Data Scientist Salary Prediction

A Project Report

submitted in partial fulfillment of the requirements

of

Industrial Artificial Intelligence

with Cloud Computing

by

Aryan Thengdi, 20010456

Ashish Amirneni, 20010247

Divyam Padole, 20010331

Piyush Kawale, 20010586

Tushar Charde, 20010903

Under the Esteemed Guidance of

Mr. Akshay Chaskar

ACKNOWLEDGEMENT

We would like to take this opportunity to express our profound gratitude to all individuals who have contributed directly or indirectly to the completion of this thesis.

Foremost, our deepest appreciation goes to our supervisor, Mr. Akshay Chaskar, whose guidance and mentorship have been invaluable throughout this journey. Mr. Chaskar's expertise, encouragement, and constructive criticism have been instrumental in shaping this dissertation and pushing us towards excellence. His unwavering confidence in our abilities has served as a constant source of motivation and inspiration. Working under his mentorship for the past months has been an enriching experience, and we are grateful for his consistent support not only during the thesis but also in various academic pursuits. His mentorship extends beyond the confines of academic work, instilling in us qualities that transcend the boundaries of the classroom and contribute to our growth as responsible professionals. Also, we extend our thanks to Tech Saksham for their support and resources, which have facilitated the research process and enabled us to achieve our goals. In conclusion, we are deeply grateful to all those who have played a part, big or small, in bringing this thesis to fruition. Your contributions have not gone unnoticed, and we are indebted to you for your support and guidance.

TABLE OF CONTENTS

GitHub Link: https://github.com/Divyam-Padole/Data-Scientist-Salary-Prediction

ABSTRACT

This project presents a comprehensive analysis of data scientist salaries, utilizing a dataset containing pertinent information such as experience level, employment type, company size, and geographical location. Initial data preprocessing involves handling missing values and encoding categorical variables for subsequent analysis. Exploratory data visualizations uncover insights into salary distributions across various experience levels, employment types, and company locations. Additionally, the study investigates salary trends over time, both in the United States and India, highlighting median salary fluctuations and their relationship with factors like company size and remote work arrangements.

Furthermore, the project employs predictive modeling using a Random Forest Regressor to predict salaries based on specified parameters such as year, experience level, employment type, company size, and remote work ratio. Evaluation metrics including Mean Absolute Error, Mean Squared Error, and R2 Score are employed to assess the model's predictive performance. The findings from this analysis provide valuable insights for data scientists and employers, facilitating informed decision-making regarding salary negotiations, job opportunities, and strategic workforce planning. Moreover, the predictive model serves as a practical tool for estimating salaries, aiding in budgeting and resource allocation within organizations. Overall, this project contributes to a deeper understanding of data scientist salaries and their underlying dynamics, empowering stakeholders to navigate the landscape of data science employment effectively.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. Problem Statement:

To create a prediction model for the data scientist salaries using random forest model.

1.2. Problem Definition:

The primary objective of this project is to explore and understand the factors influencing data scientist salaries and their variations across different parameters. This involves conducting exploratory data analysis, visualizing salary distributions, investigating trends over time, and building predictive models to forecast salaries based on relevant factors. By addressing these aspects, the project aims to provide actionable insights into data scientist compensation dynamics, aiding stakeholders in navigating the complexities of the data science job market.

1.3. Expected Outcomes:

The expected outcomes of this project include:

- Insights into the distribution of data scientist salaries across various experience levels, employment types, and company sizes.
- Identification of salary trends over time, highlighting factors influencing salary fluctuations.
- Development of predictive models to estimate salaries based on specified parameters.
- Evaluation of model performance using relevant metrics such as Mean Absolute Error, Mean Squared Error, and R2 Score.
- Provision of actionable insights for data scientists and employers to facilitate salary negotiations, job searches, and strategic workforce planning.
- Enhancement of understanding regarding data scientist compensation dynamics, empowering stakeholders to make informed decisions in the data science employment landscape.

CHAPTER 2

LITERATURE SURVEY

2.1. Paper-1

Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits – A Literature Review by Tee Zhen Quan Mafas Raheem,School of Computing Asia Pacific University of Technology and innovation (APU) Kuala Lumpur, Malaysia

2.1.1. Brief Introduction of Paper:

The paper aims to review recent methodologies for developing a salary prediction model tailored to the Data Science field. It emphasizes the importance of specialized skills and job benefits as input variables for accurate salary predictions. The study also identifies existing human resource challenges in the data science field and determines the most demanded skill sets. The experimental dataset incorporates skill-based and job benefits factors to enhance prediction accuracy.

2.1.2. Techniques used in Paper:

The paper categorizes benchmarking machine learning methodologies into three main categories:

1. Statistical Methods: These methods are effective in presenting variable relationships, especially with parameter tuning potential if linearity is present. Techniques like linear regression and polynomial regression are explored for salary prediction.

2. Ensemble Machine Learning Methods: Ensemble methods, such as Random Forest, combine multiple classifiers for stable and accurate predictions. They excel in handling non-linear relationships in the data.

3. Deep Learning Neural Networks: Deep learning techniques, including neural networks, have a strong specialty in handling unlabeled data and framework modifications. Models like Bidirectional Gated Recurrent Unit (GRU), Conventional Neural Network (CNN), and Graph Convolutional Network (GCN) are employed for salary prediction.

Each category of techniques offers unique strengths under different scenarios and requirements, contributing to the development of a comprehensive salary prediction model for the data science field.

2.2. Paper-2

Study of Employment Salary Forecast using KNN Algorithm by Junyu Zhang and Jinyong Cheng*School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences) Jinan, China

2.2.1 Brief Introduction of Paper:

This paper delves into the realm of employment salary forecasting, focusing on the application of the KNN algorithm. By examining the discrepancies between college students' abilities and their salary expectations, the study aims to provide insights into predicting employment salaries accurately. The introduction outlines the motivation behind the research, emphasizing the need to address the imbalance between student capabilities and salary aspirations in the job market.

2.2.2 Techniques used in Paper:

The primary technique employed in this paper is the KNN (K-Nearest Neighbor) algorithm, a non-parametric supervised learning method. KNN operates by identifying the k nearest data points to a given input, classifying the input based on the most common class among its neighbors. The paper utilizes KNN to build a predictive model for employment salary levels, leveraging seven key factors that influence salaries for Java back-end engineers. Additionally, cross-validation is employed to determine the optimal value of k, ensuring the accuracy of the predictive model.

2.3. Paper-3

Salary Prediction Analysis for the ‘Slow Employment’ Phenomenon - Based on Random Forest Algorithm by Liang Ye1, SiYi Fu2, GuangYuan Chen2, and Jin Lu

2.3.1 Brief Introduction of Paper:

This paper addresses the contemporary issue of "slow employment" among college graduates, focusing on enhancing students' employability through salary prediction analysis. It investigates the causes of slow employment and proposes countermeasures to mitigate its effects. By utilizing the random forest algorithm, the study aims to develop a predictive model to assist both companies and job seekers in understanding market value and facilitating smoother transitions for graduating students into the workforce.

2.3.2 Techniques used in Paper:

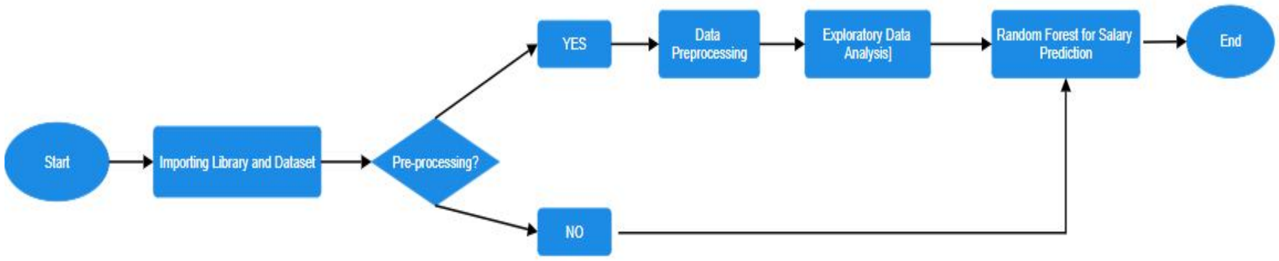
The primary technique employed in this paper is the KNN (K-Nearest Neighbor) algorithm, a non-parametric supervised learning method. KNN operates by identifying the k nearest data points to a given input, classifying the input based on the most common class among its neighbors. The paper utilizes KNN to build a predictive model for employment salary levels, leveraging seven key factors that influence salaries for Java back-end engineers. Additionally, cross-validation is employed to determine the optimal value of k, ensuring the accuracy of the predictive model.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Data Flow Diagram

A Data Flow Diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).


Start
Importing Library and Dataset
Pre-processing?
YES
NO
Data Preprocessing
Exploratory Data Analysis]
Random Forest for Salary Prediction
End

3.3.1. DFD Level 0

Fig 1. Flow Diagram

3.2 Methodology

Start: This marks the initiation of our data science pipeline, where we commence the process of analyzing salaries.

Importing Library and Dataset: At this stage, we import essential libraries and load the dataset containing information about salaries. This step ensures that we have access to the necessary tools and data for our analysis.

Data Preprocessing: In this crucial step, we preprocess the dataset to prepare it for analysis. Tasks such as handling missing values, encoding categorical variables, and scaling numerical features are performed here. The objective is to clean and transform the data to make it suitable for further analysis and modeling.

Pre-processing?: This serves as a decision point where we assess whether additional preprocessing steps are required based on the initial analysis of the dataset. It allows us to iteratively refine our preprocessing approach to ensure data quality and integrity.

Exploratory Data Analysis (EDA): EDA involves analyzing and visualizing the dataset to gain insights and identify patterns. Through various statistical techniques and visualization methods, we explore relationships between variables and uncover potential trends or anomalies. EDA helps us understand the underlying structure of the data and inform subsequent modeling decisions.

Random Forest for Salary Prediction: With a thorough understanding of the dataset achieved through EDA, we proceed to build a predictive model using the Random Forest algorithm. This machine learning model leverages decision trees to predict salaries based on features such as experience level, employment type, company size, etc. By training the model on the dataset, we aim to accurately forecast salaries and uncover factors influencing compensation.

End: This signifies the culmination of our data science pipeline, where we have successfully developed a predictive model for salary prediction. The model undergoes evaluation to assess its performance and validity, concluding our analysis process.

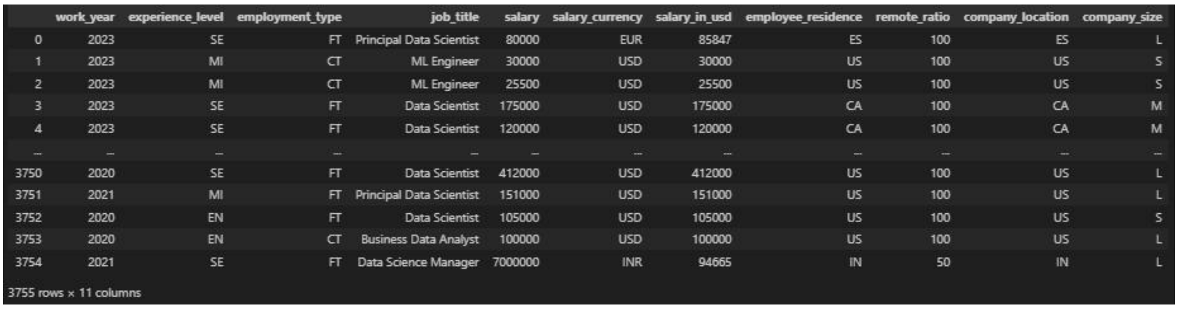
By following this systematic approach, we establish a structured framework for analyzing salaries and building predictive models, enabling informed decision-making and actionable insights.

CHAPTER 4

IMPLEMENTATION and RESULT

4.1 Dataset Overview

The table below provides an overview of the dataset used in this analysis. It contains information about various factors such as work year, experience level, employment type, company size, remote work ratio, and salary in USD. Each row represents a data entry, capturing the details of individuals in the data science field. This dataset serves as the foundation for our analysis and predictive modeling, enabling us to explore the relationships between different variables and predict salaries effectively.

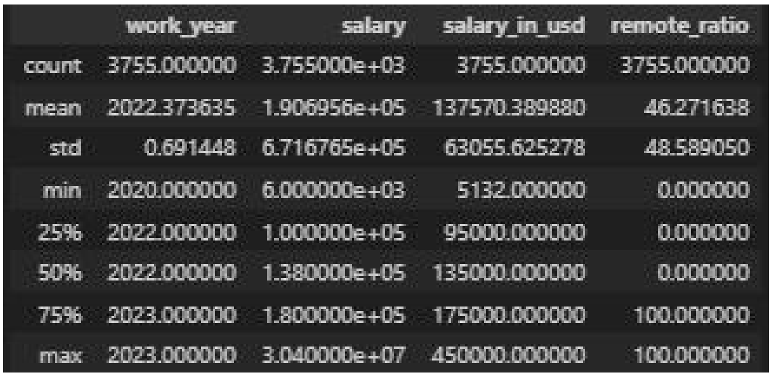


	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES	L
1	2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US	S
2	2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US	S
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA	M
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA	M
...	...	...	...	...	...	...	...	...	...	...	...
3750	2020	SE	FT	Data Scientist	412000	USD	412000	US	100	US	L
3751	2021	MI	FT	Principal Data Scientist	151000	USD	151000	US	100	US	L
3752	2020	EN	FT	Data Scientist	105000	USD	105000	US	100	US	S
3753	2020	EN	CT	Business Data Analyst	100000	USD	100000	US	100	US	L
3754	2021	SE	FT	Data Science Manager	7000000	INR	94665	IN	50	IN	L

3755 rows × 11 columns

Table 1. Table Overview

The descriptive statistics table summarizes key numerical variables in the dataset, including work year, salary, salary in USD, and remote work ratio. It provides insights into the central tendency, variability, and distribution of these variables. Notably, the mean salary in USD is approximately $137,570, with a standard deviation of around $63,055. The remote work ratio ranges from 0 to 100%, with a mean of approximately 46.3%. These statistics offer a concise overview of the dataset's numerical features, aiding in understanding its characteristics.



	work_year	salary	salary_in_usd	remote_ratio
count	3755.000000	3.755000e+03	3755.000000	3755.000000
mean	2022.373635	1.906956e+05	137570.389880	46.271638
std	0.691448	6.716765e+05	63055.625278	48.589050
min	2020.000000	6.000000e+03	5132.000000	0.000000
25%	2022.000000	1.000000e+05	95000.000000	0.000000
50%	2022.000000	1.380000e+05	135000.000000	0.000000
75%	2023.000000	1.800000e+05	175000.000000	100.000000
max	2023.000000	3.040000e+07	450000.000000	100.000000

Table 2. Numerical Data Description

4.2 Results of EDA

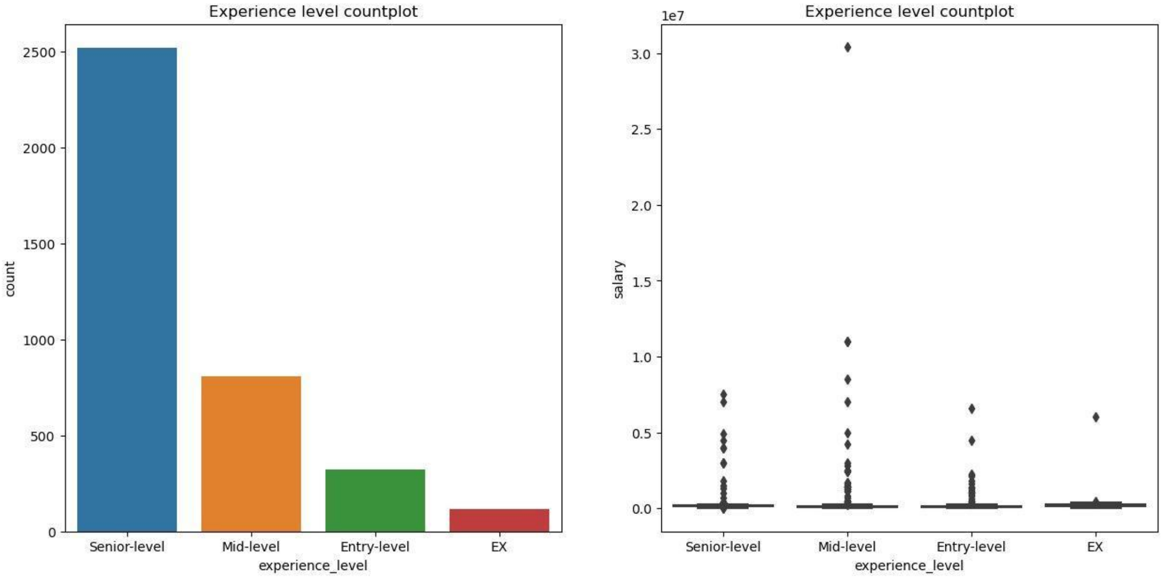


Fig 2. Experience Level Count Plot

The first subplot shows a countplot of experience levels, indicating the distribution of data scientists across various experience levels.

The second subplot presents a boxplot comparing salary distributions for different experience levels, allowing for an analysis of salary trends and variations based on experience level.

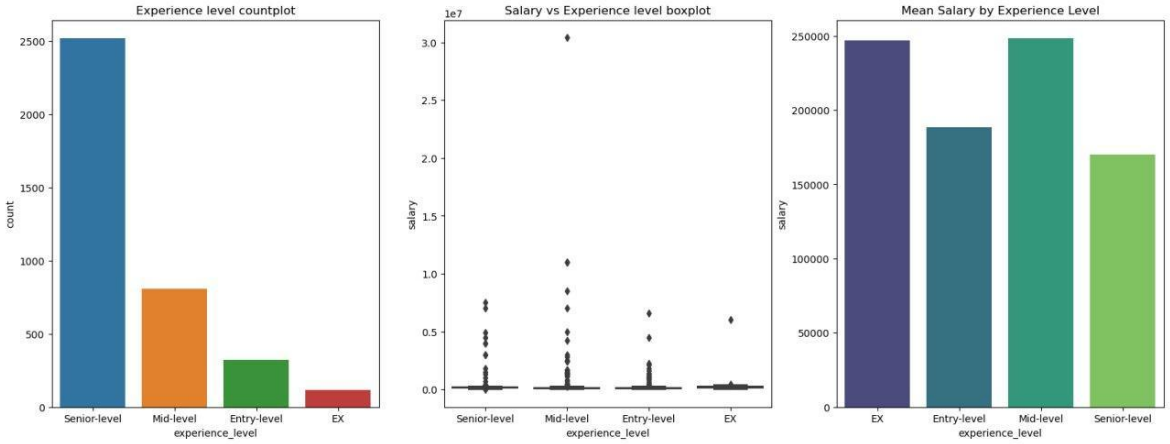


Fig 3. Mean Salary by Experience Level

The third subplot presents a barplot showing the mean salary for each experience level. This visualization offers insights into the average salary trends across various experience levels within the dataset.

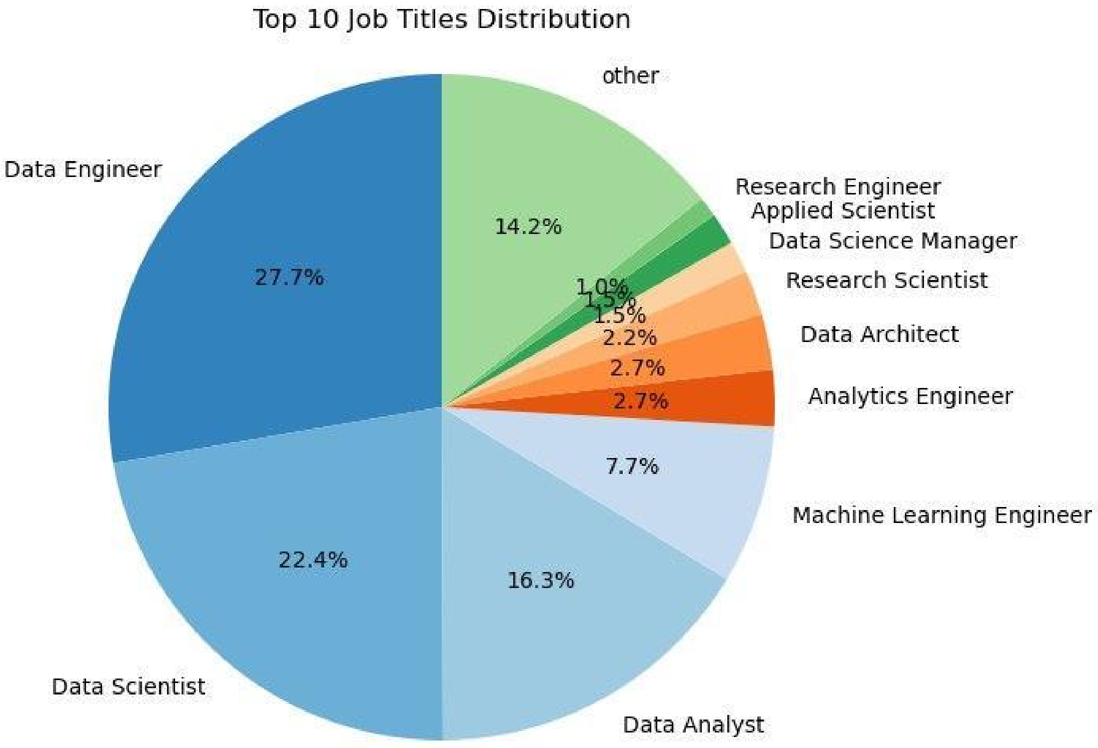


Fig 4. Top 10 Job Titles Distribution

The visualization showcases the distribution of job titles among data scientists. It focuses on the top 10 most common job titles through a pie chart, with each slice representing a job title's proportion. The "other" category groups less common titles. Percentage labels offer insight into the relative frequency of each title, while a color scheme enhances clarity. This concise representation provides an overview of prevalent job titles in the dataset.

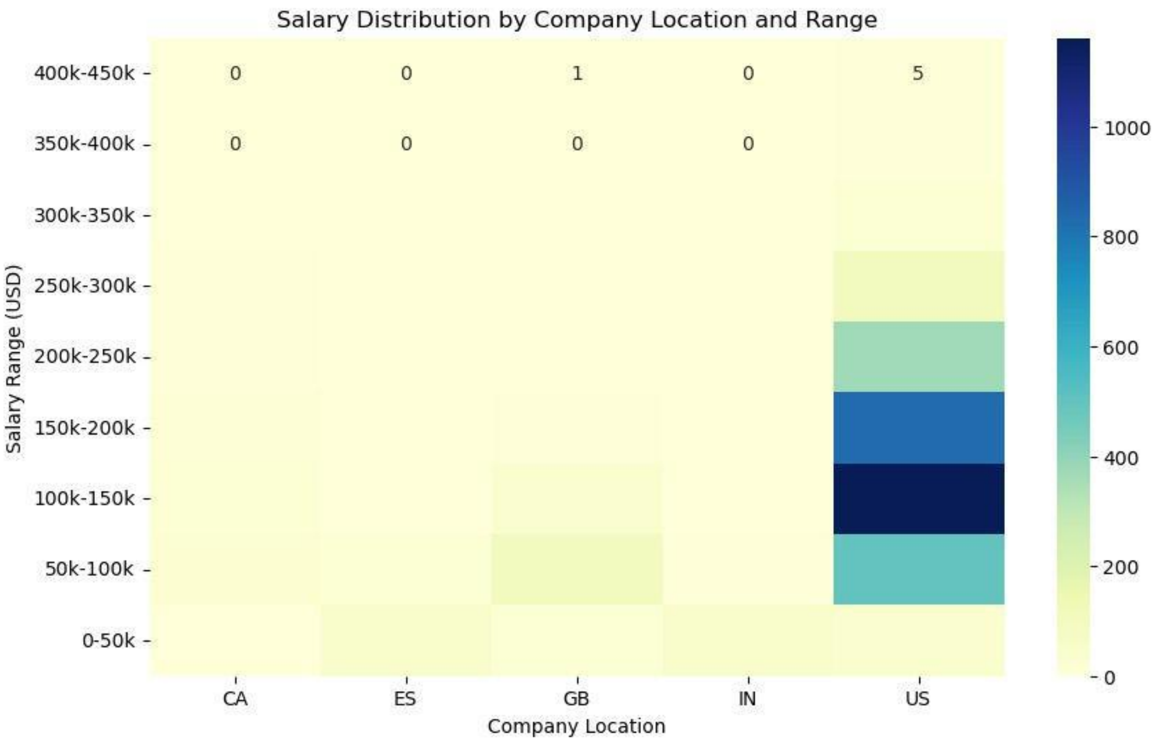


Fig 5. Salary Distribution by Company Location and Range

This heatmap visualizes salary distributions across different company locations (US, GB, CA, ES, IN) categorized into salary ranges. Darker colors indicate higher frequency of salaries falling within each range, offering a succinct overview of salary distributions by location and range.

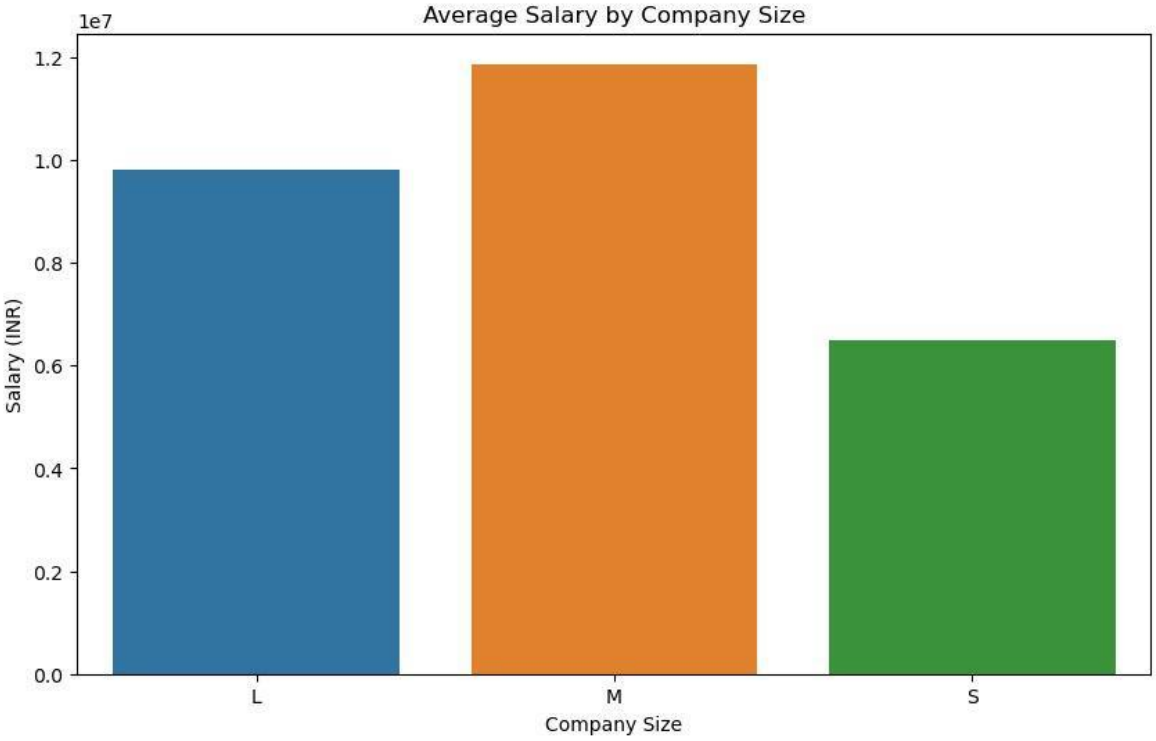


Fig 6. Average Salary by Company Size

This bar plot illustrates the average salary (in INR) categorized by company size. It provides a clear comparison of salary levels across different company sizes, aiding in understanding the salary dynamics based on the scale of the employing organizations.

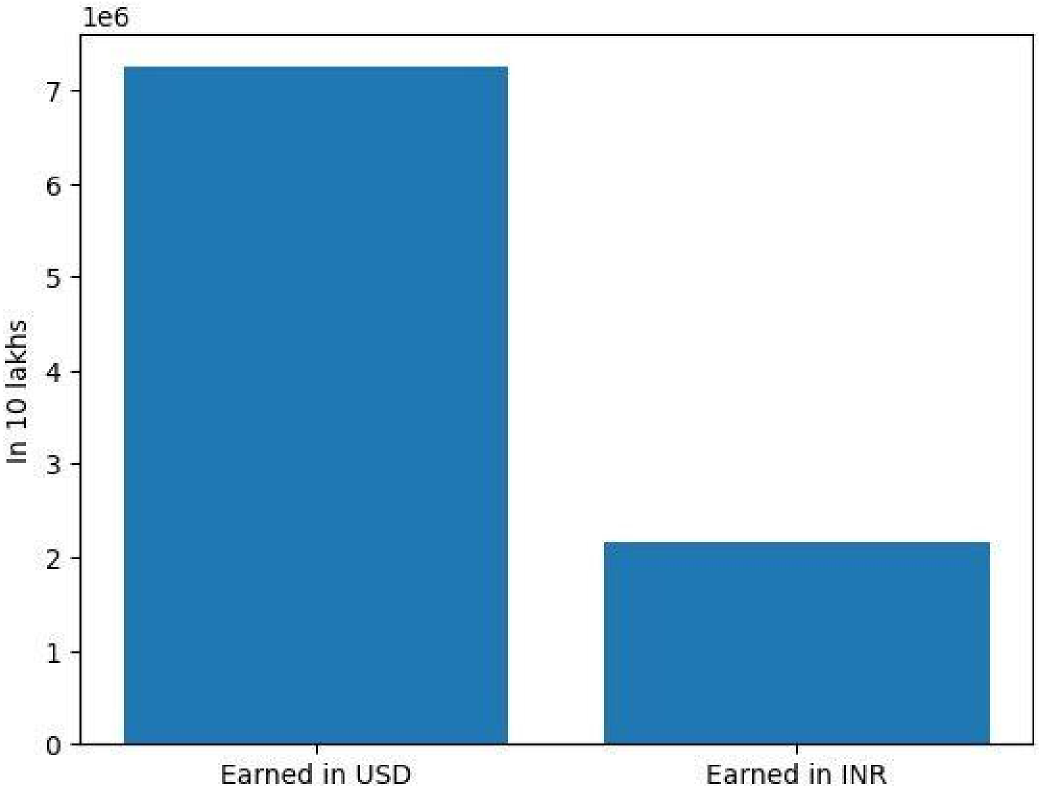


Fig 7. Salary Comparison USD vs INR

This bar plot compares the average salary earned by individuals in India, categorized by the currency in which they earn their salary. It provides insights into the average earnings of individuals in India, distinguishing between salaries earned in USD and INR, facilitating a comparison of earnings based on currency denomination.

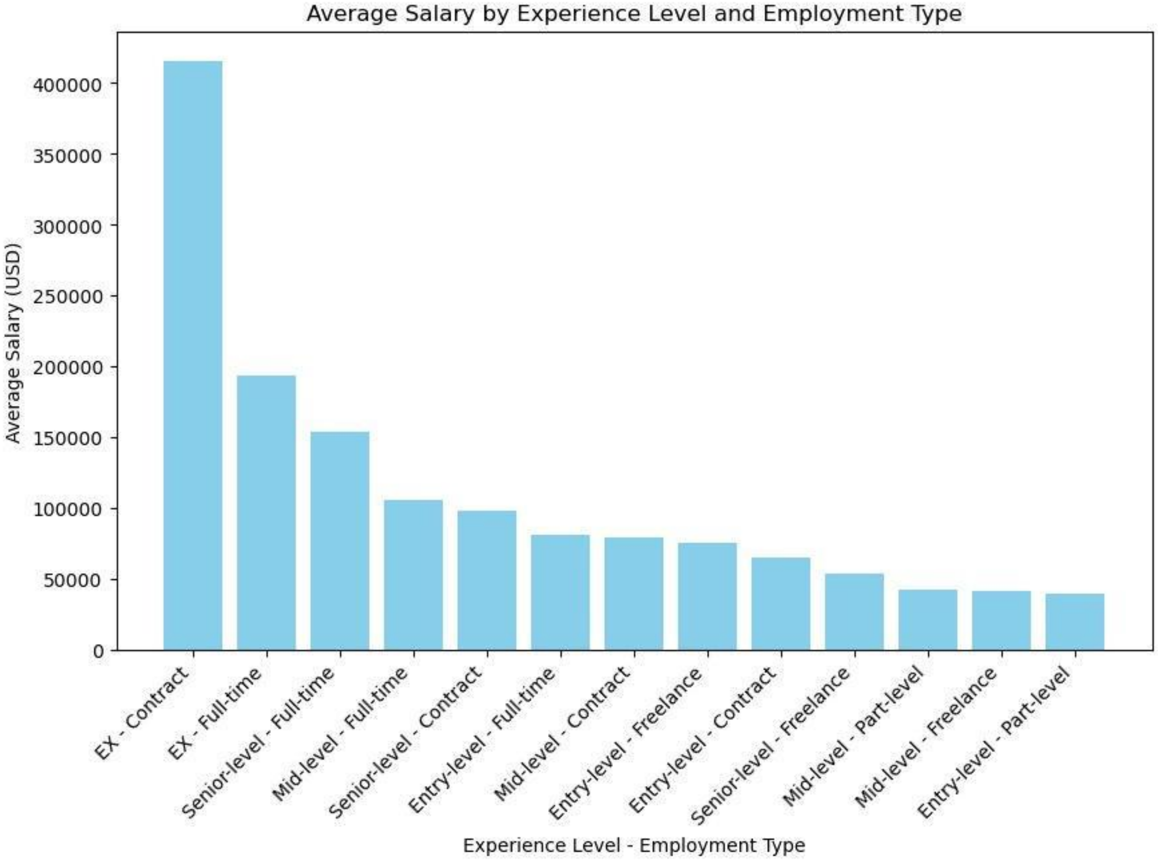


Fig 8. Average Salary by Experience Level and Employment Type

This bar plot illustrates the average salary based on different combinations of experience level and employment type. It provides a visual representation of how average salaries vary across various experience levels and types of employment.

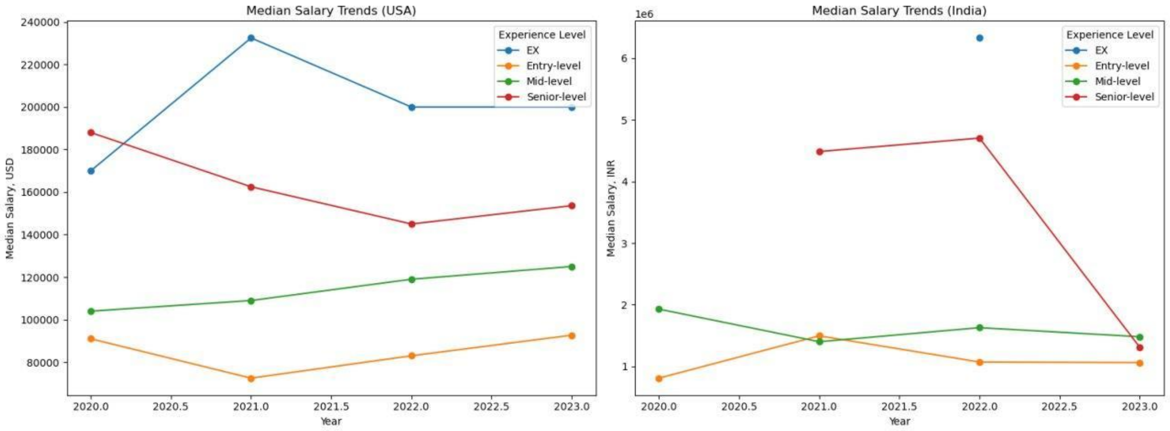


Fig 9. Comparison of Salary Trends in USA and India

This visualization depicts the median salary trends over the years for different experience levels in the USA and India. The left subplot illustrates the trends for the USA, while the right subplot displays the trends for India. Each line represents a specific experience level, showing how median salaries have evolved over time in both countries.

4.3 Result Of Model Analysis

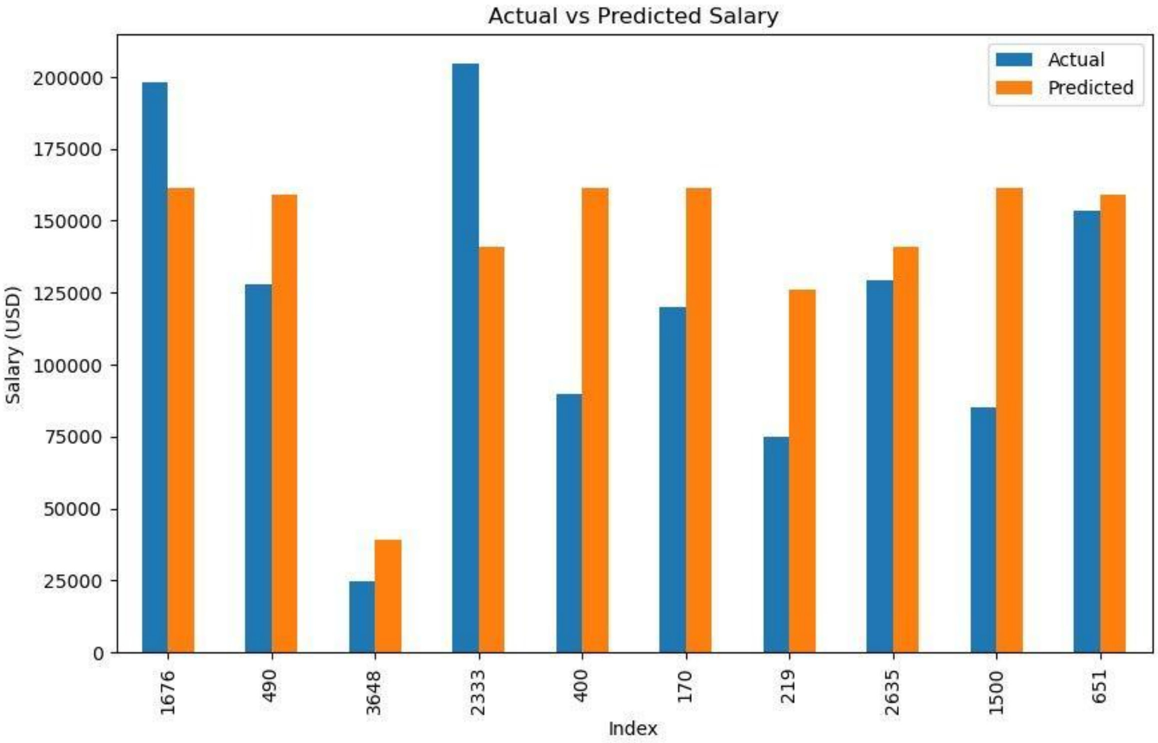


Fig 10. Actual vs Predicted Salary using Random Forest

Actual vs Predicted Salary: It visualizes the comparison between actual and predicted salaries for the first 10 data points. The plot is a bar chart with the index on the x-axis and salary in USD on the y-axis.

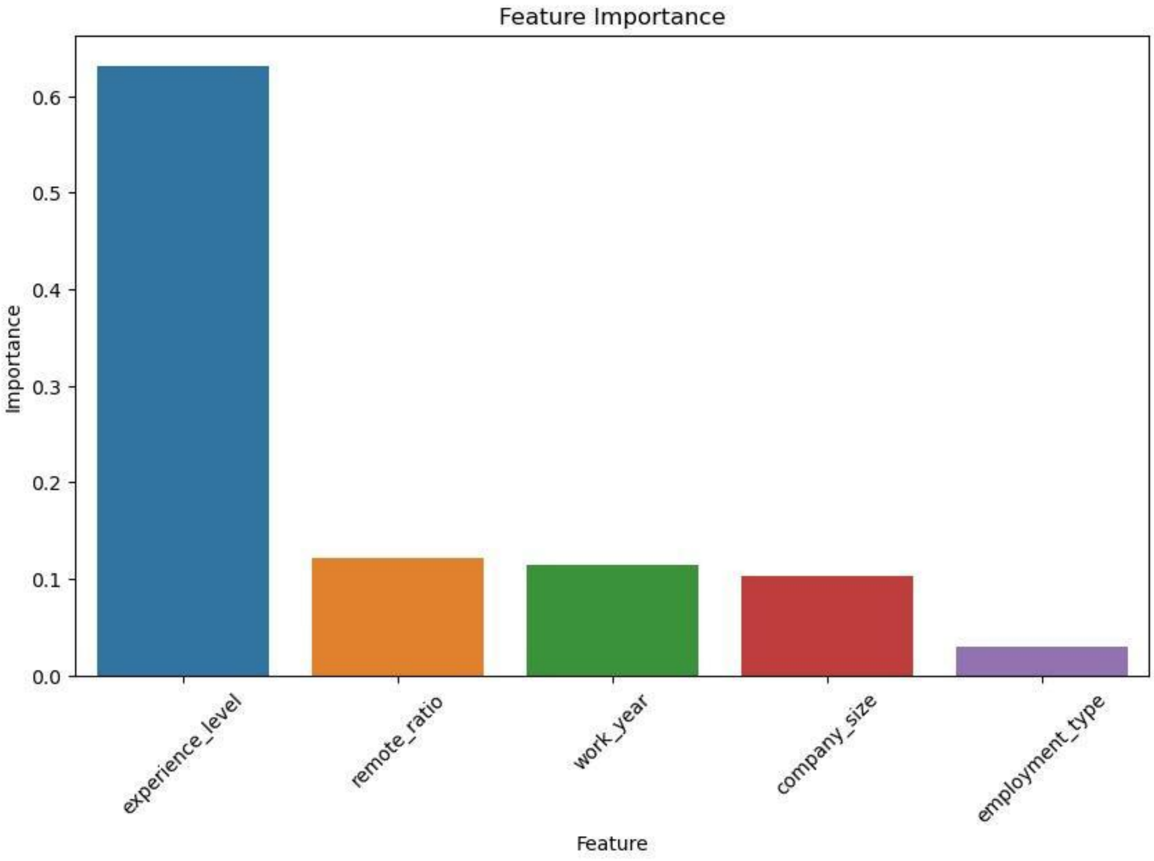
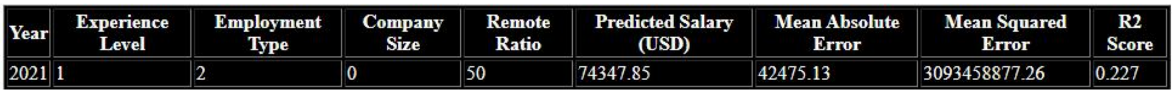


Fig 11. Feature Importance

Feature Importance: It illustrates the importance of each feature in predicting the salary. The plot shows the importance of each feature using a bar chart, with features on the x-axis and their importance scores on the y-axis. Features are sorted in descending order based on their importance.

4.3.1 Error Analysis

This table provides a structured overview of the predicted salary along with the associated evaluation metrics for the specified input parameters. The metrics include Mean Absolute Error, Mean Squared Error, and R2 Score, providing insights into the performance of the predictive model.



Year	Experience Level	Employment Type	Company Size	Remote Ratio	Predicted Salary (USD)	Mean Absolute Error	Mean Squared Error	R2 Score
2021	1	2	0	50	74347.85	42475.13	3093458877.26	0.227

Table 3. Calculating MAE, MSE, R2 Score for model accuracy

CHAPTER 5

CONCLUSION

ADVANTAGES:

Through our project on data scientist salary prediction, we have identified several key advantages and benefits:

Informed Decision-Making: Our analysis provides valuable insights into the factors influencing data scientist salaries, enabling both job seekers and employers to make informed decisions. By understanding the trends and predictors of salary variations, individuals can negotiate salaries effectively, while employers can strategically plan their workforce and budget allocation.

Predictive Modeling: The development of a predictive model using the Random Forest algorithm offers a practical tool for estimating salaries based on specified parameters. This model can assist organizations in budgeting and resource allocation, as well as aid data scientists in salary negotiation and job search.

Enhanced Understanding: Our project contributes to a deeper understanding of data scientist compensation dynamics, including variations across different experience levels, employment types, and company sizes. By uncovering these insights, stakeholders can gain a clearer picture of the data science job market and position themselves accordingly.

Scope for Future Research: The findings from our project open up avenues for future research and exploration in the field of data science employment. Further studies could delve into more granular analyses, explore additional factors influencing salaries, or develop advanced predictive models for salary estimation.

SCOPE:

The scope of our project extends beyond the current analysis and offers opportunities for future research and application.

Further Analysis: While our project provides comprehensive insights into data scientist salaries, there is scope for further analysis to explore additional factors and their impact on compensation. This could include factors such as educational background, specialized skills, industry verticals, and geographic location.

Model Refinement: Although our predictive model demonstrates promising results, there is room for refinement and optimization. Future research could focus on enhancing the accuracy and robustness of the model by incorporating advanced machine learning techniques, optimizing hyperparameters, and refining feature engineering.

Application in Other Fields: While our project focuses on data scientist salaries, similar methodologies and models could be applied to other domains and professions. This opens up opportunities for extending our analysis to other job roles, industries, and regions, providing valuable insights into salary dynamics across diverse sectors.

In conclusion, our project lays the groundwork for understanding and predicting data scientist salaries, offering valuable insights and opportunities for future research and application. By leveraging data-driven approaches and advanced analytics, we can continue to explore and address challenges in the dynamic landscape of data science employment.

GitHub Link : https://github.com/Divyam-Padole/Data-Scientist-Salary-Prediction

REFERENCES

[1]. Quan, T.Z. and Raheem, M., 2022. Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits–A Literature. Journal of Applied Technology and Innovation, 6(3), pp.70-74.

[2]. Zhang, J. and Cheng, J., 2019, August. Study of Employment Salary Forecast using KNN Algorithm. In 2019 International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019) (pp. 166-170). Atlantis Press.

[3]. Ye, L., Fu, S., Chen, G. and Lu, J., 2023, September. Salary Prediction Analysis for the 'Slow Employment'Phenomenon-Based on Random Forest Algorithm. In 2023 4th International Conference on Big Data and Informatization Education (ICBDIE 2023) (pp. 298-303). Atlantis Press.

[4]. Kaggle Notebook https://www.kaggle.com/datasets/henryshan/2023-data-scientists-salary