, and future research implications. J Oper Manag 29:329–342. https://doi.org/10.1016/j.jom.2010.06.002

Bharadwaj A, Sawy OE, Pavlou P, Venkatraman N (2013) Digital Business Strategy: Toward a Next Generation of Insights. MIS Q 37:471–482

Bharadwaj AS (2000) A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation. MIS Q 24:169–196. https://doi.org/10.2307/3250983

Božič K, Dimovski V (2019a) Business intelligence and analytics use, innovation ambidexterity, and firm performance: A dynamic capabilities perspective. J Strateg Inf Syst 28:101578. https://doi.org/10.1016/j.jsis.2019.101578

Božič K, Dimovski V (2019b) Business intelligence and analytics for value creation: The role of absorptive capacity. Int J Inf Manag 46:93–103. https://doi.org/10.1016/j.ijinfomgt.2018.11.020

Brassington F, Pettitt S (2006) Principles of marketing, 4th ed. Prentice Hall, New York

Breiman L (2001) Random forests. Mach Learn 45:5–32

Brinch M (2018) Understanding the value of big data in supply chain management and its business processes: Towards a conceptual framework. Int J Oper Prod Manag 38:1589–1614. https://doi.org/10.1108/IJOPM-05-2017-0268

Brodley CE, Rebbapragada U, Small K, Wallace B (2012) Challenges and Opportunities in Applied Machine Learning. AI Mag 33:11. https://doi.org/10.1609/aimag.v33i1.2367

Brynjolfsson E, Rock D, Syverson C (2017) Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. National Bureau of Economic Research, Cambridge, MA

Chen DQ, Preston DS, Swink M (2015) How the Use of Big Data Analytics Affects Value Creation in Supply Chain Management. J Manag Inf Syst 32:4–39. https://doi.org/10.1080/07421222.2015.1138364

Chen H, Chiang R, Storey V (2012) Business Intelligence and Analytics: From Big Data to Big Impact. MIS Q 36:1165–1188

Chiang RHL, Grover V, Liang T-P, Zhang D (2018) Special Issue: Strategic Value of Big Data and Business Analytics. J Manag Inf Syst 35:383–387. https://doi.org/10.1080/07421222.2018.1451950

Coates TT, McDermott CM (2002) An exploratory analysis of new competencies: a resource based view perspective. J Oper Manag 20:435–450. https://doi.org/10.1016/S0272-6963(02)00023-2

Constantiou ID, Kallinikos J (2015) New games, new rules: big data and the changing context of strategy. J Inf Technol 30:44–57. https://doi.org/10.1057/jit.2014.17

Cui G, Wong ML, Wan X (2012) Cost-Sensitive Learning via Priority Sampling to Improve the Return on Marketing and CRM Investment. J Manag Inf Syst 29:341–374

Davenport TH, Harris JG (2007) Competing on Analytics: The New Science of Winning. Harvard Business Press

Dennehy D, Kasraian L, O'Raghallaigh P, et al (2019) A Lean Start-up approach for developing minimum viable products in an established company. J Decis Syst 28:224–232. https://doi.org/10.1080/12460125.2019.1642081

Donoho D (2017) 50 Years of Data Science. J Comput Graph Stat 26:745–766. https://doi.org/10.1080/10618600.2017.1384734

Elgendy N, Elragal A, Päivärinta T (2021) DECAS: a modern data-driven decision theory for big data and analytics. J Decis Syst 0:1–37. https://doi.org/10.1080/12460125.2021.1894674

Elia G, Polimeno G, Solazzo G, Passiante G (2020) A multi-dimension framework for value creation through big data. Ind Mark Manag 90:508–522. https://doi.org/10.1016/j.indmarman.2019.08.004

Fan S, Lau RYK, Zhao JL (2015) Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. Big Data Res 2:28–32. https://doi.org/10.1016/j.bdr.2015.02.006

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15:3133–3181

Fosso Wamba S, Akter S, Edwards A, et al (2015) How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. Int J Prod Econ 165:234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

Fosso Wamba S, Gunasekaran A, Akter S, et al (2017) Big data analytics and firm performance: Effects of dynamic capabilities. J Bus Res 70:356–365. https://doi.org/10.1016/j.jbusres.2016.08.009

Fredriksson C (2018) Big data creating new knowledge as support in decision-making: practical examples of big data use and consequences of using big data as decision support. J Decis Syst. https://doi.org/10.1080/12460125.2018.1459068

Frick NRJ, Mirbabaie M, Stieglitz S, Salomon J (2021) Maneuvering through the stormy seas of digital transformation: the impact of empowering leadership on the AI readiness of enterprises. J Decis Syst 0:1–24. https://doi.org/10.1080/12460125.2020.1870065

Galetsi P, Katsaliaki K, Kumar S (2020) Big data analytics in health sector: Theoretical framework, techniques and prospects. Int J Inf Manag 50:206–216. https://doi.org/10.1016/j.ijinfomgt.2019.05.003

George G, Haas MR, Pentland A (2014) Big Data and Management. Acad Manage J 57:321–326. https://doi.org/10.5465/amj.2014.4002

Ghasemaghaei M (2019) Does data analytics use improve firm decision making quality? The role of knowledge sharing and data analytics competency. Decis Support Syst 120:14–24. https://doi.org/10.1016/j.dss.2019.03.004

Ghasemaghaei M, Ebrahimi S, Hassanein K (2018) Data analytics competency for improving firm decision making performance. J Strateg Inf Syst 27:101–113. https://doi.org/10.1016/j.jsis.2017.10.001

Gillon K, Aral S, Lin C-Y, et al (2014) Business Analytics: Radical Shift or Incremental Change? Commun Assoc Inf Syst 34:. https://doi.org/10.17705/1CAIS.03413

Grønsund T, Aanestad M (2020) Augmenting the algorithm: Emerging human-in-the-loop work configurations. J Strateg Inf Syst 101614. https://doi.org/10.1016/j.jsis.2020.101614

Grover V, Chiang RHL, Liang T-P, Zhang D (2018) Creating Strategic Business Value from Big Data Analytics: A Research Framework. J Manag Inf Syst 35:388–423. https://doi.org/10.1080/07421222.2018.1451951

Günther WA, Rezazade Mehrizi MH, Huysman M, Feldberg F (2017) Debating big data: A literature review on realizing value from big data. J Strateg Inf Syst 26:191–209. https://doi.org/10.1016/j.jsis.2017.07.003

Gupta M, George JF (2016) Toward the development of a big data analytics capability. Inf Manage 53:1049–1064. https://doi.org/10.1016/j.im.2016.07.004

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer, New York, NY

Heiskanen E, Matschoss K (2017) Understanding the uneven diffusion of building-scale renewable energy systems: A review of household, local and country level factors in diverse European countries. Renew Sustain Energy Rev 75:580–591. https://doi.org/10.1016/j.rser.2016.11.027

Hopf K (2019) Predictive Analytics for Energy Efficiency and Energy Retailing, 1st edn. University of Bamberg, Bamberg

Hopf K, Riechel S, Sodenkamp M, Staake T (2017) Predictive Customer Data Analytics – The Value of Public Statistical Data and the Geographic Model Transferability. In: ICIS 2017 Proceedings. AIS electronic library, Seoul, South Korea

Hopf K, Sodenkamp M, Staake T (2018) Enhancing energy efficiency in the residential sector with smart meter data analytics. Electron Mark 28:. https://doi.org/10.1007/s12525-018-0290-9

Indira V, Vasanthakumari R, Sugumaran V (2010) Minimum sample size determination of vibration signals in machine learning approach to fault diagnosis using power analysis. Expert Syst Appl 37:8650–8658. https://doi.org/10.1016/j.eswa.2010.06.068

Johnston WJ, Leach MP, Liu AH (1999) Theory Testing Using Case Studies in Business-to-Business Research. Ind Mark Manag 28:201–213. https://doi.org/10.1016/S0019-8501(98)00040-6

Kaisler S, Armour F, Espinosa JA, Money W (2013) Big Data: Issues and Challenges Moving Forward. In: 2013 46th Hawaii International Conference on System Sciences. pp 995–1004

Kamioka T, Tapanainen T (2014) Organizational Use of Big Data and Competitive Advantage – Exploration of Antecedents. In: PACIS 2014 Proceedings. p 372

Kaur H, Pannu HS, Malhi AK (2019) A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Comput Surv 52:79:1-79:36. https://doi.org/10.1145/3343440

Kitchens B, Dobolyi D, Li J, Abbasi A (2018) Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data. J Manag Inf Syst 35:540–574. https://doi.org/10.1080/07421222.2018.1451957

Kohli R, Grover V (2008) Business Value of IT: An Essay on Expanding Research Directions to Keep up with the Times. J Assoc Inf Syst 9:. https://doi.org/10.17705/1jais.00147

Korcaj L, Hahnel U, Spada H (2015) Intentions to adopt photovoltaic systems depend on homeowners' expected personal gains and behavior of peers

Krafft M, Kumar V, Harmeling C, et al (2021) Insight is power: Understanding the terms of the consumer-firm data exchange. J Retail 97:133–149. https://doi.org/10.1016/j.jretai.2020.11.001

Kraus M, Feuerriegel S, Oztekin A (2020) Deep learning in business analytics and operations research: Models, applications and managerial implications. Eur J Oper Res 281:628–641. https://doi.org/10.1016/j.ejor.2019.09.018

Krippendorff K (2018) Content analysis: an introduction to its methodology, Fourth Edition. SAGE, Los Angeles

Kristoffersen E, Mikalef P, Blomsma F, Li J (2021) The Effects of Business Analytics Capability on Circular Economy Implementation, Resource Orchestration Capability, and Firm Performance. Int J Prod Econ 108205. https://doi.org/10.1016/j.ijpe.2021.108205

Kühl N, Hirt R, Baier L, et al (2021) How to Conduct Rigorous Supervised Machine Learning in Information Systems Research: The Supervised Machine Learning Reportcard. Commun Assoc Inf Syst

Kuhn M, Johnson K (2013) Applied Predictive Modeling. Springer New York, New York, NY

LaValle S, Lesser E, Hopkins MS, Kruschwitz N (2011) Big Data, Analytics and the Path From Insights to Value. MIT Sloan Manag Rev 52:

Lehrer C, Wieneke A, vom Brocke J, et al (2018) How Big Data Analytics Enables Service Innovation: Materiality, Affordance, and the Individualization of Service. J Manag Inf Syst 35:424–460. https://doi.org/10.1080/07421222.2018.1451953

Lockett A, Thompson S (2001) The resource-based view and economics. J Manag 32

Loureiro ALD, Miguéis VL, da Silva LFM (2018) Exploring the use of deep neural networks for sales forecasting in fashion retail. Decis Support Syst 114:81–93. https://doi.org/10.1016/j.dss.2018.08.010

Lycett M (2013) 'Datafication': making sense of (big) data in a complex world. Eur J Inf Syst 22:381–386. https://doi.org/10.1057/ejis.2013.10

Makadok R (2001) Toward a Synthesis of the Resource-Based and Dynamic-Capability Views of Rent Creation. Strateg Manag J 22:387–401

Marabelli M, Newell S, Vaast E (2017) Boundary Objects Survival Over Time: Insights from the Field of Healthcare Coordination. Acad Manag Annu Meet Proc 2017:1–6. https://doi.org/10.5465/AMBPP.2017.144

Martens D, Provost F, Clark J, de Fortuny EJ (2016) Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics. MIS Q 40:869–888

McAfee A, Brynjolfsson E (2012) Big Data: The Management Revolution. Harv Bus Rev 90:60–68

Melville N, Kraemer K, Gurbaxani V (2004) Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value. MIS Q 28:283–322. https://doi.org/10.2307/25148636

Mikalef P, Boura M, Lekakos G, Krogstie J (2019) Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment. Br J Manag 30:272–298. https://doi.org/10.1111/1467-8551.12343

Mikalef P, Krogstie J (2020) Examining the interplay between big data analytics and contextual factors in driving process innovation capabilities. Eur J Inf Syst 29:260–287. https://doi.org/10.1080/0960085X.2020.1740618

Mikalef P, Pappas IO, Krogstie J, Giannakos M (2017) Big data analytics capabilities: a systematic literature review and research agenda. Inf Syst E-Bus Manag 1–32. https://doi.org/10.1007/s10257-017-0362-y

Motrenko A, Strijov V, Weber G-W (2014) Sample size determination for logistic regression. J Comput Appl Math 255:743–752. https://doi.org/10.1016/j.cam.2013.06.031

Müller O, Fay M, Brocke J vom (2018) The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. J Manag Inf Syst 35:488–509. https://doi.org/10.1080/07421222.2018.1451955

Narasimhan R (2014) Theory Development in Operations Management: Extending the Frontiers of a Mature Discipline via Qualitative Research. Decis Sci 45:209–227. https://doi.org/10.1111/deci.12072

Olson DL, Chae B (2012) Direct marketing decision support through predictive customer response modeling. Decis Support Syst 54:443–451. https://doi.org/10.1016/j.dss.2012.06.005

Passlick J, Guhr N, Lebek B, Breitner MH (2020) Encouraging the use of self-service business intelligence – an examination of employee-related influencing factors. J Decis Syst 29:1–26. https://doi.org/10.1080/12460125.2020.1739884

Rialti R, Zollo L, Ferraris A, Alon I (2019) Big data analytics capabilities and performance: Evidence from a moderated multi-mediation model. Technol Forecast Soc Change 149:119781. https://doi.org/10.1016/j.techfore.2019.119781

Roberts N, Galluch PS, Dinger M, Grover V (2012) Absorptive capacity and information systems research: Review, synthesis, and directions for future research. MIS Q 625–648

Rogiest W, Laevens K, Walraevens J, Bruneel H (2015) When random-order-of-service outperforms first-come-first-served. Oper Res Lett 43:504–506

Roh Y, Heo G, Whang SE (2021) A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. IEEE Trans Knowl Data Eng 33:1328–1347. https://doi.org/10.1109/TKDE.2019.2946162

Rouhani S, Rotbei S, Hamidi H (2017) What do we know about the big data researches? A systematic review from 2011 to 2017. J Decis Syst 26:368–393. https://doi.org/10.1080/12460125.2018.1437654

Sambamurthy V, Bharadwaj A, Grover V (2003) Shaping Agility through Digital Options: Reconceptualizing the Role of Information Technology in Contemporary Firms. MIS Q 27:237–263. https://doi.org/10.2307/30036530

Sandrin E (2017) Synergic Effects of Sales-Configurator Capabilities on Consumer- Perceived Benefits of Mass-Customized Products. Int J Ind Eng Manag 8:177–1888

Schryen G (2013) Revisiting IS business value research: what we already know, what we still need to know, and how we can get there. Eur J Inf Syst 22:139–169. https://doi.org/10.1057/ejis.2012.45

Seel J, Barbose GL, Wiser RH (2014) An analysis of residential PV system price differences between the United States and Germany. Energy Policy 69:216–226. https://doi.org/10.1016/j.enpol.2014.02.022

Sharma R, Mithas S, Kankanhalli A (2014) Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. Eur J Inf Syst 23:433–441. https://doi.org/10.1057/ejis.2014.17

Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. J Data Warehous 5:13–22

Shmueli G, Koppius OR (2011) Predictive Analytics in Information Systems Research. MIS Q 35:553–572

Shrivastava U, Jank W (2015) A data driven framework for early prediction of customer response to promotions. In: AMCIS 2015 Proceedings. AIS electronic library, Puerto Rico

Sirmon DG, Hitt MA, Ireland RD (2007) Managing firm resources in dynamic environments to create value: Looking inside the black box. Acad Manage Rev 32:273–292

Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of Big Data challenges and analytical methods. J Bus Res 70:263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

Smith A-S, Stöber J, Ulrich J (2020) How data analytics helps sales reps win more deals. McKinsey & Company, Vienna and Munich

Someh I, Wixom B, Zutavern A (2020) Overcoming Organizational Obstacles to Artificial Intelligence Value Creation: Propositions for Research. In: Proceedings of the 53rd Hawaii International Conference on System Sciences

Tambe P (2014) Big Data Investment, Skills, and Firm Value. Manag Sci 60:1452–1469. https://doi.org/10.1287/mnsc.2014.1899

Tarafdar M, Gordon SR (2007) Understanding the influence of information systems competencies on process innovation: A resource-based view. J Strateg Inf Syst 16:353–392. https://doi.org/10.1016/j.jsis.2007.09.001

Teece DJ, Pisano G, Shuen A (1997) Dynamic capabilities and strategic management. Strateg Manag J 18:509–533

Tremblay MC, Kohli R, Forsgren N (2021) Theories in Flux: Reimagining Theory Building in the Age of Machine Learning. MIS Q 45:455–459

van den Broek E, Sergeeva A, Huysman M (2022) When the Machine Meets the Expert: An Ethnography of Developing Ai for Hiring. MIS Q 45:1557–1580. https://doi.org/10.25300/MISQ/2021/16559

Vapnik VN, Vapnik V (1998) Statistical learning theory. Wiley New York

Walczak S (2001) An Empirical Analysis of Data Requirements for Financial Forecasting with Neural Networks. J Manag Inf Syst 17:203–222

Wilson HJ, Daugherty PR (2020) Small Data Can Play a Big Role in AI. Harv. Bus. Rev.

Winter SG (2003) Understanding dynamic capabilities. Strateg Manag J 24:991–995. https://doi.org/10.1002/smj.318

Woerner SL, Wixom BH (2015) Big data: extending the business strategy toolbox. J Inf Technol 30:60–62. https://doi.org/10.1057/jit.2014.31

Wolske KS, Stern PC, Dietz T (2017) Explaining interest in adopting residential solar photovoltaic systems in the United States: Toward an integration of behavioral theories. Energy Res Soc Sci 25:134–151. https://doi.org/10.1016/j.erss.2016.12.023

Wu L, Hitt L, Lou B (2019) Data Analytics, Innovation, and Firm Productivity. Manag Sci 66:2017–2039. https://doi.org/10.1287/mnsc.2018.3281

Yasmin M, Tatoglu E, Kilic HS, et al (2020) Big data analytics capabilities and firm performance: An integrated MCDM approach. J Bus Res 114:1–15. https://doi.org/10.1016/j.jbusres.2020.03.028

Yin RK (2018) Case study research and applications, Sixth Edition. Sage, Los Angeles ; London ; New Delhi ; Singapore ; Washington DC ; Melbourne

Yoo Y (2015) It is not about size: a further thought on big data. J Inf Technol 30:63–65. https://doi.org/10.1057/jit.2014.30

Zeng J, Glaister KW (2018) Value creation from big data: Looking inside the black box. Strateg Organ 16:105–140. https://doi.org/10.1177/1476127017697510

Appendix A: Conducted interviews and focus group discussions

We analyzed seven interviews and nine focus group discussions that we held during the research project (see Table A.1). The interviews were semi-structured to ensure that the relevant topics for the respective activities of the project are covered (#3-#7 focused on the sales process at the vendor and the system requirements, #11 and #12 focused on the development of the software prototype), but all interviews remained open to the exploration of participants' responses when they raised a different area of investigation. In the focus group meetings (except #16 and #17 who were only focused on the evaluation of the software prototype with sales agents), the whole project consortium participated. This means that at least two managers of the RES vendor, one or two representatives of the software company (consultant or developer), as well as the two authors of this paper were present in the meeting. We recorded all interviews and almost all focus group discussions (only #1, #8, #9, #14, #18 could not be recorded due to the organizational setting) and then transcribed verbatim and translated the most important quotations into English. One of the participating authors led the group meetings while the other participating author followed the discussion and took field notes that we later analyzed.

Table A.1: Meetings with informants, either in the form of interviews (I) or focus group discussions (FGD)

Item	Date	Type	Participants	Content
1	29.08.2018	FGD	Project consortium	Project initiation
2	25.02.2019	FGD	Project consortium	Project scoping and system requirements
3		I	Marketing manager	Sales process, system requirements
4		I	Head of RES sales department	
5		I	Sales agent 1	
6		I	Sales agent 2	
7		I	Sales agent 3	
8	20.08.2019	FGD	Project consortium	Discussion of first predictive models
9	31.10.2019	FGD	Project consortium, research grant officials	Project review and closing meeting
10	11.11.2019	FGD	Project consortium	Discussion of predictive models
11	22.11.2019	I	CTO of software company	Software development and design principles
12	28.11.2019	I	Consultant of software company	
13	05.12.2019	FGD	Project consortium	Discussion of explainable models
14	31.01.2020	FGD	Project consortium	Discussion of explainable models
15	04.03.2020	FGD	Project consortium	Evaluation of software prototype
16		FGD	Sales agent 2, Head of RES sales	
17		FGD	Sales agent 3, Head of RES sales	
18	18.03.2020	FGD	Project consortium, research grant official	Project review and closing meeting
19	10.12.2020	FGD	Data scientist 1 and 2	Status of data analytics application

Appendix B: Survey data

We conducted two online surveys to collect training data for the machine learning models. Both surveys run in parallel in June 2019. In order to guarantee anonymity in the review process (especially the research group and corporate partners that are named in the full survey), we only include the parts that are relevant for the research presented in this paper. The complete survey can be made available on request.

To the first survey, we invited all past users of the online configurator for heating systems and asked them: “Have you ordered a heating system in the meantime?” (Figure B.1). If the answer to this question was “Yes”, we asked participants: “Which heating system did you choose?” and presented them a choice between possible heating systems, where one was “heat pump” (Figure B.2). As a positive example for the machine learning model training, we considered the combination of both questions. In the second survey, we invited all past users of the photovoltaic online configurator to state whether they have already purchased a photovoltaic system with the question “Have you ordered a photovoltaic system in the meantime?” (Figure B.3). These questions were all mandatory to complete the survey.

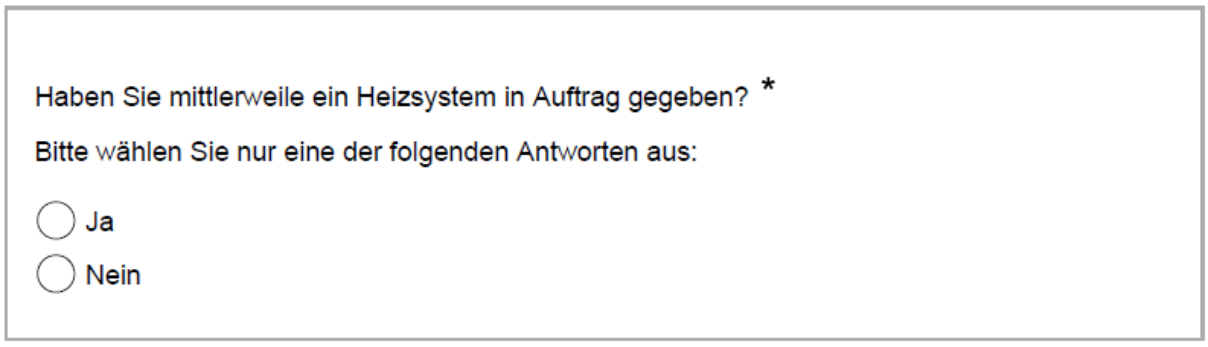

Figure B.1: Question for users of the heating online configurator regarding the purchase of a heating system

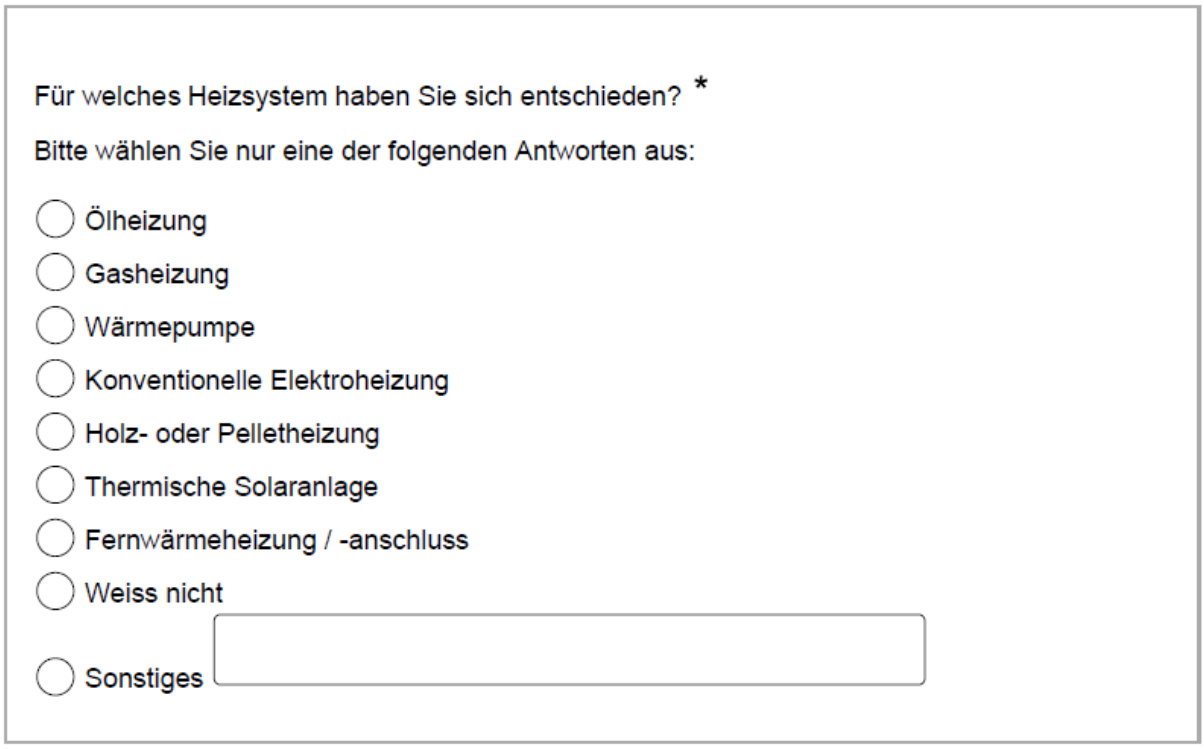

Figure B.2: Question for users of the heating online configurator regarding heating system

Figure B.3: Question for users of the photovoltaic online configurator regarding the purchase of a photovoltaic system

Appendix C: Details on machine learning results

In the main body of our paper, we applied machine learning algorithms with default hyperparameters and the entire dataset available. Below, we document two sensitivity analyses to show how the prediction performance changes with different hyperparameters and different amounts of input data.

Sensitivity analysis I: Hyperparameter tuning

In the first analysis, we test different hyperparameters to identify possible improvements for the machine learning algorithms Random Forest (RF), and Support Vector Machine (SVM). For RF, we follow Liaw and Wiener (2002)[1] and estimate the parameter *mtry* for the *default* model as the number of variables randomly sampled as candidates at each split with $mtry = \sqrt{number\_of\_features}$. We follow the default procedure implemented in the R package caret[2] to derive hyperparameters for a *grid search*: We use three values out of a simple sequence between 2 and the number of available features. In this analysis, the number of features increased compared to the number stated in Table 2 (of the main article) due to the data preparation step dummy encoding. For SVM, we have two parameters Cost (C), and the σ that must be determined. In case of the *default* model, we use C=1, and estimate sigma based on a Gaussian Radial Basis kernel with the standard procedure described by Karatzoglou et al. (2004)[3]. For C we use {0.25, 0.5, 1.0} and estimate three sigma values accordingly what results in a 3×3 *grid search*. There are no parameters to be tuned for the statistical method Logistic Regression (LR).

Table C.1 shows the prediction performance of all tested algorithms with the default parameter and using a grid search for the two dependent variables (DV1, DV2) and two types of renewable energy systems (photovoltaic, heat pump). The values correspond to the mean values of ten cross-validation predictions, the standard deviations (SD) are shown in brackets. The best prediction performance of each machine learning algorithm is printed bold. The samples in each cross fold were kept constant to guarantee comparability between the respective parameter combinations and machine learning methods.

[1] Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by RandomForest." R News 2 (3): 18–22.

[2] Kuhn M (2020) caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret

[3] Karatzoglou, Alexandros, Alex Smola, Kurt Hornik, and Achim Zeileis. 2004. "Kernlab – An S4 Package for Kernel Methods in R." Journal of Statistical Software 11 (9): 1–20.

In case of RF, prediction quality in terms of AUC does infrequently increase for different values of mtry. However, in case of DV2 (heat pump), larger mtry values increase prediction performances and lowers the standard deviation. In case of SVM, all prediction problems could benefit from a hyperparameter tuning as we found in all cases a variant that either increases the AUC or decreases the standard deviation.

In summary, there are some cases where hyperparameter tuning would improve predictive performance.

Table C.1: Prediction performance results (default and grid search) for lead scoring using two dependent variables

			Random Forest				Logistic Regression	Support Vector Machine									
DV1=Sales talk initiated	photovoltaic	search type	default	grid search			default	default	grid search								
		parameter	mtry=5	mtry=2	mtry=22	mtry=42	none	σ=0.009 C=1	σ=0.009 C=0.25	σ=0.009 C=0.5	σ=0.009 C=1	σ=0.025 C=0.25	σ=0.025 C=0.5	σ=0.025 C=1	σ=0.042 C=0.25	σ=0.042 C=0.5	σ=0.042 C=1
		AUC	0.68 (0.07)	**0.68 (0.06)**	0.67 (0.07)	0.66 (0.07)	0.64 (0.06)	0.72 (0.07)	0.71 (0.05)	**0.72 (0.06)**	0.71 (0.07)	0.68 (0.08)	0.68 (0.07)	0.68 (0.07)	0.69 (0.07)	0.71 (0.07)	0.66 (0.08)
		Sensitivity	0.67 (0.10)	0.66 (0.07)	0.56 (0.12)	0.59 (0.09)	0.65 (0.09)	0.79 (0.09)	0.73 (0.08)	0.71 (0.06)	0.73 (0.12)	0.69 (0.12)	0.62 (0.11)	0.68 (0.11)	0.61 (0.1)	0.67 (0.09)	0.67 (0.15)
		Precision	0.10 (0.02)	0.09 (0.01)	0.09 (0.02)	0.09 (0.01)	0.08 (0.01)	0.10 (0.01)	0.10 (0.01)	0.10 (0.02)	0.10 (0.01)	0.09 (0.02)	0.09 (0.02)	0.10 (0.02)	0.10 (0.02)	0.10 (0.02)	0.09 (0.02)
		F2	0.31 (0.05)	0.29 (0.03)	0.27 (0.06)	0.28 (0.04)	0.27 (0.03)	0.32 (0.04)	0.33 (0.04)	0.32 (0.04)	0.32 (0.04)	0.29 (0.04)	0.29 (0.06)	0.31 (0.05)	0.30 (0.05)	0.31 (0.05)	0.29 (0.06)
	heat pump	parameter	mtry=5	mtry=2	mtry=14	mtry=27	none	σ=0.011 C=1	σ=0.011 C=0.25	σ=0.011 C=0.5	σ=0.011 C=1	σ=0.030 C=0.25	σ=0.030 C=0.5	σ=0.030 C=1	σ=0.049 C=0.25	σ=0.049 C=0.5	σ=0.049 C=1
		AUC	0.56 (0.05)	0.56 (0.05)	0.55 (0.09)	0.55 (0.06)	0.56 (0.05)	0.55 (0.10)	0.54 (0.09)	0.58 (0.06)	**0.60 (0.07)**	0.55 (0.11)	0.60 (0.06)	0.55 (0.10)	0.59 (0.08)	0.59 (0.09)	0.56 (0.06)
		Sensitivity	0.53 (0.09)	0.54 (0.05)	0.53 (0.12)	0.52 (0.07)	0.50 (0.08)	0.50 (0.09)	0.46 (0.09)	0.52 (0.13)	0.54 (0.09)	0.47 (0.14)	0.54 (0.12)	0.48 (0.13)	0.57 (0.14)	0.52 (0.15)	0.53 (0.08)
		Precision	0.11 (0.02)	0.12 (0.02)	0.12 (0.02)	0.11 (0.01)	0.12 (0.02)	0.12 (0.03)	0.12 (0.04)	0.13 (0.02)	0.13 (0.02)	0.12 (0.03)	0.13 (0.02)	0.12 (0.03)	0.12 (0.02)	0.13 (0.03)	0.13 (0.02)
		F2	0.31 (0.05)	0.31 (0.03)	0.31 (0.06)	0.29 (0.04)	0.30 (0.04)	0.30 (0.06)	0.29 (0.06)	0.33 (0.06)	0.33 (0.05)	0.29 (0.07)	0.32 (0.05)	0.29 (0.07)	0.33 (0.06)	0.33 (0.08)	0.32 (0.05)
DV2=Actual purchase	photovoltaic	parameter	mtry=5	mtry=2	mtry=22	mtry=42	none	σ=0.007 C=1	σ=0.007 C=0.25	σ= 0.007 C=0.5	σ=0.007 C=1	σ= 0.021 C=0.25	σ= 0.021 C=0.5	σ= 0.021 C=1	σ= 0.036 C=0.25	σ= 0.036 C=0.5	σ= 0.036 C=1
		AUC	0.77 (0.08)	0.77 (0.08)	0.77 (0.08)	0.77 (0.08)	0.75 (0.06)	0.78 (0.08)	**0.78 (0.07)**	0.78 (0.08)	0.78 (0.08)	0.78 (0.08)	0.77 (0.08)	0.77 (0.08)	0.77 (0.07)	0.77 (0.07)	0.76 (0.08)
		Sensitivity	0.72 (0.09)	0.71 (0.10)	0.68 (0.10)	0.69 (0.11)	0.67 (0.10)	0.65 (0.10)	0.64 (0.09)	0.65 (0.11)	0.65 (0.10)	0.64 (0.08)	0.65 (0.09)	0.65 (0.08)	0.66 (0.06)	0.66 (0.08)	0.66 (0.08)
		Precision	0.73 (0.08)	0.70 (0.06)	0.75 (0.08)	0.75 (0.06)	0.75 (0.06)	0.82 (0.07)	0.78 (0.06)	0.81 (0.07)	0.82 (0.07)	0.77 (0.05)	0.78 (0.08)	0.76 (0.07)	0.75 (0.06)	0.75 (0.05)	0.75 (0.05)
		F2	0.72 (0.08)	0.70 (0.08)	0.69 (0.09)	0.7 (0.10)	0.69 (0.09)	0.67 (0.10)	0.66 (0.08)	0.67 (0.10)	0.67 (0.10)	0.66 (0.08)	0.67 (0.08)	0.67 (0.08)	0.68 (0.06)	0.67 (0.07)	0.67 (0.07)
	heat pump	parameter	mtry=5	mtry=2	mtry=15	mtry=28	none	σ=0.010 C=1	σ=0.010 C=0.25	σ= 0.010 C=0.5	σ= 0.010 C=1	σ=0.031 C=0.25	σ=0.031 C=0.5	σ=0.031 C=1	σ=0.051 C=0.25	σ=0.051 C=0.5	σ=0.051 C=1
		AUC	0.67 (0.13)	0.68 (0.13)	**0.72 (0.10)**	0.73 (0.12)	0.63 (0.13)	0.62 (0.09)	0.65 (0.09)	0.66 (0.07)	0.64 (0.12)	0.66 (0.09)	0.69 (0.10)	**0.69 (0.08)**	0.67 (0.11)	0.61 (0.13)	0.60 (0.12)
		Sensitivity	0.68 (0.18)	0.67 (0.19)	0.73 (0.12)	0.71 (0.19)	0.67 (0.22)	0.63 (0.16)	0.83 (0.18)	0.67 (0.19)	0.65 (0.19)	0.65 (0.22)	0.70 (0.18)	0.64 (0.22)	0.51 (0.3)	0.51 (0.23)	0.55 (0.23)
		Precision	0.35 (0.08)	0.35 (0.08)	0.41 (0.07)	0.39 (0.10)	0.34 (0.09)	0.34 (0.07)	0.34 (0.05)	0.35 (0.08)	0.35 (0.04)	0.35 (0.09)	0.37 (0.08)	0.36 (0.08)	0.38 (0.12)	0.35 (0.13)	0.35 (0.08)
		F2	0.57 (0.13)	0.56 (0.14)	0.63 (0.08)	0.61 (0.15)	0.55 (0.16)	0.53 (0.11)	0.64 (0.12)	0.56 (0.14)	0.55 (0.13)	0.55 (0.16)	0.59 (0.13)	0.54 (0.14)	0.46 (0.23)	0.46 (0.18)	0.48 (0.17)

Sensitivity analysis II: Reduction of the training size

In the second analysis, we tested how gradually reducing the training data size affects prediction performance in order to examine the lower bound of data necessary for the predictive analytics cases under study. To this end, we reduced the available data for the training phase in 1% increments from all available data (100%) to half of the available data (50%) and applied machine learning using RF with default parameters as described above. We held the remaining experimental parameters (e.g., seeds, data preparation, cross-validation) constant.

Figure C.2 shows the prediction performance for the different training sizes in terms of the average AUC on the vertical axis (in a 10-fold cross-validation) for each tested share of the training data, shown on the horizontal axis. The 95% confidence interval of the prediction is depicted in gray.

Figure C.2: Prediction performance results when reducing available training data size from 100% to 50%

Our first observation was that as the training data were reduced, the AUCs for all DVs decreased. This decrease is moderate for most of the predictions but considerably for the dependent variable *Actual purchase (DV2)–heat pump*. We particularly observed this decrease in prediction performance when using fewer than 80% of the data (see the data left of the dashed blue lines in Figure C.2). To examine this observation, we tested if AUC values for dataset sizes with fewer than 80% vs. dataset sizes were lower than those with 80% of the data or more (Figure C.3) and found support for this hypothesis for all dependent variables using Welch's two sample t-tests (Table C.4).

Figure C.3: Boxplots showing the distribution of the prediction performance in terms of AUC means for dataset sizes with <80% vs. dataset sizes ≥80%

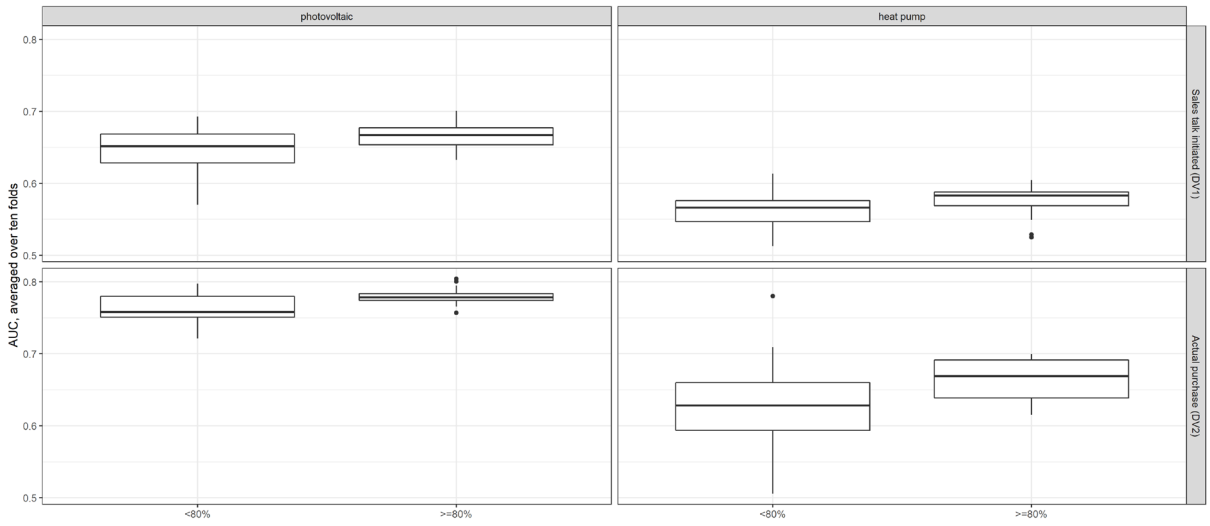

Table C.4: Welch's two sample t-test statistics indicate lower performance in terms of AUC means for dataset sizes with <80% vs. dataset sizes ≥80%

Dependent variable	Models trained on fewer than 80% of the entire dataset		Models trained on more than 80% of the entire dataset		Welch's two sample t-test statistics
	Mean AUCs	SD AUCs	Mean AUCs	SD AUCs	
Sales talk initiated (DV1)–photovoltaic	0.646	0.033	0.665	0.019	$t(47.46) = -2.71$, $p < .001$
Actual purchase (DV2)–photovoltaic	0.764	0.021	0.779	0.011	$t(45.53) = -3.36$, $p < .001$
Sales talk initiated (DV1)–heat pump	0.564	0.025	0.576	0.021	$t(46.88) = -1.78$, $p < .05$
Actual purchase (DV2)–heat pump	0.630	0.056	0.665	0.029	$t(45.98) = -2.93$, $p < .01$

Our second observation was that variances of the prediction performances (when looking at the individual cross-folds) increased with the decrease in the amount of training data, which can be seen with the larger confidence interval. We particularly observed this increase in variance when using fewer than 80% of the data (see the data left of the dashed blue lines in Figure C.2). To examine this observation, we compared the distribution of the standard deviation of the prediction performance for dataset sizes with fewer than 80% vs. dataset sizes with 80% or more (Figure C.5) and found support for this hypothesis for all dependent variables using Welch's two sample t-tests (Table C.6).

To sum up, the models reported in this study can also be trained with slightly fewer training data. Yet, the loss of performance (in terms of averaged AUC) and the increased uncertainty (in terms of standard deviation of AUC) already occur with 20% fewer training data. Therefore, we cannot recommend building the presented models with fewer training data. It is important to emphasize that the estimation of the amount of training data is always case-

specific and the results presented above may only be valid for the application case and the corresponding dependent variables at hand.

Figure C.5: Boxplots showing the distribution of the prediction performance in terms of AUC SDs for dataset sizes with <80% vs. dataset sizes ≥80%

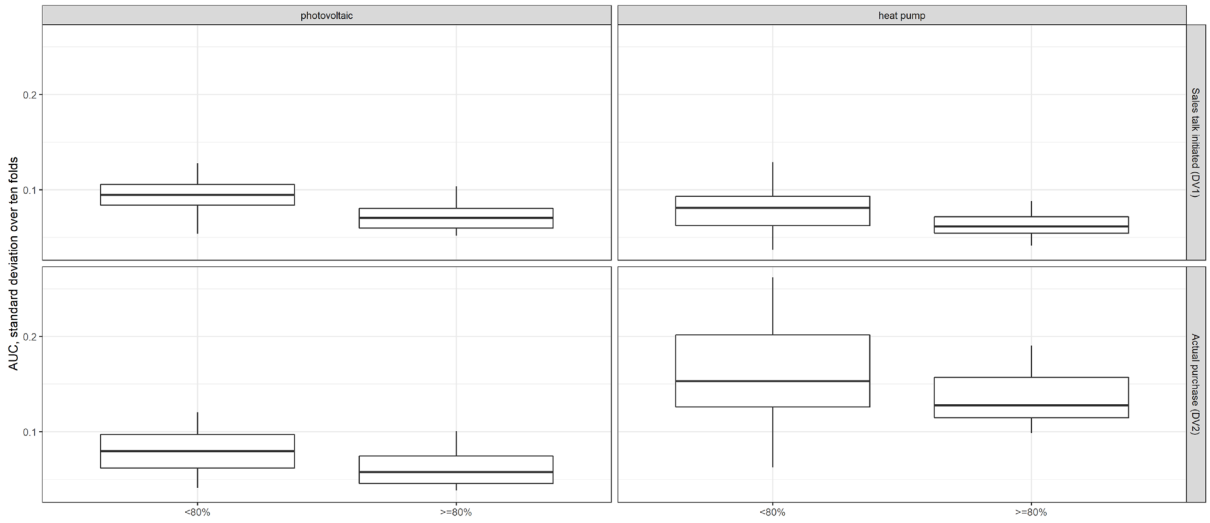

Table C.6: Welch's two sample t-test statistics indicate lower variance in terms of AUC SDs for dataset sizes with ≥80% vs. dataset sizes <80%

Dependent variable	**Models trained on fewer than 80% of the entire dataset**		**Models trained on more than 80% of the entire dataset**		**Welch's two sample t-test statistics**
	Mean AUC SDs	**SD AUC SDs**	**Mean AUC SDs**	**SD AUC SDs**	
Sales talk initiated (DV1)–photovoltaic	0.094	0.019	0.073	0.015	$t(48.09) = 4.27$, $p < .001$
Actual purchase (DV2)–photovoltaic	0.081	0.022	0.061	0.017	$t(48.59) = 3.58$, $p < .001$
Sales talk initiated (DV1)–heat pump	0.080	0.024	0.063	0.013	$t(45.82) = 3.31$, $p < .001$
Actual purchase (DV2)–heat pump	0.160	0.046	0.136	0.025	$t(46.48) = 2.38$, $p < .05$