Data Scientist Job Salary Prediction

Saksham Raj 2005473

Introduction:

The field of data science has grown exponentially in recent years, with many companies relying on data scientists to derive insights from large amounts of data. One important aspect of any job is the salary, and predicting a data scientist's salary can be a challenging task due to several factors involved. In this project report, we have discussed how we used machine learning techniques to predict the salary of a data scientist.

Topic : Data Scientist jobs salary prediction using different ml models

Area of Topic : Data Science, ml, data analysis.

Tools to be used :

There are several tools that we have used in data science job salary prediction using ML. That are:

Programming languages: We have used programming language Python. Python languages have extensive machine learning libraries and frameworks that can be used to build models.

Machine learning libraries: There are many machine learning libraries available for Python, including scikit-learn, TensorFlow, and PyTorch. These libraries provide algorithms and tools for building various types of models, such as linear regression, decision trees, and neural networks that are needed to do our project.

Data visualization tools: We have used Tools like Matplotlib and Seaborn to visualize the data and identify patterns and trends related to salary.

Integrated development environments (IDEs): We have used IDEs Jupyter Notebook to write, test, and debug code for this project.

Problem statement : develop a model that can accurately predict the salary of a data scientist based on various factors such as work\_year, experience\_level,employment\_type,salary\_currency,

employee\_residence, company\_location and company\_size. The objective is to create a reliable and accurate salary prediction tool that can help both employers and job seekers in negotiating job offers and evaluating their compensation packages. This requires collecting and analyzing a large amount of data to identify patterns and trends that can be used to make accurate predictions. Additionally, the model should be regularly updated to reflect changes in the job market and ensure its continued accuracy over time.

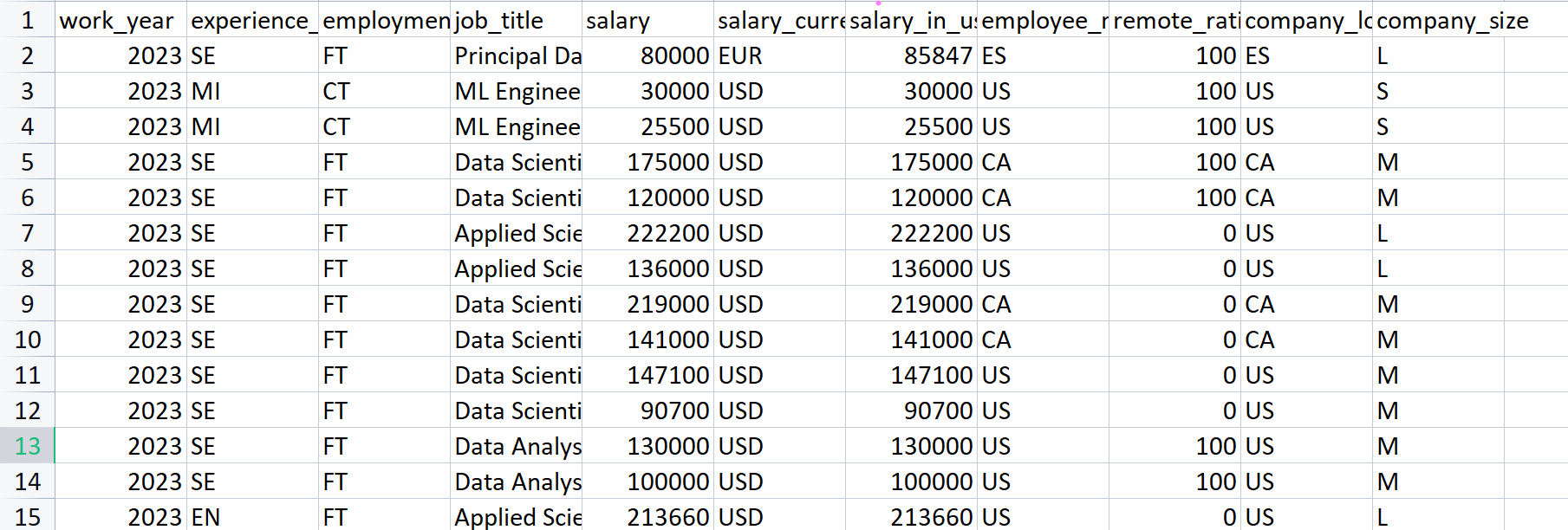
Data Collection:

We gathered data on data scientist salaries from kaggle, we searched for relevant data on various sources such as Glassdoor, Indeed, and Payscale. The dataset contains information such as work\_year, experience\_level, employment\_type, salary\_currency,employee\_residence, company\_location and company\_size.

Data set link is :

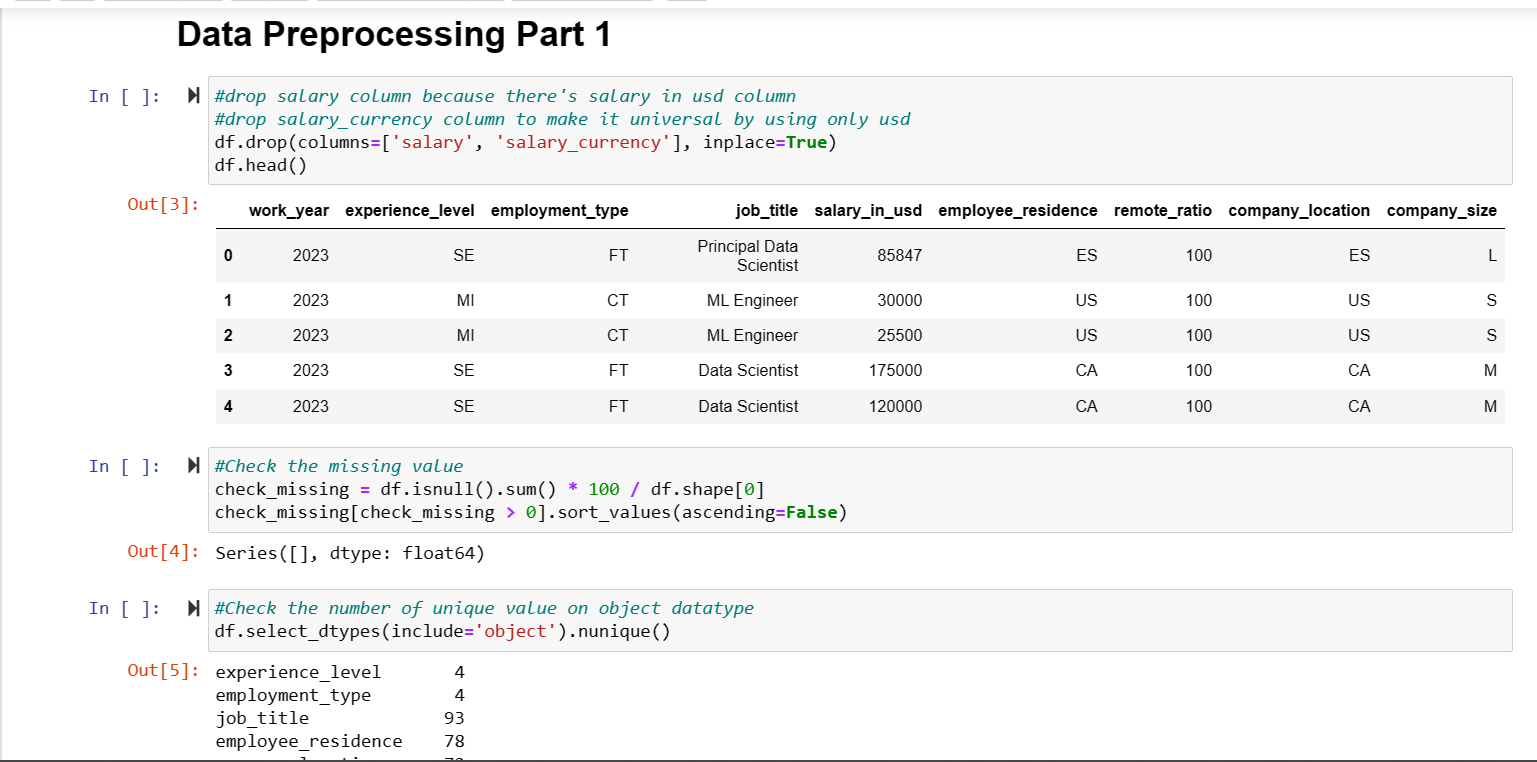
<https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023>

Dataset content is :



Data Cleaning:

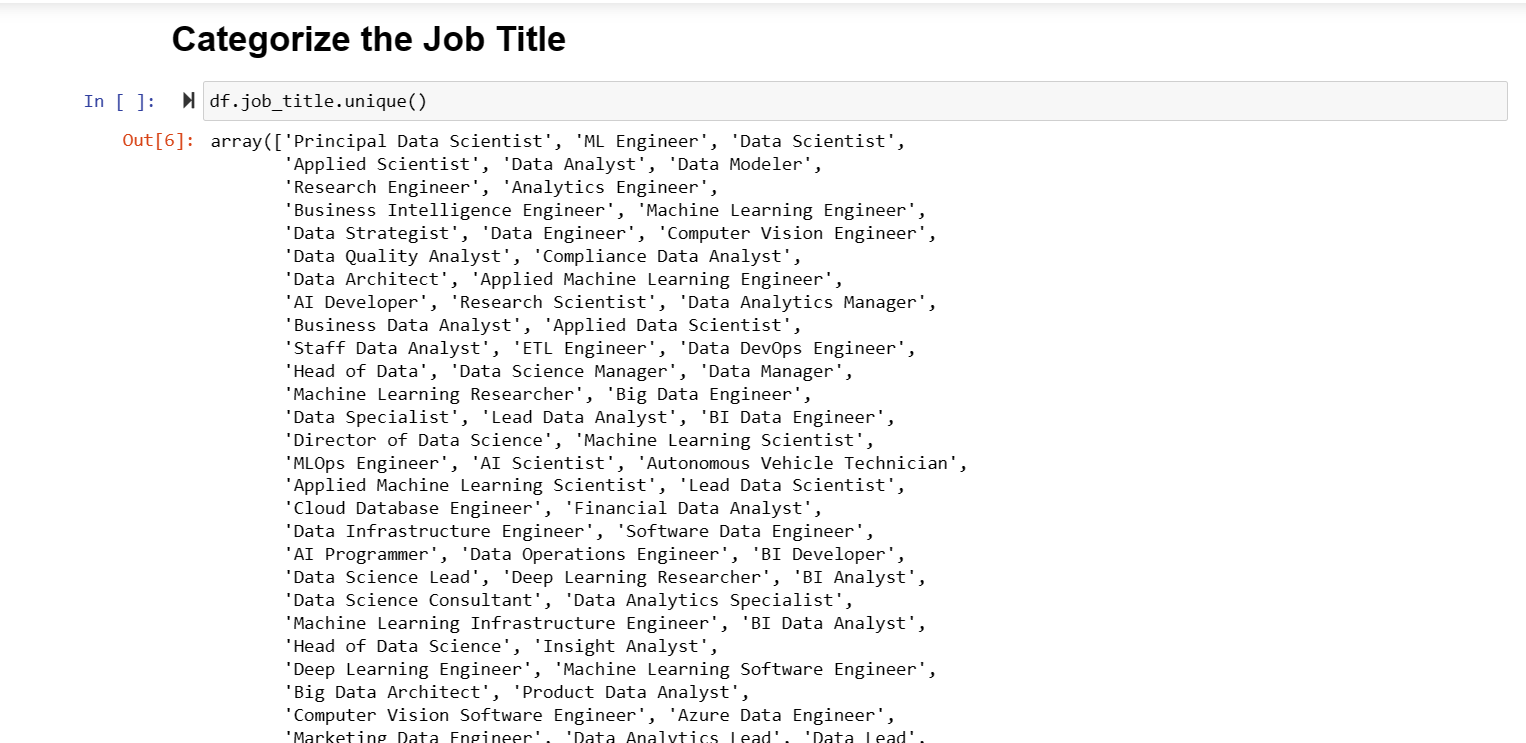
As with any data analysis project, cleaning the data was an essential step. We removed any duplicate or irrelevant entries and handled any missing values.



We also converted categorical variables into numerical values. We Categorize different job titles under data\_scientist, machine\_learning, data\_analyst, data\_engineer, bi\_analytics, other.

We Categorize different employee\_residence into different continents.

We Categorize different company\_locations into different continents.





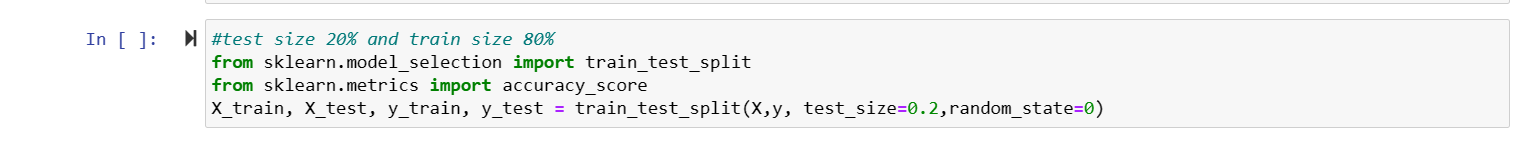




We then drawn pairwise plot and correlation matrix to find correlation between 2 features and if the 2 features are highly correlated then we dropped one of them.

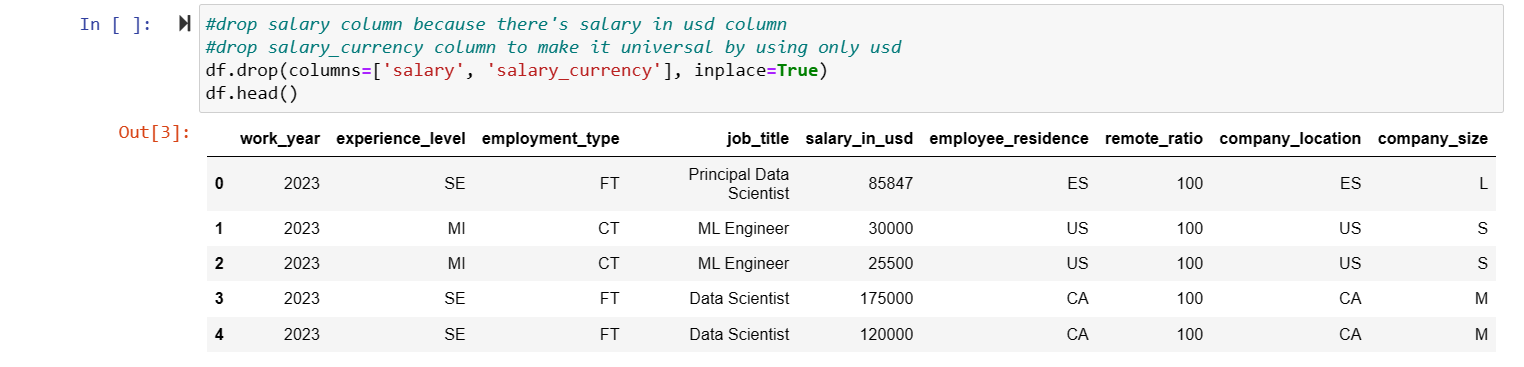


we divided our datasets into training and testing data into 80, 20 ratio



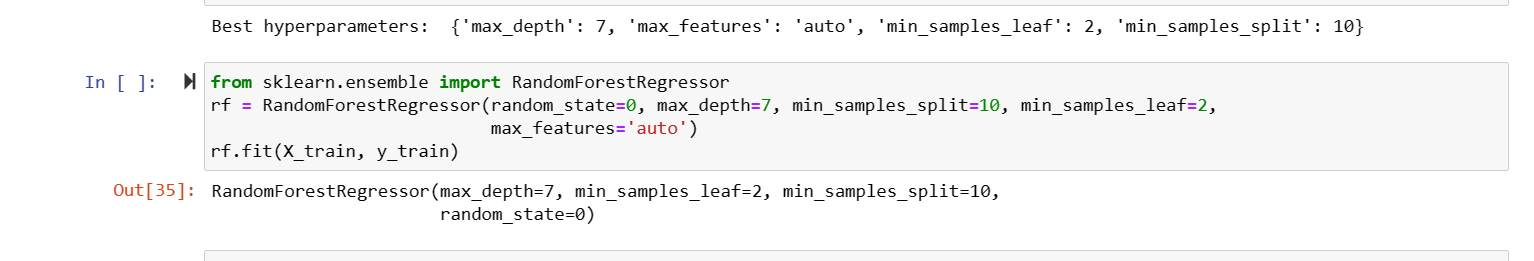
Feature Engineering:

To improve the accuracy of our model, we created new features based on the existing data. We also droped some features like we dropped salary\_currency to make it universal by using only usd



Model Selection:

We experimented with various machine learning models such as linear regression, decision trees, random forests, and gradient boosting. After evaluating their performance, we decided to use the decision tree and Random Forest algorithm as it provided the best results.





Model optimazitation:

After developing initial models, we focused on optimizing them further to achieve

higher accuracy and reduce overfitting.

I employed regularization techniques to prevent overfitting and improve

generalization performance.

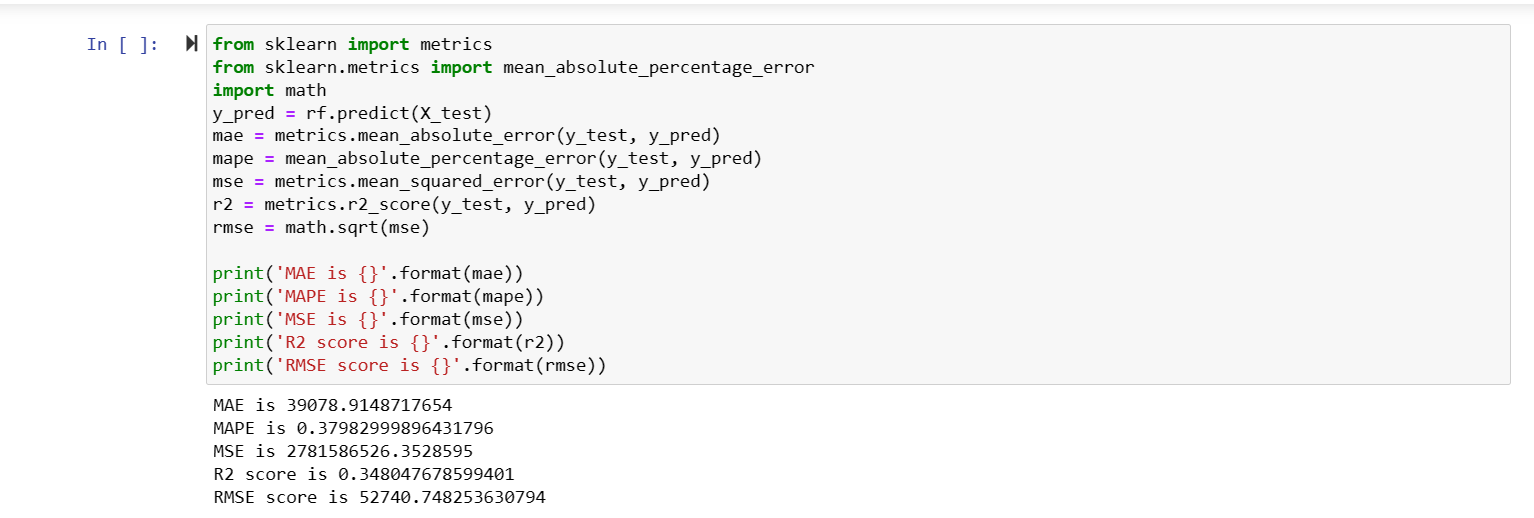
we have used grid search to find optimal values.





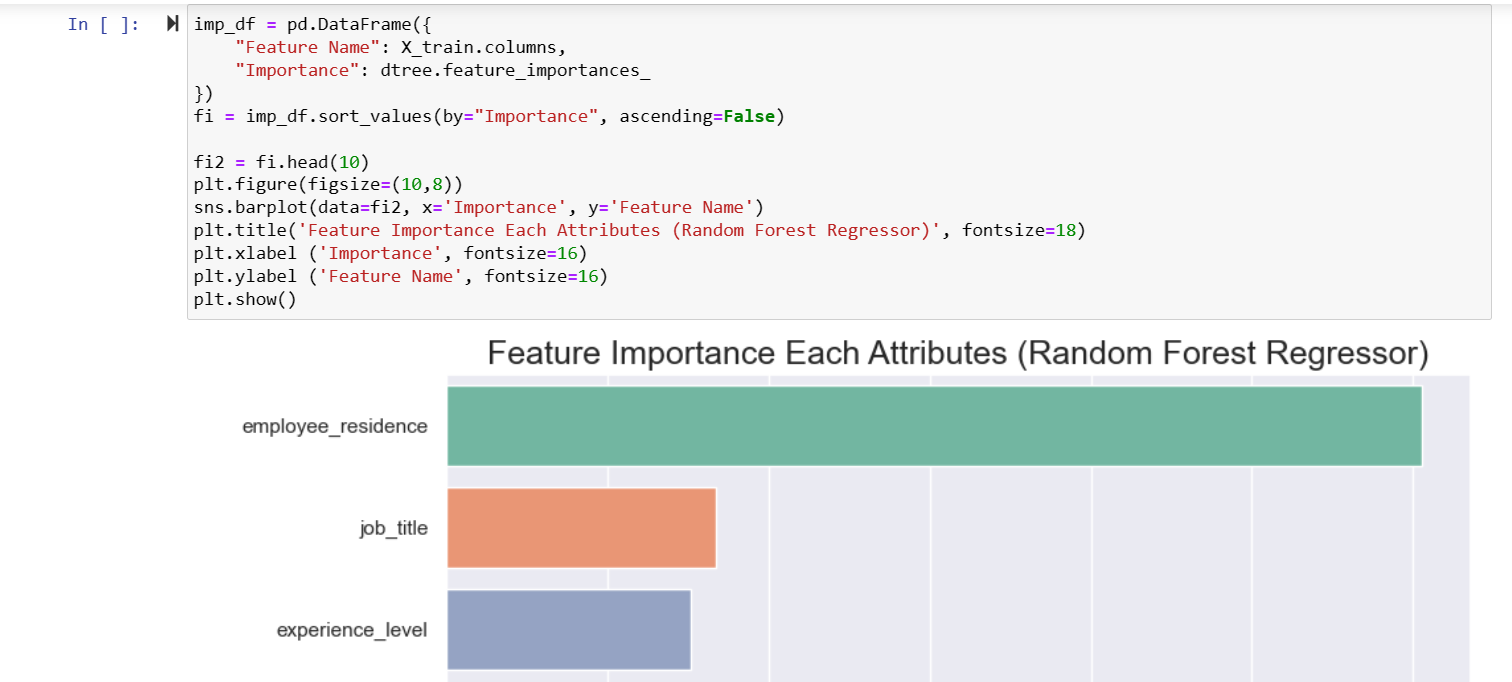
Model Evaluation:

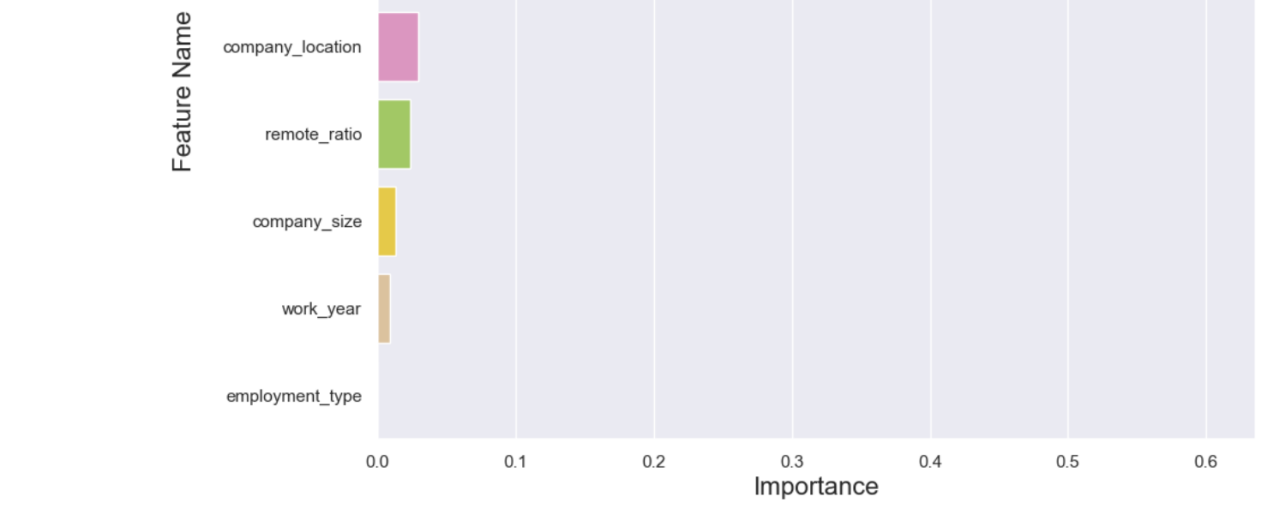
We split the data into training and testing sets and evaluated the model's performance using metrics such as mean absolute error (MAE), mean squared error (MSE), MAPE, R2 and root mean squared error (RMSE). Our model achieved an RMSE of $53764, which indicates that the predicted salaries were within $53763 of the actual salaries.



Visualization :

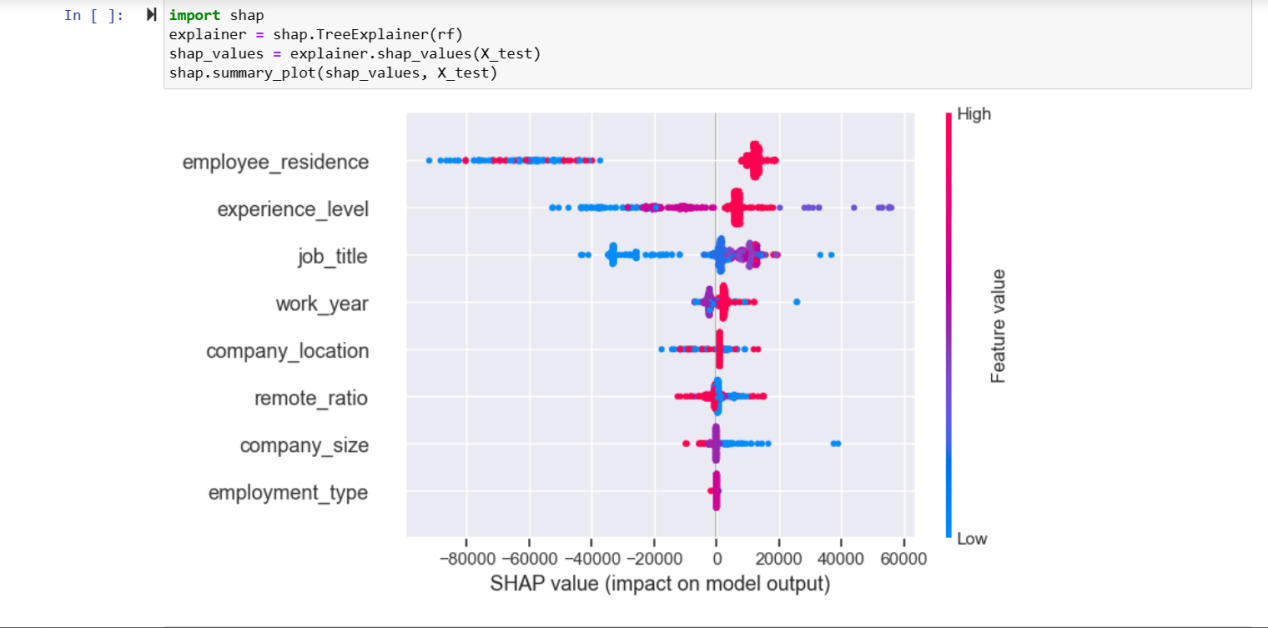
We have drawn feature importances curve to determine which features are important in deciding the salary/ with changing which feature salary will change most





Conclusion:

In conclusion, we successfully built a machine learning model that can predict the salary of a data scientist based on various factors such as location, education level, years of experience, company size, and industry. The model achieved good accuracy, indicating that it can be useful for companies or individuals to estimate the salary range of a data scientist. However, it is essential to note that the model's predictions are based on historical data and may not accurately reflect current market conditions or future trends.





Objectives achieved in Milestone 1: Data collection, Data Preprocessing.

Objectives achieved in Milestone 2: Data vizualization, model Training, model optimization, Testing, Result Analysis, Result vizualization.