



Hochschule Reutlingen
Reutlingen University



Fallstudie Smart Banking

vorgelegt von

Thi Yen Vy Huynh

Matrikelnummer: 800584

Hochschule Reutlingen

Masterstudiengang Digital Business Engineering

Sommersemester 2021

Modul: DBE24 Artificial Intelligence


Dozent: Prof. Dr. Alexander Roßmann

Abgabe: 11.07.2021

Eidesstaatliche Erklärung

Ich erkläre eidesstattlich, dass ich die vorliegende Arbeit selbständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet habe.

Reutlingen, den 11.07.2021

A handwritten signature in blue ink, appearing to read 'J. Huynh', is written on a light blue rectangular background.

Thi Yen Vy Huynh

Inhaltsverzeichnis

Eidesstaatliche Erklärung	I
Abbildungsverzeichnis	III
Tabellenverzeichnis	III
Abkürzungsverzeichnis	III
1 Einleitung	1
1.1 Problem- und Fragestellung	1
2 Einordnung in den Bereich Machine Learning	2
3 Datenmodellierung nach dem CRISP-DM-Modell	4
3.1 Business Understanding	5
3.2 Data Understanding	5
3.3 Data Preparation	8
3.4 Modeling und Evaluation	9
3.5 Deployment	12
4 Services und Effekte / Ausblick und Implikationen	14
Literaturverzeichnis	IV

Abbildungsverzeichnis

Abbildung 1: Das CRISP-DM-Modell (Quelle: CRISP DM: Das Data Mining Modell einfach erklärt Data Driven Company).....	4
Abbildung 2: Histogramm der Attribute	6
Abbildung 3: Displot am Beispiel CreditScore	6
Abbildung 4: Korrelationsmatrix	7
Abbildung 5: Countplot der Attribute Geography	7
Abbildung 6: Identifizierung von Ausreißern am Beispiel CreditScore	8
Abbildung 7: VIF	9
Abbildung 8: Klassifikationsreport Decision Tree	10
Abbildung 9: Klassifikationsreport Logistic Regression	11
Abbildung 10: Manuelle Deployment Validation	13
Abbildung 11: Einfluss der Features auf die Kundenabwanderung	14

Tabellenverzeichnis

Tabelle 1: Machine Learning Algorithmen (Quelle: Eigene Darstellung in Anlehnung an Döbel, S. 10; Udemy Kurs)	2
Tabelle 2: Variablen des Datensatzes (Quelle: Eigene Darstellung)	5
Tabelle 3: Entfernung nicht relevanter Spalten	8
Tabelle 4: Confusion Matrix Decision Tree	10
Tabelle 5: Confusion Matrix Logistic Regression	11

Abkürzungsverzeichnis

Abkürzung	Bedeutung
AUC	Area under Curve
CRISP-DM-Modell	Cross Industry Standard Process for Data Mining-Modell
KI	Künstliche Intelligenz
ROC	Receiver Operating Characteristic
VIF	Variance Inflation Factor

1 Einleitung

Unternehmensbereich: Bankenbranche

Die Kundenabwanderungsvorhersage stellt in vielen Unternehmen eine Herausforderung dar und nimmt weiterhin an Bedeutung zu. Dies trifft auch in der Bankenbranche zu (Vgl. Neustadt, o. S.).

Der Grund liegt in der stetig wachsenden Digitalisierung und Erfindung neuer Technologien. In dem Zusammenhang steigt aus Bankensicht die Erkenntnis der Wichtigkeit, bestehende Kunden beizubehalten (Vgl. AI United, o. S.). Dies führt dazu, dass Banken versuchen bestehenden Kunden davon abzuhalten, das Unternehmen zu verlassen, anstatt neue Kunden zu erwerben.

Mithilfe von Machine Learning soll es den Banken ermöglicht werden, Predictions in Hinblick auf die Kundenabwanderung vorherzusagen. Basierend darauf können Banken den Vorhersagen entgegenwirken und Maßnahmen bzw. Strategien einsetzen, um ihre Kunden zu überzeugen weiterhin ihre Mitgliedschaft beizubehalten (Vgl. AI United, o. S.).

1.1 Problem- und Fragestellung

Zusammenfassend lassen sich folgende Punkte für die Fallstudie zusammenfassen:

- Der Erwerb von neuen Kunden ist für die Bank teurer, als bestehende Kunden zu halten.
- Für die Bank ist es von Vorteil zu wissen, welche Eigenschaften die Entscheidung eines Kunden beeinflussen, das Unternehmen zu verlassen.
- Die Abwanderungsvorhersage ermöglicht es den Banken, frühzeitige Kampagnen, Marketingmaßnahmen und -strategien sowie Treureprogramme zu entwickeln und einzusetzen, um die meisten Kunden von einer Abwanderung abzuhalten.

Im Hinblick auf die zusammengefassten Punkte und zur Entwicklung eines solchen Modells, ergeben sich die folgenden Fragestellungen mit der sich die Fallstudie beschäftigt:

- Kann mithilfe von Machine Learning Modellen und Algorithmen, basierend auf vorhanden Daten, die Abwanderungen von bestehenden Kunden vorhergesagt werden?
- Kann mit Machine Learning analysiert werden, welche Variablen die Abwanderung beeinflussen?

Dazu wird in der vorliegenden Fallstudie ein Datensatz von Kaggle verwendet. Im Datensatz sind die Bankenkunden mit ihren Eigenschaften aufgelistet. Durch den Einsatz von Machine Learning Ansätzen können diese Eigenschaften analysiert und überprüft werden. Dies führt dazu, dass ein Kunde identifiziert werden kann, der vermutlich das Unternehmen verlassen wird. Durch die Vorhersage, haben die Banken einen Überblick ihrer Kunden und können dementsprechend Angebote machen, um den Kunden von einer Abwanderung zu verhindern.

2 Einordnung in den Bereich Machine Learning

Die Methodik der Datenanalyse, welche die analytischen Modellbildungen automatisiert wird im Bereich der Künstlichen Intelligenz (KI) unter dem Begriff Machine Learning („Maschinelles Lernen“) zusammengefasst. Im Machine Learning können IT-Systeme durch den Einsatz von Algorithmen automatisch Erkenntnisse und Einblicke aus vorherigen Erfahrungen erzeugen. Mit anderen Worten wird die Generierung von Wissen, das Training von Algorithmen und die Identifizierung von Zusammenhängen aus bekannten Datensätzen durch die Verwendung von Modellen mit Lernalgorithmen automatisiert (Vgl. Wuttke, o. S).

Grundsätzlich werden drei Arten von Machine Learning Algorithmen unterschieden, die in der folgenden Tabelle 1 mit ihren Modellen dargestellt werden.

Lernstil	Beschreibung	Lernaufgabe	Modell
Supervised Learning („überwachtes Lernen“)	Gekennzeichnete Daten sollen mithilfe von bekannten Features vorhergesagt werden	Regression	Lineare Regression
			Klassifikations- und Regressionsbaumverfahren
		Klassifikation	Logistische Regression
			Iterative Dichotomizer
			Stützvektormaschine
Unsupervised Learning („unüberwachtes Lernen“)	Nicht-Gekennzeichnete Daten, die hinsichtlich bestimmter Features ähnlich sind, werden gruppiert	Clustering	K-Means
		Dimensionsreduktion	Kernel Principal Component Analysis
Reinforcement Learning („verstärkendes Lernen“)	Die Ausführung einer bestimmten Aufgabe erfolgt, indem die Algorithmen Erfahrungen sammeln	Sequentielles Entscheiden	Q-Lernen

Tabelle 1: Machine Learning Algorithmen (Quelle: Eigene Darstellung in Anlehnung an Döbel, S. 10; UdeMY Kurs)

Beim Supervised Learning verwendet der Algorithmus einen Datensatz. Dieser besteht aus den Input Daten mit dem zugeordneten, richtigen Output-Daten. Der Algorithmus lernt, indem er seinen tatsächlichen Output mit dem erwarteten Output vergleicht und verbessert dementsprechend sein Modell. Besonders häufig wird der Algorithmus verwendet, wenn Aussagen über Zukunftsdaten getroffen werden sollen, die auf historischen Daten basieren.

Im Gegensatz zu dem vorherigen Algorithmus, befinden sich im Datensatz des Unsupervised Learnings keine zugeordneten Output-Daten. Hier liegt der Fokus auf der Erkundung der Daten und verbunden damit das Erkennen von zugrundeliegenden Mustern. Mit Unsupervised Learning Algorithmen können z. B. Kundensegmente voneinander unterschieden werden oder Produktempfehlungen erfolgen. Das Reinforcement Learning besteht aus einem Agenten, einer Umgebung und einer Aktion. Der Agent erlernt dabei selbständig eine Strategie, um die erhaltenen

Belohnungen zu maximieren. Durch trial-and-error ermittelt der Algorithmus, welche Aktionen die besten Ergebnisse liefert.

Im Hinblick auf die vorliegende Fallstudie und des Datensatzes wird der Supervised Learning Algorithmus näher betrachtet. In Tabelle 1 ist grün markiert, welches Modell für die Vorhersage von Kundenabwanderungen in der Bankenbranche verwendet wird. Der Grund für die Logistische Regression ist, dass in dem ausgewählten Datensatz bereits die Zielvariable „Exited“ existiert. Diese beschreibt, ob ein Kunde in der Vergangenheit abgewandert ist oder nicht. In Anbetracht der Klassifikation wird auch das Modell Decision Tree Classifier auf seine Performance untersucht.

3 Datenmodellierung nach dem CRISP-DM-Modell

Zur Datenmodellierung wird das Cross Industry Standard Process for Data Mining-Modell (CRISP-DM-Modell) verwendet (s. Abb. 1). Dieses Modell unterteilt sich in sechs Phasen, die im Folgenden näher beschreiben werden.

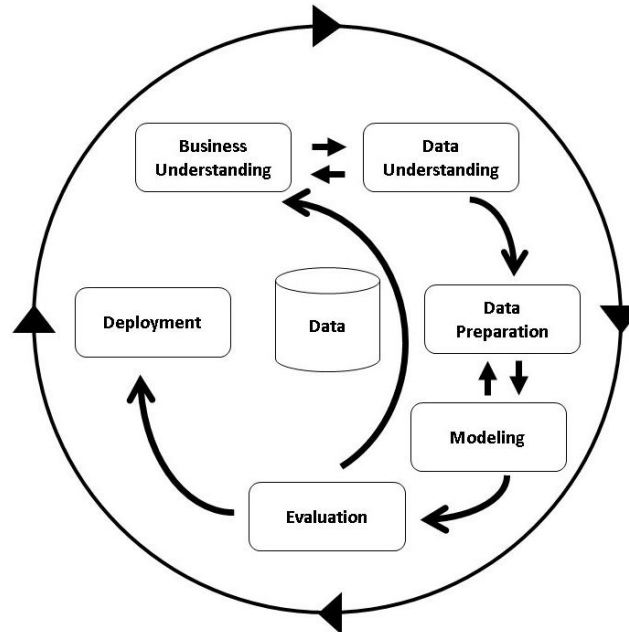


Abbildung 1: Das CRISP-DM-Modell (Quelle: CRISP-DM Definition & Erklärung | Datenbank, DWH & BI Lexikon (datenbanken-verstehen.de))

Der Ausgangspunkt des CRISP-DM-Modells bildet der „[...] Datenstand, der in den Phasen **Business understanding**, **Data understanding**, **Data preparation**, **Modeling**, **Evaluation** und **Deployment** zur Erfüllung der Projektanforderungen und -ziele bearbeitet und ausgewertet wird“ (Gabriel et al., S. 125).

Das Business understanding beschäftigt sich mit der Festlegung von Zielen und Anforderungen sowie die Erstellung eines Projektplans. In der zweiten Phase liegt der Fokus auf dem Datenverständnis, in dem ein Überblick über die verfügbaren Daten verschafft sowie die Datenqualität überprüft und bewertet wird. Darauf folgt die Data preparation, welche für die nächsten Phasen von hoher Relevanz ist. In dieser Phase wird ein finaler Datensatz für das Modeling erstellt. Dabei werden die Daten bereinigt, aufbereitet und in einer geeigneten Form für die nachfolgenden Analysen bereitgestellt. Im Modeling werden Methoden und Modelle zur Klassifizierung, Prognose, Kategorisierung und Abhängigkeitsanalyse aus dem Bereich des Machine Learnings auf den zuvor erstellten finalen Datensatz verwendet. Die Phase Evaluation setzt sich mit der Bewertung und Beurteilung der Ergebnisse auseinander. Abschließend wird das CRISP-DM-Modell mit dem Deployment beendet. Diese Phase umfasst die Aufbereitung der Ergebnisse, um die anschließend zu präsentieren und bei einem Entscheidungsprozess zur Verfügung zu stellen (Vgl. Luber und Litzel, o. S.).

3.1 Business Understanding

In der Bankenbranche stellt die Kundenabwanderung ein wichtiges Thema da. Um in Zukunft die Quote der Kundenabwanderung zu reduzieren, hat die Bankenbranche eine Kundenbindungskampagne entwickelt. Ziel der Kampagne ist es, die Kunden zu identifizieren, welche mit hoher Wahrscheinlichkeit das Unternehmen verlassen werden. Basierend darauf, können Marketingmaßnahmen entwickelt werden und die betroffenen Kunden angesprochen und eine Abwanderung verhindert werden. Damit verbunden werden auch die Kosten reduziert, die für den Erwerb neuer Kunden benötigt werden.

3.2 Data Understanding

Der Datensatz besteht zu Beginn aus 14 Spalten und 10000 Zeilen. In Tabelle 2 ist dargestellt, welche Variablen der Datensatz beinhaltet.

Variable	Beschreibung
RowNumber	Entspricht der Datensatzzeilennummer.
CustomerId	Enthält zufällige Werte, die einem Kunden zur Identifizierung zugeordnet werden.
Surname	Der Nachname eines Kunden.
CreditScore	Kreditscore des Kunden in der Bank.
Geography	Der Standort eines Kunden. Dieser ist entweder Germany, France oder Spain.
Gender	Geschlecht des Kunden (W/M).
Age	Das Alter der Kunden.
Tenure	Bezieht sich auf die Anzahl der Jahre, die der Kunde bereits Kunde der Bank ist.
Balance	Ist ebenfalls ein sehr guter Indikator für die Kundenabwanderung, da Kunden mit einem höheren Guthaben auf ihren Konten die Bank seltener verlassen als Kunden mit einem niedrigeren Guthaben.
NumOfProducts	Bezieht sich auf die Anzahl der Produkte, die ein Kunde über die Bank gekauft hat.
HasCrCard	Gibt an, ob ein Kunde eine Kreditkarte besitzt oder nicht (0 = Nein, 1 = Ja).
IsActiveMember	Mitgliedsstatus (0 = Nein, 1 = Ja).
EstimatedSalary	Höhe des Gehaltseinkommen der Kunden.
Exited	Ob der Kunde die Bank verlassen hat oder nicht (0 = Nein, 1 = Ja).

Tabelle 2: Variablen des Datensatzes (Quelle: Eigene Darstellung)

Die in Tabelle 2 erwähnten Variablen sind dabei die Spalten des Datensatzes, während jeder einzelne Zeileneintrag einen Kunden darstellt. Mit Hilfe dieser Daten soll im weiteren Verlauf klassifiziert werden, ob ein Kunde abspringt oder nicht. Hierzu enthalten die historischen Daten die Zielvariable „Exited“. Diese Spalte gibt die Auskunft darüber, ob ein Kunde abgewandert ist. Der Datensatz beinhaltet keine Duplikate und keine fehlenden Werte (s. Jupiter Notebook, Zeile 9 & 11).

Für das weitere Data Understanding wird eine explorative Datenanalyse durchgeführt. Die Analyse erfolgt separat für die numerischen und für die kategorischen Attributen.

Durch Plotten eines Histogramms aller Attributen, wird die Normalverteilung untersucht (s. Abb. 3).



Abbildung 2: Histogramm der Attribute

Die in Abbildung 2 grün markierten Diagramme zeigen dabei die numerischen Attributen und die anderen die kategorischen Attributen. Im nächsten Schritt werden die numerischen Attributen einzeln untersucht, in dem jeweils ein normaler Distplot und im Anschluss ein Distplot mit der Unterteilung „Exited“ durchgeführt wird. Ein Beispiel ist in Abbildung 3 gezeigt.

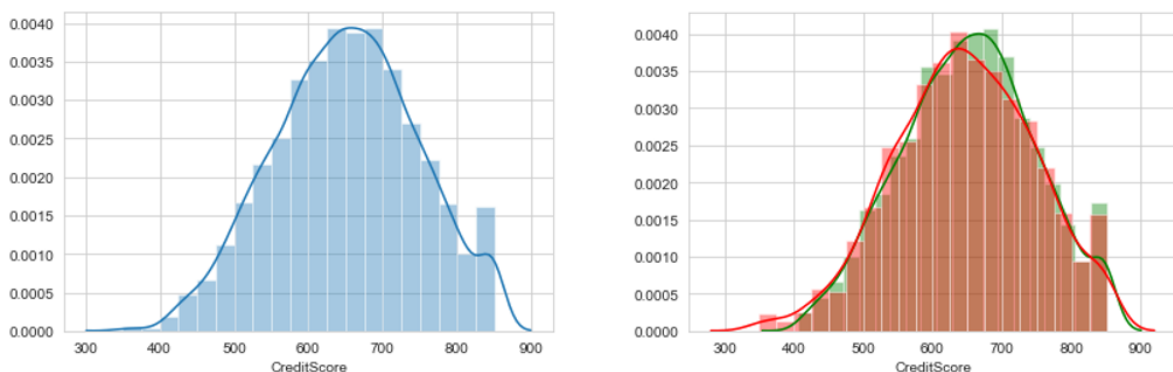


Abbildung 3: Displot am Beispiel CreditScore

Im linken Diagramm (blau) ist ein normaler Displot des CreditScores abgebildet. Auf den ersten Blick sieht der Verlauf wie eine Normalverteilung aus. Es stellt sich bei genauerem Hinsehen die Frage, ob sich im niedrigeren Bereich einige Ausreißer befinden. Die Untersuchung der Ausreißer sind Bestandteil des Kapitels Data Preparation. In der rechten Abbildung ist der Displot aufgeteilt in „Exited“ und „No Exited“ dargestellt. Beide Verläufe sind bei diesem Attribut sehr ähnlich und zeigt, dass die Höhe des CreditScores eines Kunden keinen Einfluss auf die Kundenabwanderung hat.

Die Distplots zu den anderen Attributen sind im Jupyter Notebook unter 3.2.1 zu finden. Bei der Untersuchung der Attribute „Tenure“ stellt sich heraus, dass dieser auch als kategoriale Attribute untersucht werden kann.

Nach Darstellung aller Displots der numerischen Attributen wird eine Korrelationsmatrix veranschaulicht, um die Multikollinearität zu überprüfen (s. Abb. 4).

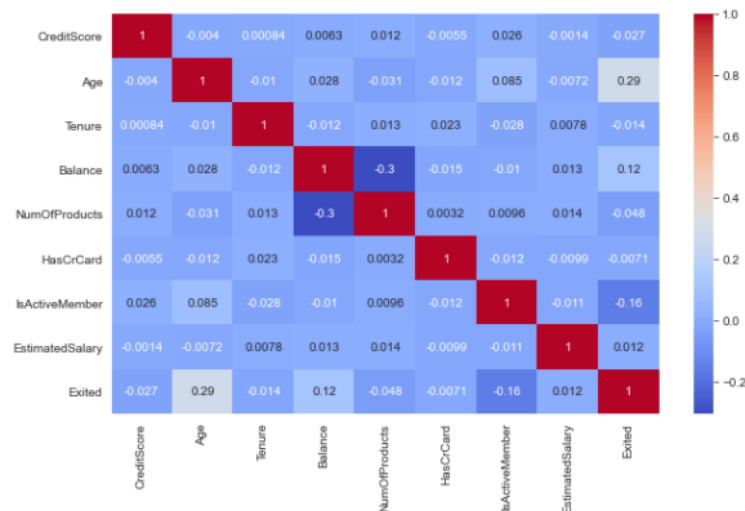


Abbildung 4: Korrelationsmatrix

In der Korrelationsmatrix sind die höchsten Korrelationwerte zwischen NumOfProducts und Balance mit einem Wert von -0,3 sowie die Korrelation zwischen Age und Exited mit einem Wert von +0,29. Basierend auf den niedrigen Korrelationswerten fällt keine eindeutige Multikollinearität auf. In der Data Preparation wird diese nochmal mit dem Variance Inflation Factor (VIF) überprüft.

Das Data Understanding der kategorischen Attributen wird mit Hilfe von Countplots untersucht. In der Folgenden Abbildung 5 ist dies an der Attribute „Geography“ dargestellt.

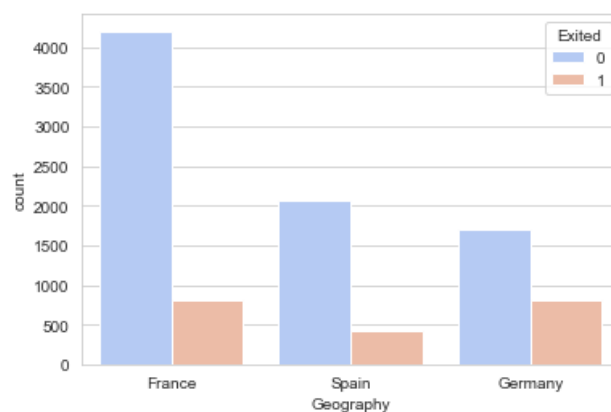


Abbildung 5: Countplot der Attribute Geography

Im Countplot fällt auf, dass die meisten Kunden aus Frankreich kommen und die wenigsten aus Germany. Die Abwanderung jedoch ist bei beiden Ländern gleich groß. Die weiteren Countplots sind im Jupyter Notebook vorzufinden unter dem Punkt 3.2.2.

Bei der Attribute NumOf Products ist zu sehen, dass die Kunden, die vier Produkte verwendet haben, alle abgewandert sind (60 von 60 Kunden mit vier Produkten sind abgewandert).

3.3 Data Preparation

Nachdem Data Understanding werden zu Beginn der Data Preparation die Spalten entfernt, die für die weitere Analyse keinen Einfluss auf die Kundenabwanderung haben und nicht benötigt werden. Dazu gehören:

1. RowNumber	Die RowNumber entspricht dem Zeilenindex und wird nicht benötigt.
2. CustomerID	Sowohl die CustomerID als auch der Surname sind keine Einflussfaktoren, ob der Kunde die Bank verlassen wird.
3. Surname	
4. Tenure	Im Data Understanding ist zu erkennen, dass die meisten Kunden zwischen einem und neun Jahren Kund in der Bank sind und die Verteilung der Attribute „Exited“ ziemlich gleich ist. Es ist somit kein Zusammenhang zwischen dem Tenure und dem Exited zusehen.

Tabelle 3: Entfernung nicht relevanter Spalten

Danach erfolgt die nähere Untersuchung der Ausreißer, die im Data Understanding zu sehen sind und erwähnt werden (s. Jupyter Notebook ab Zeile 44). Dazu werden Boxplots verwendet, welche die Ausreißer abbilden. Ein Beispiel zum Credit Score ist in der folgenden Abbildung 6 dargestellt.

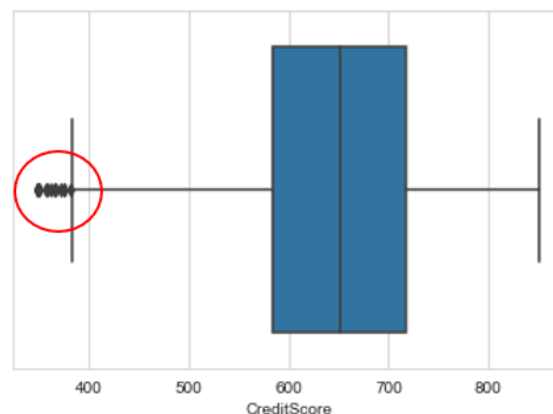


Abbildung 6: Identifizierung von Ausreißern am Beispiel CreditScore

Es ist zu erkennen, dass sich im vorderen Bereich des CreditScores Ausreißer befinden. Aus diesem Grund werden diese entfernt, indem das 1%-Quantil entfernt wird. Weitere Ausreißer sind bei den Attributen Age und NumOfProducts zu finden (s. Jupyter Notebook, unter 3.3.1).

Nach Entfernung der Ausreißer wird nochmal eine Korrelationsmatrix geplottet, die erneut keine eindeutige Multikollinearität ausweist. Diese wird nun mit dem VIF überprüft. In Abbildung 7 ist die VIF Tabelle zusehen.

	feature	VIF
0	CreditScore	20.710187
1	Age	15.804177
2	Balance	3.388799
3	EstimatedSalary	3.882472
4	Geography_Germany	1.790507
5	Geography_Spain	1.491191
6	Gender_Male	2.176027
7	NumOfProducts_2	2.179566
8	NumOfProducts_3	1.066362
9	HasCrCard_1	3.275108
10	IsActiveMember_1	1.993649

Abbildung 7: VIF

Der höchste VIF-Wert ist beim CreditScore mit 20,7 und wird aus dem Datensatz entfernt. Nach erneuter Überprüfung des VIF sind keiner weiteren Features vorhanden, die einen VIF-Wert größer 10 ausweisen. Zudem findet in der Data Preparation das Feature Engineering statt, in dem die kategorischen Attributen in Dummy_Variablen umgewandelt werden. Die kategorischen Variablen sind die Geography, das Gender, die NumofProducts, der HasCrCard und IsActiveMember.

Am Ende der Data Preparation besteht der finale Datensatz aus 9323 Zeilen und 11 Spalten.

Die Zielvariable y der Fallstudie ist Exited, während die restlichen Variablen zum Input X gehören.

Bevor die Phase des Modeling beginnt muss das Feature Scaling durchgeführt werden. Dazu werden die numerischen Daten (Age, Balance und EstimatedSalary) standardisiert. Im Anschluss wird der Datensatz in Test- und Trainingsdaten aufgeteilt. Hier wird eine Testsize von 20 % und ein random_state von 365 verwendet.

3.4 Modeling und Evaluation

Das vorliegende Machine Learning Modell ist eine binäre Klassifizierung. Zur Untersuchung werden Klassifizierungsmodelle verwendet. Für die vorliegende Fallstudie wird zum einen der Decision Tree Classifier und zum anderen die Logistische Regression verwendet. Aufgrund des unbalancierten Datensatzes reicht die Metrik Accuracy alleine nicht aus, weshalb weitere Metriken zur Evaluation der Performance betrachtet werden. Das ist zu einen der Classification Report und zum anderen die Confusion Matrix.

Decision Tree Classifier

Nach der Instanziierung des Decision Tree Classifiers und des Fittings der Trainingsdaten, wird die folgende Klassifikationsmatrix ausgegeben:

Trainingsdaten:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	6022
1	1.00	1.00	1.00	1436
accuracy			1.00	7458
macro avg	1.00	1.00	1.00	7458
weighted avg	1.00	1.00	1.00	7458
Testdaten:				
	precision	recall	f1-score	support
0	0.87	0.88	0.87	1502
1	0.48	0.47	0.47	363
accuracy			0.80	1865
macro avg	0.68	0.67	0.67	1865
weighted avg	0.80	0.80	0.80	1865

Abbildung 8: Klassifikationsreport Decision Tree

Die Performance auf den Trainingsdaten beträgt eine 1.00 wohingegen die Performance auf den Testdaten 20 % geringer ist. Aus diesem Grund ist ein deutliches und starkes Overfitting des Modells zu erkennen.

Die Ergebnisse aus der Confusion Matrix sind wie folgt:

<p><u>True Negative (TN) = 1316</u></p> <p>1316 Datenpunkte der negativen Klasse (keine Kundenabwanderung) werden vom Modell richtig klassifiziert.</p>	<p><u>False Positive (FP) = 186</u></p> <p>186 Datenpunkte der negativen Klasse werden fälschlicherweise vom Modell als positive Klasse klassifiziert. Das heißt, dass die Kunden nicht abgewandert sind, aber vom Modell als abgewandert klassifiziert werden.</p>
<p><u>False Negative (FN) = 185</u></p> <p>185 Datenpunkte der positiven Klasse werden fälschlicherweise vom Modell als negative Klasse klassifiziert. Das heißt, dass die Kunden abgewandert sind, aber vom Modell als keine Abwanderung klassifiziert werden.</p>	<p><u>True Positive (TP) = 178</u></p> <p>178 Datenpunkte der positiven Klasse (Kundenabwanderung) werden richtig von Modell klassifiziert.</p>

Tabelle 4: Confusion Matrix Decision Tree

Darüber hinaus beträgt die Precision 49 %. Diese gibt an, wie viele der positiv vorhergesagten Fälle sich am Ende als wirklich als positive Fälle ergeben haben ($TP/(TP + FP)$). Genauso groß ist Wert beim Recall (49 %), der angibt, wie viele richtige positive Fälle von unserem Modell erfolgreich vorhergesagt werden konnten ($TP/(TP+FN)$). Die Accuracy gibt an, wie oft der vorhergesagte Wert dem realen Zielwert entspricht. Bei dem vorliegenden Modell beträgt dieser Wert nur 80%, weshalb im Folgenden noch das Logistic Regression Modell gebildet wird.

Logistic Regression

Wie bei dem Modell zuvor wird zu Beginn das LogisticRegression Modell instanziiert und im Anschluss auf das Trainingsset (Y_train, X_train) trainiert und gefittet.

Daraufhin erfolgen die Vorhersagen für X_test und X_train. Die Accuracy auf dem Testdatensatz beträgt 0.85 und zeigt im Vergleich zum Decision Tree mit Wert von 0.78 einen besseren Accuracy.

Der Klassifikationsreport sieht wie folgt aus:

Trainingsdaten:					
	precision	recall	f1-score	support	
0	0.87	0.96	0.91	6022	
1	0.71	0.41	0.52	1436	
accuracy			0.85	7458	
macro avg	0.79	0.69	0.72	7458	
weighted avg	0.84	0.85	0.84	7458	
Testdaten:					
	precision	recall	f1-score	support	
0	0.87	0.96	0.91	1502	
1	0.71	0.42	0.53	363	
accuracy			0.85	1865	
macro avg	0.79	0.69	0.72	1865	
weighted avg	0.84	0.85	0.84	1865	

Abbildung 9: Klassifikationsreport Logistic Regression

Die Accuracy als auch die anderen Werte bei dem Test- und Trainingsdatensatz sind sehr ähnlich. Aus diesem Grund wird von keinem Overfitting ausgegangen. Die Confusion Matrix bei der Logistischen Regression zeigt folgende Werte:

<p><u>True Negative (TN) = 1439</u></p> <p>1439 Datenpunkte der negativen Klasse (keine Kundenabwanderung) werden vom Modell richtig klassifiziert.</p>	<p><u>False Positive (FP) = 63</u></p> <p>63 Datenpunkte der negativen Klasse werden fälschlicherweise vom Modell als positive Klasse klassifiziert. Das heißt, dass die Kunden nicht abgewandert sind, aber vom Modell als abgewandert klassifiziert werden.</p>
<p><u>False Negative (FN) = 209</u></p> <p>209 Datenpunkte der positiven Klasse werden fälschlicherweise vom Modell als negative Klasse klassifiziert. Das heißt, dass die Kunden abgewandert sind, aber vom Modell als keine Abwanderung klassifiziert werden.</p>	<p><u>True Positive (TP) = 154</u></p> <p>154 Datenpunkte der positiven Klasse (Kundenabwanderung) werden richtig von Modell klassifiziert.</p>

Tabelle 5: Confusion Matrix Logistic Regression

Precision und Recall zeigen ein realistischeres Bild des Modells. Dieses erzielt eine Precision von rund 71% und einen Recall von 42%. Aufgrund des niedrigen Recall-Werts, der aber für den Anwendungsfall deutlich wichtiger ist, sollte der Recall auf

Kosten der Precision erhöht werden. Dazu sollte der Threshold der logistischen Regression analysiert und dementsprechend angepasst werden.

Im Notebook werden für beide Modelle die Receiver Operating Characteristic (ROC) Kurven abgebildet, die das Verhältnis zwischen den True Positives und den False Positives darstellt. Die Area under Curve (AUC) bei dem Decision tree Classifier zeigt einen Wert von 0.68 auf während die Logistische Regression mit einem Wert von 0.83 ein besseres Ergebnis liefert.

Aufgrund der besseren Ergebnisse bei dem Modell der Logistischen Regression mit einer Accuracy von 85 % wird dieses für die Klassifizierung herangezogen. Bevor das Deployment erfolgt, muss die Recall-Rate erhöht werden. Dazu werden verschiedene Threshold untersucht.

3.5 Deployment

Das Deployment kann über zwei verschiedene Arten erfolgen. Zum einen über die IBM Cloud und zum anderen Manuell im Notebook.

IBM Cloud

Das finale Model wird im Rahmen dieser Hausarbeit als Web-Service über die IBM Cloud – Watson Machine Learning bereitgestellt. Dazu wird ein API-Key für das Modell generiert. Im Jupyter Notebook wird dieser zusammen mit der URL der Region, in welcher das Modell bereitgestellt werden soll, dem neuen API-Client übergeben. Im Anschluss wird das Modell registriert und ein API-Endpoint wird erstellt. An dem API-Endpoint kann anschließend per http-post Werte an das Modell übermittelt werden. Als Antwort wird die Vorhersage des Modells übergeben. Der ausführliche Code ist im ersten als auch im zweiten Jupyter Notebook unter dem Punkt 3.5 Deployment und Anwendung des Modells zu finden. Im zweiten Jupyter Notebook, welches von **Watson Studios** heruntergeladen wurde, um die Ergebnisse des Deployments und der Validation zu zeigen, werden über die IBM Cloud von zwei Kunden die Abwanderungsquote vorhergesagt. Die Ergebnisse werden nach einer lokalen Überprüfung richtig vorhergesagt (s. Watson Jupyter Notebook, Zeile 119).

Manuelle Deployment Validation

Die Manuelle Deployment Validation wird im Jupyter Notebook durchgeführt (s. Jupyter Notebook ab Zeile 104). Basierend auf einem beliebig ausgewählten Kunden aus dem Datensatz wird ausgewählten Modell vorhergesagt, ob der Kunde abwandern wird oder nicht. Dies sieht wie folgt aus:


```
# Überblick über ausgewählten Kunden
customer_df

Age          -0.400692
Balance      -1.224498
EstimatedSalary  1.444329
Geography_Germany  0.000000
Geography_Spain  1.000000
Gender_Male    1.000000
NumOfProducts_2  1.000000
NumOfProducts_3  0.000000
HasCrCard_1    0.000000
IsActiveMember_1  1.000000
Name: 9409, dtype: float64

# Prediction ausführen
cust_pred = logistic_model.predict([customer_df])

# Ergebnis interpretieren
def check_prediction(pred):
    if pred[0] == 1:
        print("Der Kunde wird vermutlich abwandern! Customer Re
    else:
        print("Der Kunde wird vermutlich nicht abwandern.")

check_prediction(cust_pred)

Der Kunde wird vermutlich nicht abwandern.
```

Abbildung 10: Manuelle Deployment Validation

4 Services und Effekte / Ausblick und Implikationen

Im Rahmen der vorliegenden Fallstudie konnte dargelegt werden, dass eine binäre Klassifikation mit Hilfe von Kundeneigenschaften in der Bankenbranche möglich ist. Die Accuracy die hier erreicht werden konnte beträgt 85 % und zeigt ein sehr gutes Ergebnis. Darüber hinaus konnte analysiert werden, welche Features besonders zu einer Kundenabwanderung führen und welche Features dazu beitragen, dass die Kunden gegenüber der Bank loyal bleiben und das Unternehmen nicht verlassen. In Abbildung 11 ist dargestellt, wie die Features die Abwanderung beeinflussen.

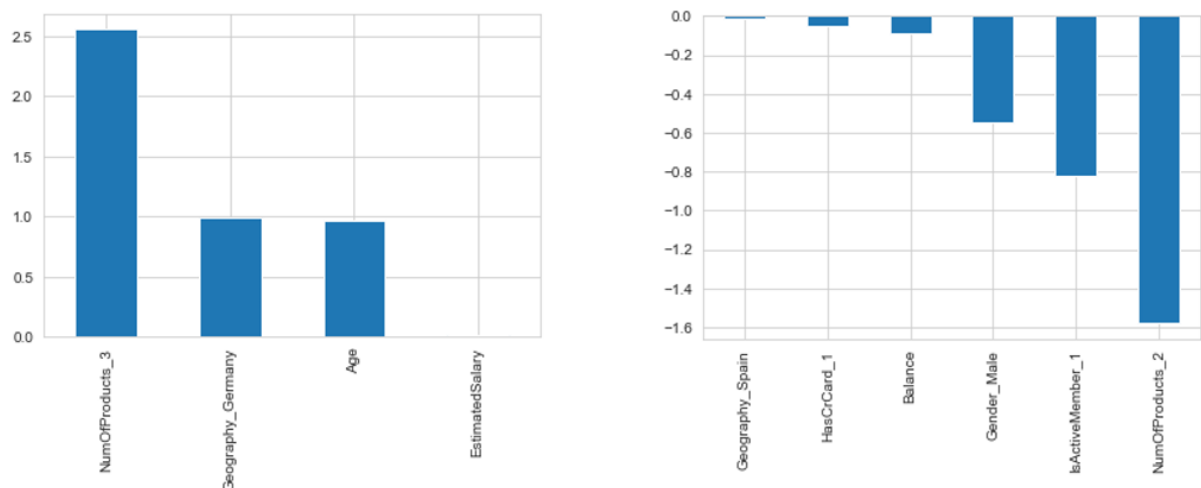


Abbildung 11: Einfluss der Features auf die Kundenabwanderung

Das linke Diagramm zeigt die Features, auf die sich Bank fokussieren sollte. Besonders wichtig für die Bank sind Kunden, die drei oder mehr Produkte bei der Bank kaufen. Denn das Feature NumOfProducts_3 zeigt den größten Einfluss auf die Kundenabwanderung an. Die Bank sollte zukünftig nicht mehr den Fokus darauflegen, den Kunden mehr verkaufen zu wollen. Ein weiteres Feature, dass die Abwanderung stark beeinträchtigt ist die Herkunft aus Deutschland. Im Hinblick auf diesen Aspekt, sollte die Bank näher analysieren, weshalb die Abwanderung der deutschen Kunden im Vergleich zu den anderen beiden Ländern so hoch ist. Features die Kunden von einer Abwanderung abhalten sind vor allem, wenn die Kunden ein bis zwei Produkte der Bank gekauft haben und besitzen. Zusätzlich sollten Banken ihre Kunden dazu verleiten, ihre Mitgliedschaft aktiver zu verwenden, da dieser ebenfalls ein Grund dafür ist, dass die Kundenabwanderung verringert.

Durch die Identifikation und Klassifikation bestimmter Kundeneigenschaften die möglicherweise zu einer Kundenabwanderung führen, können Banken effizientere Marketingmaßnahmen und -strategien entwickeln und verhindern, dass Kunden das Unternehmen verlassen, um den Gewinn zu verbessern.

Durch das Deployment in der Cloud, kann das Modell als Service angeboten werden, in dem zukünftige Kundendaten eingeben und direkt ohne technische Infrastruktur ausgewertet werden können. Features die einen großen Einfluss haben, können schneller identifiziert werden und schnellere Marketingmaßnahmen können entwickelt und umgesetzt werden.

Literaturverzeichnis

AI United (o. J.): *KI für Banken*. Verfügbar unter <http://www.ai-united.de/ki-fuer-banken/>. [Letzter Zugriff: 05.07.2021]

CRISP-DM Definition & Erklärung | Datenbank, DWH & BI Lexikon (datenbanken-verstehen.de) [Letzter Zugriff: 05.07.2021]

Datensatz (2020): *Churn for Bank Customers*. Verfügbar unter <https://www.kaggle.com/mathchi/churn-for-bank-customers>. [Letzter Zugriff: 05.07.2021]

Döbel, Inga, et al.: *Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung*. Herausgeber: Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. München, 2018. Verfügbar unter: <http://publica.fraunhofer.de/dokumente/N-497408.html> [Letzter Zugriff: 05.07.2021]

Gabriel, Roland; Gluchowski, Peter; Pastwa, Alexander (2009): *Data warehouse & data mining*. Verfügbar unter: [Data warehouse & data mining - Google Books](#) [Letzter Zugriff: 05.07.2021]

Luber und Litzel (2019): *Was ist CRISP-DM*. Veröffentlicht am 10.04.2019. Verfügbar unter: [Was ist CRISP-DM? \(bigdata-insider.de\)](#) [Letzter Zugriff: 05.07.2021]

Neustadt (2019): *Künstliche Intelligenz gegen Kundenabwanderung (Customer Churn)*. Veröffentlicht am 29. Juli 2019. Verfügbar unter <https://www.linkedin.com/pulse/k%C3%BCnstliche-intelligenz-gegen-kundenabwanderung-churn-sergej-neustadt/>. [Letzter Zugriff: 05.07.2021]

Wuttke, Laurenz (o. J.): *Machine Learning: Definition, Algorithmen, Methoden und Beispiele*. Verfügbar unter: <https://datasolut.com/was-ist-machine-learning/> [Letzter Zugriff: 05.07.2021]