Digital Business University of Applied Sciences
Data Science & Business Analytics
DMI01 Data Mining
Prof. Daniel Ambach

Datengetriebene Prognosen im Hochschulbereich: Ein Data-Mining-Ansatz zur Vorhersage von Abschlussquoten an universitären Hochschulen und Fachhochschulen in der Schweiz

Studienarbeit

Eingereicht von Sabine Wildemann
Matrikelnummer 190297
Datum 03.03.2025

Inhaltsverzeichnis

1. Betriebswirtschaftliche Problemstellung und Analyseproblem

Im Rahmen dieses Data-Mining-Projekts soll für ein Schweizer EdTech-Startup ein prädiktives Modell entwickelt werden, das die Abschlussquoten von Studierenden an universitären Hochschulen und Fachhochschulen in der Schweiz prognostiziert. Das Startup bietet eine Lernplattform, die sich auf die innere Entwicklung von Studierenden konzentriert und Bildungseinrichtungen dabei unterstützt, diese Kompetenzen zu fördern. Die sogenannten Inner Development Goals (IDGs) bilden dabei einen Rahmen für persönliche und kollektive Entwicklung, der Fähigkeiten wie Selbstreflexion, kritisches Denken, emotionale Intelligenz und Zusammenarbeit umfasst. Diese Fähigkeiten sind entscheidend, um nachhaltige Entwicklungsziele, wie die UN-Nachhaltigkeitsziele (SDGs), zu erreichen.

Die betriebswirtschaftliche Problemstellung für das Startup besteht darin, fundierte Entscheidungen bezüglich der Produktentwicklung und des Vertriebs zu treffen. Durch die Vorhersage der Abschlussquoten können gezielte Strategien entwickelt werden, um die Lernplattform effektiver zu gestalten und den Bedürfnissen der Studierenden besser gerecht zu werden. Dies ermöglicht es dem Startup, seine Ressourcen optimal zu allozieren und maßgeschneiderte Lösungen anzubieten, die die Erfolgsquote der Studierenden erhöhen.

Das Analyseproblem, das mit diesem Projekt adressiert wird, lautet: „Können historische Daten zu Studierenden und Abschlüssen an universitären Hochschulen und Fachhochschulen genutzt werden, um zukünftige Abschlussquoten vorherzusagen?“
Dabei sollen verschiedene Faktoren wie Studienfeld, Hochschultyp und Geschlecht berücksichtigt werden, um ein umfassendes und aussagekräftiges Modell zu entwickeln. Die Ergebnisse dieses Projekts sollen dem Startup Einblicke liefern, die für strategische Entscheidungen und die Weiterentwicklung der Lernplattform von entscheidender Bedeutung sind.

2. Data Mining und CRISP-DM

Data Mining ist ein zentraler Bestandteil der Datenanalyse und des maschinellen Lernens, der darauf abzielt, verborgene Muster, Zusammenhänge und nützliche

Informationen aus großen Datenmengen zu extrahieren. Dabei werden algorithmische Methoden eingesetzt, um aus strukturierten und unstrukturierten Daten aussagekräftige Erkenntnisse abzuleiten, die für strategische Entscheidungsprozesse in Wissenschaft und Wirtschaft genutzt werden können (Fayyad et al., 1996).

Der Data-Mining-Prozess umfasst mehrere aufeinanderfolgende Phasen:

Geschäftsproblem verstehen & Anforderungen definieren: Ausgangspunkt ist eine betriebswirtschaftliche oder wissenschaftliche Fragestellung, die in eine analytische Problemstellung überführt wird.
Datenakquise und -vorverarbeitung: Die Daten werden aus unterschiedlichen Quellen gesammelt, bereinigt und transformiert, um für die Analyse vorbereitet zu werden.
Datenbankintegration: Die aufbereiteten Daten werden in geeigneten Datenbanksystemen gespeichert, um effiziente Abfragen und Analysen zu ermöglichen.
Statistische Analyse und Modellbildung: Durch den Einsatz statistischer Verfahren, Machine-Learning-Algorithmen und heuristischer Methoden werden Muster erkannt und Modelle trainiert.
Evaluation und Interpretation: Die generierten Modelle werden hinsichtlich ihrer Genauigkeit und Aussagekraft bewertet, um zuverlässige Erkenntnisse für Entscheidungsprozesse abzuleiten.

Data Mining ist eng mit dem Konzept der Wissensentdeckung in Datenbanken (Knowledge Discovery in Databases, KDD) verbunden und findet Anwendung in Bereichen wie Finanzanalyse, Medizin, Marketing und Industrie 4.0 (Han, Kamber & Pei, 2011).

Der CRISP-DM-Ansatz (Cross Industry Standard Process for Data Mining)) wurde in den späten 1990er Jahren von einer Gruppe führender Unternehmen im Bereich Data Mining entwickelt. Zu den Hauptentwicklern gehörten SPSS (heute Teil von IBM), Daimler AG, NCR Corporation und OHRA. Das Framework wurde im Rahmen eines EU-finanzierten Projekts entwickelt, um einen einheitlichen, branchenübergreifenden Standard für Data Mining-Prozesse bereitzustellen. Es ist

heute einer der am weitesten verbreiteten Ansätze für analytische Projekte und maschinelles Lernen.

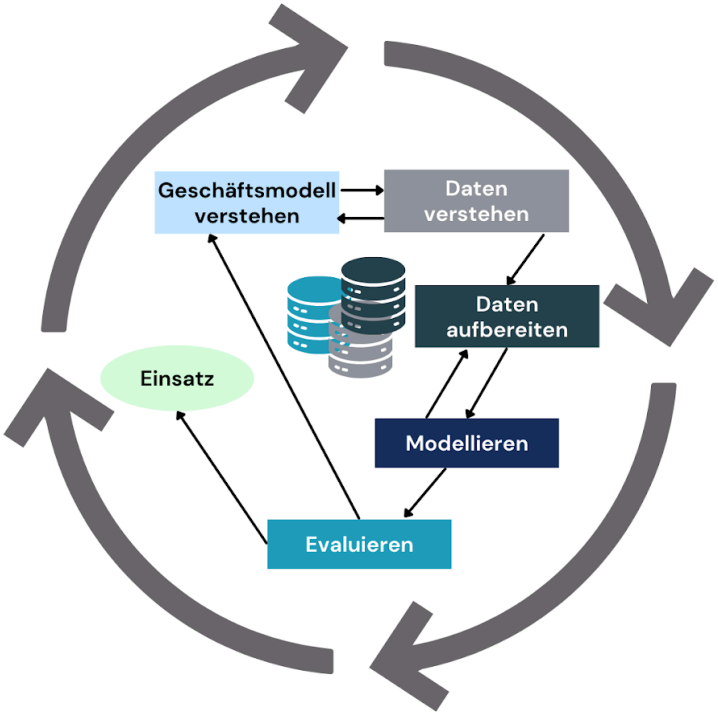

Abbildung 1: CRISP-DM Prozess (eigene Darstellung)

Eine aktuelle Studie von Shimaoka et al. (2024) hat 16 relevante Studien identifiziert und bewertet, um zu untersuchen, wie CRISP-DM mit verschiedenen Methoden, Frameworks und Prozessmodellen angepasst werden kann. Diese Analyse zeigt nicht nur die Entwicklung von CRISP-DM von starren zu agilen Ansätzen auf, sondern identifiziert auch spezifische Anpassungsmuster, die die Flexibilität und Anpassungsfähigkeit des Modells für unterschiedliche Geschäftsbereiche unterstreichen.

3. Datenauswahl, Datenverständnis und Datenvorbereitung

Die Datenauswahl bildet die Grundlage für das vorliegende Data-Mining-Projekt und ist entscheidend für die Qualität und Aussagekraft der nachfolgenden Analysen. Es werden Daten zu Abschlüssen von Studierenden an Schweizer Hochschulen verwendet, die auf der Plattform opendata.swiss zur Verfügung stehen. Die Datensätze umfassen sowohl Abschlüsse an Fachhochschulen und pädagogischen Hochschulen als auch an universitären Hochschulen.

a) Datenquellen

Die Studienarbeit basiert auf Daten, die auf opendata.swiss, dem zentralen Portal für offene Daten der Schweizer Behörden, zur Nutzung bereitgestellt werden. Die 6 Dateien enthalten Daten zu Fachhochschul- und Universitätsabschlüssen von 34 Institutionen, die sich über den Zeitraum von 1980 bis 2023 erstrecken. Detaillierte Angaben zu den Datensätzen sind im Anhang erläutert.

b) Datenformate und Strukturierung

Die Daten liegen in strukturierten .csv-Formaten vor, die eine einfache Integration und Verarbeitung ermöglichen. Die ursprünglichen Appendix-Dateien im .ods-Format werden in ein .xls-Format konvertiert, um eine einheitliche Datenstruktur zu gewährleisten.
Um die Daten aus den verschiedenen Quellen in eine relationale Datenbank zu überführen, wird eine normalisierte SQLite-Datenbankstruktur erstellt.

Da die beiden Appendix-Dateien (abschluesse-FH-APPENDIX.xls und abschluesse-HS-APPENDIX.xls) teilweise gleiche Entitäten mit denselben Schlüsseln (z. B. SEX, LEVEL, FIELD) und teilweise unterschiedliche Schlüsselsysteme (z. B. UNI) haben, erfordert das eine differenzierte Strategie für den Datenbankaufbau.

Da die Kodierungen für Universitäten (UNI), Studienfelder (FIELD) und Abschlusslevel (LEVEL) zwischen Fachhochschulen (FH) und universitären Hochschulen (HS) unterschiedlich sind, werden separate Tabellen für University_FH, University_HS, Field_FH und Field_HS angelegt. Für die Abschlusslevel wird eine einheitliche Level-Tabelle erstellt, die durch die zusätzliche Zuordnungstabelle Level_Mapping die spezifischen FH- und HS-Codes referenziert, sodass redundante Daten vermieden und eine flexible Abfrage ermöglicht wird.
Die Graduations-Tabelle ist bis zur fünften Normalform (5NF) normalisiert, einschließlich Boyce-Codd Normalform (BCNF), da keine mehrwertigen Abhängigkeiten oder Join-Abhängigkeiten vorhanden sind. Dies bedeutet, dass die Tabelle frei von Redundanzen ist und eine effiziente Datenstruktur aufweist. Die Normalisierung wurde durch die Auslagerung von Attributen in separate

Tabellen (Period, UniversityType, University_Mapping, Field_Mapping, Level, Gender) und die Verwendung von Fremdschlüsseln erreicht. Es ist anzumerken, dass die berechneten Spalten die Normalform verletzen.

c) Datenvorbereitung

In der Datenvorbereitung wird zunächst sichergestellt, dass alle notwendigen Dateien vorhanden sind, indem CSV- und Excel-Daten extrahiert und auf fehlende Dateien überprüft werden. Anschließend werden die Rohdaten bereinigt und transformiert – dazu zählen das Aufsplitten von kommagetrennten Werten, das Entfernen unerwünschter Zeichen und die Umwandlung von Datentypen.
Zur besseren Übersicht und Analyse werden Profiling-Berichte für beide Hochschul-Typen erstellt und interpretiert.

d) Erfassungs- und Speichermethode

Zur effizienten Verarbeitung und Analyse der Daten werden diese in eine SQLite-Datenbank importiert, wobei Lookup-Tabellen und Mapping-Tabellen zur Sicherstellung der Datenintegrität angelegt und befüllt werden.
Die Datenbereinigung und -transformation erfolgt mittels Python und der Bibliothek Pandas, um die Daten für die nachfolgenden Analysen vorzubereiten.

e) Explorative Datenanalyse (EDA)

Im Rahmen der explorativen Datenanalyse (EDA) werden verschiedene Visualisierungstechniken eingesetzt, um erste Einblicke in die Datenstruktur zu gewinnen, Muster zu identifizieren und Hypothesen für die Modellierung zu generieren. Ein grundlegendes Werkzeug der EDA ist der Scatterplot, der, wie in Han, Kamber und Pei (2011) beschreiben, Beziehungen zwischen zwei numerischen Variablen aufzeigt. Ein Scatterplot stellt Wertepaare als Punkte in einem Koordinatensystem dar, wodurch Cluster, Ausreißer und Korrelationen sichtbar werden (siehe Abbildung 2).

Obwohl Scatterplots für die bivariate Analyse *zweier* Variablen optimiert sind, können andere Visualisierungen wie Balkendiagramme nützlich sein, um *eine* Variable (hier: Gesamtzahl der Abschlüsse) über verschiedene Kategorien (hier: Universitäten/Hochschulen) zu vergleichen. Abbildung 3 zeigt ein solches

Balkendiagramm, das die kumulierten Absolventenzahlen jeder Universität/Hochschule im Zeitraum 2000-2023 darstellt. Diese Visualisierung ermöglicht einen schnellen Vergleich der Hochschulen hinsichtlich ihres gesamten Abschlussvolumens und liefert Hinweise auf potenzielle Ausreißer oder Gruppenunterschiede, die in nachfolgenden Analysen und Modellierungen berücksichtigt werden sollten.

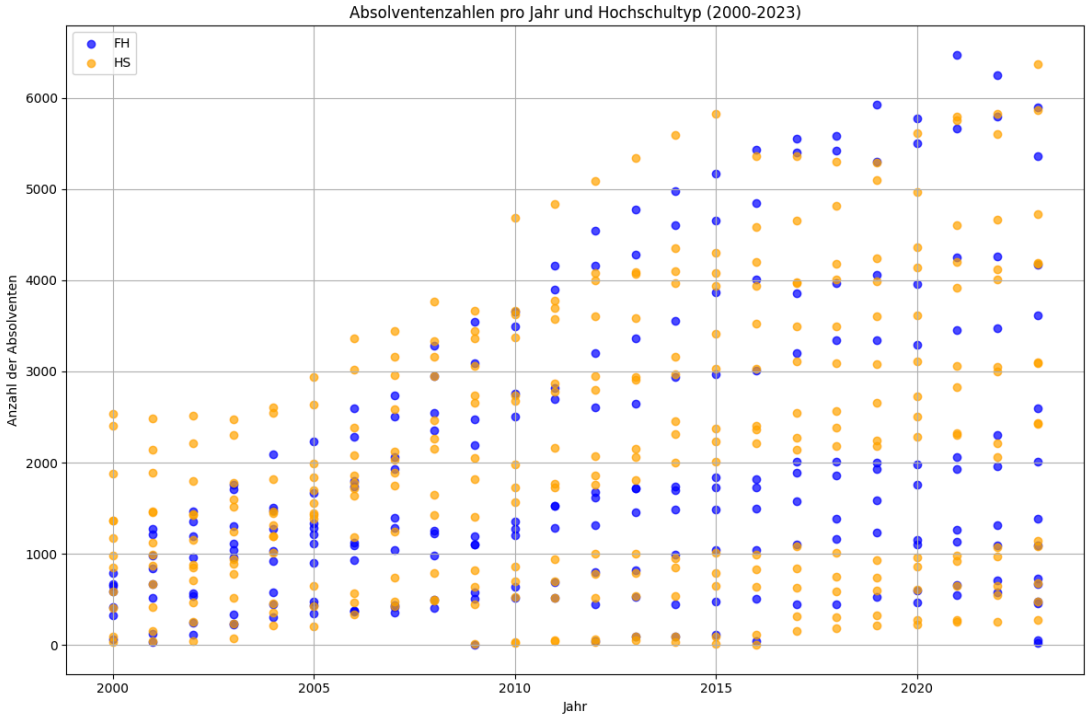

Abbildung 2: Anzahl der Abschlüsse pro HS/FH für den Zeitraum 2000-2023

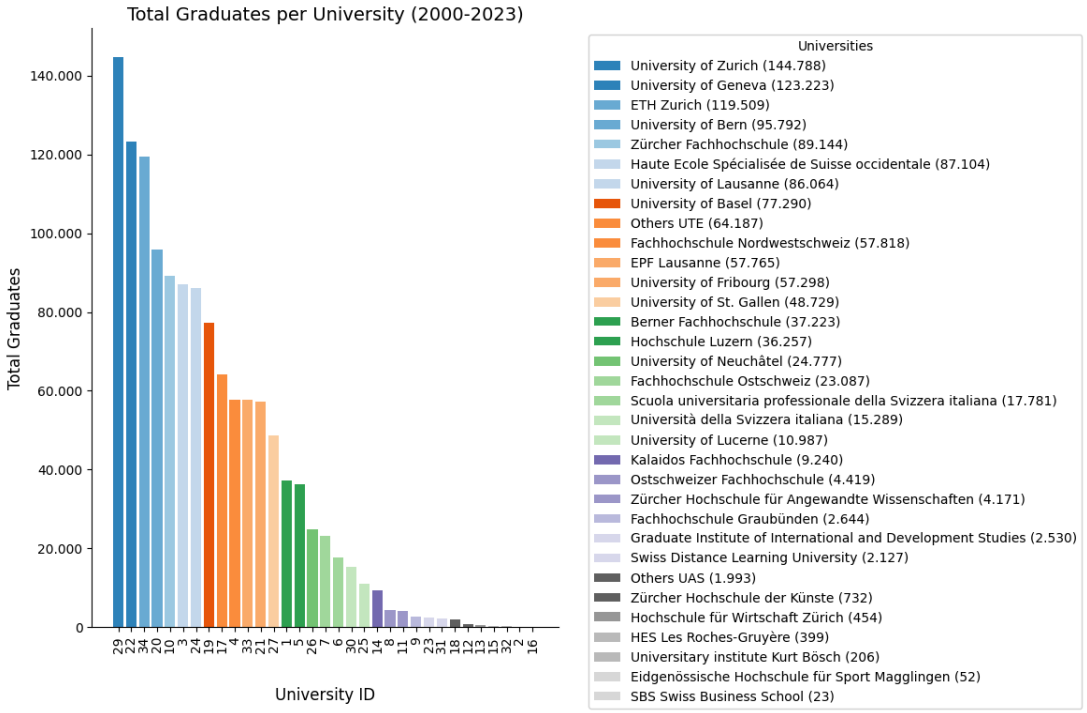

Abbildung 3: Anzahl der Abschlüsse pro HS/FH für den Zeitraum 2000-2023

4. Methodik und Einsatz von generativer KI

Durch den Einsatz verschiedener analytischer und maschineller Lernmethoden wird eine Vorhersage von Abschlussquoten an Hochschulen angestrebt. Die Tabelle 1 gibt einen Überblick über die zentralen methodischen Schritte und die verwendeten Technologien:

Tabelle 1: Zusammenfassung der angewandten Methoden

Schritt	**Methode**	**Technologie**
Datensammlung	Datenextraktion/Bereinigung	Python (pandas)
Datenintegration	Speicherung und Verwaltung	SQLite
EDA & Visualisierung	Deskriptive Statistik, Charts	Python (pandas, matplotlib)
Modellbildung	Regression, Zeitreihenanalyse, ML-Modelle	Lineare Reg, ARIMA, Neuronales Netzwerk, GBM, Decision Tree Regressor LSTM, Prophet
Feature Engineering	Interaktionsmerkmale, Aggregation	Python (pandas, scikit-learn)
Hyperparameter Tuning	GridSearchCV	Python (pandas, scikit-learn)
Evaluierung	Fehleranalyse, Kreuzvalidierung	RMSE, R^2

Diese Studienarbeit profitiert maßgeblich von den Fortschritten im Bereich der generativen Künstlichen Intelligenz (KI), insbesondere von den Fähigkeiten großer Sprachmodelle (Large Language Models, LLMs) als "Zero-Shot Reasoners," wie von Kojima et al. (2022) in "Large Language Models are Zero-Shot Reasoners" gezeigt. LLMs sind in der Lage, komplexe Aufgaben ohne explizite Beispiele durch "Chain-of-Thought-Prompting" zu lösen. Diese Fähigkeit wird in vier Kernbereichen der datengetriebenen Analyse genutzt, um den Forschungsprozess zu optimieren und die Ergebnisse zu verbessern:

Verbessertes Feature Engineering: Anstatt manueller Merkmalskonstruktion wird generative KI eingesetzt, um automatisiert relevante Muster und Beziehungen in den Daten zu identifizieren und daraus potenziell informative Merkmale abzuleiten. Dieser Ansatz ermöglicht die Entdeckung nicht-trivialer Zusammenhänge, die sonst möglicherweise übersehen worden wären.

Optimierte Hyperparameter-Suche: Die zeitaufwändige manuelle Optimierung von Modellhyperparametern wird durch KI-gestützte Verfahren ersetzt. Generative KI-Modelle können den Suchraum effizienter explorieren und Konfigurationen identifizieren, die zu einer höheren Prognosegenauigkeit führen.

Vertiefte Ergebnisinterpretation: Über die reine Modellprognose hinaus unterstützt generative KI die Interpretation der Ergebnisse. Durch die Generierung von prägnanten Zusammenfassungen, Erklärungen von Modellentscheidungen und Visualisierungshilfen wird ein tieferes Verständnis der zugrunde liegenden Datenmuster und -treiber ermöglicht.

Beschleunigte Code-Entwicklung: Die Implementierung der Analysepipelines und Modelle wird durch die Fähigkeit generativer KI zur Code-Generierung erheblich beschleunigt. Verschiedene Modelle (OpenAI, ChatGPT o3-mini, Gemini 2.0 Pro Experimental, Mistral AI, Claude 3.7 Sonnet) wurden genutzt, um Teile des Codes, oder ganze Skripte, effizient und konsistent zu erstellen, wodurch die Entwicklungszeit reduziert und die Reproduzierbarkeit erhöht wurde.

Der Einsatz generativer KI in diesen Bereichen führt zu einer Effizienzsteigerung im gesamten Modellierungsprozess, einer verbesserten Modellperformance und einer tiefergehenden, verständlicheren Ergebnispräsentation. Letztendlich unterstützt dies fundierte, datengetriebene Entscheidungen.

5. Modellauswahl und -beschreibung

5.1. Eingesetzte Machine Learning Modelle

Um die Abschlussquoten an Schweizer Hochschulen und Fachhochschulen zu prognostizieren, werden verschiedene Modellierungsansätze evaluiert. Tabelle 2 stellt die ausgewählten Modelle und die jeweilige Begründung für ihre Auswahl zusammenfassend dar.

Tabelle 2: Auswahl Machine Learning-Modelle

Modell	Begründung der Auswahl
Lineare Regression	Basismodell zur Vorhersage von Abschlussquoten basierend auf linearen Beziehungen zwischen den Merkmalen.
ARIMA (Autoregressive Integrated Moving Average)	Modellierung von Zeitreihendaten, um Trends und saisonale Muster in den Abschlussquoten zu erfassen.
Neuronal Network	Erfassen komplexer, nicht-linearer Beziehungen zwischen den Merkmalen und den Abschlussquoten.
Decision Tree Regressor	Vorhersage von Abschlussquoten durch Entscheidungsbäume, die auf bedingten Regeln basieren.
Gradient Boosting Regressor w. Hyperparameter Tuning	Kombination mehrerer schwacher Lerner (Bäume) zur Verbesserung der Vorhersagegenauigkeit.
LSTM – Long Short Term Memory	Rekurrentes neuronales Netz (RNN), speziell für Zeitreihen. Kann langfristige Abhängigkeiten lernen, geeignet für die Erfassung nichtlinearer Muster.
Prophet	Spezialisiertes Zeitreihenmodell (Trend + Saisonalität). Robust, einfach zu verwenden, gut für Daten mit klarem Trend. Weniger datenintensiv als LSTM.

5.2 Feature Engineering

Das Feature Engineering zielt darauf ab, die Rohdaten in ein Format zu transformieren, das für die verwendeten Modelle informativer ist. Durch die Erstellung neuer Merkmale, die spezifische Aspekte der Daten hervorheben (z.B. Interaktionen zwischen Variablen oder aggregierte Statistiken), können Modelle komplexe Beziehungen besser erfassen und potenziell genauere Vorhersagen treffen. In Tabelle 3 wird das in diesem Projekt durchgeführte Feature Engineering erläutert.

Tabelle 3: Feature Engineering: Interaktionsmerkmal und aggregierte Merkmale

Feature	Erläuterung
Interaktionsmerkmal: field_university_interaction	kombiniert zwei kategoriale Variablen (field_id und university_type_id), um spezifische Effekte bestimmter Fachrichtungen an bestimmten Universitäten zu modellieren.
Aggregiertes Merkmal: avg_graduates_per_field	stellt durchschnittliche Anzahl der Absolvent:innen für jede field_id/Fachrichtung dar
Aggregiertes Merkmal: avg_graduates_per_university	stellt durchschnittliche Anzahl der Absolvent:innen für jede university_type_id dar

Die Interaktionsmerkmale (field_university_interaction, field_level_interaction, field_gender_interaction) ermöglichen es einem Modell, Wechselwirkungen zwischen kategorialen Variablen, wie z.B. Fachrichtung und Hochschultyp oder Geschlecht und Studienfach, zu berücksichtigen. Zusätzlich werden aggregierte Merkmale in Form von durchschnittlichen Absolventenzahlen pro field_id (avg_graduates_per_field) und university_type_id (avg_graduates_per_university_type) hinzugefügt, wodurch das Modell gruppenspezifische Unterschiede erkennen kann.

6. Modellanwendung

a) Lineare Regression

Mean Squared Error: Root Mean Squared Error: R^2 Score:	8326.730965132621 91.25092309194807 0.3289495907953002

Das lineare Regressionsmodell wird verwendet, um die Anzahl der Absolvent:innen vorherzusagen. Als unabhängige Variablen dienen period_id, university_mapping_id, university_type_id, field_id, level_id und gender_id, wobei die kategorialen Variablen mittels One-Hot-Encoding transformiert werden. Das Modell erzielt einen Mean Squared Error (MSE) von 8326.73 und ein Bestimmtheitsmaß (R^2) von 0.32. Es erklärt somit 32% der Varianz in den Absolventenzahlen. Trotz der verbesserten Anpassung durch das

One-Hot-Encoding zeigt das mäßige R² Verbesserungspotential auf.
Das Modell dient als Baseline für weitere Modellentwicklungen. Verbesserungen sind möglich, etwa durch komplexere Modelle oder zusätzliche Merkmale. Das Modell bietet einen ersten Referenzpunkt zur Bewertung weiterführender Ansätze.

b) ARIMA (Autoregressive Integrated Moving Average)

Mean Squared Error: Root Mean Squared Error: R^2 Score:	8348641.337864024 2889.401553585798 -0.11390491670921143

Das ARIMA-Modell (5,1,0) dient hier zur Prognose der jährlichen Absolventenzahlen, wobei die letzten fünf Jahre als Testzeitraum abgetrennt werden. ACF- und PACF-Plots unterstützen die Parameterwahl. Mit einem RMSE von 2889,40 und einem negativen R² von -0,11 liegt das Modell deutlich hinter einer einfachen Mittelwert-Prognose zurück. Die hohen Abweichungen zwischen Vorhersage und Ist-Werten verdeutlichen, dass ARIMA in dieser Form das zugrunde liegende Muster nur unzureichend erfasst.

c) Neuronales Netzwerk

Mean Squared Error: Root Mean Squared Error: R^2 Score:	7341.39697265625 (100 Epochen) 7142.10302734375 (Feature Engineering) 1148.7884521484375 (One-Hot Encoding) 33.893781909790434 0.9081645011901855

Das neuronale Netzwerk wurde entwickelt, um komplexe Zusammenhänge zwischen den Variablen (z. B. period_id, university_mapping_id, university_type_id, field_id, level_id, gender_id) und den Absolventenzahlen zu erfassen. Vor dem Training werden die numerischen Daten standardisiert; das Modell besteht aus zwei Dense-Schichten (ReLU) plus linearer Ausgabeschicht und wird mit dem Adam-Optimizer trainiert.

Ohne Feature Engineering liegt der RMSE bei etwa 85,67 (100 Epochen). Mit zusätzlichen Interaktionsmerkmalen verbessert sich der RMSE nur leicht auf rund 84,55. Erst durch One-Hot-Encoding der kategorialen Variablen sinkt der RMSE deutlich auf 33,89, bei einem R² von 0,91.

Die Ergebnisse zeigen, wie entscheidend eine passende Repräsentation kategorialer Daten für die Modellgüte ist.

e) Decision Tree Regressor

Mean Squared Error: Root Mean Squared Error: R^2 Score:	1310.8891700404859 36.206203474549575 0.8952059496166167

Entscheidungsbäume basieren auf einer Baumstruktur, die aus Knoten und Kanten besteht. In jedem Knoten wird eine spezifische Frage gestellt, und die Antwort auf diese Frage bestimmt, welcher Pfad entlang der Kanten im Baum weiterverfolgt wird. Am Ende eines solchen Pfades steht dann der Output, der die Entscheidung des Baumes repräsentiert. Entscheidungsbäume bieten somit eine intuitive und transparente Möglichkeit, Entscheidungsprozesse abzubilden und Vorhersagen zu treffen.

Der Entscheidungsbaum nutzt sowohl die ursprünglichen Merkmale (period_id, university_mapping_id, usw.) als auch Interaktionsmerkmale (field_university_interaction, field_level_interaction, field_gender_interaction). Er wird per 80/20-Train/Test-Split (random_state=42) trainiert. Ein RMSE von rund 36.2 deutet auf eine noch ausbaufähige Prognosegenauigkeit hin, betrachtet man die Größenordnung der Absolventenzahlen. Die Feature Importance zeigt, dass university_mapping_id am stärksten wirkt, gefolgt von period_id und level_id. Interaktionsmerkmale wie field_gender_interaction und field_level_interaction sind ebenfalls relevant, während field_university_interaction und university_type_id weniger ins Gewicht fallen. field_id und gender_id spielen nur eine untergeordnete Rolle.

f) Gradient Boosting Regressor mit Hyperparameter Tuning

Mean Squared Error (MSE):	1395.28206345168 (beide Universitätstypen) FH/Fachhochschulen: 938.4469994711192 HS Universitäten: 540.9796616564906

Root Mean Squared Error:	30.63408231808355 (FH) 23.25896948827464 (HS)
R^2 Score:	0.9477638276161258 (FH) 0.9442092410085745 (HS)

Ein Gradient Boosting Regressor kombiniert sukzessive Entscheidungsbäume, um Prognosen zu verbessern. Mithilfe von GridSearchCV werden Hyperparameter (z. B. learning_rate=0.2, max_depth=5) optimiert. Das gemeinsame Modell für beide Hochschultypen erzielt einen RMSE von etwa 37,35. Separate Modelle für Fachhochschulen (FH) erreichen einen RMSE von 30,63 (R^2=0,95), während Hochschulen (HS) auf 23,26 (R^2=0,94) kommen.

Diese Werte scheinen zunächst auf eine gute Modellanpassung innerhalb der Kreuzvalidierung hinzuweisen. Es ist jedoch wichtig zu betonen, dass GBM, wie auch Entscheidungsbäume, nicht für Zeitreihenprognosen im engeren Sinne geeignet ist, da die zeitliche Reihenfolge der Daten nicht explizit in die Modellstruktur eingeht. Die ermittelten Metriken (MSE, R^2) beziehen sich auf die Fähigkeit des Modells, die gegebenen Daten zu erklären, nicht auf seine Fähigkeit, in die Zukunft zu extrapolieren.
Die zuvor beobachtete unrealistisch hohe Vorhersage für 2024 (39188.20) vor der Kreuzvalidierung bestätigt diese Einschränkung. Das Modell wird daher nicht für die eigentliche Prognose verwendet, sondern dient primär der Analyse der Feature Importance.

g) LSTM — Long short-term memory

Root Mean Squared Error:	Train RMSE: 995.74 Test RMSE: 586.94
R^2 Score:	Train R^2: 0.95 Test R^2: 0.67

Das LSTM (Long Short-Term Memory)-Netzwerk, entwickelt von Hochreiter und Schmidhuber (1997) wird wegen seiner Fähigkeit genutzt, zeitliche Abhängigkeiten zu erfassen. Es besteht aus einer LSTM-Schicht mit 10 Einheiten, gefolgt von einer Dropout-Schicht und einer Dense-Ausgabeschicht. Trainiert wird mit Adam (MSE-Loss) und Early Stopping.

Trotz dieser reduzierten Architektur (eine LSTM-Schicht, wenige Einheiten, Dropout) und einer Validierung (look_back=2) ist kein klassisches Overfitting erkennbar: Der Train-RMSE (995,74) liegt sogar über dem Test-RMSE (586,94). Auch verläuft die Validierungskurve nah an der Trainingskurve, ohne deutliche Abweichungen. Die geringe Datenmenge (12 Trainings- und 4 Validierungssequenzen) und die hohe Variabilität erschweren allerdings die Interpretation. Aufgrund der unzureichenden Modellleistung und spezieller Anforderungen bei Zeitreihendaten wird das LSTM zugunsten von Prophet nicht weiterverfolgt.

h) Prophet

Root Mean Squared Error:	Test-RMSE: 805.88 (FH) Test-RMSE: 1409.47 (HS)
R^2 Score:	Test R^2: 0.44 (FH) Test R^2: -0.11 (HS)
Nach Kreuzvalidierung (FH/HS): Root Mean Squared Error: R^2 Score:	FH / HS 457.90 / 1409.64 0.99 / 0.96

Das Prophet-Modell, konzipiert für Zeitreihenprognosen mit Trend- und Saisonalitätskomponenten, wird auf die jährlichen FH- und HS-Absolventenzahlen (2000-2023) angewendet. Nach Konvertierung der Jahreszahlen in datetime-Objekte und einem 85/15-Train/Test-Split ergibt sich bei den FH-Daten ein Test-RMSE von 805,88; die Prognose für 2024 liegt bei etwa 29.896 Absolventen. Anfänglich zeigte sich bei den HS-Daten ein negatives R^2, das durch eine präzisere Zeitrahmenfilterung und eine 5-fache Kreuzvalidierung auf einen mittleren RMSE von 1409,64 und ein mittleres R^2 von 0,96 verbessert werden.

Bei der Kreuzvalidierung wird der Datensatz mehrfach in Trainings- und Validierungsanteile unterteilt (Folds). Das Modell wird pro Falte neu trainiert und auf dem jeweiligen Validierungsanteil geprüft. So erhält man robustere Schätzungen der Modellgüte und vermeidet Overfitting, da das Modell nicht nur auf einer einzelnen Aufteilung basiert. Gerade bei knappen Daten liefert dieser Ansatz zuverlässige Kennzahlen (z. B. RMSE, R^2) und ermöglicht eine fundierte Beurteilung der Vorhersageleistung und Stabilität des Modells.

7. Ergebnisse

7.1 Evaluationsmetriken

Die Vorhersagegenauigkeit eines Modells ist entscheidend für die Vertrauenswürdigkeit seiner Prognosen. In dieser Studienarbeit werden sieben Modelle (Lineare Regression, ARIMA, Neuronales Netzwerk, Decision Tree, Gradient Boosting, LSTM und Prophet) anhand von zwei Kenngrößen beurteilt:

- **Root Mean Squared Error (RMSE)**: Quantifiziert den mittleren Vorhersagefehler in den Einheiten der Zielvariable (Absolventenzahl). Ein niedrigerer RMSE weist auf genauere Prognosen hin.
- **R^2 (Bestimmtheitsmaß)**: Zeigt an, wie gut das Modell die Streuung der Daten erklärt. Höhere Werte deuten auf eine bessere Modellanpassung hin.

Ein Modell mit sehr niedrigem RMSE, aber schwachem R^2 (oder umgekehrt) kann in der Praxis weniger nützlich sein als ein Modell, das beide Metriken solide erfüllt. Deshalb werden RMSE und R^2 stets gemeinsam betrachtet. Während manche Modelle in puncto RMSE vorn liegen, zeigt sich beim Prophet-Modell eine insgesamt überzeugende Kombination aus Fehlermaß und erklärter Varianz – insbesondere unter Berücksichtigung der zeitlichen Struktur der Daten.

7.2 Das Gewinner-Modell: Prophet

Für die Prognose der jährlichen Absolventenzahlen an Fachhochschulen (FH) und Universitäten (HS) erweist sich das Prophet-Modell als überlegene Wahl gegenüber anderen evaluierten Modellen (siehe Abschnitt 7.1). Prophet, entwickelt von Sean J. Taylor und Benjamin Letham und vorgestellt in ihrer Arbeit "Forecasting at Scale" (Taylor & Letham, 2017), ist ein Open-Source-Verfahren, das speziell für die Analyse von Geschäftszeitreihen mit Saisonalität und Trendänderungen konzipiert wurde. Es basiert auf einem additiven Modell, das nicht-lineare Trends mit jährlicher, wöchentlicher und täglicher Saisonalität sowie Feiertagseffekten berücksichtigt.

Es werden separate Modelle für FH und HS trainiert (Datenbasis: 2000-2023) und mittels 5-facher Kreuzvalidierung validiert, um die Generalisierbarkeit sicherzustellen.

Prognosen für 2024:

- FH: Das Modell prognostiziert ca. 28.965 Absolvent:innen.
- HS: Es werden ca. 40.462 Absolvent:innen erwartet.

Metrik	**FH**	**HS**	**Interpretation und Implikationen**
Mittlerer RMSE	457.90	1409.64	Der durchschnittliche Prognosefehler ist bei FHs deutlich geringer. Dies deutet auf eine höhere Vorhersagegenauigkeit für diesen Hochschultyp hin. Der höhere RMSE bei HS spiegelt größere Schwankungen wider.
Mittleres R^2	0.99	0.96	Beide Modelle erklären einen sehr hohen Anteil der Varianz in den Daten. Das FH-Modell zeigt eine nahezu perfekte Anpassung, während das HS-Modell nur geringfügig abfällt.

Die Kombination aus Kreuzvalidierung und der inhärenten Regularisierung von Prophet minimiert das Risiko von Overfitting. Die Modelle generalisieren gut auf neue, ungesehene Daten.

Die Visualisierungen in Abbildung 4 zeigen die prognostizierten Werte (Linien) und die zugehörigen Konfidenzintervalle (schattierte Bereiche), die ein Maß für die Prognoseunsicherheit darstellen.

FH: Ein klarer, nahezu linearer Aufwärtstrend wird fortgeschrieben. Die hohe Modellgüte (niedriger RMSE, R^2 nahe 1) unterstreicht die Verlässlichkeit dieser Prognose.
HS: Ein genereller Aufwärtstrend ist ebenfalls erkennbar, jedoch mit einer markanten Zunahme um 2012. Diese Nichtlinearität könnte auf strukturelle Veränderungen im Hochschulbereich oder veränderte Studienneigungen hindeuten. Obwohl die Prognosegenauigkeit etwas geringer ist als beim FH-Modell, ist sie immer noch hoch ($R^2 = 0.96$).

Das Prophet-Modell bietet eine überzeugende Kombination aus hoher Genauigkeit und Interpretierbarkeit für die Prognose von Absolventenzahlen. Es bestätigt die

zentrale Hypothese: Historische Zeitreihendaten ermöglichen *präzise* Vorhersagen zukünftiger Abschlusszahlen, insbesondere für Fachhochschulen. Der etwas höhere RMSE und das leicht niedrigere R^2 des HS-Modells deuten auf eine größere inhärente Variabilität in den Universitätsdaten hin, was weitere qualitative Analysen anstoßen könnte, um die Ursachen dieser Schwankungen zu verstehen (z.B. Reformen, veränderte Studienfachwahl). Insgesamt liefert das Prophet Modell eine belastbare Grundlage für strategische Entscheidungen im Hochschulbereich.

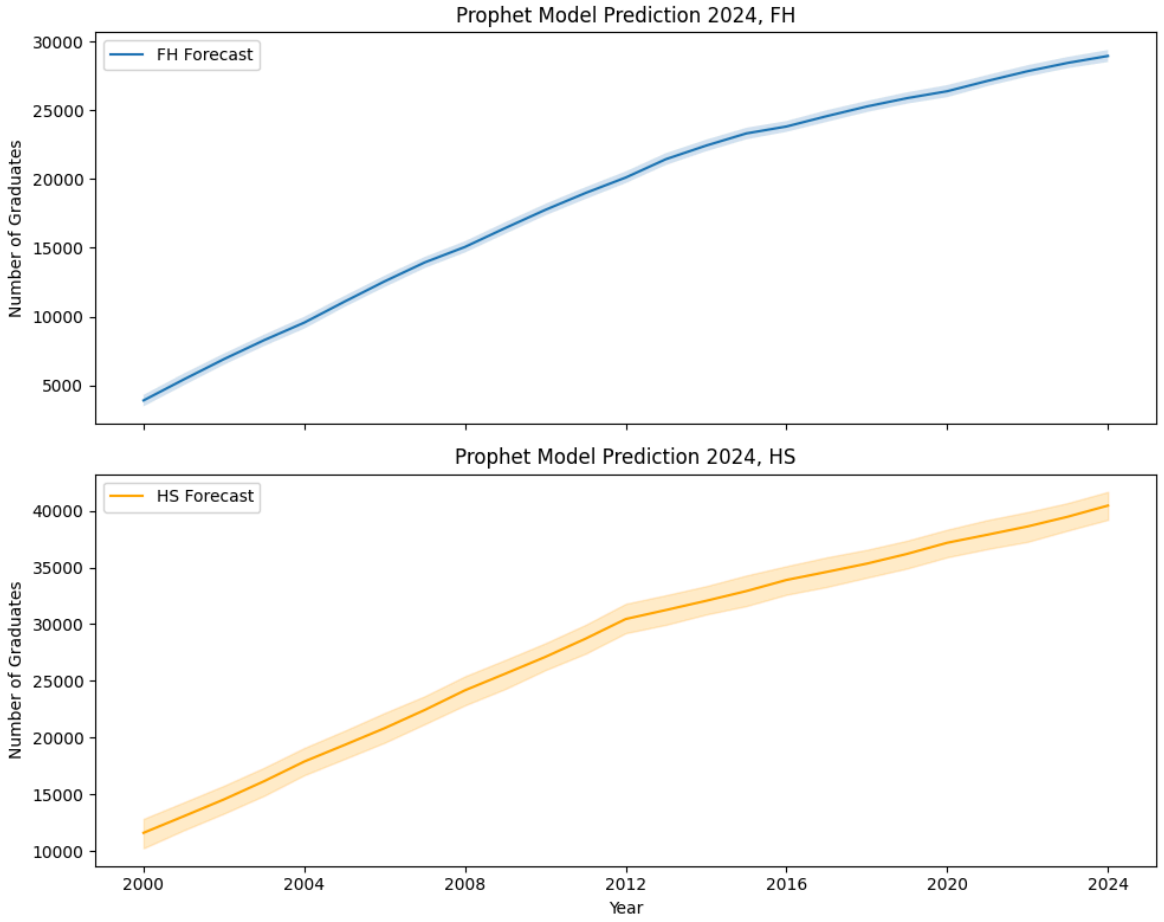

Abbildung 4: Prophet-Modellprognosen für Absolventenzahlen an FH und HS (2024)

9. Fazit

Insgesamt bestätigt die Studienarbeit die zentrale Hypothese, dass historische Abschlussdaten eine belastbare Grundlage für prädiktive Modelle im Hochschulbereich darstellen. Gleichzeitig wird deutlich, dass unterschiedliche Modellierungsansätze je nach Datencharakteristika und Hochschultyp variierende

Prognosegüten liefern. Zukünftige Arbeiten könnten durch die Erweiterung der Datenbasis und zusätzliche qualitative Analysen – beispielsweise zur Untersuchung struktureller Veränderungen im Hochschulbereich oder demographische Entwicklungen – weitere Verbesserungen erzielen und so noch fundiertere Handlungsempfehlungen bieten.

Anhänge und Verzeichnisse

Das Projekt ist in Github unter folgender URL abgelegt:
https://github.com/pluzgi/1_DataMining_Pruefung

Begleitend zu dieser Studienarbeit wurden folgende Dateien erstellt:
- Jupyter Notebook mit Python Code (“main.ipynb”)
- Screencast (Teams-Link),
(“250303_DataMining_LinkScreencast_SabineWildemann.pdf”)
- Eigenständigkeitserklärung
(“250303_Eigenständigkeitserklärung_SabineWildemann.pdf”)
- Zip-Datei mit allen Dateien (“250303_DataMining_Prüfungsleistung_
SabineWildemann.zip”)

Datensätze

1) Abschlüsse der Fachhochschulen und pädagogischen Hochschulen nach Jahr, Hochschule, Fachbereich, Examensstufe und Geschlecht, 27. Juni 2024

Zeitliche Abdeckung 1. Januar 2000 - 31. Dezember 2023
https://opendata.swiss/de/dataset/abschlusse-der-fachhochschulen-und-padagogischen-hochschulen-nach-jahr-hochschule-fachbereich-e1

2) Abschlüsse der universitären Hochschulen nach Jahr, Hochschule, Fachbereichsgruppe, Examensstufe und Geschlecht, 27. Juni 2024

Zeitliche Abdeckung 1. Januar 1980 - 31. Dezember 2023
https://opendata.swiss/de/dataset/abschlusse-der-universitaren-hochschulen-nach-jahr-hochschule-fachbereichsgruppe-examensstufe-u1

Dateien

1) CSV-Dateien mit Einzeldaten:

Abschlüsse der Fachhochschulen und pädagogischen Hochschulen:

- Zeitliche Abdeckung: 1. Januar 2000 bis 31. Dezember 2023
- Datenfelder: Kalenderjahr, Hochschule, Fachbereich, Studienstufe, Geschlecht und Anzahl der Studierenden
- Datenformat: CSV („abschluesse_FH.csv“)

2) Abschlüsse der universitären Hochschulen:

- Zeitliche Abdeckung: 1. Januar 1980 bis 31. Dezember 2023
- Datenfelder: Kalenderjahr, Hochschule, Fachbereichsgruppe, Studienstufe, Geschlecht und Anzahl der Studierenden
- Datenformat: CSV („abschluesse_HS.csv“)

3) Excel-Dateien mit detaillierten Attribut-Informationen:

Diese Dateien enthalten die detaillierten Attribut-Informationen in separaten Arbeitsblättern, wie z.B. „LEVEL“.

- „abschluesse-FH-APPENDIX.xls“
- „abschluesse-HS-APPENDIX.xls“

4) XML-Dateien mit Details zu den Datensätzen:

Diese Dateien bieten zusätzliche Informationen zu den Datensätzen.

- „abschluesse_FH.xml“
- „abschluesse_HS.xml“

Quellenverzeichnis

Bramer, M. (2007). *Principles of data mining*. Springer Science & Business Media.

Caelen, O., & Blete, M. (2025). *Anwendungen mit GPT-4 und ChatGPT entwickeln: Intelligente Chatbots, Content-Generatoren und mehr erstellen*. O’Reilly Media, Inc.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media, Inc.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35* (pp. 22199-22213). https://doi.org/10.48550/arXiv.2205.11916

Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks. *SBC Reviews on Computer Science*, *4*(1), 28–43. https://doi.org/10.5753/reviews.2024.3757

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080