

Duale Hochschule Baden-Württemberg Mannheim

Report

Data Exploration

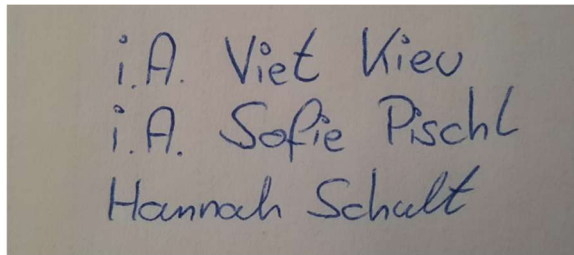
Studiengang Wirtschaftsinformatik

Data Science

Verfasser(in):	Viet Kieu (9588548), Sofie Pischl (3943911), Hannah Schult (5373022)
Kurs:	WWI-20-DSB
Vorlesung:	Data Exploration Projekt
Dozent:	Simon Poll
Bearbeitungszeitraum:	11.05.2022 – 13.07.2022

Ehrenwörtliche Erklärung

Wir versichern hiermit, dass wir die vorliegende Arbeit mit dem Titel "Report – Data Exploration Projekt" selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt haben.



i.A. Viet Kieu
i.A. Sofie Pischl
Hannah Schult

Inhaltsverzeichnis

Ehrenwörtliche Erklärung	ii
Inhaltsverzeichnis.....	iii
Abkürzungsverzeichnis.....	iv
1. Motivation und Zielsetzung	5
2. Related Work	5
3. Wirtschaftlicher Kontext.....	5
4. Verwendete Technologien und Bibliotheken.....	6
5. Präsentation der Ergebnisse	7
6. Kritische Bewertung	8
7. Quellen.....	10

Abkürzungsverzeichnis

CSV:	Comma-separated values
SQL:	Structured Query Language
JSON:	JavaScript Object Notation
ACF:	Autocorrelation function
PACF:	Partial autocorrelation
ARIMA:	Auto-Regressive Integrated Moving Average
MSE:	Mean squared error, mittlerer quadratischer Fehler

1. Motivation und Zielsetzung

Es gibt verschiedene Beweggründe, weswegen sich die Studierenden des Kurses WWI20DSB für ihr Studium entschieden haben. Ein Beweggrund kann auf jeden Fall sein, dass ein Abschluss in Wirtschaftsinformatik mit der Vertiefung Data Science gute Jobaussichten bieten¹. Zusätzlich hat man als Informatiker eins der höchsten Einstiegsgehälter². Deshalb ist es sehr interessant zu sehen, wie sich die Gehälter im IT-Sektor innerhalb von Deutschland entwickeln und von welchen Faktoren das Gehalt abhängig ist.

Diese Arbeit beschäftigt sich also mit der Gehaltsentwicklung von IT-Gehältern in Deutschland. Zusätzlich werden die Faktoren analysiert, welche das Gehalt beeinflussen. Am Ende steht dann eine Anwendung, die anhand von einigen Faktoren (wie beispielsweise Geschlecht, Alter und Berufserfahrung) das Gehalt vorhersagt.

2. Related Work

Es gibt im Internet mehrere Anwendungen, die ihren Nutzern Gehaltsvorschläge geben bzw. auch als Gehaltvergleichsportale dienen. Darunter sind beispielsweise glassdoor.com, kununu oder auch Gehalt.de.

Außerdem gibt es verschiedene Paper zu diesen Themen. Beispielsweise erscheint jährlich der Report „Global Knowledge IT Skills and Salary Report“ von Global Knowledge (Unternehmen für IT und Technologien Trainings), welches z.B. von Trendentwicklungen, Challenges in der Branche und den höchstbezahltesten Zertifikate handelt³.

3. Wirtschaftlicher Kontext

Unsere Anwendung hat zwei Zielgruppen. Auf der einen Seite ist unsere Anwendung für Unternehmen interessant, die überprüfen wollen, in welchem Bereich die Gehälter für ihre Mitarbeiter liegen sollen. Auf der anderen Seite ist die Anwendung auch für berufstätige Menschen sehr interessant. Möchte ein Arbeitnehmer das Unternehmen wechseln und in diesem Rahmen seinen Marktwert, also wie viel

¹ Koch, Julia/Joachim Mohr (2006): A-4d55a003-0002-0001-0000-000049849307, in: *DER SPIEGEL, Hamburg, Germany*, 10.12.2006, [online] <https://www.spiegel.de/politik/gute-faecher-schlechte-faecher-a-4d55a003-0002-0001-0000-000049849307> [abgerufen am 13.07.2022].

² Bestbezahlte Berufe 2022 (2022): UNICUM Karrierezentrum, [online] <https://karriere.unicum.de/erfolg-im-job/gehalt-finanzen/bestbezahlte-berufe> [abgerufen am 13.07.2022].

³ IT Skills and Salary Report (o. D.): skillsoft global knowledge, [online] <https://www.globalknowledge.com/us-en/content/salary-report/it-skills-and-salary-report/> [abgerufen am 13.07.2022].

Gehalt ihm zusteht, überprüfen, dann kann er die Anwendungen nutzen und anhand seiner persönlichen Faktoren eine Vorhersage seines Gehaltes erhalten. Ein anderer Anwendungsfall ist es, wenn eine Gehaltsverhandlung ansteht. Auch in diesem Falle kann der Mitarbeiter sich sein Gehalt vorhersagen lassen, um einen Richtwert zu haben, an dem er sich in der Gehaltsverhandlung orientieren kann.

4. Verwendete Technologien und Bibliotheken

Verwendete Bibliotheken:

- **Pandas:** Pandas leitet sich von „panel data“ (strukturierte, multidimensionale Daten) ab und unterstützt die Verwaltung, Modellierung und Analyse von Daten. Die Bibliothek basiert auf Numpy und Matplotlib. Vor allem Daten im CSV-, Excel-, SQL oder JSON Format lassen sich leicht einlesen und manipulieren.
- **Numpy:** Numpy steht für „Numeric Python und unterstützt numerische Berechnungen und mathematische Verfahren in Python. Arrays lassen sich schnell erstellen und berechnen.
- **Matplotlib:** Matplotlib ist eine umfangreiche Bibliothek zur Erstellung von Diagrammen. Es werden unter anderem Linien-, Stab- oder Kuchendiagramme, Histogramme, Boxplots, Kontourdiagramme, aber auch dreidimensionale Diagramme und Funktionenplots unterstützt. Das Modul pyplot ermöglicht den Zustand von Grafiken zu ändern.
- **Datetime:** Das Modul Datetime ist für die Verarbeitung von Datum und Zeit notwendig. Der Fokus liegt dabei auf dem Extrahieren von Attributen und der Manipulation des Outputs.
- **Adfuller, ACF, PACF from statsmodels.tsa.stattools:** Statsmodels.tsa.stattools bietet statistische Werkzeuge zur Analyse von Time Series Daten. Adfuller wird für den Augmented Dickey-Fuller-Test verwendet und testet auf eine Einheitswurzel in einem univariaten Prozess bei Vorhandensein einer seriellen Korrelation. ACF und PACF werden verwendet, um die Autokorrelationsfunktion und die partielle Autokorrelationsfunktion zu berechnen.
- **seasonal_decompose from statsmodels.tsa.seasonal:** Seasonal decompose berechnet die saisonale Zerlegung mit Hilfe des gleitenden Durchschnittswerts.
- **ARIMA from statsmodels.tsa.arima_model:** ARIMA steht für Autoregressive Integrated Moving Average. Es ist die einfachste Version der ARIMA-type Models und ermöglicht die Beschreibung und Analyse von Zeitreihen. Es beinhaltet einen autoregressiven Teil (AR-Modell) und einen gleitenden Mittelwertbeitrag (MA-Modell).
- **scikit-learn:** scikit-learn bietet einfache und effiziente Tools für vorhersagende Datenanalysen. Die Bibliothek bietet beispielsweise Algorithmen zur Regression oder Modelauswahl an.

Als Modell zur Vorhersage der Gehaltsentwicklung innerhalb Deutschlands wurde ARIMA (Auto-Regressive Integrated Moving Average) genutzt. ARIMA ist eine Modellklasse der Zeitreihenanalyse. Die beobachtete Zeitreihe wird unter Verwendung der gegebenen Werte (autoregressiv, AR) und des gleitenden Durchschnitts (moving average, MA) der Werte angenähert. Dieser auch im ARMA Modell vorkommende Teil wird im ARIMA Modell um die Trendbeseitigung und Herstellung der Stationarität ergänzt, sodass die zu analysierende Zeitreihe auch Trendverläufe aufweisen kann. Auf Basis des Modells lassen sich kurzfristige Vorhersagen treffen, sowie Schätzungen und Validierungen vornehmen.

Die Modelle für die Salary Prediction waren Lineare Regression, Polynominal Features Transformation, Ridge Regression und Random Forest.

Lineare Regression ist ein statistisches Verfahren, das eine abhängige Variable durch eine oder mehrere unabhängige Variablen erklärt. Das Ziel ist es die MSE zu minimieren.

Polynominal Feature Transformation ist eine Art Feature-Engineering bei der neue Eingabe Features basierend auf den bereits vorhandenen Features erzeugt werden. Der Grad des Polynoms wird genutzt, um die Anzahl der neuen Features zu steuern. Wenn der Grad 3 verwendet wird, entstehen für jedes Feature zwei weitere neue Features⁴.

Ridge Regression ist eine Weiterentwicklung der Linearen Regression. Hierbei gibt es den Bestrafungsparameter Alpha, der dafür sorgt, dass das Modell Overfitting vermeidet. Dies geschieht, indem es den Ridge Term unterschiedlich stark bestraft. Die Grundidee ist Bias zu schaffen, da so die Varianz reduziert wird und damit ein niedriger MSE entsteht⁵.

Bei einem Random Forest existieren mehrere unkorrelierten Entscheidungsbäume, die eine Vorhersage treffen. Von allen getroffenen Vorhersagen wird dann der Mittelwert gebildet.

5. Präsentation der Ergebnisse

Der erste Wert in den unterstehenden Tabellen ist der durchschnittliche Fehler (mittlere absolute Abweichung). Der zweite Wert berechnet sich aus der prozentualen Abweichung der Predictions zum

4

Polynomial Feature Transform in Machine Learning - (2021): steps, [online] <https://blog.stepskochi.com/polynomial-feature-transform-in-machine-learning/#:%7E:text=Polynomial%20feature%20Transformation%20is%20a%20type%20of%20feature,add%20two%20new%20variables%20for%20each%20input%20variable>. [abgerufen am 13.07.2022].

5

Ridge regression (o. D.): StatLect, [online] <https://www.statlect.com/fundamentals-of-statistics/ridge-regression#:~:text=Ridge%20regression%20is%20a%20term%20used%20to%20refer, but%20has%20lower%20variance%20than%20the%20OLS%20estimator>. [abgerufen am 13.07.2022].

Originalwert. Also liegt der vorhergesagte Wert 10% über oder unter dem Originalwert, dann definieren wir das als eine 90 %ige Genauigkeit.

	Training	Test
Lineare Regression	7752,16 € 88,1 %	8936,75 € 86,06 %
Polynomial Features Transformation	5777,83 € 91,36 %	11351,54 € 81,2 %
Ridge Regression	6251,07 € 90,58%	9163,56 € 85, 53%
Random Forest	3732,54€ 94,15%	9024,41 € 85,46%

Hyperparameter Tuning auf dem Random Forest. Zuerst wurden die Parameter mit Random Search bestimmt und danach wurde mit Grid Search die optimalen Parameter auf Basis der Ergebnisse des Random Search gesucht.

	Training	Test
Random Search	6121, 86 € 90, 54 %	8760, 57 € 85, 96 %
Grid Search basierend auf Random Search	6987, 28 € 88,94 %	8966, 07 € 85, 62 %

Das beste Ergebnis mit 86,06 % auf den Testdaten wird mit der linearen Regression erzeugt. Allerdings wurden mit den anderen Modellen relativ ähnliche Ergebnisse erzielt.

6. Kritische Bewertung

Insgesamt ist der Datensatz mit rund 2600 Daten sehr klein. Auch einzelne Features konnten nicht richtig genutzt werden aufgrund der zu geringen Datenmenge. So wurde beispielsweise kaum eine Korrelation zwischen Gehalt und Unternehmensbranche errechnet, obwohl dort in der Realität eine Korrelation besteht. Das kommt daher, dass die einzelnen Branchen in unserem Datensatz zu wenige Daten enthielten.

Wir haben versucht durch Anfragen bei Gehaltsvergleichsportalen unsere Datenbasis zu vergrößern, allerdings bekommen wir nur negative oder gar keine Rückmeldungen.

Auch konnten wir nicht alle am Anfang gesteckten Ziele erreichen, da unsere Daten auch dafür nicht geeignet waren. Beispielsweise wollten wir zu Beginn herausfinden, in welchem Bundesland das höchste Gehalt möglich ist, während die Lebensnebenkosten am geringsten sind, um so den optimalen Arbeitsplatz zu finden. Aber auch die Daten zum Wohnort der Arbeitnehmer waren zu ungleich verteilt, um daraus gute Rückschlüsse ziehen zu können.

Aufgrund dessen haben wir gelernt, dass die Größe des Datensatzes sowie die Mengen der Daten pro Feature sehr relevant ist. Auch haben unsere Ergebnisse gezeigt, dass ein komplexeres Modell nicht

unbedingt ein besseres Endergebnis liefern muss. Ein weiterer kursübergreifender Punkt ist, dass wir Ridge Regression praktisch anwenden konnten, welche wir bereits in dem Modul Applied Optimization Techniques kennengelernt haben.

Anmerkungen zum Ausführen des Quellcodes sind in der README Datei in GitHub zu finden.

7. Quellen

Koch, Julia/Joachim Mohr (2006): A-4d55a003-0002-0001-0000-000049849307, in: *DER SPIEGEL, Hamburg, Germany*, 10.12.2006, [online] <https://www.spiegel.de/politik/gute-faecher-schlechte-faecher-a-4d55a003-0002-0001-0000-000049849307> [abgerufen am 13.07.2022].

Bestbezahlte Berufe 2022 (2022): UNICUM Karrierezentrum, [online] <https://karriere.unicum.de/erfolg-im-job/gehalt-finanzen/bestbezahlte-berufe> [abgerufen am 13.07.2022].

IT Skills and Salary Report (o. D.): skillsoft global knowledge, [online] <https://www.globalknowledge.com/us-en/content/salary-report/it-skills-and-salary-report/> [abgerufen am 13.07.2022].

Polynomial Feature Transform in Machine Learning - (2021): steps, [online] <https://blog.stepskochi.com/polynomial-feature-transform-in-machine-learning/#:%7E:text=Polynomial%20feature%20Transformation%20is%20a%20type%20of%20feature,add%20two%20new%20variables%20for%20each%20input%20variable.> [abgerufen am 13.07.2022].

Ridge regression (o. D.): StatLect, [online] <https://www.statlect.com/fundamentals-of-statistics/ridge-regression#:~:text=Ridge%20regression%20is%20a%20term%20used%20to%20refer,but%20has%20lower%20variance%20than%20the%20OLS%20estimator.> [abgerufen am 13.07.2022].