
LOONEY TUNES
& © Warner Bros. Entertainment Inc. (s18)

Gehälter im IT Sektor

Payio

Viet Kieu, Hannah Schult, Sofie Pischl

Our Goals

Zeitreihenprognose

Featureanalyse

Persönliche Prognose

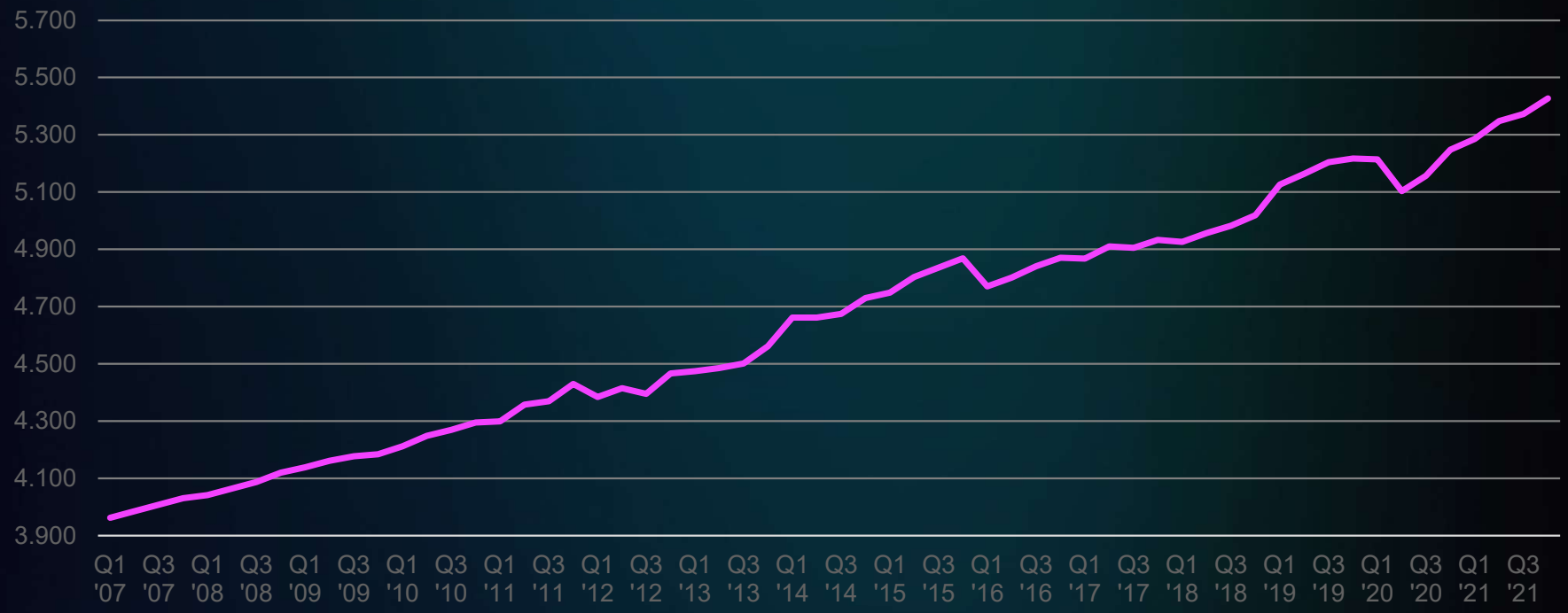
TABLE OF CONTENTS

01

Zeitreihenprognose

Entwicklung der IT Gehälter

Unsere Daten



Durchschnittlicher Monatsverdienst vollzeitbeschäftigter Arbeitnehmer in der ITK-Branche in Deutschland vom 1. Quartal 2007 bis zum 4. Quartal 2021 (in Euro)

Vorgehensweise

Stationarität

Testen, ob sich statistische Werte verändern

Trend

Identifizieren des Trendes

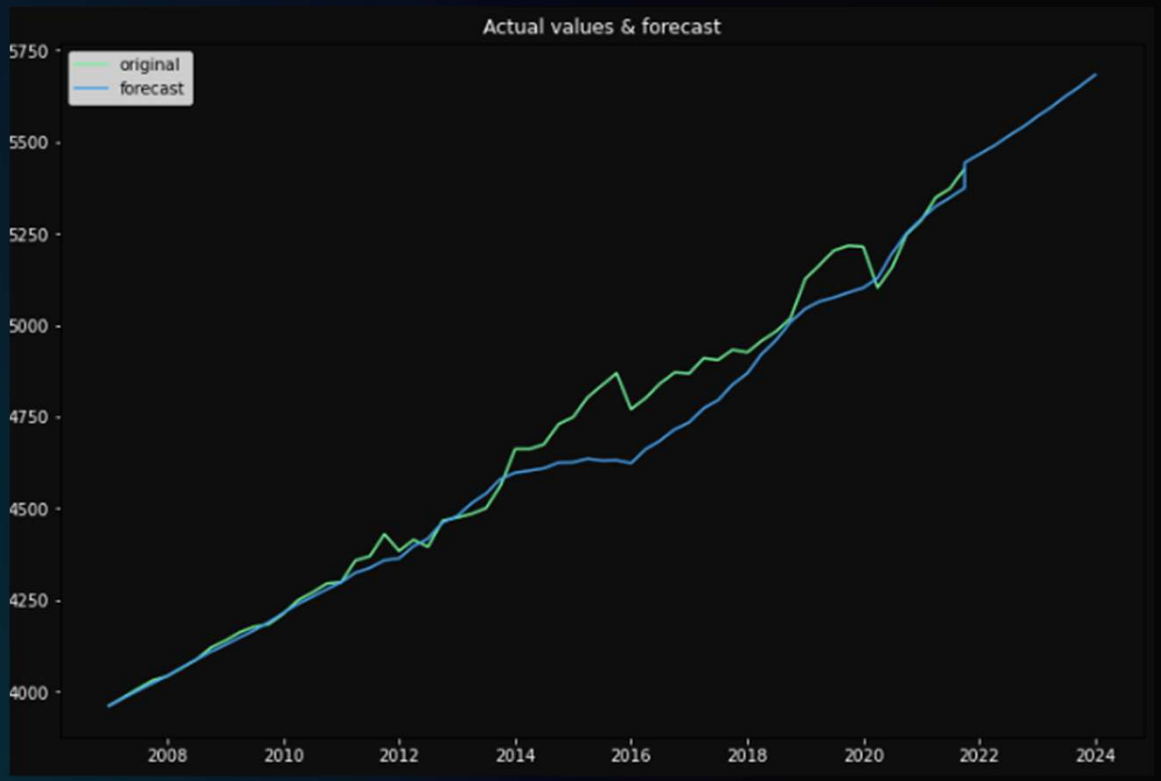
Saisonalität

Prüfen, ob zyklische Schwankungen vorliegen

Prognose

Vorhersage neuer Werte

Zeitreihenprognose



Originale Daten

Vorhergesagtes Gehalt

02

Gehaltsprognose

Business Use Case

Für Unternehmen

Für Privatpersonen

Daten - Gehaltsprognose

3 Jahre

Mehr Daten, weniger Features

Nur 2020

Weniger Daten, mehr Features

Datapreprocessing

Spaltennamen anpassen

Mergen

Outliner entfernen

0 Values / NAN Werte entfernen

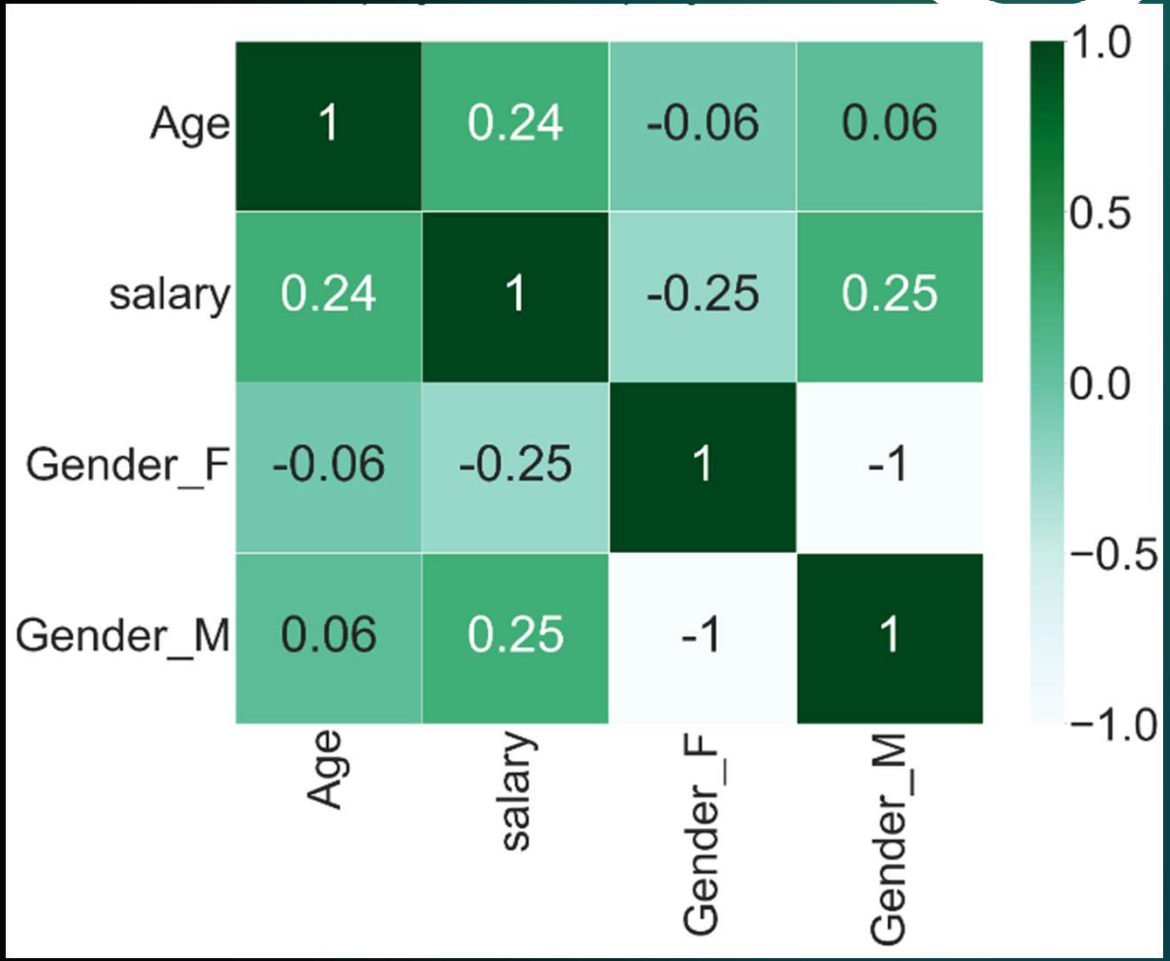
Wir haben 405 Datenzeilen verloren

Unser Dataframe

	Age	Gender	City	yearsExperience	SeniorityLevel	salary	MainLanguage	CompanySize	CompanyType
0	43	M	München	11	Senior	77000	Deutsch	100-1000	Product
1	33	F	München	8	Senior	65000	Deutsch	50-100	Product
2	32	M	München	10	Senior	88000	Deutsch	1000+	Product
3	25	M	München	6	Senior	78000	English	1000+	Product
4	39	M	München	10	Senior	69000	English	100-1000	Ecom retailer
...	...	...	...	...	...	...	...	...	...
3004	31	M	Berlin	9	Senior	70000	English	51-100	Product
3005	33	M	Berlin	10	Senior	60000	English	1000+	Product
3006	39	M	Munich	15	Lead	110000	English	101-1000	eCommerce
3007	26	M	Saarbrücken	7	Middle	38350	German	101-1000	Product
3008	26	M	Berlin	2	Middle	65000	English	51-100	Startup

2603 rows × 9 columns

Korrelationsmatrix



Unsere Features

- Gender
- Unternehmenstyp
- City
- Alter
- Seniority Level
- Berufserfahrung
- Unternehmens-größe
- Hauptsprache

Unsere Modelle

Lineare Regression

Ridge Regression

Polynomial Features

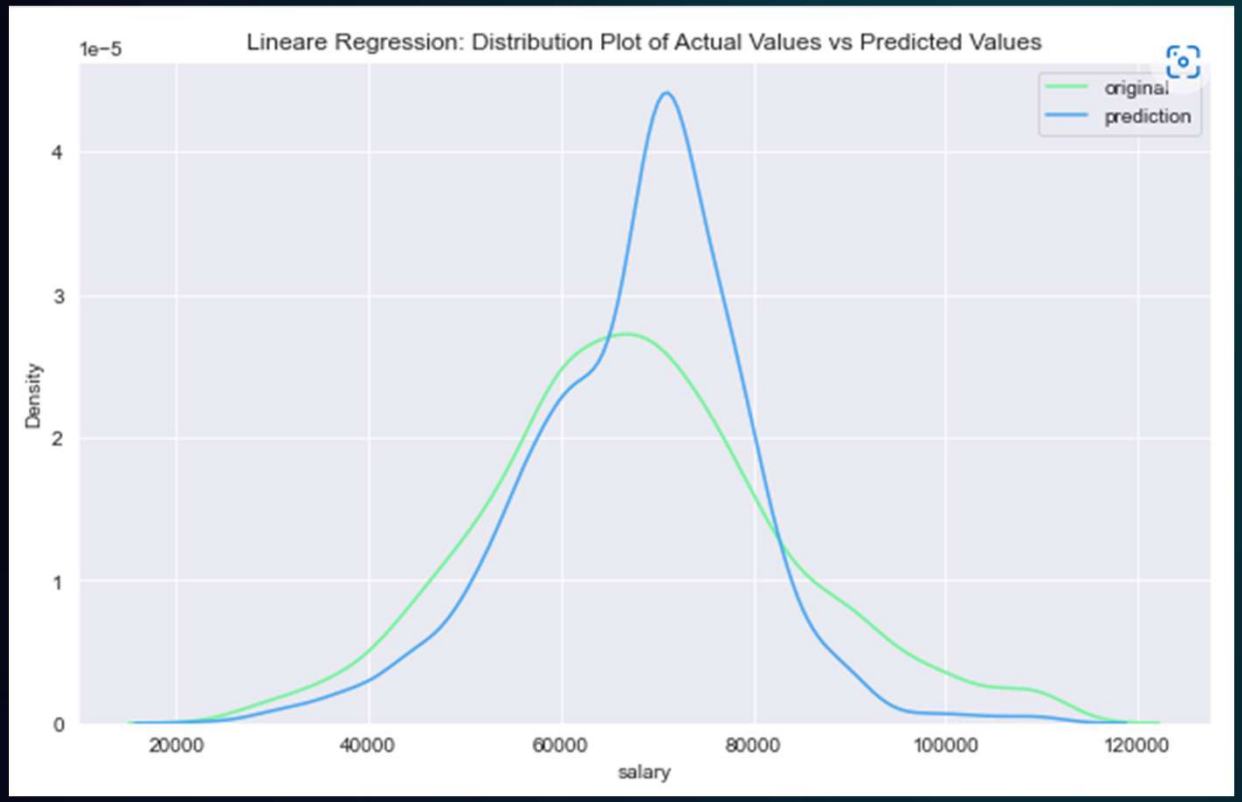
Random Forest

Welches Gütemaß?

$$\mathrm{d} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

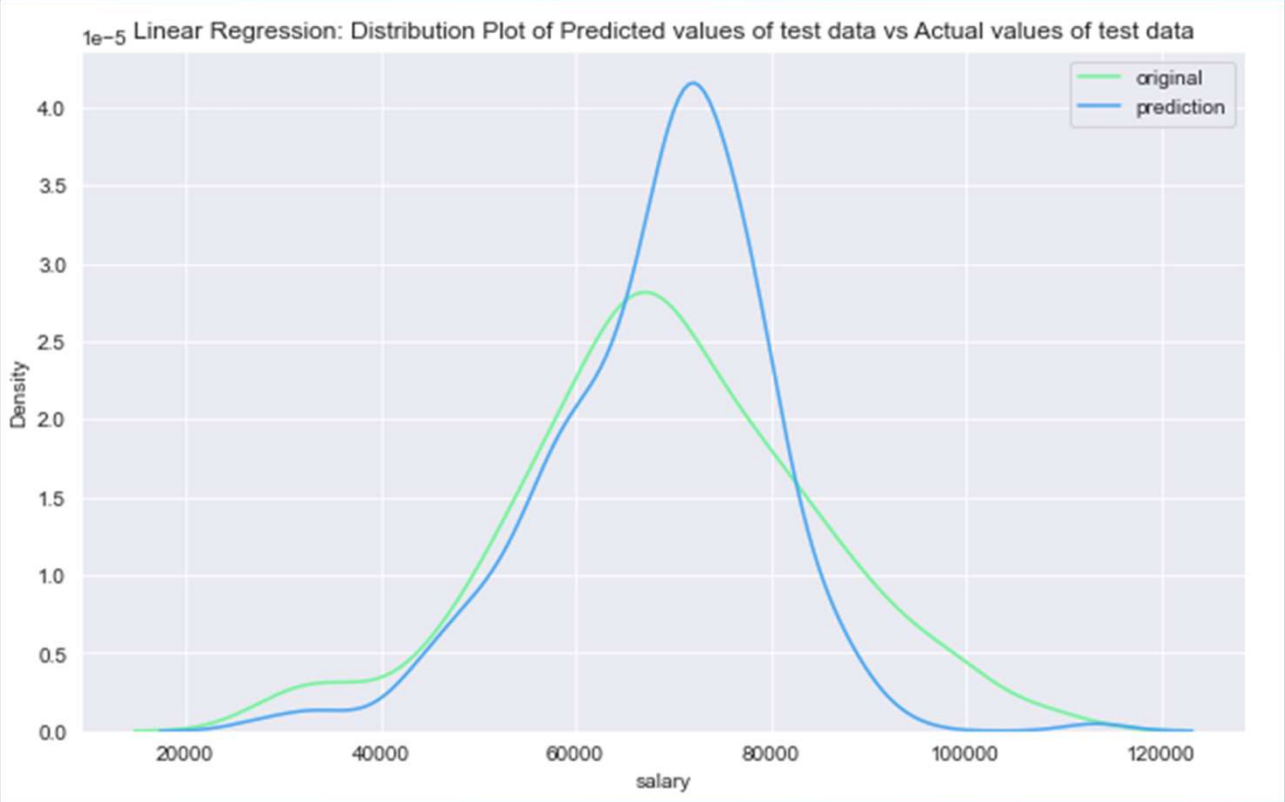
mittlere absolute Abweichung

Lineare Regression



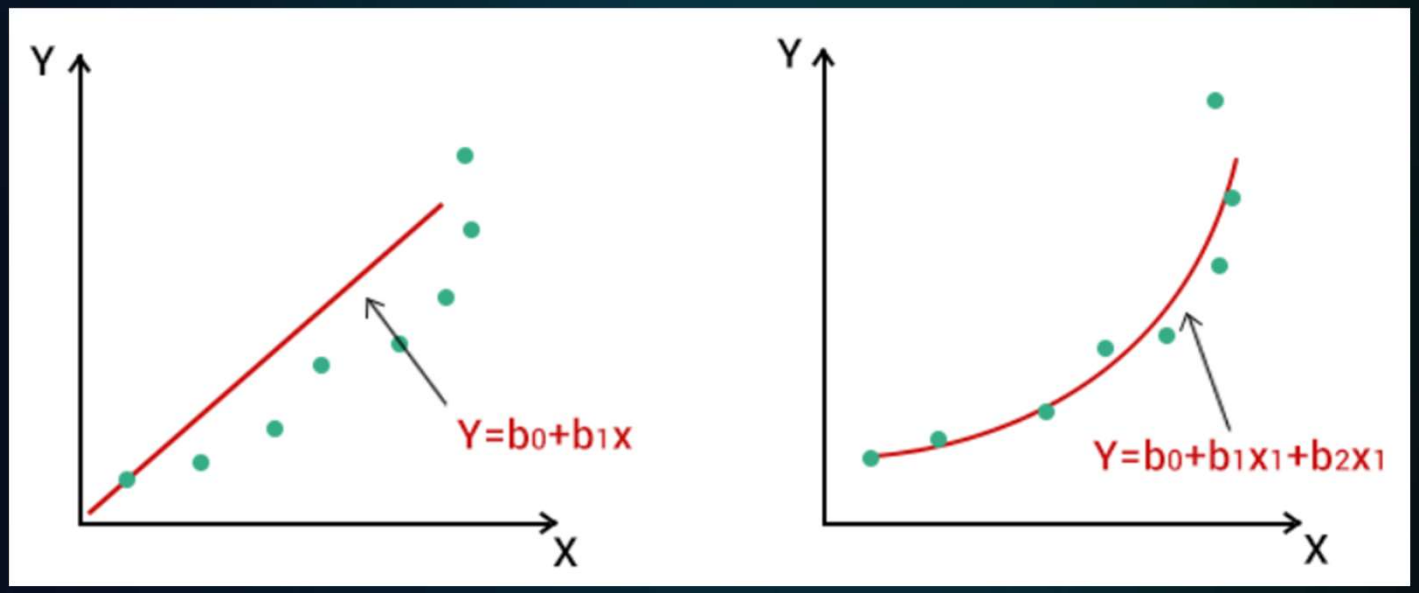

```
Model Performance
Average Error: 7752.1617 €.
Accuracy = 88.10%.
```

Lineare Regression

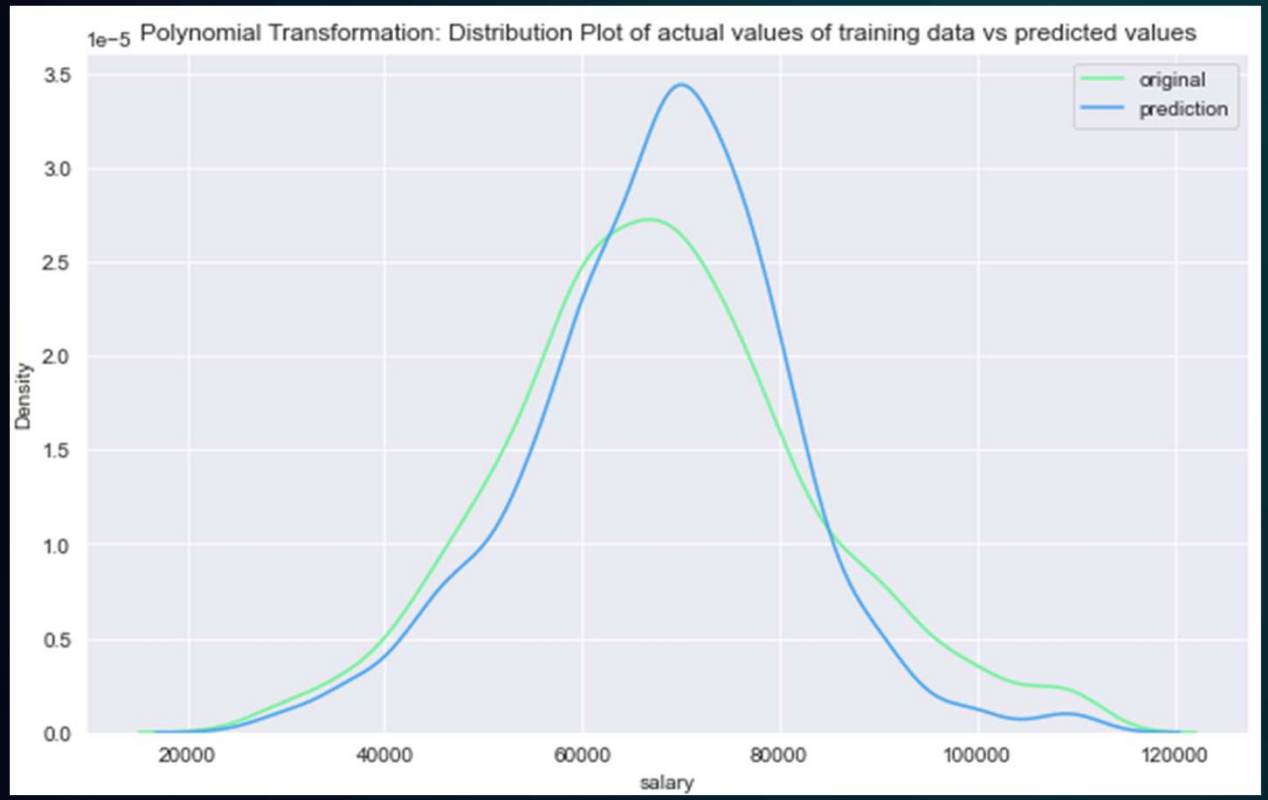



```
Model Performance
Average Error: 8936.7533 degrees.
Accuracy = 86.06%.
```

Polynomial Features

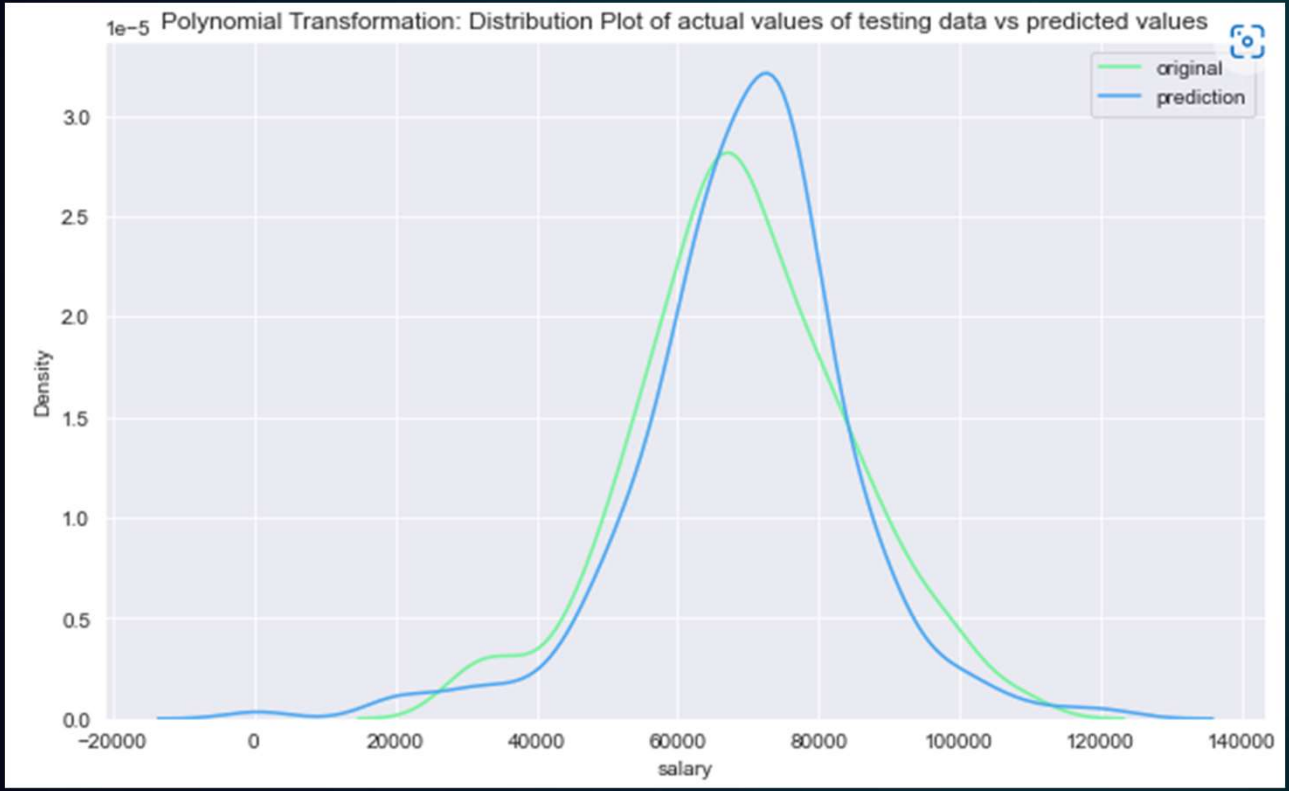


Polynomial Features



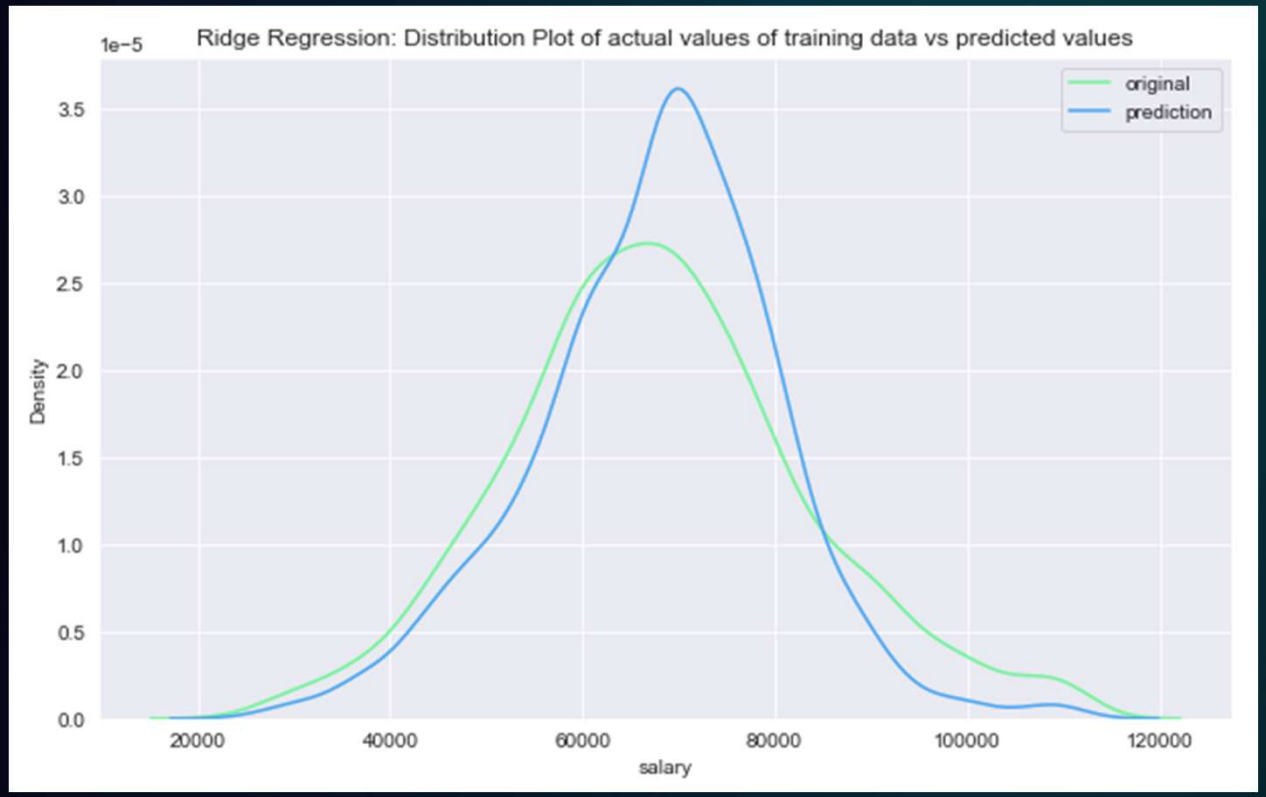

```
Model Performance
Average Error: 5777.8281 €.
Accuracy = 91.36%.
```

Polynomial Features



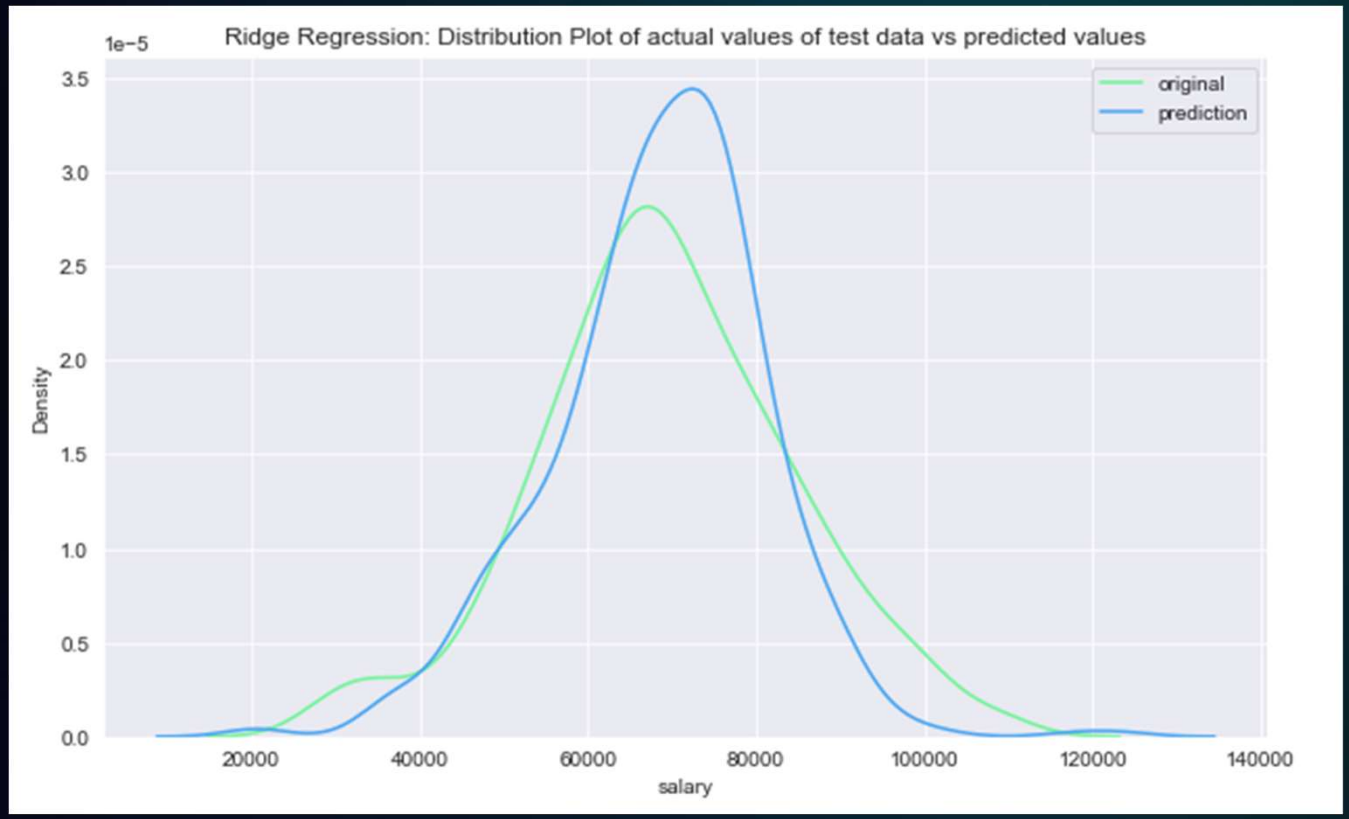

```
Model Performance
Average Error: 11351.5377 €.
Accuracy = 81.20%.
```

Ridge Regression



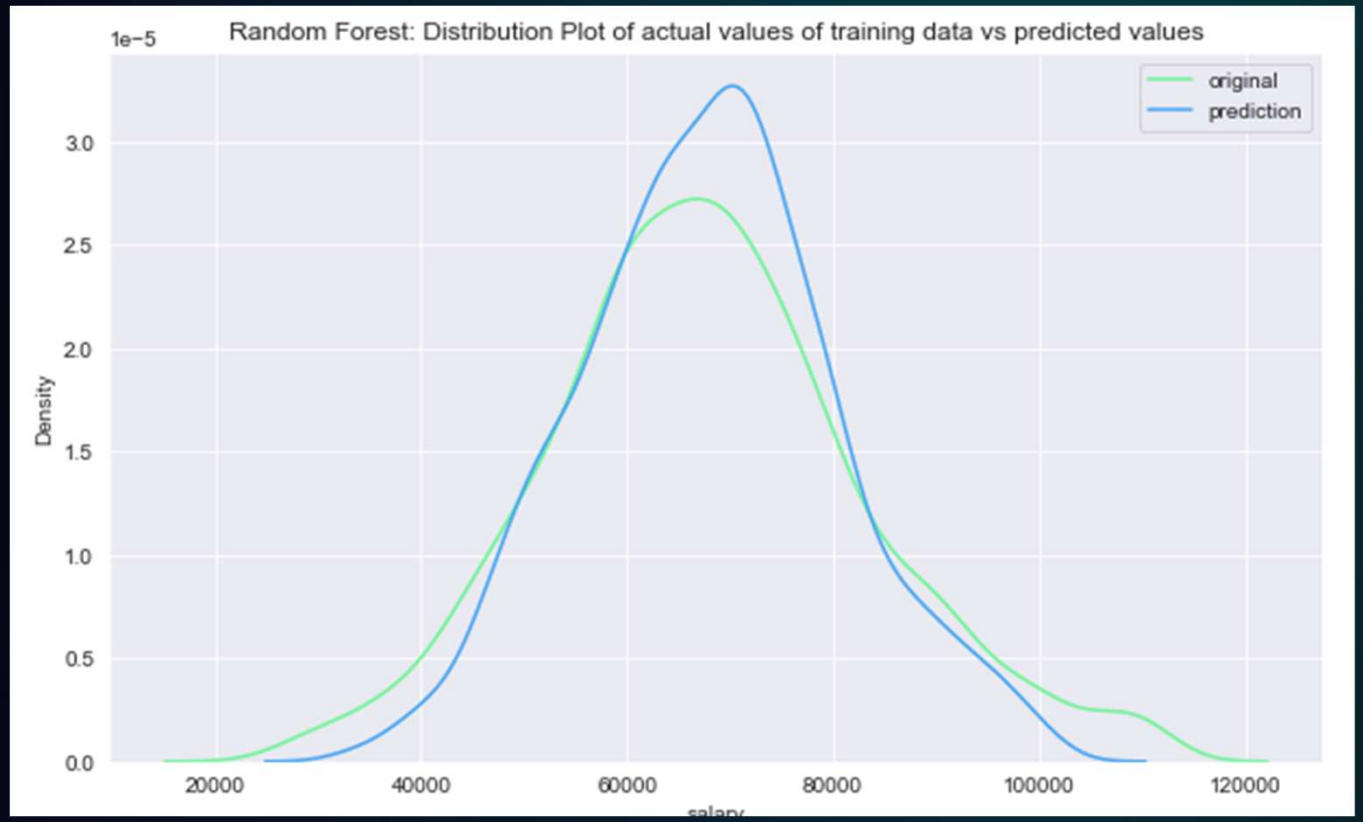

```
Model Performance
Average Error: 6251.0682 €.
Accuracy = 90.58%.
```

Ridge Regression



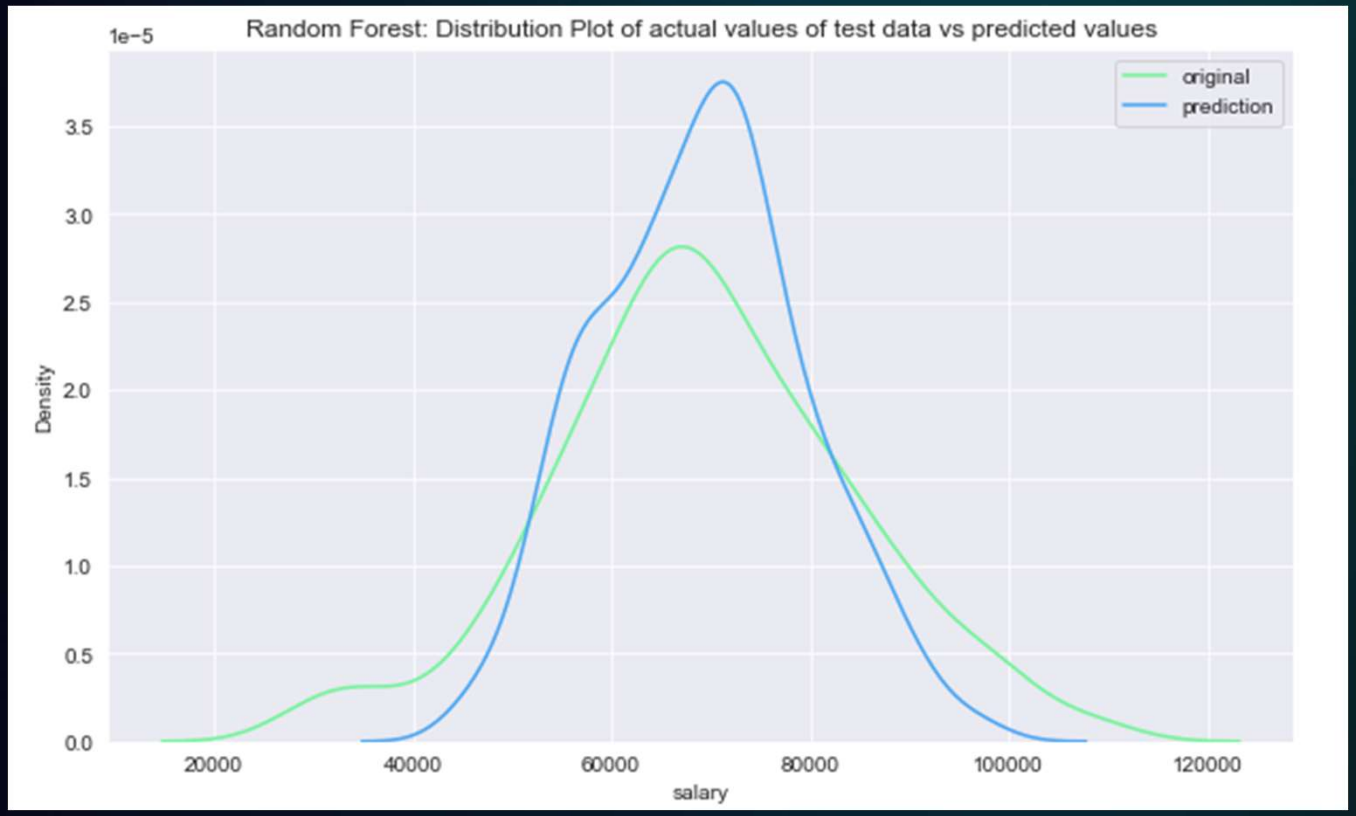

```
Model Performance
Average Error: 9163.5630 €.
Accuracy = 85.53%.
```

Random Forest




```
Model Performance
Average Error: 3732.5367 €.
Accuracy = 94.15%.
```

Random Forest




```
Model Performance
Average Error: 9024.4113 €.
Accuracy = 85.46%.
```

Hyperparameter

01

Random Search

02

Grid Search mit Ergebnissen des Random Searchs

Random Forest Hyperparameter

- Random Search

```
Model Performance
Average Error: 6121.8609 €.
Accuracy = 90.54%.
```

```
Model Performance
Average Error: 8760.5704 €.
Accuracy = 85.96%.
```

- Grid Search

```
Model Performance
Average Error: 6987.2766 €.
Accuracy = 88.94%.
```

```
Model Performance
Average Error: 8966.0671 €.
Accuracy = 85.62%.
```

Modelle im Vergleich

	Average Error	Accuracy
Lineare Regression	7753 €	86,06%
Polynomial Features	11352 €	81,2%
Ridge Regression	9164 €	85,53%
Random Forest	8760 €	85,96%

Endergebnis

Lineare Regression mit 86,06 % accuracy auf den Testdaten das beste Model

Kritische Reflektion

Zu kleiner Datensatz

Ohne ausreichende Recherche einfach Modelle ausgetestet und dann von den Ergebnissen verwundert gewesen

Nicht alle von uns am Anfang gestellten Ziele erreicht

Lessons learned

Größe des Datensatzes und Menge der Daten pro Feature extrem wichtig

→ Kann sonst Ergebnisse verfälschen

Komplexe Modelle bedeutet nicht gleich bessere Ergebnisse

Ridge Regression für AOT verstanden ☺

THANKS

Do you have any questions?