MACHINE LEARNING MODELS FOR DATA SCIENTIST SALARY PREDICTION

Data Sience - Binus University

- Renata Aqila R P - 2702244284
- Naira Faizanoor - 2702241465
- Farren Angelica D - 2702243546
- Vionita Lesia - 2702238312

INTRODUCTION

Problem Understanding

Understanding the factors that affect data scientist salaries is essential for professionals and employers alike in a job market that is changing quickly. However, accurate prediction is difficult due to the interplay of various features such as job roles, company size, and remote work, as well as the variability of salaries. The goal of this project is to develop a trustworthy salary prediction model by using advanced analytics to identify these relationships.

Objectives

- Create a predictive model that can be broadly applied.
- Reduce errors (low RMSE) and provide an explanation for the salary variation (high R^2).
- Determine which factors have the biggest effects on salary.

Data Description

Summary of Data Scientist Salary Dataset

The Data Scientist Salary dataset provides information about data scientist salaries from 2020 to 2023. It is used to analyze salary trends, work conditions, and industry dynamics in data science. This dataset offers insights into salary trends, differences by experience or location, and supports accurate salary predictions in the data science industry.

Key Variables:

- work_year
- experience_level
- employment_type
- job_title
- salary & salary_in_usd
- employee_residence
- company_location
- remote_ratio
- company_size

Applications:

- Salary prediction using machine learning models like linear regression.
- Trend analysis based on year, job type, experience, location, and remote work ratio.

→ Cleaning data by removing missing values and duplicates. Feature engineering to enhance data quality before analysis and prediction.

DATA UNDERSTANDING

- Target variable (Y) = salary_in_usd
- Ordinal columns = experience_level, employment_type, company_size

Data Understanding

During the data understanding phase, the target variable (Y) is identified as salary_in_usd. Ordinal encoding is implemented for the columns experience_level, company_size, and employment_type.

MACHINE LEARNING MODELLING

For the modeling phase, linear regression is utilized as the baseline model, while the proposed model employs the random forest regressor.

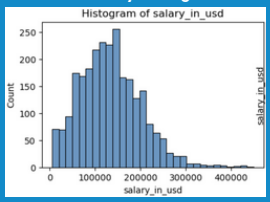
- Baseline model analysis : The linear regression model underperformed as it assumes linear relationships, making it less effective for capturing the non-linear patterns present in the data.
- Proposed model analysis : The proposed Random Forest Regressor significantly reduced prediction errors and effectively captured the variance in the data, highlighting its suitability for modeling complex, non-linear relationships.

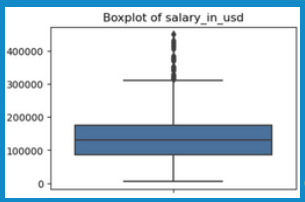
```
Training Results:
MAE: 42572.5879
MSE: 3160764674.5009
RMSE: 56220.6784
R2: 0.3035

Testing Results:
MAE: 43601.9997
MSE: 3103590386.3110
RMSE: 55709.8769
R2: 0.2905
```

```
Proposed Model Evaluation:
Training Results:
MAE: 785.5695
MSE: 22854766.2229
RMSE: 4780.6659
R2: 0.9950

Testing Results:
MAE: 1966.0524
MSE: 198054138.0149
RMSE: 14073.1709
R2: 0.9547
```

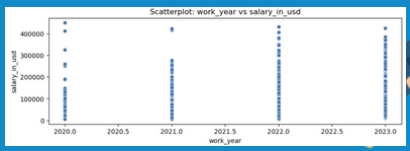
DATA PREPARATION

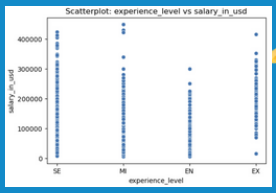
1. Check Anomaly in Target Variable





2. Check Correlation





3. Encoding

	work_year	experience_level	employment_type	salary	salary_in_usd	remote_ratio	company_size
0	2023	2	0	80000	85847	100	2
1	2023	1	2	30000	30000	100	0
2	2023	1	2	25500	25500	100	0
3	2023	2	0	175000	175000	100	1
4	2023	2	0	120000	120000	100	1

4. Splitting Data

```
Training set shape: (2067, 10)
Test set shape: (517, 10)
```

5. Scalling

```
Index(['work_year', 'experience_level', 'employment_type', 'job_title',
       'salary', 'salary_currency', 'employee_residence', 'remote_ratio',
       'company_location', 'company_size'],
      dtype='object')
set()
```

EVALUATION AND ANALYSIS

The evaluation shows that the Random Forest Regressor outperformed Linear Regression by capturing complex non-linear relationships, resulting in lower prediction errors and higher variance explanation. Effective data preparation, including cleaning and ordinal encoding, further improved performance and revealed key factors influencing data scientist salaries.

CONCLUSION

The goal of this project was to develop a trustworthy salary prediction model for data scientists by leveraging machine learning techniques.

- Experience level, employment type, and company size are critical factors impacting salaries.
- The ability to work remotely and company location also contribute to salary variation.