Data Science Salarys Exploration

Clemens Gulde

April 3, 2024

Abstract

Die vorliegende Studie untersucht Gehaltsdaten im Bereich Data Science anhand eines Datensatzes mit 6.599 Einträgen. Die Analyse zeigt eine deutliche geografische Konzentration auf die USA sowie eine Präferenz für Vollzeitbeschäftigung und Senior-Positionen. Mithilfe von Feature Engineering wurden relevante Merkmale identifiziert und für die Vorhersage vorbereitet. Die Auswahl und Schulung eines XGBoost-Modells ergab die beste Leistung für die Gehaltsvorhersage, wobei Hyperparametertuning die Modellgenauigkeit weiter verbesserte. Ein Residual Plot zeigt, dass das Modell die Gehälter gut vorhersagen kann, ohne systematische Fehler aufzuweisen.

Alle Dateien und Ergebnisse in folgendem GitHub-Repositorie einsehbar:
https://github.com/wi22215/DS-Salary-Exploration

Contents

1 Methodik **2**
- 1.1 Datenexploration . . . 2
  - 1.1.1 Beschreibung der Daten . . . 2
  - 1.1.2 Charakterisierung des Datensatzes . . . 2
- 1.2 Feature Engineering . . . 3
- 1.3 Modellauswahl und Training . . . 4
  - 1.3.1 Split des Datensatzes . . . 4
  - 1.3.2 Auswahl der Metriken . . . 4
  - 1.3.3 Auswahl und Beschreibung der ML-Methode . . . 4
  - 1.3.4 Hyperparametertuning . . . 5
- 1.4 Evaluation . . . 5

2 Ergebnisse **5**

1 Methodik

1.1 Datenexploration

1.1.1 Beschreibung der Daten

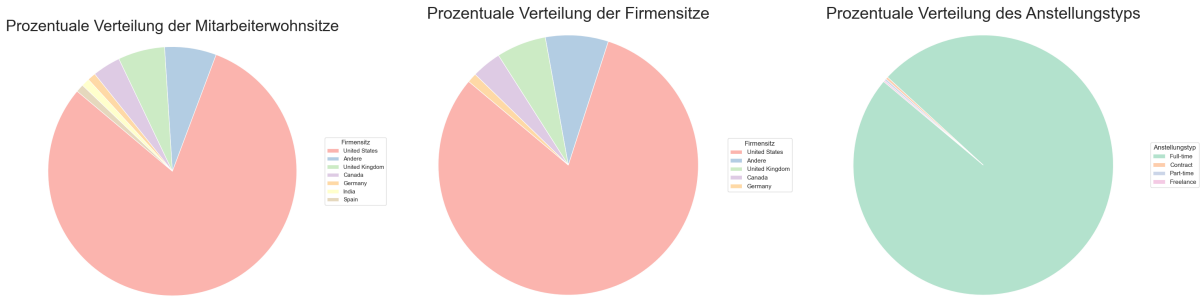
Das untersuchte Datenset enthält spezifische Daten zu Gehältern im Bereich Data Science. Es umfasst 6.599 Datensätze mit folgenden Schlüsselattributen:

Table 1: Übersicht der Datensatzmerkmale

Merkmal	Beschreibung
job_title	132 einzigartige Rollen im Bereich Data Science.
experience_level	Kategorisiert die Erfahrung des Mitarbeiters in vier Stufen.
employment_type	4 Arten der Anstellung, wie z.B. Vollzeit, Teilzeit, etc.
work_models	Remote, On-site oder in hybrides Modell.
work_year	2020 bis 2024
employee_residence	Das Land, in dem der Mitarbeiter ansässig ist.
salary	Das jährliche Grundgehalt des Mitarbeiters in der lokalen Währung.
salary_currency	Die Währung, in der das Gehalt ausgezahlt wird.
salary_in_usd	Grundgehalt umgerechnet in US-Dollar.
company_location	Das Land, in dem das Unternehmen seinen Sitz hat.
company_size	Eine Einteilung der Unternehmensgröße in Klein, Mittel und Groß.

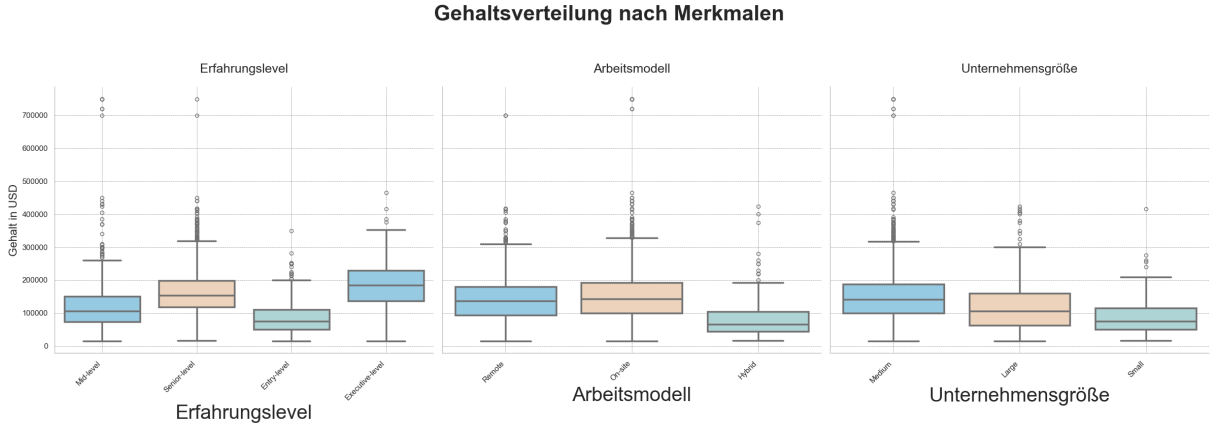
Die Datenqualität kann hinsichtlich der Konsistenz hochwertig bewertet werden, da weder NULL-Werte noch duplizierte Zeilen vorhanden sind.

1.1.2 Charakterisierung des Datensatzes



Analyse der Balanciertheit Eine signifikante geographische Konzentration auf die USA, das Vereinigte Königreich und Kanada erkennbar, mit einer überwältigenden Mehrheit der Daten aus den USA. Dies deutet auf eine potenzielle Verzerrung in geographischer Hinsicht hin, die die Generalisierbarkeit der Ergebnisse auf eine globale Ebene einschränken könnte. Des Weiteren zeigt die Analyse eine starke Präferenz für Vollzeitbeschäftigungen (99 Prozent) und eine ungleiche Verteilung über verschiedene Karrierestufen, wobei Senior Positionen dominieren (60 Prozent). Auch im Hinblick auf die Arbeitsmodelle zeigt sich eine Tendenz zu traditionellen On-Site Arrangements. Diese Unausgeglichenheiten könnten die Relevanz und Anwendbarkeit der aus dem Datensatz gewonnenen Einsichten für bestimmte Gruppen oder Szenarien limitieren.

Analyse der Gehälter Executive- und Senior-level Positionen haben im Allgemeinen höhere Gehälter, was durch höhere Mediane und Quartilswerte im Vergleich zu anderen Erfahrungsstufen deutlich wird. Entry-level Positionen haben erwartungsgemäß die niedrigsten Gehälter, mit einer engeren Verteilung und niedrigeren Ausreißern. Die Gehaltsverteilungen zwischen den verschiedenen Arbeitsmodellen sind im Remote als auch On-Site Modell sehr ähnlich, das Hybrid Modell schneidet eher schlechter ab. Interessanterweise gibt es eine Tendenz, die zeigt, dass eine mittlere Unternehmensgröße systematisch höhere Gehälter bietet mit mehreren und deutlich höheren Ausreißer. Kleine Unternehmen schneiden unerwarteter Weise am schlechtesten ab.



Analyse der Ausreißer Die hohen Ausreißer im Gehalt scheinen hauptsächlich bei Mitarbeitern in Mid- und Senior-Level Positionen aufzutreten und sind geographisch ausschließlich auf die USA beschränkt. Diese Mitarbeiter arbeiten überwiegend in mittelgroßen Unternehmen und in On-site-Arbeitsmodellen. Die Gehälter dieser Ausreißer variieren erheblich, wobei das höchste festgestellte Gehalt bei 750.000 USD liegt.

Bewertung der Daten Der analysierte Datensatz zeigt eine deutliche geographische Konzentration auf die USA, was zu einer Verzerrung führt und die Generalisierbarkeit der Ergebnisse limitiert. Die Daten sind stark auf Vollzeit- und Senior-Positionen fokussiert, mit einer Präferenz für On-Site-Arbeitsmodelle. Mittelgroße Unternehmen in den USA neigen dazu, höhere Gehälter zu zahlen, besonders für Mid- und Senior-Level Positionen, was signifikante Gehaltsausreißer nach oben zur Folge hat. Diese Faktoren beeinträchtigen die Datenqualität insofern, als sie eine begrenzte Sicht auf die globale Gehaltslandschaft bieten und spezifische Verzerrungen enthalten, die bei der Interpretation und Anwendung der Ergebnisse berücksichtigt werden müssen.

1.2 Feature Engineering

One-Hot-Encoding wird eingesetzt, um kategoriale Daten in eine Form umzuwandeln, die von maschinellen Lernalgorithmen effektiv verarbeitet werden kann und um Korrelationen zwischen den kategorischen Features und der Zielvariablen herzustellen.
In folgender Tabelle wurden die relavantesten Features zum einen mit ihrer Korrelation mit der Zielvariablen und dem F-Score dargestellt. Dieser hilft dabei, die relative Wichtigkeit von Features in Modellen des maschinellen Lernens zu bestimmen, indem er misst, wie gut jedes Feature die Variabilität der Zielvariable unter Berücksichtigung der Gruppenunterschiede erklärt.

Feature	F-Score	Korrelation mit salary_in_usd
salary_currency_USD	976.46	0.359
employee_residence_United States	913.69	0.349
company_location_United States	842.55	0.337
experience_level_Senior-level	645.71	0.299
experience_level_Entry-level	492.62	-0.264
salary_currency_EUR	395.07	-0.238
experience_level_Mid-level	329.80	-0.218
job_title_Data Analyst	328.46	-0.218

Table 2: F-Scores und Korrelationen ausgewählter Features mit salary_in_usd

1.3 Modellauswahl und Training

1.3.1 Split des Datensatzes

Der Datensatz wurde in ein Trainingsset und ein Testset mit einem Verhältnis von 80 Prozent zu 20 Prozent aufgeteilt. Diese Aufteilung ermöglicht es, das Modell mit einer umfangreichen Datenmenge zu trainieren, während ein signifikanter und repräsentativer Datenanteil für die unabhängige Bewertung der Modellleistung zurückgehalten wird, was eine ausgewogene Basis für das Training und die Validierung bietet.

1.3.2 Auswahl der Metriken

Für die Bewertung des Modells wurden der Mittlere Quadratische Fehler (MSE) und der R^2-Score ausgewählt. Der MSE misst die durchschnittliche quadratische Abweichung zwischen den vorhergesagten und den tatsächlichen Werten, was ein direktes Maß für die Vorhersagegenauigkeit darstellt. Der R^2-Score gibt an, welcher Anteil der Varianz in der Zielvariable durch das Modell erklärt wird, und bietet somit Einblick in die Anpassungsgüte des Modells an die beobachteten Daten.

1.3.3 Auswahl und Beschreibung der ML-Methode

Modell	MSE	R^2-Score
Random Forest	4079052085.233043	0.306
XGBoost	3975755486.1515627	0.324
LightGBM	4042935304.3532634	0.312

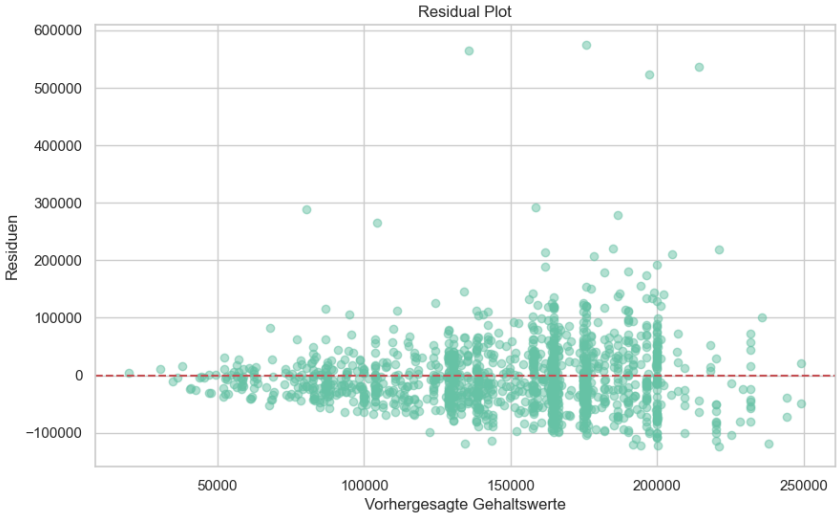
Table 3: Vergleich der Modellleistungen

Während die Modelle eine starke Leistung zeigen, übertrifft XGBoost den Random Forest und den LightGBM in diesem spezifischen Vorhersagekontext deutlich, sowohl in Bezug auf die Vorhersagegenauigkeit (niedrigerer MSE) als auch die Modellanpassung (höherer R^2-Score). XGBoost, kurz für Extreme Gradient Boosting, ist ein fortschrittliches Ensemble-Lernverfahren, das auf der Methode des Gradient Boosting basiert und darauf abzielt, die Leistung und Geschwindigkeit der Modellbildung und Vorhersage zu optimieren. Es kombiniert eine große Anzahl von schwachen Entscheidungsbäumen zu einem starken Modell durch sequenzielle Modellierung von Fehlern vorheriger Bäume, wobei es sich auf die Korrektur von Fehlern konzentriert und dabei sowohl Regularisierungstechniken zur Vermeidung von Overfitting als auch eine hoch effiziente Implementierung für schnelles Training nutzt.

1.3.4 Hyperparametertuning

Nach der Optimierung der Hyperparameter mittels eines Parametergrids erreichte das verbesserte XGBoost-Modell einen MSE von 3924475499.101123 und einen R^2-Score von 0.332. Diese Ergebnisse zeigen eine signifikante Verbesserung im Vergleich zum ursprünglichen XGBoost-Modell, was auf eine effektive Anpassung der Hyperparameter hinweist. Die Reduzierung des MSE und die Erhöhung des R^2-Scores deuten darauf hin, dass das Modell präzisere Vorhersagen liefert und einen größeren Anteil der Varianz in der Zielvariable erklären kann.

1.4 Evaluation



Der Residual Plot dient zur Bewertung der Leistung des erstellten Regressionsmodells zur Vorhersage der Gehälter basierend auf den Merkmalen im Datensatz. Die horizontalen Achsen repräsentieren die vorhergesagten Gehaltswerte, während die vertikalen Achsen die Residuen darstellen, d.h. die Differenzen zwischen den tatsächlichen Gehaltswerten und den vom Modell vorhergesagten Werten. Idealerweise sollten die Residuen zufällig um die Nulllinie herum verteilt sein, was darauf hinweist, dass das Modell die Daten gut modelliert und keine systematischen Fehler aufweist. In diesem Plot sind die Residuen weitgehend zufällig um die Nulllinie verteilt, was darauf hindeutet, dass das Modell die Gehaltswerte gut vorhersagen kann und die Unterschiede zwischen den vorhergesagten und tatsächlichen Werten zufällig sind.

2 Ergebnisse

Die detaillierte Untersuchung und Analyse des Datensatzes im Bereich Data Science hat eine hohe Datenqualität bestätigt. Die geografische Konzentration auf Länder wie die USA, das Vereinigte Königreich und Kanada deutet auf eine potenzielle Verzerrung hin, die bei der Interpretation der Ergebnisse berücksichtigt werden muss. Trotz dieser Einschränkung bot die Analyse wertvolle Einblicke in die Verteilung der Gehälter basierend auf verschiedenen Faktoren. Durch das Feature Engineering konnte eine effektive Verarbeitung der kategorialen Daten für maschinelles Lernen gewährleistet und die Wichtigkeit der Features im Hinblick auf die Zielvariable aufgedeckt werden. Insbesondere zeichnete sich das XGBoost-Modell durch eine überlegene Vorhersagegenauigkeit und Modellanpassung aus und übertraf damit andere Modelle wie Random Forest und LightGBM. Nach der Optimierung der Hyperparameter wies das verbesserte XGBoost-Modell einen MSE von 3924475499.101123 und einen R^2-Score von 0.332 auf, was eine deutliche Leistungssteigerung darstellt.

Evaluation

Die Testdaten umfassten folgende Merkmale:

Merkmal	**Wert**
Jobtitel	Machine Learning Engineer
Erfahrungslevel	Senior
Beschäftigungsart	Vollzeit
Arbeitsmodell	Vor Ort
Arbeitsjahr	2024
Wohnsitz des Mitarbeiters	Vereinigte Staaten
Gehaltswährung	USD
Unternehmensgröße	Mittel
Unternehmensstandort	Vereinigte Staaten

Das Modell prognostizierte ein Gehalt von **199.767,73 USD**. Zur Überprüfung dieser Vorhersage wurde der Durchschnitt aller Gehälter aus dem Datensatz gebildet, auf die eben diese Merkmale zutreffen. Der durchschnittliche Gehaltswert betrug **198.257,49 USD**. Die Vorhersage des Modells zeigt eine hohe Übereinstimmung mit den tatsächlichen Gehaltsdaten, was auf eine effektive Leistung des Modells bei der Abschätzung von Gehältern für Machine Learning Engineers unter den definierten Bedingungen hindeutet.