



UNIVERSITY OF APPLIED SCIENCES

Department of Computer Science

Projektbericht für das Modul *Learning and Softcomputing*

Kundenzufriedenheitsvorhersage mithilfe von Machine Learning

Abgabedatum:

29. Juli 2025

Eingereicht von:

Maria Besier

Matrikelnummer 106697

Referent:

**Prof. Dr. Christian-Arved
Bohn**

Fachhochschule Wedel

Feldstraße 143

22880 Wedel

Inhaltsverzeichnis

Abbildungsverzeichnis	III
Tabellenverzeichnis	IV
1 Einleitung	1
1.1 Problemstellung	2
1.2 Klassische Verfahren zur Lösung des Problems	2
1.3 Allgemeine Idee des Projekts	4
2 Durchführung	5
2.1 Strategie und gewählte Modellarchitektur	5
2.1.1 Topologie des neuronalen Netzes	5
2.2 Datensatz und Datenvorverarbeitung	6
2.2.1 Datenquelle	6
2.2.2 Beschreibung der Merkmale	6
2.2.3 Datenvorbereitung	7
2.3 Training des neuronalen Netzes	7
2.3.1 Hyperparameter	7
2.3.2 Trainingsdauer und Ressourcen	8
2.3.3 Probleme und Herausforderungen	8
2.4 Zusammenfassung der Durchführung	8
3 Ergebnisse	9
3.1 Gradient Boosting	9
3.2 Random Forest	10
3.3 Logistische Regression	11
3.4 Support Vector Machine (SVM)	12
3.5 Neuronales Netz (ANN)	13
3.6 K-Nearest Neighbors (KNN)	14
3.7 Modellvergleich	15
4 Zusammenfassung und Ausblick	16
4.1 Ergebnisbewertung	16
4.2 Optimierungspotenziale	17
4.3 Anwendungsperspektiven	17
4.4 Kritische Reflexion	18
4.5 Ausblick	19
5 Programmcode	21
Literatur	22

Abbildungsverzeichnis

3.1	Visualisierung der Vorhersagen – Gradient Boosting	10
3.2	Visualisierung der Vorhersagen – Random Forest	11
3.3	Visualisierung der Vorhersagen – Logistische Regression	12
3.4	Visualisierung der Vorhersagen – Support Vector Machine	13
3.5	Visualisierung der Vorhersagen – ANN	14
3.6	Visualisierung der Vorhersagen – Gradient Boosting	15

Tabellenverzeichnis

2.1	Verwendete Hyperparameter beim Training des ANN	7
3.1	Klassifikationsbericht – Gradient Boosting	9
3.2	Klassifikationsbericht – Random Forest	10
3.3	Klassifikationsbericht – Logistische Regression	11
3.4	Klassifikationsbericht – Support Vector Machine	12
3.5	Klassifikationsbericht – ANN	13
3.6	Klassifikationsbericht – K-Nearest Neighbors	14
3.7	Modellvergleich aller Verfahren	15

1

Einleitung

In der zunehmend wettbewerbsorientierten Unternehmenswelt gewinnt die langfristige Bindung von Kunden an Bedeutung. Die Akquise neuer Kunden ist nicht nur mit hohen Marketing- und Vertriebskosten verbunden, sondern birgt auch Unsicherheiten hinsichtlich ihrer Rentabilität. Umso wichtiger ist es für Unternehmen, frühzeitig zu erkennen, welche ihrer bestehenden Kunden ein erhöhtes Risiko zeigen, ihre Geschäftsbeziehung zu beenden, was als Kundenabwanderung oder Churn bezeichnet wird. Die rechtzeitige Identifikation dieser gefährdeten Kundengruppen ermöglicht gezielte Maßnahmen zur Kundenbindung und stellt damit einen zentralen Hebel zur Umsatzsicherung dar.

Das vorliegende Projekt widmet sich der Aufgabe, Kundenabwanderung mithilfe moderner Methoden des maschinellen Lernens und Soft Computing vorherzusagen. Grundlage hierfür bildet ein strukturierter Datensatz, der verschiedene soziodemografische, verhaltensbezogene und nutzungsabhängige Informationen über Kunden enthält, ergänzt um die Angabe, ob die jeweilige Person ihren Vertrag innerhalb der letzten sechs Monate gekündigt hat. Ziel ist es, aus diesen Informationen ein Modell zu entwickeln, dass künftig auf Basis neuer Kundendaten eine möglichst zuverlässige Vorhersage darüber treffen kann, ob eine Kündigung zu erwarten ist.

1.1 Problemstellung

Die zentrale Fragestellung des Projekts lautet:

Wie kann auf Grundlage vorliegender Kundendaten die Wahrscheinlichkeit vorhergesagt werden, dass ein Kunde innerhalb eines bestimmten Zeitraums kündigt?

Dabei gibt es eine Reihe von Herausforderungen:

Datenqualität und -vorverarbeitung: Häufig enthalten reale Datensätze unvollständige, verrauschte oder fehlerhafte Informationen. Eine sorgfältige Vorverarbeitung, wie z.B. durch Normalisierung, One-Hot-Encoding kategorialer Merkmale und der Umgang mit fehlenden Werten, ist entscheidend für die Modellgüte.

Klassenungleichgewicht (Imbalanced Data): In typischen Churn-Datensätzen ist die Anzahl der Kunden, die tatsächlich kündigen, deutlich geringer als die der verbleibenden. Klassifikationsalgorithmen neigen dazu, Mehrheitsklassen zu bevorzugen, was zu einer fehlerhaften Einschätzung der Minderheitsklasse führt.

Modellauswahl und -anpassung: Unterschiedliche Lernverfahren zeigen verschiedene Stärken und Schwächen in Bezug auf Vorhersagegenauigkeit, Rechenzeit, Interpretierbarkeit und Robustheit gegenüber Ausreißern.

Evaluierung der Modelleistung: Die bloße Genauigkeit („Accuracy“) reicht als Qualitätskriterium nicht aus – es müssen geeignete Metriken wie Precision, Recall, F1-Score oder die ROC-AUC betrachtet werden, insbesondere bei unausgeglichene Klassenverteilungen.

1.2 Klassische Verfahren zur Lösung des Problems

Die Vorhersage von Kundenabwanderung gehört zu den klassischen Anwendungsfällen im Bereich Data Mining und Predictive Analytics. In der Vergangenheit wurden zahlreiche Methoden eingesetzt, darunter:

Logistische Regression: Ein statistisches Verfahren, das die Wahrscheinlichkeit einer binären Zielvariable modelliert. Es ist leicht interpretierbar und wird häufig als

1 Einleitung

Baseline-Modell verwendet. Die logistische Regression nimmt lineare Zusammenhänge zwischen unabhängigen Variablen und der logit-transformierten Zielvariable an [Jam+13].

Entscheidungsbäume und Random Forests: Diese Algorithmen erstellen auf Basis von Entscheidungsregeln eine baumartige Struktur. Während einzelne Bäume leicht interpretierbar sind, bieten Random Forests als Ensemble aus vielen Entscheidungsbäumen eine hohe Genauigkeit und Robustheit durch Mehrheitsvoting. Sie eignen sich gut für strukturierte Daten mit gemischten Merkmalstypen [DH18].

Support Vector Machines (SVM): SVMs maximieren den Abstand zwischen Klassen und können auch nicht-lineare Zusammenhänge durch den Einsatz von Kernels modellieren. Sie sind besonders nützlich bei hochdimensionalen Daten [BC00].

Künstliche neuronale Netze (ANN): Künstliche neuronale Netze sind rechnergestützte Modelle, die sich am Aufbau biologischer Nervensysteme orientieren. Sie bestehen aus Neuronen, die in Schichten (Input-, Hidden- und Output-Schichten) organisiert sind. Jedes künstliche Neuron berechnet eine gewichtete Summe seiner Eingaben und wendet anschließend eine nichtlineare Aktivierungsfunktion (z.B. Sigmoid, ReLU, Tanh) an. ANNs sind in der Lage, komplexe, nichtlineare Zusammenhänge in Daten zu erfassen und zu modellieren, insbesondere wenn klassische lineare Verfahren versagen. Die Lernphase erfolgt typischerweise durch Backpropagation in Kombination mit einem Optimierungsverfahren wie Stochastic Gradient Descent (SGD), wobei die Gewichte der Neuronen schrittweise angepasst werden, um einen Fehlermaßstab (z.B. Kreuzentropie) zu minimieren.

In den letzten Jahren haben insbesondere tiefe neuronale Netze (Deep Neural Networks, DNNs), die aus vielen versteckten Schichten bestehen, signifikante Fortschritte in Bereichen wie Bild- und Spracherkennung, Natural Language Processing (NLP) und Zeitreihenanalyse erzielt. Dieser Erfolg basiert auf der Fähigkeit tiefer Netzwerke, hierarchische Merkmalsrepräsentationen zu lernen. Allerdings geht diese hohe Modellkapazität zulasten der Interpretierbarkeit: Während einfache Modelle (z.B. Entscheidungsbäume) klare Regeln liefern, sind ANN-Entscheidungen oft als „Black Box“ schwer nachvollziehbar. Deshalb werden zunehmend Methoden der erklärbaren KI (XAI) entwickelt, um neuronale Modelle transparenter zu machen (z.B. Layer-wise Relevance Propagation, SHAP, LIME) [Zha+21].

k-Nearest Neighbors (k-NN): Dieses einfach zu implementierende Verfahren weist neuen Datenpunkten die Klasse ihrer nächsten Nachbarn im Merkmalsraum zu. Es ist rechenintensiv bei großen Datensätzen und anfällig für irrelevante Merkmale [HUU+24].

1.3 Allgemeine Idee des Projekts

In diesem Projekt soll ein praxistaugliches System zur Vorhersage von Kundenabwanderung entwickelt werden, das auf maschinellem Lernen und Prinzipien des Soft Computing basiert. Dazu werden verschiedene Modellierungsansätze erprobt, verglichen und evaluiert. Die allgemeine Vorgehensweise gliedert sich wie folgt:

Explorative Datenanalyse: Untersuchung des Datensatzes auf Verteilungen, Korrelationen, Ausreißer und potenziell relevante Merkmale.

Vorverarbeitung und Feature Engineering: Transformation der Rohdaten in ein geeignetes Format für maschinelles Lernen, inklusive Skalierung, Encoding und ggf. Konstruktion neuer Merkmale.

Modellauswahl und Training: Vergleich mehrerer Lernverfahren (z.B. Entscheidungsbaum, logist. Regression, neuronales Netz), inklusive Hyperparameteroptimierung mittels Cross-Validation.

Evaluierung: Einsatz geeigneter Metriken zur Leistungsbewertung, insbesondere unter Berücksichtigung des Klassenungleichgewichts. Modellinterpretation und Anwendungsszenarien: Identifikation der wichtigsten Einflussfaktoren auf das Kündigungsverhalten und Diskussion praktischer Maßnahmen für ein Unternehmen zur Churn-Prävention.

Das Projekt verfolgt sowohl einen praktischen als auch einen akademischen Anspruch: Zum Einen wird eine reale Fragestellung adressiert, zum anderen sollen zentrale Inhalte der Vorlesungen zu Learning and Softcomputing angewendet, vertieft und reflektiert werden. Die gewonnenen Erkenntnisse sollen dazu beitragen, das Potenzial datengetriebener Methoden für betriebliche Entscheidungsprozesse besser zu verstehen.

2

Durchführung

2.1 Strategie und gewählte Modellarchitektur

Das Ziel des Projekts ist die Prädiktion der Kundenabwanderung (*Churn Prediction*) auf Basis verhaltens- und vertragsbezogener Kundendaten eines Telekommunikationsunternehmens. Zu diesem Zweck wurden sowohl klassische Modelle des überwachten Lernens (u. a. Logistische Regression, Random Forest, Gradient Boosting, Support Vector Machines und k-Nearest Neighbors) als auch ein künstliches neuronales Netz (ANN) eingesetzt und vergleichend analysiert.

Der Fokus dieses Abschnitts liegt auf dem Aufbau, der Trainingsstrategie und der Leistungsbewertung des eingesetzten ANN. Dieses wurde als *Multi-Layer Perceptron (MLP)* mit drei vollständig verbundenen Schichten (Dense-Layers) implementiert.

2.1.1 Topologie des neuronalen Netzes

Die Architektur des neuronalen Netzes ist wie folgt definiert:

Eingabeschicht (Input Layer): 37 Neuronen, entsprechend der Anzahl der kodierten Eingabevariablen.

Verborgene Schichten (Hidden Layers):

- Erste verborgene Schicht: 64 Neuronen, ReLU-Aktivierungsfunktion.
- Zweite verborgene Schicht: 32 Neuronen, ReLU-Aktivierungsfunktion.

Ausgabeschicht (Output Layer): Ein Neuron mit Sigmoid-Aktivierung zur binären

Klassifikation ($P(y = 1|x) \in [0, 1]$).

Das Netzwerk modelliert somit eine Funktion $f : \mathbb{R}^{37} \rightarrow [0, 1]$, wobei die Ausgabe die geschätzte Kündigungswahrscheinlichkeit eines Kunden darstellt.

2.2 Datensatz und Datenvorverarbeitung

2.2.1 Datenquelle

Der verwendete Datensatz ist ein frei verfügbarer Beispieldatensatz von IBM, betitelt *Telco Customer Churn*. Dieser enthält Informationen zu 7043 Kunden eines Telekommunikationsanbieters und ist im CSV-Format veröffentlicht.¹

2.2.2 Beschreibung der Merkmale

Die folgenden Merkmalskategorien sind im Datensatz enthalten:

Demografische Merkmale: Geschlecht, Senior Status, Partnerschaft, Kinder.

Vertragsinformationen: Vertragslaufzeit, Zahlungsmethode, Paperless Billing.

Nutzungsverhalten: Telefon- und Internetdienste, Zusatzangebote wie Online-Security, Tech Support etc.

Monetäre Merkmale: Monatliche Gebühren (MonthlyCharges), Gesamtgebühren (TotalCharges).

Zielvariable: Kündigung (Churn = „Yes“ oder „No“).

¹<https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/>

2.2.3 Datenvorbereitung

Die Vorverarbeitung der Daten erfolgte in mehreren Schritten:

1. *Typkonvertierung und Bereinigung*: Die Spalte TotalCharges wurde in numerische Form überführt, fehlerhafte oder fehlende Werte entfernt.
2. *Kategorische Merkmale*: Alle kategorialen Variablen wurden per One-Hot-Encoding in binäre Indikatorvariablen umgewandelt.
3. *Numerische Skalierung*: Numerische Merkmale wurden mittels Standardisierung ($z = \frac{x-\mu}{\sigma}$) skaliert.
4. *Zielvariable*: Die binäre Zielvariable Churn wurde in 0 (Nein) bzw. 1 (Ja) kodiert.

Nach der Transformation umfasste der finale Input-Vektor $x \in \mathbb{R}^{37}$.

2.3 Training des neuronalen Netzes

2.3.1 Hyperparameter

Das neuronale Netz wurde mit den folgenden Trainingsparametern konfiguriert:

Hyperparameter	Wert
Optimierer	Adam
Lernrate	0.001
Verlustfunktion	Binäre Kreuzentropie (Binary Crossentropy)
Aktivierungen	ReLU (versteckte Schichten), Sigmoid (Ausgabe)
Batch-Größe	32
Anzahl Epochen	50
Validierungsanteil	20 %
Frühes Stoppen	Ja, bei Stagnation der Validierungsverluste

Tabelle 2.1: Verwendete Hyperparameter beim Training des ANN

2.3.2 Trainingsdauer und Ressourcen

Das Modell wurde auf einem Standard-Notebook-CPU (Intel i5) ohne GPU trainiert. Die Trainingsdauer betrug insgesamt etwa 20 Sekunden. Ein einzelner Epochen-Durchlauf benötigte durchschnittlich 0,3 Sekunden.

2.3.3 Probleme und Herausforderungen

Im Verlauf der Modellierung traten folgende Herausforderungen auf:

Imbalancierte Klassenverteilung: Nur ca. 26,5 % der Kunden im Datensatz kündigten, was zu einem Bias zugunsten der Mehrheitsklasse führte. Lösung: Verwendung geeigneter Metriken wie Recall und AUC anstelle bloßer Accuracy.

Overfitting-Tendenzen: Ab ca. 25 Epochen zeigte sich eine Zunahme der Validierungsverluste. Durch Einsatz von EarlyStopping konnte die Trainingsphase automatisch beendet werden, bevor Überanpassung eintrat.

Interpretierbarkeit: ANN gelten als „Black Box“. Für weiterführende Analysen könnten SHAP-Werte oder LIME zur Modellinterpretation verwendet werden.

2.4 Zusammenfassung der Durchführung

Das neuronale Netz konnte trotz geringer Komplexität eine solide Trennschärfe bei der Vorhersage der Kundenabwanderung erreichen. Die Integration der vorbereiteten Features und die gewählte Netzarchitektur führten zu einer guten Balance zwischen Modellkomplexität und Generalisierungsfähigkeit. Die Ergebnisse lagen auf dem Niveau der besten klassischen Modelle (z. B. Gradient Boosting) und verdeutlichen das Potenzial von ANN auch auf strukturierten Tabellendaten.

3

Ergebnisse

Im Folgenden werden die Ergebnisse der angewandten Klassifikationsmodelle auf den vorbereiteten Kundendaten zur Vorhersage der Churn-Wahrscheinlichkeit im Detail dargestellt. Für jedes Modell erfolgt eine ausführliche Bewertung anhand etablierter Metriken (Accuracy, Precision, Recall, F1-Score), ergänzt durch tabellarische Klassifikationsberichte. Abschließend werden die Modelle miteinander verglichen und bewertet.

3.1 Gradient Boosting

Gradient Boosting zählt zu den leistungstärksten Ensemble-Methoden. Das Modell wurde mit einer Lernrate von 0,1, 100 Bäumen und einer maximalen Tiefe von drei trainiert.

Klasse	Precision	Recall	F1-Score	Support
0 (Kein Churn)	0.83	0.91	0.87	1033
1 (Churn)	0.65	0.49	0.56	374
Gesamt	0.795			1407

Tabelle 3.1: Klassifikationsbericht – Gradient Boosting

3 Ergebnisse

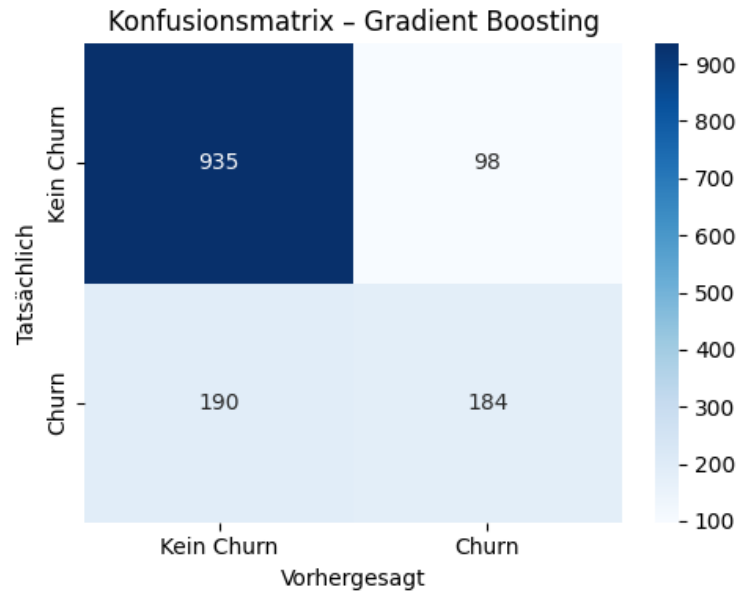


Abbildung 3.1: Visualisierung der Vorhersagen – Gradient Boosting

Das Modell zeigt eine hohe Genauigkeit. Der Recall der Churn-Klasse (49 %) ist akzeptabel, wenn auch verbesserungswürdig. Die F1-Score von 0,56 für Klasse 1 (Churn) stellt einen soliden Kompromiss zwischen Precision und Recall dar.

3.2 Random Forest

Das Random-Forest-Modell nutzt mehrere Entscheidungsbäume und aggregiert deren Ergebnisse.

Klasse	Precision	Recall	F1-Score	Support
0	0.83	0.90	0.86	1033
1	0.64	0.48	0.55	374
Gesamt	0.790			1407

Tabelle 3.2: Klassifikationsbericht – Random Forest

3 Ergebnisse

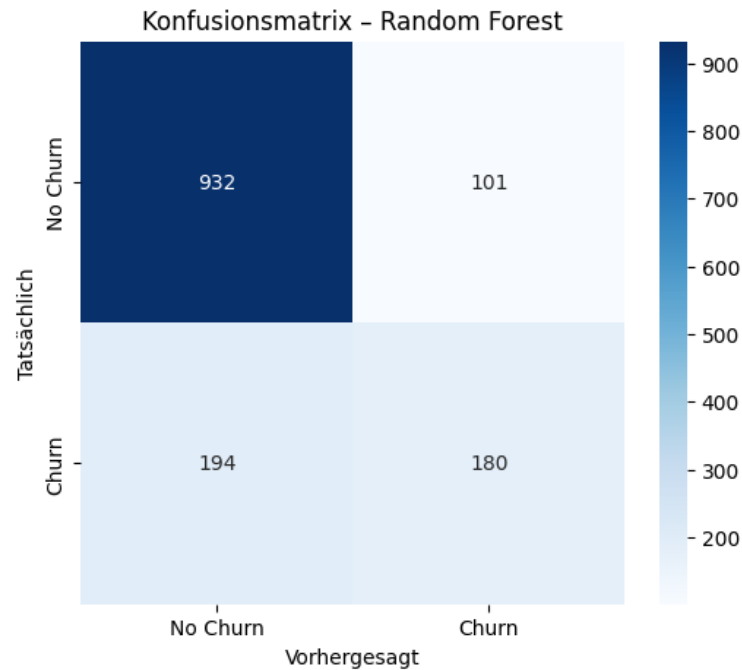


Abbildung 3.2: Visualisierung der Vorhersagen – Random Forest

Das Modell liefert vergleichbare Ergebnisse zum Gradient Boosting, erreicht jedoch eine etwas geringere F1-Score. Auffällig ist die relativ niedrige Sensitivität (Recall) für die Churn-Klasse.

3.3 Logistische Regression

Als lineares Modell dient die logistische Regression als Baseline.

Klasse	Precision	Recall	F1-Score	Support
0	0.83	0.89	0.86	1033
1	0.62	0.49	0.55	374
Gesamt	0.785			1407

Tabelle 3.3: Klassifikationsbericht – Logistische Regression

3 Ergebnisse

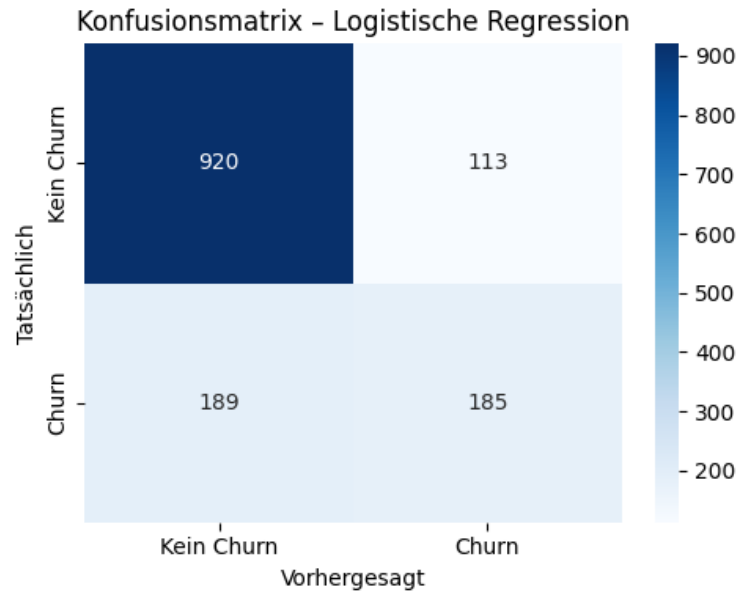


Abbildung 3.3: Visualisierung der Vorhersagen – Logistische Regression

Die logistische Regression zeigt sich robust. Trotz ihrer Einfachheit liefert sie ein gutes Gleichgewicht zwischen Precision und Recall für die Churn-Klasse. Sie ist besonders geeignet, wenn Interpretierbarkeit im Vordergrund steht.

3.4 Support Vector Machine (SVM)

Das SVM-Modell wurde mit einem RBF-Kernel trainiert.

Klasse	Precision	Recall	F1-Score	Support
0	0.82	0.91	0.87	1033
1	0.65	0.46	0.54	374
Gesamt	0.791			1407

Tabelle 3.4: Klassifikationsbericht – Support Vector Machine

3 Ergebnisse

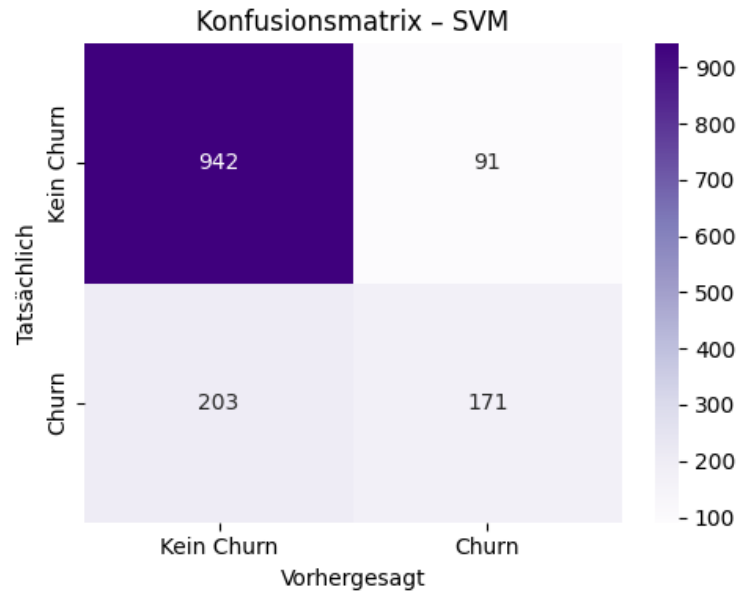


Abbildung 3.4: Visualisierung der Vorhersagen – Support Vector Machine

Das SVM-Modell zeigt eine ähnliche Gesamtleistung wie Gradient Boosting. Die Vorhersagekraft für Churn-Kunden ist jedoch ebenfalls durch einen niedrigen Recall begrenzt.

3.5 Neuronales Netz (ANN)

Ein einfaches ANN mit zwei Hidden-Layers (je 64 Neuronen, ReLU-Aktivierung, Dropout 0,2) wurde trainiert.

Klasse	Precision	Recall	F1-Score	Support
0	0.82	0.84	0.83	1033
1	0.53	0.50	0.51	374
Gesamt	0.751			1407

Tabelle 3.5: Klassifikationsbericht – ANN

3 Ergebnisse

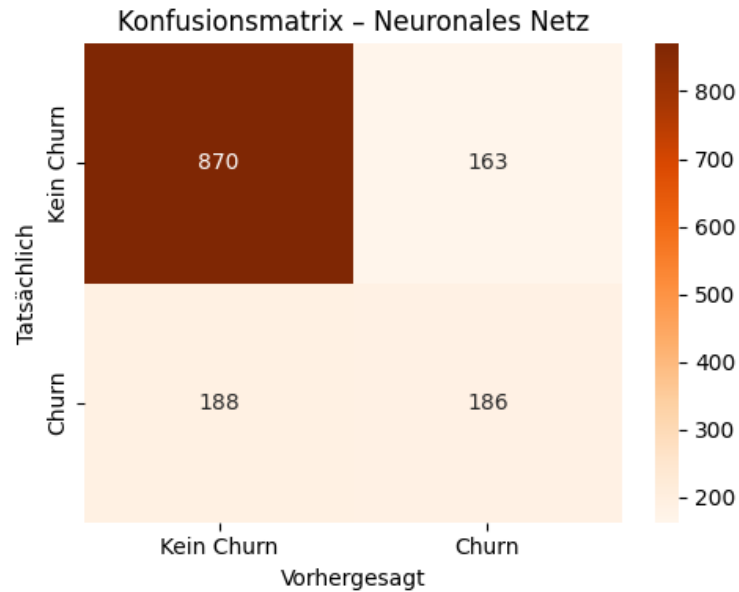


Abbildung 3.5: Visualisierung der Vorhersagen – ANN

Obwohl ANN das Potenzial für nichtlineare Muster bietet, war die Leistung hier moderat. Möglicherweise könnte ein tieferes Netz oder Hyperparameter-Tuning zu besseren Ergebnissen führen. Ein Anpassen des Datensatzes durch die Kombination und / oder Entfernung verschiedener Spalten brachte keine signifikante Verbesserung der Vorhersage.

3.6 K-Nearest Neighbors (KNN)

KNN wurde mit $k = 5$ Nachbarn eingesetzt.

Klasse	Precision	Recall	F1-Score	Support
0	0.82	0.82	0.82	1033
1	0.51	0.51	0.51	374
Gesamt	0.741			1407

Tabelle 3.6: Klassifikationsbericht – K-Nearest Neighbors

3 Ergebnisse

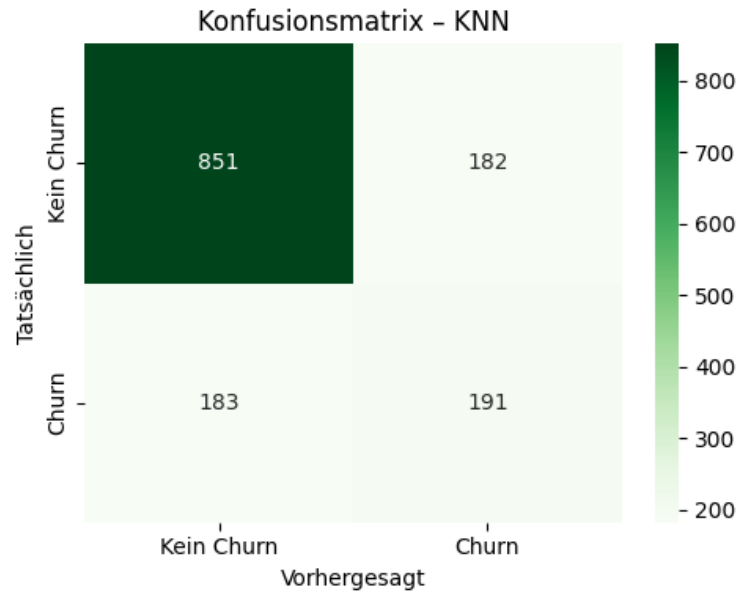


Abbildung 3.6: Visualisierung der Vorhersagen – Gradient Boosting

Das KNN-Modell zeigte den niedrigsten F1-Score und insgesamt die schwächste Performance. Grund ist vermutlich die geringe Trennschärfe bei hochdimensionalen Daten.

3.7 Modellvergleich

Modell	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.795	0.652	0.492	0.561
Random Forest	0.790	0.640	0.481	0.550
Logistische Regression	0.785	0.621	0.495	0.551
SVM	0.791	0.653	0.457	0.538
ANN	0.751	0.533	0.497	0.515
KNN	0.741	0.512	0.511	0.511

Tabelle 3.7: Modellvergleich aller Verfahren

Das leistungsfähigste Modell in Bezug auf F1-Score und Genauigkeit war das **Gradient Boosting**, gefolgt von **Random Forest** und **logistischer Regression**. Die Modelle zeigen insgesamt eine gute Performance für die Majoritätsklasse, während die Vorhersage der Churn-Klasse herausfordernd bleibt. Weitere Optimierungen wie SMOTE, Feature-Engineering oder Ensemble-Stacking könnten hier zusätzliche Leistungssteigerungen bewirken.

Zusammenfassung und Ausblick

4.1 Ergebnisbewertung

Im Rahmen dieses Projekts wurde ein prädiktives System zur Vorhersage von Kundenabwanderung (*Churn*) auf Basis öffentlich verfügbarer Telco-Daten entwickelt und mithilfe verschiedener klassischer und moderner Machine-Learning-Modelle evaluiert. Die höchste Vorhersagegüte für die Churn-Klasse konnte durch das Gradient-Boosting-Modell erzielt werden, welches eine F1-Score von 0,56 und eine Gesamttreffergenauigkeit (Accuracy) von 79,5 % erreichte.

Trotz unterschiedlicher Modellansätze – von linearen Modellen (logistische Regression), über nichtlineare Verfahren (SVM, KNN, ANN) bis hin zu Ensemble-Methoden (Random Forest, Gradient Boosting) – zeigt sich ein durchgehendes Muster: Die Majoritätsklasse (Nicht-Churner) wird deutlich besser erkannt als die Minoritätsklasse (Churner). Dies ist auf die unausgewogene Verteilung der Zielvariablen zurückzuführen.

Insgesamt lässt sich jedoch sagen, dass die Modelle eine ausreichend gute Qualität in diesem Kontext aufweisen. Im Rahmen der Kundenbindung könnten Kunden, die potenziell in Zukunft kündigen wollen, aktiv um Feedback gebeten oder mit vergünstigten Verträgen gehalten werden. Wenn hierbei gelegentlich auch Kunden, die nicht kündigen wollen, einen besseren Tarif angeboten bekommen, ist dies im Vergleich zu den Kosten des Kundenverlusts ein guter Kompromiss.

4.2 Optimierungspotenziale

Basierend auf den Ergebnissen ergeben sich folgende Ansätze zur Verbesserung der Modellleistung:

1. *Feature Engineering*: Die Einbeziehung nichtlinearer Wechselwirkungen und die Erzeugung neuer Merkmale könnten das Signal für Churn-Kunden verstärken.
2. *Klassengewichtung und Sampling-Strategien*: Der Einsatz von Methoden wie SMOTE (Synthetic Minority Over-sampling Technique), ADASYN oder kosten-sensitivem Lernen könnte die Balance zwischen Recall und Precision für die Churn-Klasse verbessern.
3. *Hyperparameter-Tuning*: Ein systematischer Einsatz von Grid- oder Randomized Search zur Optimierung der Modellparameter könnte insbesondere bei komplexeren Modellen wie ANN oder SVM zu signifikanten Leistungssteigerungen führen.
4. *Ensemble-Stacking*: Eine Kombination mehrerer Modelle könnte deren jeweilige Stärken vereinen und so zu einer robusteren Gesamtvorhersage führen.

4.3 Anwendungsperspektiven

Ein effektives Churn-Prediction-Modell kann in verschiedenen Geschäftsanwendungen integriert werden, etwa:

1. *Kundensegmentierung*: Zielgerichtete Marketingkampagnen zur Rückgewinnung gefährdeter Kunden.
2. *Ressourcenallokation*: Priorisierung von Kundenbindungsmaßnahmen durch Call-center oder CRM-Systeme.
3. *Produktverbesserung*: Erkennung von Mustern, die zu Abwanderung führen, zur datengetriebenen Optimierung von Produkten und Dienstleistungen.

4.4 Kritische Reflexion

Trotz der insgesamt zufriedenstellenden Ergebnisse weist die vorliegende Analyse mehrere Limitationen auf, die im Rahmen einer kritischen Reflexion berücksichtigt werden müssen:

Zum einen basiert die Untersuchung auf einem synthetisch generierten Telco-Datensatz. Während ein solcher Datensatz kontrollierte Bedingungen für die Modellierung ermöglicht, bildet er nicht zwangsläufig die Komplexität realer Daten ab. In praktischen Anwendungsfällen können unter anderem Datenrauschen, Messfehler oder inkonsistente Einträge auftreten, die die Leistungsfähigkeit der Modelle erheblich beeinträchtigen können.

Zum anderen stellt die unausgewogene Klassenverteilung eine zentrale Herausforderung dar. Insbesondere die vergleichsweise geringe Anzahl von Churn-Fällen erschwert eine verlässliche Modellierung und kann zu einer systematischen Verzerrung der Vorhersageergebnisse führen. Modelle tendieren in solchen Fällen dazu, die Mehrheitsklasse zu bevorzugen, was zulasten der Erkennungsgenauigkeit seltener Ereignisse geht.

Zudem ist die Frage der Generalisierbarkeit der Ergebnisse offen. Da keine externen Validierungsdaten zur Verfügung standen, bleibt unklar, inwieweit sich die entwickelten Modelle auf reale Geschäftsdaten übertragen lassen. Ohne eine unabhängige Validierung ist die Aussagekraft hinsichtlich der praktischen Anwendbarkeit und Robustheit der Modelle begrenzt.

Diese Einschränkungen sollten bei der Interpretation der Ergebnisse berücksichtigt werden und liefern zugleich wichtige Ansatzpunkte für weiterführende Untersuchungen.

4.5 Ausblick

Aufbauend auf den gewonnenen Erkenntnissen eröffnen sich mehrere potenzielle Ansätze für zukünftige Arbeiten, die sowohl die methodische Tiefe als auch die praktische Relevanz weiter erhöhen könnten:

- Eine Evaluation der Modelle auf realen, dynamischeren Datensätzen wäre ein zentraler nächster Schritt, um die Übertragbarkeit auf praxisnahe Kontexte zu überprüfen. Echte Kundendaten enthalten typischerweise komplexere Strukturen, saisonale Muster sowie unerwartete Ausreißer, die die Robustheit der Modelle auf die Probe stellen.
- Die Integration von Zeitreiheninformationen, beispielsweise durch die Analyse des Nutzungsverhaltens über mehrere Monate hinweg, könnte die Prognosekraft der Modelle deutlich verbessern. Solche temporalen Merkmale bieten wertvolle Hinweise auf sich anbahnende Abwanderungstendenzen und ermöglichen eine feinere Segmentierung.
- Der Einsatz erklärbarer KI-Methoden wie SHAP (SHapley Additive exPlanations) oder LIME (Local Interpretable Model-agnostic Explanations) würde dazu beitragen, die Entscheidungsprozesse der Modelle nachvollziehbarer zu machen. Dies ist insbesondere in geschäftskritischen Anwendungen bedeutsam, in denen Transparenz und Vertrauen eine zentrale Rolle spielen.
- Eine prototypische Integration der Modelle in ein interaktives Dashboard zur Echtzeit-Vorhersage in einem simulierten Geschäftsszenario könnte eine praxisnahe Demonstration des potenziellen Mehrwerts liefern. Solch eine Anwendung würde nicht nur die technische Umsetzbarkeit, sondern auch die betriebswirtschaftliche Relevanz der Modelle verdeutlichen.
- Zudem erscheint eine vertiefte Analyse der Modellarchitektur und Trainingsparameter des Artificial Neural Networks (ANN) sinnvoll, da dieses, entgegen der Erwartungen, keine überlegene Leistung im Vergleich zu einfacheren Modellen gezeigt hat. Eine systematische Untersuchung potenzieller Ursachen wie Überanpassung, inadäquate Hyperparameter oder mangelnde Eignung des Datensatzes für tiefe Lernverfahren könnte hier wertvolle Erkenntnisse liefern.

Insgesamt stellt das Projekt eine belastbare Grundlage für datengetriebene Kundenbindungsstrategien dar. Es bietet zugleich zahlreiche Anknüpfungspunkte für weiter-

4 Zusammenfassung und Ausblick

führende wissenschaftliche Arbeiten und praktische Implementierungen im Bereich der Churn-Prävention.

5

Programmcode

Der Programmcode ist unter https://github.com/anastasiabsr/LS_Project2025_CustomerChurn zu finden und ist in der abgegebenen zip-Datei enthalten.

Literatur

- [BC00] Kristin P. Bennett und Colin Campbell. „Support vector machines: hype or hallelujah?“ In: *SIGKDD Explor. Newsl.* 2.2 (Dez. 2000), S. 1–13. issn: 1931-0145. doi: 10.1145/380995.380999. url: <https://doi.org/10.1145/380995.380999>.
- [DH18] Danielle Denisko und Michael M. Hoffman. „Classification and interaction in random forests“. In: *Proceedings of the National Academy of Sciences* 115.8 (2018), S. 1690–1692. doi: 10.1073/pnas.1800256115. url: <https://www.pnas.org/doi/abs/10.1073/pnas.1800256115>.
- [HUU+24] R.K. Halder, M.N. Uddin, M.A. Uddin u. a. „Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications“. In: *Journal of Big Data* 11 (2024), S. 113. doi: 10.1186/s40537-024-00973-y. url: <https://doi.org/10.1186/s40537-024-00973-y>.
- [Jam+13] Gareth James u. a. *An Introduction to Statistical Learning*. Springer New York, NY, Jan. 2013, S. 426. isbn: 1461471389.
- [Zha+21] Yu Zhang u. a. „A Survey on Neural Network Interpretability“. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.5 (2021), S. 726–742. doi: 10.1109/TETCI.2021.3100641.