Band 30

ifes Schriftenreihe

Bianca Krol (Hrsg.)

Machine Learning Modelle zur Vorhersage von Zahlungsausfällen im Energiemarkt

~

Rouven Stecker, Frank Lehrbass

FOM Hochschule
ifes
Institut für Empirie & Statistik
der FOM Hochschule
für Oekonomie & Management

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

Akademie
Verlags- und Druck-
Gesellschaft mbH

MA Akademie Verlags- und Druck-Gesellschaft mbH
Leimkugelstraße 6, 45141 Essen
info@mav-verlag.de

Rouven Stecker, Frank Lehrbass

Machine Learning Modelle zur Vorhersage von Zahlungsausfällen
im Energiemarkt

ifes Institut für Empirie & Statistik
der FOM Hochschule für Oekonomie & Management

ifes Schriftenreihe
Band 30, 2023

ISBN (Print) 978-3-89275-435-0 ISSN (Print) 2191-3366
ISBN (eBook) 978-3-89275-436-7 ISSN (eBook) 2569-5355

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AdaBoost	Adaptive Boosting
AIC	Akaike Information Criterion
AGB	Allgemeine Geschäftsbedingungen
AUPRC	area under the precision-recall curve
AUROC	area under the receiver operating characteristic
BEV Energie	BEV Bayerische Energieversorgungsgesellschaft mbH
BIC	Bank Identifier Code
CAC	Customer Acquisition Cost
CRISP-DM	Cross-Industry Standard Process for Data Mining
EEX	European Energy Exchange
EnBW	EnBW Energie Baden-Württemberg AG
EnWG	Energiewirtschaftsgesetz
EU	Europäische Union
GBM	Gradient Boosting Machine
GPU	Graphics Processing Unit
iMSys	intelligentes Messsystem
kWh	Kilowattstunde
LRT	Likelihood Ratio Test
MCC	Matthews Correlation Coefficient
MDI	Mean Decrease in Impurity
NAV	Niederspannungsanschlussverordnung
NDAV	Niederdruckanschlussverordnung
OOB	out-of-bag
PPV	positive predictive value
PR	Precision-Recall
ROC	receiver operating characteristic
SCHUFA	Schufa Holding AG
XGBoost	eXtreme Gradient Boosting

Formelverzeichnis

Listingverzeichnis

1 Einleitung

Mit der Liberalisierung des Energiemarkts wurde der Wettbewerb zwischen Energielieferanten der Sparten Strom und Gas ermöglicht. Seitdem werben die Lieferanten mit unterschiedlichsten Methoden um Neukunden. Energiekonzerne stärken ihre eigene Marke oder das Produkt mit einer Farbe, schließen Sponsorenverträge mit Sportvereinen ab oder werben mit einer regionalen Identifikation. Hinzu kommen die Vergleichsportale, die mittlerweile die Hauptinformationsquelle für Haushaltskunden darstellen, oder auch Discount-Anbieter, die dauerhaft niedrige Preise und Preisgarantien anbieten.[1] Der Preisdruck im Markt erhöht sich zudem, da Energielieferanten hohe Bonuszahlungen versprechen und kurzfristige Beschaffungen durchführen, sodass im ersten Belieferungsjahr negative Deckungsbeiträge hingenommen werden oder durch Schwankungen im Strom- und Gasmarkt die kurzfristige Beschaffung zu Insolvenzen führt.[2,3,4] Durch das Angebot niedriger Energiepreise kommt der Kostendruck auf Energielieferanten dazu. Provisionszahlungen an Vergleichsportale, Kundenservice und manuelle Prozesse müssen zur Abwicklung eines Energiekunden bezahlt und einkalkuliert werden. Einen wesentlichen Kostenträger stellen Zahlungsausfälle dar. In der Sparte Strom werden Kündigungen bei einem Zahlungsrückstand von durchschnittlich 176 € und bei Gas von durchschnittlich 170 € ausgesprochen.[5] Im Zusammenspiel mit den hingenommenen negativen Deckungsbeiträgen wird ein noch größerer Verlust generiert oder gar die Deckungsbeiträge einiger ordnungsgemäßer Vertragsverhältnisse vernichtet.

Zur Verhinderung von Zahlungsausfällen können Auskunfteien, die die Bonität eines Kunden prüfen, eingesetzt werden. Mit einer Auskunftei kann gesteuert werden, ob und mit welcher Bonität ein Vertragsschluss mit einem Kunden zustande kommt. Die Anfrage einer Bonitätsauskunft erhöht wiederum die Kosten des Energielieferanten. Pro Bonitätsprüfung ist eine Zahlung zu entrichten. Zudem ist nicht bekannt, ob der Vertrag durch einen Widerruf oder marktspezifische Ursachen überhaupt zustande kommt und Erlöse generiert werden. Die Kosten aller Bonitätsprüfungen sind somit in allen Verträgen, die sich aktiv in Belieferung befinden, einzupreisen.

---

1 Vgl. Schiffer, H.-W., 2019, S. 248, 255, 257.
2 Vgl. Stiftung Warentest, 2021.
3 Vgl. Güßgen, F., 29.10.2021.
4 Vgl. Lohse, L., Künzel, M., 2011, S. 386.
5 Vgl. Bundesnetzagentur und Bundeskartellamt, 2021, S. 267, 433.

Um den Kostendruck eines Energielieferanten zu senken, gilt es die IT sowie die Digitalisierung und Automatisierung von Prozessen zu stärken. Insbesondere der Einsatz von Big Data bietet viele Chancen im und vor dem Lebenszyklus eines Vertrags.[6] In einer mexikanischen Einzelhandelskette konnte durch Daten zum Einkaufsverhalten das Kreditausfallrisiko annähernd bestimmt werden.[7] Bei der Vergabe von Privatkrediten hat das Internetverhalten wie Interessen oder installierte Apps eine signifikante Vorhersagekraft zur Berechnung eines Kreditausfallrisikos.[8] Ebenso besitzen Energielieferanten durch bestehende Verträge Daten zum und über den Kunden. Diese lassen sich in zwei Kategorien unterteilen: Daten, die vor Vertragsschluss bekannt sind, und Daten, die während der Energiebelieferung erfasst werden. Zu Letzterem zählt insbesondere das Zahlungsverhalten eines Kunden. Durch eigene Mahn- oder Inkassoprozesse ist bekannt, welche Kunden in einem Zahlungsausfall münden. Zur direkten Gewinnung von Erkenntnissen werden dazu die Daten eines Energielieferanten analysiert. Die Daten stammen von einem Energielieferanten, der wettbewerblich im deutschen Energiemarkt im Haushaltskundensegment tätig war. Es werden Machine-Learning-Modelle mit den Daten und gewonnenen Erkenntnissen zur Vorhersage von Zahlungsausfällen entwickelt. Die Modelle werden zudem für einen Einsatz im Energiemarkt auf ihre Wirtschaftlichkeit untersucht.

1.1 Zielsetzung

Um Zahlungsausfälle zu verhindern, scheint es aufgrund der spezialisierten Algorithmen und der riesigen Datenmenge unausweichlich zu sein, eine Auskunftei in die Prozesse einzubinden. Kontrahierend dazu sind Prozesse, mit denen Kosten so gering wie möglich gehalten werden sollen. In Bezug auf einen Energielieferanten müsste demnach für jeden Vertragsabschluss eine entgeltliche Bonitätsauskunft angefragt werden, ohne dass ein erfolgreicher Vertragsabschluss garantiert ist, beispielsweise im Falle eines Widerrufs oder bei marktspezifischen Hindernissen. Einem Energielieferanten, der am Markt aktiv Energie vertreibt, liegen Daten von Kunden und deren Verträgen sowie Daten zu Zahlungsaktivitäten oder Mahn- und Inkassoprozessen vor. Daher stellen sich die zentralen Fragen:

---

6 Vgl. Kammel, Eike and Hollmann, Maik, 2016, S. 41 ff.
7 Vgl. Vissing-Jorgensen, A., 2011.
8 Vgl. Wu, W. et al., 2020.

- Welche Erkenntnisse lassen sich aus den Energielieferanten-eigenen Daten ziehen? Welche Datenfelder stehen im Zusammenhang mit Zahlungsausfällen?
- Reichen die Energielieferanten-eigenen Daten über einen Kunden aus, die beim Vertragsabschluss zur Verfügung stehen, um eine präzise Vorhersage über einen Zahlungsausfall und somit eine Aussage über die Bonität eines Privatkunden zu treffen? Welche Datenfelder sind für eine Vorhersage besonders wertvoll?
- Bieten Verhaltensdaten eines Kunden, die innerhalb des ersten Jahres nach Belieferungsbeginn anfallen, zur Erkennung von Zahlungsausfällen einen Mehrwert und lohnt sich eine Überwachung der Verhaltensdaten?
- Wie können die entwickelten Modelle eingesetzt werden und welche wirtschaftlichen Vorteile werden erzielt?

Zur Beantwortung der Fragen liegt ein Datensatz eines Energielieferanten vor, der am deutschen Energiemarkt aktiv war. Mit diesen Daten wird gemäß Cross-Industry Standard Process for Data Mining (CRISP-DM) ein Zyklus durchlaufen. In jedem Schritt werden neue Erkenntnisse zur Beantwortung der Fragen gehoben. Der CRISP-DM wird nachfolgend erläutert.

1.2 Aufbau und Vorgehensweise

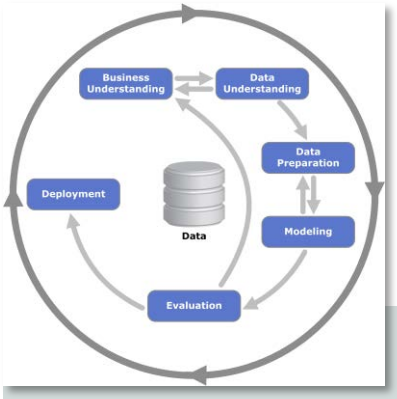


Abb. 1: Cross-Industry Standard Process for Data Mining (CRISP-DM)[9]

[9] Jensen, K., 2012.

Der Aufbau orientiert sich am Zyklus des CRISP-DM, der von Chapman et al. entwickelt wurde.[10] Die unterschiedlichen Phasen sind adaptiv und bauen auf das Ergebnis der vorherigen Phase auf. Die Phasen sind in Abbildung 1 dargestellt. Darin ist ersichtlich, dass auch Rückschritte möglich sind. Das kann beispielsweise der Fall sein, wenn durch Erkenntnisse einer Modellierung weitere Datenvorbereitungen auffallen. Jeder Schritt ist deshalb gleich wichtig und wirkt sich auf die Qualität des Ergebnisses aus. Der äußere Ring des CRISP-DM deutet auf die iterative Entwicklung hin. Nachdem ein Modell evaluiert und bereitgestellt wurde, werden neue Erkenntnisse erlangt, wodurch der Zyklus von neuem startet.[11] Mit den zuvor beschriebenen Aspekten sowie für die erstmalige Entwicklung und mit der Möglichkeit, die Entwicklung fortzuführen, wird sich für diesen Zyklus entschieden.

In Kapitel 2 findet die Geschäftsverständnisphase (*Business understanding*) statt. Darin wird kurz die Historie der Energiemarktliberalisierung geschildert. Anschließend werden der starke Wettbewerb zwischen Energielieferanten sowie der Preis- und Kostendruck und die derzeitigen Herausforderungen und Chancen eines Energielieferanten herausgestellt. Darin werden insbesondere die Zahlungsausfälle und deren Auswirkungen gezeigt. Die Kosten für eine Bonitätsauskunft und auch die Auswirkungen eines Zahlungsausfalls werden aus Anfragen, Statistiken und an einem konstruierten *Worst-Case-Kunden* deutlich gemacht.

In Kapitel 3 wird an die Verständnisphase angeknüpft, indem die Grundlagen zum Machine Learning dargelegt werden. Aufbau und Funktionsweise der logistischen Regression, klassischer Entscheidungsbäume sowie der Ensemble Methoden – insbesondere Random Forest und eXtreme Gradient Boosting (XGBoost) – werden für einen späteren Einsatz vermittelt.

Kapitel 4 stellt die Datenverständnisphase (*Data understanding*) dar. Darin wird der Datensatz, der von einem Energielieferanten vorliegt, vorgestellt sowie deskriptiv und explorativ analysiert. Aus den Erkenntnissen des Datenverständnisses folgt in Kapitel 5 die Datenvorbereitungsphase (*Data preparation*). Darin werden für eine Modellentwicklung unterschiedliche Datenfelder und Werte behandelt. Dazu gehören beispielsweise die Entfernung von Ausreißern, die Umwandlung kategorialer Variablen oder auch die Aufteilung in Trainings- und Testdatensätze.

Mit den vorbereiteten Daten wird in Kapitel 6 die Modellierungsphase (*Modeling*) beschritten. Mithilfe der logistischen Regression, dem Random Forest- und dem

[10] Vgl. Chapman, P. et al., 2000, S. 13.
[11] Vgl. Larose, D. T., 2015, S. 6 ff.

eXtreme Gradient Boosting (XGBoost)-Algorithmus werden die Modelle entwickelt. Es findet zudem eine Auswertung und Filterung der Variablen statt, die für eine Zahlungsausfallvorhersage am wichtigsten sind.

Kapitel 7 behandelt die Evaluierungsphase (*Evaluation*). In der Evaluierung werden die verschiedenen Modelle miteinander verglichen und das Modell ausgewählt, welches die besten Ergebnisse erzielt. Mit dem besten Modell findet für einen Einsatz die wirtschaftliche Betrachtung statt. Es wird herausgestellt, welchen monetären Nutzen dieses Modell hat.

In Kapitel 8 wird das Fazit mit den Erkenntnissen und Ergebnissen zum Durchlauf des CRISP-DM gezogen. Kapitel 9 schließt mit der Betrachtung von Bereitstellungsmöglichkeiten (*Evaluation*) ab.

1.3 Setup der Testumgebung

Für die Datenanalyse, Datenvorbereitung, Modellentwicklung und Evaluierung wurde `Python` in Version `3.9.7` verwendet.[12] Die genutzten Python-Bibliotheken sowie deren Versionen und Verwendungszwecke sind in Tabelle 1 aufgeführt. Zusätzlich wurden für die Visualisierung der Deutschlandkarten aufbereitete Postleitzahl-Ort-Polygon Daten von *suche-postleitzahl.org*[13], basierend auf Daten von *OpenStreetMap*[14] genutzt.

[12] Vgl. Python Software Foundation, 2022.
[13] Vgl. Schwochow, M., 2022.
[14] Vgl. OpenStreetMap contributors, 2017.

Bibliothek	Version	Verwendungszweck
pandas[15,16]	1.3.4	Datenanalyse- und manipulation
GeoPandas[17]	0.9.0	Datenanalyse- und manipulation raumbezogener Daten
Shapely[18]	1.7.1	Zusammenfassung von Polygonen
NumPy[19]	1.21.2	Matrizen- und Array-Handhabung/-Operationen
matplotlib[20]	3.5.0	Visualisierungen
seaborn[21]	0.11.2	Konfusionsmatrixvisualisierung
scikit-learn[22]	1.0.1	Datenvorbereitung und Entwicklung Random Forest-Modelle
xgboost[23]	1.5.1	Entwicklung XGBoost-Modelle

Tab. 1: Übersicht der verwendeten Python-Bibliotheken

Ausschließlich für die Entwicklung des logistischen Regressionsmodells wurde R in Version 4.1.2 verwendet.[24] Es wurden die R-Bibliotheken pscl[25] in Version 1.5.5 und lmtest[26] in Version 0.9-31 verwendet.

Zur Sicherstellung der Reproduzierbarkeit der Ergebnisse wurde der Startpunkt der Zufallszahlengenerierung (seed (R) und random_state (Python)) auf einen

15 The pandas development team, 2020.
16 Vgl. McKinney, W., 2010.
17 Vgl. Jordahl, K. et al., 2020.
18 Vgl. Gillies, S. et al., 2007.
19 Vgl. Harris, C. R. et al., 2020.
20 Vgl. Hunter, J. D., 2007.
21 Vgl. Waskom, M. L., 2021.
22 Vgl. Pedregosa, F. et al., 2011.
23 Vgl. Chen, T., Guestrin, C., 2016.
24 The R Foundation, 2022.
25 Vgl. Zeileis, A., Kleiber, C., Jackman, S., 2008.
26 Vgl. Zeileis, A., Hothorn, T., 2002.

festen Wert gesetzt. Zu beachten ist, dass bei einer (Neu-)Berechnung der XGBoost-Modelle trotzdem leicht abweichende Ergebnisse erzielt werden.[27]

Die Entwicklungen und Berechnungen wurden auf einem Computer mit Windows 11 (Build 22538.101) als Betriebssystem, einem Intel Core i7-9700K und 32 Gigabyte Arbeitsspeicher durchgeführt.

[27] Vgl. xgboost developers, 2022a, Slightly different result between runs.

2 Relevanz für Energielieferanten

Die Europäische Union (EU) setzte im Jahr 1996 den Grundstein für die Liberalisierung des Energiemarkts. Am 19. Dezember 1996 wurde die Richtlinie 96/92/EG des Europäischen Parlaments und des Rates veröffentlicht, die gemeinsame Vorschriften für den Elektrizitätsbinnenmarkt enthält. Die Maßnahmen und Anforderungen sollen das einwandfreie Funktionieren des Binnenmarkts gewährleisten. Waren, Dienstleistungen oder Kapital sind nur einige Elemente, deren freier Verkehr durch den Raum des Binnenmarkts sichergestellt werden. Zur Verwirklichung wird ein wettbewerbsorientierter Elektrizitätsmarkt gefordert, um die Vollendung des Energiebinnenmarkts zu erreichen.[28] Neben der Elektrizität wurde mit der Richtlinie 98/30/EG weniger als zwei Jahre später auch der wettbewerbsorientierte Erdgasmarkt gefordert.[29] Nach diesen zwei Richtlinien folgten im Laufe der Zeit noch weitere Richtlinien, um Sorge für gleiche Wettbewerbsbedingungen in den Energiemärkten zu tragen. Es wurden gleiche Bedingungen für den Zugang zu den Energiemärkten geschaffen und ebenfalls ein *legal unbundling* (Entflechtung) der Energieversorgungsunternehmen angestoßen. Die am Markt agierenden Unternehmen mussten durch das legal unbundling ihre einzelnen Marktrollen in Gesellschaften auslagern, die rechtlich selbstständig am Markt agieren.[30] Das bisherige Monopol im Energiemarkt wurde durch diese Richtlinien aufgebrochen. Vor der Liberalisierung übernahmen die Energieversorgungsunternehmen die Erzeugung, die Übertragung und die Verteilung, folglich die Netze der Endkunden sowie den Vertrieb an die Endkunden. Somit übernahmen sie die gesamte Handels- und Wertschöpfungskette der Commodities Strom und Gas. Mit der Liberalisierung stehen nun Energieerzeugungsunternehmen im Wettbewerb zueinander und verkaufen ihre Energiemengen an Börsen wie der European Energy Exchange (EEX). Die Übertragungs- und Verteilnetzbetreiber wurden reguliert, jedoch als natürliches Monopol im Energiemarkt belassen, da aus volkswirtschaftlicher Sicht der Bau und parallele Betrieb von mehreren Strom- und Gasnetzen nicht sinnvoll wäre. Zu diesen Netzen konnten Energieerzeugungsunternehmen durch das Gesetz zur Neuregelung des Energiewirtschaftsrechts einen regulierten Zugang erhalten.[31] Neben dem Verkauf der eigenen Energie durch Erzeugungsunternehmen an Endkunden konnten nun Vertriebsunternehmen

[28] Vgl. Richtlinie 96/92/EG Europäisches Parlament und Europäischer Rat, 1996, Abs. 1 und 2.
[29] Vgl. Richtlinie 98/30/EG Europäisches Parlament und Europäischer Rat, 1998, Abs. 3.
[30] Vgl. Pfaffenberger, W., Hille, M., 2004, S. 3–37 f.
[31] Vgl. Bundesgesetzblatt, 1998, §6, §7.

ohne eine eigene Energieerzeugung Energie am Markt und an den Börsen einkaufen, über Netze transportieren und einem Haushalts- oder Geschäftskunden in ganz Deutschland verkaufen. Abbildung 2 stellt den Marktwandel durch die Liberalisierung dar.

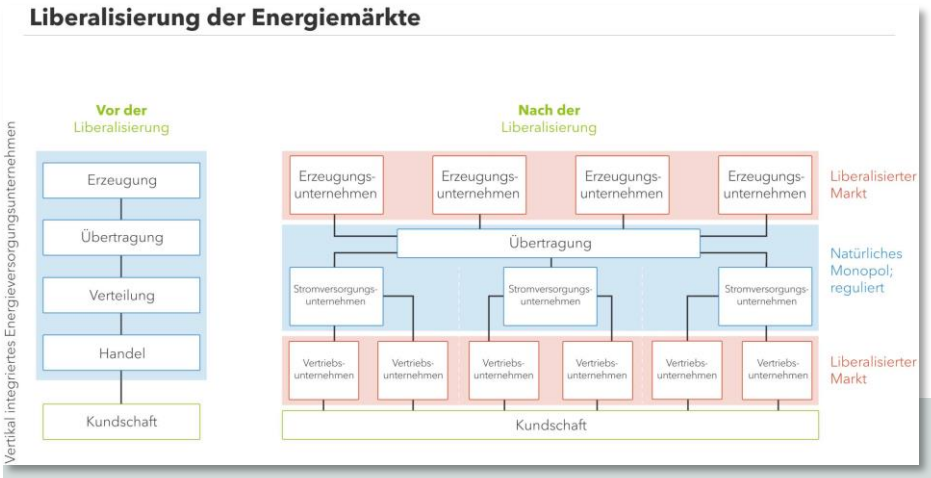


Abb. 2: Unternehmen und Rollen im Energiemarkt vor und nach der Liberalisierung[32]

2.1 Marktübergreifender Wettbewerb auf vielen Ebenen

Mit dem Fokus auf Haushaltskunden stehen die Unternehmen seit der Liberalisierung im stetigen Wettbewerb zueinander. Bereits Ende 1999 und nur wenige Jahre nach den ersten Richtlinien und Gesetzen zur Liberalisierung hatten die großen Energieversorger in Deutschland eigene, neue Strommarken und Tochterunternehmen ausgegründet. Die EnBW Energie Baden-Württemberg AG (EnBW) gründete die Tochter Yello mit der prägnanten gelben Farbe, die RWE AG warb mit der Strommarke Avanza und der Farbe Blau und die E.ON verband ihre Strommarke mit der Farbe Rot. Es folgten zahlreiche Werbeauftritte und Sportvereine trugen den Namen des Energieversorgers auf Trikots. Werbeausgaben stiegen zwischen 1998 und 2001 um durchschnittlich 39 % pro Jahr. Im Jahr 2007 stiegen die Werbeaufwendungen um rund 227 Mio. € und haben sich somit fast verdoppelt.[33] Dahinter stand die Doppelstrategie, einerseits Bestandskunden zu halten und andererseits Neukunden in anderen Netzen und

[32] Next Kraftwerke GmbH, 2021.
[33] Vgl. Lohse, L., Künzel, M., 2011, S. 386.

Teilen Deutschlands zu gewinnen. Bei E.ONs Kampagne *Mix it* belaufen sich die geschätzten Werbeausgaben auf 22,5 Mio. € bei gerade einmal 1.100 gewonnenen Neukunden. Bei einem durchschnittlichen Ertrag von 600 Euro und Akquisitionskosten von 20.500 € je Kunde stellen diese Werte kein wirtschaftlich positives Ergebnis dar. Ob eine etwaige Steigerung der Markenbekanntheit bei 1.100 gewonnenen Kunden stattgefunden hat, kann nicht beantwortet werden. Die Hoffnungen waren im Nachhinein größer als der tatsächliche Erfolg dieser Maßnahmen.[34,35] Gewinnung von Neukunden wie auch Halten von Bestandskunden gestalten sich im deutschen Energiemarkt besonders schwierig. Hier wirkt insbesondere die Beziehung zwischen Lieferanten und Kunden. Wahrnehmbare Unterscheidungsmerkmale in Verwendung und Qualität bei der Lieferung von Strom oder Gas liegen nicht vor. Die Lieferung und Abrechnung erfolgt nahezu identisch.[36] Ebenso kommen nach Lohse und Künzel weitere Merkmale erschwerend zur Kundenbeziehung hinzu:[37]

Die Omnipräsenz

Gas und besonders Strom sind für Konsumenten gefühlt immer und überall verfügbar.

Die Immaterialität und die mittelbare Nutzenstiftung

Energie spricht bei Vorhandensein und Verbrauch keine Sinne an und kann nur schwer vom Kunden wahrgenommen werden. Die Emotionen und der Nutzen sind nur sekundär vorhanden, da Energie selbst keinen Kundennutzen darstellt. Im Vordergrund stehen die Verwendung eines Geräts wie Herd, Smartphone oder Heizung mit direktem Kundennutzen.

Produkthomogenität und Austauschbarkeit

Energie ist in ihrer Darreichungsform reguliert. Strom wird mit einer normierten Spannung und Frequenz in jeden Haushalt geliefert. Gas wird je nach Darreichungsform in Energie bemessen, um eine Gleichwertigkeit verschiedener Temperaturgebiete und Gase herzustellen.[38] Durch das Energiewirtschaftsge-

---

34 Vgl. Schiffer, H.-W., 2019, S. 248.
35 Vgl. Meffert, H., Schröder, J., Perrey, J., 2002, S. 28 f.
36 Vgl. Wiedmann, K.-P., Ludewig, D., 2011, S. 100.
37 Vgl. Lohse, L., Künzel, M., 2011, S. 384 f.
38 In Deutschland zählt ein Gaszähler die gelieferte Gasmenge in Kubikmeter (m^3). In Deutschland sind noch zwei Gasnetze vorhanden, in denen L- oder H-Gas (low/high calorific gas) geliefert wer-

setz (EnWG) ist ebenso die Ersatzversorgung und – bis auf wenige Ausnahmen – die Grundversorgung in Deutschland gesichert.[39]

Low-Involvement

Die kontinuierliche Verfügbarkeit der Energie lässt die Produkterfahrung allgegenwärtig erscheinen und die Wahrnehmung verblassen. Erst bei einem Ausfall wird das Interesse der Kunden geweckt.

Neben der Schwierigkeit, eine gute Kundenbeziehung mit Commodityprodukten herzustellen, folgt der Preisdruck durch neue Stadtwerke und digitale Energielieferanten. Insbesondere Discountanbieter versuchen mit einer aggressiven Preisstrategie Neukunden zu gewinnen. In diesem Zuge bieten Energielieferanten Produkte mit attraktiven Boni im ersten Belieferungsjahr an. Somit ist die Energielieferung für den Verbraucher im ersten Jahr besonders günstig. Hingegen nimmt der Energielieferant zum Teil Verluste für die Neukundenakquisition hin und vermarktet bewusst mit negativen Deckungsbeiträgen. Dabei setzt der Energielieferant auf die Trägheit des Kunden und darauf, dass die Belieferung mehr als nur ein Jahr anhält, sodass der Lieferant anschließend in die Gewinnzone gelangt. In den folgenden Belieferungsjahren ist der Preisvorteil für den Kunden weitgehend nicht mehr vorhanden.[40,41,42] Zum Preis- und Kostendruck durch massiv gesenkte Produktpreise im ersten Belieferungsjahr kommen die tatsächlichen Roherträge je kWh für die Commodityprodukte Strom und Gas hinzu. In Tabelle 2 sind die einzelnen Preisbestandteile je kWh in Bezug auf Haushaltskunden aufgeführt.

den. Diese Gase haben einen unterschiedlichen Brennwert. Zusätzlich ist die Temperatur ausschlaggebend, da dadurch der Druck des Gases beeinflusst wird. Die Abrechnung von Gas in Kilowattstunde (kWh) erfolgt schlussfolglich aus der verbrauchten Gasmenge, dem Brennwert des Gases und einer temperaturabhängigen Zustandszahl.

39 Vgl. EnWG, 2021, §36 - §38.

40 Vgl. Stiftung Warentest, 2021.

41 Vgl. Lohse, L., Künzel, M., 2011, S. 386.

42 Vgl. Schiffer, H.-W., 2019, S. 252.

Preisbestandteil	Strom (Abnahme zwischen 2.500 und 5.000 kWh)	Gas (Abnahme zwischen 5.556 und 55.556 kWh)
Energiebeschaffung, Vertrieb und Marge	7,97	3,12
Netzentgelt	7,14	1,47
Entgelt für Messstellenbetrieb	0,36	0,07
Entgelt für Messung	-	0,02
Konzessionsabgabe	1,64	0,08
Umlagen	7,78	-
Strom-/Gassteuer	2,05	0,55
Umsatzsteuer	5,12	1,01
Gesamt	**32,05**	**6,31**

Tab. 2: Durchschnittlicher mengengewichteter Preis für Haushaltskunden (in Ct je kWh)[43]

Der Preis einer kWh Strom liegt bereits bei durchschnittlich 32,05 Cent. Die Strompreise in Deutschland zählen somit zu den höchsten Strompreisen in ganz Europa.[44,45] Von den 32,05 Cent je kWh entfallen rund 75,1 % der Kosten auf nicht beeinflussbare Preisbestandteile wie Netzentgelte, Steuern und Umlagen. Die restlichen 24,9 % des Strompreises kommen dem Lieferanten für Energiebeschaffung, Vertrieb und Marge zugute. Zwar scheinen die 24,9 % ein prozentual großer Anteil zu sein, jedoch sind das absolut und monetär betrachtet nur 7,97 Cent je kWh. Bei Haushaltskunden mit einer Stromabnahme zwischen 2.500 und 5.000 kWh ergeben das Einnahmen von rund 199,25 € bis 398,50 € pro Jahr.[46] Zum Vergleich: Im ersten Belieferungsjahr werden Boni zwischen 5 € und 300 € gewährt, was die effektiven Einnahmen für Beschaffung, Vertrieb und Marge schmälert und die Spanne der Einnahmen im ersten Belieferungsjahr auf -100,75 € bis zu 393,50 € weitet.[47] Von diesen Einnahmen müssen Akquisition, Prozesse, Service, Abrechnung und Sonderfälle bezahlt sowie die tatsächliche Strommenge beschafft werden. In Bezug auf die Sparte Gas ist der prozentuale Anteil des Lieferanten am Preis mit 49,4 % zwar deutlich höher, jedoch ist auch hier der absolute Wert mit 3,12 Cent je kWh gering. Bei einer Gasabnahme zwischen 6.000 und 10.000 kWh sind Einnahmen zwischen 187,20 €

43 Bundesnetzagentur und Bundeskartellamt, 2021, S. 276, 441.
44 Vgl. ebd., S. 305.
45 Vgl. Mihm, A., 20.08.2017.
46 Vgl. Bundesnetzagentur und Bundeskartellamt, 2021, S. 275 f.
47 Vgl. ebd., S. 288.

und 312,00 € pro Jahr für Beschaffung, Vertrieb und Marge zu erwarten und mit den Einnahmen der Sparte Strom zu vergleichen. Ebenfalls sind die Bonusauszahlungen miteinander zu vergleichen.[48]

Während die Lieferanten mit ihrer Preisstrategie im ersten Belieferungsjahr nur wenig Gewinn erzielen oder teils sogar Verluste machen, sind weitere Akteure im Markt vorhanden, die den Wettbewerb unter Lieferanten anheizen. Vergleichsportale bieten Endkunden über die Eingabe ihrer Postleitzahl, ihres Wohnorts und ihrem Energieverbrauch eine Übersicht der verfügbaren Lieferanten, Preise, Bonuszahlungen und weiterer Vertragsbestandteile und -merkmale wie Laufzeit, Kündigungsfrist oder ob es ein reiner Online-Tarif ist. In ca. 72,7 % aller Netzgebiete sind mehr als 100 Stromlieferanten und in ca. 49,5 % mehr als 100 Gaslieferanten tätig, was auf eine große Lieferantenauswahl für den Kunden hindeutet.[49] Dabei ist zu beachten, dass hauptsächlich teilnehmende Lieferanten bzw. Lieferanten, die eine Provision zahlen, bei Vergleichsportalen gelistet sind.[50] Die Vergleichsportale bieten dem Endkunden somit eine hohe, aber nicht vollständige Transparenz im Energiemarkt und gelten mittlerweile als Hauptinformationsquelle für Haushaltskunden. Insbesondere der Wechsel ist durch die Eingabe der Stammdaten, Angaben zur Lieferstelle und Zahlungsdaten besonders komfortabel. Der Kunde muss im Idealfall nichts weiter tun, da die Kündigung beim Altlieferanten und die Anmeldung beim zuständigen Netzbetreiber vom zukünftigen Lieferanten durch die marktregulierten Prozesse übernommen werden.[51] Vergleichsportale verdienen mit Provisionen pro abgeschlossenem Vertrag und sind dadurch an hohen Wechselquoten interessiert.[52] Die durchschnittlichen Provisionen liegen bei rund 40 bis 60 € für einen Energievertrag.[53] Diese Provisionen sind von den Einnahmen des Lieferanten im ersten Belieferungsjahr zusätzlich abzuziehen. Im Mittel wurden in den Jahren 2016 und 2017 rund 4.648.010 Strom- und rund 1.499.965 Gasverträge durch Einzüge oder Lieferantenwechsel abgeschlossen. Zusammen ergeben das 6.147.975 Vertragsabschlüsse.[54] In einer Strukturbefragung hat sich herausgestellt, dass im Zeitraum November 2016 bis

48 Vgl. ebd., S. 441, 451.
49 Vgl. ebd., S. 254.
50 Vgl. Bundeskartellamt, 2017, S. 3.
51 Sollte der Kunde zum Beispiel falsche Daten eingegeben haben oder noch längerfristig an einen bestehenden Energieliefervertrag gebunden sein, erfolgen daraus Rückfragen an den Kunden oder/und die Vertragsablehnung.
52 Vgl. Schiffer, H.-W., 2019, S. 257.
53 Vgl. Verivox, 2021.
54 Vgl. Bundesnetzagentur und Bundeskartellamt, 2021, S. 261, 426.

Oktober 2017 rund 3,5 Mio. Energieverträge (Strom- und Gasverträge) von Haushaltskunden über Vergleichsportale abgeschlossen wurden.[55] Daraus folgt, dass im deutschen Energiemarkt mehr als 56 % aller Energieverträge über Vergleichsportale abgeschlossen werden.

2.2 Möglichkeiten zur Kostenreduktion für Energielieferanten

Um dem marktübergreifenden Druck aus Wettbewerb und Preisgestaltung standzuhalten, gibt es für Energielieferanten einige Optionen. Zum einen gilt die Kundenzufriedenheit für Energielieferanten als wichtiger Schlüsselfaktor. Durch die Zufriedenheit und eine positive Kundenerfahrung wird sich eine höhere Loyalität und weniger Kundenabwanderung erhofft, ergo eine nachhaltige Kundenbeziehung aufgebaut. Durch eine nachhaltige Beziehung findet die Energiebelieferung unter Umständen weitaus länger als ein Jahr statt oder es können beispielsweise weitere Produkte verkauft werden (Cross-Selling). Um die Zufriedenheit eines Kunden zu steigern, werden diverse Maßnahmen ergriffen. Diese umfassen eine stets verfügbare Kundenbetreuung, schnelle Reaktionszeiten und effiziente Lösungen, wie auch Self-Service-Funktionen, damit der Kunde seine Anliegen eigenständig lösen kann.[56,57] Neben dem Service für Kunden stellen die Prozesse, die Effizienz und die Kostenstruktur des Vertriebs einen großen Hebel dar. Die insbesondere durch Vergleichsportale hervorgerufenen sehr hohen Wechselquoten bei Strom und Gas als Commodityprodukt stellen eine Herausforderung dar. Aufgrund dessen muss die Vertriebsstrategie ein differenziertes Leistungsspektrum für unterschiedliche Kundensegmente anbieten. Ein besonderes Augenmerk liegt zudem auf der IT sowie der Digitalisierung und Automatisierung von Prozessen. Rund 30 % aller Energielieferanten bieten einen reinen Online-Tarif an. Mit einer Betrachtung auf die Lieferanten, die mehr als 80 % aller Haushaltskunden beliefern, bieten sogar 77 % aller Lieferanten einen Online-Tarif an.[58,59] Online-Tarife eliminieren durchweg viele manuelle und kostspielige Prozesse. Anstelle von Briefen werden E-Mails versandt, um Druck- und Portokosten einzusparen. Die Zählerablesung wird beim Kunden mittels Benachrichtigungs-E-Mail eingefordert und zudem wird die Eingabe in einem Web-Portal ermöglicht, um herkömmlich postalisch zugestellte Ablesekarten mit Rückversand, Scanning und ggf. manuellen Klärfällen bei unleserlicher

[55] Vgl. Bundeskartellamt, 2017, S. 24.
[56] Vgl. Kammel, Eike and Hollmann, Maik, 2016, S. 42.
[57] Vgl. Rücker, Markus, 2016, S. 33.
[58] Vgl. Schiffer, H.-W., 2019, S. 249.
[59] Vgl. Bundesnetzagentur und Bundeskartellamt, 2021, S. 267.

Schrift zu erübrigen. Alle Funktionalitäten, die der Kunde benötigt, angefangen von der Änderung seines Abschlags bis zum Widerruf oder der Kündigung, werden in einem Kundenportal bereitgestellt, um die Anzahl der Kontaktpunkte mit dem Kundenservice stetig zu verringern. Und falls der Kunde doch weitere Hilfe benötigt, werden Lösungen wie Chatbots und Frequently Asked Questions (FAQ)-Seiten vorgeschaltet, um erst im Nachhinein einen kostengünstigen Kontaktpunkt (z. B. via Chat) mit einem Kundenservicemitarbeiter anzubieten.

Ein weiteres Handlungsfeld ist die Nutzung der Daten. Daten entstehen bei der Akquisition oder bei der Belieferung eines Kunden fortwährend. Wenn der Kunde allein ein intelligentes Messsystem (iMSys) verbaut hat[60], sendet der Zähler Messwerte wie die derzeitige Leistung oder den Gesamtenergieverbrauch in einem Intervall von 15 Minuten. In der Summe sind das 35.000 Datenpunkte pro Jahr und pro Zähler. Energielieferanten haben heute die Herausforderungen und die darin enthaltenen Chancen von Big Data. Nicht nur in Bezug auf die viertelstündlichen Messwerte kann ein Lieferant durch Analyse und Vorhersagen die Kundenzufriedenheit und die Effizienz steigern. Die weiteren Chancen durch Big Data sind in Tabelle 3 aufgeführt.[61]

[60] Die Zähler sind bei einem Jahresstromverbrauch über 6.000 kWh verpflichtend. Bei einem Verbrauch unter 6.000 kWh kann der Einbau auf freiwilliger Basis über den Messstellenbetreiber erfolgen.

[61] Vgl. Kammel, Eike and Hollmann, Maik, 2016, S. 41 ff.

Prozessschritt	Chancen durch Big Data
Akquisition & Marketing	• Kundengruppenbezogene Angebote und Empfehlungen • Gezielter Einsatz von Boni (Geldbonus oder Warenbonus) • Personalisierte Kundenkommunikation • Identifizierung und Verhinderung des Kundenabsprungs während der Akquisition
Belieferung & Service	• Beschaffungsoptimierung durch Verbrauchsanalyse • Erkennung der Kundenanliegen sowie Anbieten von Hilfestellung und Lösungsmöglichkeiten zum Kunden-Self-Service • Cross-Selling und Aktionshinweise für Kundenservicemitarbeiter bei Kontakt durch den Kunden • Proaktives Fehlermanagement, um Kundenaufstörungen zu verhindern
Zahlung & Kündigung	• Risikominimierung von Zahlungsausfällen durch z. B. Bonitätsprüfungen • Analyse und Optimierung des Beschwerdemanagements Erkennung von Fehlern, die zu Rechnungskorrekturen führen • Reduktion der Kundenabwanderung

Tab. 3: Chancen durch Big Data für Energielieferanten[62]

2.3 Zahlungsausfälle als besondere Herausforderung

Ungeachtet dessen, wie der Kunde akquiriert wurde oder wie hart der Wettbewerb im Energiemarkt ist, besteht zwischen Energielieferanten und Kunden in der Regel ein einfaches Vertragsverhältnis. Der Lieferant sorgt für die Energiebeschaffung und die Abwicklung aller (Kundenservice-)Prozesse von der Auftragsannahme bis zur Vertragsbeendigung. Hingegen ist der Kunde dazu angehalten, seine Zählerstände für eine genaue Abrechnung zu übermitteln sowie Zahlungen der monatlichen Abschläge oder auch einer etwaigen Nachzahlung durch eine Abrechnung nachzukommen.

[62] In Anlehnung an Kammel, Eike and Hollmann, Maik, 2016, S. 43 ff.

Wie in jeder vertraglichen Beziehung kann es zu unvorhergesehenen oder ungewünschten Ereignissen kommen, die zu Beeinträchtigungen der vertraglich vereinbarten Pflichten führen. Dabei ist die Sicht nicht immer nur auf den Kunden zu richten. Im Dezember 2018 und Januar 2019 hat die BEV Bayerische Energieversorgungsgesellschaft mbH (BEV Energie) für große Aufmerksamkeit gesorgt, da trotz gültiger Preisgarantien teils erhebliche Preiserhöhungen um ca. 700 % kommuniziert wurden. Im Zuge dessen reichten die Verbraucherzentralen Beschwerden und Unterlassungsverpflichtungen ein. Ebenfalls leitete die Bundesnetzagentur Ende Januar 2019 ein Aufsichtsverfahren gegen die BEV Energie ein.[63] Anschließend meldeten im Februar 2019 sowohl die BEV Energie wie auch deren schweizerische Muttergesellschaft Genie Holding AG Insolvenz an. Zu dieser Zeit standen mehr als 100 Millionen Euro an Forderungen gegenüber der BEV Energie offen, die teils aus Vorauszahlungen wie Abschlägen zur Strom- oder Gaslieferung bestanden.[64,65,66] Es ist zu vermuten, dass BEV Energie mit ihrer radikalen Discounter-Preisstrategie den Versuch unternahm, Neukunden u. a. über Vergleichsportale zu gewinnen und eine langfristige Beziehung aufzubauen, sodass die Gewinnzone erreicht wird.[67] Ebenso mussten ab Oktober 2021 mehrere Strom- und Gaslieferanten Insolvenz anmelden, da sie durch den starken Preisanstieg von Strom und Gas entsprechend höhere kurzfristige Beschaffungskosten hatten. Bei Vertragsschluss zwischen Lieferanten und Kunden wird meist ein Festpreis für den Belieferungszeitraum ausgehandelt, worin solch ein Energiepreisanstieg nicht mit eingepreist war.[68] Die Vorfälle und Insolvenzen an Beispielen der BEV Energie und weiterer Energielieferanten zeigen, mit welchen hohen oder gar geschäftsgefährdenden Risiken die stark kompetitiven Preisstrategien verbunden sind.

Rundum verdeutlichen die Ereignisse und der bestehende Wettbewerb im Energiemarkt den Bedarf von reibungslos verlaufenden Vertragsverhältnissen. Idealerweise dauern die Verhältnisse über ein Jahr an, damit die teils akzeptierten negativen Deckungsbeiträge ausgeglichen werden können. Aber auch von Seiten des Kunden können bei einem Vertragsverhältnis zur Lieferung von Strom oder Gas Probleme auftreten. Im Fokus steht hierzu der Zahlungsausfall, der durch eine

63 Vgl. Bayrische Staatsregierung, 22.01.2019.
64 Vgl. Köhler, M., 02.02.2019.
65 Vgl. Deutsche Presseagentur, 21.02.2019.
66 Vgl. Hoyer, N., 21.12.2018.
67 Vgl. Jung, M., 21.01.2020.
68 Vgl. Güßgen, F., 29.10.2021.

Vielzahl von Gründen[69] mit einer gewissen Ausfallwahrscheinlichkeit entstehen kann. Der Zahlungsausfall inkludiert nicht nur den Vollausfall (der Kunde zahlt gar nicht), sondern auch jegliche Schwierigkeiten, die bei Zahlungen auftreten können. Aus den Schwierigkeiten resultierende Maßnahmen sind ebenfalls mit Kosten verbunden. Dazu gehören u. a. die Erhöhung der Abschlagszahlung, um einem Zahlungsausfall vorzubeugen, sowie teils manuelle Prozesse im Mahn- oder Inkassowesen. Besonders im Bankenwesen bei der Vergabe von Krediten sind die Risiken und deren Präventionen durch höhere Summen stärker ausgeprägt. Zwischen einem Zahlungsausfall bei einer Energielieferung oder einem Zahlungsausfall bei einem Kredit können – wenn auch in einer anderen Tiefe und Komplexität – durchaus Parallelen gezogen werden.[70] Bei der Unterteilung der Forderungen gegenüber dem Kunden können diese jeweils einer von drei Kategorien nach Szczesny und Kaiser zugeordnet werden:[71]

Kein Ausfall

Der Kunde kommt den Forderungen aus dem Vertragsverhältnis nach und zahlt immer innerhalb der vorgegebenen Fristen.

Teilausfall

Der Kunde kommt den Forderungen aus dem Vertragsverhältnis nicht sofort nach. Es treten somit Probleme bei der Erfüllung des Vertragsverhältnisses auf. Hieraus resultieren Maßnahmen von Seiten des Lieferanten zur Zahlungserinnerung, Mahnung oder Übergabe der Forderung an ein Inkassounternehmen. Das Einfordern einer höheren Abschlagszahlung zählt auch hierzu. Ein Teilausfall birgt durch die zusätzlichen Prozesse weitere Kosten, jedoch wird die Forderung zu einem gewissen Zeitpunkt ausgeglichen.

Vollausfall

Der Vollausfall tritt ein, wenn der Kunde während oder nach der gesamten Prozessvielfalt zur Zahlungsaufforderung seine offenen Forderungen weiterhin nicht bezahlen kann oder nicht bezahlen möchte. Nach einem gerichtlichen Prozess kann es durch eine Pfändung und Langzeitüberwachung zu einem gewissen Zeitpunkt zu einer Zahlung kommen, wodurch der Vollausfall

---

69 Darunter fallen u. a. der Widerruf einer Lastschrift, der fehlgeschlagene Lastschrifteinzug mangels Kontodeckung, offene und nicht bezahlte Forderungen aus Zwischen-, Turnus- oder Endabrechnungen.

70 Vgl. Szczesny, Andrea and Kaiser, Ulrich, 2012, S. 316 f.

71 Vgl. ebd., S. 317.

nicht mehr vorhanden wäre. Dennoch werden diese Fälle nach den diversen Prozessen, der verstrichenen Zeit oder auch der Berichtigung der Forderung gegenüber dem Geschäftsergebnis zu den Vollausfällen gezählt.

Sollte es zu einem Zahlungsausfall kommen, haben wettbewerblich agierende Energielieferanten zwei Möglichkeiten: Entweder veranlassen sie eine Sperrung des Zählers (nach § 24 Niederspannungsanschlussverordnung (NAV) für Strom oder nach § 24 Niederdruckanschlussverordnung (NDAV) für Gas) oder sie handeln gemäß ihrer Allgemeinen Geschäftsbedingungen (AGB), wonach der Vertrag meistens gekündigt wird. Für die Prozesse Sperrung und Entsperrung eines Zählers entscheiden sich heute nur wenige Lieferanten, die nicht Grundversorger sind, denn sie müssen ebenfalls die Kosten, die durch den Netzbetreiber für die Sperrung und Entsperrung anfallen, übernehmen und dem Kunden in Rechnung stellen.[72] Aus diesen Gründen entscheiden sich wettbewerbliche Energielieferanten oftmals für die Kündigung des Energieliefervertrags, sodass die offenen Forderungen nicht weiter ansteigen.[73]

Im Jahr 2019 haben Grundversorger und wettbewerbliche Lieferanten insgesamt 221.209 Kündigungen (2018: 185.989) in der Sparte Strom und 54.463 (2018: 54.377) in der Sparte Gas gegenüber ihren Kunden ausgesprochen.[74] Die Kündigungen wurden bei einem Zahlungsrückstand von durchschnittlich 176 € (Strom) und durchschnittlich 170 € (Gas) ausgesprochen.[75] Ob in den genannten Zahlungsrückständen bereits eine Ablesung sowie die Endabrechnung der verbrauchten Energiemenge bis zum Lieferende enthalten ist, ist unklar. Daher kann angenommen werden, dass die Summe von der Kündigung bis zum Lieferende durch die verbrauchte, bisher noch nicht erfasste Energie höher sein kann. Ebenso ist unklar, ob Klärungs- und Prozesskosten enthalten sind. Zu den offenen Forderungen kommen die Kosten, die im Laufe des Prozesses oder der Bedienung des Kunden anfallen. Diese Kosten umfassen die Zeit von Fachmitarbeitern oder auch die Beauftragung von externen Dienstleistern. Zur Veranschaulichung dient exemplarisch ein konstruierter *Worst-Case-Kunde* zur gesamtheitlichen monetären Betrachtung entlang der diversen Prozessschritte. Dabei wird bei manuellen

[72] Durchschnittlich kosteten Sperrungen in 2019 53 € (Strom) und 47 € (Gas). Die durchschnittlichen Wiederherstellungskosten lagen bei 56 € (Strom und Gas). (Vgl. Bundesnetzagentur und Bundeskartellamt, 2021, S. 266, 433).

[73] Vgl. Der Tagesspiegel (o.V.), 09.06.2021.

[74] Grundversorger haben in ihrer Rolle weitaus höhere Anforderungen und Auflagen zur Kündigung einer Lieferstelle in der Grundversorgung, weshalb der Anteil der Kündigungen durch Grundversorger geringer ausfallen dürfte.

[75] Vgl. Bundesnetzagentur und Bundeskartellamt, 2021, S. 267, 433.

Prozessen eine ungefähre Bearbeitungszeit festgelegt. Bei den Kosten wird der Mindestlohn in Höhe von 9,82 € je Stunde zugrunde gelegt[76]; die Kosten eines Fachmitarbeiters im Bereich Accounting & Finance können weitaus höher sein.

Ausgangspunkt für den Worst-Case-Kunden ist ein bestehendes Vertragsverhältnis zwischen ihm und einem Energielieferanten. Die Verbrauchsprognose seitens des Kunden lag bei 3.500 kWh mit einem Abschlag von 117 €/Monat bei einer Vertragslaufzeit von 12 Monaten und 12 Abschlägen. Zusätzlich gibt es einen 100 € Neukundenbonus mit der ersten Rechnung. In der Summe zahlt der Kunde 1.404 € ohne Bonus und 1.304 € mit Bonus im ersten Belieferungsjahr.

Der Kunde zahlt seine Abschläge 10 Monate lang mittels Überweisung

Durch die Zahlung des Kunden mittels Überweisung wird eine Überweisung im Kontoauszug des Lieferanten angezeigt. Die Zuordnung der Buchung zu der offenen Forderung geschieht teilautomatisiert durch Texterkennung, ob die Kundennummer oder Vertragsnummer im Verwendungszweck der Überweisung auftaucht. Kann keine automatisierte Zuordnung stattfinden, muss ein Fachmitarbeiter den Kunden ausfindig machen und die Überweisung manuell einer offenen Forderung zuordnen (sofern möglich).

Zeitaufwand: ca. 10 Minuten/Monat, insgesamt 100 Minuten

Der Kunde zahlt seine letzten 2 Monate mittels Lastschrift

Mit einem vorliegenden Lastschriftmandat darf der Lieferant die Forderungen mittels Lastschrift vom Konto des Kunden einziehen. Die Zahlungszuordnung geschieht durch die Zahlungsreferenz automatisiert und es bedarf keines manuellen Eingriffs. Eine Lastschrift kann aber nicht durchgeführt werden, wenn die Deckung des Kontos fehlt. In solchen Fällen kommt es zu einer Rücklastschrift und der Lieferant muss Rücklastschriftgebühren von ca. 3 € zahlen. Die Rücklastschriftgebühren werden dem Kunden in Rechnung gestellt. Hinzu kommen die offenen Forderungen aus den Abschlägen von 117 €/Monat. Ebenfalls wird ein Fachmitarbeiter über die Rücklastschrift informiert, sodass eine Analyse des Sachverhalts stattfinden kann.

Offene Forderungen: 117 € Abschlagszahlung und 3 € Rücklastschriftgebühren je Monat, 10 Minuten zur Analyse durch einen Fachmitarbeiter, insgesamt 240 € und 10 Minuten

[76] Vgl. Statistisches Bundesamt, 2022.

Der Kunde hat betrügerische Absichten und wartet bis zur Jahresabrechnung

Schlimmer ist es, wenn der Kunde betrügerische Absichten verfolgt und bis zur Jahresabrechnung wartet. Bei der Jahresabrechnung wird der Bonus verrechnet. Zusätzlich gibt der Kunde einen Zählerstand an, mit dem er vorgibt, weniger Energie als in seiner Verbrauchsprognose verbraucht zu haben. Dadurch wird eine Rückerstattung der zu viel gezahlten Abschläge veranlasst. Im gleichen Zuge widerruft der Kunde die letzten zwei gezahlten Abschläge durch den Widerruf der Lastschrift.[77] Dadurch entstehen die Kosten aus dem vorangegangenen Schritt.

100 € aus Bonusauszahlung, ca. 150 € Rückerstattung durch Angabe eines falschen Zählerstands, insgesamt 250 €

Drei Stufen im nicht gerichtlichen Mahnprozess

Der Kunde wird vollautomatisiert über den Zahlungsrückstand hingewiesen und mit einer Frist zur Zahlung aufgefordert. Je Mahnstufe wird ein Brief und eine SMS an den Kunden versandt. Die letzte Mahnstufe beinhaltet zudem eine Kündigungsandrohung des Vertragsverhältnisses. Durch die Briefe und SMS findet eine Kundenaufstörung statt, wodurch der Kunde im Kundenservice anruft und weitere Kosten verursacht. Zudem muss ein Fachmitarbeiter als Business Process Owner die Aufsicht aller Fälle übernehmen.

0,80 € Portokosten für Briefversand/Mahnstufe, 0,08 € SMS-Kosten/Mahnstufe, 15 Minuten durch Anrufe im Kundencenter, 10 Minuten zur Aufsicht der Zahlungsausfälle, insgesamt 2,64 € und 25 Minuten

Kündigung des Vertrags und Übergabe an Inkasso

Nachdem der Kunde innerhalb des Mahnprozesses und auch auf die Kündigungsandrohung nicht reagiert hat, wird das Vertragsverhältnis beendet. Hierzu muss das Lieferende manuell an den zuständigen Netzbetreiber kommuniziert werden. Anschließend wird die Schlussrechnung erstellt und es findet eine Übergabe der schlussgerechneten Forderungen an ein Inkassounternehmen statt. Durch das Inkassounternehmen entstehen – je nach vertraglicher Vereinbarung – keine zusätzlichen Kosten, da die Unternehmen ihre Gebühren zusätzlich zur Forderung aufschlagen dürfen. Beim Lieferanten wird

[77] Der Lastschriftwiderruf ist nur möglich, wenn die Lastschrift durchgeführt werden konnte. Die Kosten und offenen Forderungen werden daher aus dem Punkt zuvor übernommen und nicht ein weiteres Mal aufgelistet.

nur Zeit von Fachmitarbeitern zur Betreuung des Inkassounternehmens (Datenbereitstellung o.ä.) gebunden.

10 Minuten durch manuelle Abmeldung der Lieferstelle, 15 Minuten zur Betreuung des Inkassounternehmens, insgesamt 25 Minuten

Einleitung des gerichtlichen Mahnverfahrens

Wenn das Inkassounternehmen keinen Erfolg vermerken konnte, findet vorab eine Kosten-Nutzen-Abwägung statt. Es muss die Frage beantwortet werden, ob die Prozesskosten – bestehend aus Anwaltskosten und gerichtlichen Kosten – die mögliche Begleichung der offenen Forderungen rechtfertigen. Sollte sich der rechtliche Weg nicht lohnen, findet eine Forderungsnachverfolgung über einen gewissen Zeitraum statt. Sollte der rechtliche Weg jedoch eingeleitet werden, wird der Kunde verklagt und im Erfolgsfall der Konto- oder Lohnpfändung unterzogen. Danach muss der Lieferant nur noch auf die Zahlung der offenen Forderungen warten. Aber selbst an diesem Prozessschritt kann der Kunde eine eidesstattliche Versicherung abgeben, dass er kein Geld hat, oder gar Privatinsolvenz anmelden. Im Zuge dessen wird entweder auf die Forderung verzichtet und gegen das betriebswirtschaftliche Ergebnis des Lieferanten abgeschrieben oder die Forderung wird mit einer Risikoklasse versehen, sodass die Forderung zumindest zum Teil beglichen wird. Je älter die Forderung ist, desto schlechter wird der Forderungswert.

ca. 453,16 € Prozesskosten[78], Forderungsminderung um bis zu 100 %, 30 Minuten Zeitaufwand zur Klärung der Forderung auf Seiten des Lieferanten

Eine Übersicht bietet die Auflistung der Kosten und Forderungen nach dem Durchlauf aller möglichen Prozessschritte eines Worst-Case-Kunden (Tabelle 4):

[78] Basierend auf dem Prozesskostenrechner von Juristische Informationssystem für die Bundesrepublik Deutschland (juris GmbH), 2021, Ausschnitt in Anhang 1.

Kostenpunkt	Forderung in Euro	Zeitaufwand in Minuten	Summe Forderung und Zeitaufwand in Euro[79]
Offene Forderung über ein Beliefe-rungsjahr inkl. Bonus	1.304	-	1.304
davon rechtzeitig ge-zahlte Abschläge durch Überweisung	-1.170	100	-1.153,63
Offene Forderung nach erhaltenen Zahlungen	**134**	-	**150,37**
Rücklastschrift von zwei Abschlägen	6	10	7,64
Betrügerische An-gabe eines niedrige-ren Verbrauchs	150	-	150
Nicht-gerichtlicher Mahnprozess	2,64	25	6,73
Kündigung und Übergabe an Inkasso		25	4,09
Gesamt ohne ge-richtliches Mahn verfahren	**292,64**	**160**	**318,83**
Gerichtliches Mahn-verfahren	453,16	30	458,07
Gesamt mit ge-richtlichem Mahn-verfahren	**745,80**	**190**	**776,90**

Tab. 4: Aufstellung der Kosten und Forderungen eines Zahlungsausfalls am Beispiel eines Worst-Case-Kunden

Die offene Forderung von anfänglich 134 € und einem bereits entstandenen Arbeitsaufwand von 100 Minuten ist nach den verschiedenen Prozessschritten auf eine Gesamtforderung von 292,64 € mit einem Arbeitsaufwand von 160 Minuten gestiegen, ohne Einleitung des gerichtlichen Mahnverfahrens. Auf Seiten des Lieferanten verursacht allein der Arbeitsaufwand 26,19 €, wobei anzunehmen ist,

[79] Der Zeitaufwand wurde mit dem aktuell gültigen Mindestlohn von 9,82 € berechnet. (Vgl. Statistisches Bundesamt, 2022).

dass Fachmitarbeiter weitaus mehr als den heute gültigen Mindestlohn verdienen. Mit gerichtlichem Mahnverfahren wächst die offene Forderung mit Weiterverrechnung der Prozesskosten auf 745,80 €. Die Kosten des Zeitaufwands beim Lieferanten steigen auf 31,10 €.

Ohne das gerichtliche Mahnverfahren steigt die initiale Forderung um rund 218 %, mit dem gerichtlichen Mahnverfahren sogar um rund 556 % (beides jeweils exkl. des Zeitaufwands).

In Anbetracht der effektiven Einnahmen für Beschaffung, Vertrieb und Marge nach Bonusauszahlung mit einer Spanne von -100,75 € bis zu 393,50 € würden durch einen Worst-Case-Kunden mindestens ein, wenn nicht sogar mehrere Deckungsbeiträge vernichtet werden. Unter Umständen wäre es daher sogar kostengünstiger, in einem frühen Stadium auf die Forderung zu verzichten, sodass das Gesamtergebnis des Lieferanten nicht zu stark beeinflusst wird.

Zur Eindämmung eines solchen Risikos und zur Beantwortung der Frage, ob ein Kunde solvent ist oder in der Vergangenheit bereits Zahlungsausfälle hatte, gibt es in Deutschland verschiedene Auskunfteien. Bekannte deutsche Auskunfteien sind die Schufa Holding AG (SCHUFA) und die Creditreform. Beide Unternehmen bieten das Scoring für Privat- und Geschäftskunden an. Da die Skalen anbieterabhängig sind, sind die Werte nur annähernd miteinander zu vergleichen. Der *Basisscore* der SCHUFA wird in Prozent gemessen und reicht von 0 bis 100. Ein Basisscore > 97,5 % deutet auf ein sehr geringes Risiko hin, wogegen ein Basisscore < 50 % auf ein sehr kritisches Risiko hindeutet. Das Risiko ist mit der Erfüllungswahrscheinlichkeit einer vertragsgemäßen Zahlung gleichzusetzen.[80] Hingegen bewertet die Creditreform mit einem Bonitätsindex, der einen Wert von 100 bis 600 annehmen kann. Der Bonitätsindex 100 steht „für eine ausgezeichnete Bonität" und der Bonitätsindex 600 für eine Zahlungseinstellung seitens des Kunden.[81] Wie das Scoring von Privatkunden stattfindet, ist bisher nicht öffentlich bekannt. Zusätzlich wurde im Jahr 2014 vom Bundesgerichtshof (BGH) in Karlsruhe geurteilt, dass trotz der größeren Transparenzschaffung das Scoringverfahren ein Geschäftsgeheimnis der jeweiligen Auskunftei ist und somit zu schützen ist.[82]

In einem Verfahren aus dem Jahr 2001 wird der SCHUFA nahegelegt, dass sie für die Berechnung des Scorewerts eine logistische Regression verwenden soll. Eine

[80] Vgl. SCHUFA Holding AG, 2021a, Frage: Wie kann ich mich selbst einschätzen?
[81] Vgl. Verband der Vereine Creditreform e.V., 2021b.
[82] Vgl. BGH, Urt. v. 28.01.2014, Az. VI ZR 156/13

Datenreife, bei der Kunden richtig klassifiziert werden können, sei frühestens nach 15 Monaten gegeben.[83] Ob dieses Verfahren oder die Dauer zur Datenreife rund 20 Jahre später weiterhin aktuell sind, kann nicht beantwortet werden. Die Auskunfteien selbst werben mit sehr großen Datenbeständen über Privatpersonen in Höhe von 68 Millionen Personen (SCHUFA)[84] oder auch mit mehr als 100 Millionen personenbezogenen Datensätzen (Creditreform)[85]. Die erhobenen und gespeicherten Informationsarten sind vielfältig und umfassen Informationen zu Ratenkrediten, Mobilfunkverträgen, Kreditkarten und deren Nutzung sowie die positive, neutrale oder negative Angabe über den jeweiligen Datensatz. Auch hier mangelt es an Transparenz über die verarbeiteten Daten zur Ermittlung des Scorewerts.

In einer Anfrage an die SCHUFA vom 09.11.2021 (siehe Anhang 2) wurden drei Fragen zum Einsatz des SCHUFA-Scorings für Energielieferanten gestellt und beantwortet:[86]

Welche Informationen sind für eine Auskunft notwendig?

Es werden Personenstammdaten und bei einem Neueinzug idealerweise die Voranschrift benötigt. Je mehr Daten vorhanden sind und wenn insbesondere das Geburtsdatum angegeben wird, erhöht sich die Trefferquote.

Welche Informationen erhält ein Energielieferant?

Der Energielieferant erhält die bei der SCHUFA hinterlegten Personenstammdaten und evtl. gespeicherte Zahlungsstörungen. Optional wird ein „Chancen-Score" angeboten. Beim Chancen-Score fließen nur die vorhandenen Zahlungsstörungen eines Kunden in die Bewertung ein. Damit sollen stark vertriebsgetriebenen Energielieferanten mit nicht so hohen Ausfallhöhen ein energiewirtschaftlich-spezifisches Ausfallrisiko und eine „maximale Annahmequote ohne hohe Ausfallkosten" geboten werden.

83 Vgl. Hüls, R., Henking, A., 2003.
84 Vgl. SCHUFA Holding AG, 2021b.
85 Vgl. Verband der Vereine Creditreform e. V., 2021a.
86 Vgl. Anhang 2, Fokus liegt hier auf Privat- bzw. B2C-Kunden; Antworten wurden zur besseren Lesbarkeit gekürzt.

Welche Kosten entstehen durch die Abfrage bei der SCHUFA? Gibt es Festpreise, Preisbänder oder können Sie mir einen Preisrahmen nennen?

> Der Preis je Kunde und Auskunft liegt – abhängig von der genutzten Schnittstelle und dem genutzten Produkt (z. B. Basisscore oder Chancen-Score) – zwischen 0,60 € und 1,20 €.

Bei der Nutzung einer externen Auskunftei müsste folglich jeder Neukunde geprüft werden. Dabei ist zu beachten, dass nicht jeder Neukunde auch in einem erfolgreichen Vertragsverhältnis mündet. Zwischen Auftragsabgabe durch den Kunden und Vertragsbestätigung des Lieferanten können einige Ereignisse auftreten, die den Vertrag nicht zustande kommen lassen. Hier könnte der Kunde von seinem gesetzlichen Widerrufsrecht Gebrauch machen oder der Lieferant hat durch marktspezifische Prozesse, wie die fehlgeschlagene Identifikation der Messstelle oder eine längerfristige Bindung der Messstelle bei einem anderen Lieferanten, nicht die Möglichkeit, den Kunden zu beliefern. Demnach entstehen Kosten für Kunden, bei denen keine Erlöse generiert werden. Daher müssten die Kosten, die bei einer Bonitätsauskunft anfallen, auf alle aktiven Kunden aufgeteilt werden, bei denen tatsächlich ein Vertrag mit erfolgreicher Belieferung zustande kommt. Die Akquisitionskosten bzw. die Customer Acquisition Cost (CAC) steigen somit unweigerlich.

Da Energielieferanten durch den Marktdruck auf verschiedenen Ebenen wettbewerbsfähig sein müssen und sich vor allem gezeigt hat, dass dafür die (IT-)Prozesse ein großer Hebel sind, stellt sich die Frage, ob die Beauftragung einer externen Auskunftei wirtschaftlich ist.

In der Marktanalyse sowie der exemplarischen Rechnung zum Worst-Case-Kunden wurde gezeigt, dass ein einzelner Kunde für kostenoptimierte Energielieferanten einen großen Einfluss hat. Der Energielieferant ist dem stetigen Wettbewerb ausgesetzt. Zudem folgt bei einem solchen Kunden – je nach Schwere – ein größerer Zeitaufwand, der folglich auch monetäre Auswirkungen hat. Die ausstehenden Forderungen wachsen mit jeder Minute und einigen durchgeführten Prozessschritten weiter an. Es wurde gezeigt, dass der Kunde mehrere Deckungsbeiträge von zahlenden Kunden vernichten kann. Deshalb muss eine Vermeidung der zahlungsunfähigen oder zahlungsunwilligen Kunden im Vordergrund stehen. Ein Modell zur Vorhersage von Kunden mit Zahlungsausfällen muss somit eine hohe Sensitivität aufweisen. Die Sensitivität stellt die Wahrscheinlichkeit dar, mit der ein Zahlungsausfall eines Kunden korrekt vorhergesagt wurde. Ein Modell mit einer besonders hohen Sensitivität könnte jedoch auch viele

Verträge ablehnen, obwohl diese nicht in Zahlungsausfällen resultieren würden. Der positive Vorhersagewert (positive predictive value (PPV)) zeigt den Anteil der richtig positiv klassifizierten Kunden unter allen positiv klassifizierten Kunden. Je höher dieser Wert ist, desto mehr Kunden werden richtig positiv klassifiziert, ergo werden weniger Kunden, die keine Zahlungsausfälle hervorrufen, abgelehnt. Es gilt daher – womöglich sogar Energielieferant-spezifisch – einen Kompromiss zwischen Sensitivität (Erkennung von Zahlungsausfällen) und PPV zu finden.

3 Grundlagen zum Machine Learning

3.1 Logistische Regression

Die logistische Regression basiert auf der logistischen Funktion. Die logistische Funktion gleicht graphisch einer Sigmoidfunktion bzw. einem S-förmigen Graphen. Ihren Ursprung findet die logistische Funktion im 19. Jahrhundert, wo sie unter anderem für die Beschreibung des Bevölkerungswachstums genutzt wurde. In Bezug auf das Bevölkerungswachstum ist – in einer Welt wie sie heute bekannt ist – kein unbegrenztes Wachstum möglich, weshalb keine lineare Funktion zur Beschreibung dessen geeignet ist. Mit mehr Menschen auf der Erde steigt ebenso die Wachstumsrate. Jedoch sinkt der verfügbare Platz auf der Erde mit jedem neuen Menschen, sodass das Wachstum eine obere Schranke benötigt. Die logistische Funktion (Formel 1) kann bei einer Sättigungsgrenze von $\Omega = 1$ einen Wertebereich von 0 bis 1 annehmen. Sie wird durch die Normierung auch logistische Verteilung genannt und stellt eine Wahrscheinlichkeitsverteilung dar. Durch die Verteilung kann eine Aussage über eine binäre Variable getroffen werden (Wahrscheinlichkeit, ob True oder False). Im Beispiel des Bevölkerungswachstums müsste Ω auf die maximale Anzahl der Menschen angepasst werden, die auf der Erde leben können. Durch den Parameter α wird die Position der Funktion auf der X-Achse bestimmt. Der Parameter β definiert die Steigung der Funktion. Ohne Verschiebung der Funktion auf der X-Achse (respektive $\alpha = 0$) schneidet die Funktion bei $x = 0$ unabhängig von der Steigung β die Y-Achse immer an derselben Stelle $f(0) = 0,5$. In Abbildung 3 werden drei normierte logistische Funktionen exemplarisch veranschaulicht.[87]

Formel 1: Die logistische Funktion mit Sättigungsgrenze Ω[88]

$$(x) = \Omega \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{\Omega}{1+e^{-\alpha-\beta x}}. \tag{1}$$

[87] Vgl. Cramer, J. S., 2003, S. 2–4.
[88] In Anlehnung an ebd., S. 4.

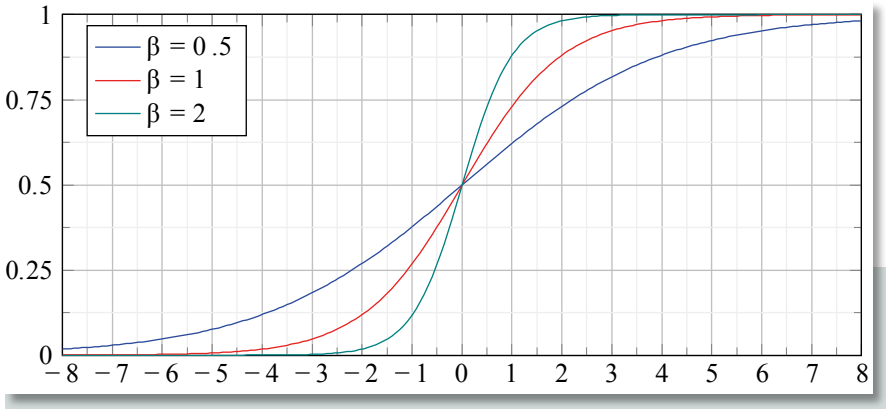


Abb. 3: Drei normierte logistische Funktionen mit Ω = 1 und α = 0

Für die Vorhersage eines Zahlungsausfalls y einer Person i ist die abhängige Variable y_i dichotom, mit $y_i = 0$ was auf keinen Zahlungsausfall und $y_i = 1$ was auf einen Zahlungsausfall hindeutet. Es wird vorab angenommen, dass das Kennzeichen *Hat Einkommen* (als unabhängige, erklärende Variable) der Person x_i mit dem Koeffizienten β, dem Achsenabschnitt α und dem Fehler ε_i zugeordnet ist und einen linearen Zusammenhang darstellt. Die abhängige Variable wird durch das Kennzeichen *Hat Einkommen* erklärt, ergo $y_i = \alpha + x_i\beta + \varepsilon_i$. Es scheint angemessen die Standardannahme zu treffen, dass die Fehler sich im Mittel ausgleichen, $E\{\varepsilon_i|x_i\} = 0$, sodass eine Wahrscheinlichkeit für einen Zahlungsausfall errechnet werden kann $y_i = \alpha + x_i\beta$. Die Wahrscheinlichkeit muss aber normiert werden und zwischen 0 und 1 (0 % und 100 %) liegen.

Formel 2: Die logistische Verteilungsfunktion[89]

$$F_{(w)} = L_{(w)} = \frac{e^w}{1+e^w} = \frac{1}{1+e^{-w}}. \tag{2}$$

Wird nun die logistische Verteilungsfunktion aus Formel 2 und für w der lineare Zusammenhang $w = \alpha + x_i\beta$ genutzt, so ergibt sich ein lineares Wahrscheinlichkeitsmodell, genauer die logistische Regression (auch Logit-Modell genannt), welche durch die Grenzen 0 und 1 beschränkt ist.[90] Um den Zusammenhang zwischen y_i und x_i festzustellen, wird anstelle von y_i die Wahrscheinlichkeit für einen Zahlungsausfall $p = P(Y = 1)$ errechnet, da y_i nur die Werte 0 und 1 annehmen

[89] In Anlehnung an Cramer, J. S., 2003, S. 2.
[90] Vgl. Verbeek, M., 2017, S. 216, 217.

kann, während p jeden Wert zwischen 0 und 1 annehmen kann. Die Chance, dass ein Ereignis eintrifft, wird durch die Wahrscheinlichkeit p (eines Zahlungsausfalls) zur Gegenwahrscheinlichkeit $1-p$ (kein Zahlungsausfall) mit $\frac{p}{1-p}$ dargestellt. Die Chance kann eine beliebige positive Zahl annehmen. Der Logit ist der natürliche Logarithmus der Chance und wird durch $log_e \frac{p}{1-p}$ beschrieben.[91] Der Logit umfasst mit seinem Wertebereich die gesamte reelle Zahlenmenge. Eine Annahme der linearen Beziehung zwischen $logit(p)$ und x_i ist oftmals sinnvoll. Damit wäre eine mathematische Äquivalenz gegeben. Dies ist in Formel 3 dargestellt. p stellt demnach eine zentrale Rolle dar, da einerseits mit dem Logit die Linearisierung der Beobachtungen erfolgt und andererseits kann p durch die logistische Regression auf einer Sigmoide dargestellt werden. Im Falle einer Erklärung der abhängigen Variable y_i durch mehrere unabhängige Variablen wird anstelle von $\alpha + \beta x_i$ eine Linearkombination eingesetzt, wodurch eine multiple logistische Regression mit $y_i = \alpha + x_1\beta_1 + \cdots + x_i\beta_i$ entsteht.[92]

Formel 3: Mathematische Äquivalenz zwischen *logit*(*p*) und x_i[93]

$$\text{logit } (p) = log_e \frac{p}{1-p} = \alpha + x_i\beta$$

$$p = F(\alpha + \beta x_i) = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha}+\beta x_i} = \frac{1}{1+e-(\alpha+\beta x_i)}. \tag{3}$$

Bei einem linearen Zusammenhang in $y_i = \alpha + \beta x_i$ wird für die Schätzung der beiden Parameter α und β insbesondere die Methode der kleinsten Quadrate genutzt. Wenn die Gauß-Markov-Annahmen[94] erfüllt sind und der Störterm normalverteilt ist, dann erzielt die Methode der kleinsten Quadrate die besten, linearen und erwartungstreuen Schätzer. Die Methode der kleinsten Quadrate wäre bei einer dichotomen abhängigen Variable (Zahlungsausfall/kein Zahlungsausfall) nicht effizient und auch der Störterm wäre nicht normal-verteilt. Aus diesem Grund wird zur Schätzung der Parameter die Maximum-Likelihood-Methode, auch Maximum-Likelihood-Schätzung genannt, herangezogen. Die Maximum-Likelihood-Schätzung wählt die Parameter aus, die unter allen Beobachtungen die

91 Vgl. Verbeek, M., 2017, S. 218.
92 Vgl. Bender, R., Ziegler, A., Lange, S., 2007, e33.
93 In Anlehnung an ebd., e33.
94 Der Störterm ergibt im Mittel 0, die Varianz aller Fehler ist gleich der Standardabweichung (Homoskedastizität), die Kovarianz zwischen zwei Störtermen ist 0 (keine Autokorrelation der Fehler) und die erklärende Variable ist unabhängig zum Fehler. (Vgl. Verbeek, M., 2017, S. 15, 16).

höchste Wahrscheinlichkeit haben, wahr ($y_i = 1$) zu sein (das *maximum likelihood*). Ferner und unter Beachtung der Gauß-Markov-Annahmen liefert eine logistische Regression, deren Parameter mit der Maximum-Likelihood-Methode geschätzt wurden, asymptotisch effiziente und konsistente Vorhersagen. Zusätzlich sind Signifikanztests anwendbar.[95,96]

Die Likelihood-Funktion erklärt sich am besten an einem Beispiel aus einem Pool von Personen mit Zahlungsausfällen ($y_i = 1$) $N_1 = \sum_i y_i$ und Personen ohne Zahlungsausfälle ($y_i = 0$) $N - N_1$. Der Anteil der Personen mit Zahlungsausfällen ist angenommen $P\{y_i = 1\} = p$. Dann ist die Wahrscheinlichkeit, eine Person mit Zahlungsausfall zu ziehen:

$$P\{N_1 Person\ mit\ Zahlungsausfall, N - N_1 Person\ ohne\ Zahlungsausfall\} \\ = p^{N_1}(1-p)^{N-N_1}$$

Nun wird mit der Maximum-Likelihood-Methode das *p* gesucht, welches maximal wird ($\hat{p}$). Zur Berechnung des Maximums eignet sich die logarithmierte Variante (Log-Likelihood) besonders, da Produkte in Summen transformiert werden und die Maximierung von Summen einfacher als die Maximierung von Produkten ist. Da die Likelihood-Funktion eine monotone Transformation ist, ist der Maximum-Likelihood-Schätzer nicht durch die Logarithmierung betroffen.[97] Die logarithmierte Variante wird auch Log-Likelihood-Funktion genannt. Die Log-Likelihood-Methode wird in Formel 4 dargestellt. Der Maximum-Likelihood-Schätzer für p ist nach Lösung $\hat{p} = N - N_1$. Der Schätzer folgt demnach der Verteilung der unterschiedlichen Ausprägungen aller Personen.[98]

Formel 4: Log-Likelihood-Funktion[99]

$$\log_e L(p) = N_1 \log_e(p) + (N - N_1)\log_e(1-p). \tag{4}$$

[95] Vgl. Eliason, S. R., 1993, V, Series Editor's Introduction.
[96] Vgl. Verbeek, M., 2017, S. 187, 188.
[97] Vgl. Kähler, J., 2012, S. 57.
[98] Vgl. Verbeek, M., 2017, S. 188 ff.
[99] In Anlehnung an ebd., S. 188.

Formel 5: Log-Likelihood-Funktion in logistischer Verteilungsfunktion eingebettet[100]

$$\log_e L(\beta) = \sum_{i=1}^{N} \gamma_i \log_e F(x_i\beta) + \sum_{i=1}^{N} (1-\gamma_i) \log_e\big(1 - F(x_i\beta)\big) \tag{5}$$

Der Maximum-Likelihood-Schätzer $\widehat{\beta}$ wird folglich durch einen Algorithmus gefunden, der die Maximierung der Log-Likelihood-Funktion anstrebt. Die Wahrscheinlichkeit, dass eine Beobachtung einen Zahlungsausfall hat ($y_i = 1$), kann nun mit Formel 6 errechnet werden.[101]

Formel 6: Berechnung der Wahrscheinlichkeit bei Vorliegen von β[102]

$$\hat{p}_i = \frac{1}{1 + e^{-(\alpha+\beta x_i)}} \tag{6}$$

Zur Bewertung der Güte eines erstellten Modells wird für eine lineare Regression oftmals R^2 genutzt. Mit R^2 wird eine Aussage über die Güte der linearen Approximation gegeben. Je kleiner die Residuen, desto größer das R^2. Bei der Vorhersage von binären Variablen (0 oder 1), wie es bei der logistischen Regression der Fall ist, werden die Werte mithilfe des Logit linearisiert. Da der Logarithmus von 0 oder 1 undefiniert ist, liegen die Beobachtungen bei $-\infty$ bzw. $+\infty$. Eine Messung der Residuen ist daher nicht möglich und es kann kein Gebrauch vom Gütemaß R^2 gemacht werden.

Formel 7: Bestimmtheitsmaß von McFadden[103]

$$R^2_{MF} = 1 - \frac{\log_e L_1}{\log_e L_0} \tag{7}$$

Eine Alternative stellt das Bestimmtheitsmaß von McFadden dar, welches in Formel 7 dargestellt ist. L_1 und L_0 sind Log-Likelihood-Funktionen. L_1 stammt vom trainierten Modell mit allen erklärenden Variablen. L_0 stammt vom untrainierten

[100] In Anlehnung an ebd., S. 220.
[101] Vgl. Verbeek, M., 2017, S. 220 f.
[102] In Anlehnung an ebd., S. 220.
[103] In Anlehnung an Szczesny, Andrea and Kaiser, Ulrich, 2012, S. 332.

Modell, welches die Wahrscheinlichkeit gleich dem arithmetischen Mittel aller Beobachtungen setzt. Somit wird der Unterschied bzw. die Verbesserung gegenüber der Wahrscheinlichkeit über alle Beobachtungen gemessen. Im Idealfall nimmt $R^2_{MF} = 1$ an, wodurch das Modell alle Beobachtungen genau vorhersagt. Wenn das Modell nicht besser als das Raten basierend auf der Verteilung aller Beobachtungen ist, dann ist $R^2_{MF} = 0$. Die Werte von R^2_{MF} reichen wie das R^2_{MF} von 0 bis 1. Ebenso kann für die Bewertung der Modellgüte eine „prediction/realization table", auch Konfusionsmatrix genannt, genutzt werden. Mithilfe der Konfusionsmatrix werden die Vorhersagen, die das Modell trifft, in eine 2x2-Matrix eingetragen. Die Konfusionsmatrix besteht aus richtig positiv und richtig falsch sowie falsch positiv und falsch negativ vorhergesagten Werten. Tabelle 5 enthält eine exemplarische Konfusionsmatrix. Die Konfusionsmatrix gibt einen direkten Aufschluss über die jeweiligen Vorhersagen (richtig oder falsch). Hieraus lassen sich weitere Parameter wie Genauigkeit, Spezifität oder Sensitivität errechnen. Es wird bei der Konfusionsmatrix jedoch nicht ersichtlich, ob eine Vorhersage bei einer Zahlungsausfallwahrscheinlichkeit von 51 % oder 99 % getroffen wurde.[104,105]

tatsächliche Werte	vorhergesagte Werte	
	Kein Zahlungsausfall	Zahlungsausfall
Kein Zahlungsausfall	Richtig negativ	Falsch positiv
Zahlungsausfall	Falsch negativ	Richtig positiv

Tab. 5: Exemplarische Konfusionsmatrix tatsächliche Werte vorhergesagte Werte

3.2 Klassische Entscheidungsbäume

Entscheidungsbäume sind im Bereich des Machine Learnings dem überwachten Lernen (Supervised Learning) zuzuschreiben. Ihr Aufbau folgt dabei einem gerichteten Baum, der hierarchisch angeordnet ist. Innerhalb des Baums dienen Entscheidungsregeln dazu, anhand verschiedener Attribute und deren Ausprägung eine Klassifizierung durchzuführen. Entscheidungsbäume können in der Regel sowohl für Klassifizierungen als auch Regressionen genutzt werden. Der Fokus liegt nachfolgend auf Klassifizierungen. Die Struktur eines Entscheidungsbaums lässt sich ideal in einem Baumdiagramm darstellen. Die Darstellungsmöglichkeit ist ins-

[104] Vgl. Verbeek, M., 2017, S. 221 ff.
[105] Vgl. Szczesny, Andrea and Kaiser, Ulrich, 2012, S. 327 ff.

besondere für fachfremde Personen geeignet, da sie den Klassifizierungen anhand der aufeinanderfolgenden Regeln ohne Vorkenntnisse folgen und diese nachvollziehen können. Klassifikationsverfahren sollten sich vorzugsweise nicht nur durch ihre akkuraten Vorhersagen auszeichnen, sondern auch durch ihre Visualisierungsmöglichkeit einen Einblick in das Verfahren geben und ein Verständnis für die prognostizierten Daten schaffen.[106]

Ein Entscheidungsbaum folgt immer derselben Grundstruktur. Die Struktur ist exemplarisch in Abbildung 4 dargestellt. Zu Anfang wird der Wurzelknoten (Root Node) definiert. Der Wurzelknoten stellt den Ausgangspunkt zur Klassifizierung dar. Er umfasst alle Daten respektive das Set X. Vom Wurzelknoten gehen Äste ab, die eine Entscheidung implizieren. Durch die Entscheidung wird das Set X nach und nach in Subsets $X_1 \dots X_n$ geteilt – der sogenannte Split.[107] Ein Algorithmus zur Erstellung eines Entscheidungsbaums zeichnet sich insbesondere durch seine Split-Funktion aus, um ein (Sub-)Set in weitere Subsets zu teilen und zu klassifizieren. Dabei muss die Aufteilung möglichst rein sein, sodass die Subsets mithilfe weniger Splits klassifiziert werden können. Ergo muss ein hoher Informationsgehalt (Information Gain) in dem Attribut enthalten sein, für welches eine Entscheidung getroffen wird.

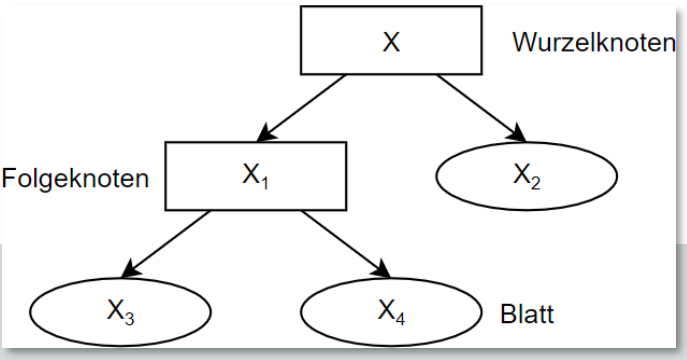


Abb. 4: Baumdiagramm eines Entscheidungsbaums[108]

Nach einem Split kann am abgehenden Ast entweder ein Folgeknoten oder ein Blatt folgen. Wenn auf einen Ast ein Knoten folgt, wird der Prozess der Entscheidung anhand eines weiteren Merkmals oder einer bestimmten Ausprägung durchgeführt. Der Prozess wiederholt sich so lange, bis auf einen Split nur noch

[106] Vgl. Breiman, L. et al., 1984, S. 7.
[107] Vgl. ebd., S. 21.
[108] In Anlehnung an Breiman, L. et al., 1984, S. 21 f.

Blätter folgen. Die Blätter sind die Endpunkte und klassifizieren die Daten anhand der vorangegangenen Entscheidungen innerhalb der Knoten.

Die Tiefe eines Entscheidungsbaums wird durch die hierarchisch aufeinander folgenden Ebenen aus Knoten und Blättern bestimmt. Der Wurzelknoten allein stellt eine Ebene dar, wodurch bereits eine Tiefe von 1 erreicht ist. Ein Baum mit einem Wurzelknoten, einer Knoten-Ebene und einer Blatt-Ebene (wie in Abbildung 4 gezeigt) besitzt eine Tiefe von 3.

Bevor mithilfe verschiedener Algorithmen Entscheidungsbäume erstellt werden können, müssen nach Larose drei Bedingungen erfüllt sein:[109]

1. Da es sich um überwachtes Lernen handelt, muss der vorliegende Datensatz im Voraus klassifiziert sein. Hieraus lassen sich zur Erstellung des Baums ein Trainings- und zur späteren Validierung des Baums ein Testdatensatz bilden.
2. Der Trainingsdatensatz sollte über alle Merkmale und deren Ausprägungen gut durchmischt sein. Der Entscheidungsbaum soll damit über alle Merkmalsausprägungen eine Klassifizierung treffen können, um auch auf unbekannte Daten gut vorbereitet zu sein.
3. Das Zielattribut, für das eine Klassifizierung stattfindet, muss ein diskretes Attribut sein. Ein Datensatz trägt entweder die Ausprägung eines gewissen Zielattributs oder nicht.

3.2.1 Information Gain und Gini Index

Für die Bestimmung an jedem Knoten, wann welches Merkmal mit welchen Merkmalausprägungen einen Split generiert, muss eine Split-Funktion ausgewählt werden. Zwei sehr bekannte und häufig verwendete Split-Funktionen sind der Information Gain und der Gini Index.

Der Information Gain (IG) wurde erstmals 1986 von Quinlan[110] verwendet. Er hat den Entscheidungsbaumalgorithmus Iterative *Dichotomiser 3* (ID3) beschrieben, der viele Attribute verarbeiten kann und zudem nicht allzu viel Rechenzeit benötigt. Für die Bestimmung des besten Splits nutzte er den Information Gain. Der

[109] Vgl. Larose, D. T., 2015, S. 319.
[110] Quinlan, J. R., 1986.

Information Gain nutzt hauptsächlich die Entropie aus Formel 8, um den Informationsgewinn zu berechnen. Dabei wird die Reinheit durch einen Split maximiert bzw. die Unsicherheit durch einen Split minimiert.

Formel 8: Entropie, genutzt im Informational Gain[111]

$$H(X) = \sum_{i=1}^{|Z|} p_i \log_2(p_i) \tag{8}$$

Der ID3-Algorithmus startet am Wurzelknoten mit dem Set S. Der Information Gain bei einem Split mit dem Attribut A mit der Menge aller Attributwerte v errechnet sich aus Formel 9. Mit anderen Worten berechnet sich der Information Gain aus der Differenz der Entropie des Sets S und der mittleren Entropie beim Split des Sets S durch das Attribut A.[112] Die Splits werden so oft auf ein (Sub-)Set durchgeführt, bis ein Set nur noch Daten einer Klasse beinhaltet und in einem Blatt endet.

Formel 9: Informational Gain[113]

$$IG(S,A) = H(S) - \sum_{i=1}^{v} \frac{|s_v|}{s} \cdot H(s_v) \tag{9}$$

Der Gini Index (Gini-Koeffizient) misst ebenso die Reinheit eines Merkmals in einem Datenset S. Im Gegensatz zum Information Gain wird der Gini Index durch die Ungleichheit der Klassen n im Datenset S mit Formel 10 berechnet. Die möglichen Werte liegen zwischen 0 und 1. Der Wert 0 sagt aus, dass alle Datensätze im Datenset S der gleichen Klasse angehören. Der Wert 0,5 zeigt hingegen eine gleichmäßige Verteilung zwischen den Klassen im Datenset S und der Wert von 1 weist eine Ungleichheit der Klassen im Datenset S auf. Es wird daher bei jedem Split das Merkmal gesucht, welches den Gini Index minimiert, um so das Datenset möglichst rein – ein Datenset mit nur einer Klasse – zu erhalten.[114] Der Gini Index

[111] In Anlehnung an ebd., S. 101.
[112] Vgl. ebd., S. 87 ff.
[113] In Anlehnung an Raileanu, L. E., Stoffel, K., 2004, S. 81.
[114] Vgl. Jain, V., Phophalia, A., Bhatt, J. S., 2018, S. 2188.

wird insbesondere vom *Classification and Regression Tree* (CART) von Breiman et al.[115] verwendet.

Formel 10: Informational Gain[116]

$$GI(S) = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{10}$$

Ob der Information Gain auf Basis der Entropie oder der Gini Index auf Basis der Klassenverteilung im Datenset für die Bestimmung des besten Splits besser geeignet ist, haben Raileanu et al. untersucht[117]. Sie kamen dabei zum Schluss, dass die beiden Algorithmen in gerade 2 % aller Fälle voneinander abweichen. Generell geben die beiden Algorithmen die gleichen Ergebnisse. Es kommt daher nur selten zu Abweichungen, bei denen in einigen Fällen der Information Gain und in einigen Fällen der Gini Index bessere Ergebnisse liefert. Daraus folgt, dass es keinen signifikanten Vorteil für einen der beiden Split-Funktionen gibt.[118]

3.2.2 Methoden zur Optimierung

Nachfolgend werden einige Methoden zur Optimierung vorgestellt, die auf Entscheidungsbäume angewandt werden können. Mit diesen Möglichkeiten wird ein Baum entweder im Voraus auf einem anderen Weg erstellt oder gar im Nachhinein verändert. Hierdurch werden sich eine bessere Vorhersagekraft und eine Senkung der Fehlerrate des Baums erhofft.

Beschränkung der maximalen Tiefe

Ein Baum wird rekursiv durch seine Algorithmen aufgebaut, bei dem immer der beste Split gesucht wird, um das Datenset am besten zu teilen. Dieser Weg kann bei einem großen oder komplexen Datenset zu ebenso großen und komplexen Bäumen mit der Gefahr zum Overfitting führen. Das Overfitting wäre spätestens dann erreicht, wenn der Baum für jeden Datensatz ein Blatt besitzt. In diesem Stadium würde der Baum eine Genauigkeit von 100 % gegenüber

[115] Breiman, L. et al., 1984.
[116] In Anlehnung an Raileanu, L. E., Stoffel, K., 2004, S. 81.
[117] Ebd.
[118] Vgl. ebd., S. 91 f.

dem Trainingsdatensatz erreichen, jedoch hätte der Baum in der Realität keine generalisierende Vorhersagekraft.

Um einem Overfitting effektiv vorzubeugen, kann bei der Erstellung eines Baums eine maximale Tiefe festgelegt werden. Durch die Festlegung einer maximalen Tiefe werden Knoten und Blätter bis zu der definierten Tiefe erstellt. Bei Erreichung dieser Tiefe wird das weitere Wachsen des Baums gestoppt. Die maximale Anzahl der Entscheidungsebenen ergibt sich durch die maximale Tiefe abzüglich der Blattebene, ergo maximale Tiefe –1.

Pruning

Das Pruning steht im deutschen Sprachraum für das Beschneiden. Angelehnt ist das Pruning für Entscheidungsbäume an den Gartenbau. Im Gartenbau ist der Beschnitt von Pflanzen für Ernte, Pflege und Form gedacht, um Ertrag oder Optimierungen in Pflege und Ästhetik zu erreichen. Im Bereich der Entscheidungsbäume wird beim Pruning nichts anderes gemacht. Im Voraus wird der Entscheidungsbaum wie üblich mit einem Algorithmus erstellt. Nachdem der Baum vollständig erstellt ist, wird der Baum mithilfe verschiedener Herangehensweisen beschnitten. Dabei werden Unterbäume identifiziert, die eine hohe Fehlerrate mit sich bringen, und anschließend durch Blätter ersetzt. Das Ziel des Prunings ist die Erhöhung der Genauigkeit, die Minimierung des Fehlers oder auch die Reduzierung der Komplexität des Baums.[119]

3.3 Ensemble Methoden

Bei jedem Knoten in einem Entscheidungsbaum unterliegt der Split einer gewissen Unsicherheit. Je mehr Entscheidungen auf dem Weg zum Blatt (bzw. zur Klassifizierung) getroffen wurden, desto höher ist die Unsicherheit einer Vorhersage, die richtige Klasse zu bestimmen. Gleichermaßen steigt die Fehlerrate des Baums. Um dem Umstand einer wachsenden oder hohen Varianz entgegenzuwirken, werden Ensemble Methoden eingesetzt. Ensemble Methoden bedienen sich anstelle nur eines Entscheidungsbaums vieler Entscheidungsbäumen. Die verschiedenen Entscheidungsbäume sind – je nach Wahl der Methode und des Algorithmus – unterschiedlich ausgeprägt und müssen nicht zwingend mit demselben Algorithmus erstellt werden. Die unterschiedliche Ausprägung der Bäume bringt unter anderem ein robusteres Ergebnis z. B. gegen Ausreißer. Die einzige und notwendige Bedingung, Ensemble Methoden einzusetzen, ist, dass die einzelnen

[119] Vgl. Mingers, J., 1989, S. 1.

Bäume eine bessere Fehlerrate als zufälliges Raten haben. Ein ungenauer, auf bestimmte Merkmale trainierter Entscheidungsbaum wird auch als *weak learner* bezeichnet. Der weak learner gibt eine schlechte Vorhersage für ein gesamtes Datenset, welche jedoch besser als zufälliges Raten ist.[120] Der Zusammenschluss aus vielen ungenaueren Entscheidungsbäumen zeigt, dass genauere Ergebnisse als mit einem einzelnen Entscheidungsbaum erzielt werden können.[121]

Die Bewertungsmethoden der verschiedenen Ergebnisse aller Bäume sind nicht festgelegt und können frei gewählt oder adjustiert werden. Klassisch kommt bei einer einfachen Klassifikation ein Mehrheitsvotum aller Bäume zustande, bei dem das Merkmal mit den meisten Stimmen gewinnt. Bei metrischen Vorhersagen ist die Bildung des arithmetischen Mittels aller Bäume üblich. Ebenfalls ist eine spezielle Gewichtung der Bäume möglich, sodass die Vorhersagen einiger Bäume mehr ins Gewicht fallen als andere Bäume.

Die *Bootstrap aggregation*, auch Bagging genannt, ist eine Ensemble Methode und beschreibt die Konstruktion von vielen verschiedenen Entscheidungsbäumen auf Basis von Stichproben aus einem Datenset. Aus der Grundgesamtheit und dem Datenset X mit N Datensätzen werden B Stichproben mit gleichem Umfang n gezogen. Für alle gezogenen Stichproben $X * b$ wird anschließend derselbe Algorithmus für die Erstellung eines Entscheidungsbaums verwendet. Hinzu können viele Variationen zur Bildung der Stichproben folgen. So kann beispielsweise eine Stichprobe nur ausgewählte Attribute oder nur Beobachtungen beinhalten, die als Ausreißer gelten. Die daraus errechneten Bäume sind auf bestimmte Attribute und Merkmale spezialisiert, weshalb sie den weak learner zugeschrieben werden können.

Die Vorhersage jedes Entscheidungsbaums auf Basis einer Stichprobe fließt üblicherweise zu einem gleichen Teil $\frac{1}{B}$ ein. Hierdurch werden die Vorhersagen bzw. das Ergebnis aller Bäume gemittelt. Es sind auch komplexere Ansätze zur Bewertung möglich wie z. B. mit einer Gewichtung der verschiedenen Bäume. Bei Klassifizierungen wird die Klasse j verwendet, welche am häufigsten von den Bäumen vorhergesagt wird ($argmax_j N_j$).[122]

[120] Vgl. Freund, Y., Schapire, R. E., 1997, S. 119.
[121] Vgl. Dietterich, T. G., 2000, S. 1.
[122] Vgl. Breiman, L., 1996, S. 123 ff.

Bagging erwirkt jedoch nur Optimierungen, wenn der *unbagged* Entscheidungsbaum noch nicht optimal ist. Zusätzlich ist das Bagging insbesondere für die Nebenläufigkeit geeignet, da verschiedene Prozesse parallel mit dem Algorithmus sowie den jeweiligen Stichproben Entscheidungsbäume errechnen können.[123]

3.3.1 Random Forest

Der Random Forest-Algorithmus gehört auch zu den Ensemblemethoden. Der Algorithmus bildet viele verschiedene Entscheidungsbäume mit einer zufälligen Ausprägung, ergo einen Random Forest (zufälligen Wald). Random Forest ist mit dem Bagging vergleichbar, jedoch liegt der fundamentale Unterschied in der Auswahl der Attribute M. Während beim Bagging alle Attribute M für einen Split zur Auswahl stehen, stehen beim Random Forest nur m Attribute, bei $m < M$ zur Auswahl. Die algorithmischen Schritte zur Erstellung eines Random Forests lauten dabei wie folgt:[124]

1. Erstelle N Bootstrap-Stichproben vom originalen Datenset.
2. Für jede Bootstrap-Stichprobe $n_1 \dots n_N$ wird ein Entscheidungsbaum trainiert. Dabei wird der beste Split nicht durch alle Attribute M, sondern nur durch eine Auswahl von m Attribute (bei $m < M$) gesucht. Der Entscheidungsbaum wird nicht durch das Pruning-Verfahren zurückgeschnitten, da keine Genauigkeit zur Klassifizierung verloren gehen soll, obwohl Pruning die Generalisierungsgenauigkeit auf ungesehenen Daten erhöht.[125]
3. Nachdem die Bäume n_{tree} erstellt wurden, kann eine einzelne Vorhersage (z. B. durch Zusammenfassung der einzelnen Vorhersagen ein Mehrheitsvotum) getroffen werden.
4. Zur Berechnung der Fehlerrate wird die *out-of-bag (OOB) error rate* (OOB-Fehlerrate) herangezogen. Hierzu werden die Daten, die nicht in der Bootstrap-Stichprobe enthalten sind, vom jeweiligen Entscheidungsbaum vorhergesagt. Die daraus resultierende Fehlerrate ist die OOB-Fehlerrate. Bei genügend trainierten Entscheidungsbäumen spiegelt die OOB-Fehlerrate den Fehler gegenüber einem Testdatensatz recht genau wider.

[123] Vgl. Breiman, L., 1996, S. 133 f.
[124] Vgl. Liaw, A., Wiener, M. et al., 2002, S. 1.
[125] Vgl. Ho, T. K., 1995, S. 278.

Durch dieses Training wird insbesondere die Vielfalt der verschiedenen Entscheidungsbäume vergrößert, da jeder Baum auf andere Merkmale trainiert wird. Zusätzlich wird jeder Baum durch andere Datensätze angelernt, weshalb sich eine hohe Diversität aller Entscheidungsbäume bzw. weak learner ergibt. Durch diesen Ansatz wird insbesondere die Genauigkeit bei zuvor ungesehener, nicht im Trainingsdatensatz enthaltenen Daten erhöht. Somit wird durch Random Forests eine bessere Generalisierbarkeit erreicht.[126] Random Forests bringen nach Breiman erstrebenswerte Charakteristiken mit sich:[127]

- Die Genauigkeit ist genauso gut wie die von Adaptive Boosting (AdaBoost)[128], in einigen Fällen sogar besser.
- Random Forests weisen aufgrund ihrer diversen Ausprägung eine Robustheit gegen Ausreißer und Rauschen auf.
- Sie sind schneller als Bagging und Boosting.
- Es werden interne Schätzungen wie Fehlerrate und Korrelation bereitgestellt.
- Random Forests bringen eine Einfachheit mit und sind ebenfalls parallelisierbar.

3.3.2 XGBoost

Bei XGBoost, ein skalierbares Entscheidungsbaum-Boosting-System, handelt es sich um eine *Gradient Boosting Machine* mit Regularisierung. In Anlehnung an Random Forests oder Bagging, bei denen verschiedene Bäume parallel nebeneinander aufgebaut werden und das Ergebnis aus allen Bäumen ermittelt wird, handelt es sich beim Boosting um einen sequenziellen Aufbau aufeinanderfolgender Entscheidungsbäume. Der sequenzielle Aufbau wird genauer additives Training genannt. Vereinfacht beschrieben erhält der nachfolgende Entscheidungsbaum dasselbe Datenset wie der vorherige Entscheidungsbaum, jedoch mit angepassten Gewichten. Die Gewichte werden, je nach richtiger oder falscher Vorhersage, angepasst, sodass der nachfolgende Baum aus den Fehlern des vorangegangenen Baums lernen kann.

[126] Vgl. -ebd., S. 278, 282.
[127] Vgl. Breiman, L., 2001, S. 5, 10.
[128] AdaBoost ist der erste Boosting-Algorithmus und wird im nachfolgenden Kapitel 3.3.2 vorgestellt.

Mit AdaBoost[129] wurde der erste Boosting-Algorithmus veröffentlicht. Die Funktionsweise von AdaBoost dient der exemplarischen Veranschaulichung eines Boosting-Algorithmus:[130,131]

Jeder einzelnen Beobachtung $\left((x_1, y_1) \dots (x_n, y_n)\right)$ im Trainingsdatensatz X wird ein initiales Gewicht $w_1 \dots w_n$ von $\frac{1}{n}$ zugeordnet.

- Die folgenden Schritte werden T-Mal ausgeführt, für $t = 1 \dots T$:
  - Es wird ein weak learner h_t basierend auf einem einzelnen Attribut erstellt, welches den Fehler in der aktuellen Iteration ε_t minimiert. Bei der Suche nach ε_t muss ein Algorithmus gewählt werden, der das Gewicht einer Beobachtung berücksichtigt und entsprechend höher oder niedriger wertet.
  - Errechne die Gewichtung des weak learner h_t mit $\alpha_1 = \frac{1}{2} \ln\left(\frac{\varepsilon_t}{1-\varepsilon_t}\right)$.
  - Anschließend werden die Gewichte der einzelnen Datensätze $w_1 \dots w_n$ für den darauffolgenden weak learner $(t + 1)$ unter Berücksichtigung der Gewichtung des aktuellen weak learner neu berechnet. Falls ein Datensatz richtig klassifiziert wurde, wird das neue Gewicht mit $w_{n,t+1} = w_{n,t} e^{-\alpha_t}$ verringert. Sollte die Klassifizierung für einen Datensatz falsch sein, wird das Gewicht mit $w_{n,t+1} = w_{n,t} e^{\alpha_t}$ erhöht.
  - Normalisiere die Gewichte aller Datensätze $w_1 \dots w_n$ von $\frac{1}{n}$, damit die Summe 1 ergibt.

Nachdem der Algorithmus durchlaufen ist, liegt ein additiv trainiertes Ensemble aus Entscheidungsbäumen vor. Dies ist in Formel 11 dargestellt. An diesem Algorithmus sind zwei Dinge besonders: Erstens, anders als beim Random Forest, bei dem nur ausgewählte Attribute und Beobachtungen zur Erstellung eines Baums genutzt werden, werden bei AdaBoost alle Beobachtungen gewichtet betrachtet und nur ein Attribut zur Erstellung eines Baums genutzt. Die Tiefe des Baums ist somit 2, da nur der Wurzelknoten und zwei Blätter vorhanden sind. Zweitens, die

[129] Freund, Y., Schapire, R. E., 1997.
[130] In Anlehnung an den AdaBoost-Algorithmus, vgl. Freund, Y., Schapire, R., Abe, N., 1999, S. 2, 3.
[131] In Anlehnung an den AdaBoost-Algorithmus, vgl. Freund, Y., Schapire, R. E., 1997, S. 126.

weak learner haben basierend auf ihrer Fehlerminimierung eine höhere Aussagekraft. Je besser sie die Fehlerrate minimieren, desto stärker (gesteuert durch α_t) fallen sie bei der Klassifizierung einer Beobachtung ins Gewicht.[132]

Formel 11: Boosting-Algorithmus von AdaBoost[133]

$$H(x) = \left| \sum_{t=1}^{T} \alpha_1 n_t(x) \right| \tag{11}$$

Aufbauend auf die Boosting-Algorithmen wurde von Friedman die Gradient Boosting Machine (GBM) entwickelt.[134] Bei der GBM handelt es sich um die Anwendung von Gradient-Boost-Algorithmen. Die Unterschiede zwischen AdaBoost und einem Gradient-Boosting-Algorithmus können wie folgt beschrieben werden:

- Die erste Iteration besteht nicht aus einem Entscheidungsbaum, sondern aus einem einzelnen Blatt. Das Blatt stellt das arithmetische Mittel aus allen Beobachtungen der abhängigen Variable dar. Bei einer Klassifizierung wird aus den vorhandenen Klassen der Logit gebildet. Zusätzlich wird mithilfe der Verlustfunktion aus Formel 12 ein Baum angestrebt, der die Verlustfunktion minimiert. Die Verlustfunktion stellt die negative Log-Likelihood-Funktion dar.[135]
- Die Entscheidungsbäume können größer sein und somit eine Tiefe $>= 2$ haben. Größere Bäume verlieren bei seltenen Beobachtungen ein wenig Genauigkeit, jedoch treffen sie insgesamt bessere und konsistentere Vorhersagen, ergo sie minimieren den Fehler besser.[136]
- Alle Entscheidungsbäume fallen zum gleichen Teil ins Gewicht. Wenn viele Bäume additiv über M Iterationen hinzugefügt werden, besteht die Gefahr zum Overfitting. Dafür wird die learning rate (Lernrate) verwendet, um die Fehlerkorrektur jedes einzelnen Baums in einem gewissen Maße zu erlau-

[132] Vergleich zu Random Forests, wo jeder Baum eine gleiche Aussagekraft hat und ein Mehrheitsvotum stattfindet.
[133] In Anlehnung an Raileanu, L. E., Stoffel, K., 2004, S. 81.
[134] Vgl. Friedman, J. H., 2001.
[135] Vgl.ebd., S. 1198 f.
[136] Vgl. ebd., S. 1215 f.

ben. Die Lernrate ist der Regularisierung zuzuordnen. Je kleiner die Lernrate, desto mehr Bäume (und Iterationen M) werden für eine bessere Approximation benötigt. Eine hohe Lernrate und die daraus folgende hohe Stimmkraft des Baums führt zu weniger benötigten Iterationen. Jedoch wird mit einer hohen Lernrate eine schlechtere Generalisierbarkeit im Vergleich zu einer kleineren Lernrate erzielt.[137]

Formel 12: Verlustfunktion der Gradient Boosting Machine (GBM)[138]

$$L(y,F) = \log_e(1 + e^{-2yF}), \text{ mit } F(x) = \frac{1}{2}\log_e\left[\frac{p_r(y=1|x)}{p_r(y=0|x)}\right]. \tag{12}$$

Mit XGBoost[139] wurde ein spezieller Gradient Boosting-Algorithmus mit Regularisierung entwickelt, der insbesondere bei großen Datensätzen mit vielen Zeilen und Spalten effizient und ressourcensparend funktioniert. Die Zielfunktion von XGBoost in Formel 13 besteht dabei aus zwei Teilen: Der Verlustfunktion und dem Regularisierungsterm. Um den Trainingsverlust zu messen, wird wie bei den Gradient Boosting-Algorithmen die negative Log-Likelihood-Funktion genutzt, siehe hierzu Formel 14. Die Log-Likelihood-Funktion wurde bereits in Kapitel 3.1 näher erläutert. Wird dieser Term alleinstehend verwendet, wird das Modell bessere Ergebnisse auf den Trainingsdaten erzielen. Demzufolge wird angenommen, dass die Grundgesamtheit eine gleichwertige Verteilung aufweist.

Formel 13: XGBoost Zielfunktion[140]

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega f(k) \tag{13}$$

Formel 14: XGBoost Verlustfunktion basierend auf der negativen Log-Likelihood-Funktion[141]

$$l(\theta) = \sum_i [y_i \log_e(1 + e^{-\hat{y}_i}) + (1 - y_i)\log_e(1 + e^{\hat{y}i})] \tag{14}$$

[137] Vgl. ebd., S. 1203 f.
[138] In Anlehnung an ebd., S. 1198.
[139] Vgl. Chen, T., Guestrin, C., 2016.
[140] In Anlehnung an ebd., S. 1198.
[141] In Anlehnung an ebd., S. 1198.

Der zweite Teil der Zielfunktion, der Regularisierungsterm, ist für die Regulierung der Komplexität des zu erstellenden Modells verantwortlich. Er bestraft, sobald das Modell komplex wird. Somit sollen weniger komplexe Bäume konstruiert werden, ergo Bäume mit wenigen Splits und Blättern entstehen. Durch die Regularisierung wird zudem eine kleine Verzerrung hingenommen, um dafür eine größere Minimierung der Varianz zu erzielen. Für Vorhersagen auf Basis ungesehener Daten geben die Modelle mit Regularisierung stabilere/konsistentere Ergebnisse. Durch die alleinige Nutzung der mittleren quadratischen Abweichung würde das Modell im schlechtesten Fall die Beobachtungen verinnerlichen und somit seine Generalisierbarkeit verlieren. Durch den Regularisierungsterm werden das Overfitting verhindert, die Generalisierbarkeit gewahrt und die Vorhersage auf bisher ungesehenen Daten nachhaltig ermöglicht. Der Regularisierungsterm setzt sich aus T für die Anzahl der Blätter und ω für die Gewichte der Blätter zusammen. Der Regularisierungsterm wird in Formel 15 dargestellt.

Formel 15: XGBoost Regulierungsterm[142]

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2. \tag{15}$$

Der Parameter γ steuert im Baum das Minimum eines Split-Verlusts, der bei einem Split erreicht werden muss, um ihn durchzuführen. Je größer γ ist, desto unwahrscheinlicher wird ein Split, ergo der Baum wird weniger komplex mit der Gefahr zum Underfitting. Der Parameter λ ist für die Regularisierung der Blattgewichte zuständig und bestimmt, wie sehr ein Blatt ins Gewicht fallen kann. Hierbei handelt es sich um die L2-Regularisierung. Ein sehr hoher Wert von λ würde die Gewichte eines Blatts sehr klein machen. Ein hoher Wert für beide Parameter, γ und λ, würde sowohl die Anforderungen für einen Split anheben sowie die Blätter bei einem Split weniger ins Gewicht fallen lassen. Dies würde zu einer hohen Verzerrung und ebenfalls zum Underfitting führen.

Neben der L2-Regularisierung verfügt XGBoost über die L1-Regularisierung, die über den separaten Parameter α gesteuert wird. Während die L2-Regularisierung die Summe der quadrierten Gewichte betrachtet, betrachtet die L1-Regularisierung die Summe der Beträge der Gewichte. Die L1-Regularisierung ist zudem für eine

[142] Chen, T., Guestrin, C., 2016, Eq. 2.

feature selection (Variablenauswahl) bekannt, da sie mit steigendem Parameter α immer mehr Variablen ein Null-Gewicht zuordnet.[143,144]

Das Training des Ensembles aus Entscheidungsbäumen verläuft ebenso additiv. Wie beim Gradient Boosting-Algorithmus, startet XGBoost mit einem Startwert, der frei gewählt werden kann. Der Startwert ist im Standard mit dem Wert 0.5 konfiguriert und kann frei gewählt werden.[145] Mit dem Startwert würde die Vorhersage einer Beobachtung bei 50 % liegen, dass $\hat{y}_i = 1$ ist. Auf den Startwert aufbauend, folgen mit jeder Iteration t weitere Entscheidungsbäume wie es in Formel 16 ersichtlich ist.[146]

Formel 16: XGBoost additives Lernen mit der Zielfunktion sowie Lernrate η (schrittweise dargestellt)[147]

$$\hat{y}_i^{(0)} = 0{,}5$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + \eta f_1(x_i)$$

$$\hat{y}_i^{(2)} = f_1(x_i) = f_2(x_i) = \hat{y}_i^{(1)} + \eta f_2(x_i)$$

$$\dots$$

$$\hat{y}_i^{(t)} = \sum_{k=l}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{16}$$

Mit dem aufeinanderfolgenden Training werden die Folgefunktionen immer mit der gleichen Lernrate η hinzugefügt. Für eine Vorhersage werden der Startwert sowie die Ergebnisse der Entscheidungsbäume samt Lernrate nach und nach zusammengerechnet. Das Ergebnis wird anschließend als Vorhersage genutzt. Durch die Addition des konstanten Startwerts sowie der aufeinanderfolgenden Entscheidungsbäume wird auch vom additiven Training gesprochen.

Mit jedem aufeinanderfolgenden Entscheidungsbaum muss die Zielfunktion (aus Formel 14) jede Iteration t und für jeden Entscheidungsbaum i mit der Funktion aus Formel 17 optimiert werden. Deshalb ist das Ziel, das f_t zu finden, welches das Ergebnis minimiert. In anderen Worten: Jeder Beobachtung wird im Schritt t das Residuum des

143 Vgl. Johnson, R., Zhang, T., 2013, S. 950.
144 Vgl. Goodfellow, I., Bengio, Y., Courville, A., 2016, S. 232–233.
145 Vgl. xgboost developers, 2022d, base_score.
146 Vgl. xgboost developers, 2022b, Additive Training.
147 In Anlehnung an ebd., Additive Training.

vorherigen Entscheidungsbaums $t - 1$ zugeordnet. Der Entscheidungsbaum korrigiert immer den Fehler des vorherigen Baums. Auf Basis der Residuen wird ein neuer Entscheidungsbaum trainiert und mithilfe einer Split-Funktion optimiert, sodass Splits auf Basis der größten Fehlerminimierung entstehen. Die Split-Funktion ist bei XGBoost der *Greedy Algorithm for Split Finding*. Jeder Entscheidungsbaum startet bei einem einzelnen Blatt, dem Startknoten, mit allen Beobachtungen. Anschließend werden iterativ alle Attribute und Schwellenwerte für einen Split der Beobachtungen getestet, um den maximalen Gain zu erhalten. Die Split-Funktion ist in Formel 18 ersichtlich. Darin stehen I_L und I_R für die Elemente im linken bzw. rechten Blatt sowie g_i und h_i für die erste und zweite Ordnung des Gradienten der Verlustfunktion. Ebenso wird der L2-Regularisierungsterm λ genutzt, um das Gewicht im Sinne eines Varianzverlusts eines Blattes zu schmälern. Das Ergebnis wird am Ende mit γ subtrahiert. γ steuert den minimalen Gain, den ein Split erreichen muss. Wird $L_{split} <= 0$, dann wird der Split verworfen, um ein Overfitting zu verhindern. Somit findet ein *ad hoc Pruning* statt.[148,149]

Formel 17: XGBoost additives Lernen mit der Zielfunktion[150]

$$L^{(t)} = \sum_{i=l}^{n} l\left(y_i{}', \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{17}$$

Formel 18: XGBoost Split-Funktion[151]

$$L_{\text{grad}} = \frac{1}{2}\left[\frac{\left(\Sigma_{i\in I_L} g_i\right)^2}{\Sigma_{i\in I_L} h_i + \lambda} + \frac{\left(\Sigma_{i\in I_L} g_i\right)^2}{\Sigma_{i\in I_L} h_i + \lambda} - \frac{\left(\Sigma_{i\in I_L} g_i\right)^2}{\Sigma_{i\in I_L} h_i + \lambda}\right] - \gamma \tag{18}$$

[148] Vgl. Chen, T., Guestrin, C., 2016, S. 787.
[149] xgboost developers, 2022b, Learn the tree structure.
[150] In Anlehnung an Chen, T., Guestrin, C., 2016, S. 786.
[151] ebd., S. 787.

4 Datengrundlage

Die Datengrundlage wird durch einen Datensatz repräsentiert, der von einem ehemaligen Energielieferanten stammt. Dieser Energielieferant war vom 22.10.2018 bis zum 01.06.2021 ausschließlich wettbewerblich und nicht als Grundversorger am deutschen Energiemarkt tätig. Anschließend hat das Mutterunternehmen eine Gesamtrechtsnachfolge angetreten und den Energielieferanten mit der eigenen Marke verschmolzen. In der Zeit des aktiven Vertriebs des Energielieferanten setzte sich das Sortiment aus den Sparten Strom und Gas zusammen. Die Sparte Strom wurde seit Markteintritt angeboten, wohingegen die Sparte Gas erst nach September 2019 in den Vertrieb aufgenommen wurde. Der Vertrieb erfolgte rein digital. Der Vertragsabschluss erfolgte Online über die eigene Bestellstrecke oder über Vergleichsportale. Die Kundenkommunikation war per AGB auf den E-Mail-Kanal beschränkt und ein Kundenportal ermöglichte die selbstständige Verwaltung des Vertrags. Falls der Kunde mit dem Kundenservice in Kontakt treten wollte, konnte er einen Rückrufwunsch einstellen, der von einem geschulten Mitarbeiter eines Servicedienstleisters erfüllt wurde. Es wurden keine Briefe versandt und es gab keinen örtlichen Servicepunkt. Die gesamte Reise des Kunden, vom Vertragsabschluss bis zur Kündigung, verlief rein digital.

Die Datenextraktion erfolgte von einer PostgreSQL-Datenbank zu einem Zeitpunkt, in der der Zustand der Daten im Rahmen der Migration in das Mutterunternehmen eingefroren wurde. Der extrahierte Datensatz umfasst insgesamt 155.926 Einträge mit jeweils 31 gleichen Datenfeldern. Jeder Eintrag stellt einen Vertrag zwischen Kunden und Energielieferanten dar. Das SQL-Skript zum Datenexport aus der Quelldatenbank ist in Listing 4 in Anhang 3 enthalten.

4.1 Struktur und Attributbeschreibung

Neben den aktuellen Daten wurden ebenfalls historische Daten miteinbezogen. Mit den historischen Daten ließen sich die Ursprungsdaten wiederherstellen, die bei einer Angebotsabgabe durch einen Kunden übermittelt wurden. Ebenfalls lassen sich aus den historischen Daten Verhaltensmerkmale ableiten und die Aktionen eines Kunden messen. Der Datensatz kann somit grundlegend in zwei Gruppen unterteilt werden:

1. Daten, die zum Zeitpunkt der Angebotsabgabe durch den Kunden zur Verfügung standen.
2. Verhaltensdaten, die mit einem Betrachtungszeitraum von einem Jahr nach Belieferungsbeginn des jeweiligen Vertrags erfasst wurden.

Von den 31 Datenfeldern sind insgesamt 18 Felder bei der Angebotsabgabe durch den Kunden bekannt und können der ersten Gruppe zugeordnet werden:

- Sparte. Bezeichnet den im Vertragsverhältnis gelieferten Energieträger.
- Anrede. Steuert die persönliche Ansprache in der Kundenkommunikation und lässt einen Rückschluss auf das Geschlecht zu.
- Alter. Wird bei Angebotsabgabe aus rechtlichen Gründen (volle Geschäftsfähigkeit) und im Kundenservice zur Verifizierung abgefragt.
- E-Mail-Domain. Die gesamte Kommunikation erfolgt digital über das E-Mail-Postfach des Kunden und ist verpflichtend, sodass vertragsrelevante Dokumente wie Rechnungen zugestellt werden können.
- Hat Telefonnummer. Stellt ein Kennzeichen dar, ob der Kunde seine Telefonnummer bei der Angebotsabgabe freiwillig eingegeben und somit übermittelt hat.
- Postleitzahl der Lieferstelle. Gibt die Postleitzahl des Ortes an, an dem die Belieferung stattfinden soll.
- Postleitzahl des Vertragsnehmers. Beim Bestellprozess wird die Möglichkeit gewährt, optional eine abweichende Rechnungs- und Postanschrift des Vertragsnehmers anzugeben. Ein Beispiel für die Nutzung dieser Möglichkeit ist die Bestellung eines Energieliefervertrags für die eigene Ferienwohnung mit Angabe des dauerhaften Wohnsitzes als abweichende Postanschrift.
- Ist Post- und Lieferanschrift identisch. Stellt ein Kennzeichen dar, ob eine Gleichheit der Post- und Lieferanschrift gegeben ist. Bei der Prüfung wurde ein Vergleich auf Feldebene zwischen Straße, Hausnummer, Postleitzahl und Ort durchgeführt.
- Bank Identifier Code (BIC). Der BIC wird bei Kunden erfasst, die ihren Energieliefervertrag mit der Zahlungsart Lastschrifteinzug bezahlen möchten. Der BIC ist ein international standardisierter Code, mit dem Geldinstitute eindeutig identifiziert werden können. Er setzt sich aus einer alphanumerischen Zeichenfolge mit 8 bis 11 Zeichen zusammen. Darin enthalten sind der Bankcode, der das Geldinstitut identifiziert, der Ländercode nach ISO 3166-1, ein

wählbares Geldinstitut-Suffix sowie optional der Branchcode, der Aufschluss über einen Ort, eine Abteilung o. ä. gibt.[152]

- Zahlungsart. Stellt die gewünschte Zahlungsart dar, die vom Kunden bei der Angebotsabgabe eingestellt wurde.
- Vertriebspartner. Bildet den Kanal ab, über den der Vertrag gewonnen wurde. Kanäle können Bestellstrecken, externe Dienstleister oder Affiliate-Programme sein.
- Hat aktiven Altlieferant. Das Kennzeichen gibt einen Aufschluss, ob der Kunde einen aktiven, laufenden und ungekündigten Vertrag bei einem Lieferanten hat. Der Kunde äußert den Wunsch, die Kündigung durch den neuen Energielieferanten vornehmen zu lassen. Sollte das Kennzeichen gepflegt sein, hat der Kunde seinem bisherigen Lieferanten nicht eigenständig gekündigt. Die Kündigung wird somit durch regulierte Marktprozesse vom neuen Lieferanten durchgeführt.
- Abrechnungsintervall. Gibt die Zeitspanne eines Abrechnungszyklus an. Der Kunde erhält nach der Zeitspanne eine Aufforderung zur Ablesung und Übermittlung seines Zählerstands, damit die Turnusrechnung auf Basis seiner verbrauchten Energiemenge erstellt wird.
- Bonuswert. Zeigt den Gesamtwert in Euro aller Bonuszahlungen an. Zu den Bonusarten zählen unter anderem der Sofortbonus (Zahlung bei Belieferungsbeginn[153]) oder der Rechnungsbonus (Auszahlung nach einer bestimmten Anzahl von Tagen in Belieferung).
- Verbrauchsprognose. Gibt den erwarteten Verbrauch in kWh an, den der Kunde bei Angebotserstellung übermittelt hat.
- Energiepreis. Gibt den im Angebot vereinbarten Energiepreis in Euro je verbrauchter kWh an. Darin enthalten sind Vertriebs- und Beschaffungspreise sowie Entgelte für die Netznutzung, Steuern und Umlagen.
- Grundpreis. Gibt den im Angebot vereinbarten Grundpreis in Euro je Monat an. Der Grundpreis umfasst unter anderem Kosten für den Betrieb und die Instandhaltung der Messstelle, den Energiezähler sowie administrative und operative Aufwände wie die Ablesung und Erfassung des Zählerstands.
- Abschlusswochentag. Stellt den Wochentag dar, an dem der Kunde sein Angebot über einen Kanal abgegeben hat.

[152] Vgl. *Society for Worldwide Interbank Financial Telecommunication (SWIFT)*, 2018, S. 5, 7, 13.
[153] Der Auszahlungszeitpunkt unterscheidet sich je nach Energielieferanten. Für üblich wird vor der Auszahlung eine gewisse Zeitspanne nach Belieferungsbeginn abgewartet, um den Kunden nicht vor Vertragsende zu verlieren (z. B. durch einen Widerruf oder marktspezifische Prozesse).

Mahn-stufe	Mahnsta-tus[154]	Aktionen
1	xx1 / xx2	Erstes Mahnschreiben mit Zahlungserinnerung
2	xx3 / xx4	Zweites Mahnschreiben mit Zahlungserinnerung und Kündigungsandrohung
3	xx5	Kündigung des Vertrags sowie Information über zeitnahe Übergabe an externes Inkassounternehmen
4	4xx	Forderungsübergabe an externes Inkassounternehmen

Tab. 6: Aufstellung der Mahnstufen und dazugehörigen Aktionen aus dem internen Mahnwesen

Der zweiten Gruppe sind die restlichen 13 Felder zuzuordnen. Diese Felder verfolgen das Verhalten der Kunden im Zeitraum von Auftragserteilung bis zu 365 Tagen, respektive einem Jahr, nach Belieferungsbeginn. Da die Erhebung der Daten erst nach einem gültigen Vertragsschluss zwischen Energielieferanten und Kunden durchgeführt wird, sind die Daten bei einem Neukunden vorab unbekannt und werden erst mit der Zeit generiert. Die Felder der zweiten Gruppe sind wie folgt definiert:

- Ist aktiv. Sobald ein Lieferende durch den Netzbetreiber gemeldet wurde, läuft der Vertrag aus und ist nicht mehr aktiv. Aktive Verträge haben im Umkehrschluss kein bestätigtes bzw. ein offenes Lieferende.
- Belieferungszeitraum. Gibt in Tagen an wie lange der jeweilige Vertrag in Belieferung ist/war.
- Erste Mahnung nach Belieferungsbeginn. Stellt in Tagen dar, wie lange der Kunde vor seiner ersten Mahnung in Belieferung war.

Höchster erreichter Mahnstatus. Die Kennzahl zeigt im internen Mahnwesen den höchsten jemals erreichten Mahnstatus zum Vertrag. Der Mahnstatus kann eine der insgesamt drei Mahnstufen annehmen. In jeder Mahnstufe wird der Kunde über unterschiedliche Kanäle kontaktiert und erhält jeweils andere Schreiben. Zusätzlich wird die Übergabe der Forderung an ein externes Inkassounternehmen als *vierte Mahnstufe* gepflegt. Die Mahnstatus, die jeweiligen Mahnstufen und

[154] Beim Mahnstatus definieren die erste Ziffer die jeweilige Zahlungsart und die zweite Ziffer den jeweiligen Status im Prozess. Die letzte Ziffer ist für die Mahnstufe ausschlaggebend. Eine Besonderheit ist bei Werten von 400 bis 499 vorhanden: Bei diesen Werten handelt es sich um die Übergabe der Forderung an ein externes Inkassounternehmen.

deren Aktionen sind in Tabelle 6 aufgeführt. Eine Mahngebühr wurde nicht erhoben.

- Anzahl der eingestellten Rückrufwünsche. Bei Kundenfragen konnte der Kunde Rückrufwünsche zu seinem Vertrag einstellen. Nach der Einstellung wurde der Kunde telefonisch kontaktiert, um ihm Fragen zu seinem Vertragsverhältnis zu beantworten sowie Hilfestellungen zu leisten.
- Anzahl der geschriebenen E-Mails. Der Kunde konnte den Kundenservice via E-Mail erreichen. Durch eine Zuordnung der E-Mailanfrage zum jeweiligen Vertrag hat der Kundenservice die Kundenanliegen vertragsspezifisch beantwortet und Hilfe geleistet.
- Anzahl der genutzten Bankdaten. Das Online-Kundenportal ermöglichte die Änderung der aktuellen Bankdaten durch den Kunden. Durch die Änderung wurde ein neues Lastschriftmandat erteilt, mit dem der Lieferant Forderungen via Lastschrift einziehen kann.
- Anzahl der eingegebenen Zählerstände. Die Zählerstandseingabe durch den Kunden wurde mittels Online-Kundenportal zu jedem Zeitpunkt auf freiwilliger Basis in und nach Belieferung ermöglicht. Nach jedem Turnus – je nach Abrechnungsintervall monatlich oder jährlich – wurde der Kunde zur Eingabe eines Zählerstands aufgefordert, damit die zu erstellende Turnusrechnung nicht auf Basis eines geschätzten Verbrauchs durch den Lieferanten geschieht.
- Anzahl der Abschlagsverminderungen. Bei Vertragserstellung wird ein Abschlagsplan basierend auf den vereinbarten Preisen und der Kundenverbrauchsprognose erstellt. Im Nachhinein besteht die Anpassungsmöglichkeit durch den Kunden. Eine Verminderung des Abschlags kommt unter anderem durch ein geändertes Verbrauchsverhalten oder den Wegfall energieintensiver Geräte seitens des Kunden zustande.
- Anzahl der Abschlagserhöhungen. Analog zu den Verminderungen kann der Kunde seinen Abschlag erhöhen. Eine Erhöhung wird unter anderem durch einen höheren Energiebedarf (z. B. durch Trocknungsmaßnahmen) ausgelöst oder wenn der Kunde in seiner Rechnung eher eine Gutschrift anstelle einer Nachzahlung wünscht.

Anzahl der Rücklastschriften. Eine Rücklastschrift ist eine fehlgeschlagene oder widerrufene Ausführung des Lastschrifteinzugs eines Geldbetrags vom Kunden

durch den Lieferanten. Die fehlgeschlagene Ausführung kann durch mehrere Gründe zustande kommen. Dazu gehören ein ungedecktes Konto des Kunden, falsche Kontodaten oder auch der Lastschriftwiderruf, bei dem das Geld erst erfolgreich eingezogen wurde, dann aber zurückgebucht wird.

- Anzahl manueller Überweisungen. Neben den automatisierten und offiziellen Prozessen wie dem Lastschrifteinzug haben Kunden ihre Zahlungen mit einer Überweisung an das Bankkonto des Lieferanten getätigt. Manuelle Überweisungen bedürfen unter Umständen einer manuellen Zahlungszuweisung auf Seiten des Lieferanten, wenn der Verwendungszweck kein identifizierendes Merkmal wie die Vertragsnummer enthält. Umso schwieriger wird die Zuordnung, wenn der Name, der Verwendungszweck oder die IBAN nicht im System wiedergefunden werden.
- Kündigungseingang nach Belieferungsbeginn. Die Kennzahl gibt in Tagen an, wann der Kunde nach seinem Belieferungsbeginn die Kündigung seines Energieliefervertrags ausgesprochen hat. Bei einer Kündigung wird unter Einhaltung der Mindestvertragslaufzeit, der Kündigungsfrist sowie des gewünschten Kündigungszeitpunkts die jeweilige Lieferstelle abgemeldet und ein Lieferende in der Zukunft vorgemerkt.

Um bereits bei der Datenextraktion nur relevante Konstrukte zu berücksichtigen, wurden Selektionskriterien definiert. Die Selektionskriterien verhelfen neben der Filterung relevanter Konstrukte auch zu einem kleineren Datensatz und zu weniger Aufwand bei der Datenaufbereitung. Konstrukte, die extrahiert wurden und für die Analyse relevant sind, wurden wie folgt definiert:

- Es war oder ist ein aktiver Vertrag, bei dem Energie an den Kunden geliefert wurde. Widerrufene und gesperrte Verträge sowie Verträge, die aufgrund einer fehlgeschlagenen Anmeldung nie in Belieferung genommen wurden, sind nicht berücksichtigt. Grund dafür sind fehlende Zahlungsvorgänge und somit auch fehlende Kennzeichen, ob ein Vertrag mit Zahlungsausfällen behaftet ist oder nicht.
- Der Belieferungsstart begann zwischen dem 1. März 2018 (inklusiv) und dem 28. Februar 2021 (inklusiv).

Es handelt sich nicht um einen Folgevertrag, der im Rahmen eines Umzugs für die neue Lieferstelle erzeugt wurde. Logisch ist es ein neues Vertragsobjekt, jedoch

bleibt das bestehende Vertragsverhältnis bestehen. Somit ist es fachlich kein neuer Vertragsabschluss und wird deshalb exkludiert.

4.2 Deskriptive und explorative Analyse

Das Datenset, mit den in Kapitel 4.1 vorgestellten Datenfeldern, wurde für die Analyse eingelesen. Die insgesamt 31 Datenfelder mit Datentypen und dem Non-Null-Zähler sind im Listing 1 aufgeführt. Von insgesamt 155.926 Datensätzen sind 9.918 Datensätze mit einem Mahnstatus (*highestdebtstatus*) versehen. Somit sind rund 6,36 % aller Verträge mit Zahlungsproblemen belastet. Mithilfe der Pandas-Methode *describe()* wird vorab ein Blick auf die deskriptiven Statistiken der Daten geworfen. Die Ausgabe ist in Tabelle 7 dargestellt und enthält unter anderem die Mittelwerte, Standardabweichungen sowie Minimum- und Maximalwerte der jeweiligen Felder. Die graphischen Darstellungen als Boxplots und Histogramme sind gesammelt für die numerisch diskret verteilten Datenfelder in Abbildung 5 vorhanden. Weitere Abbildungen nicht genannter Datenfelder befinden sich im Anhang 4.

Listing 1: Zusammenfassung der Ausgangsdaten

```
RangeIndex: 155926 entries, 0 to 155925 Data columns (total 31 columns):
#     Column                                        Non-Null Count    Dtype
---   ------                                        --------------    -----
0     section                                        155926 non-null object
1     salutation                                     155882 non-null object
2     age                                            155917 non-null float64
3     email                                          155926 non-null object
4     hasphonenumber                                 155926 non-null bool
5     deliverypostalcode                             155926 non-null object
6     ownerpostalcode                                155926 non-null object
7     ispostalanddeliveryaddressindentical           155926 non-null bool
8     bic                                            154290 non-null object
9     paymentmethod                                  155926 non-null object
10    distributionpartner                            155914 non-null object
11    hasoldsupplier                                 155926 non-null bool
12    invoiceinterval                                155926 non-null object
13    bonusvalue                                     155755 non-null float64
14    consumptionprognosis                           155926 non-null float64
15    energycostkwh                                  155921 non-null float64
16    basepricemonthly                               155921 non-null float64
17    conclusionweekday                              155926 non-null object
18    first_dunning_in_days_after_supplybegin          9872 non-null float64
```

```
19   highestdebtstatus                                9911 non-null float64
20   supplyinterval_in_days                         155926 non-null float64
21   is_active                                      155926 non-null bool
22   callcount                                       29912 non-null float64
23   emailcount                                      34722 non-null float64
24   bankaccountcount                               155718 non-null float64
25   meterreadingcount                              155926 non-null int64
26   payplanreducedcount                              4509 non-null float64
27   payplanincreasedcount                            5044 non-null float64
28   separeversedcount                              155926 non-null int64
29   cancellationreceivedaftersupplybeginindays      56788 non-null float64
30   manualsepacount                                155926 non-null int64
```

	Alter	Bonuswert	Verbrauchsprognose	Energiepreis	Grundpreis	Erste Mahnung nach Belieferungsbeginn	Höchster erreichter Mahnstatus
count	155917.0	155755.0	155926.0	155921.0	155921.0	9872.0	9911.0
mean	47.055	167.869	2988.237	641.584	649.663	155.586	199.998
std	16.923	76.506	2199.309	253248.973	253248.953	95.404	106.035
min	18.0	11.0	1.0	0.04	3.289	3.0	0.0
max	1826.0	653.0	165828.0	99999999.99	99999999.99	365.0	403.0

	Belieferungsintervall	Anzahl eingestellter Rückrufwünsche	Anzahl geschriebener E-Mails	Anzahl genutzter Bankdaten	Anzahl eingegebener Zählerstände
count	155926.0	29912.0	34722.0	155718.0	155926.0
mean	284.133	2.002	1.891	1.053	0.64
std	102.266	1.803	1.67	0.248	1.32
min	0.0	1.0	1.0	1.0	0.0
max	365.0	49.0	41.0	7.0	47.0

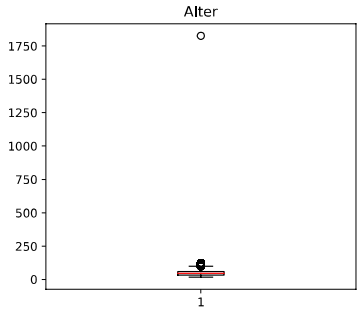
	Anzahl Abschlagsverminderungen	Anzahl Abschlagserhöhungen	Anzahl Rücklastschriften	Kündigungseingang nach Belieferungsbeginn	Anzahl manueller Überweisungen
count	4509.0	5044.0	155926.0	56788.0	155926.0
mean	1.243	1.165	0.328	254.749	0.1
std	0.743	0.506	1.473	97.395	0.623
min	1.0	1.0	0.0	-413.0	0.0
max	21.0	11.0	30.0	365.0	26.0

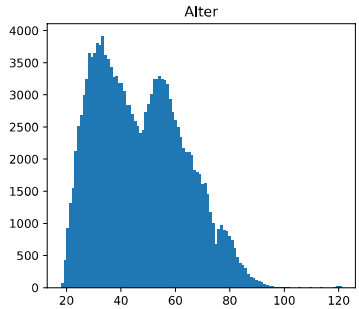
Tab. 7: Deskriptive Statistiken der Ausgangsdaten

In der Spalte Alter (*age*, Abbildung 5a) fällt der Maximalwert mit 1826 Jahren auf. Durch einen Blick auf die Quell-Datenbank ist ersichtlich, dass die Jahresangabe im Geburtsdatum nur dreistellig ist. Bei diesem Ausreißer handelt es sich sehr wahrscheinlich um einen fehlerhaften manuellen Kundenaufbau oder eine fehlerhafte manuelle Korrektur des Kundendatensatzes. Am Histogramm wird auch deutlich, dass beim digitalen Lieferanten besonders viele Kunden mit einem Alter zwischen 25 und 40 Jahren sowie zwischen 50 und 60 Jahren vertreten sind.

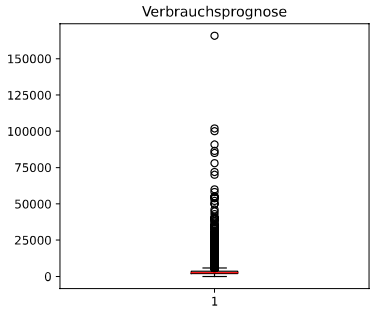
Eine Auffälligkeit stellt der Maximalwert der Spalte Verbrauchsprognose (*consumptionprognosis*, Abbildung 5b) dar. Während der Mittelwert bei ca. 2.988 kWh und das 75 %-Quantil bei 3.500 kWh liegen, ist der Maximalwert mit 165.828 kWh im Vergleich außerordentlich hoch. Es wird angenommen, dass es sich bei dem Kunden um keinen Haushaltskunden, sondern um einen Geschäftskunden handelt. Zudem ist der Minimalwert mit 1 kWh außerordentlich gering. Der Liefervertrag wird abgeschlossen worden sein, bevor eine Mindestabnahmemenge eingerichtet wurde. Bei dem Verbrauch ist nahezulegen, dass es sich um einen Leerstand ohne Verbrauch (z. B. eine leere, unbewohnte Wohnung) handelt. Als Weiteres fallen die Maximalwerte des Energiepreises und des Grundpreises (*energycostkwh* und *basepricemonthly*, Abbildungen 5c und 5d) in Höhe von 100.000.000 Cent bzw. Euro auf. Diese Werte sind im Energiemarkt nicht realistisch und eindeutig Fehlern zuzuschreiben.

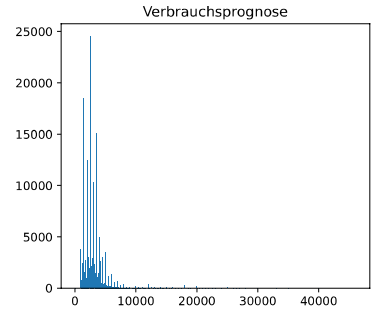
Die Anzahl der eingegebenen Zählerstände und der genutzten Bankdaten (*meterreadingcount* und *bankaccountcount*, Abbildungen 5e und 5f) sowie der der eingestellten Rückrufwünsche und geschriebenen E-Mails (*emailcount* und *callcount*, Abbildungen 5g und 5h) weisen rechtsschiefe Verteilungen auf. Sehr häufig wurde nichts geändert und auch nicht der Kundenservice kontaktiert. Bei den Zählerständen fallen die Ausreißer auf. Mit einem Maximalwert von 47 Zählerständen im Belieferungsjahr kann es sich um Kunden mit Mehrregister-Zählern, die mehrere Zählerstände haben, und/oder um intelligente Messysteme handeln, die die Zählerstände automatisiert übermitteln. In der Regel ist nur ein Zählerstand zur Erstellung der Turnusrechnung zum Ende eins Belieferungsjahres durch den Kunden zu erfassen. Von 34.722 Kunden wurden im Schnitt 1,89 E-Mails je Kunden geschrieben. Insgesamt 29.912 Kunden haben Rückrufwünsche eingestellt, die mit ca. 2 Rückrufwünschen je Kunden im Mittel etwas höher liegen. Die Ausreißer liegen mit Maximalwerten bei 49 Rückrufwünschen und 41 E-Mails.



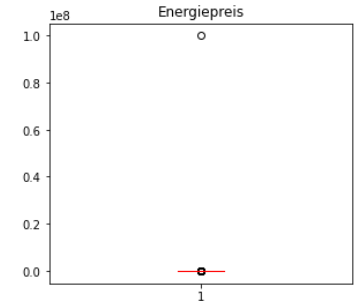


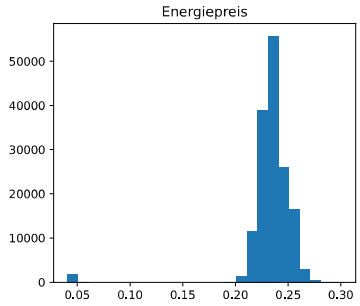
a) Alter



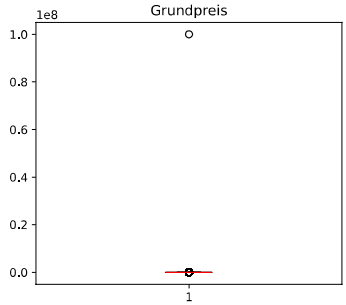


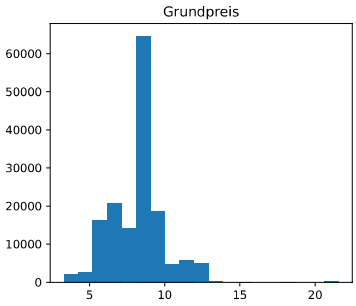
b) Verbrauchsdiagnose



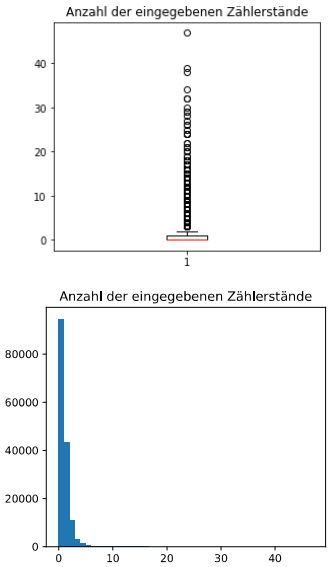


c) Energiepreis je Kilowattstunde

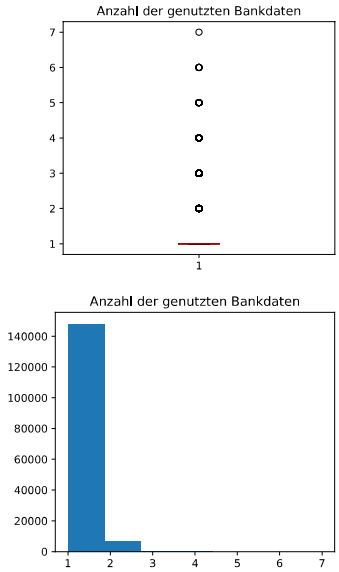




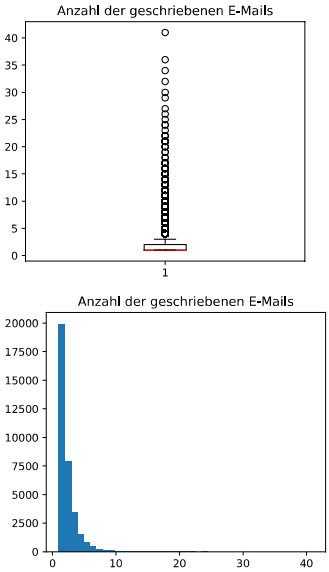
d) Grundpreis je Monat



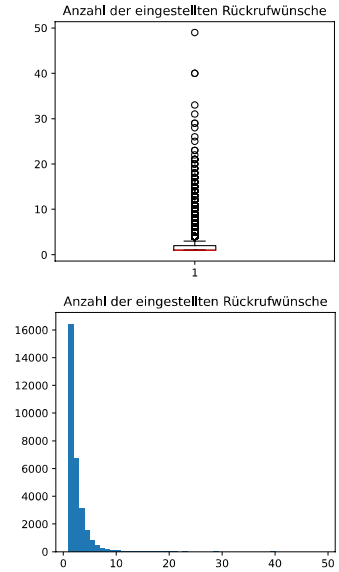
e) Anzahl der eingegebenen Zählerstände



f) Anzahl der genutzten Bankdaten



g) Anzahl der geschriebenen E-Mails



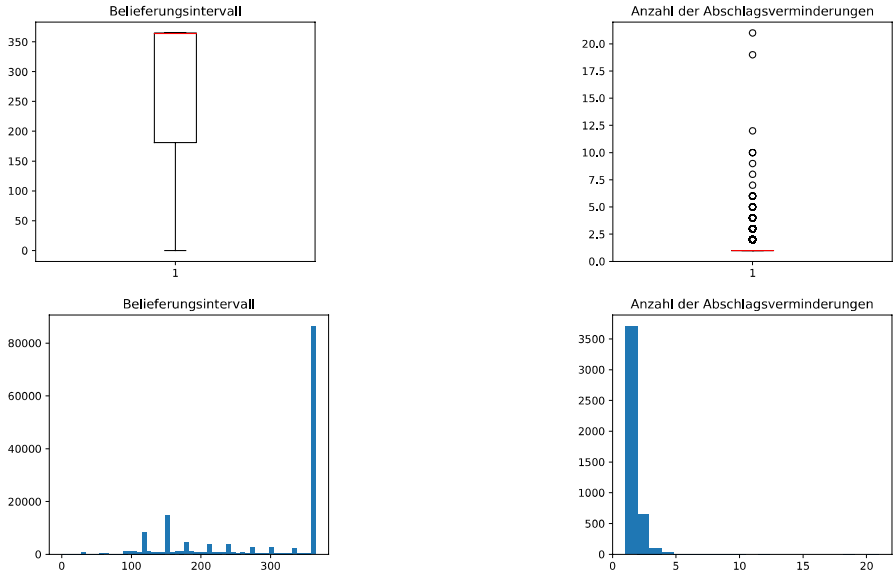
h) Anzahl der eingestellten Rückrufwünsche

Abb. 5: Boxplots (mit Ausreißern) und Histogramme (ohne Ausreißer) ausgewählter Datenfelder

Eine Belieferung dauerte im Mittel ca. 284 Tage an, wobei der Median bei 365 Tagen liegt (*supplyinterval_in_days*, Abbildung 6a). Der Median von 365 Tagen erklärt sich dadurch, dass die meisten Bonuszahlungen erst mit Vollendung des ersten Belieferungsjahres ausgezahlt werden oder die Mindestvertragslaufzeit 12 Monate beträgt. Der Minimalwert mit 0 fällt direkt auf. Das liegt an der Abgrenzung des Lieferbeginns und des Lieferendes. Die Abgrenzung geschieht im Quellsystem von 0:00 Uhr des Lieferbeginntages bis 23:59 Uhr des Lieferendetages. Sollte der Kunde nur einen Tag in Belieferung gewesen sein, dann kann der Wert 0 annehmen, da keine vollen 24 Stunden vorliegen. Dies ist bei der späteren Datenvorbereitung zu beachten.

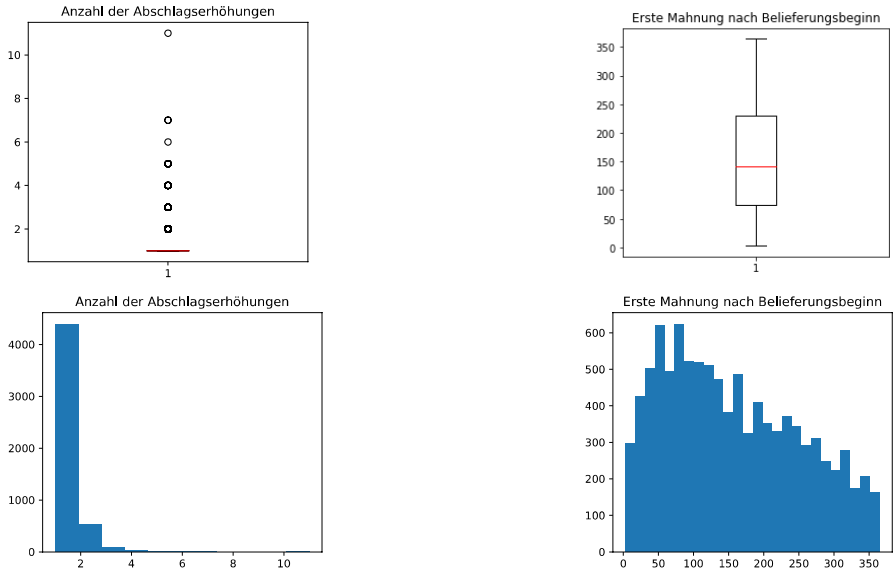
Die Anzahl, bei wie vielen Verträgen der Abschlagsplan verringert wurde, liegt bei 4.509 (*payplanreducedcount*, Abbildung 6b). Im Vergleich liegt die Anzahl, bei wie vielen Verträgen der Abschlagsplan erhöht wurde, bei 5.044 (*payplanincreasedcount*, Abbildung 6c). Die Häufigkeit der Verminderungen je Kunden ist im Vergleich zu den Erhöhungen des Abschlagsplans je Kunden leicht erhöht (Verminderungen ca. 1,24 und Erhöhungen ca. 1,16). Auffällig sind zudem die Ausreißer, die bei den Abschlagsverminderungen deutlich stärker ausgeprägt sind. Demnach werden Abschläge ca. 11,87 % öfters erhöht und nur selten mehrmals für denselben Vertrag erhöht. Hingegen werden Abschläge übergreifend seltener verringert, jedoch sind es vergleichsweise mehr Verringerungen je Vertrag.

Im Mittel wurde die erste Mahnung am 155. Tag in Belieferung versandt (*first_dunning_in_days_after_supplybegin*, Abbildung 6d), mit einer Standardabweichung von ca. 95,38. Die Spanne des 25 %-Quantil bis zum 75 %-Quantil reicht von 75 bis zu 230 Tagen. Der Minimalwert mit 3 Tagen in Belieferung zur ersten Mahnung ist besonders gering. Bei diesem Fall könnte der Lieferbeginn kurz vor Monatsende bestätigt, der Abschlagsplan für den Ersten des Folgemonats aktiviert und die Buchung durch PayPal als direktes Zahlungsmittel abgewiesen worden sein.



a) Belieferungszeitraum in Tagen

b) Anzahl der Abschlagsverminderungen



c) Anzahl der Abschlagserhöhungen

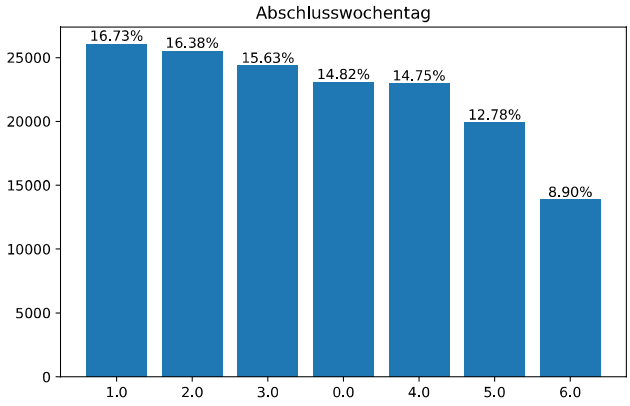
d) Erste Mahnung nach Belieferungsbeginn

Abb. 6: Boxplots (mit Ausreißern) und Histogramme (ohne Ausreißer) ausgewählter Datenfelder (fortgesetzt)

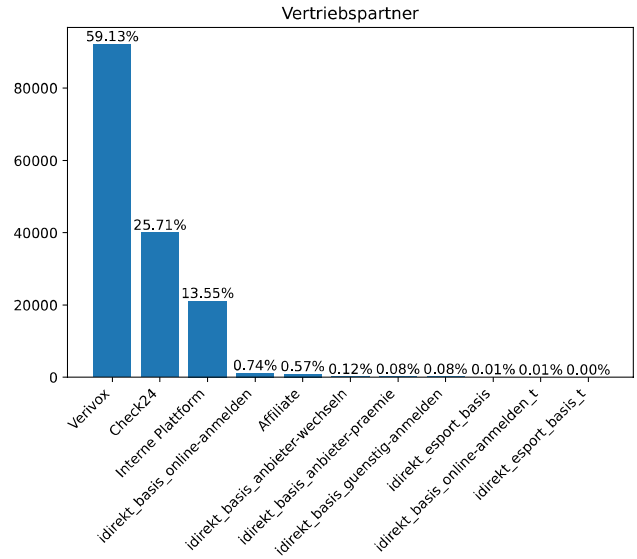
Die Darstellungen der kategorialen Datenfelder sind in Abbildung 7 sowie im Anhang 5 enthalten. Unter den Abschlusswochentagen (*conclusionweekday*, Abbildung 7a) ist der Montag mit 16,73 % am häufigsten vertreten. Dicht dahinter folgen die Wochentage Dienstag (16,38 %), Mittwoch (15,63 %) und Sonntag (14,82 %). Die weniger frequentierten Tage für Vertragsabschlüsse sind der Freitag (12,78 %) und der Samstag (8,9 %).

Beim Vertriebspartner (*distributionpartner*, Abbildung 7b) ist das Vergleichsportal Verivox mit 59,13 % deutlich stärker vertreten als das Vergleichsportal Check24 mit 25,71 %. Über die interne Plattform bzw. die eigene Bestellstrecke wurden lediglich 13,55 % der Kunden akquiriert. Mit weniger als 1 % sind verschiedene Affiliate-Maßnahmen vertreten. Die Ausprägungsunterschiede der verschiedenen Anteile können aufgrund unterschiedlicher Marketingmaßnahmen je Vertriebspartner/-kanal entstanden sein. Bei dieser Grafik wird deutlich, dass mehr als 84 % aller Verträge über ein Vergleichsportal akquiriert wurden. Bei 84 % aller Verträge fielen somit zusätzliche Kosten für die Provisionszahlung an die Vergleichsportale an.

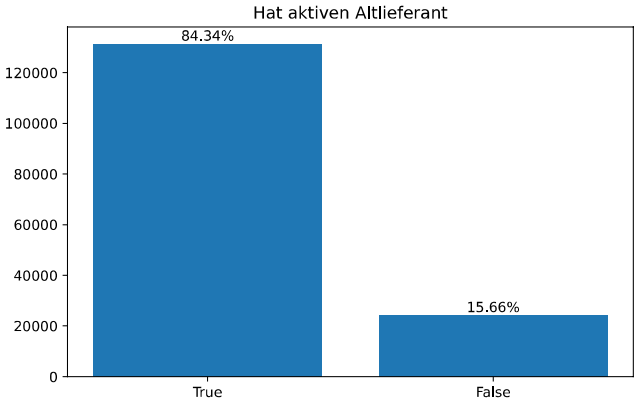
Durch die regulierten Marktprozesse sind Lieferantenwechsel für Kunden im Energiemarkt besonders komfortabel. Dies spiegelt sich auch im Kennzeichen *Hat aktiven Altlieferant* wider (*hasoldsupplier*, Abbildung 7c). Über 84 % aller Kunden geben einen Auftrag zur Belieferung mit der Vollmacht ab, den bestehenden Altlieferanten zu kündigen. Ob die 84 % der Kunden mit einem aktiven Vorlieferanten in Korrelation mit den 84 % der Verträge über Vergleichsportale stehen, beantwortet Abbildung 8. Sehr deutlich wird, dass ca. 84 % aller Kunden, die über ein Vergleichsportal abschließen, auch einen aktiven Altlieferanten haben.



a) Abschlusswochentag (0=Sonntag, …, 6=Samstag)



b) Vertriebspartner



c) Hat aktiven Altlieferanten

Abb. 7: Säulendiagramme mit prozentualen Anteilen ausgewählter Datenfelder

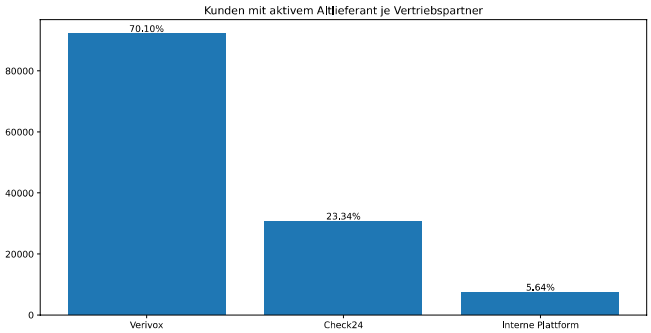


Abb. 8: Verträge mit aktivem Altlieferant je Vertriebspartner

Im Vergleich der Verträge mit und ohne Zahlungsausfälle nach Vertriebspartner ist eine weitere Auffälligkeit vorhanden. In Abbildung 9 sind die Vertragsanteile je Vertriebspartner aller Verträge und der Verträge mit einem Zahlungsausfall dargestellt. Im direkten Vergleich ist Verivox mit den Anteilen führend, während der Anteil der Verträge mit Zahlungsausfällen um ca. 14 % niedriger ist. Anders sieht es bei dem Vertriebspartner Check24 oder der eigenen Bestellstrecke aus. Die Gesamtvertragsanteile fallen prozentual geringer aus als die Anteile der Verträge mit Zahlungsausfällen. Bei Check24 sind es prozentual ca. 9 % und bei der eigenen Bestellstrecke rund 3,5 % mehr. Daraus erschließt sich, dass Kunden, die über Verivox bestellen, tendenziell weniger Zahlungsausfälle verursachen als Kunden, die über Check24 oder der eigenen Bestellstrecke akquiriert werden.

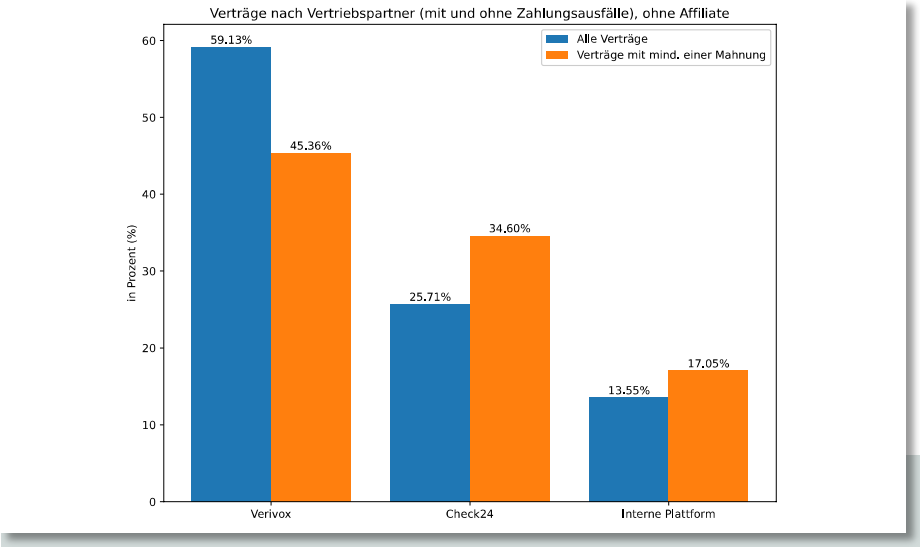


Abb. 9: Verträge nach Vertriebspartner mit und ohne Zahlungsausfälle, ohne Affiliate-Kunden

Die Verteilung der höchsten erreichten Mahnstatus ist in Abbildung 10 dargestellt. Über 93 % aller Verträge haben keine negativen Zahlungsauffälligkeiten. Rund 2,84 % aller Verträge haben die 1. Mahnstufe erreicht und wurden einmal oder mehrmals gemahnt. Die 3. Mahnstufe mit einer Kündigungsandrohung des laufenden Vertrags wurde von 1,32 % aller Verträge erreicht. Hiernach folgte die *Abgabe an das zuständige Inkassounternehmen. Im Inkassoprozess befanden sich nur noch 0,34 % aller Verträge. Das Verhältnis von Zahlungsrückständen mit Zählersperrungen als letztes Mittel (Kapitel 2.3) zu Gesamtvertragsabschlüssen (Kapitel 2.1) zeigt weitaus höhere prozentuale Anteile mit 4 % in der Sparte Strom und 3,6 % in der Sparte Gas. Der Anteil der Inkassoübergaben im vorliegenden Datensatz macht nur ein Zehntel der Anteile vom Gesamtmarkt aus. Der Anteil der Inkassoübergaben ist somit verhältnismäßig gering. Gründe könnten u. a. der fehlende Grundversorgerstatus sein. Generell ist der Datensatz besonders unausgeglichen. Die Klasse *Keine Zahlungsprobleme* ist besonders dominant. Für eine zukünftige Modellerstellung ist diese Unausgewogenheit zu beachten und es sind ggf. Maßnahmen zur Herstellung einer Balance zu unternehmen.

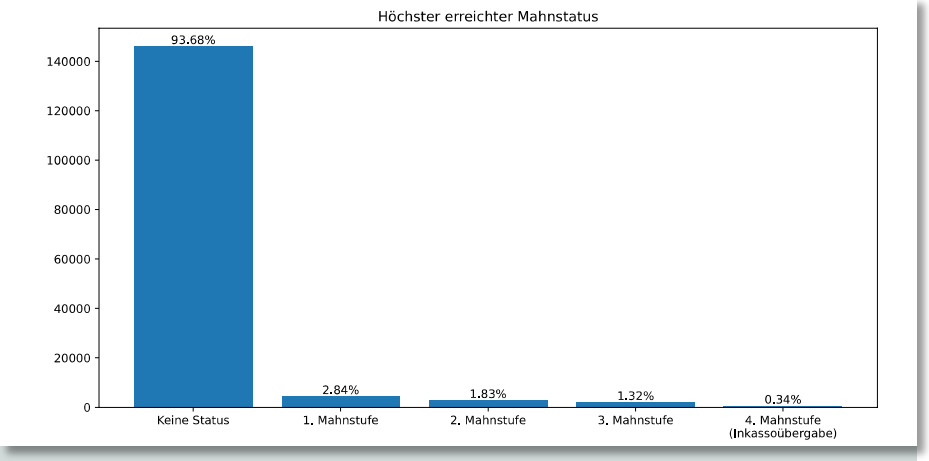


Abb. 10: Prozentuale Anteile der höchsten erreichten Mahnstatus

Bei der Zahlung mit Lastschrift ist durch die Angabe der IBAN immer die Bank des Kunden bekannt. Eine Auswertung der am meist verwendeten BICs sowie die BICs der Verträge, mit den meisten Zahlungsausfällen, befinden sich in den Abbildungen 11a und 11b. Dabei wurden nur die ersten fünf Stellen der BIC verwendet. Die letzten Stellen (Orts- und Abteilungsbezeichnung) wurden abgeschnitten, damit Banken auf Bundesebene verglichen werden können.

Mit 21,6 % ist die BBBank besonders stark vertreten. Darauf folgen die Bayrische Landesbank (9,7 %), die Commerzbank (9,1 %), die Sparkassen (8,1 %), ING-DiBa (7,7 %) und Postbank (6,7 %). Alle anderen Banken sind mit <=4 % vertreten.

Besonders auffallend sind die Banken, die in Verträgen mit Zahlungsausfällen genutzt wurden (Abbildung 11b). Die BBBank führt erneut und hat einen Anteil von 15 %. Danach folgen die Sparkassen (11,8 %), die Postbank (10,5 %), die Commerzbank (8,2 %), die Norddeutsche Landesbank (6,4 %) und die Deutsche Bank (4,2 %). Alle weiteren Banken haben einen Anteil <4 %. Auffällig ist, dass einige Banken, die unter allen Verträgen nicht führend waren, bei den Verträgen mit Zahlungsausfällen führend sind.

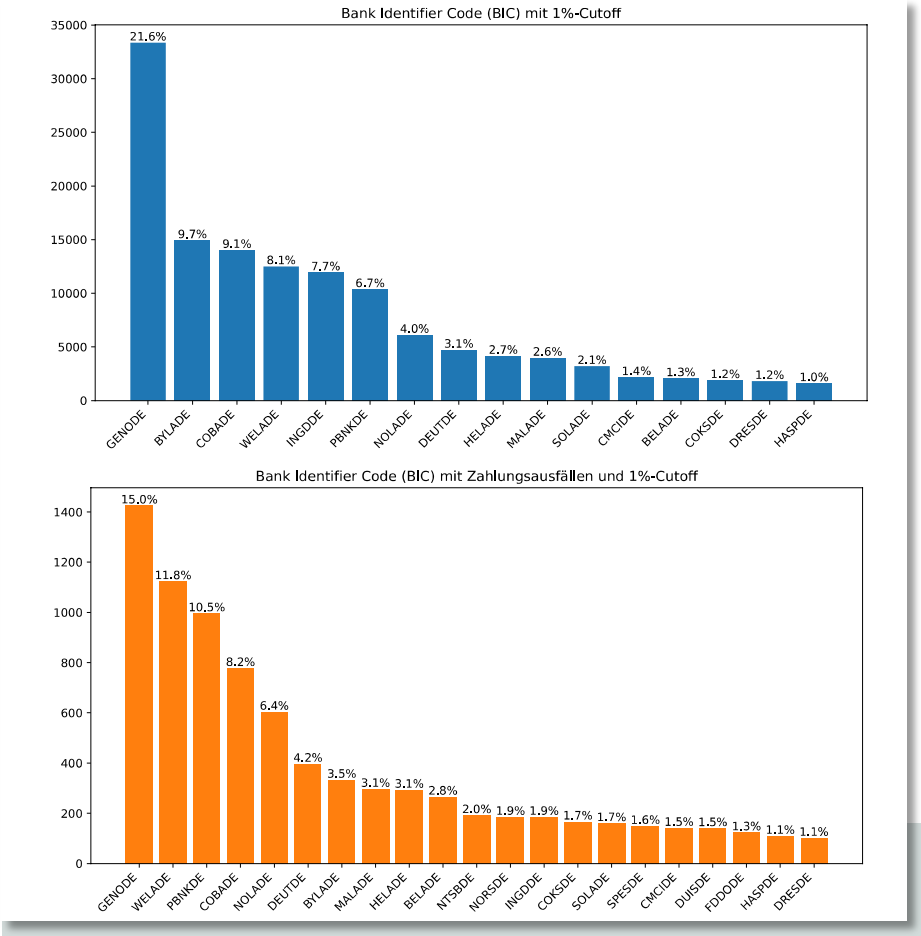


Abb. 11: Säulendiagramme prozentualer Anteile des Bank Identifier Code (BIC) mit 1 %-Cutoff

Eine weitere Erkenntnis bietet die Sicht auf die Vertragsanteile nach Alter aller Verträge und der Verträge mit Zahlungsausfällen. Diese sind im Histogramm in Abbildung 12 übereinandergelegt. Beide Histogramme sind rechtsschief. Verträge sind – wie zuvor genannt – mit einem Kundenalter zwischen 25 und 40 Jahren sowie zwischen 50 und 60 Jahren am stärksten vertreten. Besonders fallen hier jedoch die Verträge mit Zahlungsproblem der Kunden auf, die zwischen 20 und 45 Jahre alt sind. Hier übersteigt der Vertragsanteil teils das doppelte der Vertragsverteilung aller Verträge. Umgekehrt ist es ab einem Alter von 50 Jahren: Die Vertragsanteile mit Zahlungsausfällen macht teils weniger als die Hälfte der Gesamtvertragsanteile

aus. Daraus folgt, dass junge Kunden oder Kunden mittleren Alters sehr viel häufiger Zahlungsausfälle haben, als Kunden ab 50 Jahren. Je älter die Kunden werden, desto geringer wird die Wahrscheinlichkeit eines Zahlungsausfalls.

Bei einem Kundenalter von 120 Jahren ist eine Besonderheit vorhanden. Auf der Welt lebten bisher nur wenige Menschen, die ein solches Alter erreicht haben. Gerade bei diesen Kunden sind die Gesamtvertragsanteile sehr gering, jedoch sind dafür die Anteile der Verträge mit Zahlungsausfällen besonders hoch. Die Auswahl des Geburtsdatums wird bei einem Vertragsabschluss nicht verifiziert. Daher wird angenommen, dass diese Kunden absichtlich, gegebenenfalls sogar spaßeshalber, ein extrem hohes Kundenalter angegeben haben. Für eine Modellierung bietet sich die gesonderte Betrachtung der Eigenschaft an.

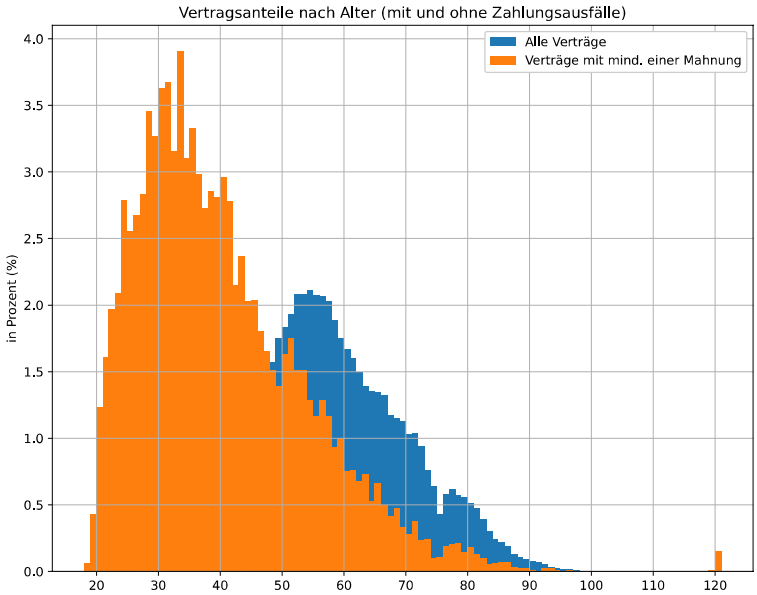


Abb. 12: Vertragsanteile nach Alter (mit und ohne Zahlungsausfälle)

Mithilfe der Korrelationsmatrix numerischer und boolescher Werte können erste Beziehungen zwischen zwei Variablen erkannt werden. Die Korrelationsmatrix ist in Abbildung 13 dargestellt. Zusammenfassend sind die folgenden Beziehungen auffällig:

- Hat aktiven Altlieferant und Hat Telefonnummer: Wenn der Kunde einen Altlieferanten hat, dann gibt er weniger wahrscheinlich eine Telefonnummer

bei Auftragsabgabe an. Dies könnte mit einem schnellen Vertragsabschluss über ein Vergleichsportal zusammenhängen.

- Verbrauchsprognose und Bonuswert: Je höher der Verbrauch (und demnach auch die Marge des Lieferanten) ist, desto höher fällt auch die Bonuszahlung aus.
- Erste Mahnung nach Belieferungsbeginn und Bonuswert: Kunden, die eine Mahnung später erhalten, haben eine geringere Bonuszahlung.
- Belieferungsintervall und Bonuswert: Hier deutet sich eine Scheinkorrelation an, da der Bonuswert mit Vertragsschluss festgelegt wird. Mit einer steigenden Tagesanzahl in Belieferung sinkt nicht gleichzeitig der Bonuswert für das erste Belieferungsjahr.
- Belieferungsintervall und Höchster erreichter Mahnstatus: Mit einer höheren Tagesanzahl in Belieferung sinkt das Risiko, dass ein Kunde einen sehr hohen Mahnstatus (z. B. den Inkassoprozess) erreicht.
- Ist aktiv und Höchster erreichter Mahnstatus: Sollte der Vertrag nicht aktiv sein, so erhöht sich der Wert für den höchsten erreichten Mahnstatus. Verträge werden vor der Inkassoübergabe, ergo dem höchsten erreichten Mahnstatus, gekündigt. Aus diesem Grund ist eine Korrelation ersichtlich, dass inaktive bzw. gekündigte Verträge mit dem Mahnstatus negativ korrelieren.
- Anzahl der Rücklastschriften und Höchster erreichter Mahnstatus: Eine leichte Korrelation ist zu erkennen, dass bei einem Anstieg der Anzahl der Rücklastschriften ebenfalls der höchste erreichte Mahnstatus steigt.
- Kündigungseingang nach Belieferungsbeginn und Erste Mahnung nach Belieferungsbeginn: Bei einer frühen Kündigung, nachdem die Belieferung begonnen hat, ist eine Korrelation mit einer ebenso frühen Mahnung nach dem Belieferungsbeginn vorhanden.
- Kündigungseingang nach Belieferungsbeginn und Höchster erreichter Mahnstatus: Die Kündigung nach Belieferungsbeginn korreliert negativ mit dem höchsten erreichten Mahnstatus. Eine frühe Kündigung kann in einem höheren Mahnstatus resultieren.

- Kündigungseingang nach Belieferungsbeginn und Belieferungsintervall: Da bei den meisten Verträgen eine Mindestvertragslaufzeit von 12 Monaten besteht, das Belieferungsintervall nur eine Momentaufnahme beim Datenabzug ist und das Belieferungsintervall kontinuierlich steigt, kann keine direkte Erkenntnis gewonnen werden.
- Anzahl der Rücklastschriften und Anzahl manueller Überweisungen: Bei steigender Anzahl von Rücklastschriften steigt ebenso die Anzahl von manuellen Überweisungen.

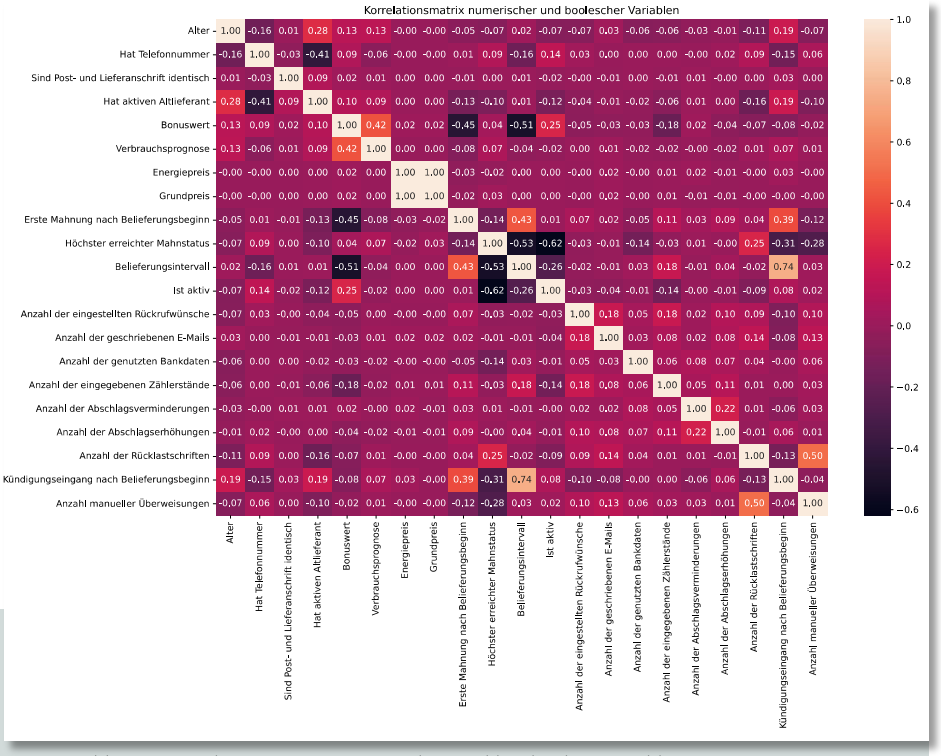
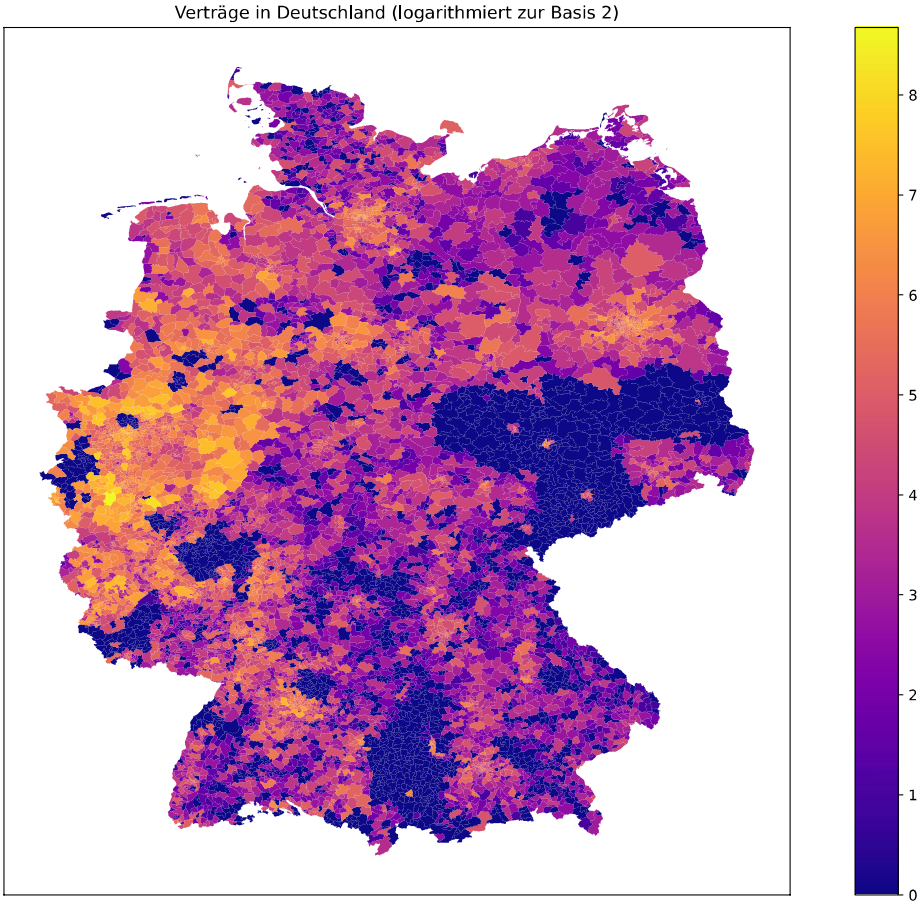


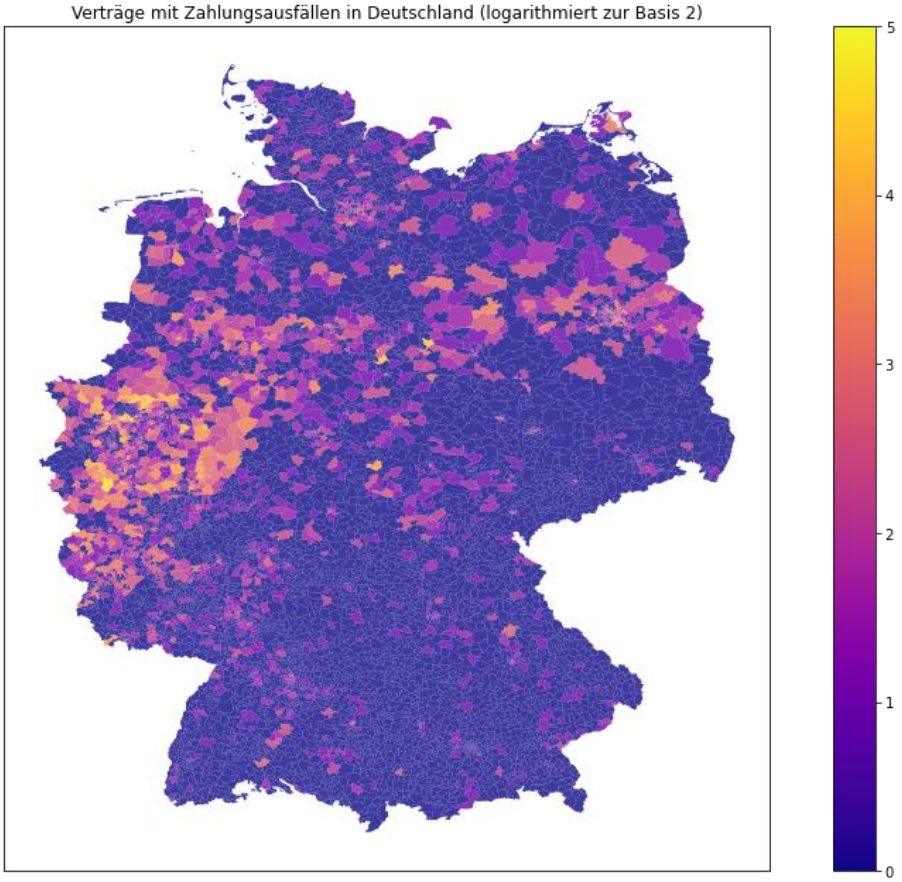
Abb. 13: Korrelationsmatrix numerischer und boolescher Variablen

In Abbildung 14 sind zwei separate Deutschlandkarten abgebildet. Abbildung 14a stellt die logarithmierte Anzahl aller Verträge dar, während Abbildung 14b die logarithmierte Anzahl aller Verträge mit Zahlungsausfällen innehat. Die Verträge wurden auf Postleitzahlen-Ebene zusammengefasst. Zuerst fällt auf, dass weite Teile der Bundesländer Thüringen und Sachsen nicht vom Energielieferanten beliefert wurden. Generell sind die Verträge jeweils in Westen, Norden und Süden gleichmäßig verteilt. Im Westen von Deutschland sind besonders viele Kunden, die

Verträge beim Energielieferanten haben. Im Vergleich zur Karte, auf der die Verträge mit Zahlungsausfällen gezeigt werden (Abbildung 14b), stechen Nordrhein-Westfalen und einige Orte in der Mitte Deutschlands sowie das Umland von Berlin heraus. Im Süden und Norden Deutschlands sind Kunden mit Zahlungsausfällen nicht so stark repräsentiert. Auf eine genauere Betrachtung der Orte und Abwägung, welche Orte die höchsten Zahlungsausfallraten haben, wird verzichtet.



a) Alle Verträge (logarithmiert)



b) Verträge mit Zahlungsproblemen (logarithmiert)

Abb. 14: Darstellung aller Verträge und der Verträge mit Zahlungsausfällen in zwei separaten Deutschlandkarten

5 Datenvorbereitung

Durch die Datenvorbereitung werden die diversen Datenfelder, -typen und -arten aufbereitet, sodass sie für Modelle verarbeitbar sind. Zudem muss sichergestellt werden, dass die Daten fachlich korrekt sind, damit ein Modelltraining stattfindet, welches auf realitätsnahen Daten basiert. Die Datenvorbereitung stellt somit einen essenziellen Schritt für das Machine Learning dar. Hierbei werden insbesondere die Erkenntnisse aus Kapitel 4.2 genutzt, um Anpassungen vorzunehmen.

5.1 Behandlung von Ausreißern und fehlenden Daten

Im ersten Schritt werden die Ausreißer entfernt. Als Ausreißer gelten hier extreme Datenpunkte, die fachlich und technisch offensichtlich auf einen Fehler zurückzuführen sind. Konkret handelt es sich um die Variablen *Alter*, *Grundpreis* und *Energiepreis*. In der vorangegangenen Analyse hat sich herausgestellt, dass ein Kunde weitaus älter als 1000 Jahre ist. Ebenso wurden Energiepreise und Grundpreise in Höhe von 100.000.000 Cent bzw. Euro identifiziert. Die Daten sind sehr wahrscheinlich durch manuelle Bearbeitungen entstanden. Der Ursprung der Daten lässt sich nicht wiederherstellen. Ebenfalls fehlen bei einigen Verträgen die Daten der Felder *Anrede*, *Grundpreis* und *Energiepreis*. Generell sind verschiedene Möglichkeiten vorhanden, um mit fehlenden Daten (Missing Data) umzugehen. Larose beschreibt zusammengefasst die folgenden vier Punkte:[155]

- Fehlende Daten mit einem konstanten Wert ersetzen, entweder frei definiert oder basierend auf einer Kennzahl wie dem Median oder dem Mittelwert aller Beobachtungen.
- Fehlende Daten zufällig aus der Menge aller Beobachtungen ziehen.
- Imputation fehlender Werte, indem der Wert in Abhängigkeit zu anderen Eigenschaften der Beobachtungen hergeleitet wird.
- Entfernen der Beobachtungen, unter Beachtung, dass bei 30 Variablen und einem Anteil von 5 % fehlender Werte über 80 % der Beobachtungen betroffen sind.

Nun handelt es sich bei den beschriebenen Variablen *Alter*, *Grundpreis* und *Energiepreis* um metrische Werte und bei der Variable *Anrede* um nominale Werte. Nicht alle zuvor genannten Möglichkeiten sind auf alle Variablen gleichermaßen

[155] Vgl. *Larose, D. T.*, 2015, S. 23.

anwendbar wie die Imputation der nominal-skalierten Variable *Anrede*. Zudem ist bekannt, dass es sich teilweise um manuelle Fehler handelt. Aus diesem Grund werden die fehlenden oder fehlerbehafteten metrischen Werte mit dem Durchschnitt der entsprechenden Variable ersetzt. Bei der Variable *Anrede* wird sich für eine Entfernung der Datensätze entschieden, da eine Annäherung des Geschlechts durch einen Mittelwert nicht möglich ist und nur ein zufälliges Ziehen aus der Gesamtmenge möglich wäre.

Nach der Durchführung dieser Schritte wurden von den initial 155.926 Vertragsdaten insgesamt 36 Einträge entfernt, womit die Anzahl der Vertragsdaten auf 155.890 verringert wurde.

Einige Variablen wie der *Bonuswert* oder *Anzahl der geschriebenen E-Mails* enthalten null-Werte. Dies resultiert aus einem Vertrag ohne Bonus oder einem Vertrag, zu dem nie eine E-Mail eingegangen ist. Da ein null-Wert nicht vergleichbar ist, werden die Werte einheitlich mit dem Wert 0 überschrieben. Dies gilt für die folgenden Variablen:

- Erste Mahnung nach Belieferungsbeginn
- Höchster erreichter Mahnstatus
- Anzahl der eingestellten Rückrufwünsche
- Anzahl der geschriebenen E-Mails
- Anzahl der genutzten Bankdaten
- Anzahl der Abschlagsverminderungen
- Anzahl der Abschlagserhöhungen
- Kündigungseingang nach Belieferungsbeginn

5.2 Anpassung diverser Variablen

Belieferungszeitraum

Der Belieferungszeitraum in Tagen hat in einigen Fällen den Wert 0 angenommen. Dieser Wert ist durch die Extrahierung aus der Datenquelle entstanden. In der Datenquelle wird der Lieferbeginn immer am ersten Belieferungstag mit der Uhrzeit 0:00 Uhr versehen. Das Lieferende wird am letzten Belieferungstag mit der Uhrzeit 23:59 Uhr abgegrenzt. Bei einer Ein-Tages-Belieferung vom 01.01.2021 0:00 Uhr bis zum 01.01.2021 23:59 Uhr ist kein voller Tag mit 24

Stunden erreicht, ergo nimmt der Belieferungszeitraum den Wert 0 an. Ebenso wird jeder andere Belieferungszeitraum mit einem Tag zu wenig angegeben. Alle Werte der Variable *Belieferungszeitraum* werden deshalb um den Wert 1 erhöht.

E-Mail-Domain

Einige Kunden nutzen E-Mail-Adressen mit einer eigens registrierten Domain im Format wie `Vorname@Nachname.de`. Zusätzlich sind Domains von E-Mail-Adressen vorhanden, die nur sehr selten vertreten sind. Hierbei ist eine Gefahr für falsche Klassifizierungen durch ein Modell gegeben. Wenn eine Domain nur selten genutzt wird und die Verträge zudem die gleiche Zahlungsausfallausprägung haben, dann wird diese Domain automatisch mit dieser Zahlungsausfall-Schwere assoziiert. Dies sollte jedoch nur geschehen, wenn auch ausreichend Verträge mit derselben Domain vorhanden sind, sodass eine repräsentativere Entscheidung erfolgen kann. Aus diesem Grund werden E-Mail-Domains, die weniger als 20-mal im vorliegenden Datensatz vorkommen, durch die künstliche Domain `custom.overwrite.de` überschrieben.

Bank Identifier Code (BIC)

Wie zuvor in Kapitel 4.1 beschrieben, besteht die BIC nach ISO-3166-1 aus einer alphanumerischen Zeichenfolge. Die Zeichenfolge gibt Aufschluss über das Geldinstitut, das Land sowie optionale Abteilungs- und Ortsangaben. Die Geldinstitute können durch die optionalen Abteilungs- und Ortsangaben sehr divers ausgeprägt sein. Um eine Vereinheitlichung auf Ebene des Bundes und des Geldinstituts herzustellen, werden von der alphanumerischen Zeichenfolge nur die ersten 6 Zeichen beibehalten (wie auch bei der Visualisierung in Kapitel 4.2 durchgeführt). Die restlichen Zeichen werden verworfen. Damit bleibt bei der Variable *BIC* nur das Geldinstitut zusammen mit dem Ländercode übrig wie z. B. `COBADE` für die Commerzbank AG aus Deutschland.

Zusätzlich zahlen wenige Kunden nicht via Lastschrift, sondern eigenständig mittels Überweisung oder via PayPal. Bei diesen Kunden werden der Variable *BIC* die konstanten Werte `UEBERWEISER` bzw. `PAYPAL` entsprechend zugeordnet. Ein kleiner Seiteneffekt ist vorhanden, da nun die beiden Ausprägungen der Variable *Zahlungsmethode* (`SEPA` und `PAYPAL`) durch einen Wert in der Variable *BIC* abgebildet werden. Die Variable *Zahlungsmethode* wird daher gelöscht.

BICs, die weniger als 20-mal genutzt wurden, werden mit dem konstanten Wert `SONSTIGES` überschrieben, um eine repräsentativere Entscheidung durch ein Modell zu vollziehen (analog zur Variable E-Mail-Adresse).

Vertriebspartner

Bei einigen Verträgen ist durch einen manuellen Vertragsaufbau der Vertriebspartner nicht gepflegt. Dies kann unter anderem vorkommen, wenn der Vertrag durch einen Sachbearbeiter wegen Kundenanliegen oder marktspezifischer Gründe neu aufgebaut werden musste. Da der Vertrag über einen internen Kanal erstellt wurde, wird hier der Wert der Bestellstrecke `Interne Plattform` gesetzt.

Verschiedene Affiliate-Kampagnen, die im vorliegenden Datensatz nur sehr gering repräsentiert sind, werden ganzheitlich mit dem Wert `Affiliate` überschrieben. Nach diesen Maßnahmen sind nur noch vier Vertriebspartner vorhanden: Interne Plattform, Check24, Verivox und Affiliate.

Abschlusswochentag

Der Abschlusswochentag wird durch eine numerische Variable dargestellt. Der Wert 0 stellt den Sonntag dar. Der Wert 6 stellt den Samstag dar. Die Variable *Abschlusswochentag* ist nominal skaliert, obwohl der Wert numerisch ist. Die Wochentage sind zwar geordnet, jedoch kann keine Wertung stattfinden, dass der Samstag *höhergestellt* ist als der Sonntag. Mit Blick auf immer wiederkehrende Wochentage fehlt demnach eine Rangfolge.

Für eine bessere Zuordenbarkeit werden die Werte um Eins erhöht. Der Sonntag trägt den Wert 1 und der Samstag den Wert 7. Zusätzlich wird die Variable in den Datentypen `string` konvertiert, sodass kein numerischer Vergleich mehr stattfinden kann.

5.3 Umwandlung kategorialer Daten

Wie beim Abschlusswochentag sind auch andere Variablen kategorialer Natur und können nicht direkt miteinander verglichen werden. Vielmehr sind die Variablen

- Sparte
- Anrede
- Abschlusswochentag
- E-Mail-Domain
- BIC
- Vertriebspartner
- Postleitzahl der Lieferstelle
- Postleitzahl des Vertragsnehmers

nominal skaliert und teilweise nicht durch numerische Werte repräsentiert. Zu trainierende Modelle benötigen jedoch numerische Werte. Daher werden die Daten in *flag variables* bzw. Dummy-Variablen umgewandelt. Die Variable *Sparte* mit den Ausprägungen Strom und Gas wird in zwei Variablen `Sparte_Strom` und `Sparte_Gas` aufgespalten. Die Sparte kann mithilfe der neuen Variablen mit booleschen oder numerischen Werten (True/- False oder 0/1) spezifiziert werden. Zudem sind eigene Gewichtungen der Variablen mit eigenem Schätzer möglich.[156]

Die Variable *Abrechnungsintervall* ist in natura mit den Werten 1 (monatliche Abrechnung) und 12 (jährliche Abrechnung) versehen. Die Variable wird in eine boolesche Variable umgewandelt und in `isAnuallyBilled` (wird jährlich abgerechnet) umbenannt. Durch die Angabe des Werts True wird die jährliche Abrechnung und mit Angabe des Werts `False` die monatliche Abrechnung ausgedrückt.

In der Analyse der Variable *Alter* sind die Verträge mit einem Kundenalter um ca. 120 Jahre durch ihren besonders starken Anteil der Verträge mit Zahlungsausfall aufgefallen. Aus diesem Grund wird eine neue Variable `ageOver110` eingefügt und auf `True` gesetzt, sobald das Alter des Kunden größer als 110 Jahre ist.

Nach der Hinzufügung der zusätzlichen Variablen ist die Größe des Datensatzes von 31 Variablen auf 13.320 Variablen angestiegen. Dies resultiert insbesondere aus der Hinzufügung der Dummy-Variablen für die Variablen *BIC*, *E-Mail-Domain*, *Postleitzahl der Lieferstelle* und *Postleitzahl des Vertragsnehmers*. Unter Umständen könnte es für ein Modelltraining von Vorteil sein, die Postleitzahl-Variablen zu entfernen, da bereits in der visuellen Analyse keine großen Unterschiede festgestellt wurden und da daraus ein kleinerer Datensatz entsteht, welches dem Modelltraining zugutekommen kann.

5.4 Umwandlung der abhängigen Variable Zahlungsausfall

Die Variable *Höchster erreichter Mahnstatus* gibt die höchste jemals erreichte Stufe im Mahnwesen an. Dabei ist bei der Betrachtung zum Zeitpunkt der Datenextraktion nicht ersichtlich, ob ein derzeitiger Teilausfall nicht auch in einem Vollausfall mündet. Daher wird ein Zahlungsausfall als Vollausfall betrachtet. Für eine

[156] Vgl. *Larose, D. T.*, 2015, S. 39 ff.

binäre Vorhersage zu einem Vertrag mit den möglichen Werten `Zahlungsausfall` oder `Kein Zahlungsausfall` wird eine neue Spalte `binary-debtstatus` hinzugefügt. Sobald ein Vertrag einmal gemahnt wurde, ist der Wert `True`. Diese Modelle sind anfällig für Alpha- und Beta-Fehler und lassen sich diesbezüglich bzw. mit der Konfusionsmatrix miteinander vergleichen. Ein trennschärferes Modell mit drei Ausprägungen wird nicht betrachtet.[157]

Nach dieser finalen Datensatzanpassung sind 155.882 Vertragsdaten mit insgesamt 13.321 Variablen im Datensatz enthalten. Davon stellt eine Variable die abhängige Variable dar. Verträge mit Zahlungsausfällen sind mit einem Anteil von ca. 6,38 % enthalten.

5.5 Gruppierung und Aufteilung der Datensätze

Der Datensatz wird für die nachfolgende Modellierung in drei verschiedene Sub-Datensätze unterteilt. Die fachliche Gruppierung orientiert sich an der fachlichen Art und Herkunft, die in Kapitel 4.1 beschrieben ist. Die Sub-Datensätze umfassen nur wenige bis alle Variablen und sind wie folgt fachlich zu beschreiben:

1. **Angebotsdaten mit Postleitzahlen**. Der Datensatz enthält alle Daten, die durch die initiale Angebotsabgabe über den Kunden vorhanden sind, inklusive der abhängigen Variable. *Variablen insgesamt: 13.309*
2. **Angebotsdaten ohne Postleitzahlen**. Der Datensatz enthält alle Daten, die durch die initiale Angebotsabgabe über den Kunden vorhanden sind, inklusive der abhängigen Variable und abzüglich der Dummy-Variablen, die die Postleitzahl der Lieferstelle oder des Vertragsinhabers inne haben. *Variablen insgesamt: 253*
3. **Verhaltensdaten**. Der Datensatz enthält alle Daten, die während der aktiven Belieferung bis 365 Tage nach Belieferungsbeginn erfasst wurden. Hinzu kommt die abhängige Variable. *Variablen insgesamt: 13*

Es findet eine Aufteilung in Trainings- und Testdatensätze statt, der *Train-Test-Split*. Der Train-Test-Split wird mit dem Verhältnis 70:30 gewählt. Der Trainingsdatensatz beinhaltet 70 % und der Testdatensatz beinhaltet 30 % aller Daten.

[157] Vgl. *Krämer, W.*, 2002, S. 2 f.

Hierbei wurde insbesondere auf das Verhältnis Zahlungsausfall-zu-Kein Zahlungsausfall geachtet. Alle Trainings- und Testdatensätze aller Gruppen beinhalten 6,31 % Verträge mit Zahlungsausfällen.

Die sechs verschiedenen Datensätze (3 Gruppen mit je einem Trainings- und Testdatensatz) wurden in Comma-separated values (CSV)-Dateien gespeichert. Für die Modellierung mit R wurden die Datensätze ohne Dummy-Variablen gespeichert, da in R ein entsprechender Datentyp (`factor`) für die Handhabung mit kategorialen Daten vorhanden ist.

6 Modellierung

6.1 Logistische Regression

Auf Basis der vorbereiteten Datensätze findet zuerst die Modellierung der logistischen Regressionen statt. Hierzu wird als Erstes der Datensatz Angebotsdaten mit Postleitzahlen genutzt. Nachdem die richtigen Datentypen aller Variablen zugeordnet und verifiziert wurden, wurde die erste logistische Regression erstellt. Als abhängige Variable wurde die zuvor erstellte Variable `binarydebtstatus` gesetzt. Als erklärende Variablen wurden alle anderen im Datensatz vorhandenen Variablen herangezogen. Zur Erstellung wurde vorerst die Funktion `glm` genutzt. Da die Funktion sehr lange Laufzeiten aufwies, wurde die Funktion `logistf` als Alternative genutzt. Dabei wurde ein Fehler erkannt, der auf einen unzureichenden Arbeitsspeicher hindeutet. Der Datensatz mit den Dummy-Variablen der Postleitzahlen ist zu groß, weshalb eine Verarbeitung in einer logistischen Regression in der vorhandenen Systemumgebung nicht möglich ist. Die weitere Modellierung der logistischen Regressionen mit den Postleitzahl-Variablen wird unterbrochen. Es werden nur noch die Datensätze ohne Postleitzahl-Variablen zur Modellierung genutzt.

Erneut wird eine logistische Regression analog zum ersten Versuch erstellt, jedoch unter der Nutzung der Angebotsdaten ohne Postleitzahlen. Die Erstellung ist erfolgreich, was indirekt die Problematik mit den Postleitzahl-Variablen bestätigt. Zuvor wurde in Kapitel 4 herausgestellt, dass die Verteilung der beiden Klassen *Zahlungsausfall* und *Kein Zahlungsausfall* unausgeglichen ist. Bei unausgeglichenen Datensätzen wird bei einem logistischen Regressionsmodell lediglich der Achsenabschnitt eines Schätzers beeinflusst.[158] Die Folge ist eine gleichmäßig höhere oder niedrigere Zahlungsausfallwahrscheinlichkeit für jede Beobachtung. Für eine Zahlungsausfallvorhersage wird später der beste Cutoff gesucht, ab dem eine Beobachtung als Zahlungsausfall klassifiziert wird. Die gleichmäßig höheren und niedrigeren Wahrscheinlichkeiten sind somit hinnehmbar und es wird mit dem unausgeglichenen Datensatz fortgefahren. Eine Zusammenfassung und Informationen zu den Koeffizienten des erstellten Modells befinden sich gekürzt in Listing 2. Unter anderem befinden sich darin p-Werte je Koeffizienten. Die p-Werte geben einen Aufschluss über die Verlässlichkeit und die Aussagekraft

[158] Vgl. *King, G., Zeng, L.*, 2001, S. 153 ff.

eines Koeffizienten. Sie basieren auf dem Wald-Test, bei der der Schätzer (*Estimate*) durch die Standardabweichung (*Std. Error*) geteilt wird.[159] Mit dem daraus resultierenden z-Wert kann der p-Wert in einer Standardnormalverteilungstabelle ermittelt werden. Die Koeffizienten mit niedrigen p-Werten sind in Listing 2 außerdem bei einem Konfidenzintervall $\alpha = 0,1\,\%$ mit drei Sternen (***) versehen. Bei $\alpha = 1\,\%$ sind es zwei Sterne (**) und bei $\alpha = 5\,\%$ ist es nur ein Stern (*). Die Schätzer enthalten die Logit-Werte (Logarithmus einer Chance), welche sich in Wahrscheinlichkeiten umrechnen lassen (beschrieben in Kapitel 3.1).

Aus Listing 2 ist ersichtlich, dass unter anderem das Alter signifikant zur Vorhersage eines Zahlungsausfalls beiträgt. Dies wurde auch schon visuell in Kapitel 4.2 festgestellt. Mit jedem Altersjahr sinkt der Logit um rund -0,01951. Rein auf dem Alter basierend ergibt sich eine Ausfallwahrscheinlichkeit bei einer 20-jährigen Person von ca. 40,3 % und bei einer 50-jährigen Person von ca. 27,3 %. Ebenfalls signifikant sind die Koeffizienten, ob der Kunde eine Telefonnummer angibt oder seine Liefer- und Postadresse übereinstimmen. Je Übereinstimmung erhöht sich die Wahrscheinlichkeit, dass ein Zahlungsausfall eintritt.

[159] Vgl. *Wasserman, L.*, 2004, S. 153.

Listing 2: Gekürzte Zusammenfassung der logistischen Regression basierend auf den Angebotsdaten ohne Postleitzahlen

```

Call:
glm(formula = binarydebtstatus ~ ., family = binomial(link = "logit"),
    data = offerdata_woplz_train)

Deviance Residuals:
    Min   1Q       Median   3Q       Max
-4.0859   -0.3374  -0.2360  -0.1636  3.6835

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
 (Intercept)                 -1.559e+01 8.284e+02 -0.019 0.984988
 sectionGas                  -5.153e-02 2.851e-01 -0.181 0.856557
 salutationFrau               1.301e+01 8.284e+02  0.016 0.987471
 salutationHerr               1.276e+01 8.284e+02  0.015 0.987713
 age                         -1.951e-02 1.035e-03 -18.854  < 2e-16  ***
 [...]
 emaildomainewetel.net       -2.029e+00 1.026e+00 -1.977 0.048028    *
 [...]
 emaildomainschutzmail.de    -1.732e+00 8.638e-01 -2.005 0.045015    *
 [...]
 hasphonenumberTRUE           2.139e-01 4.609e-02  4.642 3.45e-06  ***
 ispostalanddeliveryaddressindenticalTRUE6.739e-01 1.018e-01  6.619 3.61e-11  ***
 bicAUDFDE                   -1.307e+01 3.565e+02 -0.037 0.970764
 bicAUGBDE                    1.508e+00 5.040e-01  2.992 0.002772   **
 [...]
 bicBELADE                    1.357e+00 2.589e-01  5.240 1.61e-07  ***
 bicBEVODE                    1.609e+00 2.927e-01  5.497 3.87e-08  ***
 bicBILADE                   -4.000e-01 1.064e+00 -0.376 0.706824
 bicBRLADE                    9.722e-01 2.855e-01  3.405 0.000661  ***
 [...]
 bicCHEKDE                    1.421e+00 5.504e-01  2.582 0.009821   **
 bicCMCIDE                    8.173e-01 2.670e-01  3.061 0.002204   **
 [...]
 bicDEUTDE                    8.219e-01 2.533e-01  3.245 0.001175   **
 bicDGPBDE                    5.594e-01 4.122e-01  1.357 0.174741
 bicDORTDE                    9.661e-01 3.124e-01  3.093 0.001982   **
 bicDRESDE                    8.448e-01 2.736e-01  3.088 0.002017   **
 bicDUISDE                    1.667e+00 2.781e-01  5.993 2.06e-09  ***
 [...]
 bicFDDODE                    2.970e+00 3.031e-01  9.800< 2e-16***
 [...]
 bicHELADE                    7.868e-01 2.568e-01  3.064 0.002186   **
 bicHYVEDE                    2.292e-01 3.120e-01  0.735 0.462516
 bicINGDDE                   -7.757e-01 2.609e-01 -2.973 0.002951   **
 [...]
 bicNASSDE                    1.017e+00 3.795e-01  2.680 0.007358   **
 bicNBAGDE                    3.773e-01 5.940e-01  0.635 0.525382
 bicNOLADE                    1.115e+00 2.504e-01  4.453 8.48e-06  ***
 bicNORSDE                    1.769e+00 2.659e-01  6.651 2.92e-11  ***
 bicNTSBDE                    1.575e+00 2.667e-01  5.907 3.48e-09  ***
 [...]
 bicPAGMDE                    3.423e+00 4.217e-01  8.119 4.72e-16  ***
```

```
bicPAYPAL                    1.083e+00  2.569e-01   4.215 2.49e-05  ***
bicPBNKDE                    1.233e+00  2.480e-01   4.970 6.71e-07  ***
bicPZHSDE                    1.085e+00  3.713e-01   2.922 0.003479   **
bicSABADE                    1.218e+00  6.986e-01   1.743 0.081247    .
bicSAKSDE                    1.251e+00  4.034e-01   3.102 0.001925   **
[...]
bicSLZODE                    8.944e-01  3.074e-01   2.910 0.003618   **
bicSOBKDE                    2.922e+00  3.850e-01   7.591 3.18e-14  ***
bicSOLADE                    4.966e-01  2.665e-01   1.864 0.062351    .
bicSOLSDE                    1.463e+00  4.032e-01   3.628 0.000286  ***
bicSONSTIGES                 2.018e+00  3.311e-01   6.093 1.11e-09  ***
bicSPBIDE                    1.360e+00  3.641e-01   3.734 0.000188  ***
bicSPESDE                    7.996e-01  2.673e-01   2.991 0.002777   **
bicSPKHDE                    6.086e-01  3.291e-01   1.849 0.064414    .
bicSPKRDE                    1.114e+00  3.140e-01   3.546 0.000391  ***
bicSPMHDE                    1.002e+00  3.313e-01   3.025 0.002487   **
[...]
bicUEBERWEISER               1.380e+00  5.162e-01   2.674 0.007491   **
bicULMVDE                    1.447e+00  5.426e-01   2.667 0.007658   **
bicVBMHDE                    1.432e+00  3.640e-01   3.933 8.39e-05  ***
[...]
bicWBAGDE                    1.075e+00  5.081e-01   2.116 0.034378    *
bicWELADE                    8.176e-01  2.476e-01   3.302 0.000961  ***
bicWIBADE                   -3.331e-01  1.056e+00  -0.315 0.752442
bicWIREDE                    2.944e+00  6.364e-01   4.626 3.72e-06  ***
bicWLAHDE                    1.302e+00  3.534e-01   3.685 0.000229  ***
bicWUPSDE                    5.964e-01  3.591e-01   1.661 0.096756    .
distributionpartnerCheck24  -1.080e+00  8.927e-02 -12.098  < 2e-16  ***
distributionpartnerInterne Plattform4.099e-01  8.397e-02   4.881 1.05e-06  ***
distributionpartnerVerivox  -8.458e-01  9.269e-02  -9.126  < 2e-16  ***
hasoldsupplierTRUE          -5.226e-02  3.895e-02  -1.342 0.179669
isanuallybilledTRUE         -1.154e+00  7.110e-02 -16.233  < 2e-16  ***
bonusvalue                  -1.689e-03  2.212e-04  -7.637 2.22e-14  ***
consumptionprognosis         7.375e-05  8.013e-06   9.204  < 2e-16  ***
energycostkwh                3.705e+00  1.397e+00   2.652 0.008007  **
basepricemonthly             5.058e-02  8.597e-03   5.883 4.02e-09  ***
conclusionweekday2           5.079e-01  5.321e-02   9.546  < 2e-16  ***
conclusionweekday3           5.538e-01  5.310e-02  10.429  < 2e-16  ***
conclusionweekday4           5.319e-01  5.362e-02   9.921  < 2e-16  ***
conclusionweekday5           5.158e-01  5.427e-02   9.503  < 2e-16  ***
conclusionweekday6           4.302e-01  5.688e-02   7.563 3.95e-14  ***
conclusionweekday7           1.880e-01  6.554e-02   2.868 0.004133  **
ageover110TRUE               2.295e+00  3.841e-01   5.975 2.31e-09  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51422  on 109116  degrees of freedom
Residual deviance:41486  on 108870  degrees of freedom
AIC: 41980

Number of Fisher Scoring iterations: 15
```

Viele weitere signifikante Koeffizienten sind durch die kategoriale Variable BIC vorhanden. Ob ein Kunde ein Konto bei einer bestimmten Bank führt, lässt die Wahrscheinlichkeit eines Zahlungsausfalls steigen oder sinken. Mit einem Konfidenzintervall von $\alpha = 1\,\%$ haben Kunden bei den folgenden Banken ein höheres Zahlungsausfallsrisiko (sortiert nach BIC):

- Augsburger Aktienbank AG
- Berliner Sparkasse
- Berliner Volksbank
- Norddeutsche Landesbank
- Sparkasse Chemnitz
- Targobank
- Deutsche Bank
- Sparkasse Dortmund
- Dresdner Bank
- Sparkasse Duisburg
- Fidor Bank AG
- Landesbank Hessen-Thüringen
- Nassauische Sparkasse
- norisbank
- N26 Bank GmbH
- PayCenter GmbH
- Postbank
- Sparkasse Pforzheim Calw
- Sparkasse Saarbrücken
- Landessparkasse zu Oldenburg
- Solarisbank AG
- Sparkasse Solingen
- Sparkasse Bielefeld
- Sparkasse Essen
- Sparkasse Krefeld
- Sparkasse Mülheim an der Ruhr
- Volksbank Ulm-Biberach
- Volksbank Mittelhessen
- Sparkasse Westmünsterland
- Wirecard Bank AG
- Sparkasse Herford

Ebenfalls zählen Kunden, die mit PayPal oder Überweisung gezahlt haben, dazu sowie die Kunden, die eine BIC haben, die weniger als 20-mal im vorliegenden Datensatz repräsentiert war. Die ING-DiBa ist die einzige Bank, die mit einer verringerten Zahlungsausfallwahrscheinlichkeit, bei einem Konfidenzintervall von $\alpha = 1\,\%$, einhergeht.

Zudem sind die Variablen Vertriebspartner, Abrechnungsintervall, Verbrauchsprognose, Bonuswert und Grundpreis sowie das Kennzeichen *Kunde ist über 110 Jahre alt* signifikant ausschlaggebend. Rein auf das Alter des Kunden bezogen, steigt die Zahlungsausfallwahrscheinlichkeit um 90 %, wenn der Kunde über 110 Jahre alt ist.

Auffällig ist die Variable Wochentag, da jede Ausprägung als signifikant gekennzeichnet wird. Es wird jedoch deutlich, dass die Tage Samstag und Sonntag mit niedrigeren Zahlungsausfallwahrscheinlichkeiten in Verbindung stehen. Die Variablen Sparte, E-Mail-Adresse, das Kennzeichen *Hat Altlieferant* und viele andere BICs haben nur leicht bis gar keinen signifikanten Einfluss.

Basierend auf den neuen Erkenntnissen wurde ein verschlanktes Modell erstellt. Das verschlankte Modell beinhaltet alle Variablen wie das vorherige Modell, mit Ausnahme der nicht signifikanten Variablen Sparte, E-Mail-Adresse und Kennzeichen *Hat Altlieferant*. Generell eignet sich der Log-Likelihood zum Vergleich von Modellen. Jedoch kann bei der Hinzunahme von weiteren Parametern der Log-Likelihood nicht sinken. Daher droht eine Überparametrisierung, sollte nur ein unbereinigter Log-Likelihood zum Vergleich herangezogen werden. Im Folgenden wird das Akaike Information Criterion (AIC), definiert in Formel 19 mit n Beobachtungen und k verfügbaren Variablen, zum Modellvergleich genutzt. Der letzte Term dient als Straffunktion. Das AIC bietet sich daher zum Vergleich komplexer und weniger komplexer Modelle an, da eine Überparametrisierung bestraft wird. Konkret zeigt das AIC den Abstand zwischen dem vorhandenen Modell und dem *wahren/realen* Modell auf. Je niedriger dieser Wert ist, desto näher befindet sich das Modell an der Abbildung der Realität.[160,161] Beim direkten Vergleich der Zusammenfassungen beider Modelle fällt auf, dass das größere bzw. komplexere Modell ein niedrigeres AIC liefert. Das große Modell hat ein AIC von 41.980 und das verschlankte Modell ein AIC von 43.056. Ebenfalls erzielt der Likelihood Ratio Test (LRT) zwischen beiden Modellen einen sehr kleinen p-Wert, weshalb H_0 (das kleinere Modell) mit einem Konfidenzintervall von $\alpha = 0,1\,\%$ verworfen und sich für H_A (das komplexere Modell) entschieden wird. Das komplexere Modell weist ein $R^2_{MF} = 19{,}32\%$ auf.

Formel 19: Akaike information criterion (AIC)[162]

$$A|C = \frac{2\log_e L}{n} + \frac{2k}{n} \tag{19}$$

[160] Vgl. *Kähler, Jürgen*, 2012, S. 61, 62.
[161] Vgl. *Vrieze, S. I.*, 2012, S. 7.
[162] In Anlehnung an *Kähler, Jürgen*, 2012, S. 61.

Für die Nutzungsdaten wird analog zu den Angebotsdaten ein logistisches Regressionsmodell erzeugt. Dabei wird die Variable *Erste Mahnung nach Belieferungsbeginn in Tagen* dem Modell vorenthalten, da diese Variable nur gefüllt ist, wenn auch ein Mahnstatus bzw. Zahlungsausfall vorliegt und somit mit der abhängigen Variable einhergeht. Die Zusammenfassung des Modells ist in Listing 3 ersichtlich.

Listing 3: Gekürzte Zusammenfassung der logistischen Regression basierend auf den Nutzungsdaten ohne Variable Erste Mahnung nach Belieferungsbeginn in Tagen

```

Call:
glm(formula = binarydebtstatus ~ ., family = binomial(link = "logit"),
    data = usagedata_train[, -match(c("first_dunning_in_days_after_supplybegin",
      "ternarydebtstatus"), names(usagedata_train))])

Deviance Residuals:
    Min   1Q       Median   3Q       Max
-7.5619   -0.1505  -0.1415  -0.1355  3.3983

Coefficients:
                              Estimate  Std.Error  z value Pr(>|z|)
 (Intercept)                -3.8058098  0.0948593  -40.121  < 2e-16  ***
 supplyinterval_in_days     -0.0004750  0.0002342   -2.028 0.042549   *
 is_activeTRUE              -1.2603410  0.0737087  -17.099  < 2e-16  ***
 callcount                   0.0587856  0.0159663    3.682 0.000232  ***
 emailcount                  0.1446991  0.0135374   10.689  < 2e-16  ***
 bankaccountcount            0.5531916  0.0572146    9.669  < 2e-16  ***
 meterreadingcount          -0.0331447  0.0170454   -1.944 0.051836   .
 payplanreducedcount         0.0438864  0.0741772    0.592 0.554090
 payplanincreasedcount      -0.0683485  0.0982413   -0.696 0.486604
 separeversedcount           1.5163697  0.0237614   63.817  < 2e-16  ***
 cancellationreceivedaftersupplybeginindays-0.0039837  0.0003034  -13.130  < 2e-16  ***
manualsepacount              1.3072283  0.0447814   29.191  < 2e-16  ***
---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 (Dispersion parameter for binomial family taken to be 1)

         Null deviance: 51422   on 109116  degrees of freedom
     Residual deviance: 17423   on 109105  degrees of freedom
 AIC: 17447

 Number of Fisher Scoring iterations: 7
```

Die Variablen Kennzeichen *Ist aktiv*, Anzahl eingestellter Rückrufwünsche, Anzahl geschriebener E-Mails, Anzahl genutzter Bankdaten, Eingang der Kündigung nach Belieferungsbeginn sowie die manuellen Überweisungen und Rücklastschriften sind mit einem Konfidenzintervall von $\alpha = 0{,}1\,\%$ signifikant für die Vorhersage eines Zahlungsausfalls ausschlaggebend. Exemplarische Ausprägungen der verschiede-

nen Koeffizienten und deren Effekte auf die Zahlungsausfallwahrscheinlichkeit werden in Tabelle 8 gezeigt. Besonders auffällig sind die Variablen Anzahl manueller Überweisungen und Anzahl Rücklastschriften. Der Kunde hat bei Nichtvorhandensein eine Zahlungsausfallwahrscheinlichkeit von 18 % (Rücklastschriften) bzw. 21,3 % (manuelle Überweisungen). Sollte jedoch nur eine Rücklastschrift bzw. manuelle Überweisung vorhanden sein, so steigt die Wahrscheinlichkeit auf 82 % (Rücklastschriften) bzw. 78,7 % (manuelle Überweisungen) an. Entgegengesetzt ist es beim Eingang der Kündigung nach Belieferungsbeginn. Sollte der Kunde bereits 20 Tage nach seinem Belieferungsbeginn kündigen, so liegt die Zahlungsausfallwahrscheinlichkeit bei 48 %. Kündigt er erst 300 Tage nach dem Belieferungsbeginn, liegt die Wahrscheinlichkeit bei 23,23 %.

Die Variable, ob der Vertrag aktiv ist oder nicht, wird an dieser Stelle zur Diskussion gestellt. Ein Vertrag ist aktiv, wenn er in Belieferung ist. Der Vertrag kann keine Probleme oder einen ersten Zahlungsausfall (z. B. eine Mahnung) vorweisen. Hingegen sind die nicht aktiven Verträge all die Verträge, die gekündigt wurden. Dies kann einerseits durch den Kunden als auch im Rahmen des Inkassoprozesses passieren. Daher fehlen in der Menge der aktiven Verträge die Verträge, die den Inkassoprozess durchlaufen. In allen nicht aktiven Verträgen sind jedoch lediglich 0,9 % Verträge enthalten, die den Inkassoprozess durchlaufen (Vergleich prozentualer Anteile aus Abbildung 35g in Anhang 4 zu Abbildung 10 in Kapitel 4.2). Da die Schlussabrechnung erst mit der Beendigung der Belieferung stattfindet, werden die Kunden auch erst dann mit einer etwaigen Nachzahlung konfrontiert. Aus mehreren fachlichen Perspektiven und Interpretationen ist die Nutzung der Variable, ob der Vertrag aktiv ist oder nicht, vor- oder auch nachteilhaft. An dieser Stelle wird mit der Nutzung der Variable fortgefahren.

Koeffizent	Niedrige Ausprägung		Hohe Ausprägung	
	Wert	Wahrscheinlichkeit	Wert	Wahrscheinlichkeit
Ist aktiv	Nein	77,9 %	Ja	22,1 %
Anzahl eingestellter Rückrufwünsche	1	51,46 %	5	57,29 %
Anzahl geschriebener E-Mails	1	53,61 %	5	67,33 %
Anzahl genutzter Bankdaten	2	75,14 %	3	84,01 %
Anzahl Rücklastschriften	0	18 %	1	82 %
Anzahl manueller Überweisungen	0	21,3 %	1	78,7 %
Eingang der Kündigung nach Belieferungsbeginn (in Tagen)	20	48 %	300	23,23 %

Tab. 8: Ausgewählte Koeffizienten mit unterschiedlichen Ausprägungen und den Auswirkungen auf die Zahlungsausfallwahrscheinlichkeit in Anlehnung an Listing 3

Die Variablen Anzahl eingegebene Zählerstände sowie Erhöhungen und Verminderungen der Abschlagspläne sind nicht signifikant ausschlaggebend. Daher wird auch hier ein verschlanktes Modell erstellt, welches die Variablen nicht enthält. Der AIC-Wert verbessert sich leicht von 17.447 (initiales und komplexeres Modell) zu 17.445 (verschlanktes Modell). Mit dem LRT beider Modelle wurde ein hoher p-Wert ermittelt. H_0 wird nicht verworfen. Daher findet die weitere Auswertung mit dem verschlankten Modell statt. Das R^2_{MF} des verschlankten Modells beträgt ca. 66,11 %.

6.2 Random Forest

Mit der Modellierung von Random Forest-Modellen wird anders als mit den logistischen Regressionsmodellen fortgefahren. Zuerst ist nun die Verarbeitung des großen und komplexen Datensatzes, Angebotsdaten mit Postleitzahlen, möglich. Bei über 13.000 Variablen sind Herausforderungen für das Parameter-Tuning gegeben. Üblicherweise müssen die optimalen Parameter zum Bau der Random Forests gefunden werden. Hierzu zählen je Baum die maximale Tiefe, die maximale Anzahl der Beobachtungen zum Training, Werte, bei denen ein Split geschehen

darf, oder übergreifend die maximale Anzahl der Entscheidungsbäume im Random Forest. Da in Kapitel 3.2.1 gezeigt wurde, dass weder der Gini Index noch der Information Gain einen signifikanten Vorteil gegenüber der anderen Split-Funktion haben, wird der Gini Index als Standardeinstellung und Split-Funktion beibehalten. Random Forests ermöglichen bei unausgewogenen Datensätzen die Gewichtung der Beobachtungen. Bei einem Split wird der minderheitlich vertretenen Klasse ein höheres Gewicht bzw. ein höherer Gain zugeschrieben, wodurch sie bei einem Split eine gleichwertige Bedeutung gegenüber der mehrheitlich vertretenen Klasse erlangt. Gewichtete Random Forests benötigen zum Training den gesamten Datensatz und sind anfälliger für falsch klassifizierte Beobachtungen.[163] Da die Quelldaten keine falsch klassifizierten Beobachtungen enthalten, wird mit den gewichteten Random Forests fortgefahren. Bei dem vorliegenden großen Datensatz hat die Erstellung eines ersten Random Forests mit 1.000 Entscheidungsbäumen mit je einer maximalen Tiefe von 40 und 10 % der Trainingsdaten rund 4 Minuten gedauert. Jeder Test mit einem angepassten Parameter lässt die benötigte Zeit um ungefähr die gleiche Zeit ansteigen. Bei exemplarischen fünf Parametern mit je drei zu testenden Ausprägungen würde ein Zeitaufwand von ca. 75 Minuten entstehen. Da dies ein übliches Problem mit großen Datensätzen darstellt, wird an dieser Stelle mit der Identifizierung und Entfernung unwichtiger Variablen fortgefahren, um den rechnerischen Aufwand eines Parameter-Tunings zu minimieren.[164] Wie in Kapitel 3.2 beschrieben, wird innerhalb eines Entscheidungsbaums immer die Variable gesucht, die ein Set in zwei Sets mit der höchsten Reinheit teilt. Je Knoten wird immer der beste Split gesucht, bis das Set rein genug ist, um eine Klassifizierung durchzuführen. Mithilfe des Mean Decrease in Impurity (MDI) wird über alle Entscheidungsbäume im Random Forest der *durchschnittliche Verlust an Unreinheit* jeder Variable gemessen. Eine Variable mit einem hohem MDI verringert die Unreinheit besonders, ergo sorgt die Variable für die größte Reinheit.[165]

[163] Vgl. *Chen*, *C.*, *Liaw*, *A.*, *Breiman*, *L.* et al., 2004, S. 8.
[164] Vgl. *Biau*, *G.*, *Scornet*, *E.*, 2016, S. 10.
[165] Vgl. ebd., S. 26 f.

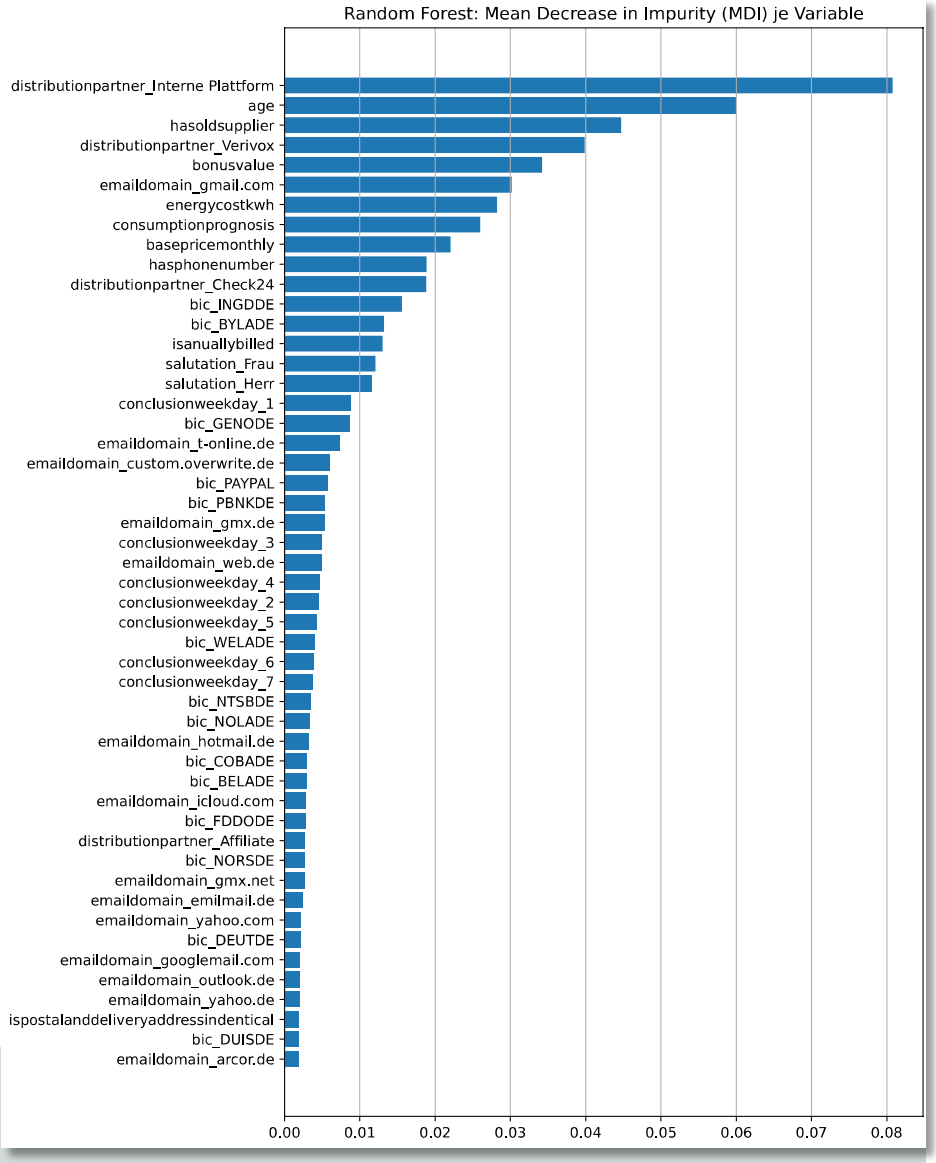


Abb. 15: Random Forest: Variablen mit dem größten MDI (trainiert mit Angebotsdaten inkl. Postleitzahlen)

Eine grafische Aufstellung der Variablen mit dem höchsten MDI befindet sich in Abbildung 15. Für den Random Forest stellen die Variablen Vertriebspartner *Interne Plattform*, Alter und Kennzeichen *Hat aktiven Altlieferant* den höchsten

Reinheitsgewinn dar. Im Vergleich mit dem zuvor entwickelten logistischen Regressionsmodell (Kapitel 6.1) wurde das Kennzeichen *Hat aktiven Altlieferant* als nicht signifikant dargestellt, jedoch steht es beim Random Forest an dritter Stelle, um die Reinheit im Datensatz herzustellen. Darauffolgend sind andere Variablen wie der Bonuswert, Verbrauchsprognose und vereinzelte E-Mail-Domains oder BICs mit einem höheren MDI versehen. Die Werte streben schnell nahe 0. Variablen mit einem sehr niedrigen MDI werden aussortiert, sodass die weiteren Tests zum Parameter-Tuning stattfinden können. Der Cutoff wird auf 0,0001 gesetzt. Von den ursprünglich 13.308 Variablen bleiben 907 Variablen übrig (inklusive der abhängigen Variable).

Mit dem verschlankten Datensatz finden Tests zum Parameter-Tuning statt. Hierzu werden kombinatorisch alle Parameter mit allen Werten der Reihe nach getestet, die in Tabelle 9 aufgeführt sind. Mithilfe der Parameter werden weitere Entscheidungsbäume zum Random Forest hinzugefügt, die maximale Tiefe oder die maximale Anzahl der Blätter der jeweiligen Bäume begrenzt sowie die notwendigen Beobachtungen für die Erstellung eines Blatts erhöht. Dabei ist zu beachten, dass der Parameter zur Beschränkung der Blätteranzahl nur angewandt wurde, wenn der Wert kleiner oder gleich der maximal verfügbaren Blätteranzahl ($2^{\text{Maximale Tiefe}}$) ist. Das Training und der Vergleich basieren somit auf 30 Random Forests.

Parameter	Wert
Maximale Tiefe	8, 10, 12, 14
Minimum Beobachtungen je Blatt	1, 3, 5
Maximale Anzahl Blätter	64, 256, 1024, 4096, 16384
Anzahl Entscheidungsbäume	200, 600, 1000, 1400

Tab. 9: Random Forest: Parameter und Werte zum Parameter-Tuning-Test

Formel 20: F1-Score[166]

$$F_1 = \frac{TP}{TP + \frac{1}{2} \cdot (FP + FN)} \tag{20}$$

[166] In Anlehnung an *Chicco, D., Jurman, G.*, 2020, S. 5.

Für eine Evaluierung der verschiedenen Random Forests, die beim Test zum Parameter-Tuning erstellt werden, werden 5 % der Trainingsdaten in einen separaten OOB-Datensatz geschrieben. Mithilfe des OOB-Datensatzes, der beim Training vorenthalten wurde, findet vorab eine Vorhersage der abhängigen Variable statt. Die Auswertung der Vorhersage findet mit dem F1-Score statt. Der F1-Score ist in Formel 20 definiert. Er stellt damit ein harmonisches Mittelmaß zwischen der Sensitivität und der PPV dar.[167] Die Ergebnisse des Parameter-Tuning-Tests mit den zehn besten Konfigurationen befinden sich in Tabelle 10. Eine Visualisierung als Diagramm der neun besten Random Forests ist in Abbildung 16 enthalten sowie ein Diagramm mit allen Entscheidungsbäumen in Abbildung 37 in Anhang 7. Auffällig unter den zehn besten Parameterkonfigurationen sind die Minima der Beobachtungen je Blatt, die immer bei 1 liegen sowie die Überlagerung der Linien im Graphen. Ebenso ist die maximale Tiefe auffällig, die bei den besten Konfigurationen immer bei 14 liegt. Die Anpassung der maximalen Blätteranzahl oder auch die Anpassung der Anzahl der Entscheidungsbäume je Random Forest bringen nur geringfügige Änderungen am F1-Score mit sich. Generell liegen die zehn höchsten F1-Scores mit einer Abweichung von maximal 0,3179 relativ dicht beieinander. Der Random Forest mit einer maximalen Tiefe von 14, den Minimum Beobachtungen je Blatt von 1, der maximalen Blätteranzahl von 1024 und insgesamt 1000 Entscheidungsbäumen erzielt den höchsten Wert mit $F_1 = 0,328597$. Dieser Random Forest wird für die Evaluierung der Angebotsdaten mit Postleitzahlen ausgewählt und weiterverwendet, da er zum Vergleich mit der gleichwertigen Konfiguration an Stelle zwei weniger komplex ist.

[167] Vgl. *Sasaki, Y.*, *Fellow, R.*, 2007, S. 3.

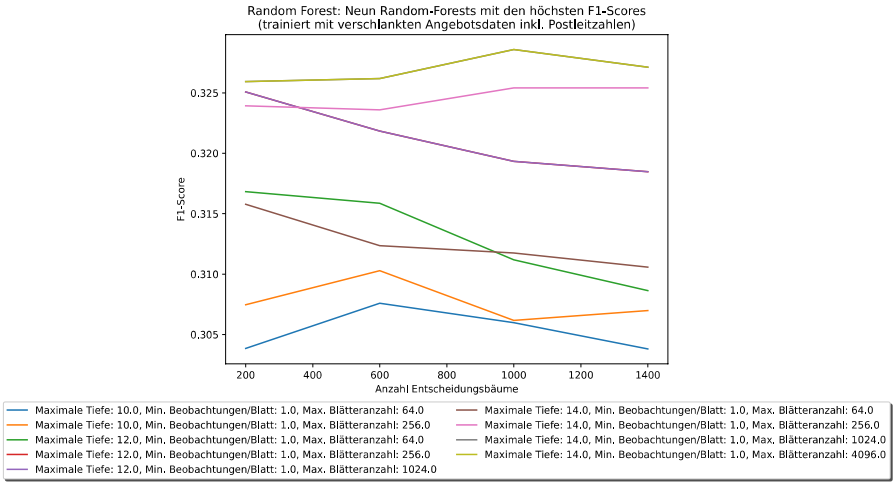


Abb. 16: Random Forest: Neun Random Forests mit den höchsten F1-Scores basierend auf Parametern aus Tabelle 9 (trainiert mit verschlankten Angebotsdaten inkl. Postleitzahlen)

Max. Tiefe	Min. Beobachtungen je Blatt	Max. Blätter	Anzahl Entscheidungsbäume	F1-Score
14	1	1024	1000	0,328597
14	1	4096	1000	0,328597
14	1	1024	1400	0,327144
14	1	4096	1400	0,327144
14	1	1024	600	0,326203
14	1	4096	600	0,326203
14	1	1024	200	0,325952
14	1	4096	200	0,325952
14	1	256	1400	0,325418
14	1	256	1000	0,325418

Tab. 10: Random Forest: Zehn Parameterkonfigurationen mit den höchsten F1-Scores (trainiert mit verschlankten Angebotsdaten inkl. Postleitzahlen)

Für den Datensatz Angebotsdaten ohne Postleitzahlen wird analog zum vorherigen Datensatz ein ähnliches Prozedere durchgeführt. Nach dem Einlesen des Datensatzes wird direkt mit der Modellierung und der Suche nach der Parameterkonfigura-

tion, die den besten F1-Score liefert, begonnen. Da der Datensatz weniger Variablen enthält, ist das Training schneller geschehen und es liegen keine Einschränkungen in Bezug auf die Rechenleistung vor. Zudem lassen die Entscheidungsbäume durch die diversen Parameter automatisch Variablen zur Klassifizierung aus. Ebenso kann ein direkter Vergleich zur logistischen Regression stattfinden, bei der das Modell ebenfalls mit allen Variablen trainiert wurde.

Durch die vorherigen Erfahrungen wurde die Suche nach der besten Parameterkonfiguration eingeschränkt, sodass nur noch 20 Random Forests trainiert werden müssen. In Abbildung 17 sind alle Parameterkonfigurationen aufgezählt und in einem Diagramm visualisiert sowie die Parameterkonfigurationen mit den zehn besten F1-Scores in Tabelle 11 aufgeführt. Ähnlich wie in Abbildung 16 steigen und sinken die F1-Scores mit der Hinzunahme weiterer Entscheidungsbäume. Ebenso beinhalten die zehn besten Konfigurationen in Tabelle 11 für den Parameter Minimum Beobachtungen je Blatt immer den Wert 1. Dies wurde ebenfalls zuvor in Tabelle 10 beobachtet. Zum Datensatz Angebotsdaten ohne Postleitzahlen erzielt ein Random Forest mit einer maximalen Tiefe von 14, einer Beobachtung als Minimum je Blatt, maximal 1024 Blätter sowie insgesamt 1400 Entscheidungsbäume den höchsten F1-Score mit $F_1 = 0,357672$. Eine weitere Konfiguration erreicht denselben F1-Score, jedoch wird diese Konfiguration wegen ihrer höheren Komplexität nicht gewählt.

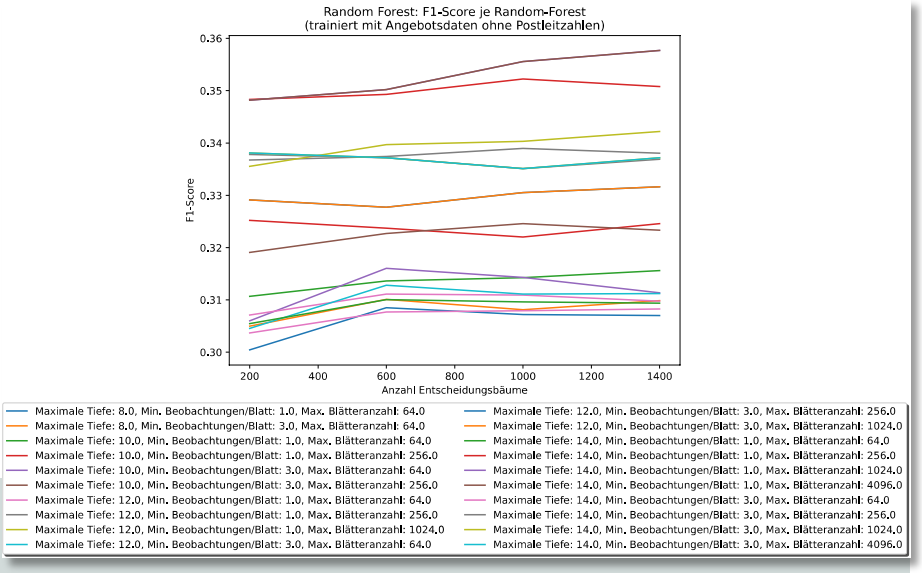


Abb. 17: Random Forest: Random Forests mit den höchsten F1-Scores trainiert mit Angebotsdaten ohne Postleitzahlen

Max. Tiefe	Min. Beobachtungen je Blatt	Max. Blätter	Anzahl Entscheidungsbäume	F1-Score
14	1	4096	1400	0,357672
14	1	1024	1400	0,357672
14	1	1024	1000	0,355556
14	1	4096	1000	0,355556
14	1	256	1000	0,352250
14	1	256	1400	0,350775
14	1	1024	600	0,350211
14	1	4096	600	0,350211
14	1	256	600	0,349268
14	1	256	200	0,348337

Tab. 11: Random Forest: Zehn Parameterkonfigurationen mit den höchsten F1-Scores (trainiert mit Angebotsdaten ohne Postleitzahlen)

Für die Entwicklung eines Modells auf Basis der Verhaltensdaten wird für die vorliegenden Variablen der MDI mithilfe eines ersten exemplarischen Random Forests berechnet. Dabei wird, wie bei der logistischen Regression, die Variable *Erste Mahnung nach Belieferungsbeginn in Tagen* dem Modell vorenthalten, da die Variable nur gesetzt ist, wenn auch die abhängige Variable `True` ist. Die berechneten MDI je Variable auf Basis der Verhaltensdaten befinden sich in Abbildung 18. Ähnlich wie bei beim logistischen Regressionsmodell aus Listing 3 haben die Variablen Anzahl Rücklastschriften sowie Anzahl manueller Überweisungen den größten Stellenwert und bringen im Random Forest die größte Reinheit im Datensatz. Ebenso gleichen die Modelle sich in der Bewertung der Variablen Anzahl der Abschlagsplanverminderungen oder -erhöhungen. Beide Variablen weisen einen sehr niedrigen MDI auf und haben keinen großen Einfluss auf die Klassifizierung eines Zahlungsausfalls. Da bei der logistischen Regression ein verschlanktes Modell gegenüber dem Modell mit allen Variablen vorgezogen wurde, werden die Random Forests im folgenden Schritt auf Basis beider Datensätze trainiert. Der eine Datensatz beinhaltet alle Variablen exkl. Variable *Erste Mahnung nach Belieferungsbeginn in Tagen*. Im verschlankten Datensatz fehlen zudem die beiden Variablen mit dem schwächsten MDI, Anzahl Abschlagsplanverminderungen sowie Anzahl Abschlagsplanerhöhungen. Zu beiden Datensätzen werden die

Parameter aus Tabelle 12 kombinatorisch und sequenziell durchgetestet. Das Training umfasst insgesamt 24 Random Forests, deren Entscheidungsbaumanzahl in fünf Schritten erhöht wird.

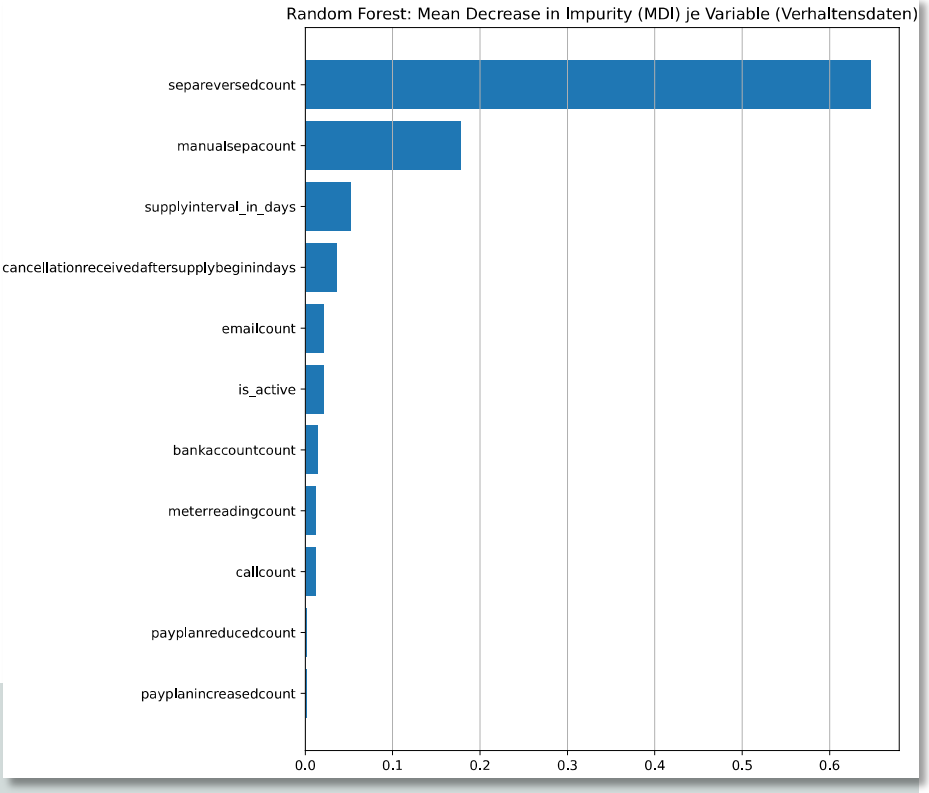


Abb. 18: Random Forest: Variablen mit gemessenen MDI (trainiert mit den Verhaltensdaten, exkl. Variable *Erste Mahnung nach Belieferungsbeginn in Tagen*)

Parameter	Wert
Maximale Tiefe	2, 4, 6, 8
Minimum Beobachtungen je Blatt	1, 3
Maximale Anzahl Blätter	4, 16, 64, 256
Anzahl Entscheidungsbäume	200, 400, 600, 800, 1000

Tab. 12: Random Forest: Parameter und Werte zum Parameter-Tuning-Test für Verhaltensdaten

Die Ergebnisse des Trainings sind reduziert auf die zwölf Random Forests mit den höchsten F1-Scores in Abbildung 19 dargestellt. Ein Graph über alle trainierten

Random Forests befindet sich in Abbildung 38 in Anhang 8. Zusätzlich sind die zehn Parameterkonfigurationen mit den höchsten F1-Scores in Tabelle 13 aufgeführt.

Auch hier ist direkt ersichtlich, dass die F1-Scores aller Random Forests sehr nah beieinander liegen. Die Differenz zwischen dem kleinsten und größten F1-Score beträgt lediglich rund 0,005. Ebenso überlagern die verschiedenen Linien der einzelnen Random Forests in Abbildung 19 einander, da oftmals dieselben F1-Scores erreicht werden. Besonders in Tabelle 13 wird deutlich, dass die besten Ergebnisse mithilfe des verschlankten Datensatzes erzielt werden. Wie bei der logistischen Regression wird das Random Forest Modell auf Basis des verschlankten Datensatzes (exkl. Variable *Erste Mahnung nach Belieferungsbeginn in Tagen*) trainiert und für die Vorhersage eines Zahlungsausfalls genutzt. Die drei besten Parameterkonfigurationen erzielen alle denselben F1-Score von $F_1 = 0,843945$ und unterscheiden sich nur in der Anzahl der Entscheidungsbäume. Für eine Evaluierung findet nun ein Kompromiss zwischen F1-Score und Komplexität statt. Da die F1-Scores sehr dicht beieinander liegen, wird der Random Forest mit der geringsten Komplexität aus den zehn besten Parameterkonfigurationen gewählt (aus Tabelle 13). Der ausgewählte Random Forest beinhaltet 200 Entscheidungsbäume, davon haben die Entscheidungsbäume eine maximale Tiefe von 8, drei Beobachtungen als Minimum zur Erzeugung eines Blatts und maximal 64 Blätter.

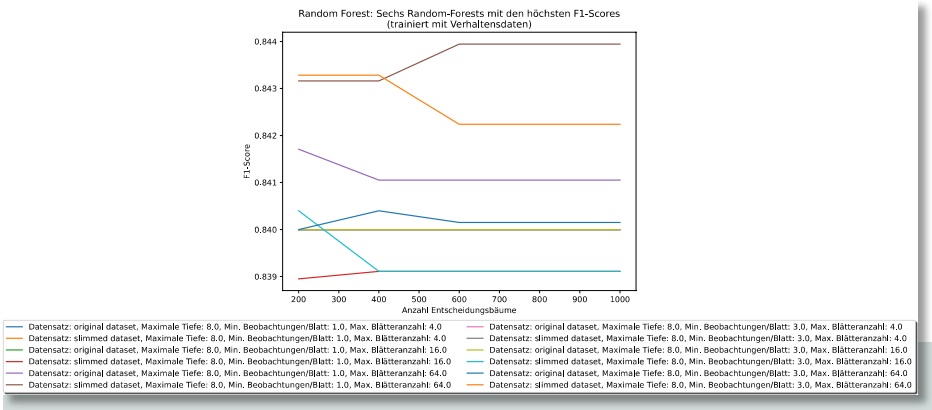


Abb. 19: Random Forest: Sechs Random Forests mit den höchsten F1-Scores basierend auf Parametern aus Tabelle 12 (trainiert mit den Verhaltensdaten, exkl. Variable *Erste Mahnung nach Belieferungsbeginn in Tagen*)

Max. Tiefe	Min. Beobachtungen je Blatt	Max. Blätter	Anzahl Entscheidungsbäume	Datensatz	F1-Score
8	1	64	1000	Verschlankt	0,843945
8	1	64	800	Verschlankt	0,843945
8	1	64	600	Verschlankt	0,843945
8	3	64	400	Verschlankt	0,843284
8	3	64	200	Verschlankt	0,843284
8	1	64	400	Verschlankt	0,843162
8	1	64	200	Verschlankt	0,843162
8	3	64	1000	Verschlankt	0,842236
8	3	64	800	Verschlankt	0,842236
8	3	64	600	Verschlankt	0,842236

Tab. 13: Random Forest: Zehn Parameterkonfigurationen mit den höchsten F1-Scores (trainiert mit den Verhaltensdaten, exkl. Variable *Erste Mahnung nach Belieferungsbeginn in Tagen*)

6.3 XGBoost

Mit XGBoost werden auf Basis der verschiedenen Datensätze die letzten Modelle zur Vorhersage von Zahlungsausfällen entwickelt. Da es sich bei XGBoost wie bei den Random Forests um eine Ensemble Methode der Entscheidungsbäume handelt, ähnelt sich der Entwicklungsablauf stark. Analog zu den Random Forest-Modellen wird die minderheitlich vorhandene Klasse (*Zahlungsausfall*) mit einem höheren Gewicht versehen, damit die Verlustfunktion die Klasse *Zahlungsausfall* nicht vernachlässigt. Große Unterschiede sind exemplarisch, dass die Entscheidungsbäume mit XGBoost sequenziell trainiert werden und durch das Boosting weitere Parameter zur Verfügung stehen. Da XGBoost mit einem neuen Entscheidungsbaum immer den Fehler des vorangegangenen Entscheidungsbaums verbessert, kann die Fehlerrate während des Trainings gemessen werden. Die Verbesserung kann durch den sequenziellen Aufbau kontinuierlich beobachtet werden. Hierzu werden von jedem Datensatz, ähnlich wie bei den Random Forests, 5 % der Trainingsdaten zur Validierung des besten XGBoost-Modells entnommen. Die gleichmäßige Verteilung der Klassen (Zahlungsausfall/Kein Zahlungsausfall) ist berücksichtigt.

Zu Beginn wird ein exemplarisches Modell basierend auf dem Datensatz mit Postleitzahlen trainiert. Bei dieser Modellierung werden die Standardparameter genutzt.

Das exemplarische Modell dient erneut der Identifizierung der wichtigsten Variablen, um die Tests zum Parameter-Tuning mit einer verminderten Variablenanzahl bzw. einem verschlankten Datensatz durchzuführen. Anders als bei den Random Forests wird beim XGBoost-Modell der durchschnittliche Gain je Variable herangezogen. Der Gain impliziert jedoch nichts anderes, als dass eine bestimmte Reinheit herbeigeführt wird. Die 50 Variablen mit dem größten durchschnittlichen Gain sind in Abbildung 20 dargestellt. Im direkten Vergleich mit Abbildung 15 der Random Forests ist der Vertriebspartner Interne Plattform nach wie vor an erster Stelle. Hingegen sind diverse Dummy-Variablen der E-Mail-Domains und der BICs mit einem höheren durchschnittlichen Gain eingestuft.

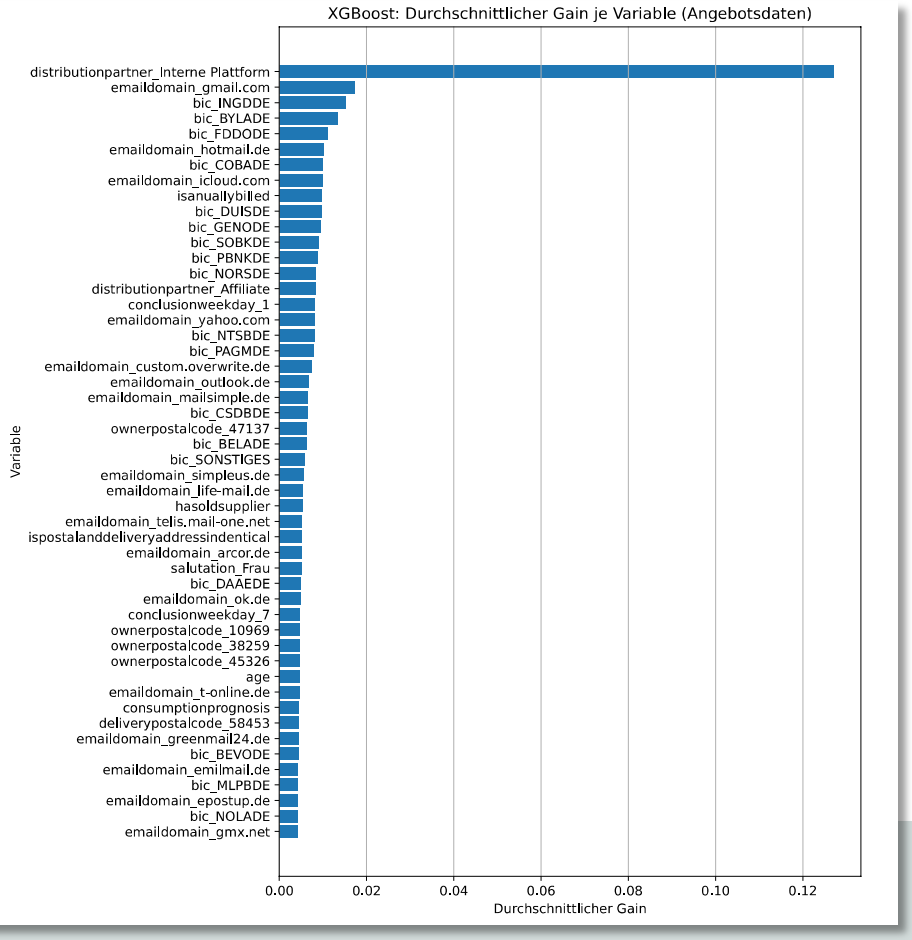


Abb. 20: XGBoost: Variablen mit den höchsten durchschnittlichen Gain (trainiert mit Angebotsdaten inkl. Postleitzahlen)

Für den verschlankten Datensatz werden nur Variablen behalten, die einen durchschnittlichen Gain > 0,0001 haben. Der daraus resultierende Datensatz umfasst nach der Bereinigung nur noch 296 Variablen, inklusive der abhängigen Variable. Mit dem neuen verschlankten Datensatz werden Tests zum Parameter-Tuning durchgeführt. Für den Test werden vorab drei Parameter herangezogen: Die maximale Tiefe eines Entscheidungsbaums, der L1-Regularisierungsparameter α und der L2-Regularisierungsparameter λ. Durch die Kombination aller Parameterwerte aus Tabelle 14 werden insgesamt 60 XGBoost-Modelle trainiert. Bei der Berechnung der Modelle wird auf den Graphics Processing Unit (GPU)-Support der XGBoost-Bibliothek zurückgegriffen, um das Training mithilfe der Grafikkarte zu beschleunigen. Für die Messung des Verlusts wird der separat erstellte 5 % Validierungsdatensatz genutzt. Die Verlustfunktion entspricht der negativen Log-Likelihood-Funktion (engl. *Log-Loss*). Zusätzlich wird, in Kombination mit dem separat erstellten 5 % Validierungsdatensatz, vom Parameter *early_stopping_rounds* Gebrauch gemacht. Durch den Parameter bricht das Training weiterer Entscheidungsbäume ab, sobald auf dem Validierungsdatensatz nach einer definierten Anzahl von Iterationen keine Verbesserung festgestellt wurde. Dies verhindert zusätzlich ein Overfitting. Im Test werden der Parameter auf den Wert 30 gesetzt und maximal 2000 Iterationen (gleich aufeinanderfolgende Entscheidungsbäume) durchgeführt. Die Lernrate wird mit der höheren Anzahl an Entscheidungsbäumen beim Standardwert $\eta = 0{,}3$ belassen.

Parameter	Wert
Maximale Tiefe	6, 8, 10, 12
α (L1-Regularisierung)	0, 0,5, 1, 1,5, 2
λ (L2-Regularisierung)	1, 1,5, 2

Tab. 14: XGBoost: Parameter und Werte zum Parameter-Tuning-Test (auf Basis der Angebotsdaten)

Die Konfigurationen und Ergebnisse der acht besten Konfigurationen sind in Tabelle 15 dargestellt sowie in Abbildung 21 visualisiert. Eine Darstellung aller XGBoost-Modelle befindet sich in Anhang 9. Die XGBoost-Modelle mit den fünf niedrigsten Log-Loss sind mit einer maximalen Tiefe von 12 ausgeprägt. Anschließend folgt ein Modell mit einer maximalen Tiefe von 10. Das beste Modell wurde mit der L1-Regularisierung $\alpha = 2$, der L2-Regularisierung $\lambda = 1{,}5$ und der maximalen Tiefe von 12 trainiert. Es erreicht einen Log-Loss von ca. 0,23.

Insgesamt waren 274 Iterationen bzw. aufeinanderfolgende Entscheidungsbäume notwendig, um diesen Wert zu erreichen. Zur direkten Reduzierung der Modellkomplexität und Behandlung eines vermeintlichen Overfittings wurden die weiteren Parameter Minimum Blattgewicht und Minimum Split-Verlust γ in diversen Ausprägungen in Kombination mit der besten Modellkonfiguration aus Tabelle 15 getestet.[168] Die Ergebnisse haben keine Senkung des Log-Loss herbeigeführt und werden daher für das zu spezifizierende Modell ausgelassen. Das Modell für die Vorhersage eines Zahlungsausfalls auf Basis des verschlankten Datensatzes Angebotsdaten inkl. Postleitzahlen hat eine maximale Tiefe von 12, $\alpha = 2$ und $\lambda = 1{,}5$ mit insgesamt 274 Iterationen.

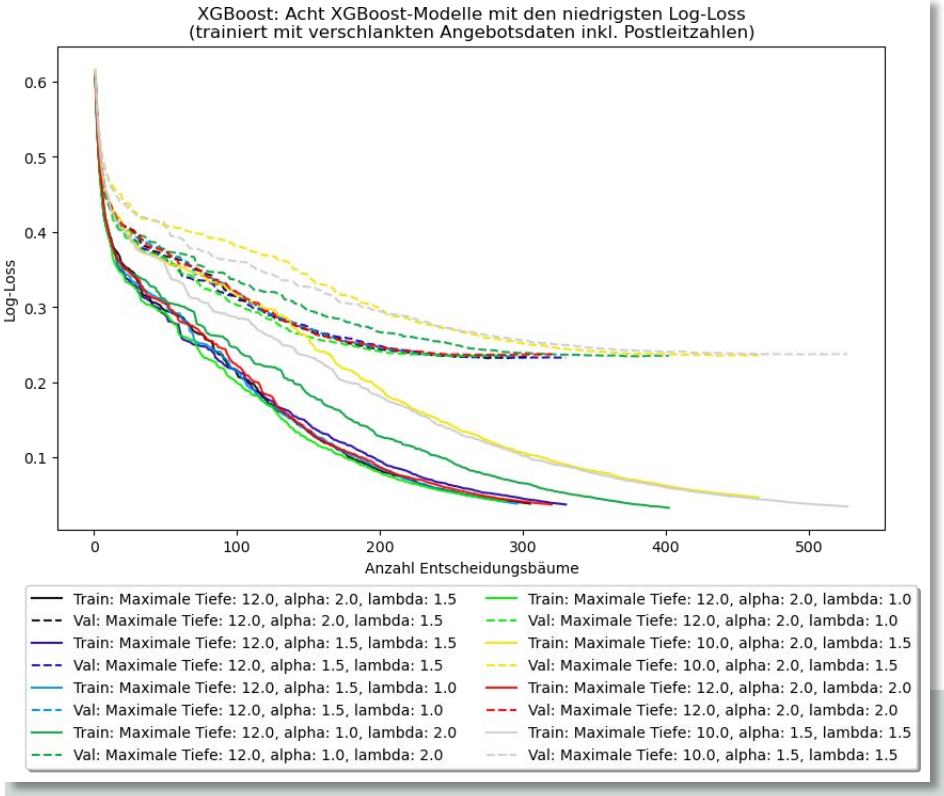


Abb. 21: XGBoost: Acht XGBoost-Modelle mit den niedrigsten Log-Loss (trainiert mit verschlankten Angebotsdaten inkl. Postleitzahlen)

[168] Vgl. Control Overfitting xgboost developers, 2022c.

Max. Tiefe	α	λ	Log-Loss	Iterationen
12	2,0	1,5	0,232085	274
12	1,5	1,5	0,232629	299
12	1,5	1,0	0,234421	265
12	1,0	2,0	0,234642	371
12	2,0	1,0	0,234933	257
10	2,0	1,5	0,235527	434
12	2,0	2,0	0,236241	289
10	1,5	1,5	0,237039	496

Tab. 15: XGBoost: Acht Parameterkonfigurationen mit den niedrigsten Log-Loss (trainiert mit verschlankten Angebotsdaten inkl. Postleitzahlen)

Für den Datensatz Angebotsdaten ohne Postleitzahlen wird dasselbe Prozedere wie mit dem zuvor genutzten Datensatz mit Postleitzahlen durchlaufen. Ebenso wird für die Suche nach den besten Parametern die Kombinationen aus Tabelle 14 durchlaufen. Die Ergebnisse sind in Tabelle 16 aufgelistet und in Abbildung 22 dargestellt. Die beste Konfiguration hat mit einer Tiefe von 12, einer L1-Regularisierung $\alpha = 2$, der L2-Regularisierung $\lambda = 1{,}5$ und mit 248 Iterationen den besten Log-Loss von rund 0,231 erzielt. Eine nachfolgende Komplexitäts- und Overfittingreduzierung mit den Parametern Minimum Blattgewicht und Minimum Split-Verlust γ hat keine Verbesserung ergeben, sodass mit den zuvor genannten Parametern das Modell zur Vorhersage von Zahlungsausfällen basierend auf dem Datensatz Angebotsdaten ohne Postleitzahlen ausgewählt wird.

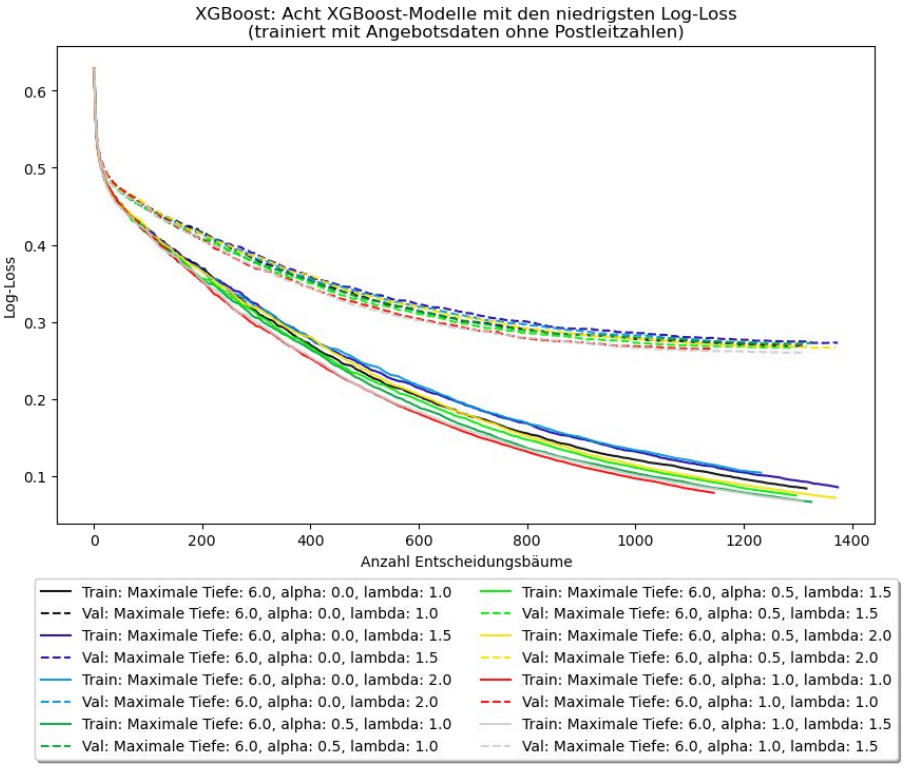


Abb. 22: XGBoost: Acht XGBoost-Modelle mit den niedrigsten Log-Loss (trainiert mit Angebotsdaten ohne Postleitzahlen)

Max. Tiefe	α	λ	Log-Loss	Iterationen
12	2,0	1,5	0,231285	248
12	2,0	1,0	0,233261	231
12	1,0	1,0	0,233406	236
12	1,0	2,0	0,233631	249
12	1,5	2,0	0,234418	262
12	2,0	2,0	0,235083	228
10	2,0	1,5	0,235689	378
12	0,5	1,0	0,236874	230

Tab. 16: XGBoost: Acht Parameterkonfigurationen mit den niedrigsten Log-Loss (trainiert mit Angebotsdaten ohne Postleitzahlen)

Zuletzt werden die Modelle auf Basis der Verhaltensdaten konzipiert. Da die enthaltenen Variablen nicht den vorherigen Variablen entsprechen, wird erneut der durchschnittliche Gain je Variable mithilfe eines exemplarischen XGBoost-Modells berechnet. Wie auch bei den Random-Forest-Modellen wurde die Variable *Erste Mahnung nach Belieferungsbeginn in Tagen* zum Training aus dem Datensatz entfernt, da diese Variable mit der abhängigen Variable einhergeht. In Abbildung 23 werden die Ergebnisse visualisiert. Ähnlich wie beim Random Forest (Vgl. Abbildung 18) sticht die Variable Anzahl Rücklastschriften mit dem größten Gain hervor. Alle anderen Variablen sind für das XGBoost-Modell mit einem niedrigen Gain versehen. Ebenfalls stellen die Variablen Anzahl der Abschlagsplanerhöhungen bzw. -verminderungen die niedrigsten Gains dar.

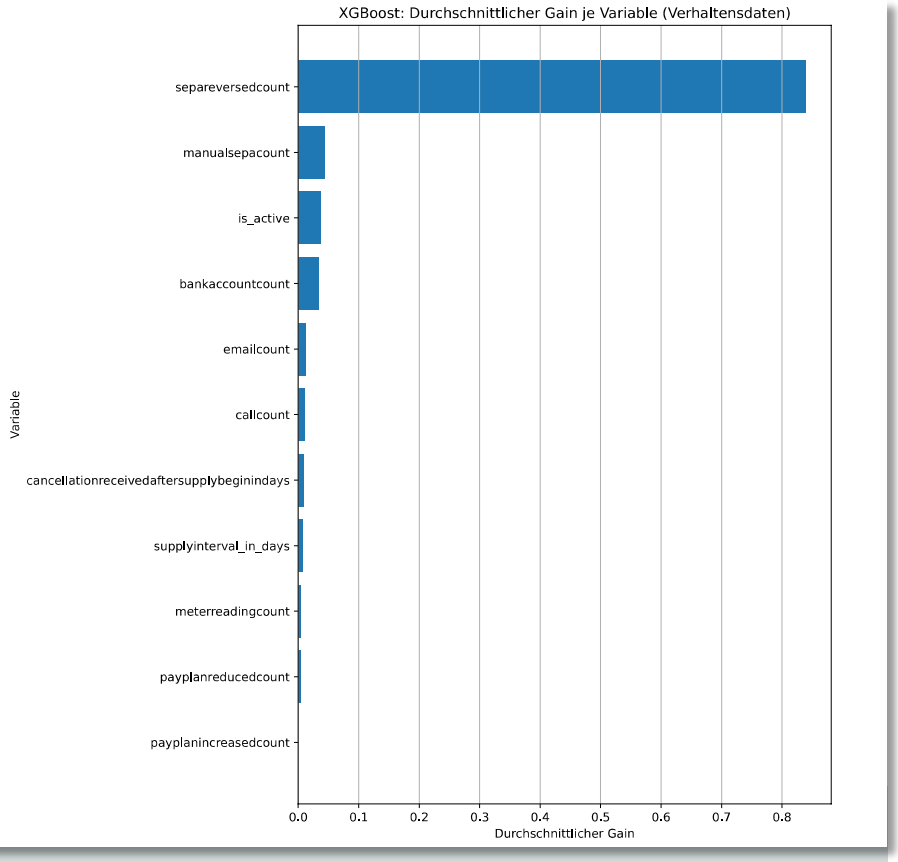


Abb. 23: XGBoost: Variablen mit dem größten durchschnittlichen Gain
(trainiert mit den Verhaltensdaten, exkl. Variable *Erste Mahnung nach Belieferungsbeginn in Tagen*)

Für den Parameter-Tuning-Test werden wegen der verminderten Variablenanzahl niedrigere Werte für die maximale Tiefe eines Entscheidungsbaums gesetzt. Da der Datensatz weniger Variablen als die vorherigen Datensätze hat, die maximale Tiefe gesenkt wird und das Training mittels GPU geschieht, verläuft das Training sehr viel schneller. Deshalb werden ebenfalls die Parameter zur direkten Komplexitäts- und Overfittingreduzierung, Minimum Blattgewicht und Minimum Split-Verlust γ mitgetestet. Alle Parameter und Werte sind in Tabelle 17 aufgelistet. Zudem wird der Parameter *early_stopping_rounds* auf den Wert `10` heruntergesetzt. In der Kombination werden 675 XGBoost-Modelle trainiert.

Parameter	Wert
Maximale Tiefe	4, 6, 8
α (L1-Regularisierung)	0, 0,5, 1, 1,5, 2
λ (L2-Regularisierung)	1, 1,5, 2
Minimum Blattgewicht	1, 3, 5
γ (Minimum Split-Verlust)	0, 0,25, 0,5, 0,75, 1

Tab. 17: XGBoost: Parameter und Werte zum Parameter-Tuning-Test (auf Basis der Verhaltensdaten)

Die XGBoost-Modelle mit den acht niedrigsten Log-Loss sind in Tabelle 18 aufgelistet sowie in Abbildung 24 visualisiert. Das Training aller Modelle wird nach maximal 212 Iterationen beendet. Die besten acht Modelle erreichen bereits bei 20 Iterationen einen sehr niedrigen Log-Loss und verbessern sich anschließend nur noch marginal. Zudem liegen alle Log-Loss sehr nahe beieinander, wodurch eine starke Überlagerung der Kurven in Abbildung 24 stattfindet. Das Delta zwischen dem niedrigsten und höchsten Log-Loss der acht besten Modelle liegt bei rund 0,014. Für die Vorhersage von Zahlungsausfällen auf Basis der Verhaltensdaten wird das XGBoost-Modell mit den folgenden Parametern ausgewählt: Maximale Tiefe von 8, Minimum Blattgewicht von 1 (Standardwert), Minimum Split-Verlust von 0 (Standardwert), $\alpha = 1{,}5$ und $\lambda = 1{,}5$ bei insgesamt 124 Iterationen. Dieses Modell erreichte den niedrigsten Log-Loss mit rund 0,076.

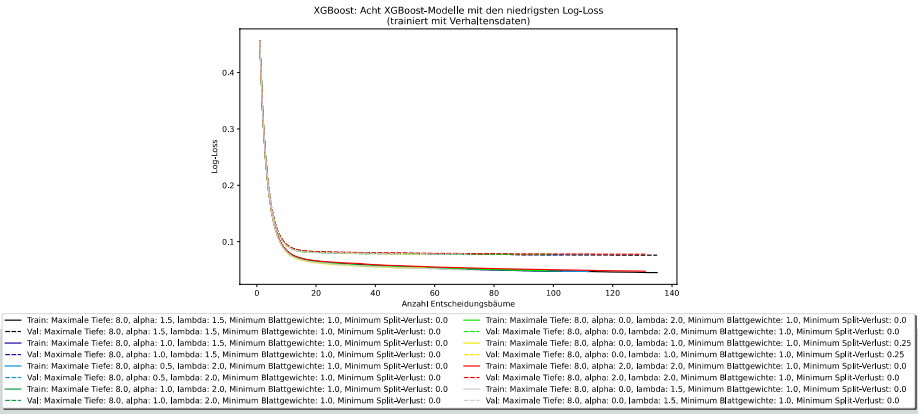


Abb. 24: XGBoost: Acht XGBoost-Modelle mit den niedrigsten Log-Loss (trainiert mit Verhaltensdaten)

Max. Tiefe	Minimum Blattgewicht	Minimum Spilt-Verlust	α	λ	Log-Loss	Iterationen
8	1	64	1,5	1,5	0,075939	124
8	1	64	1,0	1,5	0,076517	92
8	1	64	0,5	2,0	0,076861	101
8	1	64	1,0	2,0	0,077329	91
8	1	64	0,0	2,0	0,077335	87
8	1	64	0,0	1,0	0,077729	46
8	1	64	2,0	2,0	0,077740	120
8	1	64	0,0	1,5	0,077742	60

Tab. 18: XGBoost: Acht Parameterkonfigurationen mit den niedrigsten Log-Loss (trainiert mit Verhaltensdaten)

7 Evaluierung

7.1 Methodik

In Kapitel 2 wurden die aktuelle Marktlage widergespiegelt sowie die Einnahmen von erfolgreichen Belieferungsverträgen und die Verluste bei Verträgen mit Zahlungsausfällen aufgezeigt. Aus Gründen der einfacheren Darstellung und der ähnlichen Gewinnspannen liegt der Fokus in der Evaluierung der Modelle auf den Kunden der Sparte Strom. Der Betrachtungszeitraum umfasst ein Jahr. Die Einnahmen eines Stromkunden im ersten Belieferungsjahr mit Bonuszahlung reichen von -100,75 € bis zu 393,50 €. Von den Einnahmen verbleiben 24,9 % beim Stromlieferanten für Beschaffung, Vertrieb und Marge. Der Lieferant muss bei einem negativen Deckungsbeitrag für Komponenten wie Netzkosten oder Steuern aufkommen. Aus der eben beschriebenen Spanne ergibt das absolut einen Rohertrag bzw. einen Rohverlust von rund -176 € bis 98 € für den Lieferanten. Im Falle eines Zahlungsausfalls konnten mit einem Worst-Case-Kunden Kosten von rund 318 € (ohne gerichtliches Mahnverfahren) bis zu 776 € (mit gerichtlichem Mahnverfahren) aufgezeigt werden. Für die weitere Evaluierung werden die Einnahmen von 98 € für Verträge ohne Zahlungsausfälle angesetzt. Da nicht jeder Kunde einem Worst-Case-Kunden entspricht, wird der Verlust eines Zahlungsausfalls auf die Hälfte der unteren (auf den Worst-Case-Kunden bezogenen) Schranke mit 159 € gesetzt. Die Kosten liegen damit niedriger als der durchschnittliche Zahlungsrückstand, der zur Kündigung eines nicht-zahlenden Kunden führt.[169] Eine Veranschaulichung der Kosten innerhalb einer Konfusionsmatrix bietet Tabelle 19.

[169] Vgl. *Bundesnetzagentur und Bundeskartellamt*, 2021, S. 267, 433.

tatsächliche Werte	Vorhergesagte Werte		
	Kein Zahlungsausfall	Zahlungsausfall	
Kein Zahlungsausfall	98 € (Richtig Negativ)	-98€ (Falsch Positiv)	
Zahlungsausfall	-159 € (Falsch Negativ)	159 € (Richtig Positiv)	$\text{Sensitivität} = \frac{RP}{RP + FN}$
		$\text{PPV} = \frac{RP}{RP + FN}$	$F_1 = \frac{RP}{RP + \frac{1}{2}(FP + FN)}$

Tab. 19: Konfusionsmatrix mit den aus den Vorhersagen resultierenden Einnahmen bzw. Kosteneinsparungen (Werte mit negativem Vorzeichen sind entgangene Einnahmen bzw. Kosten)

Aus Sicht eines Energielieferanten sollten das normale Geschäft geschützt und Zahlungsausfälle verhindert werden. Deshalb sollte die Sensitivität eines Modells so hoch wie möglich sein, um Zahlungsausfälle zu erkennen und diesen effektiv vorzubeugen. Die alleinige Betrachtung der Sensitivität zur Beurteilung oder zur Auswahl eines Modells ist jedoch nicht zielführend. Grund dafür sind die berücksichtigten Variablen Richtig Positiv (RP) und Falsch Negativ (FN). Wenn nun alle Beobachtungen als Zahlungsausfall vorhergesagt werden, ist RP hoch, während FN den Wert 0 annimmt. Die daraus resultierende Sensitivität erreicht den Spitzenwert von 1, was ein perfektes Modell vermuten lässt. Tatsächlich schnellt bei diesem Vorgehen die Falsch Positiv (FP)-Rate hoch. Denn Verträge, die nicht in einem Zahlungsausfall gemündet hätten, würden abgelehnt werden. Damit gehen mögliche Erlöse verloren. Aus diesem Grund wird als zweite Kennzahl die PPV mitbetrachtet. Die PPV betrachtet den Anteil der korrekt als Zahlungsausfall vorhergesagten Verträge. Wenn viele Beobachtungen fälschlicherweise als Zahlungsausfall klassifiziert werden, sinkt die PPV. Im soeben beschriebenen Beispiel, bei dem die Sensitivität den perfekten Wert von 1 angenommen hat, da alle Beobachtungen als Zahlungsausfall vorhergesagt werden, sinkt hingegen die PPV drastisch bis auf den prozentualen Anteil aller Zahlungsausfälle im Datensatz. Andersherum verhalten sich die Kennzahlen, wenn von allen Beobachtungen nur vereinzelt Zahlungsausfälle klassifiziert werden. Dann schnellt die PPV hoch, da vereinzelt richtig klassifiziert wird (RP), jedoch sinkt die Sensitivität stark, da vermeintliche Zahlungsausfälle nicht mehr erkannt werden (FN). Somit reagieren die Kennzahlen Sensitivität und PPV auf einen etwaigen Fehler untereinander. Mit der Kombination beider Werte liegt der Fokus auf den vorhergesagten und den

tatsächlichen positiven Beobachtungen, da die Richtig Negativen (*RN*)-Werte in den Kennzahlen nicht berücksichtigt sind.[170]

Jedes Modell aus Kapitel 6 berechnet für eine gegebene Beobachtung eine Wahrscheinlichkeit, ob es sich um einen Zahlungsausfall handelt oder nicht. Die Wahrscheinlichkeit wird in Prozent angegeben und ist von 0 bis 1 skaliert. Demnach ist für eine Bewertung oder auch den späteren Einsatz die Berücksichtigung des Cutoffs notwendig. Mit dem Cutoff wird gesteuert, ab welcher Zahlungsausfallwahrscheinlichkeit eine Beobachtung als Zahlungsausfall bzw. nicht als Zahlungsausfall klassifiziert wird.

Für einen Vergleich mehrerer Modelle bei einem binären Klassifizierungsproblem wird oftmals die area under the receiver operating characteristic (AUROC), die Fläche unter der receiver operating characteristic (ROC)-Kurve, genutzt. Die ROC-Kurve zeigt die Veränderung zwischen Sensitivität und Falsch-Positiv-Rate für verschiedene Cutoff-Werte. Durch die Messung der Fläche unter dieser Kurve, deren Wert von 0 bis 1 reichen kann, kann die Modellbewertung stattfinden. Der Wert 1 repräsentiert ein perfektes Modell, bei dem die Kurve senkrecht steigt (Sensitivität von 1) und anschließend waagerecht zur Seite geht. Wenn das Modell gleichermaßen *RP*- und *FP*-Werte vorhersagt, steigt die Kurve diagonal und ergibt einen Wert von 0,5. Dies entspricht einem zufälligen Raten und das Modell bringt keinen Mehrwert.[171] Zuvor wurde hervorgehoben, dass für eine Bewertung die Kennzahlen PPV und Sensitivität herangezogen werden. Bei der ROC-Kurve wird hingegen die Falsch-Positiv-Rate anstelle der PPV genutzt; somit werden die zuvor ausgeschlossenen *RN*-Werte eingeschlossen. Zudem zeigte sich, dass ROC-Kurven durch unausgewogene Datensätze beeinflusst werden und das Ergebnis verzerrt ist. Aus diesen zwei Gründen wird für die Bewertung die Precision-Recall (PR)-Kurve genutzt (Precision = PPV, Recall = Sensitivität), die die zuvor herausgestellten und im F-Score enthaltenen Werte betrachtet. Ebenso wird die Fläche unter der PR-Kurve, die area under the precision-recall curve (AUPRC), gemessen. Dabei verläuft die ideale Kurve von der linken oberen Ecke (PPV = 1) bis zur oberen rechten Ecke (Sensitivität = 1), wodurch eine AUPRC von 1 gegeben ist. Anders als bei der AUROC gilt bei der AUPRC der Anteil der positiven Beobachtungen als Orientierungswert. Der Orientierungswert liegt nachfolgend bei rund 6,4 % bzw. 0,064, was dem Anteil der Zahlungsausfälle im Datensatz entspricht. Falls der Wert für die AUPRC darunter fällt, wird von einem Modell ohne Aussagekraft

[170] Vgl. *Powers*, *D.*, 2011, S. 38.
[171] Vgl. *Bewick*, *V.*, *Cheek*, *L.*, *Ball*, *J.*, 2004, S. 510 f.

ausgegangen. Es wurde zudem gezeigt, dass eine Kurve, die im PR-Raum dominiert, ebenso im ROC-Raum dominiert.[172]

Auf Basis der Testdatensätze werden für jedes Modell die AUPRC berechnet. Die Testdatensätze umfassen 30 % der Daten aus dem Gesamtdatensatz. Absolut sind dies 46.765 Beobachtungen. Sie wurden den Modellen vorenthalten, damit die Evaluierung auf ungesehenen Daten gemäß einer realen Umgebung durchgeführt werden kann. Das Modell, welches die höchste AUPRC erzielt, wird für eine wirtschaftliche Betrachtung ausgewählt. Darin wird die Wahl des optimalen Cutoffs unter Berücksichtigung der monetären Auswirkungen dargelegt.

7.2 Auswahl des besten Modells je Datensatz

Nachdem die Zahlungsausfallwahrscheinlichkeiten für jeden Datensatz durch jedes Modell vorhergesagt wurden, können die PR-Kurven mit den dazugehörigen AUPRC-Werten dargestellt werden. Für den ersten Datensatz Angebotsdaten inkl. Postleitzahlen musste die Modellierung der logistischen Regression aufgrund mangelnden Arbeitsspeichers abgebrochen werden. Somit stehen nur das ausgewählte Random Forest- und XGBoost-Modell für einen Vergleich bereit. Die PR-Kurve befindet sich in Abbildung 25. Allgemein sind die beiden Werte weit vom optimalen AUPRC-Wert 1 entfernt. Ebenso verläuft die Kurve nicht annähernd wie das Optimum waagerecht zur Seite (von Punkt (0,1) zu (1,1)). Die Modelle weisen damit ein großes Verbesserungspotenzial auf. Der AUPRC-Wert vom XGBoost-Modell liegt mit 0,2834 rund 15 % höher als der AUPRC-Wert vom Random Forest-Modell, der bei 0,2468 liegt. Damit stellt im direkten Vergleich das XGBoost-Modell das bessere Modell dar.

[172] Vgl. *Davis*, *J.*, *Goadrich*, *M.*, 2006, S. 1 f.

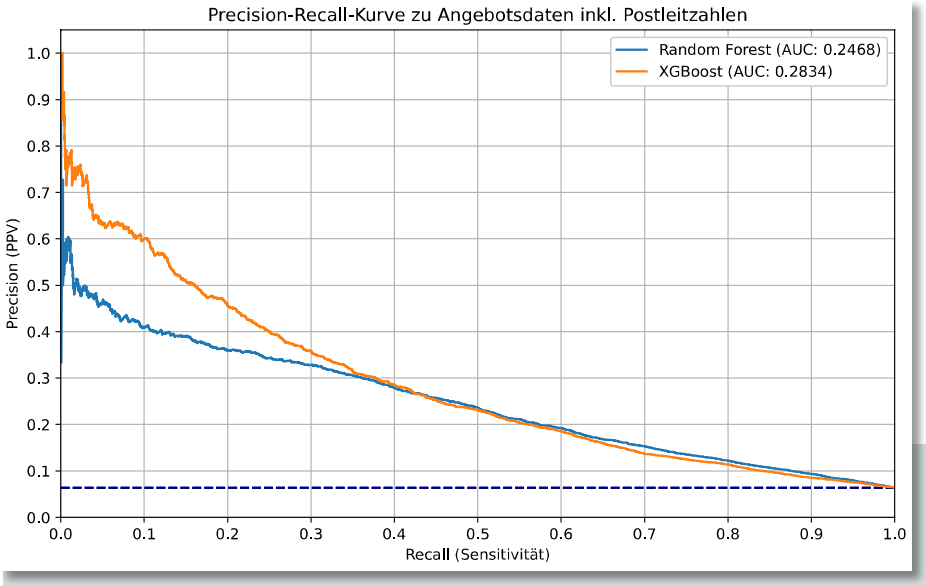


Abb. 25: Precision-Recall-Kurve zu Angebotsdaten inkl. Postleitzahlen

Für den Datensatz Angebotsdaten ohne Postleitzahlen ist die PR-Kurve in Abbildung 26 dargestellt. Die Kurve ähnelt dabei der vorherigen Kurve auf Basis der Angebotsdaten mit Postleitzahlen. Die niedrigsten AUPRC-Werte werden durch das Random-Forest-Modell mit 0,2520 und der logistischen Regression mit 0,2775 dargestellt. Mit einem AUPRC-Wert von 0,2776 stellt auch auf Basis der Angebotsdaten ohne Postleitzahlen das XGBoost-Modell das beste Ergebnis dar. Das XGBoost-Modell ist durch den geringeren Abstand nur noch rund 0,03 % besser als das nachfolgende logistische Regressionsmodell. Der Wert des XGBoost-Modells weicht rund 2 % vom erzielten Wert auf Basis der Angebotsdaten mit Postleitzahlen ab. Daraus folgt, dass die Postleitzahlen eine Verbesserung der Vorhersagekraft hervorrufen.

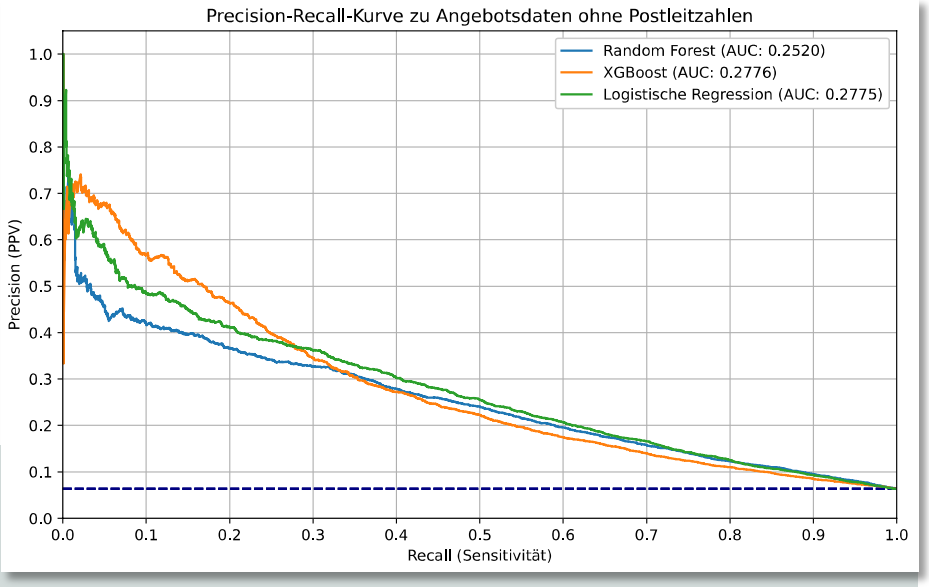


Abb. 26: Precision-Recall-Kurve zu Angebotsdaten ohne Postleitzahlen

Zum Datensatz Verhaltensdaten ist die PR-Kurve für alle drei Modelle in Abbildung 27 enthalten. Allgemein zeigt sich die größte optische Veränderung darin, dass die Kurve nun eher einer optimalen PR-Kurve gleicht. Für alle Modelle sind die AUPRC-Werte relativ hoch. Die logistische Regression erreicht einen AUPRC-Wert von 0,8846 und der Random Forest einen AUPRC-Wert von 0,9012. Auch hier ist das XGBoost-Modell führend und erzielt einen AUPRC-Wert von 0,9123. Somit wurde erneut eine Verbesserung von rund 1 % im Vergleich zum nächsthöheren Wert vom Random Forest-Modell erzielt.

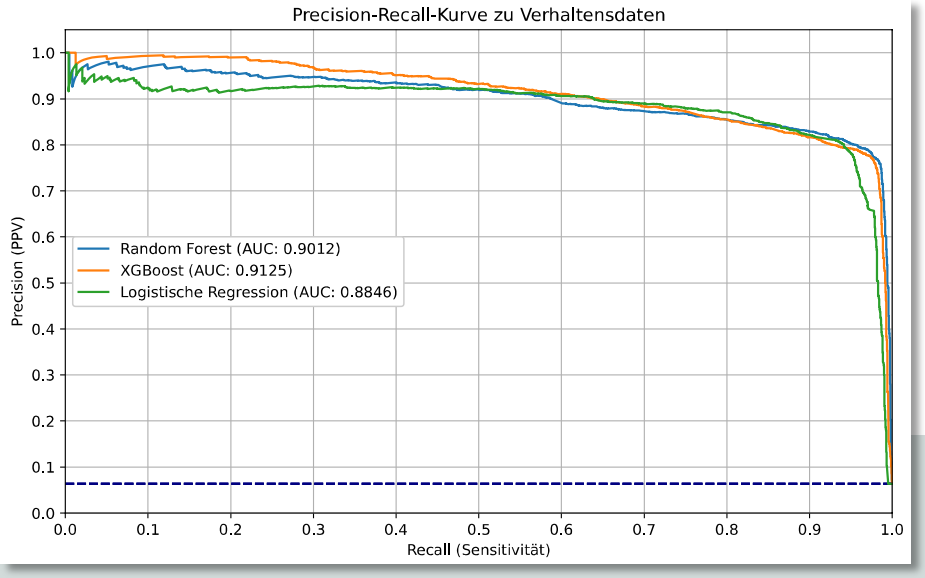


Abb. 27: Precision-Recall-Kurve zu Verhaltensdaten

Beim Vergleich der Modelle mithilfe der PR-Kurve kann die Hinzunahme eines Dummy-Modells von Vorteil sein. Das Dummy-Modell dient dem direkten Vergleich zwischen trainiertem Modell und einer konstanten Vorhersage *Zahlungsausfall* bzw. *kein Zahlungsausfall*. Exemplarisch ist für die beiden Datensätze, bei denen der Vergleich zwischen den drei vorhandenen Modellen stattfinden konnte, ein Dummy-Modell zur PR-Kurve hinzugefügt worden. Auf dem Datensatz Angebotsdaten ohne Postleitzahlen sagte das Modell immer *Zahlungsausfall* vorher. Auf dem Datensatz Verhaltensdaten wurde hingegen immer *kein Zahlungsausfall* vorhergesagt. Die beiden Kurven sind zum Vergleich in Abbildung 28 dargestellt. In beiden Kurven sind Diagonalen abgebildet, die zumindest beim Datensatz Angebotsdaten ohne Postleitzahlen in Abbildung 28a über den Kurven der trainierten Modelle liegen. Dort ist der erzielte AUPRC-Wert von 0,5316 höher als der AUPRC-Wert vom XGBoost-Modell von 0,2832. Rein auf die AUPRC zeigt sich, dass das Dummy-Modell besser als die trainierten Modelle ist. Dabei hat das Dummy-Modell immer *Zahlungsausfall* vorhergesagt. Ein Lieferant, der dieses Modell einsetzen würde, würde nach dem Dummy-Modell per se keine Verträge mehr annehmen. Anders ist es bei den Verhaltensdaten in Abbildung 28b. In dieser PR-Kurve wurde immer *kein Zahlungsausfall* vorhergesagt. Dort ist die Diagonale mit einem AUPRC-Wert von 0,5316 genauso hoch wie beim anderen

Dummy-Modell, jedoch niedriger als alle trainierten Modelle. Das Dummy-Modell ist demnach nicht den trainierten Modellen vorzuziehen.

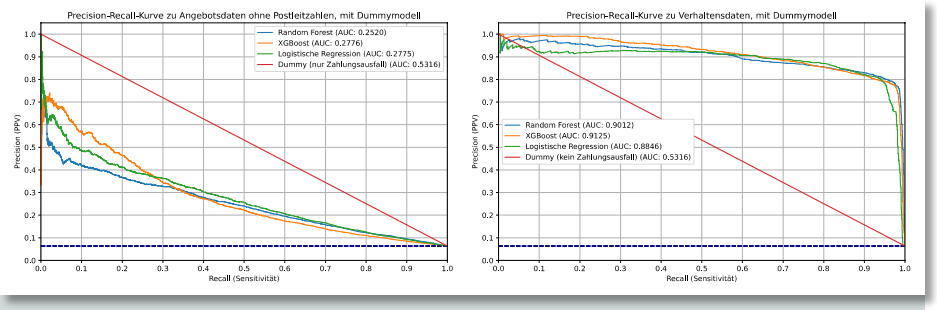


Abb. 28: Precision-Recall-Kurven mit Dummy-Modellen basierend auf zwei Datensätzen

Es stellt sich in Abbildung 28a insbesondere heraus, dass das vergleichsweise beste Modell immer die Vorhersage *Zahlungsausfall* trifft. Ein Lieferant würde demnach jeden Vertrag ablehnen, was auf ein nicht geschäftsunterstützendes Modell hindeutet. Damit kann gezeigt werden, dass die Nutzung der PR-Kurve nicht alleinstehend zur Bewertung eines guten Modells verwendet werden sollte. Deshalb wird der Matthews Correlation Coefficient (MCC) für einen weiteren Vergleich herangezogen. Der MCC ist in Formel 21 dargestellt. Er reicht von -1 (schlechtester Wert) bis 1 (bester Wert). Er berechnet sich aus der Kovarianz der wahren Werte y_{true} und der Modellvorhersagen y_{pred} geteilt durch das Produkt der Standardabweichungen von y_{true} und y_{pred}. Der MCC wird aus allen vier Werten der Konfusionsmatrix ermittelt und ist damit vollständiger als die Berücksichtigung von drei Werten in der PR-Kurve. Damit der MCC steigt, müssen hohe Werte sowohl für die richtig negativen als auch die richtig positiven Vorhersagen vorliegen.[173,174]

Formel 21: Matthews Correlation Coefficient (MCC)[175]

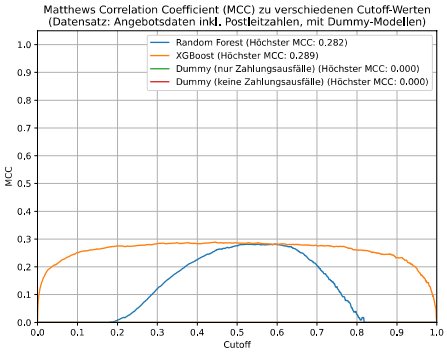
$$MCC = \frac{\mathrm{Cov}(y_{true,}, y_{true,})}{\sigma_{true} \cdot \sigma_{pred}} = \frac{RN\ RN - FP\ FN}{\sqrt{(RP + FP) \cdot (RP + FN) \cdot (RN + FP) \cdot (RN + FN)}} \tag{21}$$

Für alle Datensätze und Modelle wurde der MCC mit verschiedenen Cutoffs (ab welcher Zahlungsausfallwahrscheinlichkeit ein Zahlungsausfall vorhergesagt wird)

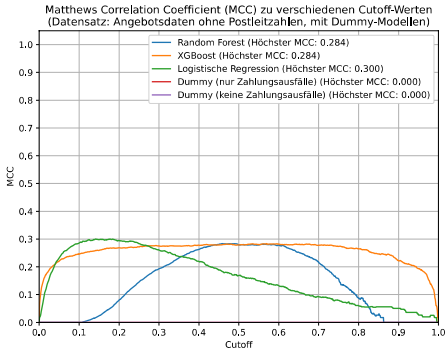
[173] Vgl. *Chicco, D., Tötsch, N., Jurman, G.*, 2021, S. 3, 15.
[174] Vgl. *Chicco, D., Jurman, G.*, 2020, S. 11.
[175] In Anlehnung an *Chicco, D., Tötsch, N., Jurman, G.*, 2021, S. 3.

berechnet. Diese sind in der Abbildung 29 zu sehen. Bei allen drei Datensätzen erreichen die Dummy-Modelle einen MCC von 0, ungeachtet dessen, ob die Modelle immer *Zahlungsausfall* oder immer *kein Zahlungsausfall* vorhersagen. Hingegen können alle trainierten Modelle einen positiven MCC aufweisen, die näher an 1 (dem besten Wert) streben.

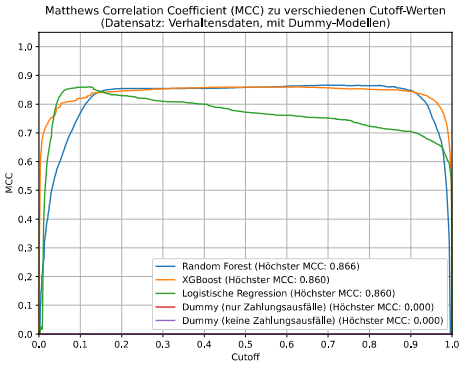
Mit dem MCC wurde nun validiert, dass die trainierten Modelle trotz einer geringeren AUPRC im Vergleich zu den Dummy-Modellen informativer im Hinblick auf alle Werte der Konfusionsmatrix sind und einen Mehrwert leisten können. Jedes Modell erzielte unter einem Datensatz den besten MCC. Lediglich das XGBoost-Modell erzielt für alle drei Datensätze die höchsten AUPRC. Für die wirtschaftliche Betrachtung werden zu allen Datensätzen die XGBoost-Modelle genutzt.



a) Datensatz: Anbegotsdaten inkl. Postleitzahlen



b) Datensatz: Anbegotsdaten ohne Postleitzahlen



c) Datensatz: Verhaltensdaten

Abb. 29: Matthews Correlation Coefficient (MCC) zu verschiedenen Cutoffs mit Dummy-Modellen basierend auf den vorhandenen Datensätzen

7.3 Wirtschaftliche Betrachtung

In Tabelle 19 wurden für eine Vorhersage die exemplarischen Kosten und Kosteneinsparungen sowie die Erlöse und Verluste gezeigt. Der Lieferant kann 159 € bei einem richtig vorhergesagten Zahlungsausfall einsparen bzw. ihm entstehen Kosten von 159 €, wenn dieser fälschlicherweise nicht vorhergesagt wird. Ein Vertragsverhältnis ohne Zahlungsausfall bringt einen Erlös von 98 € ein und, falls dieses Vertragsverhältnis fälschlicherweise als Zahlungsausfall klassifiziert wird, entsteht ein entgangener Erlös bzw. ein Verlust von 98 €. Je nach Cutoff werden die Vorhersagen unterschiedlich klassifiziert und in der Konfusionsmatrix einsortiert. Der Gewinn lässt sich für einen bestimmten Cutoff mit Formel 22 berechnen. Die Parameter werden wie folgt gesetzt: $z = 159$ und $v = 98$.

Formel 22: Gewinn bei Cutoff c, Zahlungsausfallkosten z und Vertragserlöse v

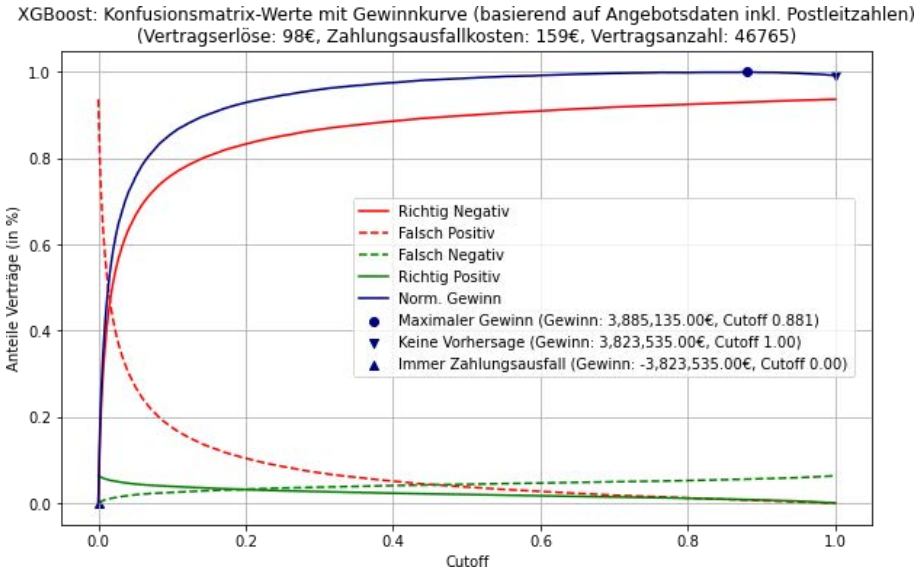
$$G_c = (RP_c - FN_c) \cdot z + (RN_c - FP_c) \cdot v \tag{22}$$

In Abbildung 30 sind auf Basis des Datensatzes Angebotsdaten inkl. Postleitzahlen zwei Abbildungen enthalten. Abbildung 30a zeigt die Verläufe der Konfusionsmatrix-Werte (RP, RN, FP, FN) mit steigendem Cutoff in der X-Achse. Die Y-Achse ist normiert und gibt den prozentualen Anteil der Verträge an. Die dunkelblaue Kurve zeigt den Gewinn auf Basis der Formel 22. Zusätzlich sind der maximale Gewinn, der Gewinn ohne Vorhersage bzw. ohne Modelleinsatz und der Gewinn mit konstanter Zahlungsausfall-Vorhersage mit Zugabe der jeweiligen Cutoffs separat markiert. In der Legende sind die Gewinne in Euro-Beträgen und die dazugehörigen Cutoffs aufgeführt. Sollte der Energielieferant kein Modell einsetzen, nimmt er auch alle Zahlungsausfälle mit auf und erzielt einen Gewinn von rund 3,824 Mio. €. Setzt der Energielieferant demgegenüber das Modell ein, welches für den maximalen Gewinn optimiert ist, dann erzielt der Lieferant einen Gewinn von rund 3,885 Mio. €. Dies ist eine relative Gewinnsteigerung von rund 1,6 % und eine absolute Steigerung des Gewinns von rund 61,6 Tsd. €. Der optimale Cutoff, mit welcher Zahlungsausfallwahrscheinlichkeit eine Beobachtung als Zahlungsausfall klassifiziert wird, liegt bei 88,1 % (0,881). In Abbildung 30b ist die Konfusionsmatrix mit den Modellvorhersagen abgebildet, die das Modell basierend auf dem Testdatensatz und dem Cutoff von 0,881 getroffen hat. Darin wurden insgesamt 378 Zahlungsausfälle richtig vorhergesagt und 2.577 Zahlungsausfälle fälschlicherweise angenommen. 43.511 Verträge

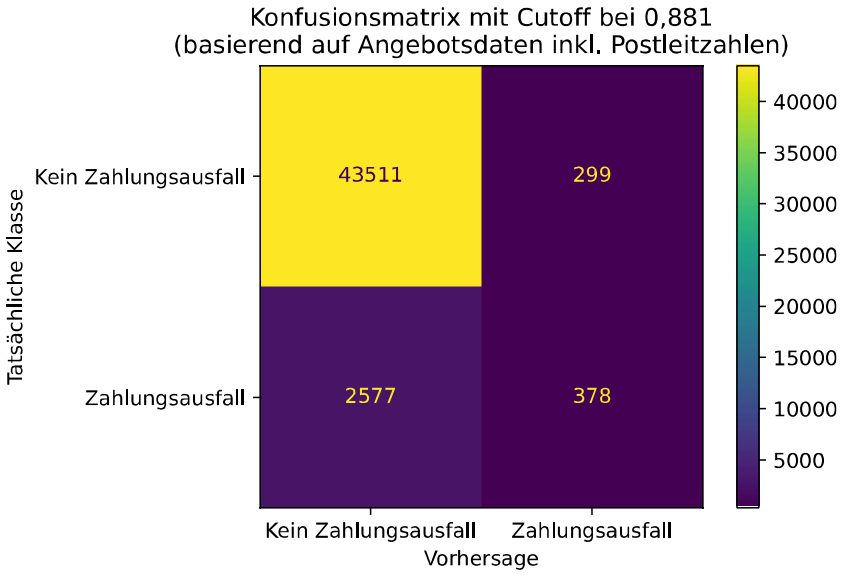
wurden richtig als kein Zahlungsausfall vorhergesagt, während 299 fälschlicherweise als Zahlungsausfall klassifiziert wurden. Demnach würden mit einem Modelleinsatz 1,4 % weniger Verträge angenommen werden. Rund 12,8 % der Zahlungsausfälle wurden richtig vorhergesagt und würden abgelehnt werden. Rund 0,7 % der Verträge würden fälschlicherweise abgelehnt werden. Die Daten sowie die Veränderungen mit und ohne Modelleinsatz sind in Tabelle 20 enthalten.

	Ohne Modell	Mit Modell	Steigerung mit Modell (gerundet)
Vertragsannahmen	46.765	46.088	-1,4 %
davon Zahlungsausfälle	2.955	2.139	-12,8 %
davon keine Zahlungsausfälle	43.810	42.611	-0,7 %
Gewinn (in Mio. €)	3,824	3,885	1,6 %

Tab. 20: Vertragsannahmen, Gewinn und der daraus resultierenden Steigerung mit und ohne Modelleinsatz (basierend auf dem Testdatensatz Angebotsdaten mit Postleitzahlen)



a) Normierte Gewinnkurve und Konfusionsmatrix-Werte



b) Konfusionsmatrixbei Cutoff 0,881

Abb. 30: Konfusionsmatrix und deren Werte mit Gewinnkurve basierend auf Formel 22 mit $z = 159$ und $v = 98$ (Datensatz Angebotsdaten inkl. Postleitzahlen)

Das Modell auf Basis der Angebotsdaten ohne Postleitzahlen ist nur mit einem Ausschnitt der Attribute des vorherigen Modells trainiert. Es erzielt auf dem Testdatensatz im Vergleich niedrigere Vertragsannahmen und höhere Gewinne. In Bezug auf den Gewinn ist ein Mehrgewinn von 1.180 € zu verzeichnen. Es wurden rund 3 % mehr Zahlungsausfälle vorhergesagt, jedoch stieg die Fehlerrate der fälschlicherweise als Zahlungsausfall klassifizierten Beobachtungen um 0,3 %. Der Cutoff mit 82 % Zahlungsausfallwahrscheinlichkeit liegt rund 6 % niedriger. Die Abbildungen und Tabellen befinden sich in Anhang 11.

Für den Testdatensatz der Verhaltensdaten wird das trainierte XGBoost-Modell herangezogen, um dieselben Werte unter Zunahme der Formel 22 zu berechnen. Es ist zu beachten, dass die Verhaltensdaten ein geschlossenes Vertragsverhältnis voraussetzen. Der Kunde wird demnach bereits aktiv mit Energie beliefert. Mit der PR-Kurve wurde eine AUPRC nahe dem optimalen Wert erreicht. Dies spiegelt sich auch in den Abbildungen 31a und 31b wider. Die Vorhersage der *RP* und *RN* ist nahezu perfekt. Dies liegt vor allem an der Variable Anzahl der Rücklastschriften, die sich als besonders informativ und wichtig während der Modellierung her-

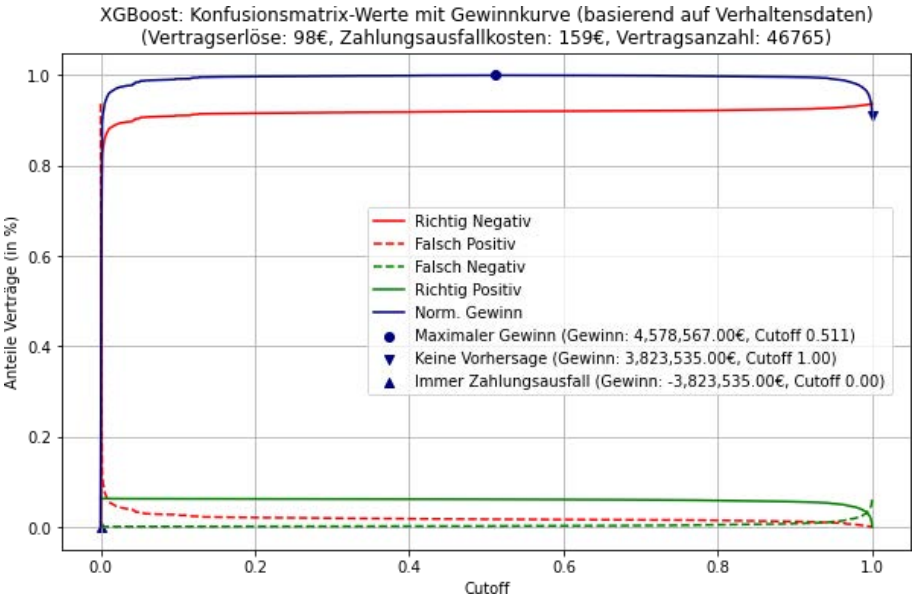
ausgestellt hat. Dieses Modell klassifiziert rund 96,8 % der Zahlungsausfälle korrekt, während lediglich 1,8 % fälschlicherweise als Zahlungsausfall klassifiziert werden. Würden die Verhaltensdaten bereits vor Vertragsschluss zur Verfügung stehen, wäre infolge einer Vertragsablehnung der Zahlungsausfälle eine Gewinnmaximierung von rund 19,7 % auf rund 4,579 Mio. € möglich. Die Ergebnisse sind in Tabelle 21 gegenübergestellt.

	Ohne Modell	Mit Modell	Steigerung mit Modell (gerundet)
Vertragsannahmen	46.765	43.117	-7,8 %
davon Zahlungsausfälle	2.955	95	-96,8 %
davon keine Zahlungsausfälle	43.810	43.022	-1,8 %
Gewinn (in Mio. €)	3,824	4,579	19,7 %

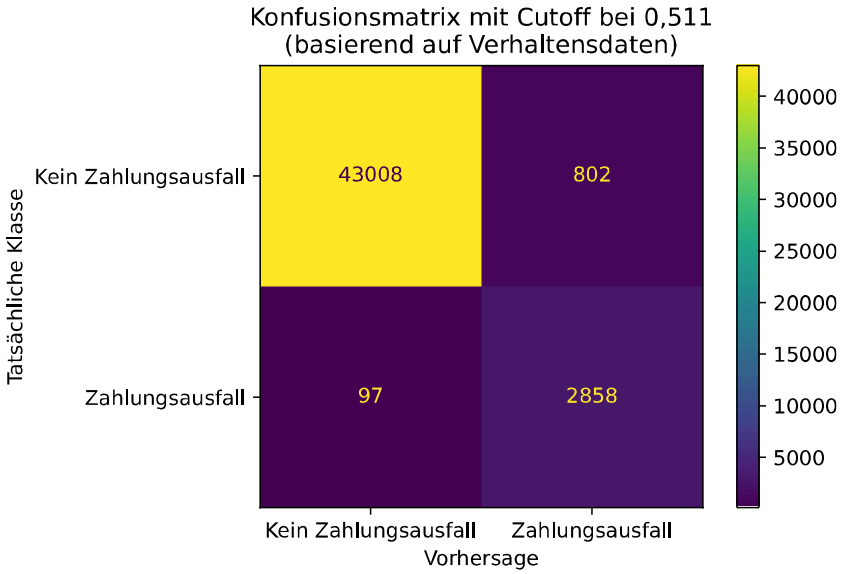
Tab. 21: Vertragsannahmen, Gewinn und der daraus resultierenden Steigerung mit und ohne Modelleinsatz (basierend auf dem Testdatensatz Verhaltensdaten)

Zusätzlich wurden exemplarisch höhere Vertragserlöse für Formel 22 mit $z = 159$ und $v = 200$ getestet. Die Erlöse für einen Vertrag wurden mehr als verdoppelt. Ein erfolgreiches Vertragsverhältnis bringt somit mehr Erlöse ein als ein richtig vorhergesagter Zahlungsausfall an Kosten einspart. Bei diesem konstruierten Fall konnte auf Basis der Angebotsdaten inkl. Postleitzahlen eine Gewinnsteigerung von rund 0,2 % erzielt werden.

Die Abbildungen zu den Vorhersagen mit den höheren Vertragserlösen befinden sich in Anhang 12.



a) Normierte Gewinnkurve und Konfusionsmatrix-Werte



b) Konfusionsmatrixbei Cutoff 0,511

Abb. 31: Konfusionsmatrix und deren Werte mit Gewinnkurve basierend auf Formel 22 mit $z = 159$ und $v = 98$ (Datensatz Verhaltensdaten)

8 Fazit

Energielieferanten sind durch den Wettbewerb im Energiemarkt einem stetigen Preisdruck ausgesetzt. Der Preisdruck wird insbesondere durch Discount-Lieferanten angefeuert, die eine radikale Preisgestaltung vornehmen, um günstige Preise anzubieten und somit Neukunden zu gewinnen. Dabei ist es für Lieferanten generell nicht unüblich, im ersten Belieferungsjahr negative Deckungsbeiträge hinzunehmen, um Neukunden zu gewinnen. Es zeigt sich, dass die Strategie der Discount-Lieferanten durch steigende Strom- und Gaspreise geschäftsgefährdend ist, da viele solcher Lieferanten durch den Anstieg kurzfristiger Beschaffungskosten Insolvenz anmelden mussten. Der Kostendruck kommt als weiterer Vektor hinzu, da mit den Erlösen für einen Energielieferanten bei 24,9 % für Strom und 49,4 % für Gas pro kWh noch Beschaffungs- und Servicekosten gedeckt werden müssen. Bei allen Maßnahmen, die dem Kostendruck entgegenwirken sollen, sind die IT sowie die Digitalisierung und Automatisierung von Prozessen die wesentlichen Schlüsselfaktoren. Es wurde gezeigt, dass insbesondere der Einsatz von Big Data in vielen Prozessschritten eine Chance für die weitere Prozessoptimierung ist. Eine Komponente des Kostendrucks eines Lieferanten stellen Kunden dar, die nicht der vertragsgemäßen Zahlung nachkommen. Es treten somit Zahlungsausfälle auf. Zahlungsausfälle sind in Teil- und Vollausfälle zu unterteilen. Bei Teilausfällen wird die Forderung später oder teilweise ausgeglichen, während bei einem Vollausfall die Wahrscheinlichkeit auf einen Forderungsausgleich u. a. mit dem Alter der Forderung oder einer Privatinsolvenz sehr gering ist. Wettbewerblich agierende Energielieferanten kündigten Kunden bei Zahlungsrückständen von durchschnittlich 176 € (Strom) und 170 € (Gas). Es blieb dabei offen, ob die verbrauchte Energiemenge bereits in den Forderungen inkludiert ist. Mit einem konstruierten Worst-Case-Kunden wurde gezeigt, dass die Prozesse im laufenden Vertragsverhältnis gestört werden, sodass bei einem Zahlungsausfall Forderungs- und Prozesskosten in Höhe von rund 318 € ohne gerichtliches Mahnverfahren anfallen. Mit einem gerichtlichen Mahnverfahren sind Forderungs- und Prozesskosten in Höhe von 776 € errechnet worden.

Ein Datensatz eines rein digitalen wettbewerblichen Energielieferanten, der bis Juni 2021 am deutschen Energiemarkt tätig war, lag für die Analyse vor. Der Datensatz umfasste insgesamt 155.926 Vertragsdaten mit 31 Datenfeldern. Davon sind 18 Felder dem Zeitpunkt der Angebotsabgabe durch den Kunden vor Vertragsschluss zuzuordnen. Die restlichen 13 Datenfelder sind Verhaltensdaten und

umfassen Daten, die während des Vertragsverhältnisses angefallen sind. Der digitale Lieferant verzeichnet die meisten Verträge mit Kunden, deren Alter zwischen 25 und 40 Jahren sowie zwischen 50 und 60 Jahren liegt. Über 75 % aller Kunden schließen ihren Vertrag über ein Vergleichsportal ab. Der Samstag ist der Tag, an dem die wenigsten Verträge abgeschlossen werden. Kunden, die den Vertrag über ein Vergleichsportal nutzen, sind besonders affin für die Lieferantenwechselprozesse und die darin enthaltene automatisierte Kündigung des Altlieferanten. Je nach Vertriebspartner wurden höhere Vertragsanteile und niedrigere Zahlungsausfallanteile (Verivox) sowie niedrigere Vertragsanteile und höhere Zahlungsausfallanteile festgestellt (Check24 und eigene Bestellstrecke). Rund 94 % aller Kunden im vorliegenden Datensatz zahlen die Forderungen des Lieferanten vertragsgemäß. Lediglich 0,34 % münden in der Abgabe an ein externes Inkassounternehmen. Die Kunden erhalten durchschnittlich am 155. Tag in Belieferung ihre erste Mahnung. Dabei reicht die Spanne von 75 bis zu 230 Tagen. Besonders jüngere Menschen zwischen 20 und 45 Jahren sind überproportional mit Zahlungsausfällen vertreten. Ab 50 Jahren sind die Zahlungsausfälle unterproportional vertreten. Eine Besonderheit ist bei Kunden mit einem Alter von 120 Jahren vorhanden, da diese Kunden sehr wenige Vertragsanteile haben, jedoch einen relativ großen Anteil der Zahlungsausfälle ausmachen.

Die Daten wurden hinsichtlich der Ausreißer und einiger fehlender Daten behandelt. Diverse Variablen wurden angepasst und kategoriale Datenfelder in Dummy-Variablen transformiert. Die Modellierung basierte auf den drei Datensätzen Angebotsdaten mit Postleitzahlen, Angebotsdaten ohne Postleitzahlen und Verhaltensdaten. Zur Modellierung wurden logistische Regression, Random Forest und XGBoost verwendet. Unter den Angebotsdaten stellte sich das Attribut Alter in allen Modellen heraus. Ebenso sind einige BICs mit einer höheren Zahlungsausfallwahrscheinlichkeit behaftet. Nur eine BIC steht im Zusammenhang mit wenigen Zahlungsausfällen. Bei Kunden, die über 110 Jahre alt sind, steigt die Zahlungsausfallwahrscheinlichkeit um rund 91 % enorm. Ebenfalls sind die Vertriebspartner sehr aussagekräftig. Unter den Verhaltensdaten sind die Anzahl der Rücklastschriften oder Anzahl der manuellen Überweisungen durch den Kunden besonders für einen Zahlungsausfall ausschlaggebend. Die Attribute Postleitzahl, E-Mail-Domain, Abschlusswochentag oder die Anzahl der Abschlagsplanverminderungen oder -erhöhungen stellten sich als nicht informationsgebend heraus.

In der Evaluierung wurden mithilfe der AUPRC die XGBoost-Modelle auf Basis der verschiedenen Datensätze als beste Modelle ausgewählt. Das XGBoost-Modell erreichte eine AUPRC von 0,2834 auf den Angebotsdaten. Der Wert ist weit vom

optimalen Wert 1 entfernt und stellt damit eine Unsicherheit bei den Vorhersagen dar. Das XGBoost-Modell auf Basis der Verhaltensdaten erzielte eine AUPRC von 0,9123 und liegt somit deutlich näher an dem optimalen Wert 1. Durch Hinzunahme von Dummy-Vorhersagen und Berechnung des MCC wurde gezeigt, dass die Modelle mit ihren Vorhersagen einen Mehrwert bringen. Unter Berücksichtigung der Vertragserlöse von 98 € und der Zahlungsausfallkosten von 159 € wurde ein gewinnoptimierter Cutoff für die Klassifizierung eines Zahlungsausfalls gewählt, ab welcher Zahlungsausfallwahrscheinlichkeit ein Zahlungsausfall als solcher klassifiziert wird. Die Evaluierung fand mit separaten Testdatensätzen über 46.765 Beobachtungen statt, die während des Trainings vorenthalten wurden. Auf Basis der Angebotsdaten und des damit trainierten XGBoost-Modells konnte eine Gewinnsteigerung von 1,6 % erzielt werden, wenn Verträge, die als Zahlungsausfall klassifiziert wurden, abgelehnt würden. Insgesamt wurden rund 12,8 % aller Zahlungsausfälle verhindert, während rund 0,7 % fälschlicherweise als Zahlungsausfall klassifiziert wurden. Mit einer Rechnung mit höheren Vertragserlösen von 200 € und gleichbleibenden Zahlungsausfallkosten von 159 € konnte eine Gewinnsteigerung von 0,2 % gemessen werden. Daraus folgt, dass mit den Angebotsdaten und vor Vertragsschluss keine hoch-präzise Vorhersage von Zahlungsausfällen stattfindet. Mit dem Modell wurde jedoch bewiesen, dass mit der Erkennung von rund einem Achtel aller Zahlungsausfälle eine Gewinnoptimierung stattfindet.

Auf Basis der Verhaltensdaten wurden rund 96,8 % aller Zahlungsausfälle richtig und nur 1,8 % aller Verträge fälschlicherweise als Zahlungsausfall klassifiziert. Die Verhaltensdaten, die während eines Vertragsverhältnisses gesammelt werden, sind für eine Klassifizierung durchaus wertvoll. Eine Überwachung der Verhaltensdaten zur Ableitung weiterer Maßnahmen bietet sich an.

9 Ausblick

Mit den entwickelten Modellen auf Basis der Angebotsdaten konnte der Gewinn um rund 62 Tsd. € gesteigert werden. Dabei sind die Gewinne und Kosten abhängig vom Energielieferanten. Entsprechend würde sich der Cutoff verschieben, ab welcher Wahrscheinlichkeit ein Zahlungsausfall klassifiziert werden würde. Ebenso handelt es sich bei der Berechnung der Vertragserlöse um die obere Grenze der Gewinne, die im ersten Belieferungsjahr erzielt werden können. Es könnte ein zusätzliches Modell entwickelt werden, welches die Daten einer Kundenabwanderung wie die Dauer der Vertragslaufzeit vorhersagt. Die Ergebnisse beider Modelle könnten kombiniert werden, um eine Schärfung der Zahlungsausfallvorhersagen zu bewirken oder eine Gewichtung der Kunden mit ihren potenziellen Vertragserlösen vorzunehmen. Zudem besteht die Möglichkeit, externe Auskunfteien mit Bonitätsprüfungen zu beauftragen. Die Auskunfteien stellen für 0,60 € bis 1,20 € (abhängig vom ausgehandelten Vertrag) eine Bonitätsauskunft oder *Chancen-Score* bereit.[176] Bei der alleinigen Verwendung einer externen Auskunftei muss ein Energielieferant ebenfalls einen Cutoff definieren, ab welchem ein Vertrag angenommen oder abgelehnt wird. Die Entscheidung eines geeigneten Cutoffs muss demnach genauso wie bei den entwickelten Modellen eruiert werden. Daten der Auskunfteien können sich aufgrund der großen Datenmengen als genauer herausstellen. Deshalb könnte ein hybrider Einsatz zum Einsparen von Kosten helfen, indem die entwickelten Modelle eine Vorabprüfung durchführen und nur Kunden mit einem gewissen Zahlungsausfallrisiko zusätzlich bei der externen Auskunftei überprüft werden.

Auf Basis der Verhaltensdaten wurden sehr gute Vorhersagen beobachtet. Rund 96,8 % aller Zahlungsausfälle wurden korrekt als solche vorhergesagt. Erst wenn der Kunde in Belieferung ist und ein Vertrag geschlossen wurde, können Verhaltensdaten gesammelt werden. Zumeist besteht eine Mindestvertragslaufzeit, die beide Vertragsparteien, und demnach auch der Lieferant, einhalten müssen. Es wäre zu testen, ob sich eine Verringerung der Mindestvertragslaufzeiten positiv auf das Geschäftsergebnis auswirken könnte, wenn Kunden mit einer gewissen Zahlungsausfallwahrscheinlichkeit in Belieferung gekündigt würden. Auch könnte für Kunden, die ein hohes Zahlungsausfallrisiko haben, ein gesondertes Mahnverfahren gewählt werden, welches weniger Mahnstufen beinhaltet, um die offene Forderung weitestgehend gering zu halten.

[176] Vgl. Anhang 2.

Eine Erweiterung der Modelle unter Analyse und Hinzunahme weiterer Variablen ist möglich. Die entwickelten Modelle treffen ihre Aussage für eine binäre abhängige Variable. Daher könnten weitere Erkenntnisse und Optimierungen mit einer höheren Trennschärfe erlangt werden. Die höhere Trennschärfe kann durch weitere Klassen erreicht werden, die dann die Schwere des Zahlungsausfalls beinhalten. Dazu wären die Modelle Random Forest und XGBoost geeignet. Bei der logistischen Regression müsste von der geordneten logistischen Regression Gebrauch gemacht werden. Neben der Klassifizierung ist eine weitere Analyse und Entwicklung von Regressionsmodellen mit anderen Algorithmen möglich, die den direkten Gewinn bzw. Verlust eines Kunden vorhersagen kann.

Auf Basis der Daten eines wettbewerblichen Energielieferanten wurden Erkenntnisse durch die Analysen gewonnen und gewinnoptimierende Modelle entwickelt. Rund 26 % aller Stromkunden und 17 % der Gaskunden befinden sich in der Grundversorgung.[177] Die Kundengruppe in der Grundversorgung umfasst Kunden, die einen Neueinzug hatten und keinen separaten Energielieferanten beauftragt haben, oder Kunden, die ihren bisherigen Liefervertrag ohne Anschlussvertrag auslaufen ließen. Diese Kundengruppe war im vorliegenden Datensatz nicht enthalten. Im vorliegenden Datensatz waren lediglich Kunden- und Vertragsdaten von Kunden enthalten, die aktiv einen Auftrag zur Belieferung abgegeben haben. Daher bietet sich eine Untersuchung der Kunden- und Vertragsdaten an, damit eine Übertragbarkeit der Erkenntnisse und der Modelle bestätigt oder ausgeschlossen werden kann.

[177] Vgl. *Bundesnetzagentur und Bundeskartellamt*, 2021, S. 255.

Anhang

Anhang 1: Prozesskostenrechner

juris Das Rechtsportal — DeutscherAnwaltVerein

Grundangaben

Streitwert	Rechtsstand	MwSt.-Satz	
292,64 €	ab 2021	19 %	☐ Außergerichtliche Vertretung
Anzahl Mandanten: 1	Anzahl Gegner: 1		☑ Gerichtliche Vertretung ☐ 2. Instanz ☑ 3. Instanz

Prozesskosten

		1. Instanz
Eigene Anwaltskosten		
Verfahrensgebühr Nr. 3100, 1008 VV RVG:	1,3	63,70 €
Anrechnung:		0,00 €
Terminsgebühr Nr. 3104 VV RVG:	1,2	58,80 €
Einigungsgebühr Nr. 1003 f. VV RVG:		0,00 €
Auslagen Nr. 7001 u. 7002 VV RVG:		20,00 €
Sonstige Kosten Nr. 7000 u. 7003 ff. VV RVG:		0,00 €
MwSt. %:	19	27,08 €
Zwischensumme:		**169,58 €**
Gegnerische Anwaltskosten		
Verfahrensgebühr Nr. 3100, 1008 VV RVG:	1,3	63,70 €
Terminsgebühr Nr. 3104 VV RVG:	1,2	58,80 €
Einigungsgebühr Nr. 1003 f. VV RVG:		**0,00 €**
Auslagen Nr. 7001 u. 7002 VV RVG:		20,00 €
Sonstige Kosten Nr. 7000 u. 7003 ff. VV RVG:		0,00 €
MwSt. %:	19	27,08 €
Zwischensumme:		**169,58 €**
Gerichtskosten KV GKG:	3,0	114,00 €
Summe Prozesskosten:		**453,16 €**

Gesamtkosten

Eigene Anwaltskosten:	169,58 €
Fremde Anwaltskosten:	169,58 €
Gerichtskosten:	114,00 €
Gesamtsumme:	**453,16 €**

Abb. 32: Prozesskosten zur offenen Forderung von einem Worst-Case-Kunden[178]

[178] *Juristische Informationssystem für die Bundesrepublik Deutschland (juris GmbH)*, 2021, mit einem Streitwert von 292,64 €, Betrachtung des Prozesses in 1. Instanz sowie Nutzung der Standardwerte.

Anhang 2: Anfrage zum Scoring bei der Schufa Holding AG

10.11.21, 11:00 E-Mail – Stecker, Rouven – Outlook

AW: Informationen zur Bonitätsauskunft für Energieversorger im Rahmen einer wissenschaftlichen Ausarbeitung

[redacted]@schufa.de>
Di, 09.11.2021 11:36
An: Stecker, Rouven <rouven.stecker@fom-net.de>

1 Anlagen (679 KB)
1703_B2B_Broschure_Online.pdf;

Hallo Herr Rouven,

viel Erfolg bei der Arbeit! Hier meine Antworten:

- Welche Informationen sind für eine Auskunft notwendig?

- **B2C**: Personenstammdaten (nach Möglichkeit mit Geburtsdatum!), Anschrift (bei „frischem" Umzug nach Möglichkeit auch die Voranschrift)
  Je besser die Qualität der Anfragedaten, je besser die Auskunft, d.h. z.B. die Angabe des Geburtsdatums erhöht die Trefferquote.
  B2B: Komplizierter ist es bei **Kleingewerbetreibenden**, aber auch hier gilt, je mehr qualifizierte Daten geliefert werden (z.B. Inhaberinformation etc.), je besser auch die Trefferquote und die Auskunftsqualität. Ein gutes Beispiel ist die Gastronomie, wo wegen der großen Fluktuation eigentlich die Angabe des Inhabers fast unerlässlich ist, um gute Bonitätsinformationen beziehen zu können. Da über die vom Versorger angebotenen Tarife nicht immer eindeutig zu erkennen ist, ob es sich bei dem Kunden um eine Privatperson oder um einen Gewerbetreibenden handelt, ist es auch wichtig, **dass diese Information beim Onboarding von Neukunden zusätzlich erhoben wird.** Falls dies ausbleibt und beispielsweise im Vornamensfeld einer B2C-Anfrage ein Firmenname übermittelt wird, läuft die Bonitätsprüfung in der Regel ins Leere. Bei B2B-Anfragen zu größeren Unternehmen mit Handelsregisternummer etc. ist die Trefferquote in der Regel fast 100 Prozent und die Auskunftsqualität sehr gut.

- Welche Informationen erhält ein Energielieferant?

- **B2C:** Versorger erhält bei SCHUFA hinterlegte Personenstammdaten, evtl. gespeicherte Zahlungsstörungen und optional einen „Chancen-Score". Da beim reinen Energievertrieb das Ausfallrisiko (inkl. Ausfallhöhe) in der Regel nicht so hoch ist (wie z.B. beim Bundle-Vertrieb in der Telekommunikation) und gerade überregionale Unternehmen sehr stark vertriebsgetrieben sind, werden oftmals Produkte bei SCHUFA gewählt, die es ermöglichen, Verbraucher auch dann anzunehmen, wenn bereits Zahlungsstörungen zum Kunden vorhanden sind. Das berechnete energiespezifische Ausfallrisiko ist dann natürlich die wesentliche Entscheidungsbasis. Ein typisches Produkt dafür ist unser sogenannter Chancenscore, bei dem **nur** zu Kunden mit vorhandenen Zahlungsstörungen ein Scorewert (und somit ein energiespezifisches Ausfallrisiko) berechnet wird. Ziel ist eine maximale Annahmequote ohne hohe Ausfallkosten. **Bitte beachten Sie**: Speziell dieser Score wird nicht auf unserer Webseite beschrieben.
  B2B: Hier wird grundsätzlich immer ein Bonitätswert beauskunftet (oft ein Wert, der am Schulnotensystem orientiert ist). Auch Zahlungsstörungen werden detailliert beauskunftet. Ansonsten hängt der Detaillierungsgrad der Auskunft vom gewählten Produkt ab. Ich habe Ihnen zur Ansicht eine **nicht mehr ganz aktuelle** SCHUFA-B2B-Broschüre beigefügt. Auf Seite 9 ist gut zu sehen, was mit Detaillierungsgrad gemeint ist.

- Welche Kosten entstehen durch die Abfrage bei der SCHUFA? Gibt es Festpreise, Preisbänder oder können Sie mir einen Preisrahmen nennen?

- **B2C**: Eine Auskunft kostet zwischen 60 Cent und € 1,20. Der Preis hängt z.B. ab von der genutzten Schnittstelle, vom genutzten Produkt (mit/ohne Score), von der angefragten Menge etc.

https://outlook.office365.com/mail/inbox/id/AAQkADhlN2I0ZGY3LTdhYTAtNGQ5Mi05M2ExLWQ2YmEyZWNjN2E3OQAQAJhy8%2F3KbURLiRp… 1/2

a) Seite 1 von 2

10.11.21, 11:00 E-Mail – Stecker, Rouven – Outlook

B2B: Die Preisspannbreite ist etwas größer und hängt sehr stark vom Detaillierungsgrad der Auskunft ab. Ganz grob kostet die Auskunft zwischen € 5.- und € 15.-. Bitte haben Sie Verständnis dafür, dass ich hier nicht weiter ins Detail gehen kann.

Beste Grüße

[redacted]

Von: Stecker, Rouven <rouven.stecker@fom-net.de>
Gesendet: Dienstag, 9. November 2021 09:20
An: [redacted]@schufa.de>
Betreff: Informationen zur Bonitätsauskunft für Energieversorger im Rahmen einer wissenschaftlichen Ausarbeitung

Hallo Herr [redacted].

soeben habe ich mit einem Ihrer Kollegen telefoniert, woraufhin ich an Sie verwiesen wurde. Zurzeit schreibe ich meine Masterarbeit über Zahlungsausfälle bei wettbewerblich agierenden Energielieferanten. Neben der Behandlung, wie schmerzhaft ein Zahlungsausfall sein kann, möchte ich gerne einen Abschnitt über die Möglichkeiten zur Verhinderung eines solchen Ausfalls beschreiben.
Zu Ihrem Scoring-Modell, welche Werte einer Privatperson oder einem Geschäftskunden zugeordnet werden können, finde ich einige Hinweise auf Ihrer Website. Leider kann ich keine Informationen über die Kosten beim Einsatz Ihrer Auskunftei finden, die für meine Ausarbeitung und Gegenrechnung (Was kostet ein Zahlungsausfall und was kostet dessen Verhinderung) sehr hilfreich sind.

Können Sie mir an dieser Stelle weiterhelfen und Informationen zum Einsatz Ihrer Auskunftei bei Energielieferanten bereitstellen?
Idealerweise würden die folgenden Punkte dadurch beantwortet werden:
- Welche Informationen sind für eine Auskunft notwendig?
- Welche Informationen erhält ein Energielieferant?
- Welche Kosten entstehen durch die Abfrage bei der SCHUFA? Gibt es Festpreise, Preisbänder oder können Sie mir einen Preisrahmen nennen?

Bei weiteren Fragen stehe ich gerne zur Verfügung. Gerne auch telefonisch.

Durch Ihre Informationen würden Sie der wissenschaftlichen Betrachtung dieses Themas sehr weiterhelfen. Vielen Dank vorab.

Viele Grüße
Rouven Stecker

https://outlook.office365.com/mail/inbox/id/AAQkADhlN2I0ZGY3LTdhYTAtNGQ5Mi05M2ExLWQ2YmEyZWNjN2E3OQAQAJhy8%2F3KbURLiRp… 2/2

b) Seite 2 von 2

Abb. 33: Ausschnitt der Anfrage mit Antwort bei der SCHUFA Holding AG zum Einsatz des Scorings bei Energielieferanten (mit Schwärzung personenbezogener Daten)

Anhang 3: SQL-Skript zum Datenexport aus Quelldatenbank

Listing 4: SQL-Skript zum Datenexport aus der Quelldatenbank

```
1  select
2  contract.section_ as section,
3  salutation.name_ as salutation,
4  date_part('year', age('2021-05-31'::date, bp.dateofbirth_)) as age,
5  substring(bp.email_ from '@(.*)$') email,
6  case when phone.businesspartner_ is null then 'False' else 'True' end
      as hasPhonenumber,
7  contractaddress.postalcode_ as deliveryPostalcode,
8  bpaddress.postalcode_ as ownerPostalcode,
9  case when
10    contractaddress.postalcode_ = bpaddress.postalcode_
11    and
12          (contractaddress.street_ = bpaddress.street_
13          or contractaddress.street_ = regexp_replace(bpad-
               dress.street_, '^(.*) str\.$', '\1straße')
14          or bpaddress.street_ = regexp_replace(contrac-
               taddress.street_, '^(.*) str\.$', '\1straße')
15          or contractaddress.street_ = regexp_replace(bpad-
               dress.street_, '^(.*) Str\.$', '\1 Straße')
16          or bpaddress.street_ = regexp_replace(contrac-
               taddress.street_, '^(.*) Str\.$', '\1 Straße')
17          or bpaddress.housenumber_ = contractaddress.housenumber_)
18    then 'True' else 'False' end as isPostalAndDeliveryAddressIndenti-
      cal,
19 bankacc.bic_ as bic,
20 case when payMethod.id_ = 10 then 'Paypal' else 'SEPA' end as pay-
      mentmethod,
21 distributionpartner.name_ as distributionpartner,
22 case when oldsupplier.name_ is null then 'False' else 'True' end as
      hasOldSupplier,
23 contract.invoicinginterval_ as invoiceinterval,
24 flba.value as bonusvalue,
25 powerconsprog.prognosis_ as consumptionprognosis,
26 snapcost.energycost_ as energycostkwh,
27 snapcost.monthlycost_ as basepricemonthly,
28 extract(dow from contract.contractconclusiondate_) as conclusion-
      Weekday, --// 0= sunday, 1=monday...
29
30 -- Erstmalig gemahnt in Tagen nach Lieferbeginn
31 case when extract(DAYS from (firstIntegerAdditional3.minia3date -
      contract. periodstart_)) > 365
32    then null
33    else extract(DAYS from (firstIntegerAdditional3.minia3date - con-
       tract.periodstart_))
34 end as first_dunning_in_days_after_supplybegin,
35
36 -- Höchster jemals erlangter integerAdditional3
37 highestIntegerAdditional3.maxia3 as highestdebtstatus,
38
39 -- Lieferzeitraum
```

```
40  case when
41     (extract(DAYS from ('2021-05-31'::date - contract.periodstart_)) >
          365 and contract.periodend_ is null)
42  or
43     (extract(DAYS from (contract.periodend_ - contract.periodstart_))
          > 365)
44  then
45     365
46  else
47     case when contract.periodend_ is null
48  then extract(DAYS from ('2021-05-31'::date - contract.periodstart_))
49  else
50     extract(DAYS from (contract.periodend_ - contract.periodstart_))
51     end
52  end as supplyinterval_in_days,
53
54  -- Belieferung aktiv?
55  case when lco.periodend_ is null
56     then 'True'
57     else 'False'
58  end as is_active,
59
60
61  ---- Verhaltensattribute nach 365 Tagen
62
63  -- Anzahl Rückrufformular
64  callbacktask.callcount,
65  -- Anzahl geschriebene E-Mails
66  emailtask.emailcount,
67  -- Anzahl genutzte Bankdaten
68  fsepachange.bankacccount as bankAccountCount,
69  -- Eingegebene Zählerstände vom Kunden
70  meterreading.meterreadingcount,
71  -- Abschlagsplanänderungen nach unten
72  payplanreduced.reducedcount as payplanReducedCount,
73  -- Abschlagsplanänderungen nach oben
74  payplanincreased.increasedcount as payplanIncreasedCount,
75  -- Anzahl Rücklastschriften
76  separeversed.countSepaReversed as sepaReversedCount,
77  -- Eingang der Kündigung/Wechselanfrage neuer Lieferant in Tagen
78  case when extract(DAYS from (gpke.firstCancellationDate - con-
       tract.periodstart_)) <= 365
79        then extract(DAYS from (gpke.firstCancellationDate - con-
             tract.periodstart_))
80        else null
81  end as cancellationReceivedAfterSupplybeginInDays,
82  -- Anzahl manueller Überweisungen
83  manualsepa.countManualSepa as manualSepaCount
84
85 from
86  fl_contract contract
87  left join fl_contractpartner cp1
88     on cp1.contract_ = contract.id_ and cp1.role_ = 1
89  left join log_businesspartner bp -- den ersten Datensatz nutzen, um
       Veränderungen nicht zu betrachten
90     on cp1.businessPartner_ = bp.id_  and bp.created_log_ = (select
          min(created_log_) from log_businesspartner lb where
          lb.email_ is not null and lb.id_ = bp.id_)
91  left join fl_businesspartnersalutati salutation
```

```
92      on salutation.id_ = bp.salutation_
93  left join (select businesspartner_, min(created_log_)
94                    from log_phonenumber lp
95                    where lp.creationdate_ is null
96                    group by businesspartner_) phone
97      on bp.id_ = phone.businesspartner_
98
99  -- Ist der Vertrag nach 365 Tagen noch aktiv gewesen?
100 left join log_contract lco
101     on contract.id_ = lco.id_
102       and lco.created_log_ = (select max(created_log_) from log_con-
          tract where id_ = contract.id_ and created_log_ < con-
          tract.periodstart_ + interval '365 days')
103
104
105 left join fl_contractpartner cpDistribPart
106     on cpDistribPart.contract_ = contract.id_ and cpDistribPart.role_
        = 12 and cpDistribPart.periodend_ is null
107 left join fl_businesspartner distributionpartner
108     on cpDistribPart.businessPartner_ = distributionpartner.id_
109
110 left join fl_contractpartner cpOldSupplier
111     on cpOldSupplier.contract_ = contract.id_ and cpOldSupplier.role_
        = 13 --and cpOldSupplier.periodend_ is null
112 left join fl_businesspartner oldsupplier
113     on cpOldSupplier.businessPartner_ = oldsupplier.id_
114
115 left join fl_pointofdelivery pointofdelivery
116     on pointofdelivery.id_ = contract.pointofdelivery_
117 left join fl_address contractaddress
118     on contractaddress.id_ = pointofdelivery.address_address_
119
120 left join log_businesspartneraddress fbpa
121     on fbpa.businepartne_businepartne_ = bp.id_ and fbpa.created_log_
        = (select min( created_log_) from log_businesspartneraddress
        lbpa where lbpa. businepartne_businepartne_ = bp.id_)
122 left join log_address bpaddress
123     on bpaddress.id_ = fbpa.address_address_ and bpaddress.cre-
        ated_log_ = (select min( created_log_) from log_address ladd
        where ladd.id_ = fbpa.address_address_ and typeofoperation_log_
        = 1)
124
125 left join fl_paymentmethodapplication payment
126     on payment.contract_contract_ = contract.id_ and payment.period-
        start_ = (select min(periodstart_) from fl_paymentmethodappli-
        cation where contract_contract_ = contract.id_ and (pay-
        ment.periodend_ > payment.periodstart_ or payment. periodend_
        is null))
127 left join fl_paymentmethod payMethod
128     on payMethod.id_ = payment.paymenmethod_paymenmethod_
129 left join fl_sepamandate sepa
130     on sepa.id_ = payment.sepamandate_sepamandate_ --and sepa.active_
        = true ergebnismenge verfälscht
131 left join fl_bankaccount bankacc on sepa.bankaccount_bankaccount_ =
    bankacc.id_
132
133 left join (select contract_contract_, sum(value_) as value from
    fl_bonusapplication group by contract_contract_) as flba
134     on flba.contract_contract_ = contract.id_
135
```

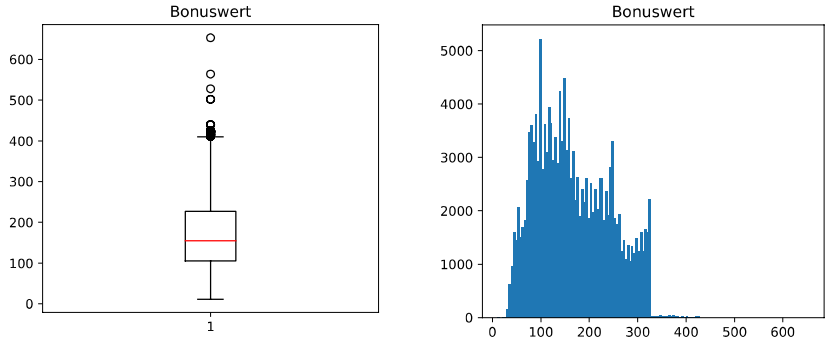
```
136 left join fl_powerconsumptprognosi powerconsprog
137    on powerconsprog.contract_ = contract.id_ and
          powerconsprog.source_ = 1 and powerconsprog.creationdate_ =
          (select min(creationdate_) from fl_powerconsumptprognosi where
          contract_ = contract.id_)
138
139 left join log_tariffapplication logtariffapp
140    on logtariffapp.contract_ = contract.id_ and logtariffapp.period-
          start_ is null and logtariffapp.typeofoperation_log_ = 1
141     --and logtariffapp.created_log_ = (select min(created_log_) from
            log_tariffapplication lt where lt.contract_ = logtar-
            iffapp.contract_)
142 left join fl_snapshotcost snapcost
143    on snapcost.tariffapplic_tariffapplic_ = logtariffapp.id_ and
          snapcost. pricecompon_pricecompon_ not in
          (120,122,124,126,128,129)
144
145 -- Höchster jemals erlangter integerAdditional3
146 left join (select lca.id_, max(lca.integeradditional3_) as maxia3
147                  from log_contractadditional lca
148                    inner join fl_contract ifc
149                      on ifc.contractadditionalid_ = lca.id_
150                  where lca.integeradditional3_ < 800
151                    -- Endstatus 9 (=bezahlt) geht immer eine Mahn-
                         stufe voraus
152                    and lca.integeradditional3_::text not like '%9' -
                         - ~ '^[0-9]{2}[0-8]{1}$'
153                    and lca.created_log_ < ifc.periodstart_ + inter-
                         val '365 days'
154                  group by lca.id_ ) as highestIntegerAdditional3
155    on contract.contractadditionalid_ = highestIntegerAdditional3.id_
156
157 -- Zeitpunkt nach Belieferung bis zur ersten Mahnung
158 left join (select id_, min(created_log_) as minia3date from log_con-
      tractadditional where integeradditional3_ >= 100 and integerad-
      ditional3_ < 800 group by id_ ) as firstIntegerAdditional3
159    on contract.contractadditionalid_ = firstIntegerAdditional3.id_
160
161 -- anzahl genutzte bankdaten
162 left join (select ifs.contract_contract_, count(distinct ifs. bankac-
      count_bankaccount_) bankacccount
163                  from fl_sepamandate ifs
164                     inner join fl_contract ifc
165                     on ifs.contract_contract_ = ifc.id_
166                       and ifs.periodstart_ < ifc.periodstart_ + in-
                            terval '365 days'
167                  group by contract_contract_) fsepachange
168    on fsepachange.contract_contract_ = contract.id_
169
170 -- häufigkeit rückrufservice
171 left join (select contract_contract_, count(contract_contract_) call-
      count from fl_task
172                     left join fl_contract on fl_contract.id_ =
                           fl_task. contract_contract_ and fl_task.peri-
                           odstart_ < fl_contract. periodstart_ + inter-
                           val '365 days'
173                     where title_ = 'Rückrufservice' group by con-
                           tract_contract_) callbacktask
174    on callbacktask.contract_contract_ = contract.id_
```

```
175 -- häufigkeit emails
176 left join (select contract_contract_, count(contract_contract_)
       emailcount from fl_task
177                        left join fl_contract on fl_contract.id_ =
                              fl_task. contract_contract_ and fl_task.peri-
                              odstart_ < fl_contract. periodstart_ + inter-
                              val '365 days'
178                        where title_ like 'E-Mail bearbeiten%' group by
                              contract_contract_) emailtask
179    on emailtask.contract_contract_ = contract.id_
180
181 -- anzahl zählerstände (summe über alle register/zähler)
182 left join (select fc.id_, count(flm.register_) as meterreadingcount
       from fl_contract fc
183                  left join fl_pointofdelivery as malo
184                        on malo.id_ = fc.pointofdelivery_
185                  left join fl_pointofdelivery as melo
186                        on melo.parent_ = malo.id_
187                  left join fl_meter as meter
188                        on meter.pointofdelivery_ = melo.id_
189                  left join fl_register as reg
190                        on reg.meter_registers_ = meter.id_
191                  left join (select register_, date_
192                             from fl_meterreading
193                             where source_ = 2
194                             ) as flm
195                        on flm.register_ = reg.id_
196                          and flm.date_ < fc.periodstart_ + interval
                             '365 days'
197                  group by fc.id_) meterreading
198    on contract.id_ = meterreading.id_
199
200 -- Abschlagsplanänderungen nach unten
201 left join (select contract_, count(diff_to_previous_row) as re-
       ducedcount
202                from (select contract_, creationreason_, creationdate_,
                         height_ - lag( height_)
203                                   over (partition by contract_ order by
                                         periodstart_, creationdate_) as
                                         diff_to_previous_row
204                        from fl_advancepayplan
205                        order by contract_, periodstart_) payplandiff
206                  inner join fl_contract fc
207                    on fc.id_ = payplandiff.contract_
208                      and payplandiff.creationdate_ < fc.periodstart_ +
                         interval '365 days'
209              where diff_to_previous_row < 0
210                and payplandiff.creationreason_ in (2, 3, 4, 5, 9, 10)
211              group by contract_) payplanreduced
212    on payplanreduced.contract_ = contract.id_
213
214 -- Abschlagsplanänderungen nach oben
215 left join (select contract_, count(diff_to_previous_row) as in-
       creasedcount
216                from (select contract_, creationreason_, creationdate_,
                         height_ - lag( height_)
```

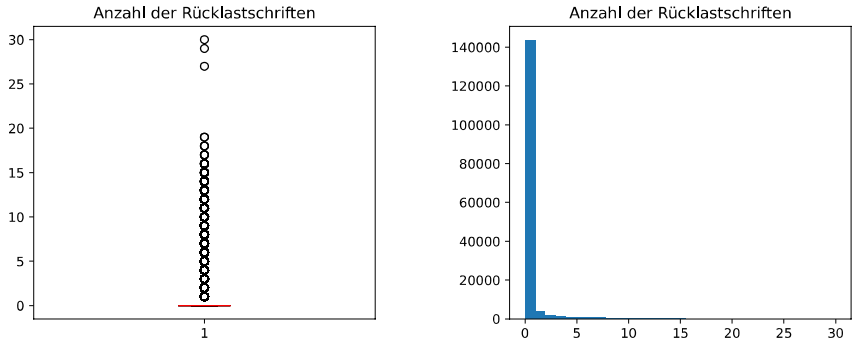
```
217                                 over (partition by contract_ order by
                                       periodstart_, creationdate_) as
                                       diff_to_previous_row
218                         from fl_advancepayplan
219                         order by contract_, periodstart_) payplandiff
220                 inner join fl_contract fc
221                   on fc.id_ = payplandiff.contract_
222                      and payplandiff.creationdate_ < fc.periodstart_ +
                         interval '365 days'
223             where diff_to_previous_row > 0
224               and payplandiff.creationreason_ in (2, 3, 4, 5, 9, 10)
225             group by contract_) payplanincreased
226    on payplanincreased.contract_ = contract.id_
227
228 -- Anzahl Rücklastschriften
229 left join (select fc.id_, count(entry.id_) as countSepaReversed
230                 from fl_contract fc
231                 left join fl_financiaaccountownershi finaccowner
232                    on finaccowner.contract_Contract_ = fc.id_
233                 inner join fl_financialaccount debitorAccount
234                    on debitorAccount.id_ = finaccowner.financAccoun_Ac-
                       count_
235                       and debitorAccount.type_ = 1
236                 left join (select financaccoun_debit_, postingdate_,
                       id_
237                               from fl_entry
238                               where subject_ like '%Rücklastschrift %'
239                               ) entry
240                 on entry.financaccoun_debit_ = debitoraccount.id_
241                    and entry.postingdate_ < fc.periodstart_ + interval
                       '365 days'
242                 group by fc.id_) separeversed
243    on contract.id_ = separeversed.id_
244
245
246 -- Eingang der Kündigung/Wechselanfrage neuer Lieferant in Tagen
247  -- In Tabelle sind nur CancellationPassive = Kündigung durch Lief
     und SupplyEndActive = Kunden Kundigung
248 left join (select contractid_, min(created_) firstCancellationDate
     from fl_gpkeprocess group by contractid_) gpke
249  on gpke.contractid_ = contract.id_
250
251
252 -- Anzahl manueller Überweisungen
253 left join (select fc.id_, count(entryManualSepa.id_) as countManu-
     alSepa
254                 from fl_contract fc
255                   left join fl_financiaaccountownershi finaccowner-
                         SepaAccount
256                   on finaccownerSepaAccount.contract_Contract_ =
                         fc.id_
257                 inner join fl_financialaccount debitorSepaAccount
258                   on debitorSepaAccount.id_ = finaccownerSepaAccount.
                         financAccoun_Account_
259                      and debitorSepaAccount.type_ = 6
260                   left join fl_financiaaccountownershi finaccowner
261                 on finaccowner.contract_Contract_ = fc.id_
262             inner join fl_financialaccount debitorAccount
```

```
263                 on debitorAccount.id_ = finaccowner.financAccoun_Ac-
                      count_
264                    and debitorAccount.type_ = 1
265                    left join (select financaccoun_debit_, financac-
                        coun_credit_, postingdate_, id_
266                                      from fl_entry) entryManualSepa
267                        on entryManualSepa.financaccoun_debit_ = debi-
                          torSepaAccount.id_
268                          and entryManualSepa.financaccoun_credit_ = deb-
                            itoraccount.id_
269                          and entryManualSepa.postingdate_ < fc.period-
                            start_ + interval '365 days'
270                    group by fc.id_) manualsepa
271      on manualsepa.id_ = contract.id_
272
273 where 1=1
274    and contract.state_ not in ('NEW', 'LOCKED', 'GPKE_ABORTED',
        'GPKE_RUNNING')
275    and (contract.periodstart_ < contract.periodend_ or contract.peri-
        odend_ is null)
276    and contract.periodstart_ > '2018-03-01'::date and contract.peri-
        odstart_ < ' 2021-03-01'::date
277    and logtariffapp.tariff_ not in (120, 124
```

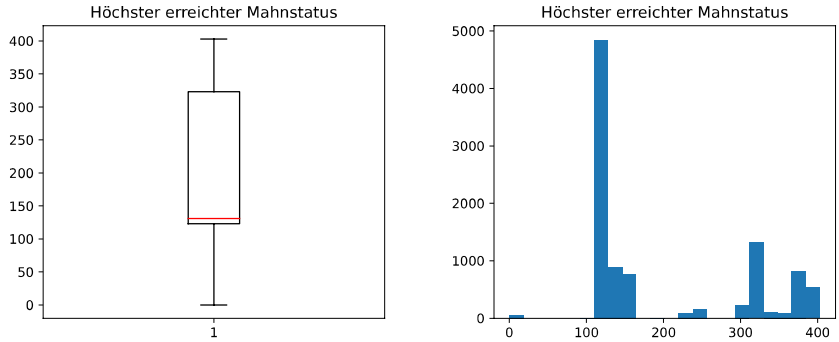
Anhang 4: Boxplots (mit Ausreißern) und Histogramme (ohne Ausreißer) ausgewählter Datenfelder



a) Bonuswert



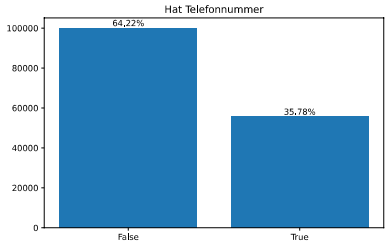
b) Anzahl der Rücklastschriften



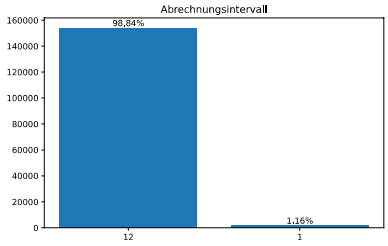
c) Höchster erreichter Mahnstatus

Abb. 34: Boxplots (mit Ausreißern) und Histogramme (ohne Ausreißer) ausgewählter Datenfelder

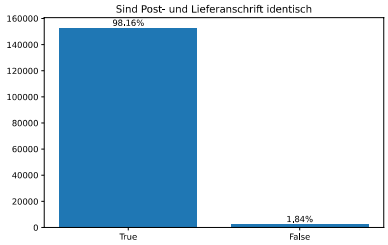
Anhang 5: Säulendiagramme mit prozentualen Anteilen ausgewählter Datenfelder



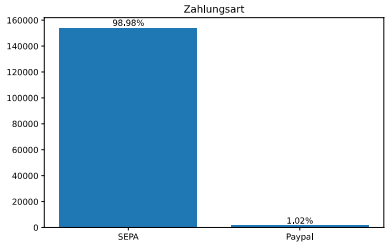
a) Hat Telefonnummer



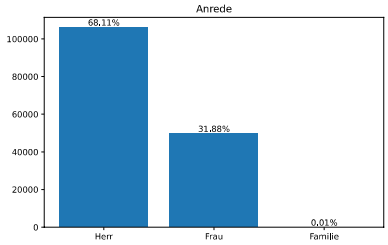
b) Abrechnungsintervall



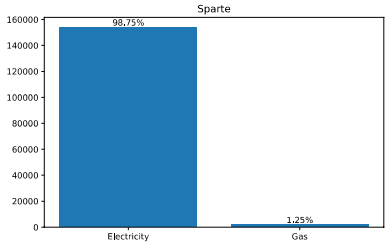
c) Sind Post- und Lieferanschrift identisch



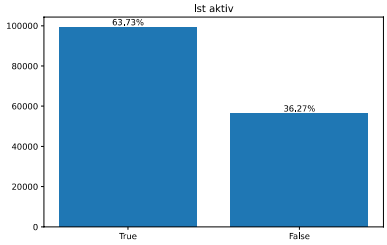
d) Zahlungsart



e) Anrede



f) Sparte



g) Ist aktiv

Abb. 35: Säulendiagramme mit prozentualen Anteilen ausgewählter Datenfelder

Anhang 6: Säulendiagramm mit Anteil aller Verträge und Anteil der Zahlungsausfälle gruppiert nach Verbrauch

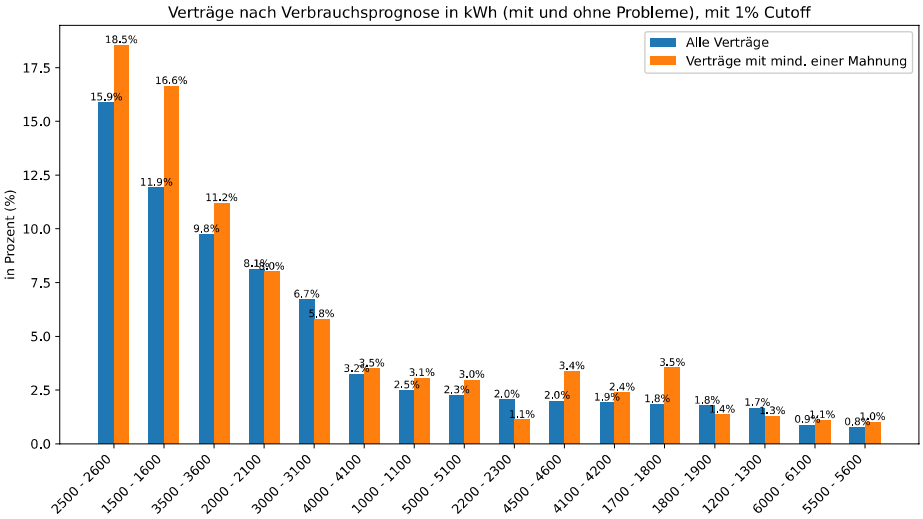


Abb. 36: Anteile aller Verträge und Verträge mit Zahlungsausfällen nach Verbrauchsprognose (gruppiert)

Anhang 7: Random Forest: F1-Score je Random Forest basierend auf Parametern aus Tabelle 9 (trainiert mit verschlankten Angebotsdaten inkl. Postleitzahlen)

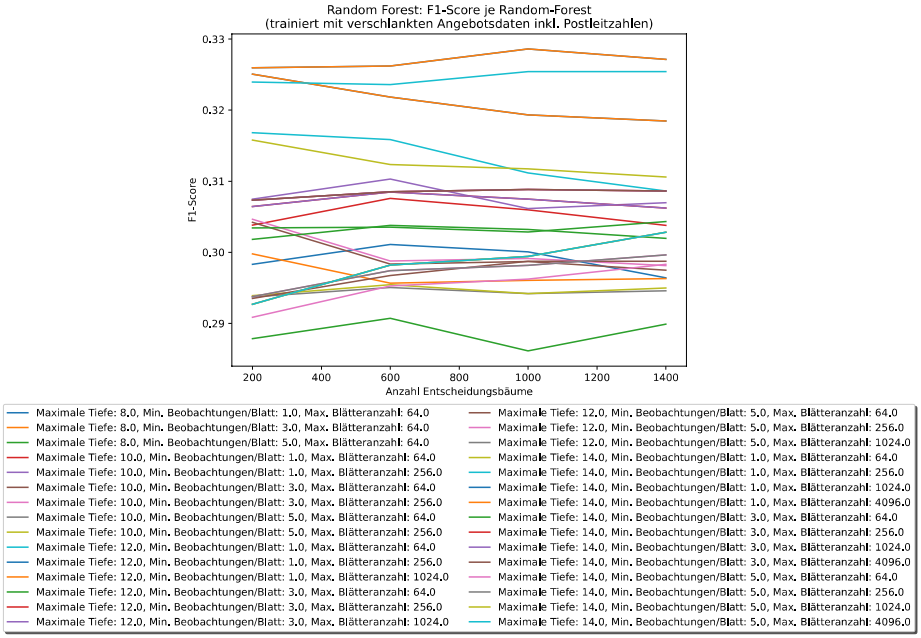


Abb. 37: Random Forest: F1-Score je Random Forest basierend auf Parametern aus Tabelle 9

Anhang 8: Random Forest: F1-Score je Random Forest basierend auf Parametern aus Tabelle 12 (trainiert mit den Verhaltensdaten, exkl. Variable Erste Mahnung nach Belieferungsbeginn in Tagen)

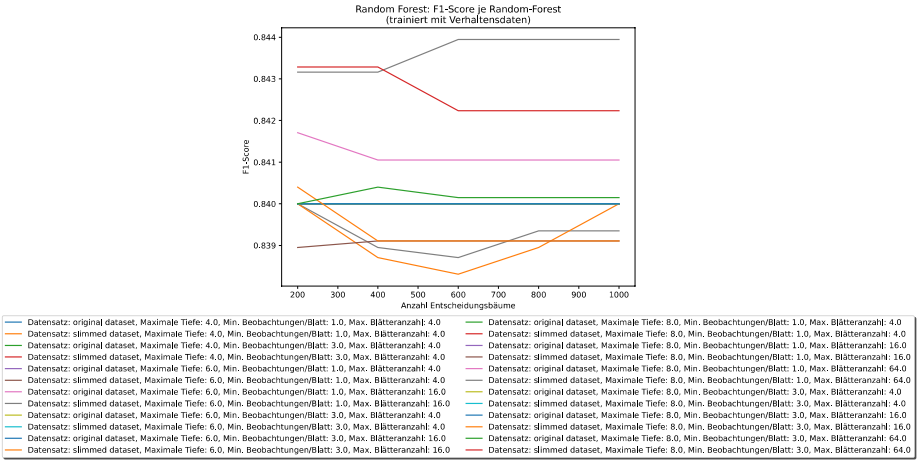


Abb. 38: Random Forest: F1-Score je Random Forest basierend auf Parametern aus Tabelle 12 (trainiert mit den Verhaltensdaten, exkl. Variable Erste Mahnung nach Belieferungsbeginn in Tagen)

Anhang 9: XGBoost: Log-Loss je XGBoost-Modell basierend auf Parametern aus Tabelle 14 (trainiert mit verschlankten Angebotsdaten inkl. Postleitzahlen)

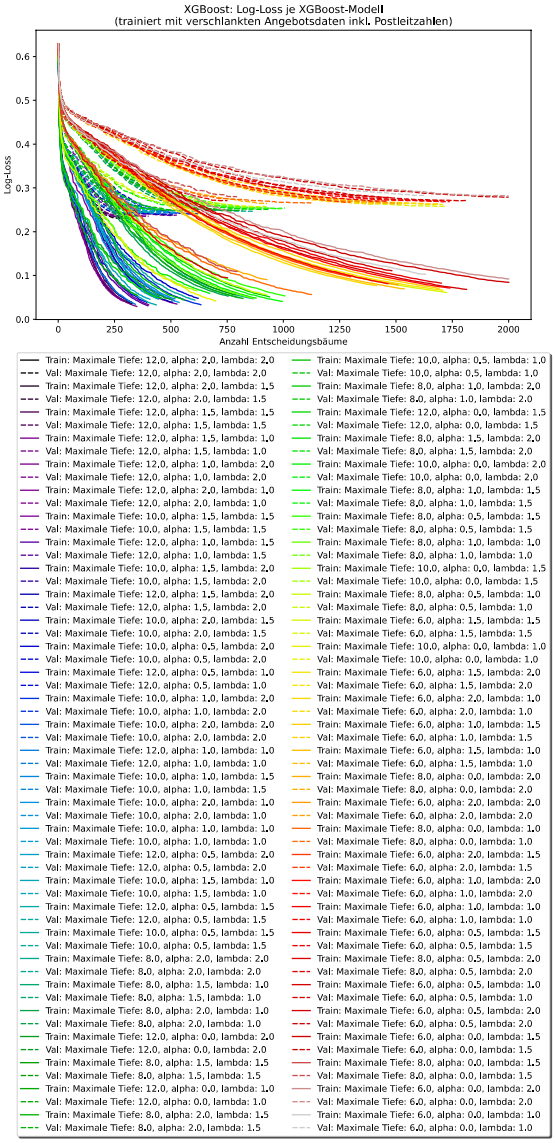


Abb. 39: XGBoost: Log-Loss je XGBoost-Modell basierend auf Parametern aus Tabelle 14 (trainiert mit verschlankten Angebotsdaten inkl. Postleitzahlen)

Anhang 10: XGBoost: Log-Loss je XGBoost-Modell basierend auf Parametern aus Tabelle 14 (trainiert mit verschlankten Angebotsdaten inkl. Postleitzahlen)

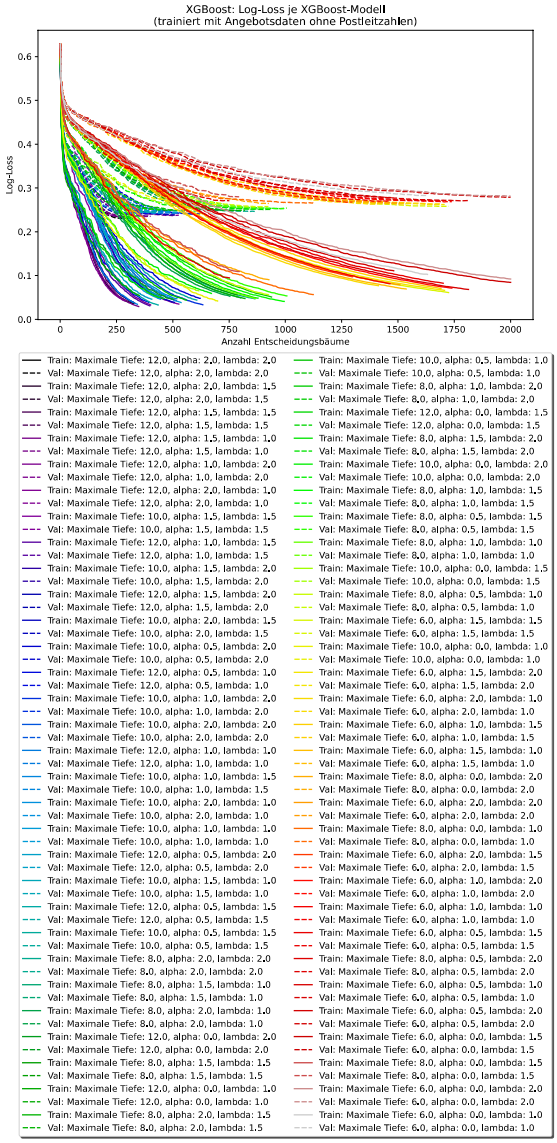
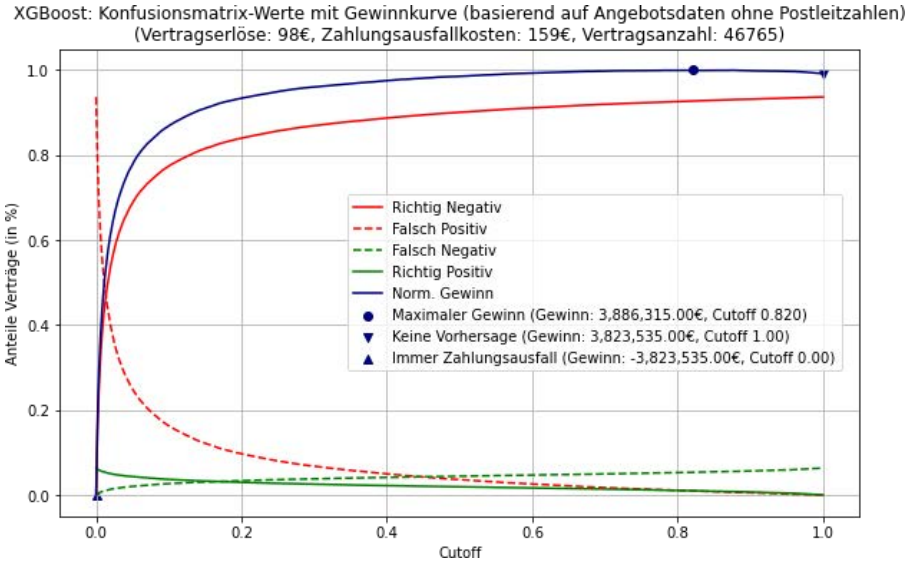
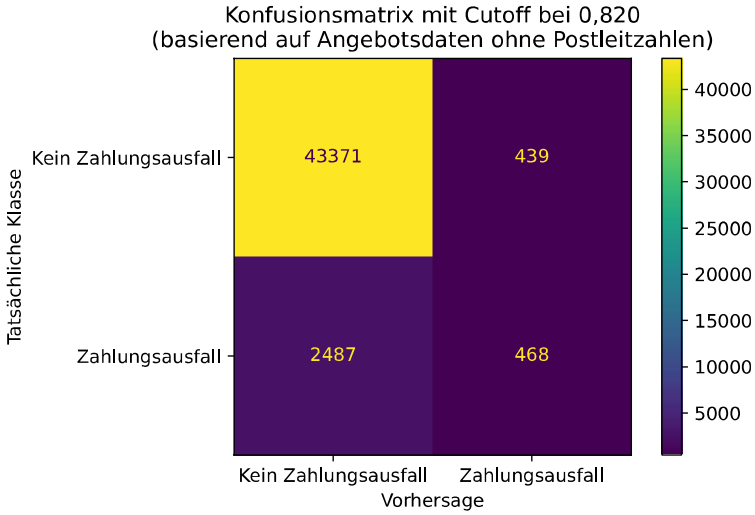


Abb. 40: XGBoost: Log-Loss je XGBoost-Modell basierend auf Parametern aus Tabelle 14 (trainiert mit verschlankten Angebotsdaten ohne Postleitzahlen)

Anhang 11: Evaluierung auf Basis des Testdatensatzes Angebotsdaten ohne Postleitzahlen mit dem ausgewählten XGBoost-Modell



a) Normierte Gewinnkurve und Konfusionsmatrix-Werte



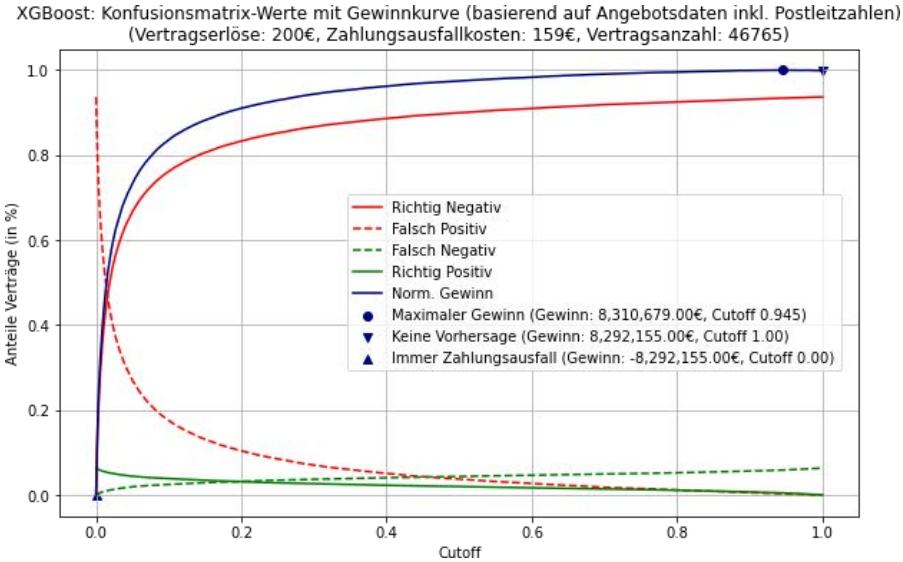
b) Konfusionsmatrix bei Cutoff 0,82

Abb. 41: Konfusionsmatrix und deren Werte mit Gewinnkurve basierend auf Formel 22 mit z = 159 und v = 98 (Datensatz Angebotsdaten ohne Postleitzahlen)

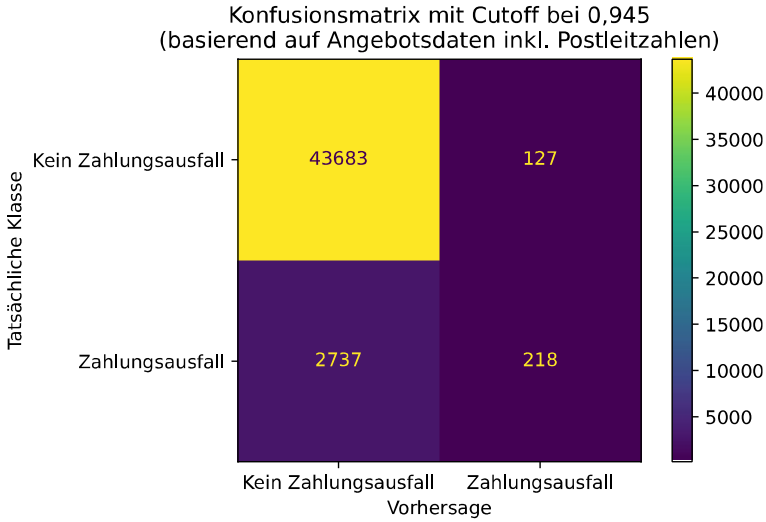
	Ohne Modell	**Mit Modell**	**Steigerung mit Modell (gerundet)**
Vertragsannahmen	46.765	44.590	-1,9 %
davon Zahlungsausfälle	2.955	2.107	-15,8 %
davon keine Zahlungsausfälle	43.810	42.483	-1 %
Gewinn (in Mio. €)	3,824	3,886	1,6 %

Tab. 22: Vertragsannahmen, Gewinn und der daraus resultierenden Steigerung mit und ohne Modelleinsatz (basierend auf dem Testdatensatz Angebotsdaten ohne Postleitzahlen)

Anhang 12: Evaluierung auf Basis der Testdatensätze mit den ausgewählten XGBoost-Modellen mit höheren Vertragserlösen für Formel 22

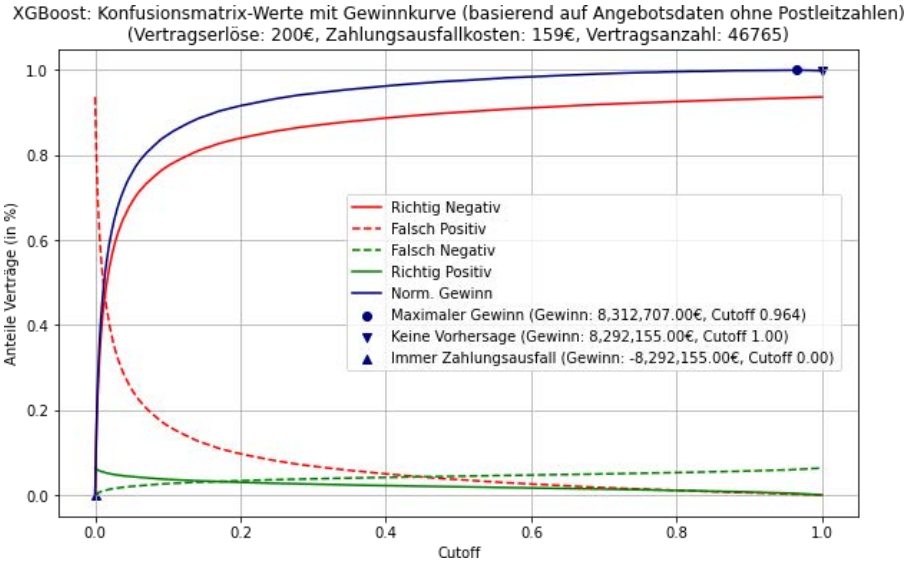


a) Normierte Gewinnkurve und Konfusionsmatrix-Werte

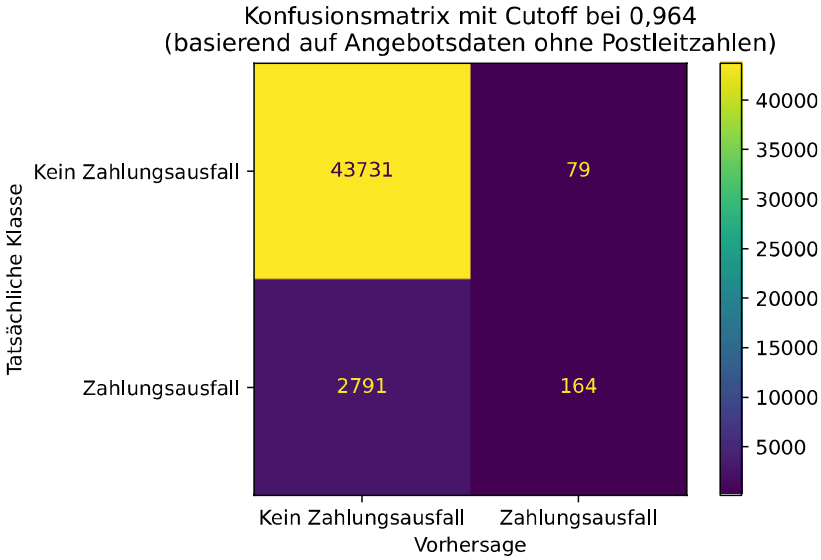


b) Konfusionsmatrix bei Cutoff 0,945

Abb. 42: Konfusionsmatrix und deren Werte mit Gewinnkurve basierend auf Formel 22 mit $z = 159$ und $v = 200$ (Datensatz Angebotsdaten inkl. Postleitzahlen)

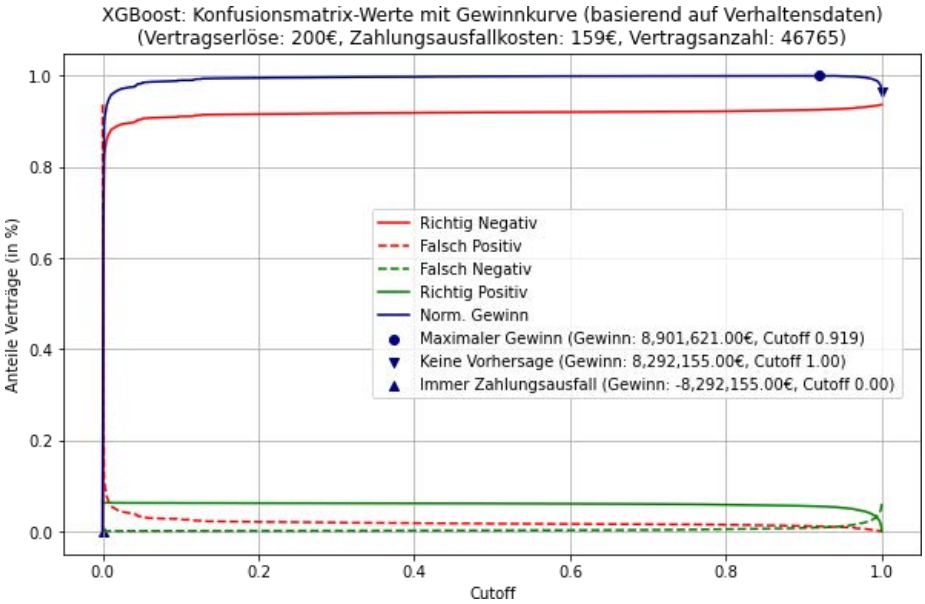


a) Normierte Gewinnkurve und Konfusionsmatrix-Werte

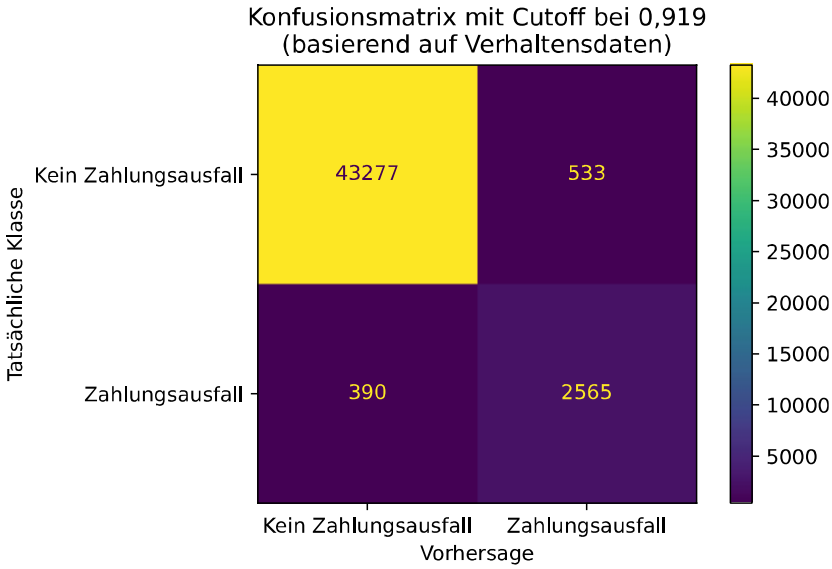


b) Konfusionsmatrix bei Cutoff 0,964

Abb. 43: Konfusionsmatrix und deren Werte mit Gewinnkurve basierend auf Formel 22 mit $z = 159$ und $v = 200$ (Datensatz Angebotsdaten ohne Postleitzahlen)



a) Normierte Gewinnkurve und Konfusionsmatrix-Werte



b) Konfusionsmatrix bei Cutoff 0,919

Abb. 44: Konfusionsmatrix und deren Werte mit Gewinnkurve basierend auf Formel 22 mit $z = 159$ und $v = 200$ (Datensatz Verhaltensdaten)

Literaturverzeichnis

Bender, *Ralf*, *Ziegler*, *Andreas*, *Lange*, *Stefan* (2007): Logistische Regression, in: DMW - Deutsche Medizinische Wochenschrift, 132 (2007), Nr. S 01, e33–e35

Bewick, *Viv*, *Cheek*, *Liz*, *Ball*, *Jonathan* (2004): Statistics review 13: receiver operating characteristic curves, in: Critical care, 8 (2004), Nr. 6, S. 1–5

Biau, *Gérard*, *Scornet*, *Erwan* (2016): A random forest guided tour, in: Test, 25 (2016), Nr. 2, S. 197–227

Breiman, *L.*, *Friedman*, *J. H.*, *Olshen*, *R. A.*, *Stone*, *C. J.* (1984): Classification and Regression Trees, Wadsworth und Brooks, 1984

Breiman, *Leo* (1996): Bagging predictors, in: Machine learning, 24 (1996), Nr. 2, S. 123–140

Breiman, *Leo* (2001): Random forests, in: Machine learning, 45 (2001), Nr. 1, S. 5–32

Bundesgesetzblatt (1998): Gesetz zur Neuregelung des Energiewirtschaftsrechts, in: Bundesgesetzblatt Jahrgang 1998 Teil I Nr. 23 (1998), S. 730–736

Bundeskartellamt (2019): Sektoruntersuchung Vergleichsportale (Az. V-21/17). Bundeskartellamt, <https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Diskussions_Hintergrundpapier/VS_SU_Vergleichsportale.pdf> [Zugriff: 2023-08-02]

Bundesnetzagentur und Bundeskartellamt (2021): Monitoringbericht 2020. Bundeskartellamt, S. 1–507, <https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Berichte/Energie-Monitoring-2020.pdf> [Zugriff: 2023-08-02]

Chapman, *Pete*, *Clinton*, *Julian*, *Kerber*, *Randy*, *Khabaza*, *Thomas*, *Reinartz*, *Thomas*, *Shearer*, *Colin*, *Wirth*, *Rudiger* et al. (2000): CRISP-DM 1.0: Step-by-step data mining guide, in: SPSS inc (2000)

Chen, *Chao*, *Liaw*, *Andy*, *Breiman*, *Leo* et al. (2004): Using random forest to learn imbalanced data, in: University of California, Berkeley, 110 (2004), Nr. 1-12, S. 24

Chen, *Tianqi*, *Guestrin*, *Carlos* (2016): XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, California, USA: ACM, 2016, S. 785–794

Chicco, *Davide*, *Jurman*, *Giuseppe* (2020): The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, in: BMC genomics, 21 (2020), Nr. 1, S. 1–13

Chicco, *Davide*, *Tötsch*, *Niklas*, *Jurman*, *Giuseppe* (2021): The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, in: BioData mining, 14 (2021), Nr. 1, S. 1–22

Cramer, *James S.* (2003): The origins and development of the logit model, <https://www.cambridge.org/ua/files/2013/6690/1022/1208_default.pdf> [Zugriff: 2023-08-02]

Davis, *Jesse*, *Goadrich*, *Mark* (2006): The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd international conference on Machine learning, 2006, S. 233–240

Deutsche Presseagentur (21.02.2019): Muttergesellschaft von BEV ebenfalls insolvent, in: Frankfurter Allgemeine Zeitung, 44 (21.02.2019), S. 18

Dietterich, *Thomas G* (2000): Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer, 2000, S. 1–15

Eliason, *Scott R* (1993): Maximum likelihood estimation: Logic and practice, 96, Sage, 1993

EnWG (2021): Energiewirtschaftsgesetz, in: Energiewirtschaftsgesetz, zuletzt geändert durch Gesetz vom 16.07.2021 (BGBl. I S. 3026) m.W.v. 01.10.2021 (2021)

Europäisches Parlament und Europäischer Rat (1996): Richtlinie 96/92/EG des Europäischen Parlaments und des Rates vom 19. Dezember 1996 betreffend gemeinsame Vorschriften für den Elektrizitätsbinnenmarkt, in: Amtsblatt Nr. L, 27 (1996), S. 0020 *Europäisches Parlament und Europäischer Rat* (1998): Richtlinie 98/30/EG des Europäischen Parlaments und des Rates vom 22. Juni 1998 betreffend gemeinsame Vorschriften für den Erdgasbinnenmarkt, in: Amtsblatt Nr. L, 204 (1998), S. 1–12

Freund, *Yoav*, *Schapire*, *Robert*, *Abe*, *Naoki* (1999): A short introduction to boosting, in: Journal-Japanese Society For Artificial Intelligence, 14 (1999), Nr. 771-780, S. 1612

Freund, *Yoav*, *Schapire*, *Robert E.* (1997): A decision-theoretic generalization of on-line learning and an application to boosting, in: Journal of computer and system sciences, 55 (1997), Nr. 1, S. 119–139

Friedman, *Jerome H* (2001): Greedy function approximation: a gradient boosting machine, in: Annals of statistics (2001), S. 1189–1232

Gillies, *Sean* et al. (2007): Shapely: manipulation and analysis of geometric objects, 2007, URL: https://github.com/shapely/shapely

Goodfellow, *Ian*, *Bengio*, *Yoshua*, *Courville*, *Aaron* (2016): Deep Learning, http://www.deeplearningbook.org, MIT Press, 2016

Harris, Charles R., Millman, K. Jarrod, van der Walt, Stéfan J., Gommers, Ralf, Virtanen, Pauli, Cournapeau, David, Wieser, Eric, Taylor, Julian, Berg, Sebastian, Smith, Nathaniel J., Kern, Robert, Picus, Matti, Hoyer, Stephan, van Kerkwijk, Marten H., Brett, Matthew, Haldane, Allan, del Río, Jaime Fernández, Wiebe, Mark, Peterson, Pearu, Gérard-Marchant, Pierre, Sheppard, Kevin, Reddy, Tyler, Weckesser, Warren, Abbasi, Hameer, Gohlke, Christoph, Oliphant, Travis E. (2020): Array programming with NumPy, in: Nature, 585 (2020), Nr. 7825, S. 357–362

Ho, *Tin Kam* (1995): Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition, Bd. 1, IEEE, 1995, S. 278–282

Hüls, *Ralf*, *Henking*, *Andreas* (2003): Privatkundengeschäft: "Mit Scoring zu mehr Ertrag". In: bank und markt + technik (2003), Nr. 03

Hunter, *J. D.* (2007): Matplotlib: A 2D graphics environment, in: Computing in Science & Engineering, 9 (2007), Nr. 3, S. 90–95

Jain, *Vikas*, *Phophalia*, *Ashish*, *Bhatt*, *Jignesh S* (2018): Investigation of a joint splitting criteria for decision tree classifier use of information gain and gini index, in: TENCON 2018-2018 IEEE Region 10 Conference, IEEE, 2018, S. 2187–2192

Johnson, *Rie*, *Zhang*, *Tong* (2013): Learning nonlinear functions using regularized greedy forest, in: IEEE transactions on pattern analysis and machine intelligence, 36 (2013), Nr. 5, S. 942–954

Jordahl, Kelsey, den Bossche, Joris Van, Fleischmann, Martin, Wasserman, Jacob, McBride, James, Gerard, Jeffrey, Tratner, Jeff, Perry, Matthew, Badaracco, Adrian Garcia, Farmer, Carson, Hjelle, Geir Arne, Snow, Alan D., Cochran, Micah, Gillies, Sean, Culbertson, Lucas, Bartos, Matt, Eubank, Nick, maxalbert, Bilogur, Aleksey, Rey, Sergio, Ren, Christopher, Arribas-Bel, Dani, Wasser, Leah, Wolf, Levi John, Journois, Martin, Wilson, Joshua, Greenhall, Adam, Holdgraf, Chris, Filipe, Leblanc, François (2020): geopandas/geopandas: v0.8.1, Version v0.8.1, 2020-07, URL: https://doi.org/10.5281/zenodo.3946761

Jung, *Marcus* (21.01.2020): Musterklage nach BEV-Insolvenz, in: Frankfurter Allgemeine Zeitung, 23 (21.01.2020), S. 21

Kähler, Jürgen (2012): Regressionsanalyse, in: *Schröder, Michael* (Hrsg.), Finanzmarkt-Ökonometrie: Basistechniken, Fortgeschrittene Verfahren, Prognosemodelle, Schäffer-Poeschel, 2012, Kap. 2, S. 29–98

Kammel, Eike and Hollmann, Maik (2016): Big Data in der Energiewirtschaft: So wird die Umwandlung in eine datengesteuerte Organisation zum Erfolg, in: *Köhler-Schute, Christiana* (Hrsg.), Die Digitalisierung der Energiewirtschaft, KS-Energy Verlag, 2016, Kap. 3, S. 41–51

King, *Gary*, *Zeng*, *Langche* (2001): Logistic regression in rare events data, in: Political analysis, 9 (2001), Nr. 2, S. 137–163

Köhler, *Manfred* (02.02.2019): BEV-Kunden müssen nicht im Dunkeln sitzen, in: Frankfurter Allgemeine Zeitung, 28 (02.02.2019), S. 37

Krämer, *Walter* (2002): Die Bewertung und der Vergleich von Kreditausfall-Prognosen, Techn. Ber., Technical Report, 2002

Larose, *Daniel T* (2015): Data Mining and Predictive Analytics, John Wiley & Sons, 2015

Liaw, *Andy*, *Wiener*, *Matthew* et al. (2002): Classification and regression by randomForest, in: R news, 2 (2002), Nr. 3, S. 18–22

Lohse, *Lutz*, *Künzel*, *Manuela* (2011): Customer Relationship Management im Energiemarkt, in: *Enke, Margit and Geigenmüller* (Hrsg.), Commodity Marketing, Bd. 148, Gabler, 2011, S. 381–400

McKinney, *Wes* (2010): Data Structures for Statistical Computing in Python, in: *Van der Walt*, *Stéfan*, *Millman*, *Jarrod* (Hrsg.), Proceedings of the 9th Python in Science Conference, 2010, S. 56–61

Meffert, *Heribert*, *Schröder*, *Jürgen*, *Perrey*, *Jesko* (2002): B2C-Märkte: Lohnt sich ihre Investition in die Marke, in: absatzwirtschaft, 45 (2002), Nr. 10, S. 28–35

Mihm, *Andreas* (20.08.2017): Deutschland hat den höchsten Strompreis in Europa - Ökostrom-Umlage treibt Preis / Energiebranche: Nicht nachvollziehbar, warum Bundesregierung nicht handelt, in: Frankfurter Allgemeine Zeitung, 131 (20.08.2017), S. 20

Mingers, *John* (1989): An empirical comparison of pruning methods for decision tree induction, in: Machine learning, 4 (1989), Nr. 2, S. 227–243

OpenStreetMap contributors (2017): Planet dump retrieved from https://planet.osm.org, https://www.openstreetmap.org, 2017

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011): Scikit-learn: Machine Learning in Python, in: Journal of Machine Learning Research, 12 (2011), S. 2825–2830

Pfaffenberger, *Wolfgang*, *Hille*, *Maren* (2004): Investitionen im liberalisierten Energiemarkt: Optionen, Marktmechanismen, Rahmenbedingungen, in: VDEW, AGFW, VDN, VGB PowerTech, VKU, VRE (2004), S. 208

Powers, *David* (2011): Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, in: Journal of Machine Learning Technologies, 2 (2011), Nr. 1, S. 37–63

Quinlan, *J. Ross* (1986): Induction of decision trees, in: Machine learning, 1 (1986), Nr. 1, S. 81–106

Raileanu, *Laura Elena*, *Stoffel*, *Kilian* (2004): Theoretical comparison between the gini index and information gain criteria, in: Annals of Mathematics and Artificial Intelligence, 41 (2004), Nr. 1, S. 77–93

Rücker, Markus (2016): Herausforderungen und Ansätze der digitalen Vertriebsoptimierung bei Energieversorgern, in: *Köhler-Schute, Christiana* (Hrsg.), Die Digitalisierung der Energiewirtschaft, KS-Energy Verlag, 2016, Kap. 2, S. 32–40

Sasaki, *Yutaka*, *Fellow*, *R* (2007): The truth of the F-measure, Manchester: MIB-School of Computer Science, in: University of Manchester (2007), S. 25

Schiffer, *Hans-Wilhelm* (2019): Energiemarkt Deutschland, 2019

Society for Worldwide Interbank Financial Telecommunication (SWIFT) (2018): ISO 9362 BIC implementation: changes and impact, in (2018): 57395, S. 22

Szczesny, Andrea and Kaiser, Ulrich (2012): Logit- und Probit-Modelle, in: *Schröder, Michael* (Hrsg.), Finanzmarkt-Ökonometrie: Basistechniken, Fortgeschrittene Verfahren, Prognosemodelle, Schäffer-Poeschel, 2012, Kap. 7, S. 313–345

The pandas development team (2020): pandas-dev/pandas: Pandas, Version latest, 2020-02, URL: https://doi.org/10.5281/zenodo.3509134

Verbeek, *Marno* (2017): A guide to modern econometrics, John Wiley & Sons, 2017

Vissing-Jorgensen, *Annette* (2011): Consumer Credit: Learning Your Customer's Default Risk from What (S) he Buys, in: Available at SSRN 2023238 (2011)

Vrieze, *Scott I* (2012): Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). In: Psychological methods, 17 (2012), Nr. 2, S. 228

Waskom, *Michael L.* (2021): seaborn: statistical data visualization, in: Journal of Open Source Software, 6 (2021), Nr. 60, S. 3021

Wasserman, *Larry* (2004): All of statistics: a concise course in statistical inference, Bd. 26, Springer, 2004

Wiedmann, *Klaus-Peter*, *Ludewig*, *Dirk* (2011): Commodity Branding, in: *Enke, Margit, Geigenmüller, Anja* (Hrsg.), Commodity Marketing, Bd. 148, Gabler, 2011, S. 81–114

Wu, *Wuqing*, *Xu*, *Dongliang*, *Zhao*, *Yue*, *Liu*, *Xinhai* (2020): Do consumer internet behaviours provide incremental information to predict credit default risk?, in: Economic and Political Studies, 8 (2020), Nr. 4, S. 482–499

Zeileis, *Achim*, *Hothorn*, *Torsten* (2002): Diagnostic Checking in Regression Relationships, in: R News, 2 (2002), Nr. 3, S. 7–10

Zeileis, *Achim*, *Kleiber*, *Christian*, *Jackman*, *Simon* (2008): Regression Models for Count Data in R, in: Journal of Statistical Software, 27 (2008), Nr. 8

Internetquellen

Bayrische Staatsregierung (22.01.2019): Pressemitteilung: Bundesnetzagentur leitet Aufsichtsverfahren gegen Energieversorger BEV Energie ein, <https://www.bayern.de/bundesnetzagentur-leitet-aufsichtsverfahren-gegen-energieversorger-bev-energie-ein/> (22.01.2019) [Zugriff: 2022-01-29]

Der Tagesspiegel (o.V.) (09.06.2021): Verbraucherschützer befürchten mehr Stromsperren, <https://www.tagesspiegel.de/wirtschaft/gravierende-verschlechterungen-verbraucherschuetzer-befuerchten-mehr-stromsperren/27270224.html> (09.06.2021) [Zugriff: 2021-11-09]

Güßgen, *Florian* (29.10.2021): Nächster Stromversorger meldet Insolvenz an, <https://www.wiwo.de/unternehmen/energie/pleitewelle-erwischt-lition-energie-naechster-stromversorger-meldet-insolvenz-an/27749844.html> (29.10.2021) [Zugriff: 2021-11-08]

Hoyer, *Niklas* (21.12.2018): Massenhaft Beschwerden über Billig-Stromanbieter BEV, <https://www.wiwo.de/unternehmen/energie/bayerische-energieversorgungsgesellschaft-massenhaft-beschwerden-ueber-billig-stromanbieter-bev/23792316.html> (21.12.2018) [Zugriff: 2021-11-08]

Jensen, Kenneth (26.04.2012): CRISP-DM Process Diagram, <https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png> (26.04.2012) [Zugriff: 2023-08-02]

Juristische Informationssystem für die Bundesrepublik Deutschland (juris GmbH) (2021): Prozesskostenrechner von juris - RVG Rechner 2021 | juris Das Rechtsportal, <https://www.juris.de/jportal/nav/juris_2015/aktuelles/rvg_rechner/rvg-rechner.jsp#> (2021) [Zugriff: 2021-11-10]

Next Kraftwerke GmbH (2021): Definition | Liberalisierung & Unbundling von Strommärkten, <https://www.next-kraftwerke.de/wissen/liberalisierung-strommaerkte> (2021) [Zugriff: 2021-10-23]

Python Software Foundation (2022): Download Python | Python.org, <https://www.python.org/downloads/> (2022) [Zugriff: 2022-01-29]

SCHUFA Holding AG (2021a): Fragen zum Thema Scoring, <https://www.schufa.de/faq/privatpersonen/scoring/> (2021) [Zugriff: 2021-11-10]

SCHUFA Holding AG (2021b): Wie funktioniert Scoring bei der SCHUFA?, <https://www.schufa.de/ueber-uns/daten-scoring/scoring/scoring-schufa/> (2021) [Zugriff: 2021-11-10]

Schwochow, *Marco* (2022): PLZ Download • Postleitzahlen als Liste und Karte, <https://www.suche-postleitzahl.org/downloads> (2022) [Zugriff: 2022-01-25]

Statistisches Bundesamt (2022): Mindestlohn - Statistisches Bundesamt, <https://www.destatis.de/DE/Themen/Arbeit/Verdienste/Mindestloehne/_inhalt.html> (2022) [Zugriff: 2022-01-29]

Stiftung Warentest (2021): Stromtarif Einfach Stromanbieter wechseln und sparen | Stiftung Warentest, <https://www.test.de/Stromtarif-Einfach-den-Stromanbieter-wechseln-und-sparen-5635723-0/> (2021) [Zugriff: 2021-10-24]

The R Foundation (2022): Download R-4.1.2 for Windows. The R-project for statistical computing. <https://cran.r-project.org/bin/windows/base/> (2022) [Zugriff: 2022-01-24]

Verband der Vereine Creditreform e. V. (2021a): Bonität Konsumenten | Creditreform, <https://www.creditreform.de/loesungen/bonitaet-risikobewertung/bonitaet-konsumenten)> (2021) [Zugriff: 2021-11-10]

Verband der Vereine Creditreform e.V. (2021b): Lösungen zu Bonität & Risikobewertung | Creditreform, <https://www.creditreform.de/loesungen/bonitaet-risikobewertung> (2021) [Zugriff: 2021-11-10]

Verivox (2021): Wir sind Verivox - 100 % unabhängig, <https://www.verivox.de/company/wir-sind-unabhaengig/> (2021) [Zugriff: 2021-10-29]

xgboost developers (2022a): Frequently Asked Questions - xgboost 1.6.0-dev documentation, <https://xgboost.readthedocs.io/en/latest/faq.html#slightly-different-result-between-runs> (2022) [Zugriff: 2022-01-29]

xgboost developers (2022b): Introduction to Boosted Trees - xgboost 1.6.0-dev documentation, <https://xgboost.readthedocs.io/en/latest/tutorials/model.html> (2022) [Zugriff: 2022-01-29]

xgboost developers (2022c): Notes on Parameter Tuning - xgboost 1.6.0-dev documentation, <https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html> (2022) [Zugriff: 2022-01-29]

xgboost developers (2022d): XGBoost Parameters - xgboost 1.6.0-dev documentation, <https://xgboost.readthedocs.io/en/latest/parameter.html> (2022) [Zugriff: 2022-01-29]

kostenloser Download
unter **fom-ifes.de**

Rojahn, J. / Schweinzger, O. / Zechser, F. (2022): Determinanten der Segmentberichtserstattungstransparenz – Eine Analyse der Variablenwichtigkeit, in: Krol, B. (Hrsg.): ifes Schriftenreihe, Band 29, 2022, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-434-3

Lehrbass, F. / Rebeggiani, L. / Schmidt, J.-S. (2022): Auswirkungen von Sponsorship-Verkündungen auf die Aktienkurse von Sportartikelherstellern, in: Krol, B. (Hrsg.): ifes Schriftenreihe, Band 28, 2022, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-432-9

Pleines, S. / Lehrbass, F. (2021): Backtesting von volatilitätsgesteuerten Aktienportfolios, in: Krol, B. (Hrsg.): ifes Schriftenreihe, Band 27, 2021, ISSN (e-Book) 2569-5355, ISBN (eBook) 978-3-89275-430-5

Maasjosthusmann, R. / Lehrbass, F. (2021): Explainable Artificial Intelligence: Analyse und Visualisierung des Lernprozesses eines Convolutional Neural Network zur Erkennung deutscher Straßenverkehrsschilder, in: Krol, B. (Hrsg.): ifes Schriftenreihe, Band 26, 2021, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-428-2

Hernes, D. / Lehrbass, F. / Maucy, K. (2021): Big Data basierte Analyse des Einflusses traditioneller und neuartiger Faktoren auf Mietpreise in Düsseldorf, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 25, 2021, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-426-8

Lehrbass, F. (2021): Deep Learning Diagnostics – How to Avoid Being Fooled by TensorFlow, PyTorch, or MXNet with the Help of Modern Econometrics, in:

Krol, B. (Hrsg.), ifes Schriftenreihe, Band 24, 2021, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-424-4

Lehrbass, F. / Wörndl, F. (2021): Was treibt die Renditen von Hedgefonds? Eine empirische Untersuchung ausgewählter Hedgefonds Strategien, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 23, 2021, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-422-0

Kladroba, A. / Friz, K. / Buchmann, T. / Wolf, P. (2020): Netzwerk- und Outputmessung – Indikatorik für transformative Technologiefelder (NEO-Indikatorik), in: Krol, B. / Kladroba, A. (Hrsg.), ifes Schriftenreihe, Band 22, 2020, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-420-6

Bähren, T. / Maasjosthusmann, R. / Walter, A. / Lehrbass, F. (2020): Praktische Umsetzung von Business Analytics im Mediensektor: Predictive Analytics im Filmgeschäft, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 21, 2020, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-418-3

Kladroba, A. (2019): Der Einfluss mathematischer Methoden auf das Ergebnis von Mannschaftswettkämpfen: Eine Simulationsrechnung, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 20, 2019, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-416-9

Raasch, A. / Lehrbass, F. (2019): Investmentstrategien im Rahmen von Übernahmen börsennotierter Gesellschaften – Merger Arbitrage und Maschinelles Lernen, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 19, 2019, ISSN 2191-3366, ISBN 978-3-89275-413-8

Hagemann, D. / Lehrbass, F. (2018): Prognosemodelle für Länderrisiken: Logit- und Deep Learning-Methoden im Vergleich, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 18, 2018, ISSN 2191-3366, ISBN 978-3-89275-411-4

Graalmann, M.-P. / Lehrbass, F. (2018): Eignung von Varianz-Kovarianz-Ansätzen und Copula-Modellen zur Risikoaggregation in bankaufsichtlichen Risikotragfähigkeitskonzepten, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 17, 2018, ISSN 2191-3366, ISBN 978-3-89275-409-1

Cox, P. / Lehrbass, F. (2018): Determinanten der Replikationsgüte von Exchange Traded Funds, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 16, 2018, ISSN 2191-3366, ISBN 978-3-89275-407-7

Lehrbass, F. / Scheipers, N. (2017): Determinanten der Höhe von Wirtschaftsprüfungshonoraren am Beispiel von gelisteten Unternehmen im Prime Standard, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 15, 2017, ISSN 2191-3366, ISBN 978-3-89275-406-0

Schwarz, J. (2017): Ergebnisse der Analyse von Studienabbrüchen, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 14, 2017, ISSN 2191-3366, ISBN 978-3-89275-405-3

Lehrbass, F. (2016): Risikomessung für den globalen Kohlehandel: Einfache und fortgeschrittene Verfahren nebst Backtesting sowie ein Vergleich mit IFRS 7, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 13, 2016, ISSN 2191-3366, ISBN 978-3-89275-404-6

Godbersen, H. (2016): Die Means-End Theory of Complex Cognitive Structures – Entwicklung eines Modells zur Repräsentation von verhaltensrelevanten und komplexen Kognitionsstrukturen für die Wirtschafts- und Sozialwissenschaften, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 12, 2016, ISSN 2191-3366, ISBN 978-3-89275-403-9

Seng, A. / Landherr, G. (2015): Vielfalt leben und Vielfalt gestalten – Diversity Management in der Lehre, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 11, 2015, ISSN 2191-3366, ISBN 978-3-89275-402-2

Gansser, O. A. / Schutkin, A. (2014): Studie zur Validierung der Persönlichkeitsmerkmale Abenteuerlust und Routineverhalten, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 10, 2014, ISSN 2191-3366, ISBN 978-3-89275-401-5

Gansser, O. A. (2014): Marketingplanung als Instrument zur Krisenbewältigung, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 9, 2014, ISSN 2191-3366, ISBN 978-3-89275-400-8

Runia, P. M. / Wahl, F. / Rüttgers, C. (2013): Das Markenimage von Hersteller- und Handelsmarken: Eine empirische Analyse der Imagekomponenten von Körperpflegemarken auf der Grundlage eines Markenidentitätskonzeptes, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 8, 2013, ISSN 2191-3366

Naskrent, J. / Rüttgers, C. (2013): Sportmonitor Essen 2013: Eine empirische Analyse über das Image regionaler Sportvereine und ihre Sponsoring- und Promotionangebote, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 7, 2013, ISSN 2191-3366

Seng, A. / Fiesel, L. / Rüttgers, C. (2013): Akzeptanz der Frauenquote, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 6, 2013, ISSN 2191-3366

Naskrent, J. / Rüttgers, C. (2012): Wahrnehmung von Werbung mit Sportereignisbezug: Eine empirische Analyse der Einschätzung von Sponsoring und Ambush-Marketing im Rahmen der Fußball-Europameisterschaft und der Olympischen Spiele im Jahr 2012, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 5, 2012, ISSN 2191-3366

Seng, A. / Fiesel, L. / Krol, B. (2012): Erfolgreiche Wege der Rekrutierung in Social Networks, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 4, 2012, ISSN 2191-3366

Heinemann, S. / Krol, B. (2011): Nachhaltige Nachhaltigkeit: Zur Herausforderung der ernsthaften Integration einer angemessenen Ethik in die Managementausbildung, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 2, 2011, ISSN 2191-3366

Hermeier, B. / Rettig, P. / Krol, B. (2010): Marken- und Produktmanagement durch Nutzung von Sportgroßereignissen: Möglichkeiten und Grenzen für Industrie und Handel, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 1, 2010, ISSN 2191-3366

ISBN (Print) 978-3-89275-435-0 ISSN (Print) 2191-3366
ISBN (eBook) 978-3-89275-436-7 ISSN (eBook) 2569-5355

Institut für Empirie & Statistik
der FOM Hochschule
für Oekonomie & Management

FOM Hochschule

FOM. Die Hochschule besonderen Formats

Mit über 50.000 Studierenden ist die FOM eine der größten Hochschulen Europas und führt seit 1993 Studiengänge für Berufstätige durch, die einen staatlich und international anerkannten Hochschulabschluss (Bachelor/Master) erlangen wollen.

Die FOM ist der anwendungsorientierten Forschung verpflichtet und verfolgt das Ziel, adaptionsfähige Lösungen für betriebliche bzw. wirtschaftsnahe oder gesellschaftliche Problemstellungen zu generieren. Dabei spielt die Verzahnung von Forschung und Lehre eine große Rolle: Kongruent zu den Masterprogrammen sind Institute und KompetenzCentren gegründet worden. Sie geben der Hochschule ein fachliches Profil und eröffnen sowohl Wissenschaftlerinnen und Wissenschaftlern als auch engagierten Studierenden die Gelegenheit, sich aktiv in den Forschungsdiskurs einzubringen.

Weitere Informationen finden Sie unter **fom.de**

ifes

Zunehmende Digitalisierung erfordert und ermöglicht datenbasierten Erkenntnisgewinn und fundiertes unternehmerisches Handeln. Um aus den allgegenwärtigen Daten die richtigen Schlüsse zu ziehen, ist überall eine kritische Methodenkompetenz erforderlich. Der wissenschaftliche Fokus der ifes-Akteure liegt dabei in den Bereichen der empirischen Unternehmens-, Markt- und Konsumentenforschung, der angewandten Statistik, des Data Minings und der Finanzstatistik.

Das ifes verfolgt das Ziel, empirische Kompetenzen an der FOM zu bündeln und die angewandte Forschung im empirischen Bereich der Hochschule weiter voranzutreiben. Damit nimmt das ifes eine zentrale Stellung im Bereich der Entwicklung und Unterstützung der Methodenausbildung in der Lehre der Bachelor- und Masterstudiengänge sowie im Promotionsprogramm der FOM ein.

Weitere Informationen finden Sie unter **fom-ifes.de**

FOM forscht.
Der Wissenschaftsblog
der FOM Hochschule

Der Wissenschaftsblog der FOM Hochschule bietet Einblicke in die vielfältigen Themen, zu denen an der FOM geforscht wird: **fom-blog.de**