DOI: https://doi.org/10.20378/irb-47756

Eurostars is a joint programme between more than 30 EUREKA member countries and the European Union

Schlussbericht zum Vorhaben

„Methoden der Energiedatenanalyse“

im Rahmen des Eurostars Projekts

E! 9859 BENginelI

„Energy Data Analytics: Increasing Service Quality and Energy Efficiency in the Residential Sector“

Konstantin Hopf, Thorsten Staake

Otto-Friedrich-Universität Bamberg, Lehrstuhl für Wirtschaftsinformatik, insb. Energieeffiziente Systeme

96047 Bamberg

Mit Beiträgen von: Tobias Graml, Ilya Kozlovskiy, Jan Marckhoff, Claire-Michelle Sévin, Mariya Sodenkamp

Gefördert durch das Bundesministerium für Bildung und Forschung (BMBF)

Förderkennzeichen: 01QE1550

Projektlaufzeit: 01.11.2015 – 31.10.2018

Inhaltsverzeichnis

I. Kurzdarstellung

Die deutsche Energiewirtschaft muss sich derzeit einer Reihe grundlegender Herausforderungen stellen. Einerseits ist die Branche mit der Umsetzung der Energiewende betraut und soll die Energieeffizienz bei Endkunden steigern, um zur Erreichung der nationalen und internationalen Klimaziele beizutragen. Andererseits führt zunehmender Wettbewerb durch die Marktliberalisierung in Europa zu sinkenden Margen bei Energieversorgungsunternehmen (EVU) und erhöht zudem den Innovationsdruck. Gleichzeitig bringt jedoch die Digitalisierung im Energiesektor, die insbesondere durch intelligente Stromnetze und die steigende Verfügbarkeit von Daten unterstützt wird, enorme Potentiale, die von Energieversorgern strategisch sinnvoll genutzt werden können.

Die Forschung in diesem Verbundvorhaben zielte darauf ab, Verfahren des maschinellen Lernens (ML) zu entwickeln, um verfügbare Daten bei EVU (Energieverbrauchsdaten, Tarifinformationen) sowie frei verfügbare Daten (Geoinformationen, Wetterdaten, usw.) so zu verarbeiten, dass wirkungsvolle, massenmarkttaugliche Energieeffizienzmaßnahmen und Dienstleistungen realisiert werden können. Hierzu wurden die Daten mit zusätzlichen Informationen aus Kundenumfragen oder früheren Kundenansprachen angereichert. Schlussendlich ermöglichen es die aus den Daten erzielten Erkenntnisse, die Servicequalität im Energievertrieb und die Energieeffizienz bei Endkunden zu steigern. Abbildung 1 illustriert die Zielsetzung des Eurostars Projekts „BENgine II“.

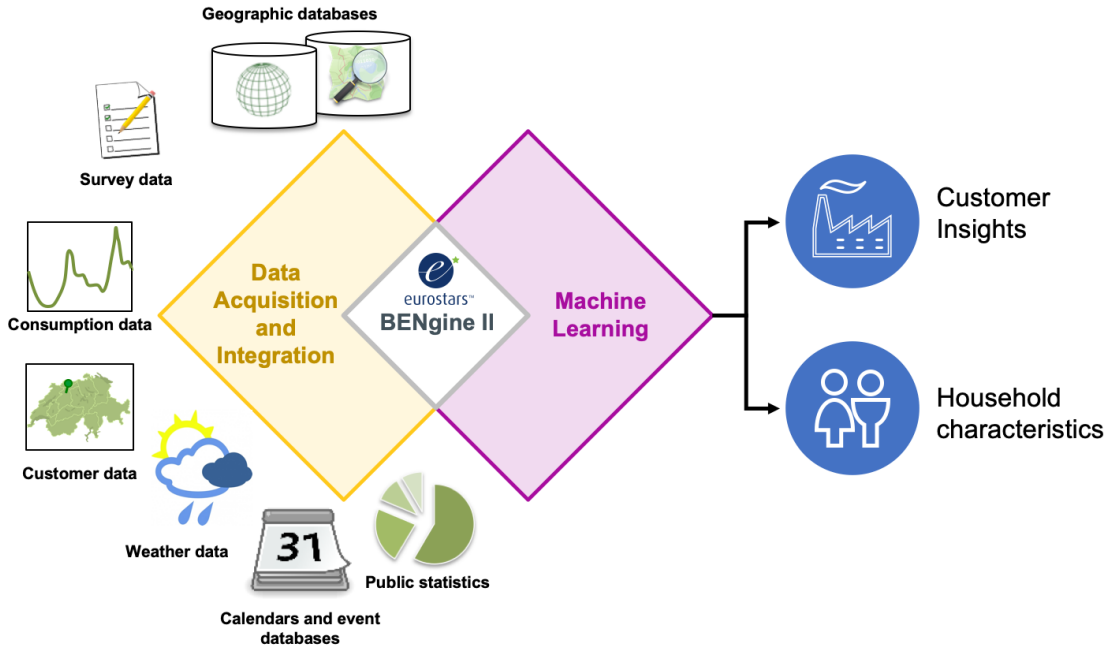


Abbildung 1: Vision des Projekts „BENgine II“

Im Teilprojekt der Otto-Friedrich-Universität Bamberg wurden Verfahren der Energiedatenanalyse entwickelt, die durch die BEN Energy AG (Zürich) in die BENgine-Softwareplattform eingebettet und zu nutzbaren Lösungen für EVU im europäischen Raum weiterentwickelt wurden. Die Lösung ist zum Zeitpunkt des Projektabschlusses bei über 30 EVU in Europa im Einsatz.

I.1 Aufgabenstellung

Ziel des Teilprojekts der Otto-Friedrich-Universität Bamberg war es, ML-basierte Verfahren der Energiedatenanalyse zu entwickeln, welche Haushaltsmerkmale aus Stromverbrauchsprofilen, Adressdaten und anderen verfügbaren Daten ableiten. Beispiele für die erzielten Informationen auf Einzelkundenebene sind Haushaltseigenschaften (z.B., Wohnungsgröße, Art der Wohnung, Art der Raumheizung und der Warmwasserbereitung), soziodemographische Kundenmerkmale (z.B. Anzahl der Bewohner im Haushalt), die Kaufbereitschaft für Up-/Cross-Selling Angebote des EVU, oder die Wechselbereitschaft zu einem anderen Energieanbieter. Die Prämisse bei der Entwicklung der Verfahren und der Implementierung in Softwarelösung war

der Einklang mit datenschutzrechtlichen Vorgaben, insbesondere den neuen gesetzlichen Rahmenbedingungen der DSGVO.

Die Projektergebnisse erlauben es EVU, Energieeffizienzmaßnahmen wirkungsvoller zu vermarkten. Dies wird möglich, da sich mit den entwickelten und getesteten Verfahren die kundenseitige Eignung und die Bereitschaft zum Kauf von effizienzsteigernden Up-/Cross-Selling Angeboten auf der Ebene einzelner Kunden vorhersagen lassen, wodurch sich Streuverluste und Prozesskosten reduzieren. Die entwickelten Verfahren wurden mit Daten von über 100.000 Haushalten validiert und in Kampagnen getestet.

I.2 Voraussetzungen, unter denen das Projekt durchgeführt wurde

Energieversorger haben einen großen Kundenstamm, ihr Wissen über konkrete Merkmale einzelner Haushalte ist jedoch gering. Dies wirkt sich sowohl auf die Entwicklung innovativer, haushaltsspezifischer Dienstleistungen als auch auf die monetären KPIs der Versorgungsunternehmen aus. Unsere maschinellen Lernwerkzeuge bieten Kunden Einblicke zu niedrigen Kosten und in großem Maßstab und lösen so ein bedeutendes Geschäftsproblem, verbessern die Effektivität von Energiesparkampagnen und erhöhen letztendlich den Kundennutzen und die Akzeptanz der damit verbundenen Dienstleistungen.

BEN Energy entwickelt und betreibt Software, die Versorgungsunternehmen dabei unterstützt, ihre Kunden in Energiesparkampagnen einzubinden und sie beim Verkauf entsprechender Dienstleistungen zu unterstützen. In diesem Projekt haben wir eine Reihe von Algorithmen des maschinellen Lernens entwickelt, die die Haushaltsmerkmale (Wohnungsgröße, Einwohnerzahl und Geräte usw.) ableiten und die Bereitschaft zur Teilnahme an Effizienz- oder Lastverlagerungskampagnen aus Lastprofilen, Standortinformationen und anderen vorhandenen Kundendaten vorhersagen.

I.3 Planung und Ablauf des Vorhabens

Das Forschungsvorhaben ist Bestandteil des Eurostars Projektes E! 9859 „BENgineII“ mit dem Titel „Energy Data Analytics: Increasing Service Quality and Energy Efficiency in the Residential Sector“. Neben der Otto-Friedrich-Universität Bamberg war die BEN Energy AG (Zürich) an dem Verbundvorhaben beteiligt. Im Einzelnen umfasste dieses Teilprojekt folgende Arbeitspakete (AP):

AP 1 – Deskriptive Verbrauchsdatenanalyse: Erstens wurden Energieverbrauchsdaten, sowie frei Verfügbare Daten erhoben. Zweitens wurden „Ground-Truth“-Daten zum Training der ML-Algorithmen mit Hilfe von Kundenumfragen, oder auf Basis von Kundenreaktionen aus der Vergangenheit, erhoben.

AP 2 – Datenschutz: Mechanismen der Datenbehandlung wurden gemäß den gesetzlichen Vorgaben entwickelt und implementiert. Darüber hinaus wurde die Datenverarbeitung so gestaltet, um einen hohen Grad an Kundenakzeptanz zu erreichen.

AP 3 – Feature-Extraktion: Die hochdimensionalen Daten wurden für die weitere Verarbeitung durch empirische Feature-Extraktion und automatische Variablenselektion reduziert und gemäß den Anforderungen der ML-Algorithmen vorbereitet.

AP 4 – Haushaltsklassifikation: Methoden des ML wurden entwickelt, um die energieeffizienzrelevanten Haushaltscharakteristika mit einer hohen Genauigkeit und Berechnungseffizienz aus den vorbereiteten Daten abzuleiten.

AP 5 – Präskriptive Verbrauchsdatenanalyse und Tool-Unterstützung: In diesem Schritt erfolgte die Entwicklung der Softwarelösung durch BEN Energy, die durch das Teilprojekt unterstützt wurde.

AP 6 – Projektmanagement: Das Konsortium wurde gemeinschaftlich durch die Partner verwaltet und es fand regelmäßiger Austausch im Rahmen der mindestens zweiwöchentlichen Telefonkonferenzen, sowie bei Workshops in Zürich und Bamberg statt.

Das Projekt war ursprünglich für den Zeitraum 01.11.2015 bis 31.10.2017 geplant. Aufgrund personeller Änderungen wurde das Projekt am 10.05.2017 ausgabenneutral um 12 Monate bis zum 31.10.2018 verlängert. Die Projektverlängerung wurde von allen Partnern sowie den nationalen Fördergebern befürwortet.

I.4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Es gibt zahlreiche Lösungsansätze, die das Problem der Erkennung von Kundenpotentialen und Haushaltseigenschaften aus Stromverbrauchsdaten adressieren. Dies ist Beleg dafür, dass das Thema für EVU von großer Wichtigkeit ist. Die rudimentären Algorithmen, die bisher in der Forschung entwickelt wurden, sind für den Praxiseinsatz bei EVU nicht geeignet und wurden deshalb für den produktiven Einsatz weiterentwickelt.

Ein Feld umfangreicher und andauernder Forschung ist das Erkennen von Geräten aus aggregierten Stromverbräuchen in Haushalten, das sogenannte „Non-Intrusive Load Monitoring" (Hart 1992; Zeifman und Roth 2011). Allerdings sind bei diesen Verfahren sehr hoch aufgelöste Verbrauchsdaten (von einer Messung pro Sekunde bis in den Megahertzbereich) nötig, welche in dieser Auflösung auch zukünftig, aufgrund der gesetzlichen Vorgaben, Privatsphäre-Aspekten und den technischen Limitationen, nicht flächendeckend zur Verfügung stehen werden. Deshalb fehlt bisher die wirtschaftliche Anwendung in der Praxis.

Andere Ansätze verwenden unüberwachte ML-Verfahren, um Segmente von Kunden zu bilden, die ähnliche Verbrauchsmuster aufzeigen (z.B. Chicco 2012; McLoughlin, Duffy, und Conlon 2012). Praktische Probleme mit diesen Lösungen sind einerseits die jeweils notwendige Interpretation der Segmente durch Experten, welche nicht automatisierbar und damit auch nicht skalierbar ist und andererseits die Instabilität solcher Segmente bei sich ändernden Daten. Zudem verwenden die Ansätze keine Ground-Truth-Daten, weshalb die Güte der Segmentierung kaum abgeschätzt werden kann.

Vielversprechender sind Ansätze, um Haushaltsmerkmale wie die durchschnittliche Anzahl der Bewohner oder deren Anwesenheitsdauer pro Tag mithilfe von überwachten Lernverfahren zu ermitteln (Beckel u. a. 2014). Unsere Forschungsgruppe hat in diesem Bereich in der letzten Zeit einige Fortschritte erzielen können. Auf Basis von 15-min Stromverbrauchsdaten intelligenter Stromzähler können beispielsweise energieeffizienz-relevante Haushaltseigenschaften, wie der Heizungstyp oder das Vorhandensein einer Photovoltaikinstallation erkannt werden (Sodenkamp u. a. 2017). Tabelle 1 listet frühere Forschungsprojekte der Forschungsgruppe auf, in denen rudimentäre Haushaltsklassifikationsalgorithmen entwickelt wurden. Diese Verfahren wurden in diesem Projekt zu prototypischen Softwarelösungen weiterentwickelt und in Marktumgebungen im Rahmen von Pilotkampagnen getestet.

Tabelle 1: Frühere Forschungsprojekte der Forschungsgruppe, in denen rudimentäre Haushaltsklassifikationsalgorithmen entwickelt wurden

Zeitraum	**Projekttitel und Ziel**	**Fördergeber**	**Partner**
2014 – 2015	**Titel: Smart-Meter-Datenanalyse für automatisierte Energieberatung** Reproduktion und Erweiterung des Basis-ML-Verfahrens zur Haushaltsklassifikation (Beckel u. a. 2014) auf Basis von 30-min Smart-Meter-Daten (aus Irland)	Bundesministerium für Energie Schweiz (CH)	ETH Zürich
2014 – 2016	**Titel: Smart Meter Data Analytics für Massenmarkttaugliche Energiedienstleistungen** Test der Praxistauglichkeit des Basis-ML-Verfahrens zur Haushaltsklassifikation (Beckel u. a. 2014) auf Basis von 30-min Smart-Meter-Daten (aus Irland)	Kommission für Technologie und Innovation (CH)	ETH Zürich, BEN Energy
2015 – 2016	**Titel: Smart-Meter-Datenanalyse für automatisierte Energieberatung II** Erweiterung des Haushaltsklassifikations-ML-Verfahrens um Wetter- und Geodaten, Erhebung von 15-min Smart-Meter-Daten und Umfragedaten in der Schweiz	Bundesministerium für Energie Schweiz (CH)	ETH Zürich, EVU

I.5 Zusammenarbeit mit anderen Stellen

Neben den Konsortialpartnern im Eurostars Projekt E! 9859 fand ein Austausch mit anderen Forschungsgruppen statt. Hierbei sind insbesondere zu nennen:

- ETH Zürich, Professur Informationsmanagement
- Universität St. Gallen, Institut für Technologiemanagement
- Copenhagen Business School, Department for Digitalization

Im Rahmen des Projekts wurden keine Unteraufträge erteilt.

II. Eingehende Darstellung

II.1 Erzielte Ergebnisse und Zielerreichung des Projekts

Nachfolgend sind die wichtigsten Ergebnisse der Projektarbeit im Teilprojekt der Otto-Friedrich-Universität Bamberg dargestellt. Die Ergebnisse der Aufgaben, welche durch den Projektpartner BEN Energy erzielt wurden, werden separat berichtet und können auf Anfrage zur Verfügung gestellt werden.

AP 1: Deskriptive Verbrauchsdatenanalyse

Ziel dieses Arbeitspakets war es, gesammelte, organisierte und synchronisierte Verbrauchsdaten, darunter Stromverbrauchsdaten (von Smart-Metern, wenn verfügbar, sonst von konventionellen Stromzählern), Wasser- und Gasverbrauchsdaten, geografische Daten (Wetter-, Satellitendaten), demographische und ökonomische Kennzahlen, Katasterdaten und andere relevante und verfügbare Daten zu sammeln, die als Trainingsdaten für die weitere Analyse genutzt werden können. Tabelle 2 zeigt eine Übersicht über die Datensätze, welche als Trainingsdaten zur Verfügung stehen.

Tabelle 2: Datensätze, die für das Forschungsprojekt zur Verfügung standen

		Number House-holds							Existing portal data / household characteristics																														
Dataset	Country	Consumption data	Survey data	E = Electricity, G = Gas, W= Water	Granularity	Time period	Geo-location	Analytical purpose	AppliancesAge	AppliancesNum	BedroomsNum	Children	Cooking	Desktop	Dishwasher	Dryer	Employment	EnergyEfficiencyMeasure	Family	Freezer	GameConsole	HeatingAge_Space	HeatingType_Space	HeatingType_Water	HeatPump	HouseAge	HouseholdType	HouseOwnership	InterestFTTH	InterestSolar	Laptop	LivingArea	ResidentsNum	Retirement	SatisfactionUtility	Single	SocialClass	SolarPanels
A	IR	6.444	4.232	E	30-min	18 month	no	Household classification		x	x	x	x	x	x	x	x		x	x	x					x	x				x	x	x	x		x	x	
B	CH	5.000	415	EGW	day	18 month	yes	Household classification															x	x			x					x				x		
C	CH	51.371	-	EG	year	3 years	yes	Upselling Green Tariff																															
D	DE	49.744	-	G	year	4 years	yes	Churn																															
E	DE	42.600	-	EG	year	4 years	yes	Churn, Cross-Selling, Clustering														x	x																
F	CH	200.846	423	E	year	4 years	yes	Household classification															x	x			x					x				x		
G	CH	3.073	3.073	E	year	3 years	yes	Household classification															x	x			x					x				x		
H	DE	1.058.972	2.074	E	year	1 year	yes	Household classification, Mailing															x	x			x					x				x		
I	BE	10.216	-	EG	year	3 years	yes	Heating Age														x	x															
J	CH	10.000	-	E	year	3 years	yes	Mailing selection																															
K	CH	10.350	-	E	year	3 years	yes	Mailing selection																															
L	CH	9.000	527	E	15-min	12 month	yes	Household classification	x	x		x	x					x	x			x	x	x	x	x	x	x	x	x	x	x		x	x		x		x

Im Ergebnis standen nach Abschluss dieses Arbeitspakets *gesammelte, organisierte und synchronisierte Verbrauchsdaten*, darunter Stromverbrauchsdaten (von Smart-Metern, wenn verfügbar, sonst von konventionellen Stromzählern), Wasser- und Gasverbrauchsdaten, *geografische Daten* (Wetter-, Satellitendaten), demographische und ökonomische *Statistiken*, Katasterdaten und andere relevante und verfügbare Zahlen, sowie *Trainingsdaten aus Kundenumfragen* für das Projekt zur Verfügung.

Datenbereinigung (Aufgabe 1.4)

Die Bereinigung der Verbrauchsdaten ist notwendig für Training und Test der Algorithmen. Im Detail haben wir Haushalte mit fehlenden Werten, solche ohne Verbrauchsmessung oder ohne Anwesenheit aus den Verbrauchsdaten identifiziert und eliminiert. Für die Umfragedaten haben wir Ausreißer und fehlende Werte identifiziert. Dieser Abschnitt beschreibt die Maßnahmen zur Datenbereinigung.

Behandlung von fehlenden Werten in Verbrauchs- und Erhebungsdaten: Der erste Schritt der Datenreinigung bestand darin, die fehlenden Werte herauszufiltern. Diese Werte stammen meist von den nicht einwandfrei funktionierenden Energiezählern. Es ist möglich, eine Interpolierung durchzuführen, um den Verbrauch für den fehlenden Zeitstempel zu schätzen. Da der Anteil der fehlenden Werte weniger als 1% beträgt und bei der Interpolierung Annahmen über die Verteilung getroffen werden müssen, haben wir die Beobachtungen ausgeschlossen.

Fehlende Werte haben wir auf zwei Arten behandelt: Erstens, Haushalte mit fehlenden Informationen in Fragebögen, die für eine spätere Analyse, die Algorithmenentwicklung und Bewertung der Algorithmen notwendig sind, wurden bei fehlenden Daten ausgeschlossen. Wenn beispielsweise Informationen über die Heizungsart fehlten, haben wir den Haushalt nur für die Analyse der Heizungsart und nicht für andere nicht fehlende Analysen ausgeschlossen. Zweitens, Fehlende oder Nullwerte in den Verbrauchsdaten wurden speziell kodiert, sodass diese Werte später in der Datenanalyse verarbeiten können. Auf Basis der Verbrauchsdaten wurden lediglich Haushalte ausgeschlossen, die nur fehlende Werte in den Verbrauchsvariablen haben, da eine Analyse der Verbrauchsdaten in diesem Fall nicht möglich ist.

Ausschluss von Ferienzeiten: Wir haben Feiertage und Schulferienzeiten in den jeweiligen Regionen der Datensätze betrachtet, um Zeiträume mit Feiertagen, besonderen Ereignissen und der Umstellung von oder auf Sommerzeit zu erhalten und diese besonderen Tage von der Analyse auszuschließen.

Erkennung von Abwesenheit: Heuristiken und Schwellwerte wurden verwendet, um Haushalte ohne Bewohner zu identifizieren. Beispiele für Heuristiken sind das Verhältnis von Verbrauchsmaximum zu –minimum oder die Anzahl der Nullwerte in einer Woche.

Identifizierung von Nicht-Haushalten: Abhängig von den verfügbaren Variablen im Datensatz, von den EVU zur Verfügung gestellt wurde, wurden durch den Projektpartner bereits Nicht-Haushalte aufgrund von Firmennamen in den Feldern Vorname, Nachname, Titel und Adresse identifiziert und entfernt.

Um weitere Nicht-Haushalte zu identifizieren, haben wir Text-Mining-Techniken auf Freitextfelder im Datensatz angewendet. Für die Variable "MeterAddres" im Datensatz F fanden wir 805 Schlüsselwörter und für "Betreff" im gleichen Datensatz 3'950 Schlüsselwörter, die Nicht-Haushalte oder besondere Haushaltsmerkmale (wie Wärmepumpe, PV-Anlage oder Stromzähler im Treppenhaus oder Keller) auszeichnen. Die häufigsten Schlüsselwörter finden Sie in der Tabelle 3.

Tabelle 3: Häufigste Zeichenketten in Freitextfeldern des Datensatzes F, identifiziert mit Text-Mining

Rank	Term	Frequency
1	whg	12’503
2	strom	10’795
3	einfamilienhaus	10’593
4	wohnung	10’296
5	allgemein	8’295
6	stock	6’184
7	[city district name]	5’702
8	efh	4’702
9	wohnungs	4’028
10	gebäude	3’297
11	allgemeinstrom	2’222
12	laden	2’163
13	parterre	1’602
14	attika	1’598
15	allgemeinelektrisch	1’531

Für das Text-Mining haben wir zunächst alle Adressinformationen eliminiert (oft wird nämlich in den Freitextfeldern nochmals die Adresse oder ein Stockwerk der Wohnung in den Häusern genannt) und die Wörter auf den Wortstamm reduziert, Leerzeichen entfernt und die restlichen Zeichen mit regulären Ausdrücken

bereinigt (mit Stemming, Stopwort-Entfernung, usw.). Dann haben wir Begriffe aus den Textkorpora extrahiert und manuell analysiert und Filter-Keywords für Nichthaushalte abgeleitet. Schließlich haben wir den Text-Mining-Filter auf alle Haushalte angewendet.

Ausreißererkennung in Verbrauchsdaten: Jährliche Verbrauchsdaten resultieren oft aus manuell erfassten Zählerständen, die in unregelmäßigen Abständen erhoben. Aus diesem Grund normieren wir die Verbrauchswerte immer auf den Tagesverbrauch für den gesamten Zeitraum, indem wir den Verbrauch durch die Anzahl der Tage während des Verbrauchszeitraums dividieren. Zusätzlich entfernen wir Haushalte mit einer Verbrauchsperiode, die kleiner als 100 Tage ist, da der Verbrauch in einem so kurzen Zeitraum hauptsächlich von der Jahreszeit abhängt.

Eine weitere Ausreißererkennung wird für die Smart-Meter-Daten durchgeführt. Dazu werden die Verbrauchsprofile mit dem mittleren Wochenverbrauch normiert, um sicherzustellen, dass die erkannten Ausreißer eine andere Form des Verbrauchs aufweisen als die anderen Haushalte und nicht nur einen hohen oder niedrigen Verbrauch haben. Für die Ausreißererkennung setzen wir Methoden ein, die lokal arbeiten (d.h. mehrere verschiedene Arten von Ausreißern finden können) und nicht von der Dimension der Eingabedaten abhängen (da wir Daten unterschiedlicher Granularität haben werden). Daher haben wir uns für die LOF-Methode (Local Outlier Factor) zur Bestimmung der Ausreißer entschieden. Die LOF-Methode ist eine Bewertungsmethode, die für jeden Punkt eine Punktzahl erzeugt, die misst, wie wahrscheinlich es ist, ein Ausreißer zu sein. Ein Beispiel für einen erkannten Ausreißer kann sein

Korrelationsanalyse (Aufgabe 1.6)

Die Korrelation zwischen externen Daten und Verbrauchsdaten ist als deskriptive Erkenntnis wichtig. Sie dient der Datenvorbereitung, Variablen aus der weiteren Analyse ausgeschlossen werden können, die nicht mit den Energieverbrauchsdaten korrelieren oder untereinander eine hohe Korrelation zeigen. Um die Variablen zu bestimmen, die Einfluss auf den Energieverbrauch V(t) haben, verwendeten wir das folgendes lineares Regressionsmodell:

$$V(t) = \Sigma_{ijk}\alpha_{ij} * W_k(t) * F_i * d_j + \Sigma_k\beta_k * S_k + \Sigma_k\gamma_k * V_k + \varepsilon$$

Dabei betrachten wir die Variablen, die als Zeitreihen zur Verfügung stehen (*W(t)* sind Wettervariablen), *S* statistischen Variablen, *V* sind Variablen aus Geoinformationssystemen) getrennt. Für die zeitabhängigen Variablen berücksichtigen wir zusätzlich die unterschiedlichen Tageszeiten (d) und andere Faktoren *F* (z.B. Arbeitstag, Wochenende oder Schulferien). Das Modell berechnet dann die Koeffizienten (α, β, γ) und den Fehlerterm ε. Die Bedeutung der Koeffizienten im linearen Modell ist für uns ein ausreichender Grund, die externen Variablen in unser Modell aufzunehmen. Basierend auf dem Datensatz B mit diesem Modell bestimmen wir beispielsweise, dass die Temperatur ein signifikanter Indikator für den Energieverbrauch ist (siehe Abbildung 2), aber die Luftfeuchtigkeit ist nur für den Verbrauch während des Tages signifikant (siehe Abbildung 3).

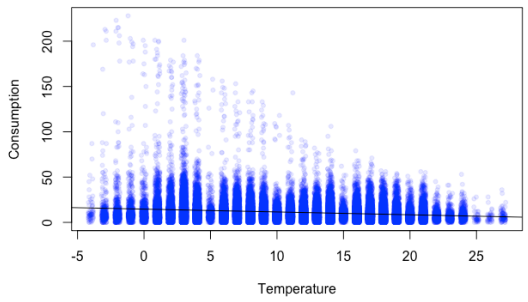


Abbildung 2: Verbrauch bei gegebener Temperatur mit geschätzter Regressionslinie

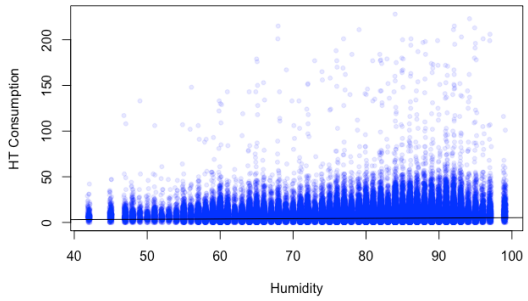


Abbildung 3: HT Verbrauch bei gegebener Luftfeuchtigkeit mit geschätzter Regressionslinie

Datensynchronisierung (Aufgabe 1.7):

Die Verbrauchskurven und die externen Daten wurden synchronisiert, wobei es folgende Hindernisse zu überwinden gab:

Zeitsynchronisation: Um verschiedene Zeitreihen in Deckung zu bringen wurden alle Messungen in Universal Time Coordinated (UTC) umgerechnet. Tage, an denen die Umstellung von oder auf Sommerzeit erfolgte, wurden von der Analyse ausgeschlossen. Neben der Synchronisation verschiedener Verbrauchszeitreihen haben wir auch Wetterdaten (Temperatur, Niederschlag, Bewölkung, usw.) mit den Verbrauchsspuren synchronisiert.

Geografische Synchronisation: Um den Energieverbrauch des Haushalts und die Haushaltseigenschaften mit externen Daten zu verbinden, ist eine Georeferenz erforderlich. Eine Georeferenz (genauer "Koordinatenbezugssystem") wird durch ein Koordinatensystem und ein räumliches Datum definiert (Becker 2012). Wir verwenden ein populäres geodätisches Referenzsystem WGS 84, das eng mit dem Global Positioning System (GPS) für die Routennavigation verbunden ist. Beide Systeme verwenden die Längen-/Breiten-/Höhenkoordinaten-Notation und sind in Web- und mobilen Anwendungen weit verbreitet. Mit einem Geokodierungsdienst haben wir die Adressdaten der Kunden in WGS 84-Koordinaten umgewandelt. Mit dieser Georeferenz kann die Verknüpfung von Geodaten erfolgen.

Die *Verknüpfung von statistischen Daten* mit Haushaltsdaten kann aufgrund der unterschiedlichen Verwaltungsaufteilung in den einzelnen Ländern nicht einfach mit den Georeferenzen erfolgen. Deshalb haben wir den offiziell veröffentlichten Ortsbezeichner für jedes Land in einer Ortsbezeichner-Metadatenbank manuell angepasst und ähnliche statistische Daten auch in diese Datenbank integriert. Details zu dieser Verknüpfung haben wir in (2017) ausgeführt. Abbildung 4 veranschaulicht das Datenbankschema in einem UML-Klassendiagramm. Für jeden Haushalt können wir die statistischen Daten

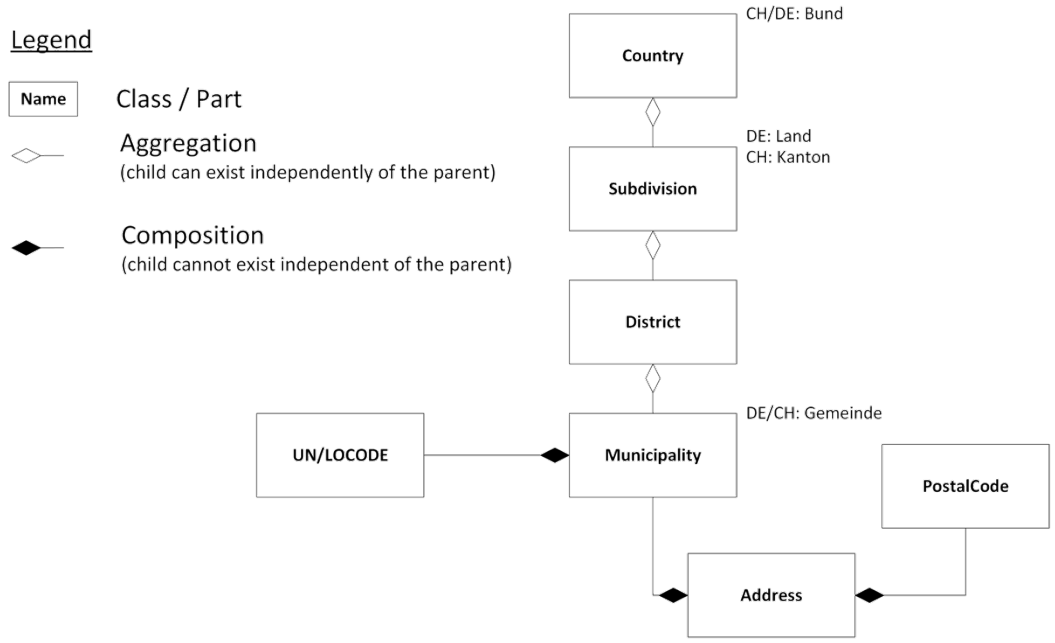


Abbildung 4: Datenbankstruktur für die Speicherung des Ortsbezeichners

Algorithmenentwicklung (Aufgabe 1.8)

Für die Datenerhebung, -bereinigung, -vorbereitung, und -synchronisierung wurden die Verfahren in Form von Algorithmen in der Statistiksoftware R implementiert.

AP 2: Datenschutz

Ziel dieses Arbeitspakets war es, sicherzustellen, dass die Datenbehandlung gemäß gesetzlichen Vorgaben entwickelt und implementiert wird (*data protection and privacy by design*). Darüber galt es, die Datenverarbeitung so gestalten, dass ein hoher Grad an Kundenakzeptanz erreicht wird. In diesem Teilproject wurden maßgeblich zwei Aufgaben erledigt.

Ergebnis dieses Arbeitspakets waren *Datenschutzvereinbarungen zwischen den Projektpartnern und eingeholte Einwilligungen der EVU-Kunden* zur Datenverarbeitung für die Forschungs- und Entwicklungsarbeit im Vorhaben und darüber hinaus. Des Weiteren wurden *Mechanismen für eine Datenbehandlung und Softwareentwicklung in einer Form, die Privatsphäre von Kunden bewahrt*, implementiert.

Techniken zum Data Hiding (Aufgabe 2.4)

Es wurden fünf Ansätze zum Data Hiding implementiert:

1. *Datenaggregation*: Innerhalb der modernen Smart Meter werden hochauflösende Daten über Verbräuche gesammelt, aber die gesetzlichen Vorgaben erlauben nur eine Übermittlung von 15-min Messintervallen oder größer. Die hoch aufgelösten Verbrauchsdaten (z.B. über 1 Hz) bleiben den Energieversorger somit verborgen. Für die Analyse stehen nur aggregierte 15-/30-Minuten-Daten zur Verfügung.
2. *Pseudonymisierung*: Hierbei werden Schlüsselbezeichnungen (z.B. Name, Vorlieben, ...) durch Pseudobezeichnungen (z.B. Alter, Postleitzahl, Geschlecht) ersetzt, sodass ein direkter Personenbezug nicht möglich ist.
3. *Verhinderung der Offenlegung der Identität*: Die Identität von Personen muss für Angreifer von Teilen des Systems verborgen bleiben. Dies wird durch die Dezentralisierung von Datenbank Teil 1 realisiert. Auf diesen Teil kann nur eine begrenzte Anzahl von Mitarbeitern von BEN zugreifen, und der Datenbankteil befindet sich an einem besonders gesicherten Ort. Forscher haben keinen Zugriff auf diesen Datenbankteil.

4. *Verhinderung der Attributoffenlegung*: Diese Technik stellt sicher, dass ein Angreifer, der ein Attribut erhalten hat (wie ein interner Identifikator), nicht in der Lage ist, die Personalinformationen zu erhalten, die zu dem Attribut gehören. Dies ist so implementiert, dass Kundenbezeichnungen des Versorgungsunternehmens durch interne, zufällig generierte Bezeichnungen ersetzt werden.
5. *Verhinderung von modellbasierter Identitätsoffenlegung*: Dies ist eine Anforderung an die zu entwickelnden maschinellen Lernalgorithmen. Angreifer dürfen nicht in der Lage sein, die Kundenidentität oder persönliche Attribute nur durch den maschinellen Lernalgorithmus zu erlangen.

Literaturrecherche zur Softwareakzeptanz bei Kunden (Aufgabe 2.7)

Wir haben eine Literaturrecherche zur Kundenakzeptanz von Haushaltsklassifizierungssystemen durchgeführt und festgestellt, dass dieses Thema in der Forschung noch nicht umfassend behandelt wurde. Der aktuelle Stand der Technik in benachbarten Disziplinen wurde identifiziert, Interviews mit Energieversorgern durchgeführt, Richtlinien für die Kundenakzeptanz von Haushaltsklassifizierungssystemen abgeleitet und auf Basis der Literatur und unseres Interviewmaterials Empfehlungen für energieeffiziente Mailingaktionen entwickelt.

Wir gliedern die damit verbundene Arbeit in zwei Teile. Der erste Teil behandelt die Akzeptanz solcher Systeme durch die Endverbraucher, der zweite Teil die Akzeptanz und den Technologieeinsatz in Versorgungsunternehmen.

Die Akzeptanz von Innovationen im Allgemeinen wurde von Rogers (1976) und Herbig & Day (1992) untersucht. Die Autoren identifizieren Faktoren, welche die Akzeptanz neuer Technologien bei Kunden beeinflussen:

- Relativer Vorteil (Wahrgenommener Innovationswert im Vergleich zu bestehenden Produkten)
- Kompatibilität (Wahrgenommener Grad der Integration von Innovation in ihre soziokulturellen Normen, Werte, Erfahrungen und Bedürfnisse)
- Komplexität (Wahrgenommene Komplexität von Innovationen, auch bestimmt durch die Anzahl der Entscheidungen, die mit dem Kauf der neuen Innovation verbunden sind)
- Testbarkeit / Teilbarkeit (Wahrscheinlichkeit, dass ein neues Produkt von einer kleinen Gruppe von Kunden getestet wird)
- Beobachtbarkeit (Die Einfachheit, die Vorteile oder Merkmale einer Innovation zu beobachten und mit anderen zu teilen)

All diese Faktoren sind schwer quantitativ zu messen. Aus diesem Grund messen wir die Kundenakzeptanz von Haushaltsklassifizierungssystemen durch die Nutzung von Online-Services, die BEN Energy für Versorgungsunternehmen und deren Kunden anbietet. Wir bewerten auch die Kundenakzeptanz der Systeme anhand der Rücklaufquoten von Mailings und Churn-Raten an einen neuen Lieferanten.

All diese Faktoren sind schwer quantitativ zu messen. Aus diesem Grund messen wir die Kundenakzeptanz von Haushaltsklassifizierungssystemen durch die Nutzung von Online-Services, die BEN Energy für Versorgungsunternehmen und deren Kunden anbietet. Wir bewerten auch die Kundenakzeptanz der Systeme anhand der Rücklaufquoten von Mailings und Churn-Raten an einen neuen Lieferanten.

Die Kundenakzeptanz von Online-Diensten war Gegenstand mehrerer Forschungsarbeiten:

- E-Commerce (Suh und Han 2003b)
- E-Banking (Kent Eriksson, Katri Kerem, und Daniel Nilsson 2005; Suh und Han 2003a)
- Soziale Netzwerke (Shin 2010; Krasnova und Veltri 2010)
- IP-TV (Jang und Noh 2011)
- Ortsbezogene Dienste (Zhou 2011)
- Biometrie (Miltgen, Popovič, und Oliveira 2013)

Eine wesentliche Determinante der Kundenakzeptanz ist an dieser Stelle die wahrgenommene Sicherheit und Privatsphäre bei der Nutzung solcher Portale (Shin 2010). Ein wichtiger Aspekt beim Aufbau internationaler Plattformen, die auf Nutzerinteraktionen abzielen, sind kulturelle Unterschiede in der Nutzung von Social Media Seiten zwischen den Ländern (Krasnova und Veltri 2010).

Ergebnisse aus der Forschung zur Akzeptanz und Nutzung von Technologie (Venkatesh, Thong, und Xu 2012) können helfen, auch die Akzeptanz von Online-Diensten zu erklären(Miltgen, Popovič, und Oliveira 2013).

Krishmurti et al. (2012) und Mah et al. (2012) haben die Akzeptanz der Endverbraucher für Smart Grid-Anwendungen empirisch untersucht und kommen zu dem Schluss, dass Privatkunden eine positive Einstellung zu Smart-Grid-Anwendungen haben (Verbrauchsrückmeldung, dynamische Preise etc.). Es ist jedoch notwendig, dass die Energieversorger ihre Kunden über den Mehrwert der Smart-Grid-Technologie und die Transparenz der Datenverarbeitung aufklären (Gangale, Mengolini, und Onyeji 2013), um Smart-Grid-Services gut am Markt zu platzieren. Ein wichtiger Treiber der Kundenakzeptanz sind auch finanzielle Anreize wie die dynamische Preisgestaltung von Energie, die durch Smart Meter ermöglicht werden (Motsch 2012).

Der Einsatz von prädiktiven Analysewerkzeugen bei EVU ist in der Literatur oft dokumentiert:

- Für die Verbrauchsprognose werden zahlreiche Methoden vorgeschlagen (z.B. multiple Regression, Exponentielle Glättung, Iterative Neugewichtete kleinste Quadrate, Adaptive Lastprognose, stochastische Zeitreihenanalyse, ARMAX-Modelle auf Basis genetischer Algorithmen, Fuzzy-Logik, Neuronale Netze und wissensbasierte Expertensysteme). Mit den Methoden können Versorgungsunternehmen ihren Strombezugsplan besser vermarkten (Alfares und Nazeeruddin 2002).
- Die Steuerung der Nachfrage der Verbraucher nach Stromdienstleistungen beim Kunden wird als Demand Side Management bezeichnet. Dieses Feld zielt auf die Motivation der Verbraucher für Energieeffizienz, Energieeinsparung und Lastverlagerung mit Hilfe von Demand-Response-Techniken ab. Oftmals eine solche Motivation durch Preisanreize oder Konsumfeedback für Verhaltensänderungen (Carley 2012; Gellings und Parmenter 2015; Strbac 2008).
- Kundenbindung und Kundenkommunikation spielen für alle Unternehmen eine große Rolle (Dixon, Freeman, und Toman 2010). Durch gezielte und korrekte Kundenansprache können die Kundenwechselraten gesenkt werden (Yang und Peterson 2004) und auch der Verkauf neuer Produkte oder Dienstleistungen wird gesteigert.

AP 3: Feature-Extraktion

Die hochdimensionalen Daten wurden für die weitere Verarbeitung durch empirische Feature-Extraktion und automatische Variablenselektion reduziert und gemäß den Anforderungen der ML-Algorithmen vorbereitet. Die drei Hauptgründe für die Definition empirischer Merkmale sind:

1. Verringerung des hohes Datenvolumens und der Dimensionalität, um die Komplexität der Datenverarbeitung zu reduzieren
2. Eliminierung von Zusammenhängen innerhalb der Variablen und Entfernen von unnötigem Rauschen
3. Modellierung von Expertenwissen (z.B. Identifizierung von Peaks aus Daten mit hoher Varianz, Feiertagen und besonderen Ereignissen)

Um die bestehenden Feature-Definitionen zu verbessern und zu erweitern und empirisch neue Features aus gekennzeichneten Stromverbrauchsdatensätzen zu entwickeln, hat BEN Energy Interviews mit drei Energieversorgern durchgeführt (Aufgabe 3.1 des Verbundvorhabens). Die Erkenntnisse aus den Interviews haben wir für die Dimensionsreduktion verwendet.

Identifikation von Zeitspannen mit homogenem Strom- Gas- und Wasserverbrauch (Aufgabe 3.3)

Wir haben typische Zeitspannen, die einen homogenen Energieverbrauch darstellen, basierend auf 30-minütigen Smart-Meter-Daten untersucht. Als Homogenitätsmaß in jedem Zeitfenster haben wir die Standardabweichung verwenden, die sich ergibt, wenn verschiedene Wochen an Verbrauchsdaten betrachtet werden.

Ergebnisse:

- Es lassen sich typische Zeitspannen des Stromverbrauchs erkennen, die den täglichen Lebenszyklus des Menschen darstellen.

- Die Standardabweichung nimmt mit jeder Erhöhung des Zeitintervalls (30-min zu Stunde, Stunde zu zwei Stunden, zwei Stunden zu drei Stunden) deutlich ab (t-Test, p-Wert < 0,0001).
- Stabile Zeiten über alle Lastkurven hinweg sind in der Nacht. Merkmale für diese Zeitspanne (zwischen 02:00 und 06:00 Uhr) sind daher für die Klassifizierung weniger relevant.

Die instabilen Zeiten treten tagsüber auf, besonders in den Spitzenzeiten (morgens und abends). Wenn der Verbrauch hoch ist, ist auch die Varianz des Verbrauchs hoch, da die Menschen nicht jeden Tag genau die gleichen Aktivitäten zur gleichen Zeit ausführen.

Vergleicht man die drei Beispiele der Haushalte (Single-Haushalt mit/ohne Beschäftigung und Nicht-Einzelhaushalt), so stellt man fest, dass die Zeiten hoher Varianz jeden Archetyp von Haushalten charakterisieren, aber es gibt - mit Ausnahme der Nacht - keine stabilen Zeiten, die für alle betrachteten Haushalte gelten. Wir empfehlen daher, Merkmale aus allen Zeitspannen zu berücksichtigen.

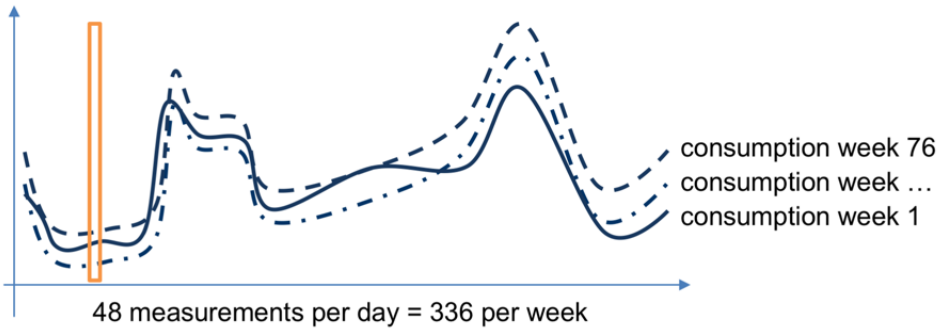
Die Zeiten hoher Varianz deuten auf eine typische Belegung im Haushalt hin, jedoch mit hohem Lärm. Wenn die Belegung vorhergesagt werden kann, empfehlen wir, den Verbrauch am Sonntag zu verwenden, da die Abweichung geringer ist.

Analyse:

Für die Analyse verwenden den kompletten Datensatz A, wobei wir Wochen mit Feiertagen, Schulferien, Zeitverschiebung, und unvollständigen Daten ausgeschlossen haben. Wie in Abbildung 5 dargestellt, haben wir eine Matrix mit $n=54$ nicht ausgeschlossenen Wochen i und $m=336$ Stromverbrauchsmessungen c_{ij}. Daraus berechnen wir die Standardabweichung als

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(c_{ij} - \bar{c_j}\right)^2}$$

wobei $\bar{c_j} = \frac{1}{n}\sum_{i=1}^{n} c_{ij}$ der Erwartungswert des Intervalls ist. Die Standardabweichung wurde fener auch über größere Zeitintervalle (1–3 Stunden) berechnet. Beispielhafte Ergebnisse für einen Haushalt (ID 5654) sind in Abbildung 6 dargestellt.



	c_1	c_2	c_3	c_4	c_5	c_6	c_7	…	c_{336}
week 1	0.080	0.110	0.075	0.030	0.180	0.151	0.031	…	0.112
week 2	0.101	0.088	0.032	0.071	0.111	0.031	0.054	…	0.124
…	…	…	…	…	…	…	…	…	…
week 76	0.094	0.098	0.029	0.089	0.305	0.380	0.092	…	0.116

Abbildung 5: Darstellung der Berechnung der Standardabweichung als Maß für die Homogenität der Zeiträume

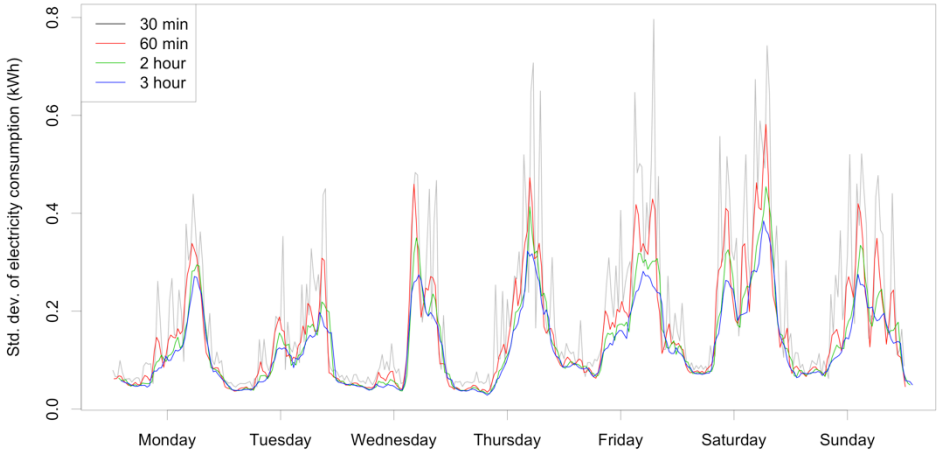


Abbildung 6: Standardabweichung einer beispielhaften Lastkurve (ID 5654) für eine Woche mit verschiedenen Messintervallen (ohne Normalisierung)

Da die Verbrauchsstandardabweichung in Zeiten mit hohem Verbrauch höher ist, normieren wir die Kennzahl mit dem erwarteten Wert des Verbrauchs im betrachteten Zeitrahmen:

$$\sigma_j^* = \frac{\sigma_j}{\bar{c}_j}$$

Die normierte Standardabweichung als Zeitreihendiagramm und als Heatmap für den Haushalt mit der ID 5654 ist in den Abbildung 7 und für zwei weitere exemplarische Haushalte in den Abbildungen 8 und 9 dargestellt.

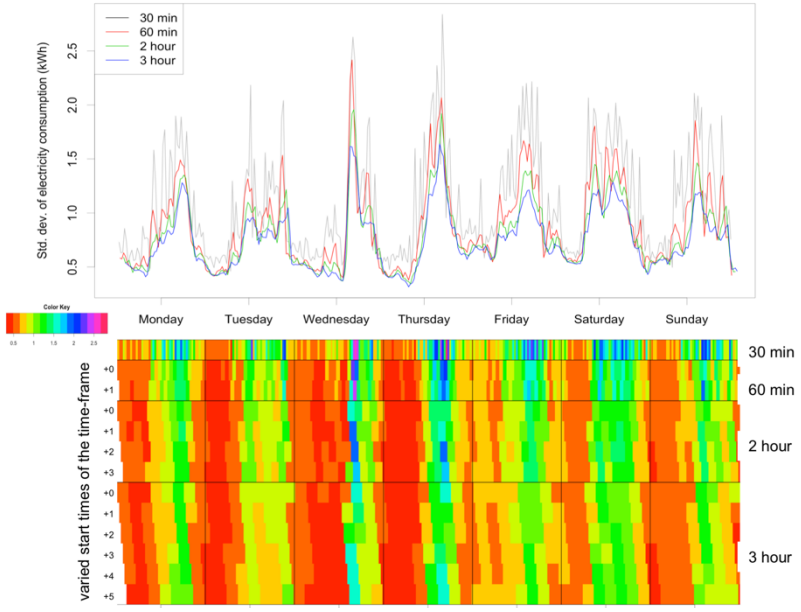


Abbildung 7: Normalisierte Lastkurve (Standardabweichung) für einen beispielhaften Haushalt (ID 5654) mit einer Person, die in Vollzeit arbeitet

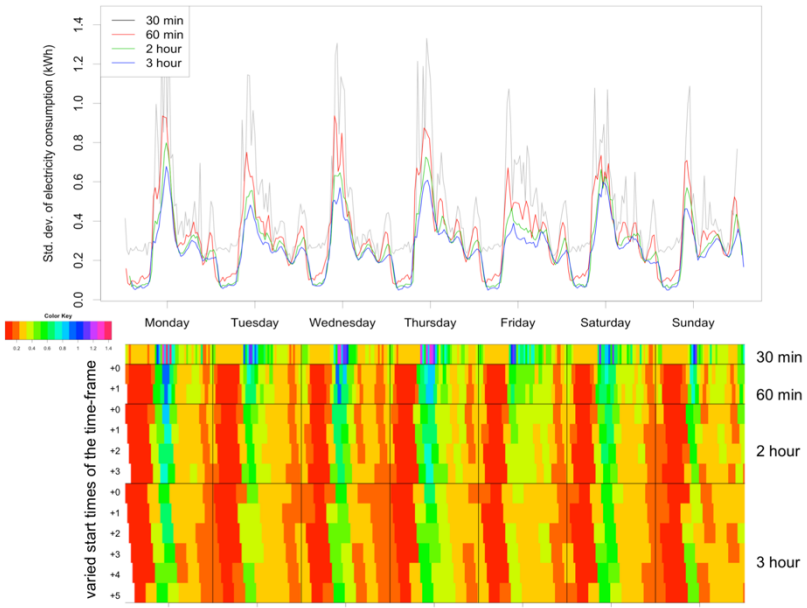


Abbildung 8: Normalisierte Lastkurve (Standardabweichung) für einen beispielhaften Haushalt (ID 6301) mit einer Person, die arbeitslos ist

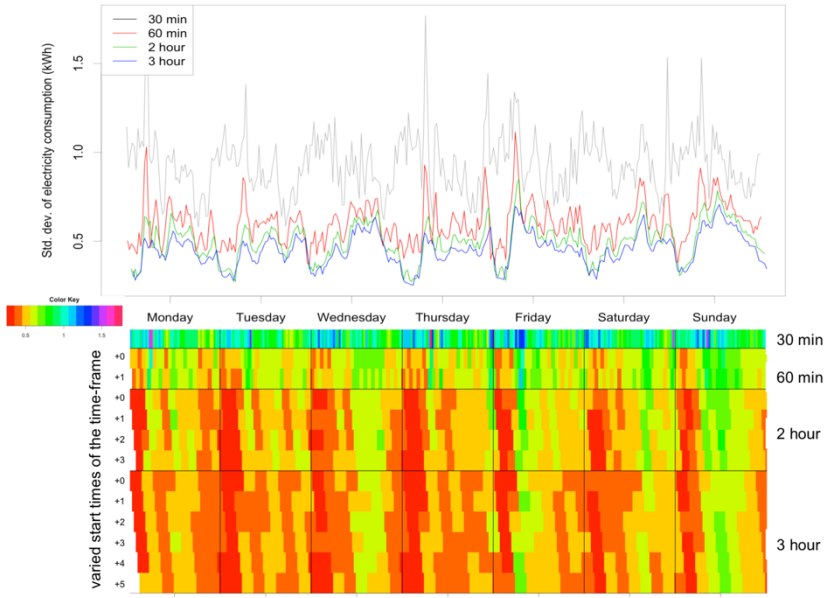


Abbildung 9: Normalisierte Lastkurve (Standardabweichung) für einen beispielhaften Haushalt (ID 6724) mit mehr als einer Person über 15 Jahre, wobei mindestens eine Person in Vollzeit beschäftigt ist

Algorithmische Zeitreihensegmentierung (Aufgabe 3.4)

Die algorithmische Zeitreihensegmentierung ist ein Ansatz der automatischen Feature-Extraktion. Hierbei haben wir Methoden der Signalverarbeitung verwendet (insb. Periodogramme und die diskrete Fourier-Transformation). Im Gegensatz zur empirischen Feature-Extraktion, unter Einbeziehung von Fachexperte, machen automatisierte Verfahren keine Annahmen über Zeitreihenprofile und sind sie in der Lage verschieden große Zeitfenster auf Zusammenhänge zu untersuchen.

Für diese Aufgabe verwenden wir tägliche Smart-Meter-Daten aus Datensatz B, welcher den Stromverbrauch von privaten Haushalten, gemessen in HT ("Hochtarif") und NT ("Niedertarif"), sowie den Gas- und Wasserverbrauch dieser Haushalte enthält.

Ergebnisse:

- In den Belastungsspuren des Stromverbrauchs existieren Wochen- und Jahreszyklen, die durch eine diskrete Fourier-Transformation erkannt werden können.
- Wöchentliche Zyklen verschwinden, indem der Verbrauch verschiedener Tarife zum Gesamtverbrauch addiert wird.
- Spitzenwerte und Statistiken (Varianz, Quantile, usw.) im Periodogramm können verwendet werden, um Merkmale automatisch aus Zeitreihendaten in unserer fortlaufenden Arbeit zu berechnen.

Analyse:

Der erste analytische Schritt war die visuelle Inspektion der verschiedenen Zeitreihen. Wir haben mehrere Haushaltsprofile analysiert und einen Haushalt mit den meisten verfügbaren Datenpunkten zur Veranschaulichung in Abbildung 10 dargestellt. Man sieht exemplarische Verbrauchszeitreihen für HT- und NT-Strom-, Gas- und Wasserverbrauch. Wöchentlich wiederkehrende Verbrauchszyklen sind in den Daten sichtbar: Der HT-Verbrauch sinkt an Wochenenden auf null (dies hat vermutlich tarifliche Gründe) und wir sehen zu diesen Zeiten Spitzenwerte im NT-Verbrauch.

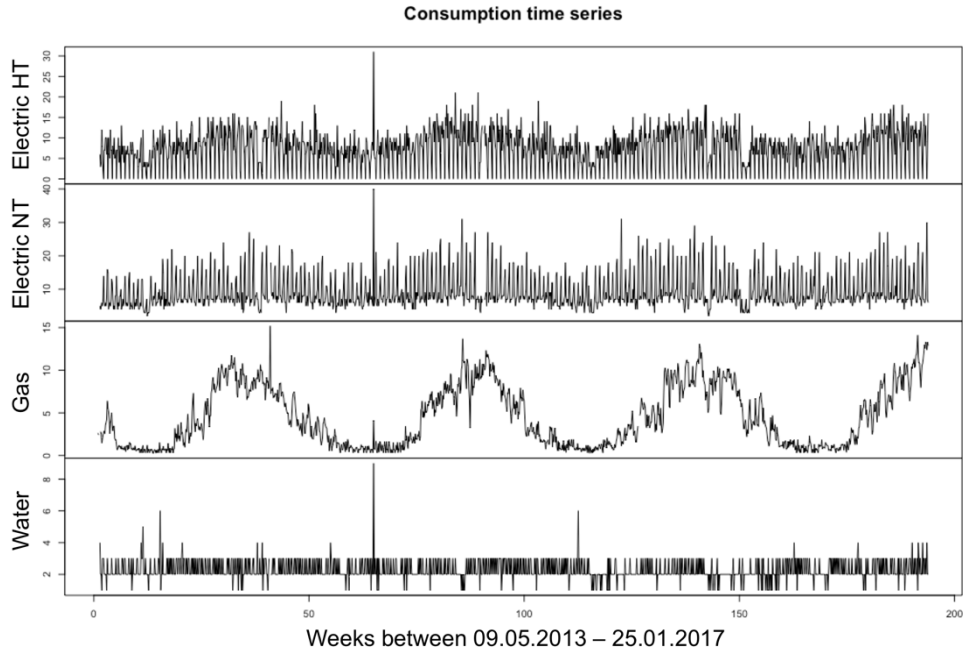


Abbildung 10: Mehrdimensionale Zeitreihendaten von täglichen Strom-, Gas- und Wasserverbrauchsmessungen (Haushalt 51912)

Periodogramme zeigen das Ergebnis einer Fast Fourier Transform (FFT) und stellen die Spektraldichte-Schätzung über die Frequenz eines periodischen Signals dar. Sie erlauben weitergehende Analysen der Verbrauchszyklen. Die Zeitreihendaten sind einmal mit dem Stromverbrauch aufgeteilt in HT und NT (Abb. 11) und einmal zusammengefasst (Abb. 12) dargestellt. Als Vorbereitung auf die FFT haben wir die Zeitreihenmessungen um das Maximum normiert, was auch ermöglicht, die Dichte verschiedener Zeitreihen zu vergleichen. Die Wochenzyklen im HT- und NT-Verbrauch sind als Spitzenwerte in Abbildung 11 für jeden vollen 7-Tage-Zeitraum (1, 2, 3, ... an der x-Achse) zu sehen. Bei der Zusammenfassung des Stromverbrauchs sind diese Spitzen kaum noch erkennbar (siehe Abbildung 12).

Die Fourier-Transformation hilft zwar, wiederkehrende Profile in Zeitseriendaten zu identifizieren, aber das Verfahren ist nicht geeignet für grobe Daten, die keine Signale nach trigonometrischen Funktionen (Sinus und Kosinus) enthalten.

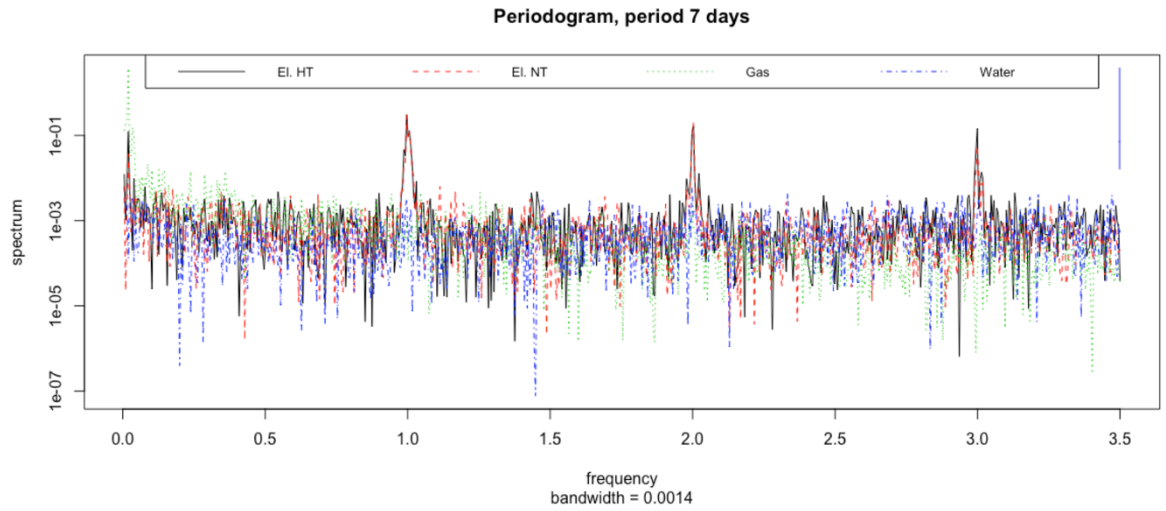


Abbildung 11: Periodogramm über das Spektrum der Frequenz (7 Tage), Stromverbrauch getrennt nach HT/NT, wöchentliche Routinen existieren an den Spitzen am Ende eines kompletten Wochenzyklus (Haushalt 51912)

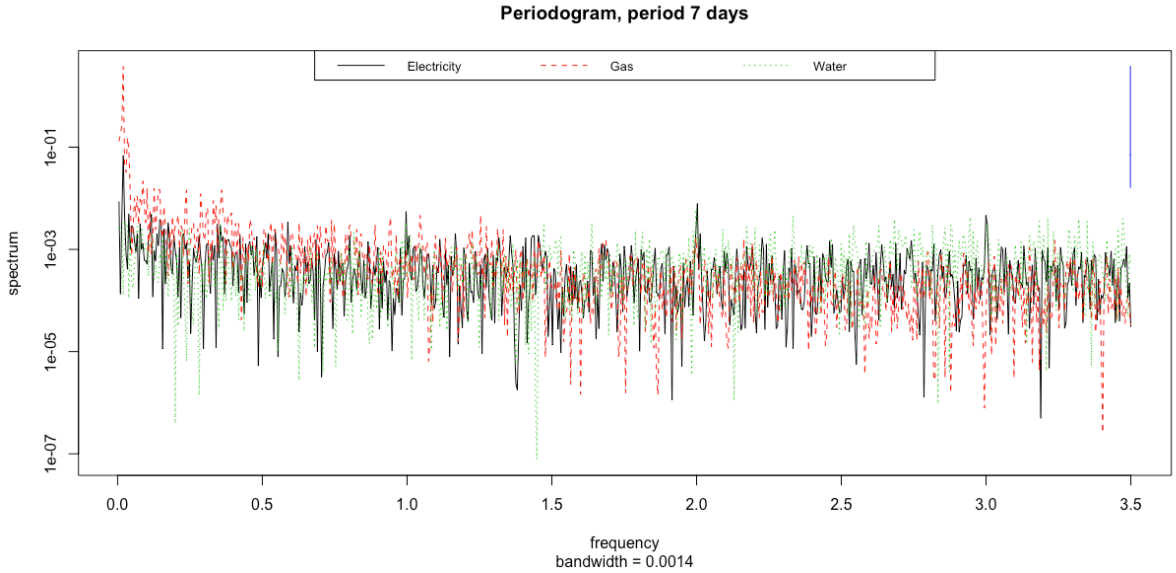


Abbildung 12: Periodogramm über das Spektrum der Frequenz (7 Tage), zusammengefasster Stromverbrauch pro Tag, wöchentliche Routinen (vgl. Spitzen) am Ende eines kompletten Wochenzyklus sind kaum noch erkennbar (Haushalt 51912)

Synchronisieren von Features basierend auf dem Elektrizitäts-, Gas- und Wasserverbrauch (Aufgabe 3.5)

BEN Energy hat in Zusammenarbeit mit uns Features für Zeitreihendaten einer einzelnen Verbrauchsart (entweder Strom, Gas oder Wasser) in Aufgabe 3.2 definiert. In dieser Aufgabe entwickelten wir weitere multidimensionale Merkmale, die die synchronisierten Messungen von zwei oder mehr Verbrauchsverläufen beschreiben. In einem ersten Schritt definierten wir Merkmale, die den Zusammenhang zwischen Strom-, Gas- und Wasserverbrauch beschreiben. Die Merkmale sind in Tabelle 4 aufgeführt.

In Kombination mit der Arbeit zur Lösung der Aufgabe 4.15 und 4.16 haben wir die Erkenntnisse aus der Zeitreihenanalyse für die Merkmalsentwicklung genutzt.

Tabelle 4: Mehrdimensionale Features für drei Verbrauchskurven (Strom, Gas, Wasser)

Name	Description
cor_el_wa (week / weekdays)	Correlation between overall electric and water consumption during the week / weekdays
cor_el_gas (week / weekdays)	Correlation between overall electric and gas consumption during the week / weekdays
cor_wa_gas (week / weekdays)	Correlation between water and gas consumption during the week / weekdays
cdc_lmCoef_gas	Coefficients of a multiple regression model explaining the electricity consumption: y_el = x_gas + x_wa + eps
cdc_lmCoef_wa	

Darüber hinaus berechneten wir Merkmale aus der Korrelation zwischen Verbrauchsdaten und Wetterbeobachtungen. Hier nutzten wir Ergebnisse aus früheren Untersuchungen, die aufzeigen, dass der Beitrag korrelationsbasierter Merkmale des Stromverbrauchs und der Wetterdaten zu Leistungssteigerungen in der Haushaltsklassifizierung führen kann (Sodenkamp, Kozlovskiy, und Staake 2016; Hopf, Sodenkamp, und Staake 2018). Anschließend betrachteten wir die folgenden Merkmale aus den Wettervariablen (Temperatur, Luftdruck, Niederschlag, Sonnenstunden, usw.):

1. cor_overall (Korrelation über die gesamte Zeitreihe)

2. cor_daily (durchschnittlicher Zusammenhang zwischen Wetter und Stromverbrauch an jedem Tag)

3-5. cor_night, cor_daytime, cor_evenening (Korrelation zu verschiedenen Tageszeiten)

6. cor_minima (Korrelation der Minima)

7. cor_maxmin (Korrelation von Wetterminima mit Verbrauchsspitzen)

8. cor_weekday_weekend (Verhältnis der Korrelation zu Wochenende / Wochentagen)

Automatische Feature-Extraktion (Aufgabe 3.6) und Komplexität der Feature-Extraktion (Aufgabe 3.10)

In dieser Aufgabe haben wir automatische Feature-Extraktionsmethoden, sowie unüberwachte Lernverfahre für die Merkmalsextraktion getestet.

Wichtigste Ergebnisse:

- Der Einsatz von Feature-Selektionsmethoden (FSM) in allen betrachteten Klassifizierungsproblemen brachte eine Verbesserung der Genauigkeit um 15,31% gegenüber dem Ergebnis, das durch die Berücksichtigung aller Merkmale im Durchschnitt erzielt wurde, wobei die Verbesserungen zwischen 3,06% und 21,85% lagen und die logistische Regression als Klassifizierer ohne interne Merkmalsumwandlung oder -auswahl betrachtet wurde.
- Die Leistung von FSM ist im Allgemeinen sehr stark von der verwendeten Datenprobe abhängig. Durch die Variation der Trainingsproben konnten wir Abweichungen in der Klassifizierungsgenauigkeit von 1% - 4% feststellen (zusammengefasst über alle getesteten Klassifizierungsinstanzen, so dass die tatsächliche Abweichung tatsächlich höher sein kann). Daher empfehlen wir, die Klassifizierungsergebnisse immer zusammen mit einem Standardfehler oder Konfidenzintervall zu berichten.
- Wir stellten fest, dass die logistische Regression zusammen mit der Feature-Auswahl Ergebnisse erzielen kann, die nahe an fortgeschrittenen Algorithmen (Random Forest, SVM) für die betrachteten Probleme liegen.
- Fünf Methoden konnten mit einer hohen Leistungssteigerung und relativ stabilen Ergebnissen identifiziert werden (F: Consistency, F:cfs, Boruta, C:Gini, C:InfGain). Wir können diese Methoden Datenanalysten als Ausgangspunkt für die Suche nach geeigneten FSM empfehlen.

Detaillierte Lösungsbeschreibung:

Die Lösung dieser Aufgabe ist zweigeteilt. Zuerst wendeten wir 44 bestehende Techniken zur Merkmalsauswahl, die irrelevante Merkmale auf unbeaufsichtigte Weise eliminieren. Zweitens verwendeten wir Erkenntnisse aus Aufgabe 3.4 und fortgeschrittene Techniken, um Merkmale aus Zeitreihendaten abzuleiten (Lebenszyklen wurden mit Fourier-Transformation, Hauptkomponentenanalyse, usw. erkannt).

Teil A: Automatische Feature-Auswahl

Es gibt eine große Anzahl von Feature-Selektionsmethoden (FSM), die bisher entwickelt wurden, und mehrere Softwarebibliotheken, die diese Methoden Datenanalysten zur Verfügung stellen (Chandrashekar und Sahin 2014; Guyon, André, und Elisseeff 2003; Saeys, Inza, und Larrañaga 2007). Wir fanden jedoch keinen umfassenden Vergleich von FSM, der hinreichend viele FSM systematisch vergleicht. Ziel dieser Aufgabe war es daher, bestehende FSM für den Einsatz in der Haushaltsklassifikation zu vergleichen. Dazu verwendeten wir einen Smart-Meter-Datensatz aus einem früheren Forschungsprojekt (Datensatz L) und haben 43 FSM nach den Kriterien (I) Verbesserung der Klassifikationsgenauigkeit, (II) Stabilität und (III) Durchschnittsgröße des resultierenden Feature-Menge verglichen.

Stand der Forschung: Bestehende FSM werden typischerweise in drei Kategorien eingeteilt: Filter, Wrapper und Embedded-Methoden. Wrapper- und Embedded-Methoden nutzen das Modelllernen und die Modellbewertung, um die Vorhersagekraft einzelner Merkmalssätze zu bewerten und Merkmalssätze mit einem lokalen Maximum der Modellleistung zu finden. Filtermethoden wählen Merkmale nur auf der Grundlage des Datensatzes aus und verlassen sich nicht auf das Erlernen und Auswerten von Modellen.

Frühe Arbeiten (Kudo und Sklansky 2000; Reunanen 2003) vergleichen Wrapper-Methoden auf der Grundlage verschiedener Datensätze in Bezug auf die Klassifizierungsgenauigkeit und fanden heraus, dass keine Methode anderen überlegen ist. Zwei aktuelle Studien bestätigten dieses Ergebnis, indem sie zehn (Hua, Tembe, und Dougherty 2009) und vier (Chandrashekar und Sahin 2014) Wrapper- bzw. Filtermethoden mit mehreren Datensätzen aus dem Bereich der Bioinformatik untersuchten. Ein großer Nachteil dieser Arbeiten ist, dass das FSM-Benchmarking nur die Klassifikationsleistung berücksichtigt. Weiterführende Studien beinhalten auch die Stabilität von FSM (Ähnlichkeit ausgewählter Feature-Sets durch leicht variierte Daten, die durch randomisiertes Sub-Sampling verursacht werden) in ihrer Analyse (Haury, Gestraud, und Vert 2011; Saeys, Inza, und Larrañaga 2007; Saeys, Abeel, und van de Peer 2008) und fanden heraus, dass Filtermethoden komplexere Wrapper- oder Embedded-Methoden übertreffen können. Vor kurzem beschrieben Lo et al. (2016) ein Verfahren zum Schätzen der Vorhersagekraft von Merkmalssätzen unabhängig vom Klassifizierer. Für die Praxis ist dieser Ansatz begrenzt, da Merkmale mit bis zu drei Kategorien kategorisch sein müssen.

Die bestehenden Studien berücksichtigen nur eine geringe Anzahl von FSM in kleinen Problembereichen. Meistens werden Probleme aus dem Bereich der Bioinformatik (Microarray, Massenspektrometrie-Daten zur Vorhersage von Krankheiten) verwendet, die es nicht erlauben, Merkmale verschiedener Skalentypen (binär, nominal, ordinal, ganzzahlig, kontinuierlich) zu umschließen. Außerdem werden nur binäre Klassifikationsprobleme (z.B. geringes oder hohes Risiko für Brustkrebs) berücksichtigt.

Metaanalyse von FSM: Unsere Arbeit erweitert die bestehende Forschung, indem wir drei Qualitätskriterien berücksichtigen und 43 FSM mit einem Empfehlungssystem des Energiehandels vergleichen. Wir berücksichtigen mehrere Klassifizierungsprobleme aus verschiedenen Bereichen (Marketing, Energieeffizienz, Lebenssituationen, usw.) mit Merkmalen mit unterschiedlichen Skalentypen und sowohl binären als auch mehrstufigen Klassifizierungsproblemen.

Eine Suche zu Softwarebibliotheken für GNU-R, die wir im Comprehensive R Archive Network (https://cran.r-project.org/, einem Software-Repository für die statistische Datenverarbeitung) durchgeführt haben, zur Durchführung der Feature-Auswahl, führte dazu, dass 43 implementierte Methoden für unsere Studie zur Verfügung standen (Kursa und Rudnicki 2010; Robnik-Sikonja und Alao 2016; Romanski und Kotthoff 2014). Wir entschieden uns, auf klassische Wrapper-Methoden (Forward-Selection, Backward-Elimination, usw.) wegen ihrer hohen Laufzeit zu verzichten (Modelltraining und Test werden mehrfach durchgeführt) und das Ergebnis früherer Studien zeigt eine geringere oder ähnliche Performance im Vergleich zu

Filtermethoden (Haury, Gestraud, und Vert 2011). Alle betrachteten FSM können in drei Kategorien eingeteilt werden (siehe Tabelle 5):

a) Einige Methoden berücksichtigen die Interdependenzen zwischen den Merkmalen. Umgekehrt werden Methoden, die Merkmale separat bewerten und den Kontext anderer Merkmale nicht berücksichtigen, als "verunreinigungsbasierte Methoden" bezeichnet (Liu und Motoda 2008).
b) Die Möglichkeit, einzelnen Klassen in einem Klassifizierungsproblem Klassenbedeutungs- (oder Kosten-) Faktoren zuzuordnen, ermöglicht eine Feinabstimmung der Klassifizierer und der FSM, um besonders kleine Klassen zu erkennen.
c) Einige Methoden haben explizite Unterstützung für Multiklassenprobleme, während andere (per Definition) für binäre Klassifizierungsprobleme konzipiert sind. In der Praxis funktionieren alle getesteten FSM für beide Arten von Problemen aufgrund der internen Klassenbinarisierung, aber es ist zu erwarten, dass Methoden mit expliziter Unterstützung für Multiklassenprobleme besser funktionieren als andere.

Tabelle 5: Kategorien der betrachteten Feature-Selektionsmethoden (FSM)

Characteristics of methods		**Frequency**
A) Interdependencies between features	Considered	16
	Not considered	27
B) Class-importance	Supported	9
	Not supported	34
C) Multi-class support	Explicit	29
	Implicit multi-class support (binary classification)	14
Method category	Filter	41
	Wrapper / embedded	2

Für die Literaturrecherche versuchten wir, die Originalreferenzen zu den Forschungsarbeiten zu finden, die die eingesetzten Methoden beschreiben. Einige Methoden basieren auf einigen bekannten Algorithmen (z.B. Entscheidungsbäumen) und unterscheiden sich nur in kleinen Komponenten der ursprünglichen Algorithmen (z.B. den ReliefF*-Methoden). Oftmals ist hier kein spezielles Papier verfügbar und wir verweisen auf das wahrscheinlichste Papier. Für einige Methoden (z.B. EqualGini) konnten wir keine Referenz finden, bei der diese Methode in Bezug auf Merkmalsauswahlmethoden, aber in einem anderen Kontext beschrieben ist. Also entschieden wir uns, diese Art von Referenzen zu ignorieren.

Methodik zum Benchmarking von FSM: Unsere Methodik zum Benchmarking von FSM und zur Auswahl der besten geeignetsten Methoden geht über bestehende Ansätze hinaus, die nur Teile der Merkmalsauswahlprobleme berücksichtigen. Wir geben einen kurzen Literaturüberblick und stellen danach die Qualitätskriterien für die in unserer Studie verwendeten FSM vor.

Qualitätskriterien für die FSM-Leistung: Wir bewerten die Leistung von FSM nach drei Kriterien: (I) Verbesserung der Klassifizierungsgenauigkeit, (II) Stabilität und (III) die Größe des resultierenden Funktionssatzes. Die Kriterien werden im Folgenden ausführlich beschrieben:

(I) Verbesserung der Klassifizierungsgenauigkeit: Die Klassifizierungsqualität ist das kritischste Kriterium für die Beurteilung von FSM und wurde in den meisten früheren Arbeiten verwendet. Wir entschieden uns für die Klassifizierungsgenauigkeit (der Prozentsatz der korrekten klassifizierten Beispiele unter allen Beispielen), da es sich um ein gut zu interpretierendes Maß handelt, das für Probleme mehrerer Klassen berechnet werden kann und einen Vergleich mit der bisherigen Forschung ermöglicht. Ein Wert von 1 bedeutet perfekte Klassifizierung, 0 eine totale Fehlklassifizierung. Die Klassifikationsleistung wird mittels k-facher Kreuzvalidierung geschätzt (Hastie, Tibshirani, und Friedman 2009). Um die Genauigkeit für verschiedene Klassifizierungsprobleme vergleichen zu können, betrachteten wir die Verbesserung der Genauigkeit in Bezug auf die Klassifizierungsleistung ohne Feature-Auswahl.

(II) Stabilität: Die Fähigkeit von FSM, ähnliche Funktionssätze unter Berücksichtigung verschiedener Teilmengen von Trainingsdaten zu finden, ist ein Indikator für die Zuverlässigkeit der Methode. Ein geeignetes Maß für die Ähnlichkeit zweier Merkmalssätze K_1 und K_2 ist der Jaccardindex (Saeys, Abeel, und van de Peer 2008):

$$Jaccard(K_1, K_2) = \left(\frac{|K_1 \cap K_2|}{|K_2 \cup K_2|}\right)$$

Die vollständige Ähnlichkeit der Mengen wird durch 1 ausgedrückt, disjunkte Merkmalssätze 0. Für jede Kombination (von Klassifizierungsproblem, FSM und Klassifikator) *c* schätzen wir die Stabilität als mittleren Jaccard-Index bei allen Permutationen von Merkmalssätzen K in der k-fachen Kreuzvalidierung (Kalousis, Prados, und Hilario 2007):

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

(III) Anzahl der ausgewählten Merkmale: Diese Zahl kann auch ein aussagekräftiges Kriterium zur Beurteilung der Qualität eines FSM sein, da einfache Modelle gegenüber komplexen Modellen bevorzugt werden (Hastie, Tibshirani, und Friedman 2009).

(IV) Laufzeit-Performance: Man kann sehen, dass die Rechenkomplexität (gemessen in der Ausführungszeit) eher ein informatives Kriterium als ein Ziel ist, das in der heutigen Rechenleistung minimiert werden soll. Bei der Durchführung der Modellauswahl und des Experiments mit verschiedenen Klassifikatoren, Merkmalsauswahlmethoden und deren Konfigurationen kann der Rechenaufwand jedoch ein kritischer Erfolgsfaktor sein. Die Laufzeit wurde in Berechnungssekunden mit gleicher Hardware gemessen und wird natürlich auf verschiedenen Computern variieren.

Ergebnisse:

Wir sahen uns hohen Varianz in den Ergebnissen konfrontiert, die von dem spezifischen Klassifizierungsproblem, dem verwendeten FSM und dem maschinellen Lernalgorithmus abhängen. Zusätzlich wurde die Performanz durch die Aufteilung der Daten in Trainings- und Testdaten beeinflusst. Doch mit der gleichen Kombination aus FSM und maschinellem Lernalgorithmus erhielten wir variable Ergebnisse aufgrund zufälliger Komponenten in den Methoden und der relativ geringen Anzahl von Trainingsbeispielen.

Daher entschieden wir uns, mit einem minimalen praktikablen Setup zum Benchmarking von FSM zu beginnen: logistische Regression mit der gleichen zufälligen Zuordnung von Datenbeispielen in die Felder für die Kreuzvalidierung. In einem zweiten Setup variierten wir die Trainingsdaten, indem wir neue Felder zufällig erstellten und ein Konfidenzintervall für die Ergebnisse schätzten.

Verbesserung der Genauigkeit von FSM in einem minimal realisierbaren Setup: Abbildung 13 zeigt die Verbesserung der Genauigkeit im Vergleich zu keiner Merkmalsauswahl mit dem logistischen Regressionsklassifizierer. Die Methoden mit Namen, die mit „F:“ beginnen, stammen aus dem Paket „FSelektor“ (Romanski und Kotthoff 2014), und „C:“ aus dem Paket „CORElearn“ (Robnik-Sikonja und Alao 2016)(Robnik-Sikonja und Alao, 2016) und „Boruta“ (Kursa und Rudnicki 2010). Mit Ausnahme von sechs verbessern alle FSM die Klassifizierungsgenauigkeit um mehr als 10%. Wir können daher Ergebnisse aus früheren Forschungen unterstützen (Haury, Gestraud, und Vert 2011), die die Gesamtklassifizierungsleistung verbessern.

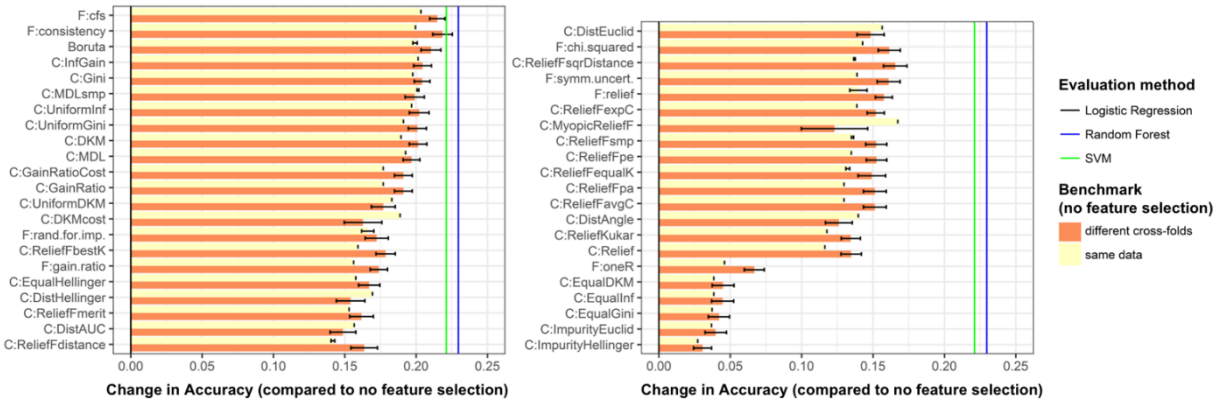


Abbildung 13: Klassifikationsgüteergebnisse verschiedener FSM in durchschnittlicher Accuracy-Abweichung gegenüber der Klassifikation ohne Feature-Selektion (Logistische Regression als Klassifikator)

Um die Ergebnisse abzuschätzen, führten wir iterativ Training und Test für jede Kombination aus FSM und Klassifikationsproblem zehnmal durch und nutzten t-Verteilung, um ein 95 % Konfidenzintervall zu schätzen. Wir schätzten die durchschnittliche Verbesserung aller FSM in zwei Setups, um die Gründe für Abweichungen in den Ergebnissen zu identifizieren: Im ersten Durchgang verteilten wir die Trainingsbeispiele mit der gleichen Zufallszuweisung auf fünf Felder (gelber Balken) und führten die FSM-Auswertung 10 mal durch (gelber Balken): Die meisten Methoden zeigen keine Varianz im Ergebnis, aber sechs Methoden haben intern zufällige Komponenten. Im zweiten Durchgang erstellten wir 10 verschiedene Verteilungen in die Felder der Kreuzvalidierung (orangefarbene Leiste). In diesem Fall sind die Konfidenzintervalle erwartungsgemäß größer.

Überraschenderweise kann die logistische Regression (als Klassifikator mit geringer Verallgemeinerungsleistung) - zusammen mit den besten FSM - Genauigkeitsergebnisse erzielen, die den Leistungsergebnissen von fortgeschrittenen Klassifikatoren (SVM, Random Forest) entsprechen.

Stabilität der Feature-Auswahl durch Variation der Trainingsdaten: Wenn man sich nur die Genauigkeitsergebnisse verschiedener FSM ansieht, kann aufgrund der hohen Varianz der Ergebnisse keine klare Entscheidung über die beste Methode getroffen werden. Daher berücksichtigen wir auch die Stabilität der Feature-Auswahl und die Anzahl der ausgewählten Features. Da ein signifikanter Zusammenhang (Pearson's $\rho = 0{,}51$, p-Wert $< 0{,}0001$, $t = 57{,}601$) zwischen Stabilität und der Anzahl der ausgewählten Merkmale besteht (die Wahrscheinlichkeit, ähnliche Merkmalssätze auszuwählen, steigt mit der Anzahl der ausgewählten Merkmale), normalisierten wir die Stabilität durch den Logarithmus der Anzahl der ausgewählten Merkmale und präsentieren die Ergebnisse in Abbildung 14. Fünf Methoden (F:consistency, F:cfs, Boruta, C:Gini, C:InfGain) haben eine hohe Leistungssteigerung (20,4% - 21,8%) und relativ stabile Ergebnisse (das 95%-Konfidenzinterval der Genauigkeitsverbesserung beträgt 1,2%). Beim Vergleich der Eigenschaften der Methoden ist kein klares Muster zu erkennen (z.B. berücksichtigen nur zwei Methoden die Interdependenzen zwischen den Merkmalen).

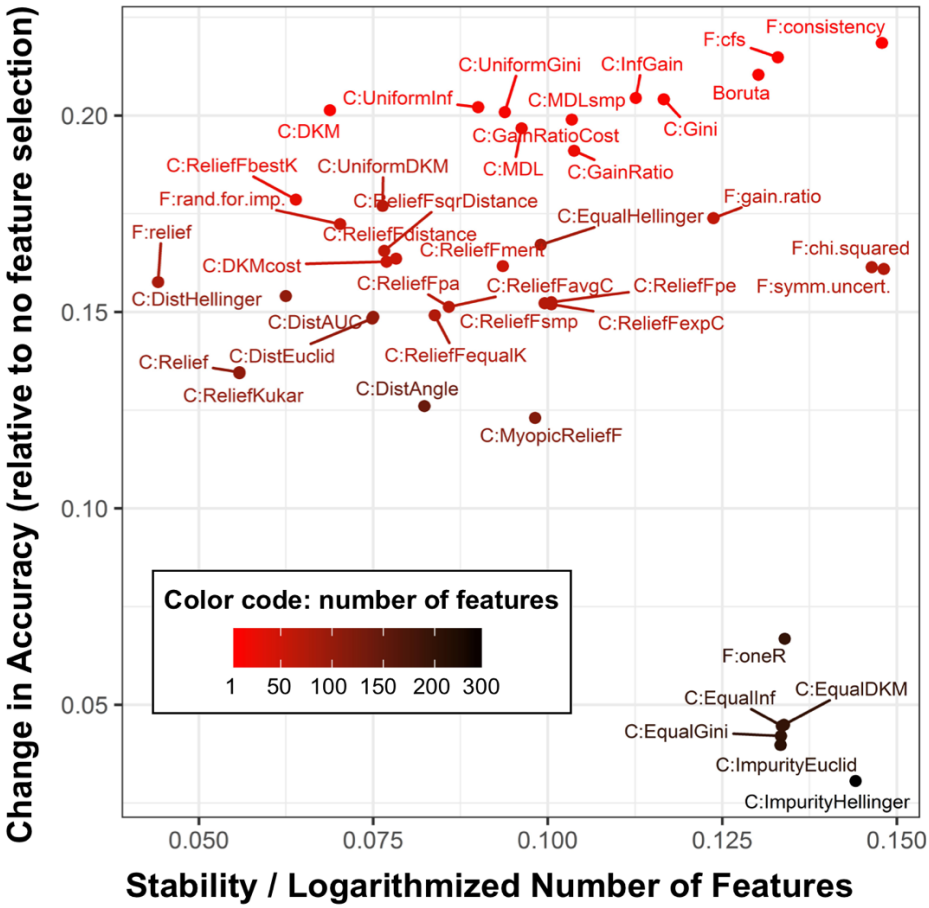


Abbildung 14: Getestete FSM mit Klassifikationsgüteverbesserung vs. Stabilität der Feature-Auswahl (normalisiert um die logarithmische Anzahl der ausgewählten Features)

Einfluss der Laufzeit auf die Genauigkeitsverbesserung: Die Motivation dieser Untersuchung basiert auf zwei Aspekten. Erstens gehen wir davon aus, dass die Laufzeit einer Merkmalsauswahlmethode als die Quantifizierung der Komplexität einer Merkmalsauswahlmethode angesehen werden kann. Die zweite Annahme ist, dass die Komplexität durch teure Rechenzeit dargestellt wird, die sich direkt auf die Laufzeit einer FSM auswirkt.

Abbildung 15 zeigt die Verbesserung der Genauigkeit aller Messungen einer einzelnen Probe, abhängig von der Laufzeit in Berechnungssekunden. Wir untersuchten mit einer einzigen Stichprobe einen möglichen Einfluss der Laufzeit einer FSM. Die meisten Messungen zeigen Werte unter 1 Sekunde, aber einige haben, verglichen mit der Vielzahl der Messungen, eine lange Laufzeit über 250 Sekunden. Der Pearson-Korrelationskoeffizient zeigt eine winzige positive Korrelation (0,01) zwischen Laufzeit der Feature-Auswahl und Genauigkeit (p-Wert < 0,01). Obwohl das Konfidenzintervall [0.02, 0.15] die positive Beziehung unterstützt, können wir aufgrund dieser sehr kleinen Beziehung keinen tragfähigen Einfluss der Laufzeit auf die Genauigkeit sehen.

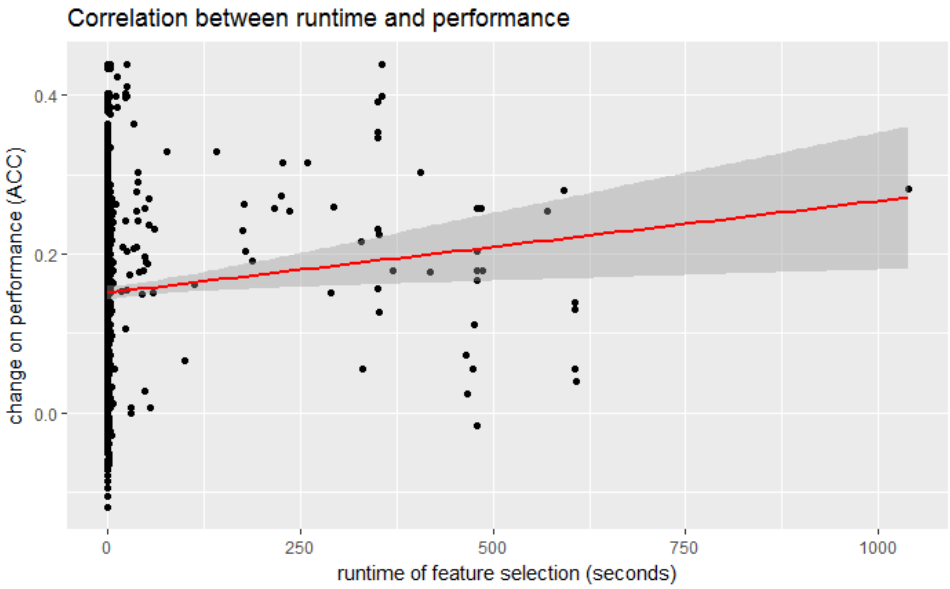


Abbildung 15: Scatterplot der FSM Laufzeitperformanz in einem einzelnen Setup; die Darstellung zeigt, dass die Klassifikationsgüte kaum einen Zusammenhang mit der Berechnungsdauer hat

Teil B: Automatisierte Merkmalsextraktion aus Zeitreihendaten

In Aufgabe 3.4 stellten wir fest, dass Wochen- und Jahreszyklen in den Stromverbrauchslastspuren existieren und durch schnelle diskrete Fourier-Transformation erkannt werden können; Spitzen und Statistiken (Varianz, Quantile, usw.) im Periodogramm können verwendet werden, um Merkmale automatisch aus Zeitreihendaten in unserer konsekutiven Arbeit zu berechnen.

k-Means Clustering wurde verwendet, um zwei Cluster mit hohem und niedrigem Stromverbrauch zu identifizieren. In Aufgabe 3.4 testen wir eine schnelle diskrete Fourier-Transformation, um Periodizitäten in Zeitreihen zu finden. In dieser Aufgabe testeten wir, inwieweit diese Technik auf automatisierte Ableitungsmerkmale angewendet werden kann. Allerdings konnten wir keine aussagekräftigen Ergebnisse erzielen, indem wir diese Technik auf die vorhandenen Daten mit 15-minütiger oder gröberer Datengranularität anwandten und uns daher für eine empirische Merkmalsextraktion entschieden.

In Kombination mit Aufgabe 4.9 und 4.10 entwickelten wir einen Algorithmus, der niedrige und hohe Verbrauchszeiten durch unbeaufsichtigtes maschinelles Lernen (Clustering) erkennt. Das Ergebnis kann für die Merkmalsextraktion verwendet werden. In diesem Ansatz identifizierten wir Zeiten, in denen der Haushalt länger (mehr als drei Tage) unbesetzt war und berechneten Merkmale anhand der Dauer der Abwesenheit. Aufgrund des Fehlens von Daten für die überwachte Auswertung (Wissen über die Belegungszeiten der Bewohner) berücksichtigten wir Visualisierungen von Lastprofilen von 12 Wochen und entwickelten einen k-Means-Clustering-basierten Algorithmus, der Zeitspannen identifiziert, die zu Clustern mit niedrigem und hohem Verbrauch gehören.

Diese Zeitspannen wurden bei der Merkmalsextraktion verwendet, um die folgenden Statistiken zu berechnen:

- Durchschnittsverbrauch in verbrauchsarmen Zeiten
- Durchschnittsverbrauch in verbrauchsintensiven Zeiten
- Verhältnis des Durchschnittsverbrauchs in High / Low Consumer Clustern
- Durchschnittliche Länge der Cluster-Zeitspanne mit niedrigem Verbrauch
- Durchschnittliche Länge der Zeitspanne der Cluster mit hohem Verbrauch

Feature-Extraktion aus verbrauchsrelevanten Zusatzdaten (Aufgabe 3.8)

Wir identifizierten sechs Kategorien von externen Zusatzdatenquellen, die wir für die räumlich-semantische prädiktive Modellierung unserer vier Anwendungsfälle (Haushaltsklassifizierung, Churn, Cross-Selling, Up-Selling) verwendeten. Aus den Datenquellen extrahierten wir relevante Merkmale.

Kategorie A (Wetter- und Klimadaten): Wetterinformationen können auf zwei Arten in die Haushaltsklassifizierung mit Smart Meter Daten einbezogen werden:

1. Normalisierung der Verbrauchslastspuren durch den Wettereinfluss - dieser Ansatz ist beschrieben in Sodenkamp, Kozlovskiy, und Staake (2016). Bei diesem Ansatz sind keine Standortdaten erforderlich und sie wurden für den Datensatz B verwendet.
2. Extraktion von Merkmalen aus der Korrelation von Lastgängen und Wettervariablen. Wir entwickelten diesen Ansatz für 15-minütige Smart-Meter-Daten (Hopf, Sodenkamp, und Staake 2018) und übernahmen die Features für 30-minütige Smart-Meter-Daten und Tagesverbrauchsdaten für Strom, Gas und Wasser.

Kategorie B (Kundenstammdaten): Kundenstammdaten enthalten Informationen, die für die Vorhersage einen wertvollen Beitrag leisten Die Tabellen 6, 7 und 8 zeigen die Merkmale, die aus den Kundenstammdaten abgeleitet sind.

Tabelle 6: Kundenstammdaten als Features für die Vorhersage von Kundenwechsel, Cross-selling und Up-selling

Feature	**Comment**
Gender (Male/Female/Unknown)	From historical data we know that men tend to churn faster and are more easily persuaded to take a cross/up sell offer
Is.Group	Groups such as home owner groups are usually more invested in choosing optimal contracts as they represent a large group of consumers
Has.Title	Consumer with academic titles are more educated and as such tend to make more rational choices concerning their contracts. In addition, education is correlated with a higher interest in green energy and as such can be an especially important feature for up-selling of green energy.
Has.Phone	Customers that have provided the utility with their phone number tend to be more invested in the relationship and as such are less likely to churn and more willing to take new offers
Credit.Or.AdditionalPayment	Customers that ended their last payment term with credit as opposed to be being hit with an additional payment generally have a more positive attitude towards the utility
Has.Traditional.EmailProvider	Customers with traditional email providers such as aol.com and t-online are more conservative and less likely to churn or to be interested in up and cross selling
Has.Uncommon.EmailProvider	Customers with customized email providers (such as a family domain) are less conservative and more technologicaly engaged. They are more likely to churn but also more interested in cross and up selling
Is.PLZDifferent	Contracts with a different billing address are less likely to churn and respond to cross/up-selling as they represent contracts that are managed for someone else
FirstProduct	Customers that have previously switched tariff within the same utility are less likely to churn and more likely to react to cross/up-sell offers
Has.Suspensions	Customers with suspensions are more likely to churn and less likely to take on cross/up-sell offers
Has.Complaints	Customers with complaints are more likely to churn and less likely to take on cross/up-sell offers
Has.Contacts	Customers with contracts are more likely to churn and more likely to take on cross/up-sell offers
Normalized.Consumption	Customers with high consumption are more likely to churn as well as more likely to take on cross-/upselling offers.
Neighborhood.Consumption.Ratio	Customers that have high consumption compared to their neighbors are more likely to churn as well as more likely to accept cross-/upselling offers.
Age.In.Years	Customers under 30 as well as customers above 70 are less likely to churn. Depending on the nature of the cross-/upselling offer different age groups may have different responses.

Tabelle 7: Kundenstammdaten als Features für die Vorhersage von Kundenwechsel

Feature	Comment
Provider.Change	If available, a known provide change is optimal training signal for churn.
Contract.Start	Because Churn follows a time pattern beginning with the contract start, we need the contract start date to normalize all event times based on when the contract began.
Contract.End	If provider change is not supplied, we may be able to deduct churn from the Contract end date.
Metering.Point.ID	When deducting churn based on the contract end, we can use the metering point ID to distinguish between moving and churning. If a new customer uses the same ID, the previous one probably moved out.

Tabelle 8: Kundenstammdaten als Features für die Vorhersage von Cross-selling und Up-selling

Feature	Comment
Tariff.History	In the absence of a controlled experiment, we can learn up- or cross-sell potential based on which customers previously switched to the tariff we are offering now.
Has.Been.Contacted	If the customer has previously not responded to a similar offer, he is unlikely to do so now.

Kategorie C (Statistische Daten): Die Statistischen Bundesämter veröffentlichen viele amtliche Statistiken auf verschiedenen geografischen Aggregationsebenen. Diese Daten können in der Datenanalyse für Strom-, Gas- oder Wasserdatenanalyse verwendet werden. Wir fanden die verfügbaren statistischen Datensätze der deutschen, schweizerischen und europäischen Behörden und beschreiben die verfügbaren Datenformate. Wir weisen auch darauf hin, welche statistischen Datensätze für das Projekt als relevant angesehen werden und warum sie als relevant angesehen werden. Darüber hinaus erklären wir, wie die verschiedenen statistischen Ämter Identifikatoren / Kodierungsschemata bereitstellen, um die Statistiken nach geografischen Regionen zu lokalisieren. Wir nennen sie Ortskennung.

Statistische Daten von deutschen Behörden:

Alle statistischen Daten für die deutschen Kommunen wurden aus dem GENESIS-Datenbankportal der „Statistischen Ämter des Bundes und der Länder“[1], auch bekannt als Regionaldatenbank Deutschland, bezogen. GENISIS ist ein statistisches Informationssystem, das vom Statistischen Bundesamt und den Statistischen Ämtern der Länder gemeinsam entwickelt wurde. Die Agentur liefert statistische Informationen mit einer sehr hohen Granularität, wobei die unterste Stufe die Gemeinden sind. Die verwendeten Statistiken sind in der Tabelle 10 aufgeführt.

Die verfügbaren Datenformate sind auf .csv,.xls und.html beschränkt. Diese Datenformate sind für den Umgang mit multidimensionalen Daten nicht sehr geeignet. Darüber hinaus sind die Datentabellen in ihrer Anzahl begrenzt, so dass Tabellen nur teilweise heruntergeladen werden können.

Die Ortskennung für die deutschen Gemeinden ist der RS (Regionalschlüssel), der eine 12-stellige Zahl ist. Sie ist in fünf Teile gegliedert: Die Ziffern 1-2 kennzeichnen das Bundesland, die Ziffer 3 gibt die Verwaltungsregion an, die Ziffern 4 und 5 repräsentieren das Land, die Ziffern 6-9 kennzeichnen den Gemeindeverband und die Ziffern 10-12 kennzeichnen die Gemeinde selbst. Der RS wurde 2009 eingeführt und ersetzt derzeit den AGS ("Amtlicher Gemeindeschlüssel"), der im Grunde genommen der gleiche Schlüssel ist, aber die vier Ziffern für Gemeindeverbände fehlen (Helmcke 2008).

Statistische Daten von Schweizer Behörden:

[1] https://www.regionalstatistik.de/genesis/online, letzter Zugriff am 13.05.16

Die statistischen Daten für die Schweiz wurden vom Datenbankportal STAT-TAB[2] des Bundesamtes für Statistik BFS heruntergeladen. Die umfangreiche Datenbank bietet viele Statistiken und ist zudem kostenlos. Die Daten haben eine hohe Granularität und sind auch für die kommunale Ebene verfügbar. Die verwendeten Statistiken sind in der Tabelle 11 aufgeführt.

Die von Stat-Tab angebotenen Datenformate sind PC-Axis (.px), Excel (.xlsx), tabulatorgetrennter Text (.tsv) und kommagetrennter Text (.csv). Das PC-Achsenformat ist für diese Aufgabe am bequemsten, da es multidimensionale Daten besser verarbeiten kann als die anderen Formate.

STAT-TAB verwendet die BFS-Nummer als Ortskennung. Die Nummer ist ein Zusammenschluss der Kantonsnummer, der Bezirksnummer und der Gemeindenummer. Sie wird vom Statistischen Bundesamt festgelegt und kann ein bis vier Stellen lang sein. Die Kantone sind so sortiert, wie sie im ersten Artikel der Schweizer Verfassung erscheinen. Die Bezirke sind nach Kantonen und die Gemeinden innerhalb der Bezirke alphabetisch sortiert (BFS 2006).

Statistische Daten aus der Europäischen Union:

Eurostat[3] ist die Quelle für statistische Daten für die Europäische Union. Es ist das statistische Amt der Europäischen Union. Die statistischen Daten werden über das ESS (Europäisches Statistisches System) aggregiert, das ist die Zusammenarbeit zwischen der Kommission (Eurostat), den nationalen statistischen Ämtern und anderen nationalen statistischen Behörden. Diese sammeln die Daten und erstellen Statistiken für die nationalen und EU-Zwecke. Die verfügbare Datenerhebung ist sehr umfangreich. Die meisten Daten basieren auf nationaler Ebene, während eine kleinere Datenmenge auf den NUTS-Regionen (Nomenclature des unités territoriales statistiques) basiert (Europäische Kommission, 2015). NUTS umfasst drei mögliche territoriale Klassifizierungen: NUTS 1, NUTS 2, NUTS 3. Für jede Klassifizierung müssen die Regionen innerhalb einer bestimmten Bevölkerungsgrenze liegen, um die Klassen einheitlicher zu gestalten. Die NUTS-1-Regionen beispielsweise müssen zwischen drei und sieben Millionen Einwohner haben und werden in der Regel durch die Staaten/Provinzen in jedem europäischen Land vertreten. Für einige Länder (z.B. Österreich, Niederlande oder Italien) sind die NUTS-1-Regionen eine Kombination aus mehreren Staaten oder Regionen. NUTS 2 und NUTS 3 sind dann kleinere Abteilungen innerhalb der NUTS 1-Regionen, in denen NUTS 3 in der Regel durch Länder innerhalb jedes europäischen Landes vertreten ist. Die verwendeten Statistiken sind in der Tabelle 12 aufgeführt.

Für die europäische statistische Datenbank ist die Anzahl der verfügbaren Datenformate groß. Neben den Standardformaten wie Excel, TSV, CSV, CSV, PDF oder HTML sind auch komplexe Formate wie PC-Axis, SPSS oder SDMX verfügbar. Aus Gründen der Kongruenz war das PC-Axis-Format für diese Implementierung die erste Wahl. Es war jedoch nicht möglich, die Dateien im PC-Axis-Format in das R-Skript zu laden, was wahrscheinlich auf Codierungsprobleme zurückzuführen ist. Weitere Probleme traten mit dem SPSS-Format auf. Deshalb wurde SDMX gewählt, da es keine Fehler hervorgerufen hat.

Die letzte NUTS-Klassifikation wurde im Januar 2015 veröffentlicht. Sie basiert auf der Verordnung Nr. 1059/2003 des Europäischen Parlaments und des Rates zur „Schaffung einer gemeinsamen Klassifikation der Gebietseinheiten für die Statistik“ (European Parliament 2003), die seit ihrem Erlass im Jahr 2003 mehrfach überarbeitet wurde. Vier verschiedene Teile bilden die NUTS-Nummern. Der erste Teil der Nummer (NUTS-0-region) ist eine zweistellige Kombination, die ein Land (z.B. DE für Deutschland) durch die internationale Norm ISO-3166 eindeutig identifiziert. Alle drei folgenden Teile fügen dem Ländercode je nach Region jeweils ein Zeichen hinzu. Das folgende Beispiel für die Stadt Leeds im Vereinigten Königreich stammt aus DESTATIS: NUTS Ebene 0 -> UK, NUTS Ebene 1 -> UKE, NUTS Ebene 2 -> UKE4, Ebene 3 -> UKE42[4].

Eurostat veröffentlicht zahlreiche statistische Daten, aber die meisten Daten beobachten nur Statistiken auf nationaler Ebene, was für unseren Zweck nicht angemessen ist. Aus diesem Grund beobachteten wir nur

[2] https://www.pxweb.bfs.admin.ch/default.aspx?px_language=de, letzter Zugriff am 13.05.16

[3] http://ec.europa.eu/eurostat/data/database, letzter Zugriff am 23.05.16

[4] https://www.destatis.de/Europa/EN/Methods/Classifications/OverviewClassification_NUTS.html, letzter Zugriff am 23.05.16

Daten aus der Kategorie „Allgemeine und regionale Statistiken - Regionalstatistiken nach NUTS-Klassifikation".

Extrahieren von statistischen Daten für Haushaltsstandorte: Abbildung 16 veranschaulicht das Verfahren zum Erhalten einer statistischen ID für Haushaltsadressen. In der rechten und linken oberen Ecke befinden sich die beiden möglichen Eingabetypen für das Skript. Die rechte Seite ist eine Liste von Koordinaten, die einer bestimmten ID zugeordnet sind, z.B. Kundennummer. Auf der linken Seite befindet sich die Eingabe einer Tabelle mit Kundendaten, die Adressen enthält. Aus diesen Eingabetabellen werden die benötigten Informationen zur weiteren Verarbeitung extrahiert. Das Verfahren zum Auffinden der statistischen IDs muss einen ID-Typ für das zu behandelnde Land haben (z.B. CH-BFS für die Schweiz). Nach Abschluss des Prozesses wird die ursprüngliche Eingabetabelle mit einer neuen Spalte für die statistische ID zurückgegeben.

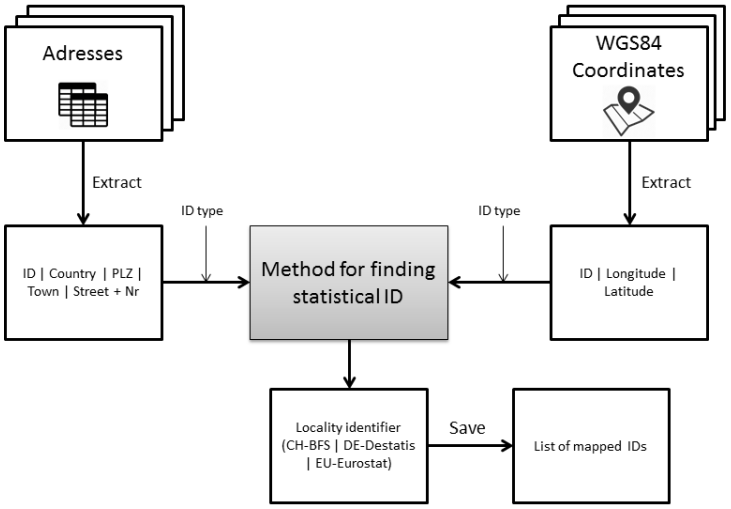


Abbildung 16: Verfahren zur Bestimmung des statistischen Verknüpfungsmerkmals eines Haushalts

Kategorie D (Geografische Informationen): Um die Adressdaten eines Haushalts zu verwenden, riefen wir geografische Informationen aus zwei beliebten VGI-Projekten ab: OpenStreetMap und GeoNames.org. Aus beiden Datenquellen leiteten wir 66 geografische Merkmale ab, die wir in vier Kategorien einteilten.

1) *Topologische Merkmale:* Beschreibung der Struktur und der Beziehungen zwischen einem Haushalt und den geografischen Nachbarn (z.B. Länge, Breite, Anzahl der geografischen Objekte in der Nachbarschaft, Entfernung zum Stadtzentrum, Fläche und Form der Polygone).
2) *Sehenswürdigkeiten und Points of Interest:* Bedeutung eines Objekts im räumlichen Kontext, in dem es erscheint (Frequenz, Entfernung und andere Maßnahmen zu Sehenswürdigkeiten, öffentlichen Einrichtungen, Haltestellen des öffentlichen Verkehrs, Sport- oder Freizeiteinrichtungen usw.).
3) *Merkmale über Gebäude* (Mittelwert, Varianz und Entfernung zu Gebäuden, Art der Gebäude in der Umgebung usw.)
4) *Merkmale der Landnutzung* (haushaltsnahes Nutzungsgebiet, Flächenverteilung in verschiedenen Landnutzungsarten, usw.)

Die definierten Merkmale basieren nicht nur auf den möglichen verfügbaren Daten in den VGI-Plattformen, sondern auch auf räumlichen Landschaftsmetriken in der Geoinformatik (Baskent und Jordan 1995; Gustafson 1998). Die Berechnung der geografischen Merkmale auf der Grundlage von OSM-Daten basiert auf der Definition des Kartenausschnitts, der ein rechteckiges Kästchen von einer zu spezifizierenden Größe ist. Die Bestimmung der richtigen Größe dieses Begrenzungsrahmens erwies sich bei dieser Arbeit als keine leichte Aufgabe, da das Rechteck frei wählbar ist. In den Geographie- und Sozialwissenschaften wird dieses Problem auf das Modifiable Area Unit Problem verwiesen (Briant, Combes, und Lafourcade 2010; Dietz 2002; Fotheringham und Wong 1991) und die Arbeiten zeigen, dass es keine feste Definition der korrekten Rahmengröße gibt. Für diese Arbeit betrachten wir drei Größen des Begrenzungsrahmens: 300m^2, 500m^2 und 1000m^2. Die größte Box ist für die Feature-Berechnung am besten geeignet und in unserer weiteren Forschung werden wir uns auf die Bestimmung der richtigen Größe der Begrenzungsbox konzentrieren.

Tabelle 9: Statistiken des deutschen Statistischen Bundesamtes und der statistischen Ämter der Länder und deren Verwendung im Projekt

Statistics type	**Used**	**Reason**
Territory		
Territory of municipalities		Has not been used, because it is not relevant for household classification
Number of municipalities		Has not been used, because it is not relevant for household classification
Population		
Population by gender and age groups on municipality level	X	The age groups of the population in a municipality can give indicators if age has an influence on energy consumption
Immigration and emigration by gender and age groups on municipality level		
Marriages, births, deaths		Not relevant for the purpose
		All other statistics are either similar to the two above mentioned or the areal granularity was to low
Working population		
Employed population which is subject to social insurance contributions		Has not been considered, because the other countries do not have comparable statistics
Unemployed population		Has not been considered, because it is not relevant for household classification
Elections		
		The category has not been considered, because it is not relevant for the purpose
Education and culture		
Schools of general education: Graduates by graduation type and year on district level	X	The educational level may can have influence on the energy consumption and thus is considered
		The other statistics are similar or not relevant
Social security		
Statistics of Recipient of social benefits		All these statistics are not relevant for household classification and/or do not have the right areal granularity
Statistics of Nursing facilities and nursing benefits		All these statistics are not relevant for household classification and/or do not have the right areal granularity
Statistics of child and teenager benefits		All these statistics are not relevant for household classification and/or do not have the right areal granularity
Health		
Basic data of hospitals	X	Number of available beds in hospitals may represent the development of the area
Basic data of rehabilitation centers		Has not been used, because other countries do not have similar statistics
Constructing and Housing		
Statistics of building permissions		Has not been used, because the other countries do not have similar statistics
Statistics of building completion		Has not been used, because the data is not relevant to household classification
Apartments in buildings with housing room by number of rooms	X	The number of rooms can give some indicators for the energy consumption

Buildings with housing space by building type, building year and type of heating	X	Very important for household classification. The building category and the age of the building can give conclusion on the energy consumption (e.g. heating)
Environment		
Statistics of waste disposal and dangerous waste		Not relevant for household classification
Statistics of water supply, purification plant, sewage		Not relevant for household classification
Statistics of land use		Has not been used, because the same category has not been used for Switzerland
Economic sectors		
Agriculture, forestry, fisheries, mining, tourism, transport		These category has been omitted completely as they do not have direct impact on the private residential sector
Businesses and craft		
Workplaces and employees by municipality, business sector and size	X	Has been used, because the number of workplaces and employed people can give conclusion on the development of an area which has an influence on the energy consumption
Statistics of trade registry		Have not been used, because the other countries do not have comparable statistics
Statistics of bankruptcy		Have not been used, as they are not relevant to household classification
Statistics of craft enterprises		Have not been used, as they are not relevant to household classification
Prices		
Statistics of buying prices for building land		Would have been interesting, but the other countries do not have comparable statistics
Public finances		
		The category has not been considered, since the comparable category from Switzerland has also not been considered
Public administration staff		
		The category has not been considered, as it is not relevant to the household classification
Economic total account		
GDP by economical sector by district	X	Has been used, because the GDP can give evidence about the development of a region which can give conclusion about the energy consumption

Tabelle 10: Statistiken des Schweizer Bundesamtes für Statistik und deren Verwendung im Projekt

Statistic type	**Used**	**Reason**
Population		
Permanent and non-permanent resident population by institutional units, sex, marital status and age class	X	The resident population of an institutional area accompanying the gender and age class can give interesting information on how age, gender and number of people in an area can have influence on the energy consumption
Migration of the permanent and non-permanent resident population by institutional units, gender, citizenship and migration type	X	
Statistics which have a low areal granularity (canton, country)		These statistics have not been included, because the granularity has to be on the level of municipalities
Live births, deaths, divorces and marriages		Have not been included, because they are not relevant for energy consumption
Territory, Environment		
		The category has not been considered, because the statistics are either not based on municipalities or the information such as land

		use (grass, agriculture, etc.) is too general to be used for house classification
Work and income		
		The category has not been considered, because the all statistics are not based on municipalities
Industry and services		
Workplaces and employees by municipality, business sector and size	X	Has been used, because the number of workplaces and employed people can give conclusion on the development of an area which has an influence on the energy consumption
New founded businesses		Was not considered because the number of new businesses is most of the time 0
		All other statistics areal granularity is to low
Agriculture and forestry		
		The category has not been considered, because the statistics are not based on municipalities and the information is not relevant for energy consumption of private households
Energy		
		The category would be interesting for the purpose, but the statistics are all on country level
Construction and Housing		
Apartments by institutional units, building category, number of rooms and building period	X	Very important for household classification. The number of rooms, the building category and the age of the building can give conclusion on the energy consumption (e.g. heating)
Building by institutional units, building category and building period	X	The same reasons apply as for the apartments
Energy: Housing by Canton, building category, type of heating, hot water, energy source and period of construction	X	Although the areal granularity is limited to cantons, this data can be used, because the heating type statistics for other countries are on the same granularity level and it is an important category
Investments by municipality, type of the contracting entity, type of buildings, type of work and year		Would be interesting, but the other countries don't have comparable statistics
		All other statistics are either similar to the above mentioned or the areal granularity is to low
Tourism		
		All statistics are limited to hotels, which is not interesting for the purpose
Mobility and Transport		
		All statistics deal with number of certain vehicles or accidents, which is all not important for the purpose
Banking, Insurance		
		The category has not been considered, because the statistics are all not based on geographical areas
Social Security		
		The category has not been considered, because the statistics are all not based on geographical areas
Health		
		The category has not been considered, because the statistics are all concerned with cancer which is not interesting for house classification
Education		
Higher education entrance qualification by type of qualification, gender, canton of school and residential canton	X	Although, the statistics are on a cantonal level, the educational level may can have influence on the energy consumption and thus is considered

Number of students at different types of colleges or the graduations		Has not been used, because all these statistics are not based on areas
Staff at colleges		Has not been used, because all these statistics are not based on areas
Costs for colleges and education		Has not been used, because all these statistics are not based on areas
Culture – Media – Information society – Sports		
		The category has not been considered, because the statistics are not based on geographical areas
Politics		
		The category has not been considered, because the statistics deal with elections and referendums which are not important for the purpose
Public administration and finance		
		The category has not been considered, because the statistics deal with public finance on a non-geographical base
Crime and criminal justice		
		The category has not been considered, because it is not relevant to the theme
Sustainable Development		
		The category has not been considered, because the statistics are not based on geographical areas
Federal population census		
Residential population at the economic place of residence		The statistics have not been considered, because the other countries do not have comparable statistics

Tabelle 11: Statistiken des europäischen Statistikamtes (EUROSTAT) und deren Verwendung im Projekt

Statistic type	**Used**	**Rationale**
Regional agriculture statistics		
Agri-environmental indicators, Agriculture production, Structure of agriculture holdings		This category has been completely omitted, since the comparable statistics for Switzerland and Germany were not available or have not been used
Regional demographic statistics		
Population on 1 January by five year age group, sex and NUTS 3 region	X	The resident population of an institutional area accompanying the gender and age class can give interesting information on how age, gender and number of people in an area can have influence on the energy consumption
Population change, Area by NUTS 3 region, Population by NUTS 2		These statistics have not been used, since they bring no more value than the Population by NUTS 3 statistics or the area is not useful for household classification
Fertility and mortality		Have not been used, as they do not have an influence on the energy consumption
Census 2001 – Population by sex, citizenship and NUTS 3; Population by sex, age group current activity status; Total and active population by sex, age, employment status, residence on year prior to the census and NUTS 3; Employed persons aged 15 and over by NUTS 3		Not used, because it is not relevant at the moment. Maybe it is useful later
Employed persons by sex, age group, educational attainment level, occupation and NUTS 3; Population by sex, age group, educational attainment level, current activity status and NUTS 3 region		Not used, because it is not relevant at the moment. Maybe it is useful later
Population by sex, age group, household status and NUTS3; Population by sex, age group, size of		Not used, because it is not relevant at the moment. Maybe it is useful later

household and NUTS 3; Private households by composition, age group of children and NUTS 3		
Private households by composition, size and NUTS 3	X	
Dwelling by type of housing, building and NUTS 3	X	
Regional economic accounts (ESA2010)		
GDP at current market prices by NUTS 3 regions	X	Has been used, because the GDP can give evidence about the development of a region which can give conclusion about the energy consumption
GDP by NUTS 2, Average annual population to calculate regional GDP, Real growth rate by NUTS 2		These statistics have not been included since they either bring the same information or are on a lower granularity level
Employment rate by NUTS 2 or 3, Gross value added by branch, Compensation of employees		
Allocation of primary income accounts of households by NUTS 2, Income of households by NUTS 2, Secondary distribution of income accounts by households by NUTS 2		
Regional economic accounts (ESA95)		
		The category has been omitted since it considers the same themes as the ESA2010 category, but is older
Regional education statistics		
Number of students by level of education by NUTS 2, Number of students by age and NUTS 2, Education indicators by NUTS 2, Participation rate in education and training by NUTS 2		
Population aged 25-64 by educational attainment level, sex and NUTS 2 regions	X	Has been used, because the educational levels of a region may indicate the possible affinity for the interest of energy saving
Population aged 30-34 by educational level, sex and NUTS 2 regions		Not used, since it represents a smaller fraction of the statistics on line above
Early leavers from education by NUTS 2, Young people neither in employment nor in education		Have been omitted, because this data has no importance for household classification
Regional science and technology statistics		
Total intramural research and development expenditure by NUTS 2	X	Has been used, since the expenditures on research by have an influence on the surrounding area
Total R&D personnel and researchers by sectors of performance and NUTS 2		Has been omitted, because it does not differ a lot from the statistics one line above
Employment in technology and knowledge-intensive sectors by NUTS 2 or NUTS 1		Has been omitted, because the data is not XXXX
Human resources in Science and technology by NUTS 2 or NUTS 1		Is not relevant enough for the topic
Intellectual property rights		Is not relevant enough for the topic
Regional structural business statistics		
SBS data by NUTS 2 regions, Multiannual statistics for distributive trades by NUTS 2, SBS data by NUTS 2, Number of local units, persons employed and wages and salaries by NUTS 2, Multi yearly statistics by NUTS 2		All statistical data are specific to the branches, which is not appropriate for the purpose
Regional business demography		
Business demography by size class and NUTS 3 regions	X	Number of businesses and their sizes may give a conclusion on the development of a region which is interesting for household classification
Business demography and high growth by NACE and NUTS 3, Employer business demography by		Either bring the same information as the statistics one line above, are too specific or too far from the topic

size class and NUTS 3, Employer business demography by NACE and NUTS 3		
Regional health statistics		
Causes of death		Not relevant to the topic
Health personnel by NUTS 2		Too specific for the topic
Hospital beds by NUTS 2	X	May give a conclusion of the local development of a region
Hospital discharges, Hospital days of in-patients, Long-term care beds, Prevalence of disability		Not relevant to the topic
Regional tourism statistics		
		The tourism category has been not considered throughout the project
Regional transport statistics		
Road freight, Rail networks, Vehicle stocks, Maritime transport, Air transport, Railway transport		All the statistics are too far from the topic and are thus not considered
Regional labor market statistics		
Regional population and economically active population		The category has not been considered, because the statistics are all concerned with cancer which is not interesting for house classification
Regional employment/unemployment		May be possible to consider later
Regional labor market disparities		At the moment, too far from the topic
Regional job vacancy statistics		Not relevant to the topic
Regional structure of earnings (2006,2010)		Not used, since the granularity is too low (NUTS 1)
Regional labor costs statistics		
Labor costs survey 2000, 2004, 2008, 2012 (NUTS 1)		Have not been used, since the granularity is too low (NUTS 1)
Regional information society statistics		
Households with access to the internet at home (NUTS 1)	X	The category is very interesting for the topic, but has a low granularity - Still, it will be used
Households with broadband access, Individuals who have never used a computer, Individuals who used the internet for different reasons		Too specific for the topic
Individuals who ordered good or services over the internet for private use (NUTS 1)	X	Also very interesting, but has a low granularity
Regional environmental and energy statistics		
Regional waste statistics		Not relevant enough to the topic
Regional water statistics (RBD)		The granularity is too low
Energy production and final consumption by NUTS 2		Interesting for the topic, but the granularity is too low
Heating degree-days by NUTS 2		Interesting for the topic, but the granularity is too low
Regional poverty and social exclusion statistics		
People at risk of poverty by NUTS 2, At-risk-of-poverty rate by NUTS 2		Not considered at the moment and the granularity is low
People living in households with very low work intensity by NUTS2		Not considered at the moment and the granularity is low
Severe material deprivation rate by NUTS 2		Too specific for the topic
Regional crime statistics		
Crimes registered by the police by NUTS 3		The category has not been considered, because it is not relevant to the topic

Kategorie E (Feiertage und besondere Ereignisse): Für die Datenbereinigung und Datenaufbereitung werden Informationen über Feiertage und Schulferien in den Regionen der Datensätze benötigt. Wir Feiertage in der Schweiz und Deutschland und verwendeten Schulferienkalender nach den Regionen unserer Datensätze.

Für jeden Datensatz erstellten wir eine Liste von Feiertagen und besonderen Ereignissen, die während der Datenaufbereitung und Merkmalsextraktion zur Kennzeichnung solcher Tage automatisch verarbeitet werden kann. Wir entschieden uns, die Tage nicht auszuschließen, sondern zu markieren, da diese Informationen für spätere Datenanalyseaufgaben wertvoll sind.

Kategorie F (Semantische Daten): Unser Ziel war es, Schnittstellen zu sozialen Netzwerken (facebook.com, twitter.com, usw.), Feeds von Nachrichtenseiten und Suchmaschinen oder Echtzeit- oder Einfügungsinformationen zu implementieren. Nach der Implementierung von Schnittstellen zu exemplarischen Online-Plattformen (facebook.com und Google News) stellten wir fest, dass der Abgleich mit Haushalten in unseren Datensätzen aus zwei Gründen nicht möglich ist:

1) In sozialen Netzwerken sind die Standortdaten aus Datenschutzgründen nicht zugänglich.
2) Die Anzahl der Einträge in Online-Portalen ist nur in Großstädten groß genug. Daher können wir keinen maschinellen Lernalgorithmus trainieren, der Informationen für Haushalte außerhalb dieser Städte voraussagt.

Wir entschieden uns, uns auf statistische, Wetter- und geografische Datenquellen zu konzentrieren und den Funktionsumfang für diese Plattformen zu erweitern.

Feature-Angemessenheit für Haushaltseigenschaften (Aufgabe 3.13)

Die Eignung der Features für die Vorhersage einzelner Haushaltseigenschaften kann durch Korrelationsanalysen abgeschätzt werden. Die berechneten Koeffizienten werden in diesem Zusammenhang auch als Effektgrößenmessungen bezeichnet, da sie einen ersten Eindruck von der Relevanz von Merkmalen für die prädiktive Analyse vermitteln.

Wir betrachten die folgenden Fälle für diese Analyse:

- Vorhersage von Haushaltseigenschaften auf der Grundlage von jährlichen Stromverbrauchsdaten und freien geografischen Informationen
    - ‚Haushaltstyp' (Abbildung 17)
    - ‚Wohnbereich' (Abbildung 18)
    - ‚Anzahl der Einwohner' (Abbildung 19)
    - ‚Raumheizungsart' (Abbildung 20)
    - ‚Art der Warmwasserbereitung' (Abbildung 21)
- Vorhersage des "Heizalters" auf der Grundlage des jährlichen Gasverbrauchs und freier geografischer Informationen (Abbildung 22)
- Vorhersage der Kundenabwanderung basierend auf Kundenstammdaten, freien geografischen Informationen und Merkmalen von zwei Datenanbietern (Abbildungen 23 und 24)

Ergebnisse aus der Korrelationsanalyse zwischen Merkmalen und Eigenschaften:

- Die Kundenstammdaten und frei verfügbaren Daten enthalten gute Variablen zur Vorhersage der Kundeneigenschaften im Haushalt. Beispiele:
    - Energieverbrauchsmerkmale (z.B. Jahresmittelwert) haben einen hohen Zusammenhang zu allen Haushaltsobjekten (z.B. Haushaltstyp, Wohnfläche, Heizungsart).
    - Für ‚Haushaltstyp' und ‚Wohnbereich' weisen die Daten über die umliegenden Gebäude (Häufigkeit der Gebäude, mittlere Grundfläche, usw.), aber auch die Anzahl der vorhandenen Einträge in der Geodatenbank (numNodes, numWays, numRelations, numNodeTags, usw.) eine hohe Korrelation auf.
- Für die Vorhersage der Kundenabwanderung scheinen nur die Kundenstammdaten relevante Variablen für die Vorhersage zu enthalten.

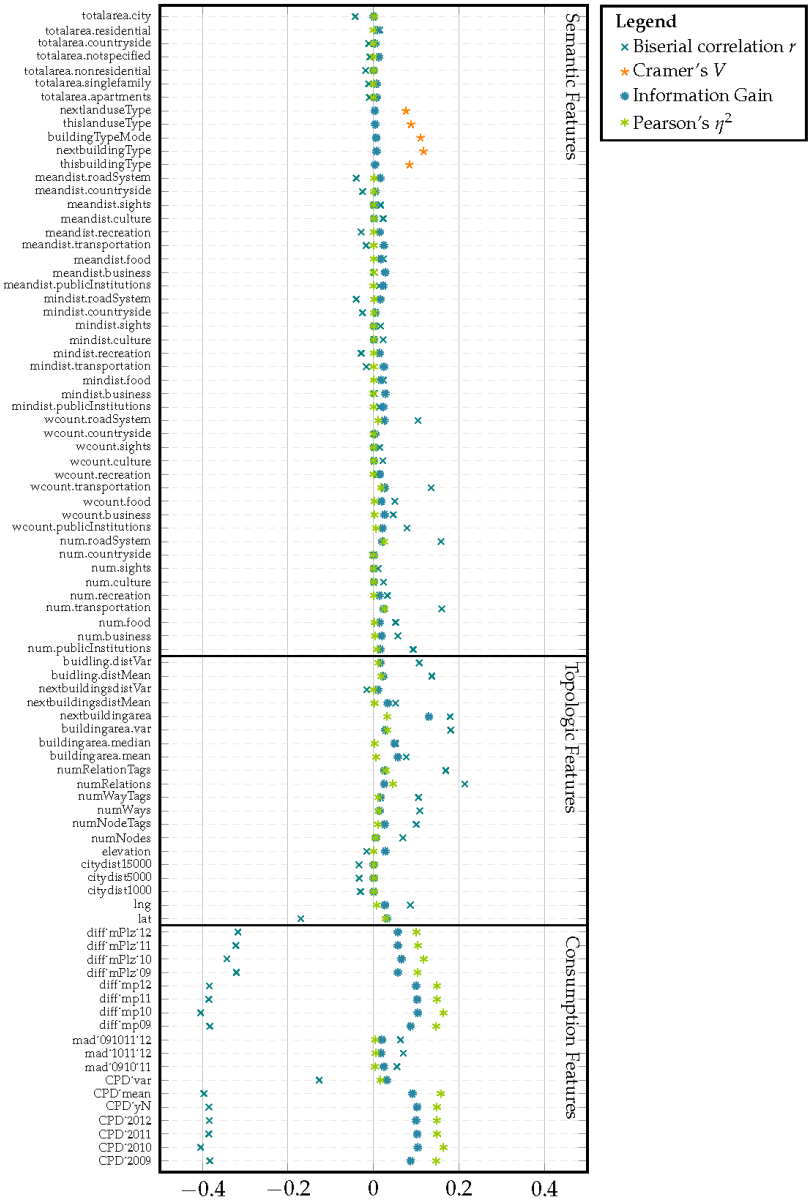


Abbildung 17: Vier Effektgrößenmaße für jährliche Verbrauchs-Features und geographische Features zur Vorhersage der Eigenschaft 'household type'

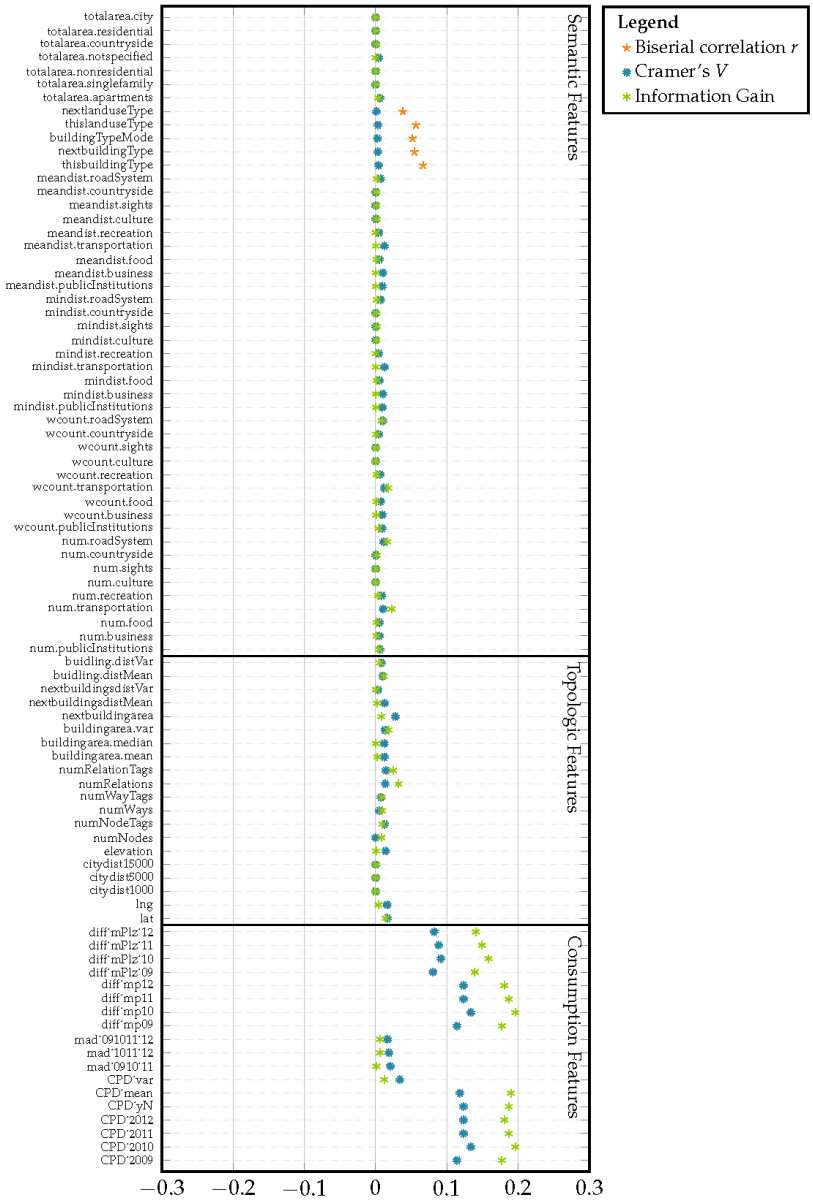


Abbildung 18: Drei Effektgrößenmaße für jährliche Verbrauchs-Features und geographische Features zur Vorhersage der Eigenschaft 'living area' (da die Eigenschaft drei Klassen hat, kann die siseriale Korrelation nicht berechnet werden)

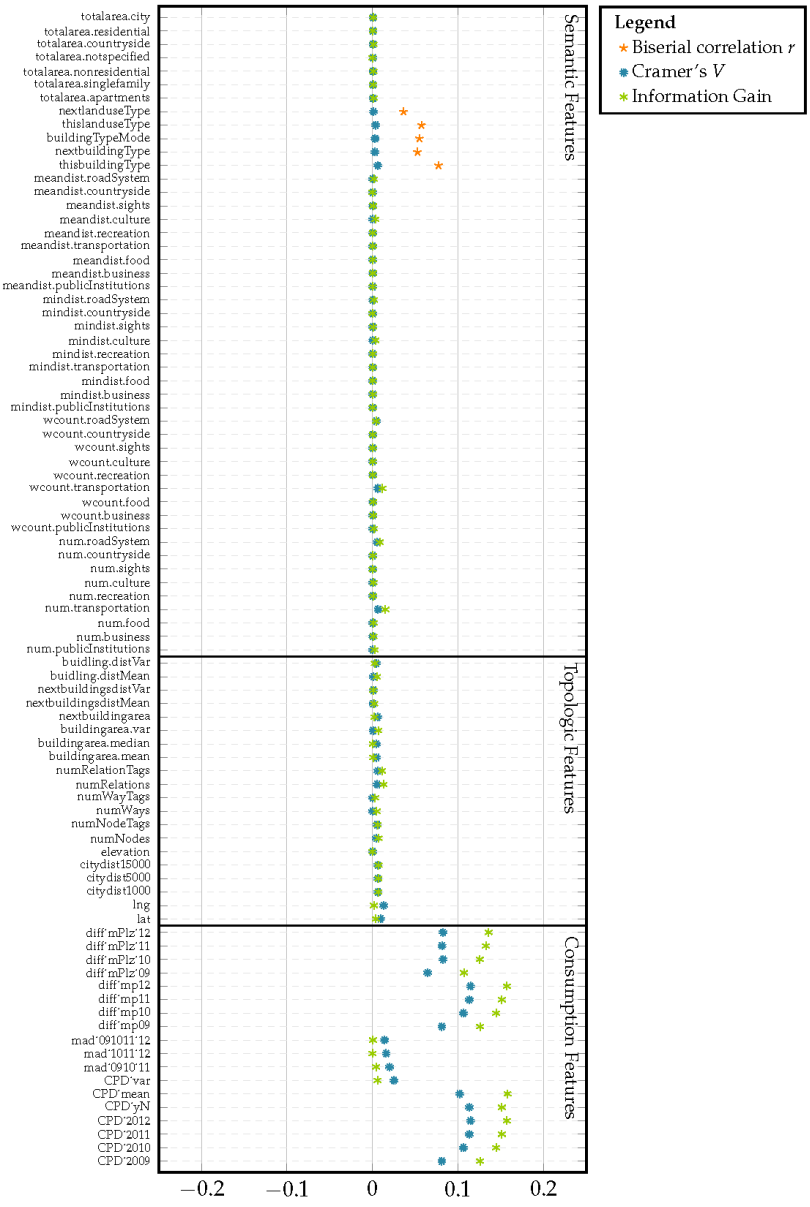


Abbildung 19: Drei Effektgrößenmaße für jährliche Verbrauchs-Features und geographische Features zur Vorhersage der Eigenschaft 'number of residents' (da die Eigenschaft drei Klassen hat, kann die biseriale Korrelation nicht berechnet werden)

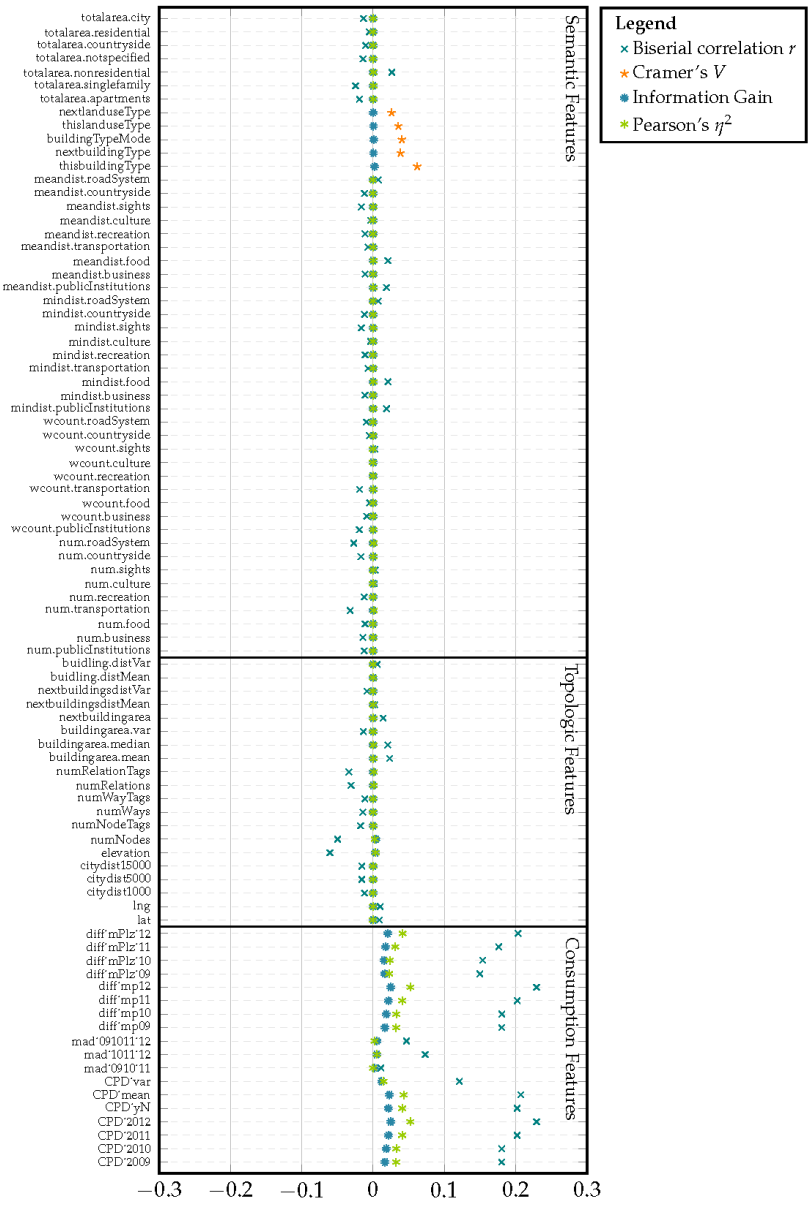


Abbildung 20: Vier Effektgrößenmaße für jährliche Verbrauchs-Features und geographische Features zur Vorhersage der Eigenschaft 'space heating type'

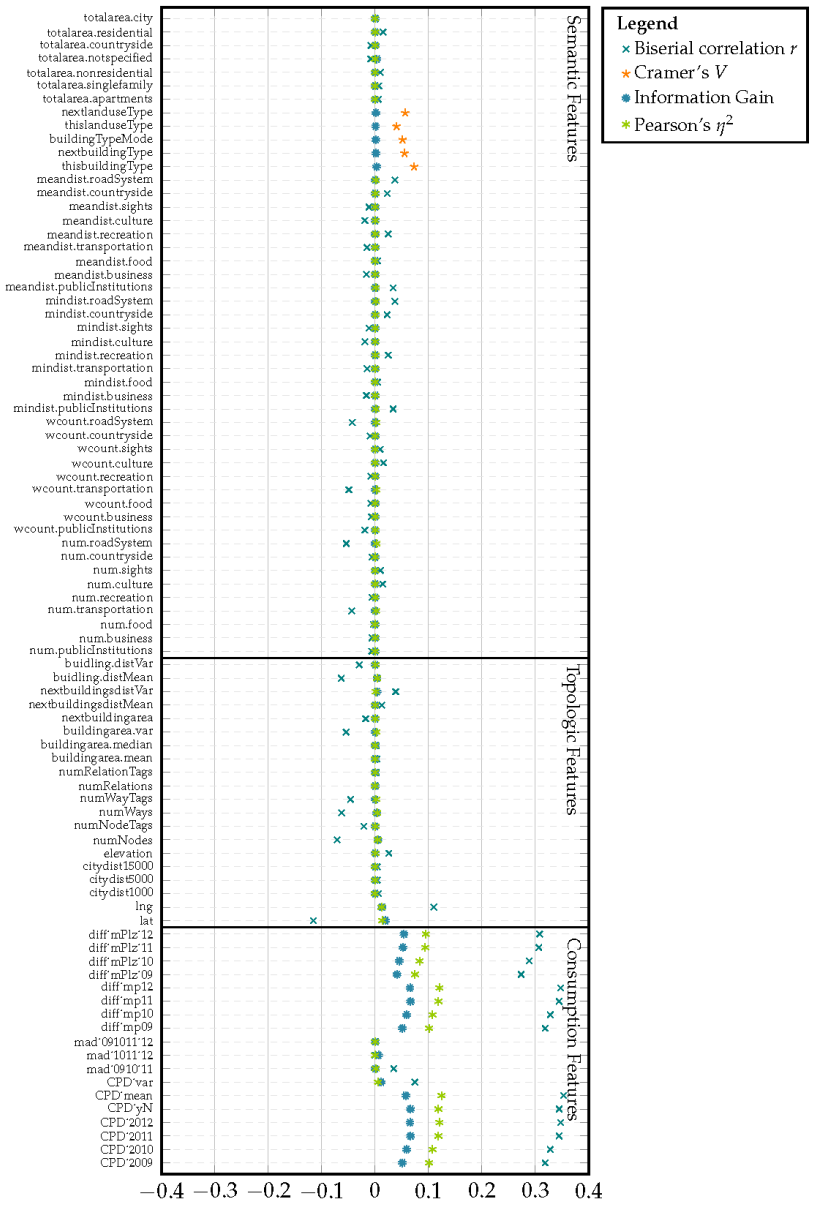


Abbildung 21: Vier Effektgrößenmaße für jährliche Verbrauchs-Features und geographische Features zur Vorhersage der Eigenschaft 'water heating type'

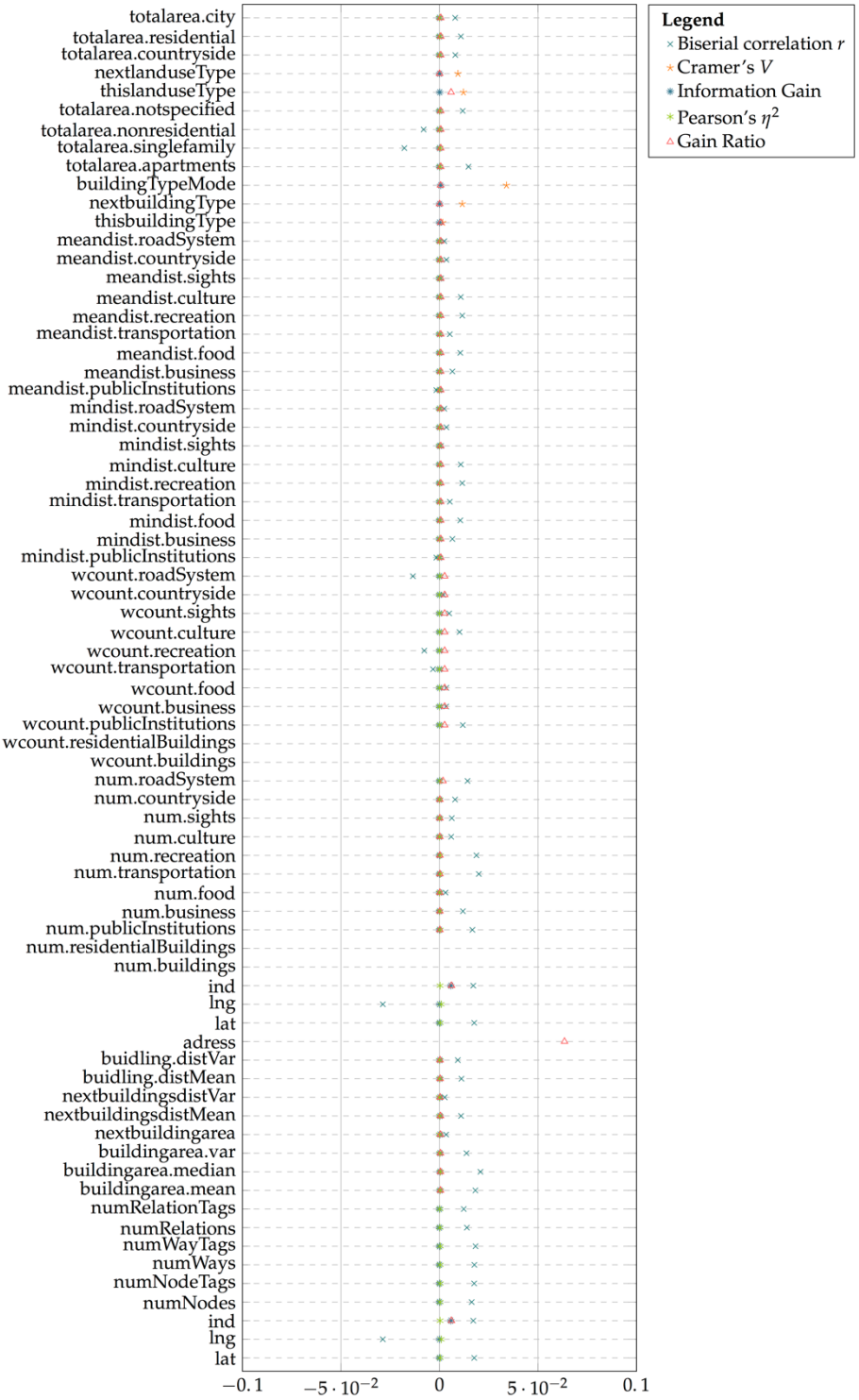
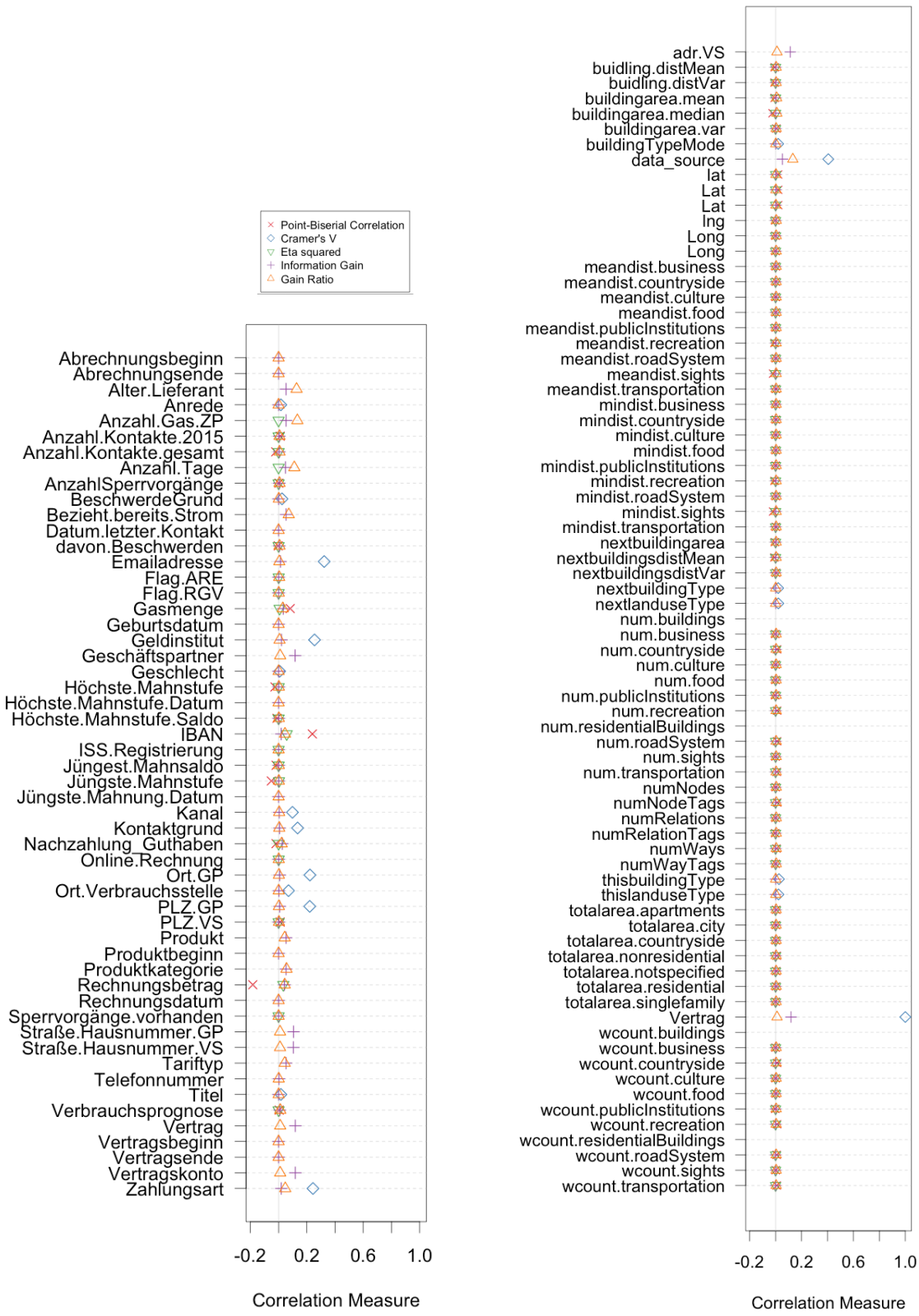


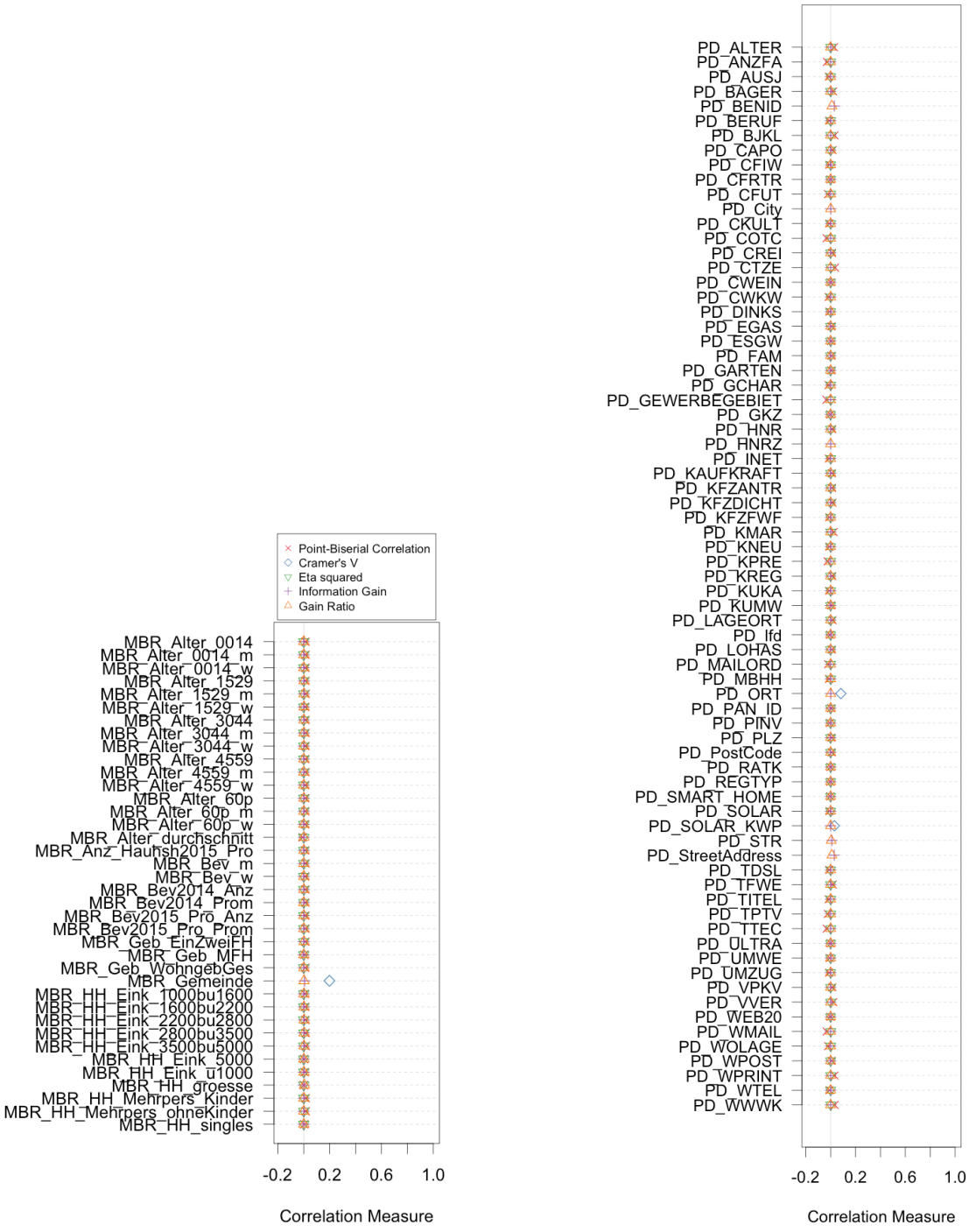
Abbildung 22: Fünf Effektgrößenmaße für geographische Features zur Vorhersage der Eigenschaft 'heating age'



(a) Dataset E customer core data

(b) Geographic features

Abbildung 23: Fünf Effektgrößenmaße für Kundenstammdaten und geographische Features zur Vorhersage von Kundenwechsel



(a) MBR purchased data

(b) Panaddress purchased data

Abbildung 24: Fünf Effektgrößenmaße für eingekaufte Daten (von zwei Datenlieferanten) für die Vorhersage von Kundenwechsel

AP 4: Haushaltsklassifikation

Auswahl und Abstimmung bestehender Klassifikatoren (Aufgabe 4.1)

Bekannte Lehrbücher und die aktuelle Machine-Learning-Literatur (Han, Pei, und Yan 2005; Hastie, Tibshirani, und Friedman 2009; Kotsiantis, Zaharakis, und Pintelas 2007; Mitchell 1997; Russell und Norvig 1995; Zaki und Meira Jr. 2014) wurden gesichtet und die wichtigsten Kategorien von maschinellen Lernalgorithmen untersucht. Die häufigsten Kategorien sind:

1. Logikbasierte Lernalgorithmen (insb. Regelbasierte Verfahren und Entscheidungsbäume)
2. Instanzbasierte Lernalgorithmen (einschließlich Nächste-Nachbarn-Klassifizierer)
3. Statistisches Lernen (z.B. Bayesscher Lerner, Naïve Bayes, Lineare Diskriminanzanalyse)
4. Support Vektor Maschines
5. Ensembles (z.B. Bagging, Boosting, Random Forest)
6. Künstliche neuronale Netze (ein- und mehrschichtiges Perzeptrone, neuronale Netze)

Alle Arten von Algorithmen haben unterschiedliche Fähigkeiten, mit den Herausforderungen des maschinellen Lernens umzugehen, z.B. hohe Dimensionalität, komplexe Entscheidungsgrenzen (d.h. nicht linear teilbare Klassen), imbalancierte Klassen.

Neben unseren Erfahrungen aus früheren Projekten (Forschungsprojekte an der Universität Bamberg und Pilotprojekte der BEN Energy AG) untersuchten wir verschiedene Arbeiten aus anderen Anwendungsbereichen, wie z.B. maschinelle Lernforschung (z.B. Fernández-Delgado u. a. 2014), Entscheidungsunterstützungssysteme (z.B. Merkert, Mueller, und Hubl 2015), Medizin (z.B. Emanet u. a. 2014) und Informationssysteme (z.B. Ram u. a. 2015), um bestehende Klassifikatoren auszuwählen und deren (Hyper-)parameter auszuwählen. Für jede Kategorie erläutern wir kurz das Funktionsprinzip der maschinellen Lernalgorithmen und diskutieren deren Eignung für die multidimensionale Haushaltsklassifizierung.

Logikbasierte Lernalgorithmen: Logikbasierte Lernalgorithmen sind eines der ersten Konzepte im maschinellen Lernen. Sie basieren auf einem Regelwerk oder Entscheidungsbäumen und haben eine geringe Verallgemeinerungsfähigkeit, sind empfindlich gegenüber Datenrauschen und fehlender Handhabung kontinuierlicher Merkmale. Es ist daher vernünftig, dass diese Methoden nicht für die Klassifizierung der Haushalte auf der Grundlage von Energieverbrauchsdaten unter Berücksichtigung der uns bekannten zugehörigen Arbeiten verwendet wurden. Selbst Mitchell (1997) weist darauf hin, dass es mehrere weitere Algorithmen gibt, die für Lernaufgaben besser geeignet sind als Algorithmen in dieser Kategorie. Daher setzen wir in diesem Projekt solche Methoden nicht ein.

Instanzbasierte Lernalgorithmen: Diese Verfahren verzögern die Verarbeitung von Lernbeispielen, bis neue Beispiele klassifiziert werden. Daher kann der Klassifizierungsschritt rechenintensiv sein. Als prominentester Vertreter dieser Kategorie wurde der k Nächste Nachbarn (kNN) in früheren Arbeiten am Haushalt verwendet (Beckel u. a. 2014; Beckel, Sadamori, und Santini 2013; Hopf u. a. 2014). Sie haben eine limitierte Fähigkeit, hochdimensionale Daten zu verarbeiten und sind daher für die multidimensionale Klassifizierung weniger geeignet. Für den Vergleich unserer Ergebnisse mit dem Stand der Technik entschieden wir uns jedoch für die Nutzung von Algorithmen dieser Kategorie.

Statistisches Lernen: Diese Algorithmen verwenden Konzepte der statistischen Analyse zur Klassifizierung. Beispiele für Algorithmen sind Logistische Regression, Naïve Bayes und Lineare Diskriminanzanalyse. Einige von ihnen wurden auch in früheren Arbeiten verwendet (Beckel u. a. 2014; Beckel, Sadamori, und Santini 2013; Sodenkamp u. a. 2017): Diese Algorithmen haben eine geringe Rechenkomplexität, aber ihre Generalisierungsleistung ist begrenzt. Die Algorithmen können komplexe und nichtlineare Entscheidungsgrenzen nur eingeschränkt verarbeiten. Außerdem sind sie anfällig für Rauschen und mehrere Eingangsvektoren.

Support Vektor Maschines: Um die Klassifizierungsprobleme bei komplexen Entscheidungsgrenzen mit nichtlinear trennbaren Klassen zu lösen, transformiert die Support Vector Machine den Eingangsvektor in den höher dimensionalen Raum und verwendet einen Soft-Margin zur Trennung von Klassen. Der SVM-Algorithmus in seiner einfachsten Form wurde in früheren Arbeiten zur Haushaltsklassifizierung verwendet.

Ensembles: Ensemble-Lernmethoden kombinieren mehrere maschinelle Lernmodelle mit dem Ziel, ein verbessertes zusammengesetztes Klassifizierungsmodell zu erstellen. Ensemble-Klassifikator-Vorhersagen basieren auf den Bewertungen der Basis-Klassifikatoren (Han, Kamber, und Pei 2012), die meist einfache Lernalgorithmen wie Entscheidungsbäume sind. Zwei Arten von Ensemble-Methoden sind bekannt:

- Bagging steht für Bootstrap-Aggregation und diese Algorithmen trainieren verschiedene Klassifikatoren unter Verwendung verschiedener Teilproben des Trainingssets, die alle neuen Beispiele klassifizieren. Mit diesem Ansatz wird die Varianz der Vorhersage reduziert und die Genauigkeit erhöht.
- Boosting trainiert auch mehrere Klassifizierungsmodelle und aggregiert die endgültige Vorhersage einschließlich eines Gewichts für jeden Basisklassifizierer, das sich aus seiner bewerteten Genauigkeit ergibt. Die folgenden Methoden sind in unserem Algorithmus implementiert:

Künstliche Neuronale Netze (ANN): ANN ist eine Klasse von maschinellen Lernalgorithmen, die Modelle aus Trainingsdaten mit Netzwerkstrukturen schätzen, die von biologischen neuronalen Netzwerken inspiriert sind. Jeder Knoten des Netzwerks (sogenannte "Neuronen") hat mehrere gewichtete Eingänge. Eine Aktivierungsfunktion wandelt die neuronengewichteten Eingänge in einen einzigen Ausgang um. Während des Lernprozesses werden die Gewichte der Eingaben für jedes Neuron gelernt. Je nach Struktur des Netzwerks, des Lernalgorithmus und der Aktivierungsfunktion gibt es viele Arten von ANN.

Implementierung bestehender Klassifikatoren:

Basierend auf der Überprüfung bestehender Klassifizierungsmethoden wählten wir einen umfassenden Methodensatz aus, den wir in diesem Projekt implementieren. Wir listen alle Algorithmen nacheinander auf und beschreiben deren Funktionsprinzip kurz:

k Nearest Neighbors (kNN): Dieser Algorithmus ordnet die Klasse unter Berücksichtigung der K-Trainingsinstanzen mit dem niedrigsten euklidischen Abstand zu dem zu klassifizierenden Beispiel zu. Wir verwendeten die Normalisierung aller Variableneingaben auf eine Codezone von [0;1], da der kNN-Klassifikator empfindlich auf die Bereiche der Eingangsgrößen reagiert (Han et al., 2012). Wir stützen uns auf die Umsetzung von (Wing et al., 2015). Die Parameter für die Abstimmung des kNN-Algorithmus sind in der Tabelle 12 aufgeführt.

Tabelle 12: Parameter des kNN Algorithmus

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

Lineare Diskriminanzanalyse (LDA): Das Ziel des LDA-Klassifikator ist es, eine Reihe von Features zu finden, die die Beispiele weitgehend zwischen den Klassen trennen.

Tabelle 13: Parameter des LDA Algorithmus

Parameter	**Description**	**Possible range**	**Considered range**
tol	A tolerance value; it will reject variables and linear combinations of unit-variance variables whose variance is less than tol^2.	Positive values	1.0e-3, 1.0e-4*, 1.0e-5
method	Estimators for mean and variance	‘moment’, ‘mle’, ‘mve’, ‘t’	
nu	Degree of freedom for method=’t’	Positive integers	1, 10, 100

Naïve Bayes (NB): Der Bayes'sche Klassifikator prognostiziert die Klassenzugehörigkeit basierend auf der Wahrscheinlichkeit, dass ein bestimmter Datenpunkt zu der Klasse gehört. Die für diese Vorhersage erforderlichen Wahrscheinlichkeiten werden mit Hilfe des Bayes-Satzes berechnet. In unserer Analyse verwendeten wir die Implementierung von (Meyer u. a. 2014).

Logistische Regression (LR): Diese statistische Technik wurde von Cox (1958) eingeführt und schätzt die Beziehung zwischen einer oder mehreren unabhängigen Variablen (in unserem Fall Merkmale) und einer der abhängigen Variablen (in unserem Fall eine Eigenschaft). Die abhängige Variable kann binomial (zwei Klassen) oder multinomial (mehr als zwei Klassen) sein.

Support Vector Machine (SVM): Der Ansatz wurde von (Vapnik und Vapnik 1998) entworfen. Der Algorithmus sucht nach einer Hyperebene im Vektorraum, die alle Trainingsbeispiele mit einem maximalen Abstand voneinander trennt. Bei nicht trennbaren Trainingsdaten wird eine Kernelfunktion verwendet, die den Trainingsvektor in eine höhere Dimension transformiert. Wir verwenden die SVM-Implementierung von (Meyer u. a. 2014). Die Parameter für die Abstimmung des SVM-Algorithmus sind in der Tabelle 14 aufgeführt. Leitfäden zur SVM-Parametereinstellung werden von Hsu et al. (2010) und Lin (2003) gegeben.

Tabelle 14: Parameter des SVM Algorithmus

Parameter	**Possible range**	**Considered range**
kernel	linear, radial, polynomial, sigmoid	
eps	Decimal	0.1
cost	Decimal	2E^2(-5, ..., 10)
deg	Decimal	2, 3, 4, 5
coef	Decimal	0, 1, 5, 10, 100

Künstliche Neuronale Netze (ANN): Es gibt keinen einzelnen Lernalgorithmus für neuronale Netzwerke, sondern eine Familie von Algorithmen, die entwickelt wurden, um überwachte Lernmodelle mit ANNs zu trainieren. Wir verwendeten die Implementierung von Fritsch und Guenther (2016), die das Training neuronaler Netze mit Backpropagation, widerstandsfähige Rückführung mit (Riedmiller 1994) oder ohne Weight Backtracking (Riedmiller und Braun 1993) oder die modifizierte global konvergente Version von Anastasiadis et al. (2005) abdeckt. Die Tabelle 15 enthält die möglichen Parameter für die ANN-Abstimmung. Wir folgen Jain et al.'s (1996) und Lawrence et al.'s (1997) Leitfäden zur Einrichtung und Einstellung des ANN.

Tabelle 15: Parameter von typischen ANNs

Parameter	**Description**	**Possible range**	**Considered range**
algorithm	The learning algorithm	backprop, rprop-, rprop+, sag, slr	
learningrate	Learning rate	Integer	0.01, 0.05, 0.1
hidden	Number of hidden layer	Integer	1*, 5, 10, 50, 100, 150
threshold	Threshold for the partial derivatives of the error function as stopping criteria.	Integer	0.01*
err.fct	Differentiable function that is used for the calculation of the error	sse, ce	sse, ce
act.fct	Differentiable function that is used for smoothing the result of the cross product of the covariate or neurons and the weights	logistic, tanh	logistic, tanh
rep	The number of repetitions for the neural network training.	Integer	1*,2,5,10

Tuning bestehender Klassifikatoren: Wir verwendeten spezifische Empfehlungen für die Abstimmung aller Klassifizierungsalgorithmen und bestimmten den Parametersatz, der bei der Parametereinstellung getestet wird, die in Tabelle 16 dargestellt ist. Die Tabelle enthält auch Algorithmen, die während der Ausführung von Aufgabe 4.2 implementiert wurden. Wir nahmen in der Tabelle auch spezielle Merkmale jedes Algorithmus auf in Bezug auf Eingabedaten, die Unterstützung von Klassengewichten und die Fähigkeit, Klassenzugehörigkeitswahrscheinlichkeiten zu geben, die für einige Methoden erforderlich sind, die bei der mehrperiodischen Aggregation verwendet werden (Aufgabe 4.3).

Tabelle 16: Überblick über die Algorithmen mit speziellen Eigenschaften und der getesteten Konfigurationen

Algorithm	Category	Input data		Supports class weights possible	Returns class-membership probabilities	Estimated number of configurations
		Numeric	Categorical			
kNN	Instance based learner	X		X	X	10
LDA	Statistical learner	X		X	X	15
NB	Statistical learner	X	X	X	X	1
LR	Statistical learner		X	-	X	1
SVM	Support Vector Machines	X		X	X	6'300
AdaBoost	Ensembles	X	X	X	X	9
RF	Ensembles	X	X	X	X	48
XGB	Ensembles	X	X	X	(some objectives)	20'000
ANN	Neural Networks	X			X	1'400

Anwendung robuster Vorhersagemethoden (Aufgabe 4.2)

Rauschen ist ein ernsthaftes Problem im Data Mining und kann die Systemleistung in Bezug auf Klassifizierungsgenauigkeit, Zeit beim Aufbau eines Klassifikators und die Größe des Klassifikators reduzieren. In diesem Abschnitt testen wir verschiedene Algorithmen, die weniger anfällig für Rauschen sind.

Robustes SVM: Neben dem Standard-SVM-Algorithmus testeten wir die robuste multivariate Klassifikation mit hoch optimierten SVM-Ensembles mit dem R classyfire-Paket (Chatzimichali und Bessant 2015). Classyfire erreicht dies durch die Bereitstellung von Funktionen, die den Aufbau und die Prüfung von Klassifikatoren so weit wie möglich automatisieren. Um zu vermeiden, dass diese Funktionen zu undurchdringlichen Blackboxen werden, werden detaillierte Informationen über die Funktionsweise dieser Funktionen bereitgestellt, und es wird voller Zugriff auf die Innenseiten aller produzierten Klassifikatoren gewährt.

AdaBoost: Der AdaBoost-Algorithmus von Freund und Schapire (1997) war der erste praktische Boosting-Algorithmus, der mehrere schwache Lerner (z.B. Entscheidungsbäume) kombiniert, um einen starken Lerner aufzubauen. Die Kombination erfolgt durch das Kräftigen der erlernten Regeln der schwachen Lerner in mehreren Iterationen. In diesem Projekt verwendeten wir die Implementierung von Alfaro et al. (2013), die in der Lage ist, mit Multiklassenproblemen umzugehen. Die für die Abstimmung verfügbaren Parameter sind in der Tabelle 17 aufgeführt.

Tabelle 17: Parameter des AdaBoost Algorithmus

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

Random Forest (RF): Dieser Algorithmus erzeugt mehrere niedrig korrelierte Entscheidungsbäume, die mit Ensemble-Methoden gelernt und bewertet werden (Breiman 2001). In unserem Artefakt verwendeten wir die Implementierung von (Hothorn u. a. 2006; Strobl u. a. 2008), die die Parameter in Tabelle 18 für das Tuning liefert. Biau (2012) gibt eine Einführung in das Random Forest Modell Tuning.

Der Random Forest liefert interne Merkmalswertmessungen, mit denen die Vorhersagekraft einzelner Merkmale beurteilt werden kann. Die Bedeutung des Merkmals wird als Mean Decrease in Accuracy (MDA) gemessen. Ein hoher MDA-Wert deutet auf eine hohe Bedeutung des Merkmals für das Klassifikationsmodell hin. Es ist wichtig zu erwähnen, dass ein MDA von Null nicht bedeutet, dass ein Merkmal keinen Einfluss

auf die Klassifizierung hat, und ein negativer MDA nicht bedeutet, dass das Merkmal einen negativen Einfluss auf die Klassifizierung hat, da die MDA-Werte nur interne Gewichte des Random Forest Klassifizierer sind. Ein alternatives Wichtigkeitsmaß für Features im Random Forest Modell ist Mean Decrease Gini (MDG). Die Interpretation dieses Maßes ist ähnlich.

Beispiele für die Bedeutung von Merkmalen für 11 energieeffiziente Haushaltseigenschaften sind in Abbildung 25 dargestellt. Man kann sehen, dass die verschiedenen Maße von Bedeutung für die Merkmale zu unterschiedlichen Interpretationen führen können.

Tabelle 18: Parameter des Random Forest Algorithmus

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

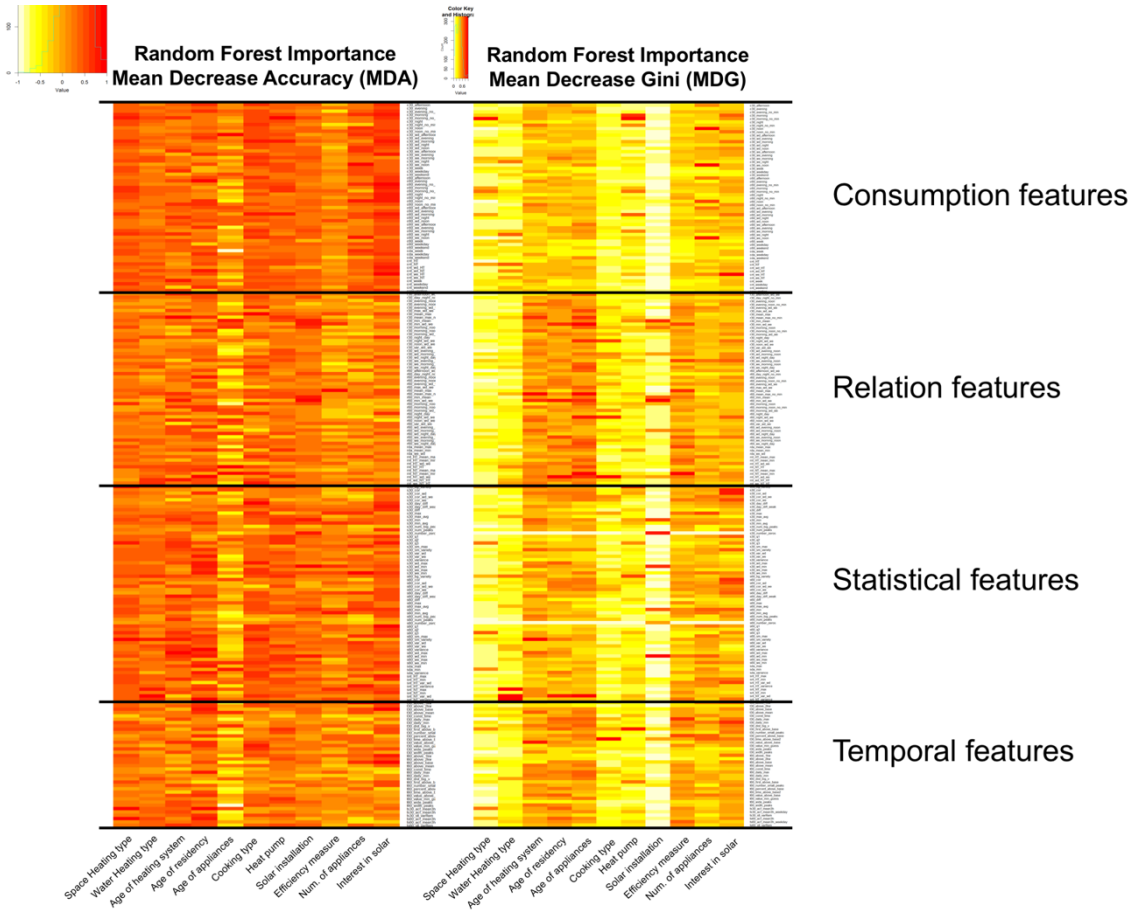


Abbildung 25: Visualisierung von zwei Random Forest Feature-Relevanzmaßen für alle Features auf Basis von 15-min Stromverbrauchsdaten für die Vorhersage von elf Haushaltseigenschaften

Im Detail zeigt die Analyse von Abbildung 26, dass sich die Feature-Importe für pApplianceAge von anderen Eigenschaften unterscheiden (siehe MDA). Bei der Vorhersage des Alters (von Raum-/Wasserheizungen und Geräten), der Frequenzen (Wirkungsgrade, Anzahl der Geräte) und des Interesses an Solaranlagen scheint

es, dass mehr Features für das Modell wichtig sind. Diese Eigenschaften haben jedoch eine geringere Genauigkeit. Dies könnte ein Hinweis darauf sein, dass Modelle mit einer hohen Anzahl wichtiger Features eine geringere Vorhersagekraft haben. Es scheint, dass Verbrauch und statistische Merkmale für die Klassifizierung wichtiger sind als die anderen Kategorien. Für MDG kann diese Beobachtung nicht bestätigt werden, im Gegenteil, die Kategorien Relationen und zeitliche Merkmale zeigen eine höhere Bedeutung.

eXtreme Gradient Boosting (XGB): Als Erweiterung des Gradient Tree Bosting-Algorithmus von Friedman (2001) beschreiben Chen und Guestrin (2016) einen skalierbaren maschinellen Lernalgorithmus, der mehrere Entscheidungsbäume aufbaut, indem er die Trainingsdaten iterativ in kleinere Teile aufteilt und die Vorhersagen aller Basis-Klassifikationsbäume aggregiert. Der Algorithmus implementiert drei Techniken, die eine Overfitting vermeiden: ein regularisiertes Lernziel, das die Komplexität des Modells bestraft, „tree shrinkage" um den Einfluss jedes einzelnen Baums zu begrenzen, und Feature-Subsampling, was bedeutet, dass nur Teilmengen von Features verwendet werden, um Bäume zu generieren. Grundsätzlich gibt es drei Booster-Algorithmen: Baum-Booster ("gbtree"), Dart-Booster ("Dart"), Linear-Booster ("gblinear"). Ein weiterer Vorteil dieses Algorithmus ist, dass er explizit mit fehlenden Werten umgehen kann. Die Anzahl der XGB-Parameter ist groß (DMLC 2016b) und wir listen sie in Tabelle 19 auf. Einige Anmerkungen des Entwicklers zur Abstimmung des XGB-Algorithmus sind online verfügbar (DMLC 2016a).

Für die Interpretation des Modells stellt XGB "Gain" als Feature Wichtigkeitswert zur Verfügung. Diese Punktzahl misst den relativen Beitrag eines Merkmals zum Modell, indem sie jeden Merkmalsbeitrag an jedem Baum im Modell berücksichtigt. Diese Punktzahl ist nur beim Vergleich von Merkmalen in einem Modell sinnvoll, und ein höherer Wert für ein Merkmal bedeutet, dass es für die Generierung einer Vorhersage in diesem Modell wichtiger ist.

Tabelle 19: Parameter des XGB Algorithmus

Parameter	**Description**	**Possible range**	**Considered range**
booster	Which booster to use	**gbtree**, gblinear, dart	
nrounds	The max number of iterations	Integer	5, 10, 20
Parameters for tree booster			
eta	After each boosting step, we can directly get the weights of new features. and eta actually shrinks the feature weights to make the boosting process more conservative.	[0,1]	0.001, 0.01, 0.05, 0.1, 0.3*
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger, the more conservative the algorithm will be.	[0,oo[	0*
max_depth	Maximum depth of a tree, increase this value will make model more complex / likely to be overfitting.	Integer	3, 7, 10, 20
min_child_weight	Minimum sum of instance weight needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning. In linear regression mode, this simply corresponds to minimum number of instances needed to be in each node. The larger, the more conservative the algorithm will be.	[0, ∞[	1*

max_delta_step	Maximum delta step we allow each tree's weight estimation to be. If the value is set to 0, it means there is no constraint. If it is set to a positive value, it can help making the update step more conservative. Usually this parameter is not needed, but it might help in logistic regression when class is extremely imbalanced. Set it to value of 1-10 might help control the update	[0, ∞[	0*
subsample	Subsample ratio of the training instance. Setting it to 0.5 means that XGBoost randomly collected half of the data instances to grow trees and this will prevent overfitting.	[0, 1]	1*
colsample_bytree	Relative number of features that are used when constructing each tree.	]0, 1]	1*, 0.8, 0.6, 0.4
colsample_bylevel	Relative number of features that are used for each split, in each level.	]0, 1]	1*, 0.8, 0.6, 0.4
lambda	L2 regularization term on weights, increase this value will make model more conservative. When the parameter is set to zero, the optimization objective equals to the gradient tree bosting (Friedman 2001)	[0, ∞[	1*
alpha	L1 regularization term on weights, increase this value will make model more conservative.	[0, ∞[	0*
tree_method	The tree construction algorithm used in XGBoost; 'auto' uses a heuristic to choose the faster one of 'exact' and 'approx'	'auto', 'exact', 'approx'	
sketch_eps	This is only used for tree_method='approx'	]0, 1[	0.03*, 0.05
scale_pos_weight	Control the balance of positive and negative weights, useful for unbalanced classes.	[0,1]	0*
Additional parameters for dart booster			
sample_type	Type of sampling algorithm	'uniform', 'weighted'	
normalize_type	Type of normalization algorithm	'tree', 'forest'	
rate_drop	Dropout rate.	[0, 1]	0*
skip_drop	Probability of skip dropout; if a dropout is skipped, new trees are added in the same manner as gbtree.	[0, 1]	0*
Parameters for linear booster			
Alpha, lambda	See above	-	-
lambda_bias	L2 regularization term on bias, default 0 (no L1 reg on bias because it is not important)	[0, ∞[	0

Leistungsbewertung von Klassifizierungsmethoden für die Haushaltsklassifizierung: Zur Identifizierung des Kaufinteresses an einer Fiber-to-the-Home (FTTH) Cross-Selling-Studie testeten wir verschiedene Klassifikatoren mit unterschiedlichen Konfigurationen. Abbildung 26 zeigt Klassifizierungsergebnisse mit verschiedenen Klassifizierern. Wir stellten fest, dass Random Forest die besten Ergebnisse erzielte.

Wir testeten den Klassifikator, den wir in Aufgabe 4.1 und 4.2 identifiziert und implementiert hatten, mit einem Smart-Meter-Datensatz, der in einem früheren Projekt erworben worden ist (Sodenkamp u. a. 2016). Die Tabelle 20 zeigt die Anzahl der getesteten Konfigurationen mit jedem Klassifizierer für jede Haushaltsimmobilie.

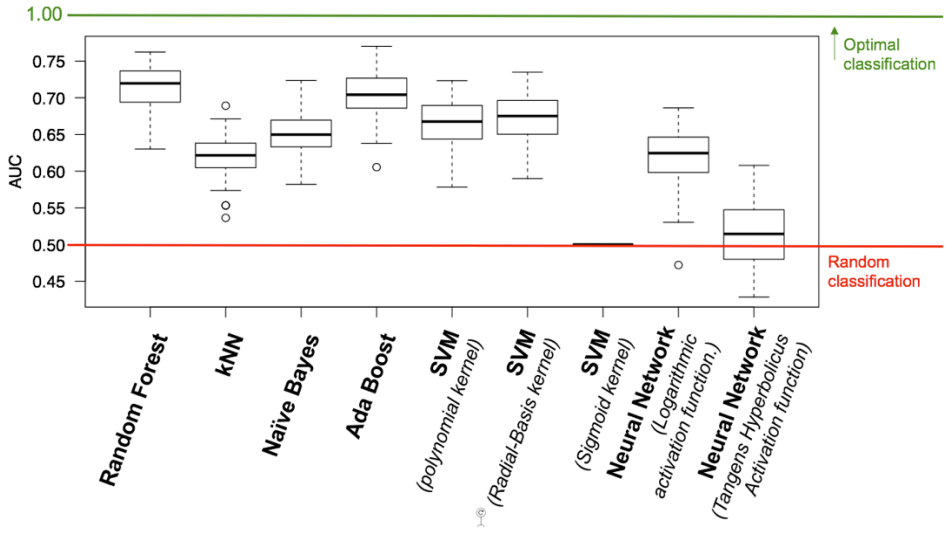


Abbildung 26: Boxplots über die AUC-Ergebnisse verschiedener Klassifizierer auf Basis verschiedener Trainingszeiträume der Smart-Meter-Daten (29 verschiedene Wochen ohne Feiertage oder unvollständige Daten)

Die Heatmap (Abbildung 27) zeigt die durchschnittliche Genauigkeit verschiedener Klassifikationsmodelle ohne Feature-Auswahl. Die Ergebnisse deuten darauf hin, dass Random Forest, Logistic Rregression und SVM bessere Ergebnisse liefern als andere Algorithmen, aber wir gehen davon aus, dass es sich um ein verzerrtes Ergebnis handelt, da wir eine ungleiche Anzahl von Klassifizierungsparametereinstellungen für die verschiedenen Klassifizierer testeten. Wenn die Klassifikationstests abgeschlossen sind, werden wir die Ergebnisse mit einer strengen Methodik bewerten.

Tabelle 20: Anzahl der getesteten Klassifikationsmodelle mit Datensatz L

	AdaBoost	kNN	Log. Regression	Naïve Bayes	Random Forest	SVM	XGBoost
ftth.pi.classesC	17	35	12	31	33	7775	14
ftth.unknown	17	35	12	31	33	7774	14
pAppliancesAge	13	35	61	31	34	7777	14
pAppliancesNum	18	35	61	31	33	7775	14
pChildren	18	35	12	31	33	7775	14
pCookingType	13	35	12	31	35	7777	14
pEnergyEfficiencyMeasure	18	35	61	31	34	7774	14
pFamily	18	35	12	31	33	7774	14
pHeatPump	13	35	12	31	33	7775	14
pHouseAge	13	35	61	31	39	7773	14
pHouseholdType	13	35	70	31	39	7772	14
pHouseholdType2	18	35	12	31	33	7774	14
pHouseOwnership	13	35	12	31	40	7773	14
pInterestSolar	18	35	61	31	33	7775	14
pLivingAreaB	13	35	68	31	35	7779	14
pNumResidentsB	15	36	69	32	36	7781	14
pSatisfactionUtility	18	35	61	31	33	7775	14
pSingle	18	35	12	31	33	7774	14
pSolar	12	35	12	31	34	7773	14
pSpaceHeatingAge	13	35	68	31	36	7779	14
pSpaceHeatingType	15	36	86	32	40	7782	14
pWaterHeatingType	15	36	28	32	37	7782	14

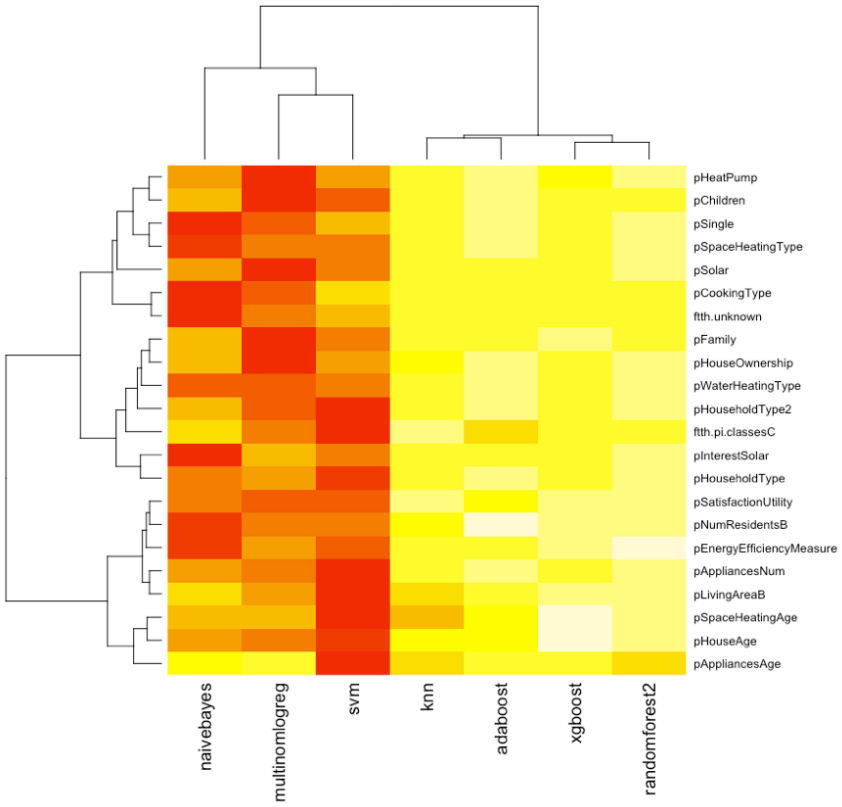


Abbildung 27: Durchschnittliche Klassifikationsgenauigkeit für verschiedene Klassifizierer zur Vorhersage von 22 Haushaltseigenschaften (ohne Feature-Selektion)

Mehrwöchige Klassifikation (Aufgabe 4.3)

Die vorhandenen rudimentären Methoden basieren auf der Analyse von Stromverbrauchsdaten einer Woche. Zur Verbesserung der Klassifikationsqualität werden alle Zeitfenster einbezogen. Das System kann selbstständig irrelevante Zeiträume von der Betrachtung automatisch ausschließen (Urlaubs- und Feiertage, usw.). Aufgrund des hohen Berechnungsumfangs wird die Klassifikation für verschiedene Zeiträume getrennt durchgeführt. Mithilfe von Ensemble- und Zeitraumaggregationsmethoden fassen wir die Vorhersagen auf Basis einzelner Wochen zusammen. Die meisten Algorithmen die Funktionalität Wahrscheinlichkeiten der Mitgliedschaft in jeder Klasse j für jeden Haushalt i für jede Woche von Verbrauchsdaten w zurückgeben können. Diese bezeichnen wir mit

$$\left\{p_i^{jw}\right\}_{i\in I, j\in C}$$

Die Werte p sind Wahrscheinlichkeiten und können daher nur Werte zwischen 0 und 1 annehmen.

$$0 \leq p_i^{jw} \leq 1, \forall\, i \in I, j \in C$$

Für jede Eigenschaft summieren sich die Wahrscheinlichkeiten jeder Klasse zu einer.

$$\Sigma_{j\in C}\left(p_i^{jw}\right) = 1, \forall i \in I, w \in W$$

Mit den daraus resultierenden Wahrscheinlichkeiten ist es möglich, eine Klassifizierung durchzuführen. Für jede Haushaltseigenschaft ordnen wir diesen Haushalt der Klasse c_i^w zu mit

$$c_i^w = argmax_{j\in C} p_i^{jw}$$

Bei mehreren Maxima (z.B. 50%/50% Wahrscheinlichkeiten für zwei Klassen) wird die Klassifizierung zufällig aus den gebundenen Klassen ausgewählt.

Majority Voting: Der einfachste Ansatz zur Aggregation der Informationen aus mehreren Wochen besteht darin, die Klassifizierung für jede Woche zu vervollständigen und die endgültige Klassifizierung mit dem Mehrheitswahl-Verfahren durchzuführen. In diesem Fall verwenden wir die Klassifizierungsfamilie

$$h_i = \{c_i^w\}_{w \in W}$$

für jeden Haushalt. Die Anzahl der Vorkommen jeder Klasse wird für jeden Haushalt gezählt und die endgültige Klassifizierungsentscheidung d_i^{mv} wird basierend auf der Klasse mit den meisten Vorkommen getroffen.

$$d_i^{mv} = argmax_{c \in C} \#\{x : x = c, x \in h_i\}$$

Im Falle eines Unentschiedens wird die Klassifizierung nach dem Zufallsprinzip aus den gebundenen Klassen ausgewählt. Dieser Prozess wird als Mehrheitsabstimmung bezeichnet, da jede Woche für eine bestimmte Klasse "stimmt" und die Klasse mit der Mehrheit der Stimmen gewinnt. Die Bewertung der Ergebnisse in Bezug auf die Genauigkeit ist in Tabelle 21 zusammengefasst.

Tabelle 21: Accuracy für die mehrwöchige Klassifikation mit Majority Voting unter Verwendung des AdaBoost Klassifikators

Properties	**Accuracy (majority voting)**
single	0.82
employment	0.72
family	0.73
children	0.73
retirement	0.74
cooking	0.71
N_residents	0.75
age_house	0.64

Aggregation mit Statistikfunktionen: Weiterhin können wir die Ergebnisse mehrerer Wochen kombinieren, indem wir die Ergebnisse mit Hilfe von Funktionen aggregieren. Die Aggregationsfunktion muss die Wahrscheinlichkeiten aus einer beliebigen Anzahl von Wochen als Input nehmen. Das bedeutet, dass die Funktionen selbst eigentlich eine Funktionsfamilie f_n sind, die Werte von einem multidimensionalen Raum $\mathbb{R}^n$ auf $\mathbb{R}$ abbildet. Um die Ergebnisse vergleichbar zu machen, sollten alle Funktionen skaliert werden, um sie zu erfüllen:

$$f(0, \ldots, 0) = 0$$

$$f(1, \ldots, 1) = 1$$

Weiterhin wollen wir, dass die Aggregationsfunktionen symmetrisch sind:

$$f(x, y) = f(y, x)$$

und monoton steigend:

$$\text{if } x_i > y_i \; \forall i \text{ then } f(x_1, \ldots, x_n) > f(y_1, \ldots, y_n)$$

Diese Eigenschaft stellt sicher, dass jede Erhöhung der Wahrscheinlichkeit in einer einzigen Woche die Gesamtwahrscheinlichkeit verbessern sollte. Die Funktionen sollten steigen oder fallen, wenn ein hoher oder niedriger Wert an den Eingang angelegt wird:

$$f(x_1, \ldots, x_n) < f(x_1, \ldots, x_n, x), \text{if } x > x_i \; \forall i$$

$$f(x_1, \ldots, x_n) > f(x_1, \ldots, x_n, x), \text{if } x < x_i \; \forall i$$

Das arithmetische Mittel ist die am häufigsten verwendete Funktion zur Aggregation. Wir führen unsere Analyse mit Mittelwerten und anderen häufig verwendeten Funktionen durch: Minimum, Quartil, Median und geometrischer Mittelwert.

Die Anwendung der Aggregationsfunktionen auf die wöchentlichen Wahrscheinlichkeiten führt zu einem realen Wert zwischen 0 und 1 für jede Klasse und jeden Haushalt:

$$r_i^{j,f} = f\left(\{p_i^{j,w}\}_{w \in W}\right)$$

Der Haushalt wird dann der Klasse mit dem größten Ergebniswert zugeordnet.

$$c_i^f = argmax_{j \in C} r_i^{j,f}$$

Die Ergebnisse für die Genauigkeit und die MCC-Werte sind in den Tabellen 22 und 23 zusammengefasst. Zwischen den verschiedenen Aggregationsfunktionen gibt es nur geringe Unterschiede. Insgesamt liefert das Minimum die besten Ergebnisse für die Genauigkeit (außer der Eigenschaft single). Für MCC ist der Mittelwert die leistungsstärkste Funktion (mit Ausnahme der Eigenschaften Kochen und Familie).

Tabelle 22: Accuracy nach der Aggregation verschiedener Wochen bei der Verwendung der statistischen Aggregation (AdaBoost Klassifizierer)

	Aggregating functions				
Property	**mean**	**quartile**	**min**	**geometric mean**	**median**
single	0,84	0,84	0,83	0,84	0,84
employment	0,73	0,73	0,73	0,73	0,72
family	0,76	0,76	0,77	0,76	0,75
children	0,74	0,74	0,75	0,74	0,73
retirement	0,69	0,69	0,73	0,7	0,67
cooking	0,63	0,63	0,66	0,63	0,62
Number of residents	0,77	0,77	0,78	0,78	0,77
House age	0,64	0,64	0,64	0,64	0,64

Tabelle 23: MCC nach der Aggregation verschiedener Wochen bei der Verwendung der statistischen Aggregation (AdaBoost Klassifizierer)

Property	**mean**	**quartile**	**min**	**geometric mean**	**median**
single	0,48	0,47	0,44	0,47	0,47
employment	0,46	0,46	0,43	0,46	0,45
family	0,42	0,42	0,37	0,42	0,43
children	0,42	0,41	0,38	0,42	0,4
retirement	0,4	0,4	0,38	0,4	0,37
cooking	0,22	0,22	0,25	0,22	0,22
Number of residents	0,55	0,54	0,55	0,55	0,54
House age	0,28	0,28	0,27	0,28	0,27

Evaluation der Klassifikationsgüte mit Performanzmetriken (Aufgabe 4.4)

Wir verwenden verschiedene Kennzahlen zur Klassifikationsqualität, um die Korrektheit unserer entwickelten Klassifikationsmodelle zu messen. Dazu zählen wir die Anzahl der korrekt erkannten Klassenbeispiele (true positives, TP), die Anzahl der korrekt erkannten Beispiele, die nicht zur Klasse gehören (true negatives, TN), und Beispiele, die entweder falsch der Klasse zugeordnet wurden (false positives, FP) oder die nicht als Klassenbeispiele erkannt wurden (false negative, FN). Diese vier Fälle stellen die in Tabelle 24 für den Fall der binären Klassifizierung (Sokolova und Lapalme 2009) und in Tabelle 25 für ein Mehrklassenproblem dar.

Tabelle 24: Konfusionsmatrix für ein binäres Klassifikationsproblem

Real values	**Predictions**		
	Positive	**Negative**	**Total**
Positive	TP	FP	P
Negative	FN	TN	N
Total	P'	N'	

Tabelle 25: Konfusionsmatrix für ein Mehrklassen-Problem

Real values	**Predictions**			
	Class A	**Class B**	**Class C**	**Total**
Class A	n_{aa}	n_{ba}	n_{ca}	n_a
Class B	n_{ab}	n_{bb}	n_{cb}	n_b
Class C	n_{ac}	n_{bc}	n_{cc}	n_c
Total	p_a	p_b	p_c	n

Wir unterscheiden zwischen Metriken für Eigenschaften (diese Metriken drücken eine Gesamtleistung für eine Eigenschaft aus) und Metriken für einzelne Klassen (diese Metriken drücken die Leistung für jede Klasse aus).

Bei der Qualitätsbewertung von Klassifizierungseinstellungen geht es nicht nur um die Berechnung von Leistungskennzahlen, sondern auch um die Interpretation der Werte. Diese Interpretation ist für jede Kennzahl spezifisch und unterscheidet sich in Bezug auf das Ziel der Nutzung und die betrachtete Referenzstatistik (Benchmark). Einen Überblick über die verwendeten Leistungskennzahlen und verwendeten Benchmarks geben wir in Tabelle 26. Die Tabelle enthält auch die Information, ob die Metrik robust gegenüber unausgewogenen Klassen ist (und somit für die Modellauswahl verwendet werden kann) und ob die Ergebnisse einer Klassifizierungseinstellung mit anderen Einstellungen (einschließlich anderer Eigenschaften mit unterschiedlichen Klassenverteilungen) verglichen werden können.

Tabelle 26: Überblick über die Performanzmaße

Measure	**Equation**	**Scope**	**Reference / Benchmark**	**Comparable to other classification settings**	**Robust against imbalanced classes**
Accuracy	1	Property	RG, BRG, Biggest class	no	no
MCC	2	Property	0	yes	yes
Precision	3	Class	Random Guess, Class size	no	no
Recall	4	Class	-	no	no
F_1	5	Class	-	yes	yes
AUC		Class	0.5	yes	Yes

Metriken auf der Ebene von Haushaltseigenschaften

Genauigkeit - Sie ist definiert als der Anteil der korrekt klassifizierten Instanzen an der Anzahl der gesamten Klassifizierungsinstanzen:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

Die Messung kann Werte zwischen 0 und 1 annehmen, wobei 1 der perfekten Vorhersage und 0 der totalen Fehlklassifizierung entspricht. Die Genauigkeit ist leicht zu interpretieren, aber in der Situation sind die

Klassen unausgewogen (d.h. eine Klasse kommt viel häufiger vor als die anderen), ein Klassifizierer, der immer davon ausgeht, dass eine Mehrheitsklasse eine hohe Genauigkeit erreichen kann. Daher kann diese Maßnahme bei Anwendung auf solche unausgewogenen Eigenschaften leicht irreführend sein.

Matthews Correlation Coefficient (MCC) - Es ist eine alternative Maßnahme, die besser für die unausgewogenen Probleme geeignet ist. Es ist ein Korrelationskoeffizient zwischen der beobachteten und der vorhergesagten Klassifizierung. Im Falle eines Problems mit binärer Klassifizierung ist es gleichbedeutend mit der phi-Statistik (Cramer 1946). Wir verwenden die Definition von MCC für Multiklassen-Klassifikationsprobleme, wie sie von Jurman et al. (2012) und Gorodkin (2004) vorgegeben wird. MCC kann Werte zwischen -1 und 1 annehmen. 1 entspricht der perfekten Klassifikation, -1 der totalen Unstimmigkeit zwischen den Vorhersagen und realen Beobachtungen und 0 der Klassifikation, die nicht besser ist als die zufällige Vorhersage. MCC fehlt die einfache Interpretierbarkeit des Genauigkeitsmessers, aber es ist robuster und eignet sich besser für den Vergleich zwischen den Klassifikatoren.

$$MCC = \frac{TP * TN \; - \; FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{2}$$

Metriken für Klassen

Precision: Dieses Maß drückt die Anzahl der korrekten klassifizierten Beispiele von allen positiv vorhergesagten aus. Das Maß wird durch die relative Klassengröße beeinflusst und es wird daher nicht empfohlen, die Genauigkeitswerte einer Klasse mit denen einer anderen zu vergleichen. Das Maß wird auch als Positiver Prädiktiver Wert (PPV) bezeichnet. Die Anzahl der FN und TN wird in dieser Metrik nicht berücksichtigt.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall: Diese Maßnahme drückt die Anzahl der korrekt identifizierten Beispiele aus allen Beispielen dieser Klasse aus. Es ist auch bekannt als Sensitivity, oder True Positive Rate (TPR). Die Anzahl der FP und TN wird nicht berücksichtigt.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F-Score: Präzision und Erinnerung sind für sich genommen nicht aussagekräftig für eine Klassifikatoroptimierung, da die Erhöhung einer dieser Messungen ohne Berücksichtigung der anderen zu verzerrten Ergebnissen führen würde. Zur Veranschaulichung kann man die folgenden Extremfälle betrachten: 1) ein Dummy-Klassifikator, der allen Beispielen eine Klasse zuordnet, hätte Recall=1 für diese Klasse und eine niedrige Genauigkeit, 2) ein Klassifikator, der nur ein Beispiel der Klasse von Interesse zuordnet und dieses Beispiel korrekt ist, würde eine Genauigkeit=1 erreichen, aber einen geringen Rückruf, da die Anzahl der FN groß ist. Der F-Score kombiniert beides mit einem gewichteten Durchschnitt und ist robust gegenüber extremen kleinen oder großen Klassen. Die Verteilung von F_1 ist in Abbildung 28 mit den entsprechenden Präzisions- und Erinnerungswerten dargestellt.

$$F_1 = \frac{2 * precision * recall}{precision + recall}, \beta \in \mathbb{R}^+ \tag{5}$$

$$F_\beta = \frac{(1 + \beta^2) * precision * recall}{(\beta^2 * precision) + recall}, \beta \in \mathbb{R}^+ \tag{6}$$

In seiner gebräuchlichen Form F_1 ist es das harmonische Mittel zwischen beiden Messungen. F_2 gibt dem Recall ein höheres Gewicht als Präzision und $F_{0,5}$ umgekehrt. Das Maß ist jedoch klassenspezifisch und seine Stärke ist der Vergleich verschiedener Klassifizierungseinstellungen. Eine einfache Interpretation der Werte für Laien ist nicht möglich.

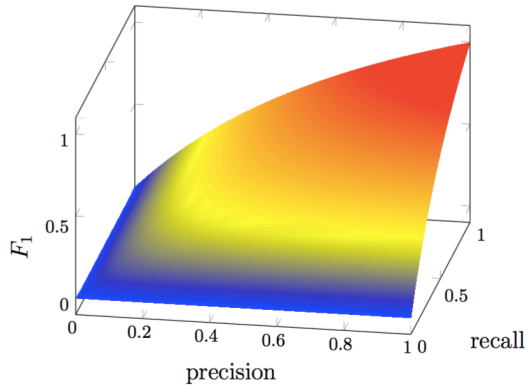


Abbildung 28: Illustration des F1 Gütemaßes

Spezifität: Dieses Maß quantifiziert die Anzahl der wahren Negative aus allen tatsächlich negativen Beispielen und quantifiziert die Vermeidung von falschen Negativen. Das Maß wird auch als True Negative Rate (TNR) bezeichnet.

$$Specificity = \frac{TN}{FP + TN} \tag{7}$$

AUC: Diese Messung basiert auf der Betriebskurve des Empfängers (ROC). ROC ist eine grafische Darstellung der Klassifikatorleistung für eine einzelne Klasse und wird durch die Darstellung des TPR (Recall) gegen den Ausfall (berechnet durch 1 Spezifität) erzeugt. Der Bereich unter dem ROC (AUC) ist daher ein unvoreingenommenes Maß für die Klassifizierungsleistung einer einzelnen Klasse. Der AUC-Wert für die Zufallsklassifizierung beträgt 0,5, was als minimale Benchmark angesehen werden kann.

Einige Studien kritisieren die AUC-Messung (Hanczar u. a. 2010; Hand 2009; Lobo, Jiménez-Valverde, und Real 2008). Der Hauptgrund ist, dass das AUC-Maß für kleine Stichprobengrößen irreführend sein könnte. Daher sollten mehrere Leistungsmessungen in wissenschaftlichen Berichten berichtet werden.

Referenzstatistiken zur Interpretation von Leistungskennzahlen

Die Möglichkeit, die Ergebnisse der Klassifikationsleistung zu vergleichen, wird tatsächlich analysiert. von Klassifikationseinstellungen, Einige Leistungskennzahlen haben keinen festen Referenzwert Wir betrachten zwei Kennzahlen für diesen Zweck:

1) Random Guess (RG) - Ohne Kenntnis der Klassenverteilung innerhalb einer Immobilie (z.B. wie viel Prozent der Kunden Einzelhaushalte sind) würde man davon ausgehen, dass die Klassenverteilungen gleichermaßen wahrscheinlich sind. Für n Klassen innerhalb einer Eigenschaft ist die zufällige Schätzmetrik daher:

$$RG = \frac{1}{n}$$

2) Biased Random Guess (BRG) - Diese Metrik wurde von Beckel et al. (2013, 2014) eingeführt und ist definiert als die Summe der quadrierten relativen Klassengrößen innerhalb einer Eigenschaft. Wenn h_k die relative Klassengröße der Klasse k bezeichnet, ist die Metrik definiert als:

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

Diese Kennzahl entspricht dem Herfindahl-Index, mit dem die Konzentration auf Monopolmärkten gemessen wird (Fahrmeir u. a. 2007 S. 86). Die Merkmale dieses Index können in die BRG-Maßnahme übernommen werden. Die Extremwerte für BRG sind also:

$$BRG_{max} = 1$$

$$BRG_{min} = \frac{1}{K}$$

Mehrdimensionale Klassifikation (Aufgabe 4.5)

Kernziel des Projektes ist die Entwicklung von multidimensionalen Klassifikationsalgorithmen, basierend auf den kompositionellen Analysen der Strom-, Gas- und Wasserverbrauchsdaten, sowie vielfältiger zusätzlicher Datenquellen. Dabei werden klassenunabhängige Variablen (externe Daten, die den Verbrauch beeinflussen) einbezogen, indem die in AP1 Aufgabe 6 abgeleiteten Korrelationskennzahlen verwendet werden.

Unser IT-Artefakt ist in Abbildung 29 schematisch dargestellt. Als Input betrachten wir Haushaltsstrom-, Gas- und Wasserverbrauchsdaten unterschiedlicher Granularität (15-min, 30-min, täglich, jährlich), sowie die Adresse (beide Informationen sind beim EVU für die Abrechnung verfügbar). Neben den Informationen, die den Unternehmen zur Verfügung stehen, haben wir kostenlose externe Datenquellen (z.B. Geodaten, statistische Daten, Daten aus Immobilienportalen) beschafft.

Als Ergebnis werden zusätzliche Kundeninformationen in Form von Haushaltsmerkmalen (Haushaltstyp, Wohnfläche, usw.) gewonnen. Diese Informationen sind einem Teil der Kunden, die das Engagement Portal nutzen, bekannt und stellen eine Grundwahrheit dar, die unserem überwachten Modell des maschinellen Lernens zugrunde liegt. Die Daten werden in einer Feature-Extraktionskomponente normiert, aufbereitet und mit Geodatendten angereichert.

Im Folgenden werden die verschiedenen Komponenten kurz erläutert und Links zu den jeweiligen Aufgaben in diesem Projekt gegeben, wo die Arbeit und die detaillierten Ergebnisse dokumentiert werden.

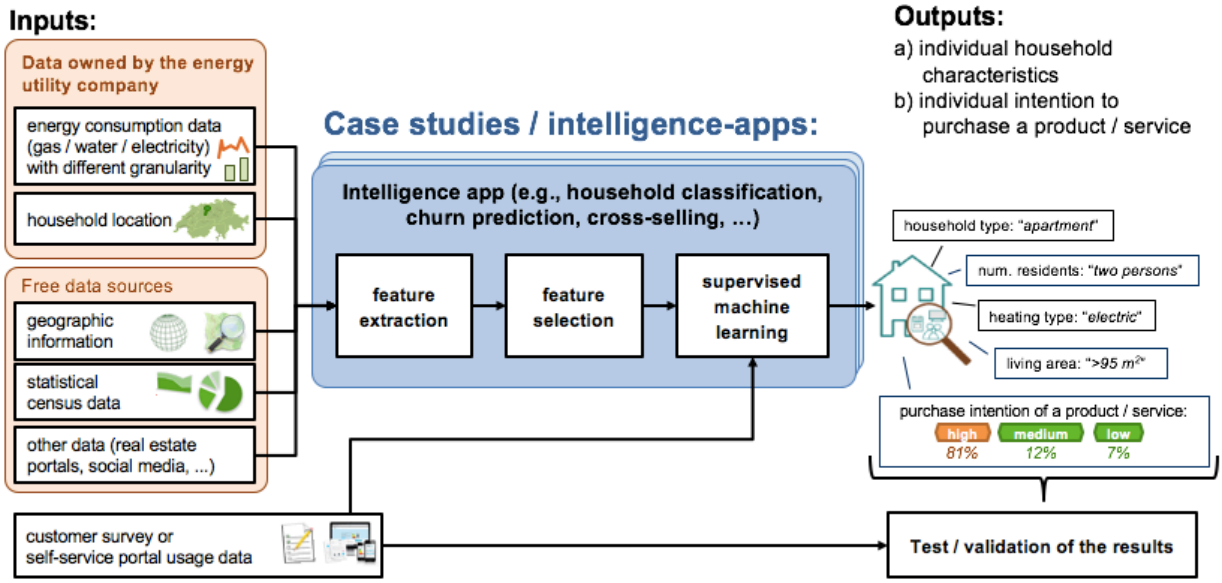


Abbildung 29: Schematische Darstellung des entwickelten mehrdimensionalen Klassifikationssystems

Die Eingangsdaten wurden in WP1 gesammelt und sind in den jeweiligen Aufgaben ausführlich beschrieben. Alle Features, die entwickelt und aus den Eingangsdaten extrahiert wurden, unterliegen WP2. Wir haben auch eine automatisierte Featureauswahl angewendet, die in Aufgabe 3.6 ausführlich beschrieben ist. Die für das maschinelle Lernen verwendeten Algorithmen sind in den Aufgaben 4.1, 4.2 und 4.3 beschrieben. Es werden Vorhersagen gemacht und das Ergebnis mit Leistungskennzahlen bewertet, die wir in Aufgabe 4.4 beschreiben.

Test des Systems auf Robustheit (Aufgabe 4.6)

Die Smart-Meter-Klassifizierung wurde auf ihre Robustheit in Bezug auf vier Einflussfaktoren getestet:

1. *Smart Meter Datenauflösung:* Die Verfügbarkeit detaillierter Stromverbrauchsdaten bei EVU ist mit den Datenschutzbedenken der Verbraucher verbunden. Abhängig von der Unternehmensrichtlinie und der lokalen Gesetzgebung nutzen EVU unterschiedliche Datengranularitäten. Die typischen Aggregationen variieren zwischen 15-Minuten, 30-Minuten, stündlichen und Tageswerten. Bei den Tagesdaten unterscheiden einige Versorgungsunternehmen zwischen dem HT und dem NT. Daten mit geringerer Häufigkeit enthalten weniger Informationen über das Verhalten der Kunden, und man kann

erwarten, dass sich die Leistung unserer Klassifizierungsalgorithmen verschlechtert, wenn sie auf diese Daten angewendet werden. Andererseits kann die Leistung des maschinellen Lernens durch eine hohe Datendimensionalität negativ beeinflusst werden, wenn die Merkmale nicht sorgfältig ausgewählt werden.

2. *Anzahl der Wochen für das Algorithmentraining:* Ein Intervall von Energieverbrauchsdaten kann durch Ausreißer beeinflusst werden. Daher ist zu erwarten, dass die Ergebnisse der Haushaltsklassifizierung, die auf dem Stromverbrauch einzelner Wochen basieren, durch Vorhersagen auf der Grundlage mehrerer Wochen Daten verbessert werden können, da dies auch die Vorhersagemodelle für den Energiebedarf verbessern könnte. Das Ausmaß der Verbesserungen wurde daher untersucht.
3. *Anzahl der Trainingsinstanzen:* Darüber hinaus untersuchten wir die Auswirkungen der Anzahl der Trainingsbeispiele mit bekannten Haushaltsmerkmalen (Ground Truth) auf die Genauigkeit der Haushaltsklassifizierung. Diese Informationen sind hilfreich für die Planung und Durchführung von Erhebungen, z.B. um Daten über neue Haushaltsinformationen zu sammeln, oder wenn die Klassifizierungsmodelle an regionalspezifische Grundwahrheitsdaten angepasst werden müssen, um bessere Vorhersagen zu treffen.
4. *Jährliche Saison der Trainingsdaten:* Wir untersuchten, wie Haushaltsmerkmale, die saisonabhängig verwendet werden, erkannt werden können (z.B. Raumheizung aufgrund niedrigerer Temperaturen und erhöhter Belegung im Winter). Die Vorhersagegenauigkeit ändert sichbei einigen Eigenschaften (insb. Raumwärmeart, Vorhandensein einer Wärmepumpe und Anzahl der Geräte) in Abhängigkeit von der aktuellen Woche des bei der Klassifizierung verwendeten Jahres, da diese Eigenschaften stark von Wettervariablen wie Außentemperatur oder Sonnenscheinstunden pro Tag abhängig sind.

Die Ergebnisse dieser Robustheitsprüfung beschreiben wir ausführlich in Hopf, Sodenkamp, und Staake (2018).

Die jährliche Klassifizierung wurde auf ihre Robustheit getestet, indem die geografische Übertragbarkeit evaluiert wurde. Konkret haben wir überprüft, inwieweit ein Klassifizierungsmodell, das auf Daten von Kunden aus einer geografischen Region trainiert wurde, zur Klassifizierung von Kunden aus einer anderen geografischen Region verwendet werden kann. Im Detail haben wir bewertet,

1. wie sich die Klassifizierungsqualität ändert, wenn ein prädiktives Modell, das mit Kunden eines Versorgungsunternehmens trainiert wurde, auf Kunden anderer Versorgungsunternehmen angewendet wird und
2. inwieweit öffentliche statistische Daten die Übertragbarkeit von Klassifizierungsmodellen zwischen Unternehmen und Regionen verbessern können.

Die Ergebnisse dieser Analyse beschreiben wir in Hopf, Riechel, Sodenkamp und Staake (2017).

Automatische Regularisierung von Parametern (Aufgabe 4.7)

Wir haben die automatische Regularisierung von Parametern (z.B. durch Ridge Regression oder LASSO Regression) geteststst, um das Gesamtsystem zu verbessern. Allerdings konnten dabei nur geringe Verbesserungen erreicht werden, die Komplexität des Systems wäre jedoch durch die Implementierung übermäßig komplexer geworden. Aus diesem Grund fanden die Ergebnisse keine Verwendung in der finalen Software.

Performanz-Evaluierung mit Kreuzvalidierung (Aufgabe 4.8)

Ziel der Klassifikatorbewertung ist es, eine unvoreingenommene Schätzung der Klassifikatorleistungs (beschrieben im Abschnitt „Evaluation der Klassifikationsgüte mit Performanzmetriken (Aufgabe 4.4)“ auf Seite 55) zu erhalten, insbesondere bei der Modellauswahl und Parametereinstellung.

Für die methodisch korrekte Berechnung ist es wichtig, dass alle Bias-Effekte (Fielding und Bell 1997) und die Überanpassung von Klassifikationsmodellen an die Daten auftreten. Dazu verwenden wir eine Kombination aus Holdout- und Kreuzvalidierungsverfahren (siehe Abbildung 30).

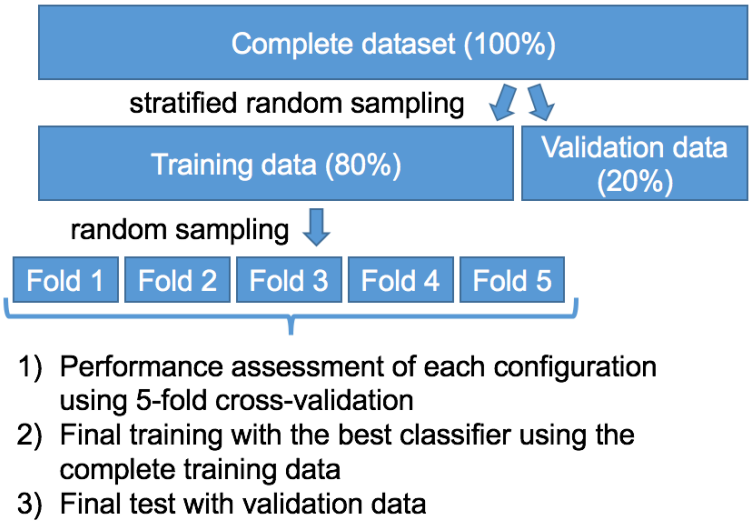


Abbildung 30: Schematische Darstellung der Klassifikations-Gütemessung

Der gesamte Datensatz wird zunächst in zwei Teile gegliedert: Trainingsdaten und Validierungsdaten. Die Trainingsdaten werden für die Modellauswahl und Parametereinstellung verwendet. Die Validierungsdaten werden für die endgültige Schätzung der Modellleistung verwendet. Die Datenaufteilung erfolgt mittels stratified sampling Stichproben (Tillé und Matei 2015) für jede Haushaltseigenschaft einzeln. Diese Methode bewahrt die Klassengrößen innerhalb jeder Eigenschaft und wir stellen mit dieser Methode sicher, dass kein degeneriertes Training / Validierungsset ausgewählt wird.

Zur Abschätzung der Performance einer Klassifikator-Konfiguration führen wir eine n-fache Kreuzvalidierung durch (mit n=4 oder n=5 Feldern, je nach Datensatzgröße). Bei dieser Technik werden die Trainingsdaten in n gleich große Teile, die sogenannten Felder, getrennt. Das Klassifikatortraining wird mit n-1 Feldern durchgeführt und das verbleibende Feld wird zur Evaluierung verwendet. Das Trainings- und Testverfahren wird n-mal durchgeführt, wobei zur Leistungsbeurteilung jeweils ein anderer Falz verwendet wird. Basierend auf den Ergebnissen der Kreurvalidierung wählen wir das Modell mit der höchsten Leistung aus und bewerten seine endgültige Leistung mit den Validierungsdaten.

Anwendung von Segmentierungsalgorithmen (Aufgabe 4.9)

Wir haben die folgenden Segmentierungsalgorithmen auf ihre Eignung für dieses Projekt untersucht:

- k-Means
- Hierarchisches Clustering
- Räumliche Clustering-Algorithmen (DBSCAN, Netzwerk-Cluster-Algorithmen)
- Selbstorganisierende Maps

Auf der Grundlage von Smart Meter Daten (15-min bis 60-minütige Messintervalle) haben bisher schon mehrere Autoren automatische, unbeaufsichtigte Clustering-Techniken zur Identifizierung von Kunden mit ähnlichem Verbrauchsmuster evaluiert (Albert und Rajagopal 2014; Al-Otaibi u. a. 2016; Beckel, Sadamori, und Santini 2012; Chicco 2012; Silva u. a. 2011; Flath u. a. 2012; Kwac u. a. 2013; McLoughlin, Duffy, und Conlon 2012). Die daraus resultierenden Kundensegmente bedürfen einer weiteren Interpretation und Behandlung durch Experten, bevor die Erkenntnisse in Energieeffizienz oder Marketingkampagnen umgesetzt werden können. Daher konzentrieren wir uns auf die Erkennung von Energieeffizienz-Merkmalen unter Verwendung von überwachten maschinellen Lernalgorithmen, da die haushaltsspezifischen Vorhersagen direkt in entsprechenden Kampagnen genutzt werden können. Jüngste Arbeiten zeigen, dass bestimmte Haushaltsmerkmale anhand von Stromverbrauchsdaten erkannt werden können.

Wir haben die folgenden Einschränkungen der Clustering-Algorithmen in unserem Fall identifiziert:

- Bei der Verwendung von räumlichen Clustering-Algorithmen wird die Annahme einer bestimmten Nachbarschaft von Datenpunkten gemacht. Bei Zeitreihendaten wird diese Tatsache nicht angegeben. Daher können wir die räumlichen Algorithmen nicht verwenden.

- Feste Annahmen über Parameter: Die meisten Algorithmen benötigen die feste Annahme über die Anzahl der Cluster im Datensatz. Dies ist eine starke Einschränkung, da wir nicht unbedingt in der Lage sind, diesen Parameter zu definieren.

Segmentierung von dynamischen Zeitreihendaten (Aufgabe 4.10)

Zur Identifizierung verbrauchsarmer Zeiten (auch als Abwesenheit von Bewohnern interpretierbar) in Haushalten verwenden wir die dynamische Zeitreihensegmentierung. Aufgrund des Fehlens von Daten für die überwachte Auswertung (Wissen über die Belegungszeiten der Bewohner) haben wir Visualisierungen von Lastprofilen von 12 Wochen berücksichtigt und einen k-Means-Clustering-basierten Algorithmus entwickelt, der Zeitspannen identifiziert, die zu Clustern mit niedrigem und hohem Verbrauch gehören.

Unser Algorithmus zur Identifizierung verbrauchsarmer Zeiten (tägliche Verbrauchsmessungen) innerhalb von 12 Wochen:

1. Verwenden von k-Means, um zwei Cluster Zeiträumen mit ähnlichen Verbrauchsmessungen zu identifizieren.
2. Wenn die Differenz zwischen den Clusterzentren größer ist als *DELTA*,
   a. werden die Tage zum *lowConsumption*-Cluster zugeordnet, welche im Clustern mit geringerem Verbrauchswerten pro Tag sind, und
   b. es wird sichergestellt, dass die Zeiten mit niedrigem Verbrauch länger als *n_days_check* sind, indem über die Zeitreihe iteriert wird und alle zusammenhängenden Zeiträume, die kleiner als s *n_days_check* sind, dem *highConsumption* Cluster zugeordnet werden.
3. Anfernfalls gibt es keine unterschiedlichen Verbrauchscluster und alle Tage werden dem *highConsumption*-Cluster eingestellt.

Wir gehen davon aus, dass *DELTA* als 1,25-fache der Standardabweichung innerhalb der Cluster eine gute Entscheidungsregel für das Vorhandensein von zwei separaten Konsumclustern gilt. Für *n_days_check* = 4 stellen wir sicher, dass die Zeiten des geringen Verbrauchs (d.h. die Abwesenheit der Bewohner) aus mindestens vier aufeinander folgenden Tagen (z.B. ein verlängertes Wochenende) bestehen müssen. Abbildung 31 und 32 zeigen exemplarische Ergebnisse des Algorithmus zur Erkennung von Abwesenheit / Anwesenheit.

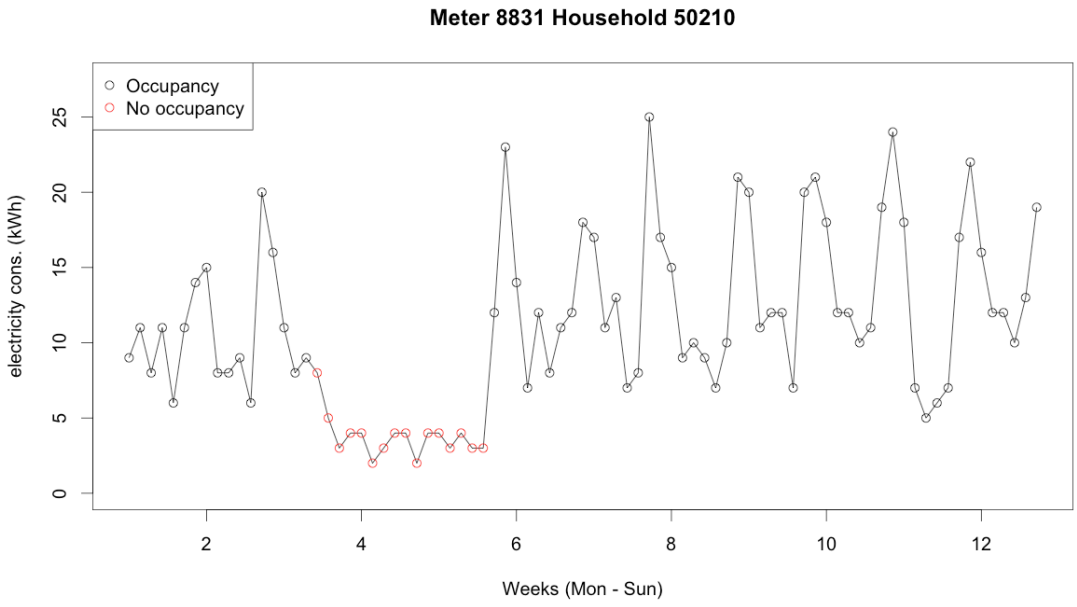


Abbildung 31: Stromverbrauchskurve mit täglichen Verbrauchsmessungen in einer 12-Wochen-Periode mit erkannter Abwesenheit

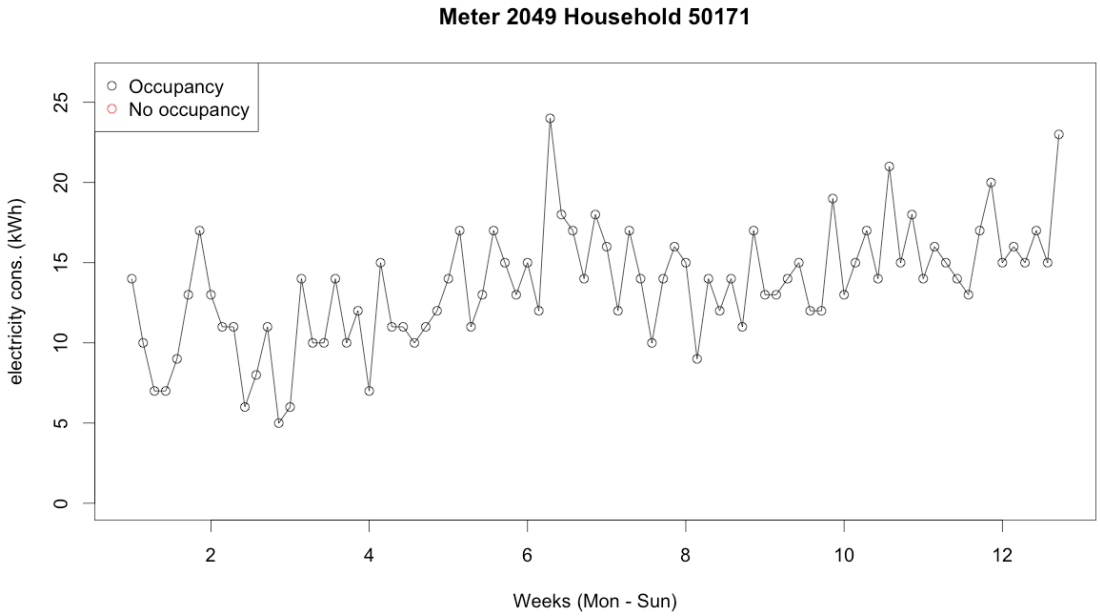


Abbildung 32: Stromverbrauchskurve mit täglichen Verbrauchsmessungen in einer 12-Wochen-Periode ohne erkennbare Abwesenheit

Hinzufügen von Gas- und Wasserverbräuchen zur Segmentierung (Aufgabe 4.11)

Gas- und Wasserverbräuche wurden mit Hilfe von Features zur Klassifikation und zur Segmentierung hinzugefügt. Alle Verbrauchszeitreihen korrelieren zueinander; da nicht nur die Werte in den gleichen Zeitabschnitten, sondern auch geschiftete Werte. Aus diesem Grund ist es notwendig, multiple Zeitreihenanalysen durchzuführen. Die detaillierte Beschreibung findet sich in Hopf (2019 Kap. 3).

Hinzufügen von verbrauchsrelevanten Daten (Aufgabe 4.12)

Features aus den externen Daten wurden zur Klassifikation und zur Segmentierung hinzugefügt. Die detaillierte Beschreibung der Features befindet sich in der Ergebnisbeschreibung zu Aufgabe 3.8.

Implementierung von Hidden-Markov-Modellen (Aufgabe 4.13)

Hidden-(Semi)-Markov-Modelle wurden getestet, um Verbrauchsprofile zu erkennen.

Wichtige Ergebnisse:

- Wir haben einen modifizierten HMM-Ansatz auf dem gegebenen Arbon SMD implementiert und angewendet, um festzustellen, ob Geräte in Haushalten vorhanden sind.
- Unser Ansatz war meist besser als Zufallsschätzungen und verzerrte Zufallsschätzmethoden und hatte einen insgesamt guten Jaccard-Index.
- Da Änderungen am HMM-Ansatz erforderlich waren, handelt es sich nicht um ein HMM.

Detaillierte Lösungsbeschreibung

Ziel war es, einen Ansatz zu entwickeln, der Hidden Markov Models (HMM) verwendet, um das aktuelle Handeln der Haushaltsbewohner zu identifizieren, d.h. welche Geräte an jeder Verbrauchsstelle verwendet werden. Die Lösung dieser Aufgabe ist in zwei Abschnitte unterteilt. Zunächst haben wir eine modifizierte Version eines HMM-Ansatzes zur Ableitung von Gerätezustandsabläufen basierend auf der individuellen Haushaltslastkurve und zusätzlichen Verbrauchsdaten der mittleren Geräte implementiert. Aufgrund von Datenbeschränkungen haben wir einen modifizierten HMM-Ansatz verwendet. Zweitens werden die Vorhersagen wurden anhand von Genauigkeit, Precision, Recall, F1 und Jaccard-Index anhand der Umfrageantworten, die zusammen mit den Smart Meter-Daten bereitgestellt werden, bewertet.

Bevor die Lösung beschrieben wird, wird ein Überblick über bestehende Ansätze zur Erkennung von Mustern im Energieverbrauch von Haushalten auf Basis von fein aufgelösten Daten gegeben. Dieser Bereich des Non-Intrusive Load Monitoring (NILM) versucht, zu erkennen, welche Geräte in einem Haushalt verwendet

werden, indem nur die in einen Haushalt fließende Spannung überwacht wird, die auch das Verbrauchsverhalten der Kunden sehr gut abdeckt und somit für unsere Aufgabe geeignet ist.

Verwandte Arbeiten:

Die verschiedenen Publikationen lassen sich in zwei Kategorien einteilen: In einem Teil der Publikationen stehen den Autoren Verbrauchsdaten mit hohen Abtastraten zur Verfügung, während die anderen Autoren nur niedrig aufgelöste Daten haben.

Ansätze für hohe Abtastraten: NILM oder Non-Intrusive Appliance Load Monitoring (NIALM) wird oft als Pionierarbeit von Hart (1992) bezeichnet. Mit der Kantenerkennung identifizierte er Zustände von Geräten in einer Lastkurve für Zustandsmodelle von Gebäudegeräten und stellte fest, ob Geräte an einzelnen Energieverbrauchsmessstellen ein- oder ausgeschaltet waren. Seine verfügbaren Daten hatten eine hohe Abtastrate von 1 Hz. Weitere Ansätze, die häufig als maschinelle Lernmethoden angesehen werden, insbesondere Hidden Markov Models (HMM) zur Modellierung des Gerätezustands. Wong et al. (2013) und Bonfigli et al. (2015) geben einen Überblick über NILM-Ansätze, einschließlich Ansätze mit Modellen des maschinellen Lernens. Die Grundidee von HMM ist es, eine Folge von Beobachtungen, in diesem Fall den Energieverbrauch, als Ausgabe einer Folge von versteckten Zuständen, in diesem Fall Gerätezuständen, zu modellieren. Das Ziel ist es, die wahrscheinlichste Folge von versteckten Zuständen für eine bestimmte Folge von Beobachtungen abzuleiten. Kim et al. (2011) haben die HMM-Methode zur effektiven Disaggregation von Verbrauchsmessungen erweitert. Ebenso erweiterten Parson et al. (2012) HMMs und kombinierten sie mit früheren Modellen von Geräten, ebenfalls mit hochgenauen Ergebnissen. Das häufigste Merkmal der NILM-Ansätze und -Methoden ist die Verwendung von Abtastraten von mehr als einmal pro Minute. Die meisten Smart Meter liefern keine Daten mit so hohen Abtastraten, da die Übertragungsraten von Powerline Communications (PLC) niedrig sind und weil Haushaltsverbrauchsdaten mit hohen Abtastraten es erlauben würden, sensible Informationen abzuleiten (Quinn 2009). In Deutschland sind daher beispielsweise Smart Meter mit Abtastraten größer als eine Messung pro 15 Minuten verboten.

Ansätze für niedrige Abtastraten: Nach unserem Kenntnisstand wurde noch nicht viel im Kontext der Geräte- oder Gerätezustandserkennung für niedrige Abtastraten publiziert, wahrscheinlich, weil es keine leichte Aufgabe ist. Bestehende Ansätze nutzen zwar kein HMM, aber werden der Vollständigkeit halber kurz darstellen. In einer Studie (Prudenzi 2002) werden 15-minütige Messungen verwendet und mit der überwachten Methode ANN analysiert, die detaillierte Vorabmessungen einzelner Geräte erfordert. Die gleiche Einschränkung gilt für die Arbeit von Kolter et al. (2010). Für deren Anwendung der „discriminative sparse coding" benötigt man auch vorherige Messungen von Einzelgeräten für das Training. Im Gegensatz dazu stellen Wytock und Kolter (2014) einen Ansatz vor, der explizit keine vorhergehende Messungen von Einzelgeräten benötigt. Stattdessen verwenden sie zusätzliche Kontextdaten (in diesem Fall Wetterdaten), weshalb sie es als „contextually supervised source separation". Die Identifizierung von Geräten und deren Verbrauch auf der Grundlage von aggregierten Daten mit geringer Stichprobengröße bleibt eine Herausforderung, für die es keinen etablierten Ansatz gibt.

Datenbeschreibung: Datensatz L besteht aus Smart-Meter-Daten für 8291 Haushalte in 15-Minuten-Zeitabständen. Der Datensatz wurde zwischen dem 04.06.2014 und dem 01.06.2015 gesammelt. Es ist in 52 wöchentliche Datendateien aufgeteilt, die jeweils 673 Messungen pro Haushalt enthalten. Zusätzlich zu den SMD wurde eine Umfrage bei den gleichen Kunden durchgeführt. Von allen Kunden wurden 109 Fragen beantwortet, darunter Fragen zu den Wohneigenschaften, Umweltansichten und Geräteinformationen. In der Umfrage konzentrierten sich 14 Fragen auf Geräteinformationen von Haushalten. Die Fragen sind in jeweils 2 Fragen (Anzahl und Alter der Geräte) für 7 Gerätetypen unterteilt, die in der Spalte "Gerätename" in Tabelle 27 zu sehen sind. Als dritte Datenquelle wurden mittlere Verbrauchsdaten von Geräten auf dem NILM-Wiki (http://wiki.nilm.eu), einer auf den NILM-Bereich spezialisierten Website, bezogen. Entweder wurde dort ein spezifischer Verbrauch für einen Gerätetyp angegeben oder ein Mittelwert berechnet, wenn für einen Gerätetyp ein Bereich angegeben wurde. In der Tabelle 27 sind der mittlere Geräteverbrauch und die Standardabweichung für die 7 Gerätetypen aus den NILM-Wiki-Daten angegeben.

Tabelle 27: Betrachtete Geräte und deren Verbrauch

Appliance name	Mean consumption (kW)	Standard Deviation
Tumble Dryer	2.50	0.10
Washing Machine	2.10	0.10
Electric Cooker	2.10	0.10
Dish washer	1.30	0.10
TV	0.16	0.01
Freezer	0.15	0.01
Fridge	0.08	0.01

HMM-Analyse Ansatz: Da die vorliegenden Daten von einer niedrigen Abtastrate sind und es keine etablierten HMM-Ansätze für niedrige Abtastraten gibt, mussten wir einen Ansatz für hohe Abtastraten verwenden und ihn an unsere Umstände anpassen. Wir haben uns für die Methode von Parson et al. (2012) entschieden, da die Grundidee immer noch mit niedrigauflösenden Daten übereinstimmt. Die Idee besteht darin, den beobachteten aggregierten Verbrauch als Einschränkung zu betrachten und den Verbrauch eines identifizierten Geräts abzuziehen. Andere komplexere Ansätze wären auf niedrigauflösende Daten nicht anwendbar. Parson et al. (2012) verwenden eine hohe Abtastrate von einer Minute und Verbrauchsmessungen einzelner Geräte für Training und Auswertung. Sie schlagen eine erweiterte Version des HMM vor, die sie „difference HMM" nennen, dargestellt in Abbildung 33. In der oberen Zeile der Abbildung stellt jeder Kreis einen Messpunkt des beobachteten aggregierten Verbrauchs dar. Der (versteckte) Gerätezustand in der mittleren Zeile ist der Zustand, den der Ansatz erkennen soll. Er kann als ein Gerät interpretiert werden, das z.B. EIN oder AUS ist. Zusätzlich wird in der unteren Zeile die Verbrauchsdifferenz zwischen zwei aufeinanderfolgenden Messpunkten als Information verwendet, um daraus abzuleiten, ob sich ein Gerätezustand zwischen den Beobachtungszuständen hätte ändern können.

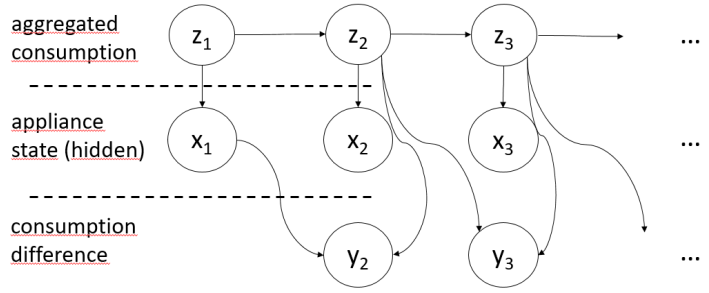
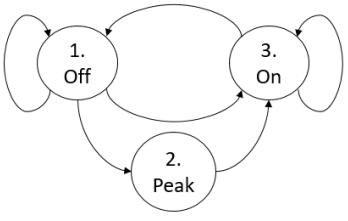


Abbildung 33: Difference HMM nach Parson et al. (2012)

Die Autoren verwenden frühere Modelle des Leistungsbedarfs und der Zustandsübergangswahrscheinlichkeiten von Gerätetypen als Parameter für das HMM, die mit Perioden trainiert wurden, in denen nur ein Gerät seinen Zustand ändert. Dieses Modell ist in Abbildung 34 übersichtlich dargestellt. (a) ist eine Darstellung der möglichen Zustände eines Kühlschranks und der möglichen Zustandsübergänge. (b) zeigt den mittleren Stromverbrauch und die Varianz des Kühlschranks, während in (c) die Zustandsübergangswahrscheinlichkeiten zu sehen sind.



(a)

z_t	μ	σ
1	0.04 kW	0.1
2	0.24 kW	0.15
3	0.19 kW	0.1

(b)

z_t	1	2	3
1	0.95	0.03	0.02
2	0	0	1
3	0.1	0	0.9

(c)

Abbildung 34: Ausgangsmodelle für einen Kühlschrank als Zustandsgraph (a), der Erwartungswert und die Standardabweichung (b) sowie die Wahrscheinlichkeiten für die Zustandsübergänge

Anhand dieser Informationen trainieren die Autoren ein difference HMM ein für jeden Gerätetyp und leiten eine Statussequenz aus den aggregierten Verbrauchsmessungen ab. Dazu kombinieren sie die Wahrscheinlichkeit, dass ein Gerät seinen Zustand ändert, basierend auf der Differenz im Verbrauch, dem Verbrauch eines Gerätetyps in jedem Zustand und der Übergangswahrscheinlichkeit. Parson et al. machen die Annahme, dass sich ein Gerät nur dann in einem Zustand befinden kann, wenn der aktuell gemessene Verbrauch deutlich höher ist als der Verbrauch des Gerätes in diesem Zustand. Mit all dem ermitteln sie den „hidden state" des Gerätes zu jedem Zeitpunkt. Wenn der Zustand bestimmt wurde, subtrahierten sie den Verbrauch des Geräts vom aggregierten Verbrauch, um die Stromeinstellung für den nächsten Gerätetyp zu korrigieren.

Modifizierter HMM-Ansatz: Da in unserem Fall das Ausgangsmodell für die Geräte und die Einzelgerätemessungen für die Haushalte nicht verfügbar sind, mussten wir diesen Ansatz auf drei Arten modifizieren:

1. Differenz im Verbrauch ausschließen,
2. Berücksichtigen von nur einem Zweistaatenmodell für alle Geräte und,
3. Verwenden eines vereinfachten Ausgangsmodells für Gerätetypen aus den mittleren Verbräuchen.

Wir haben den Unterschied im Verbrauch zwischen zwei aufeinanderfolgenden Messungen ausgeschlossen, da er nicht verwendet werden kann, um eine Wahrscheinlichkeit für eine Zustandsänderung eines Gerätes abzuleiten. Dies ist auf das geringe Abtastintervall von 15 Minuten zurückzuführen, denn in 15 Minuten ist es unwahrscheinlich, dass nur ein einziges Gerät seinen Zustand ändert. Außerdem haben wir nur ein Zwei-Zustands-Modell anstelle eines Drei-Zustands-Modells betrachtet. Dies liegt ebenfalls an der niedrigen Abtastrate. Es ist nämlich unwahrscheinlich, dass ein "Peak"-Zustand zwischen einem 15-minütigen Intervall beobachtet wird. Daher haben wir den "Peak"-Zustand aus dem Zustandsdiagramm entfernt. Die dritte Änderung war die Verwendung vereinfachter Ausgangsmodelle für die Gerätetypen, da es für die Haushalte keine solchen gab. Alles, was wir verwenden konnten, waren die mittleren Verbräuche der Gerätetypen als Ausgangsmodelle.

Zusammenfassend lässt sich sagen, dass unser Ansatz nur in den Grundüberlegungen ein HMM ist, Wahrscheinlichkeiten für den Übergang von einem Status zum nächsten können allerdings mit den niedrigen Abtastraten bei Smart-Meter-Daten nicht bestimmt werden.

Bewertung: Da keine anderen Ground Truth Daten als die Informationen darüber vorliegen, ob einzelne Geräte in den jeweiligen Haushalten vorhanden sind, konnten wir nur beurteilen, ob Geräte korrekt identifiziert wurden, die in Haushalten vorhanden sind. Dazu gingen wir davon aus, dass, wenn ein Gerät mindestens einmal im Haushalt eingeschaltet war. Wenn beispielsweise das Gerät "Wäschetrockner" in der Zustandsfolge eines Haushalts mehrfach als eingeschaltet identifiziert wurde, kann davon ausgegangen werden, dass der Haushalt einen Wäschetrockner besitzen sollte. Diese Methode wurde für jedes Gerät in jedem Haushalt angewendet. Das Ergebnis waren Konfusionsmatrizen für alle Geräte, auf die wir 4 verschiedene Gütemetriken berechnet haben: Accuracy, Precision, F1 und den Jaccard-Index. Die Ergebnisse sind in Abbildung 35, 36, 37 und 38 dargestellt.

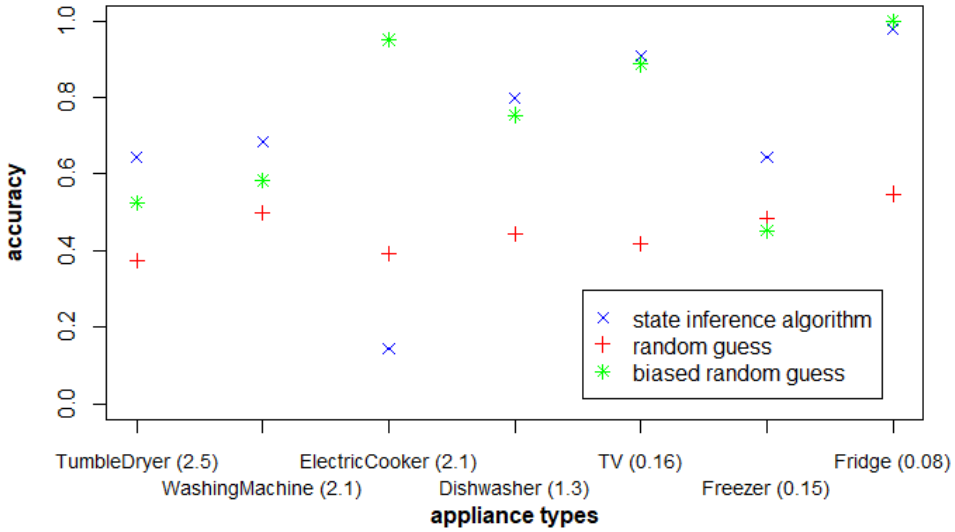


Abbildung 35: Accuracy für jedes Gerät mit Vergleichsmaßen

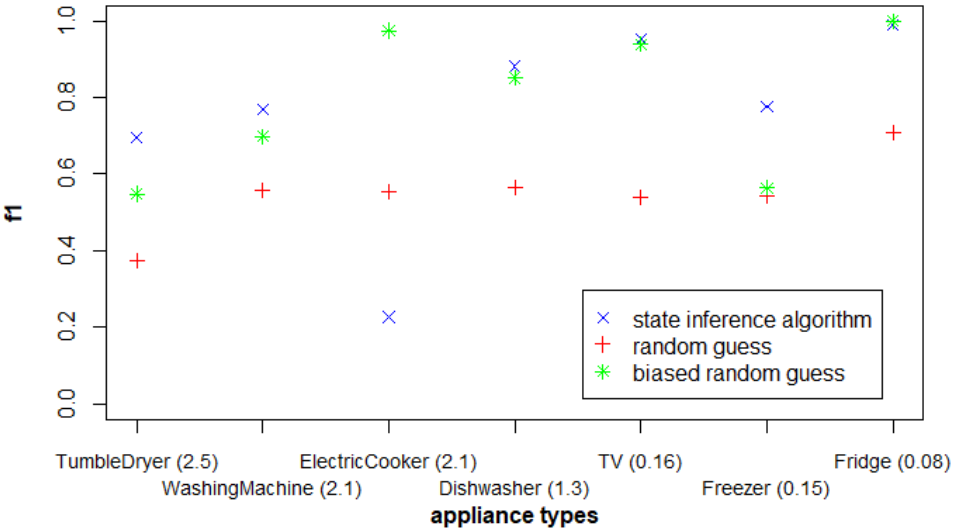


Abbildung 36: F1-Maß für jedes Gerät mit Vergleichsmaßen

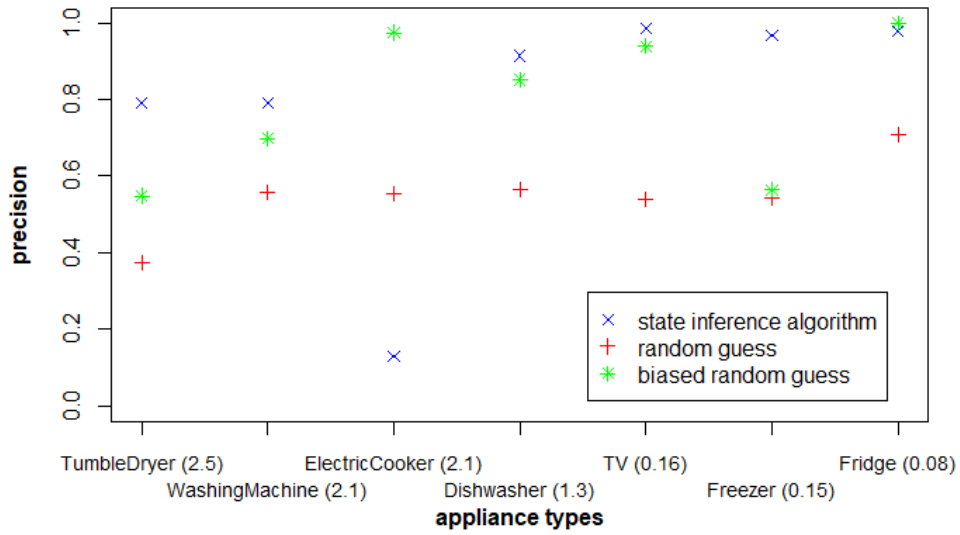


Abbildung 37: Precision für jedes Gerät mit dem BRG Vergleichsmaß

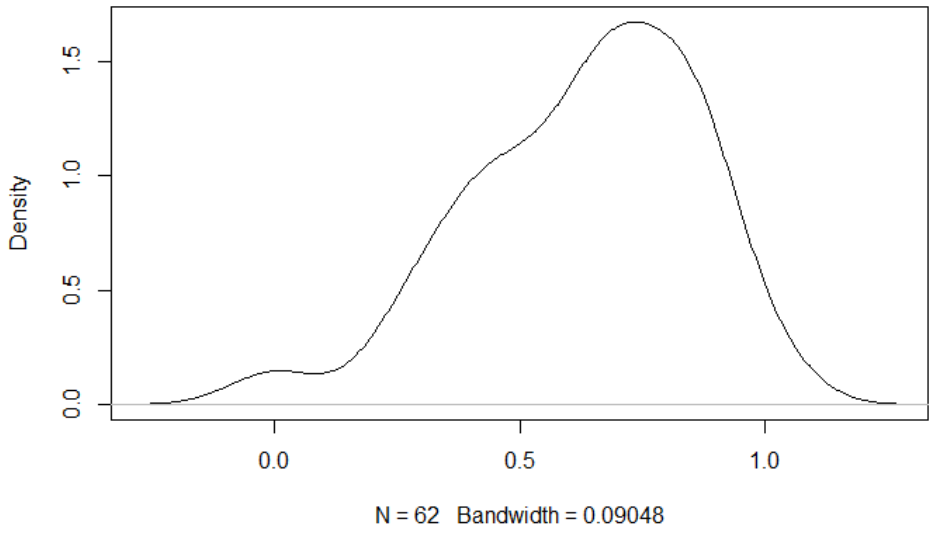


Abbildung 38: Jaccard-Index Verteilung für Woche 1

Testen von evolutionären Algorithmen (Aufgabe 4.14)

Jedes Lastprofil kann als Ansammlung von Verbrauchssegmenten gesehen werden (z.B. Grundlastverbrauch, Duschen am Morgen, Fernsehen am Abend). Wir haben die Erkennung relevanter Haushaltsimmobilien mit evolutionären Algorithmen getestet, um eine Bibliothek solcher Segmente zu generieren und eine Teilmenge der Segmente an jede Lastkurve anzupassen. Wir fanden keine signifikante Verbesserung gegenüber den anderen in diesem Projekt entwickelten Ansätzen. Darüber hinaus wurden bereits Verbrauchssegmente (z.B. Basisverbrauch, Dusche am Morgen, TV am Abend) als Merkmale in Arbeitspaket modelliert. Dies kann ein Grund für die geringe Leistung von genetischen Algorithmen für das aktuelle Problem sein.

Einbeziehung von Verbrauchstrends in die Modelle und Einsatz von Zeitreihenanalyseverfahren (Aufgabe 4.15 und 4.16)

Wir haben Wissen über die Zeitreihendaten und Verbrauchstrends in der Feature-Extraktionsphase (Arbeitspaket 3) berücksichtigt. Die Synchronisation von Strom-, Gas- und Wasserverbrauchsdaten, sowie externen

Daten wurde durchgeführt. Hierzu wurden multidimensionale Zeitreihenanalysetechniken entwickelt und angewendet, wie in den vorhergehenden Abschnitten beschrieben.

AP 5: Präskriptive Verbrauchsdatenanalyse und Tool-Unterstützung

Spezifizieren von Klassenlabels für spezielle Klassen (Aufgabe 5.3)

Haushaltseigenschaften sind die abhängigen Variablen im betrachteten Vorhersagemodell. Einige Properties können potentiell eine große Anzahl von Werten annehmen oder sind nicht diskret (z.B. Wohnfläche, Alter des Hauses). Auf der anderen Seite sind in einige Ausprägungen die Mehrheit der Haushalte enthalten und können demnach keine solide Entscheidungsfindung unterstützen (Imbalanced-Class-Problem). Diese Klassen mussten ggf. in detailliertere Klassen unterteilt werden. Tabelle 28 zeigt die Klassendefinition für Eigenschaften, die im Mittelpunkt dieses Projekts stehen, zusammen mit der Motivation für die Klassendefinition.

Wir haben die Haushaltseigenschaften detailliert in Klassen eingeteilt. Die Klassen wurden auf der Grundlage von neun Interviews (durchgeführt durch BEN Energy) definiert. Für die Klassen ist es wichtig, dass sie

a) nicht zu zahlreich sind, um sie sinnvoll und umsetzbar zu halten, und
b) dass sie widerspiegeln, was aus betriebswirtschaftlicher Sicht für den Nutzen interessant ist.

So sind beispielsweise die Versorgungsunternehmen für ihr Produkt- und Dienstleistungsangebot nicht interessiert, wie viele Kinder es in einem Haushalt gibt, sondern die Information, ob es Kinder gibt oder nicht, ist ausreichend. Die Entscheidung, welche Klassen sinnvoll sind, ist daher abhängig vom Portfolio des Versorgungsunternehmens. Außerdem sind Klassen manchmal überhaupt nicht relevant. Für Kampagnen haben Marketingabteilungen ein bestimmtes Budget und müssen die vielversprechendsten Interessenten unter ihren Kunden auswählen. Dafür benötigen sie numerische Werte und keine Klassen.

Die Klasse für die Eigenschaften „Wohnungsgröße“ und „Anzahl der Bewohner“ können ganzzahlige Werte einfügen. Wir haben die Klassen für diese Eigenschaften gemäß Hopf, Kozlovskiy, Sodenkamp (2016). Für die Eigenschaften „Haushaltstyp“, „Raumwärmetyp“ und „Wasserwärmetyp“ wurden die Klassenbezeichnungen direkt aus den Umfrageergebnissen übernommen. Wir haben auch Klassendefinitionen aus einem früheren Projekt (Sodenkamp u. a. 2016) und Haushaltsklassifizierungsarbeiten (Beckel u. a. 2014) verwendet.

Tabelle 28: Haushaltseigenschaften und Klassen

Household property	**Class**	**Definition**
Up-sell potential	Low	1. quartile of Upsell probability %
	Medium	2. quartile
	High	3. quartile
	Very high	4. quartile
Cross-sell potential	Low	1. quartile
	Medium	2. quartile
	High	3. quartile
	Very high	4. quartile
Churn risk (new supplier)	Low churn risk	<0.9% (0.5 * Avg)
	Average churn risk	0.9 – 3.8% (0.5*Avg – 2*Avg)
	High churn risk	3.8 – 10% (2*Avg – 5*Avg)
	Very high churn risk	>10% (>5*Avg)
Mailing selection	No class	Sign-up probability
Living area	≤ 95 m^2	The variable living area takes integer values in the range of 10 to 5’443. Therefore, any definition of this property is ambiguous. We defined the class borders at 95 m^2 and 145 m^2 based on the following motivation: First, the class borders are empirically defined and
	≤ 145 m^2	
	> 145 m^2	

		based on quantiles. The 33% quantile is 100m^2, the 66% quantile is 150m^2, and the 99% quantile is 400m^2. Since we assume that people estimate their living area in a survey to the next upper bound, we define the categories 5m^2 below this round number. Second, we find further evidence in our class definition in European statistics (Statistical Office of the European Communities 2014, 54): the average dwelling size in the EU-28 countries is 95.9 m^2, in Switzerland it is according to the statistics 117.1 m^2.
Number of residents	1	The number of residents in a household takes fewer values than the living area, but the variable has nevertheless a range of 1 to 10 household and the class borders can be defined ambiguously. We tested a set of definitions in the classification: a) 1 / 2 / >2, b) 1 / 2 / 3-5 / >5, c) 1 / 2 / 3 / 4 / >4, d) 1 / >1. Our results show that the definition (b) has the best trade-off between gained information, number of classes and classification performance. Therefore, we include only this definition in this paper.
	2	
	3-5	
	> 5	
Household type	house	The variables were raised with the categories and we use this definition from the survey.
	apartment	
Space heating type	electric	
	not electric	
Water heating type	electric	
	not electric	
Social class[5]	A or B	Social class of chief income earner according to NRS social grades
	C1 or C2	
	D or E	
House ownership[5, 6]	Property	Land tenur
	Rent	
Relationship status[5]	Single	Number of adults = 1 and number of children = 0
	Non- Single	Else
Children[5,6]	Children	Number of children > 0
	No children	Number of children = 0
Family[5,6]	Family	Number of adults > 1 and number of children > 0
	No Family	Else
Employment[5]	Employed	Employment of the chief income earner
	Not employed	
Retirement[5]	Retired	Retirement status of the chief income earner
	Not retired	
Number of bedrooms[5]	Very low	1 or 2 bedrooms
	Low	3 bedrooms
	High	4 bedrooms
	Very high	Num. of bedrooms > 4
Cooking type[5]	Electric	Number of electric stoves Herd > 0
	Non-electric	Number of electric stoves = 0
Desktop PC[5]	Yes	Desktop PC exixtent?
	No	

[5] This property was subject to a prior project with dataset A and is not investigated in this project

[6] This property was subject to a prior project with dataset L and is not investigated in this project

Laptop[5]	Yes	Laptop existent?
	No	
Dryer[5]	Yes	Dryer existent?
	No	
Dishwasher[5]	Yes	Dishwasher existent?
	No	
Freezer[5]	Yes	Freezer existent?
	No	
Game console[5]	Yes	Game consols existent?
	No	
House age[5]	10-29	Age of the house
	30-74	
	<10	
	>=75	
Age of heating system (space)[6]	New	Age of heating system < $q_{1/3}$
	Average	Age of heating system from $q_{1/3}$ to $q_{2/3}$
	Old	Age of heating system > $q_{2/3}$
Heat pump[6]	No	Existing heat pump
	Yes	
Number of appliances[6]	Low	Number of appliances Ø < $q_{0,25}$
	Average	Number of appliances Ø from $q_{0,25}$ to $q_{0,75}$
	Many	Number of appliances Ø > $q_{0,75}$
PV[6]	Yes	Customer indicated to have a photovoltaik solar thermal installation
	No	
Age/ efficiency of appliances[6]	New	Medium age < $q_{0,25}$
	Medium	Medium age from $q_{0,25}$ to $q_{0,75}$
	Old	Medium age > $q_{0,75}$
Satisfaction with supplier[6]	Low	Sum FN-27 < $q_{0,25}$
	Medium	Sum FN-27 from $q_{0,25}$ to $q_{0,75}$
	High	Sum FN-27 > $q_{0,75}$
Interest PV[6]	Gering	Sum FN-21 < $q_{0,50}$
	Mittel	Sum FN-21 from $q_{0,50}$ to $q_{0,75}$
	Hoch	Sum FN-21 > $q_{0,75}$
Interest Fiber to the Home[6]	High	Purchase intention >= 8
	Not high	Purchase intention < 8
Energy efficiency measure[6]	None	Number of implemented energy efficiency measures
	One	
	Multiple	

II.2 Wichtigste Positionen des zahlenmäßigen Nachweises

Der zahlenmäßige Nachweis umfasst zwei Positionen: Die Position „0812“ (Personalkosten, entstandene Ausgaben von 252.579,37 €) und die Position „0846“ (Dienstreisen, entstandene Kosten 6.067,35 €). Die Personalkosten resultieren aus dem Personalaufwand der wissenschaftlichen Mitarbeiter, die in Vollzeit an dem Projekt gearbeitet haben. Im Vorhergehenden Abschnitt II.1 wurden die Ergebnisse dieser Arbeit ausführlich dargestellt. Die Reisekosten entstanden durch Projekttreffen mit dem Forschungspartner und Forschergruppen, mit denen an der wissenschaftlichen Verwertung der Projektergebnisse gearbeitet wurde.

II.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die Implementierung von Vorhersagealgorithmen ist eine große Herausforderung und benötigt fachliches Knowhow, um die anspruchsvollen Fragestellungen aus dem Bereich der Wirtschaftsinformatik, der Datenwissenschaften und des maschinellen Lernens zu beantworten. In diesem Projekt wurden insbesondere die folgenden sieben Herausforderungen angegangen und die Grenzen der technischen Lösungen im Bereich der Vorhersagesysteme im Energievertrieb vorangetrieben:

Mining mehrdimensionaler Daten: Der von unserer Gruppe in Bamberg mit Unterstützung des Bundesamtes für Energie (Schweiz) entwickelte Algorithmus zur Klassifizierung von rudimentären Lastkurven erreicht bei 12 Haushaltsobjekten eine Genauigkeit von knapp über 60%. Um die Fehlerquote zu reduzieren und die Ergebnisqualität deutlich zu erhöhen, reichte es nicht aus, die vorhandenen Klassifikatoren zu verwenden. Daher müssen innerhalb der Projektlaufzeit neuartige multidimensionale Machine-Learning-Algorithmen entwickelt werden, um mit großen, miteinander verbundenen Datensätzen unterschiedlicher Art fertig zu werden.

Fluch der Dimensionalität: Es gibt eine Reihe von Methoden, die die Daten in eine niedrigere Dimension transformieren, indem sie "sinnvolle" Datenattribute (z.B. Minima, Maxima, Varianz) extrahieren, die in der Klassifizierung selbst verwendet werden. Die Herausforderung besteht darin, einen umfassenden Funktionsumfang zu finden, der spezifisch für den Daten- und Problemkontext ist. Der gängige empirische Ansatz hängt stark von den Expertenbeiträgen ab. Das Ergebnis kann Hunderte oder sogar Tausende subjektiv definierter Variablen beinhalten. Um unser Werkzeug vollautomatisch, schnell und wirklich präzise zu machen, haben wir neue Methoden der automatischen Merkmalsextraktion mit unbeaufsichtigtem Lernen entwickelt und getestet.

Übertragbarkeit: Die Klassifizierung basiert auf den Trainingsdaten, die mit der jeweiligen geografischen Region verknüpft sind. Die Herausforderung bestand darin, das Data-Mining-Tool für letztlich jedes Land bzw. jede Region ohne großen Anpassungsaufwand und wiederholtes Lernen einsetzbar zu machen (die Trainingsdaten sind für Versorgungsunternehmen eher nicht verfügbar). Dazu mussten wir die Ursachen für die Resonanz in den regionalen Datensätzen aufdecken und als Modellparameter einbetten.

Große Datenmengen/Skalierbarkeit: Die Menge der erfassten Verbrauchsdaten wird schnell wachsen. Datenverwaltungsmethoden (z. B. Parallelisierung, Cloud Computing, verteilte Datensysteme) sind notwendig, um auf große Datensätze zugreifen zu können. Darüber hinaus muss eine gute zeitliche, räumliche und strukturelle Skalierbarkeit des Systems gewährleistet sein, um mit potenziell noch größeren Datenmengen umgehen zu können.

Datengranularität: Die Effizienz und Effektivität neuer Klassifikationsalgorithmen hängt von der Häufigkeit der Datenerhebung und von der Jahreszeit ab. Wenn die zu analysierenden/trainierenden Daten nicht die gleiche Granularität aufweisen (was ein häufiger Fall ist), muss das System diese noch bewältigen. Um diese Schwierigkeit zu überwinden, mussten wir Strategien der Datenpartitionierung und -verknüpfung entwickeln.

Datenmanagement: Mechanismen der Datensuche, -erfassung, -aktualisierung und -zusammenstellung. Integrierter und privater Umgang mit verschiedenen Datentypen (Zeitreihen, Querschnitts- und Abfragedaten, Web-Links, usw.).

Semantische Daten: Obwohl geografische Informationssysteme und semantische Daten aus dem Internet einen potenziellen Mehrwert darstellen, ist die Erfassung und Verarbeitung schwierig und zeitaufwändig. Wir verfolgten ehrgeizige Ziele bei der Entwicklung von Methoden der Bilderkennung und des Text-Mining für die Energieverbrauchsforschung und konnten gute Fortschritte erreichen.

Schutz der Privatsphäre: Verbrauchsdaten können Haushaltsmerkmale und Verhaltensmuster aufdecken. Die Daten bergen daher potenzielle Datenschutzrisiken. Trotz dieser offensichtlichen Bedrohung berücksichtigen nur eine Handvoll verwandter Studien Datenschutzaspekte bei der Gestaltung. Dennoch wurden in anderen Bereichen eine Reihe von Techniken vorgeschlagen, um Datenanalysen auf privatwirtschaftliche Weise durchzuführen. Diese Techniken stammen aus einer Vielzahl von Themen wie Data Mining, Kryptographie

und Information Hiding. Da der Datenschutz ein weites Feld ist, mussten wir uns auf die Technologien konzentrieren, die für dieses Projekt gemäß der bestehenden und zukünftigen EU-Gesetzgebung besonders relevant waren.

II.4 Voraussichtlicher Nutzen und fortgeschriebener Verwertungsplan

Die Forschung in diesem Projekt zielte darauf ab, ein Bindeglied zwischen verfügbaren Energieverbrauchsinformationen und wirkungsvollen, massenmarkttauglichen Energieeffizienzmaßnahmen herzustellen. Wir haben Machine-Learning-Techniken entwickelt, verbessert und getestet, welche – im Einvernehmen mit den Stromkundinnen und Stromkunden – automatisch energierelevante Haushaltscharakteristiken aus Energieverbrauchsrelevanten Daten ableiten. Die Verfahren ermöglichen es schlussendlich, datenbasierte Energiedienstleistungen und Effizienz- und Kundenbindungskampagnen und großflächig und kostengünstig anzubieten.

Aus *wissenschaftlich-technischer Sicht* leistet unser Teilprojekt einen Beitrag auf zwei Ebenen:

Ebene 1: Die Entwicklung von neuen Big-Data-Analytics- und Machine-Learning-Techniken, insbesondere:

(a) Neue Methoden der multidimensionalen Klassifikation, die in der Lage sind, komplexe Zusammenhänge zwischen Variablen (wie der quantifizierte Zusammenhang zwischen Temperatur oder Haushaltstyp und Stromverbrauch) zu berücksichtigen.

(b) Verwendung von Machine-Learning-Methoden zur automatischen Feature-Identifikation und -Extraktion. Dieser Ansatz basiert auf einer empirischen Featuredefinition und automatischen Verfahren zur Merkmalsextraktion.

(c) Geographisch und semantisch unterstütztes Mining von Energiedaten, um Haushaltscharakteristiken aus externen Datenquellen zu erkennen. Beispielsweise die Abschätzung der Wohnfläche und das Alter des Hauses basierend auf den Nachbarschafts- oder Geodaten. Die Verteilung der Haushaltscharakteristiken wird aufgrund der vorhandenen Daten geschätzt und mit den erhobenen Daten verglichen.

(d) Verminderung des Bias durch die Schätzung der systematischen Fehler in den Zeitreihen und Anpassung der Klassifikationsergebnisse.

Ebene 2: Die entwickelten Tools ermöglichen gezielte Verhaltensinterventionen, indem mehr Informationen über Energiekunden verfügbar sind. So können Energieeffizienzkampagnen, wie wir in einer exemplarischen Feldstudie im Rahmen des Projekts gezeigt haben, in der Lage sein den Energieverbrauch im Privatsektor zu verringern (z.B. durch Verhaltensänderung oder Änderung an den Haushaltscharakteristika). Langfristig wird dies helfen, die Energie- und Lastverschiebungsmaßnahmen voranzubringen.

Wirtschaftliche Verwertung: Die gemeinsam entwickelten Algorithmen werden von der BEN Energy AG in bestehenden Analytics- und Smart-Metering-Produkten verwendet und als Software-as-a-Service sowie Plattform-as-a-Service Dienstleistungen verschiedenen Energieversorgern zur Verfügung gestellt. Im Rahmen des Projekts konnten fünf Pilotstudien mit Energieversorgungsunternehmen in Deutschland und der Schweiz durchgeführt werden, in denen die Algorithmen angewendet und bereits von BEN Energy vermarktet wurden.

Erfindungen bzw. Schutzrechtsanmeldungen: Die entwickelte Software fällt unter Urheberrecht. Entsprechend kann kein Patent innerhalb der EU angemeldet werden. Ein starker Schutz ergibt sich aus den im Projekt und in der späteren Anwendung gesammelten Trainingsdaten, die für die Instanziierung der Algorithmen notwendig und dem Wettbewerb nicht zugänglich sind.

Wissenschaftliche Verwertung: Wissenschaftliche Ergebnisse werden zeitnah in hochrangingen Zeitschriften und Konferenzen veröffentlicht. Die erreichten und geplanten Veröffentlichungen sind im Abschnitt II.6 genannt.

Dissemination der Projektergebnisse in die Öffentlichkeit: Die Universität Bamberg ist Mitglied des Bits to Energy Lab, einer Forschungskooperation, bestehend aus der Otto-Friedrich-Universität Bamberg, der ETH Zürich

und der Universität St. Gallen, welche sich auf Forschung im Bereich der Informationssysteme und der Verhaltensökonomie konzentriert, um energieeffizientes Verhalten bei Verbrauchern zu fördern. Das aus dem Projekt resultierte Know-how in den Bereichen Dimensionsreduktion, maschinelle Lernverfahren und deren Anwendung wird auch Basis für weitere Forschungsprojekte sein. Die erreichten Erkenntnisse werden auch in der Lehre an Studierende weitergegeben und wurden in folgenden Lehrveranstaltungen integriert:

- **Data Analytics in der Energieinformatik:** 6 ECTS Kurs auf Master-Niveau (Inhalte: Analysemethoden der Energieinformatik und deren Anwendung auf praktisch relevante Aufgabenstellungen, um wirkungsvolle Energiedienstleistungen zu entwickeln.)
- **Business Intelligence & Analytics:** 6 ECTS Kurs auf Master-Niveau (Inhalte: Fundamentale Konzepte und Methoden der Datenanalyse und der modernen Entscheidungstheorie und -praxis. Die Modulschwerpunkte liegen auf prädiktiven und präskriptiven Analysen, welche Unternehmen zu einer besseren Einsicht in Prozesse und Entscheidungen verhelfen.)
- **Smart Grid Data Analytics:** 6 ECTS Projekt für Bachelor- und Master-Studierende mit wechselnden Themen der Datenanalyse in der Energieinformatik

Anschlussfähigkeit in der Forschung: Die Projektergebnisse dienen als Grundlage für das Forschungsvorhaben „SmartLoad“ (Förderkennzeichen: 0350010), das im Rahmen der ERA-Net SG+ Initiative durchgeführt wird. Darüber hinaus laufen weitere Forschungsanträge für Vorhaben, in denen die Algorithmen, welche in diesem Projekt entwickelt wurden, weiterverwendet und verbessert werden.

II.5 Während der Durchführung bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

In verschiedenen Arbeitspaketen (Aufgaben 2.7, 3.9, 3.11 und 4.1) wurden ausführliche Literaturrecherchen in anerkannten Datenbanken, wie Google Scholar, Business Source Premier (via EBSCO Host), ACM Digital Library und ScienceDirect zum aktuellen Forschungsstand durchgeführt; der Projektpartner BEN Energy hat ausführliche Marktanalysen für alle Zielregionen durchgeführt, die für die spätere wirtschaftliche Verwertung der Projektergebnisse in Frage kommen. Der bisherige Stand der Technik, Forschung und Praxis wurde in der Projektarbeit berücksichtigt. Nach wie vor konnten wir keine Arbeiten identifizieren, welche Teilaspekte dieses Projekts negativ beeinträchtigen oder die aus dem Projekt entstehenden Ergebnisse obsolet machen.

II.6 Erfolgte und geplante Veröffentlichungen der Ergebnisse

Folgende wissenschaftliche Beiträge wurden bereits veröffentlicht:

Hopf, Konstantin. 2018. „Mining Volunteered Geographic Information for Predictive Energy Data Analytics“. *Energy Informatics*, Nr. 1:4 (July). https://doi.org/10.1186/s42162-018-0009-3.

Hopf, Konstantin. 2019. „Predictive Analytics for Energy Efficiency and Energy Retailing“. Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg, Band 36, Bamberg: University of Bamberg Press. https://doi.org/10.20378/irbo-54833

Hopf, Konstantin, Michael Kormann, Mariya Sodenkamp, und Thorsten Staake. 2017. „A Decision Support System for Photovoltaic Potential Estimation“. In *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, 3:1–3:10. IML '17. New York, NY, USA: ACM. https://doi.org/10.1145/3109761.3109764.

Hopf, Konstantin, Sascha Riechel, Mariya Sodenkamp, und Thorsten Staake. 2017. „Predictive Customer Data Analytics – The Value of Public Statistical Data and the Geographic Model Transferability“. In *ICIS 2017 Proceedings*. Seoul, South Korea: AIS electronic library.

Hopf, Konstantin, Mariya Sodenkamp, und Ilya Kozlovskiy. 2016. „Energy data analytics for improved residential service quality and energy efficiency“. In *ECIS 2016 Research in Progress Proceedings*. Istanbul, Turkey: AIS electronic library.

Hopf, Konstantin, Mariya Sodenkamp, und Thorsten Staake. 2018. „Enhancing energy efficiency in the residential sector with smart meter data analytics“. *Electronic Markets* 28 (4). https://doi.org/10.1007/s12525-018-0290-9.

Kozlovskiy, Ilya, Mariya Sodenkamp, Konstantin Hopf, und Thorsten Staake. 2016. „Energy informatics for environmental, economic and social sustainability: A case of the large-scale detection of households with old heating systems“. In *ECIS 2016 Proceedings*. Istanbul, Turkey: AIS *electronic library*.

Ein weiterer Beitrag befindet sich in Vorbereitung.

Referenzen

Albert, A., und R. Rajagopal. 2014. „Cost-of-Service Segmentation of Energy Consumers“. *IEEE Transactions on Power Systems* 29 (6): 2795–2803. https://doi.org/10.1109/TPWRS.2014.2312721.

Alfares, Hesham K, und Mohammad Nazeeruddin. 2002. „Electric load forecasting: literature survey and classification of methods“. *International Journal of Systems Science* 33 (1): 23–34.

Alfaro, Esteban, Matıas Gámez, und Noelia Garcia. 2013. „Adabag: An R package for classification with boosting and bagging“. *Journal of Statistical Software* 54 (2): 1–35.

Al-Otaibi, R., N. Jin, T. Wilcox, und P. Flach. 2016. „Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data“. *IEEE Transactions on Industrial Informatics* 12 (2): 645–54. https://doi.org/10.1109/TII.2016.2528819.

Anastasiadis, Aristoklis D., George D. Magoulas, und Michael N. Vrahatis. 2005. „New globally convergent training scheme based on the resilient propagation algorithm“. *Neurocomputing* 64: 253–270.

Baskent, Emin Z, und Glen A Jordan. 1995. „Characterizing spatial structure of forest landscapes“. *Canadian Journal of Forest Research* 25 (11): 1830–49.

Beckel, Christian, Leyna Sadamori, und Silvia Santini. 2012. „Towards automatic classification of private households using electricity consumption data“. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, herausgegeben von George J. Pappas, 169–176. Toronto: ACM.

———. 2013. „Automatic socio-economic classification of households using electricity consumption data“. In , herausgegeben von David Culler, Catherine Rosenberg, S. Keshav, und Jim Kurose, 75. Berkeley, California, USA: ACM Press. https://doi.org/10.1145/2487166.2487175.

Beckel, Christian, Leyna Sadamori, Thorsten Staake, und Silvia Santini. 2014. „Revealing household characteristics from smart meter data“. In *Energy*, 78:397–410.

Becker, Matthias. 2012. „Geodesy“. In *Springer Handbook of Geographic Information*, herausgegeben von Wolfgang Kresse und David M. Danko, 95–117. Berlin, Heidelberg: Springer.

BFS. 2006. „Amtliches Gemeindeverzeichnis der Schweiz“. Swiss Federal Statistical Office. www.bfs.admin.ch/bfs/portal/de/index/news/publikationen.Document.80465.pdf.

Biau, Gérard. 2012. „Analysis of a Random Forests Model“. *Journal of Machine Learning Research* 13 (Apr): 1063–95.

Bonfigli, Roberto, Stefano Squartini, Marco Fagiani, und Francesco Piazza. 2015. „Unsupervised algorithms for non-intrusive load monitoring: An up-to-date overview“. In *Environment and Electrical Engineering (EEEIC), 2015 IEEE 15th International Conference on*, 1175–80. IEEE.

Breiman, Leo. 2001. „Random forests“. *Machine Learning* 45 (1): 5–32.

Briant, Anthony, P-P Combes, und Miren Lafourcade. 2010. „Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?“ *Journal of Urban Economics* 67 (3): 287–302.

Carley, Sanya. 2012. „Energy demand-side management: New perspectives for a new era“. *Journal of Policy Analysis & Management* 31 (1): 6–32. https://doi.org/10.1002/pam.20618.

Chandrashekar, Girish, und Ferat Sahin. 2014. „A survey on feature selection methods“. *Computers & Electrical Engineering*, 40th-year commemorative issue, 40 (1): 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Chatzimichali, Eleni, und Conrad Bessant. 2015. *classyfire: Robust multivariate classification using highly optimised SVM ensembles*. https://CRAN.R-project.org/package=classyfire.

Chen, Tianqi, und Carlos Guestrin. 2016. „XGBoost: A Scalable Tree Boosting System“. In *arXiv:1603.02754 [cs]*, 785–94. San Francisco, USA. https://doi.org/10.1145/2939672.2939785.

Chicco, Gianfranco. 2012. „Overview and performance assessment of the clustering methods for electrical load pattern grouping“. *Energy* 42 (1): 68–80. https://doi.org/10.1016/j.energy.2011.12.031.

Cox, D. R. 1958. „The Regression Analysis of Binary Sequences“. *Journal of the Royal Statistical Society. Series B (Methodological)* 20 (2): 215–42.

Cramer, H. 1946. *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Dietz, Robert D. 2002. „The estimation of neighborhood effects in the social sciences: An interdisciplinary approach“. *Social Science Research* 31 (4): 539–75. https://doi.org/10.1016/S0049-089X(02)00005-4.

Dixon, Matthew, Karen Freeman, und Nicholas Toman. 2010. „Stop trying to delight your customers“. *Harvard Business Review* 88 (7/8): 116–22.

DMLC. 2016a. „Notes on Parameter Tuning — xgboost 0.6 documentation“. 2016. http://xgboost.readthedocs.io/en/latest//how_to/param_tuning.html.

———. 2016b. „XGBoost Parameters — xgboost 0.6 documentation“. 2016. http://xgboost.readthedocs.io/en/latest//parameter.html.

Emanet, Nahit, Halil R Öz, Nazan Bayram, und Dursun Delen. 2014. „A Comparative Analysis of Machine Learning Methods for Classification Type Decision Problems in Healthcare“. *Decision Analytics* 1 (1): 6. https://doi.org/10.1186/2193-8636-1-6.

European Parliament. 2003. *Regulation 1059/2003*. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32003R1059.

Fahrmeir, Ludwig, Rita Künstler, Iris Pigeot, und Gerhard Tutz. 2007. *Statistik*. Springer.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, und Dinani Amorim. 2014. „Do we need hundreds of classifiers to solve real world classification problems?“ *The Journal of Machine Learning Research* 15 (1): 3133–81.

Fielding, Alan H, und John F Bell. 1997. „A review of methods for the assessment of prediction errors in conservation presence/absence models“. *Environmental conservation* 24 (01): 38–49.

Flath, Christoph, David Nicolay, Tobias Conte, Clemens van Dinther, und Lilia Filipova-Neumann. 2012. „Cluster Analysis of Smart Metering Data“. *Business & Information Systems Engineering* 4 (1): 31–39. https://doi.org/10.1007/s12599-011-0201-5.

Fotheringham, A. S., und D. W. S. Wong. 1991. „The modifiable areal unit problem in multivariate statistical analysis“. *Environment and Planning A* 23 (7): 1025 – 1044. https://doi.org/10.1068/a231025.

Freund, Yoav, und Robert E Schapire. 1997. „A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting“. *Journal of Computer and System Sciences* 55 (1): 119–39. https://doi.org/10.1006/jcss.1997.1504.

Friedman, Jerome H. 2001. „Greedy Function Approximation: A Gradient Boosting Machine“. *The Annals of Statistics* 29 (5): 1189–1232.

Fritsch, Stefan, und Frauke Guenther. 2016. *neuralnet: Training of Neural Networks*. https://CRAN.R-project.org/package=neuralnet.

Gangale, Flavia, Anna Mengolini, und Ijeoma Onyeji. 2013. „Consumer Engagement: An Insight from Smart Grid Projects in Europe“. *Energy Policy* 60 (September): 621–28. https://doi.org/10.1016/j.enpol.2013.05.031.

Gellings, Clark W., und Kelly E. Parmenter. 2015. „Demand-Side-Management“. In *Energy Efficiency and Renewable Energy Handbook, Second Edition*, herausgegeben von D. Yogi Goswami und Frank Kreith, 289–310. CRC Press.

Gorodkin, J. 2004. „Comparing two K-category assignments by a K-category correlation coefficient“. *Computational biology and chemistry* 28 (5): 367–74.

Gustafson, Eric J. 1998. „Quantifying Landscape Spatial Pattern: What Is the State of the Art?“ *Ecosystems* 1 (2): 143–56. https://doi.org/10.1007/s100219900011.

Guyon, Isabelle, André, und Elisseeff. 2003. „An introduction to variable and feature selection“. *Journal of Machine Learning Research* 3: 1157–1182.

Han, Jiawei, Micheline Kamber, und Jian Pei. 2012. *Data mining: Concepts and techniques*. 3. The Morgan Kaufmann series in data management systems. Amsterdam: Elsevier.

Han, Jiawei, Jian Pei, und Xifeng Yan. 2005. „Sequential Pattern Mining by Pattern-Growth: Principles and Extensions*“. In *Foundations and advances in data mining*, herausgegeben von Wesley Chu, 180:183–220. Studies in fuzziness and soft computing. Berlin u.a: Springer.

Hanczar, Blaise, Jianping Hua, Chao Sima, John Weinstein, Michael Bittner, und Edward R. Dougherty. 2010. „Small-Sample Precision of ROC-Related Estimates“. *Bioinformatics* 26 (6): 822–30. https://doi.org/10.1093/bioinformatics/btq037.

Hand, David J. 2009. „Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve“. *Machine Learning* 77 (1): 103–23. https://doi.org/10.1007/s10994-009-5119-5.

Hart, George Wiliam. 1992. „Nonintrusive appliance load monitoring“. *Proceedings of the IEEE* 80 (12): 1870–91. https://doi.org/10.1109/5.192069.

Hastie, Trevor, Robert Tibshirani, und Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer. http://link.springer.com/10.1007/978-0-387-84858-7.

Haury, Anne-Claire, Pierre Gestraud, und Jean-Philippe Vert. 2011. „The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures". *PLOS ONE* 6 (12): e28210. https://doi.org/10.1371/journal.pone.0028210.

Helmcke, D. T. 2008. „Regionalstatistik auf europäischer und nationaler Ebene". Statistisches Bundesamt.

Herbig, Paul A, und Ralph L Day. 1992. „Customer acceptance: the key to successful introductions of innovations". *Marketing Intelligence & Planning* 10 (1): 4–15.

Hopf, Konstantin. 2019. „Predictive Analytics for Energy Efficiency and Energy Retailing". Schriften aus der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg, Band 36, Bamberg: University of Bamberg Press. https://doi.org/10.20378/irbo-54833

Hopf, Konstantin, Sascha Riechel, Mariya Sodenkamp, und Thorsten Staake. 2017. „Predictive Customer Data Analytics – The Value of Public Statistical Data and the Geographic Model Transferability". In *ICIS 2017 Proceedings*. Seoul, South Korea: AIS electronic library.

Hopf, Konstantin, Mariya Sodenkamp, und Ilya Kozlovskiy. 2016. „Energy data analytics for improved residential service quality and energy efficiency". In *ECIS 2016 Research in Progress Proceedings*. Istanbul, Turkey: AIS electronic library.

Hopf, Konstantin, Mariya Sodenkamp, Ilya Kozlovskiy, und Thorsten Staake. 2014. „Feature extraction and filtering for household classification based on smart electricity meter data". In *Computer Science-Research and Development*, (31) 3:141–48. Zürich: Springer. https://doi.org/10.1007/s00450-014-0294-4.

Hopf, Konstantin, Mariya Sodenkamp, und Thorsten Staake. 2018. „Enhancing energy efficiency in the residential sector with smart meter data analytics". *Electronic Markets* 28 (4). https://doi.org/10.1007/s12525-018-0290-9.

Hothorn, Torsten, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, und Mark J. van der Laan. 2006. „Survival Ensembles". *Biostatistics* 7 (3): 355–73. https://doi.org/10.1093/biostatistics/kxj011.

Hsu, Chih-wei, Chih-chung Chang, und Chih-jen Lin. 2010. *A practical guide to support vector classification*.

Hua, Jianping, Waibhav D. Tembe, und Edward R. Dougherty. 2009. „Performance of Feature-Selection Methods in the Classification of High-Dimension Data". *Pattern Recognition* 42 (3): 409–24. https://doi.org/10.1016/j.patcog.2008.08.001.

Jain, Anil K, Jianchang Mao, und K Moidin Mohiuddin. 1996. „Artificial neural networks: A tutorial". *IEEE computer* 29 (3): 31–44.

Jang, Hyeong Yu, und Mi Jin Noh. 2011. „Customer acceptance of IPTV service quality". *International Journal of Information Management* 31 (6): 582–92.

Jurman, Giuseppe, Samantha Riccadonna, und Cesare Furlanello. 2012. „A Comparison of MCC and CEN Error Measures in Multi-Class Prediction". *PLoS ONE* 7 (8). https://doi.org/10.1371/journal.pone.0041882.

Kalousis, Alexandros, Julien Prados, und Melanie Hilario. 2007. „Stability of Feature Selection Algorithms: A Study on High-dimensional Spaces". *Knowledge and Information Systems* 12 (1): 95–116. https://doi.org/10.1007/s10115-006-0040-8.

Kent Eriksson, Katri Kerem, und Daniel Nilsson. 2005. „Customer acceptance of internet banking in Estonia". *International Journal of Bank Marketing* 23 (2): 200–216. https://doi.org/10.1108/02652320510584412.

Kim, H., M. Marwah, M. Arlitt, G. Lyon, und J. Han. 2011. „Unsupervised Disaggregation of Low Frequency Power Measurements". In *Proceedings of the 2011 SIAM International Conference on Data Mining*, 747–58. Proceedings. Society for Industrial and Applied Mathematics.

Kolter, J Zico, Siddharth Batra, und Andrew Y Ng. 2010. „Energy disaggregation via discriminative sparse coding". In *Advances in Neural Information Processing Systems*, 1153–61.

Kotsiantis, Sotiris B., I. Zaharakis, und P. Pintelas. 2007. „Supervised machine learning: A review of classification techniques". *Informatica*, Nr. 31: 249–68.

Krasnova, Hanna, und Natasha F Veltri. 2010. „Privacy calculus on social networking sites: Explorative evidence from Germany and USA". In , 1–10. IEEE.

Krishnamurti, Tamar, Daniel Schwartz, Alexander Davis, Baruch Fischhoff, Wändi Bruine de Bruin, Lester Lave, und Jack Wang. 2012. „Preparing for Smart Grid Technologies: A Behavioral Decision Research Approach to Understanding Consumer Expectations about Smart Meters“. *Energy Policy* 41 (Februar): 790–97. https://doi.org/10.1016/j.enpol.2011.11.047.

Kudo, Mineichi, und Jack Sklansky. 2000. „Comparison of Algorithms that Select Features for Pattern Classifiers“. *Pattern Recognition* 33 (1): 25–41.

Kursa, Miron B., und Witold R. Rudnicki. 2010. „Feature Selection with the Boruta Package“. *Journal of Statistical Software* 36 (11). https://doi.org/10.18637/jss.v036.i11.

Kwac, Jungsuk, Chin-Woo Tan, Nicole Sintov, June Flora, und Ram Rajagopal. 2013. „Utility customer segmentation based on smart meter data: Empirical study“. In *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, 720–725. Vancouver, BC, Canada. https://doi.org/10.1109/SmartGridComm.2013.6688044.

Lawrence, Steve, C Lee Giles, und Ah Chung Tsoi. 1997. „Lessons in neural network training: Overfitting may be harder than expected“. In , 540–45.

Lin, Chih-Jen. 2003. „A Practical Guide to Support Vector Classification“. Freiburg, Juli 15. http://www.csie.ntu.edu.tw/~cjlin/talks/freiburg.pdf.

Liu, Huan, und Hiroshi Motoda, Hrsg. 2008. *Computational methods of feature selection*. Chapman & Hall/CRC data mining and knowledge discovery series. Boca Raton: Chapman & Hall/CRC.

Lo, Adeline, Herman Chernoff, Tian Zheng, und Shaw-Hwa Lo. 2016. „Framework for Making Better Predictions by Directly Estimating Variables' Predictivity“. *Proceedings of the National Academy of Sciences*, November, 201616647. https://doi.org/10.1073/pnas.1616647113.

Lobo, Jorge M., Alberto Jiménez-Valverde, und Raimundo Real. 2008. „AUC: A Misleading Measure of the Performance of Predictive Distribution Models“. *Global Ecology and Biogeography* 17 (2): 145–51. https://doi.org/10.1111/j.1466-8238.2007.00358.x.

Mah, Daphne Ngar-yin, Johannes Marinus van der Vleuten, Peter Hills, und Julia Tao. 2012. „Consumer Perceptions of Smart Grid Development: Results of a Hong Kong Survey and Policy Implications“. *Energy Policy* 49 (Oktober): 204–16. https://doi.org/10.1016/j.enpol.2012.05.055.

McLoughlin, Fintan, Aidan Duffy, und Michael Conlon. 2012. „Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables“. *Energy and Buildings* 48: 240–48. https://doi.org/10.1016/j.enbuild.2012.01.037.

Merkert, Johannes, Marcus Mueller, und Marvin Hubl. 2015. „A Survey of the Application of Machine Learning in Decision Support Systems“. In *ECIS 2015 Proceedings*. Münster: AIS electronic library. https://balsa.man.poznan.pl/indico/contributionDisplay.py?contribId=13&sessionId=37&confId=44.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, und Friedrich Leisch. 2014. *e1071: Misc Functions of the Department of Statistics (e1071)*. TU Wien. http://CRAN.R-project.org/package=e1071.

Miltgen, Caroline Lancelot, Aleš Popovič, und Tiago Oliveira. 2013. „Determinants of end-user acceptance of biometrics: Integrating the “Big 3” of technology acceptance with privacy context“. *Decision Support Systems* 56: 103–14.

Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill series in computer science. New York: McGraw-Hill.

Motsch, William. 2012. „Dynamische Tarife zur Kundeninteraktion mit einem Smart Grid“. In *Smart Energy*, 229–58. Wiesbaden: Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-8348-1981-9_9.

Parson, Oliver, Siddhartha Ghosh, Mark J Weal, und Alex Rogers. 2012. „Non-Intrusive Load Monitoring Using Prior Models of General Appliance Types.“ In *AAAi*.

Prudenzi, A. 2002. „A neuron nets based procedure for identifying domestic appliances pattern-of-use from energy recordings at meter panel“. In *Power Engineering Society Winter Meeting, 2002. IEEE*, 2:941–46. IEEE.

Quinn, Elias Leake. 2009. „Privacy and the new energy infrastructure“. *SSRN Electronic Journal* 2009 (02). https://doi.org/10.2139/ssrn.1370731.

Ram, Sudha, Yun Wang, Faiz Currim, und Sabah Currim. 2015. „Using Big Data for Predicting Freshmen Retention“. In *ICIS 2015 Proceedings*. Fort Worth, USA: AIS electronic library. http://aisel.aisnet.org/icis2015/proceedings/DecisionAnalytics/13.

Reunanen, Juha. 2003. „Overfitting in Making Comparisons Between Variable Selection Methods“. *Journal of Machine Learning Research* 3 (März): 1371–1382.

Riedmiller, Martin. 1994. „Rprop-description and implementation details“. Technical Report. University of Karlsruhe. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.3428.

Riedmiller, Martin, und Heinrich Braun. 1993. „A direct adaptive method for faster backpropagation learning: The RPROP algorithm“. In *Neural Networks, 1993., IEEE International Conference On*, 586–591. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=298623.

Robnik-Sikonja, Marko, und Petr Savicky with contributions from John Adeyanju Alao. 2016. *CORElearn: Classification, Regression and Feature Evaluation*. https://CRAN.R-project.org/package=CORElearn.

Rogers, Everett M. 1976. „New product adoption and diffusion“. *Journal of consumer Research*, 290–301.

Romanski, Piotr, und Lars Kotthoff. 2014. *FSelector: Selecting attributes*. http://CRAN.R-project.org/package=FSelector.

Russell, Stuart J., und Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice-Hall.

Saeys, Yvan, Thomas Abeel, und Yves van de Peer. 2008. „Robust Feature Selection Using Ensemble Feature Selection Techniques“. In *Machine Learning and Knowledge Discovery in Databases*, 313–25. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-87481-2_21.

Saeys, Yvan, Iñaki Inza, und Pedro Larrañaga. 2007. „A Review of Feature Selection Techniques in Bioinformatics“. *Bioinformatics* 23 (19): 2507–17. https://doi.org/10.1093/bioinformatics/btm344.

Shin, Dong-Hee. 2010. „The Effects of Trust, Security and Privacy in Social Networking: A Security-Based Approach to Understand the Pattern of Adoption“. *Interacting with Computers* 22 (5): 428–38. https://doi.org/10.1016/j.intcom.2010.05.001.

Silva, Daswin, Xinghuo Yu, Damminda Alahakoon, und Grahame Holmes. 2011. „A Data Mining Framework for Electricity Consumption Analysis From Meter Data“. *IEEE Transactions on Industrial Informatics* 7 (3): 399–407. https://doi.org/10.1109/TII.2011.2158844.

Sodenkamp, Mariya, Konstantin Hopf, Ilya Kozlovskiy, und Thorsten Staake. 2016. „Smart-Meter-Datenanalyse für automatisierte Energieberatungen (‚Smart Grid Data Analytics')“. Final Report 291131. Bern, Switzerland: Bundesamt für Energie. http://www.bfe.admin.ch/dokumentation/energieforschung/index.html?lang=de&publication=11372.

Sodenkamp, Mariya, Ilya Kozlovskiy, Konstantin Hopf, und Thorsten Staake. 2017. „Smart Meter Data Analytics for Enhanced Energy Efficiency in the Residential Sector“. In *Wirtschaftsinformatik 2017 Proceedings*. St. Gallen, Switzerland: AIS electronic library.

Sodenkamp, Mariya, Ilya Kozlovskiy, und Thorsten Staake. 2016. „Supervised Classification with Interdependent Variables to Support Targeted Energy Efficiency Measures in the Residential Sector“. *Decision Analytics* 3 (1). https://doi.org/10.1186/s40165-015-0018-2.

Sokolova, Marina, und Guy Lapalme. 2009. „A systematic analysis of performance measures for classification tasks“. *Information Processing & Management* 45 (4): 427–437.

Statistical Office of the European Communities. 2014. *Living Conditions in Europe: 2014 Edition*. Luxembourg: Publications Office of the European Union. http://dx.publications.europa.eu/10.2785/59473.

Strbac, Goran. 2008. „Demand side management: Benefits and challenges“. *Energy Policy*, Foresight Sustainable Energy Management and the Built Environment Project, 36 (12): 4419–26. https://doi.org/10.1016/j.enpol.2008.09.030.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, und Achim Zeileis. 2008. „Conditional Variable Importance for Random Forests“. *BMC Bioinformatics* 9 (1): 307. https://doi.org/10.1186/1471-2105-9-307.

Suh, Bomil, und Ingoo Han. 2003a. „Effect of trust on customer acceptance of Internet banking“. *Electronic Commerce research and applications* 1 (3): 247–63.

———. 2003b. „The Impact of Customer Trust and Perception of Security Control on the Acceptance of Electronic Commerce“. *International Journal of Electronic Commerce* 7 (3): 135–61.

Tillé, Yves, und Alina Matei. 2015. *sampling: Survey Sampling*. https://CRAN.R-project.org/package=sampling.

Vapnik, Vladimir Naumovich, und Vlamimir Vapnik. 1998. *Statistical learning theory*. Bd. 1. Wiley New York.

Venkatesh, Viswanath, James Y.L. Thong, und Xin Xu. 2012. „Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology“. *MIS Quarterly* 36 (1): 157–78.

Wong, Yung Fei, Y Ahmet ¸Sekerciŏglu, Tom Drummond, und Voon Siong Wong. 2013. „Recent approaches to non-intrusive load monitoring techniques in residential settings“. In *Computational Intelligence Applications In Smart Grid (CIASG), 2013 IEEE Symposium on*, 73–79. IEEE.

Wytock, Matt, und J Zico Kolter. 2014. „Contextually Supervised Source Separation with Application to Energy Disaggregation.“ In *AAAI*, 486–92.

Yang, Zhilin, und Robin T Peterson. 2004. „Customer perceived value, satisfaction, and loyalty: The role of switching costs“. *Psychology & Marketing* 21 (10): 799–822.

Zaki, Mohammed J., und Wagner Meira Jr. 2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Zeifman, Michael, und Kurt Roth. 2011. „Nonintrusive appliance load monitoring: Review and outlook“. *IEEE Transactions on Consumer Electronics*, 76–84. https://doi.org/10.1109/TCE.2011.5735484.

Zhou, Tao. 2011. „The impact of privacy concern on user adoption of location-based services“. *Industrial Management & Data Systems* 111 (2): 212–26. https://doi.org/10.1108/02635571111115146.