Data Scientist Salary Prediction

Team Members: Aryan Thengdi, 20010456
Ashish Amirneni, 20010247
Divyam Padole, 20010331
Piyush Kawale, 20010586
Tushar Charde, 20010903
Guide:Mr. Akshay Chaskar

edunet foundation

OUTLINE

- Abstract
- Problem Statement
- Aims, Objective & Proposed System/Solution
- System Design/Architecture
- System Development Approach (Technology Used)
- Algorithm & Deployment
- Conclusion
- Future Scope
- References
- Video of the Project

edunet foundation

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.

edunet foundation

Problem Statement

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

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.

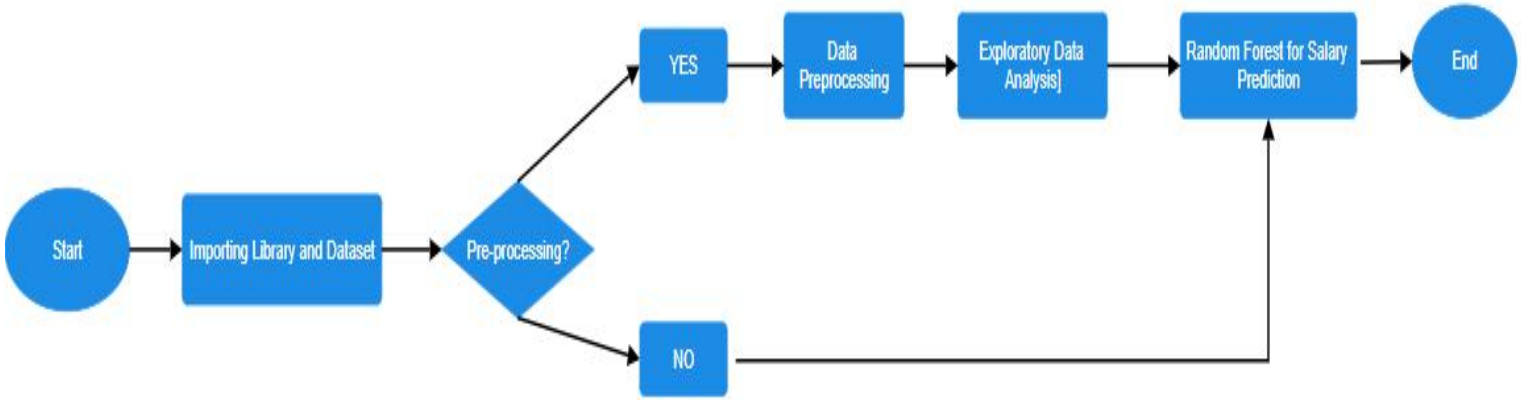
edunet foundation

Aim and Objective

- 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.

edunet foundation

System Architecture



edunet foundation

System Deployment Approach

Numerical Data Description

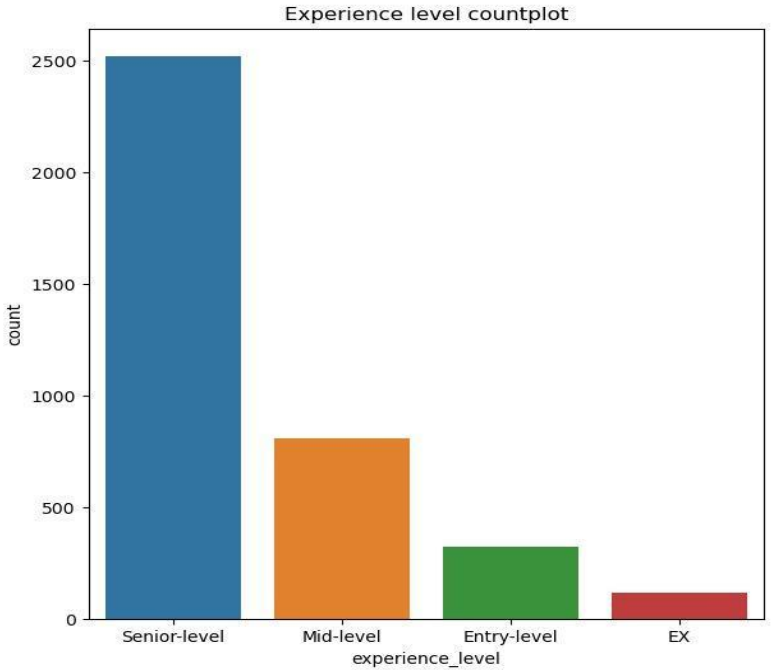
The descriptive statistics table summarizes key numerical variables in the dataset, including work year, salary, salary in USD, and remote work ratio

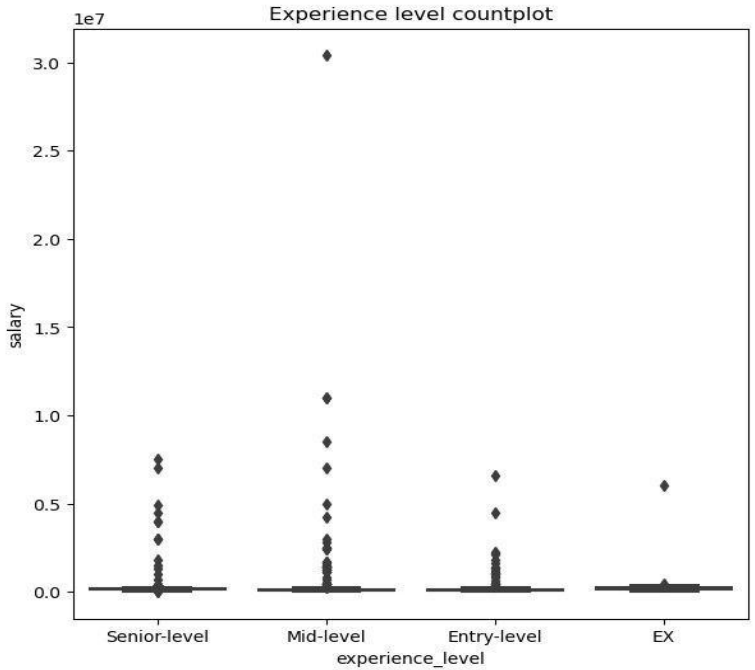
	work_year	salary	salary_in_usd	remote_ratio
count	3755.000000	3.755000e+03	3755.000000	3755.000000
mean	2022.373635	1.906956e+05	137570.389880	46.271638
std	0.691448	6.716765e+05	63055.625278	48.589050
min	2020.000000	6.000000e+03	5132.000000	0.000000
25%	2022.000000	1.000000e+05	95000.000000	0.000000
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

edunet foundation

System Deployment Approach

Experience Level Count Plot

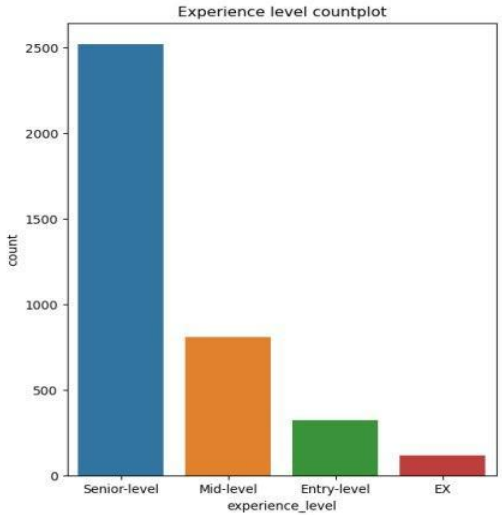


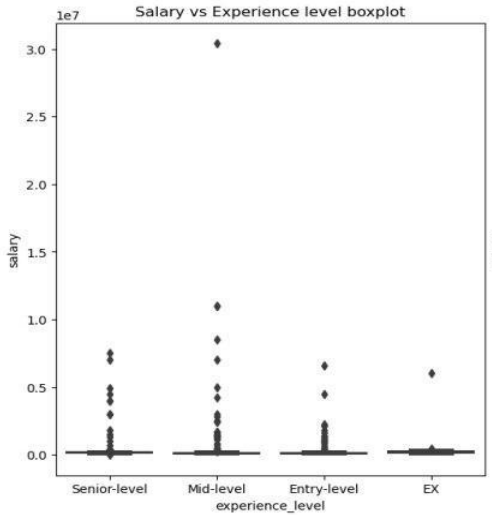


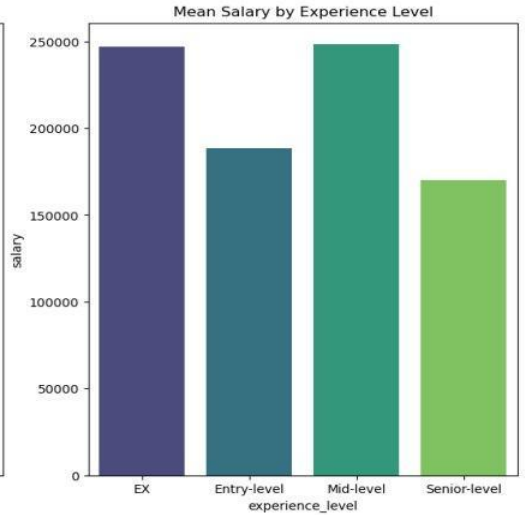
System Deployment Approach

Mean Salary by Experience Level

1. The first subplot shows a countplot of experience levels, indicating the distribution of data scientists across various experience levels.
2. 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.
3. The first subplot shows a countplot of experience levels, indicating the distribution of data scientists across various experience levels.





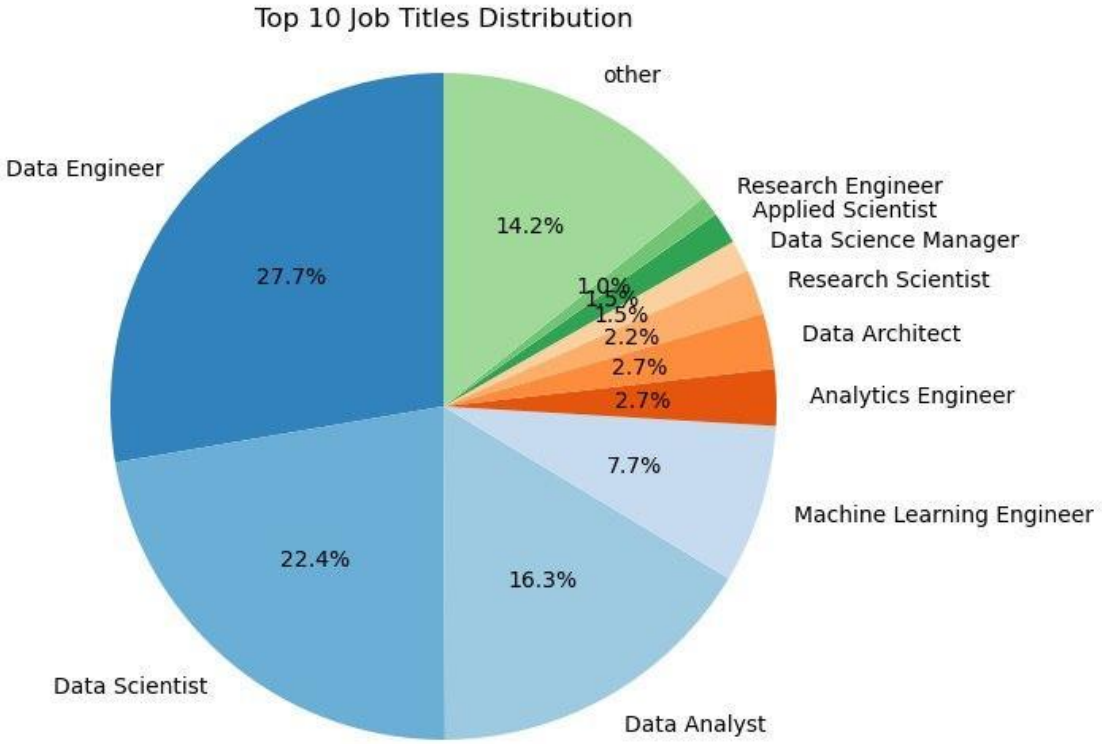


edunet foundation

System Deployment Approach

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.

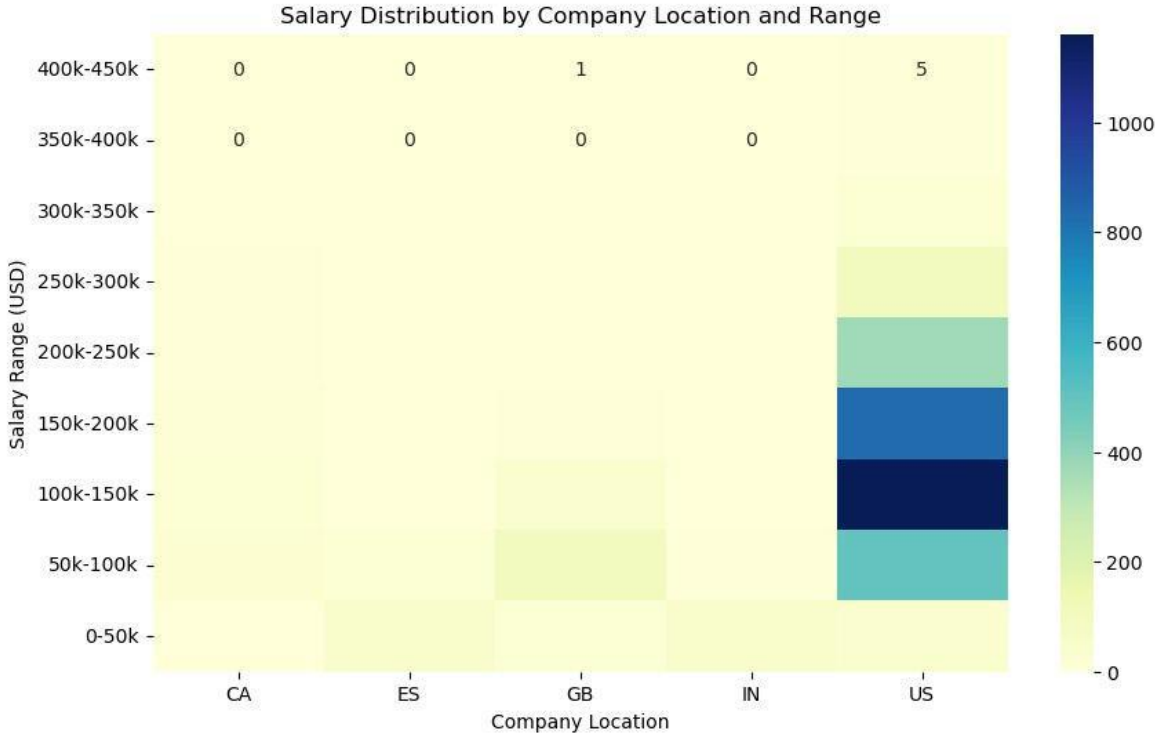


edunet foundation

System Deployment Approach

Salary Distribution by Company Location and Range

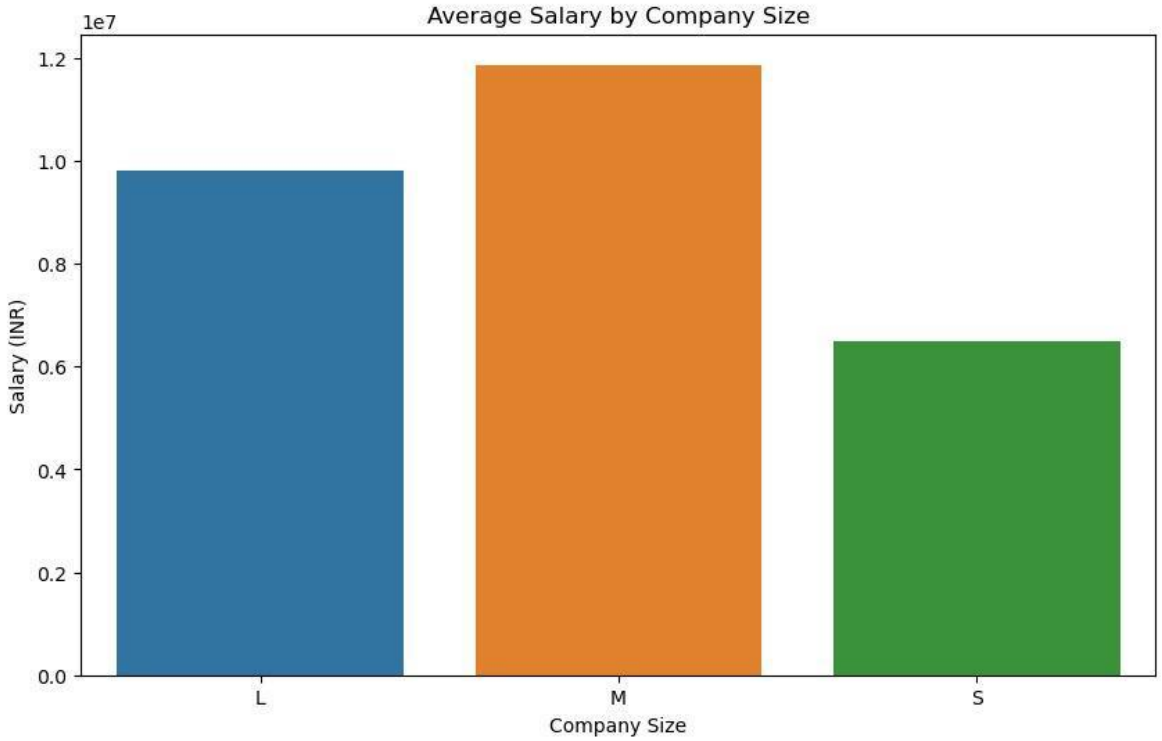
This heatmap visualizes salary distributions across different company locations (US, GB, CA, ES, IN) categorized into salary ranges. Darker colors indicate higher frequency of salaries falling within each range, offering a succinct overview of salary distributions by location and range.



System Deployment Approach

Average Salary by Company Size

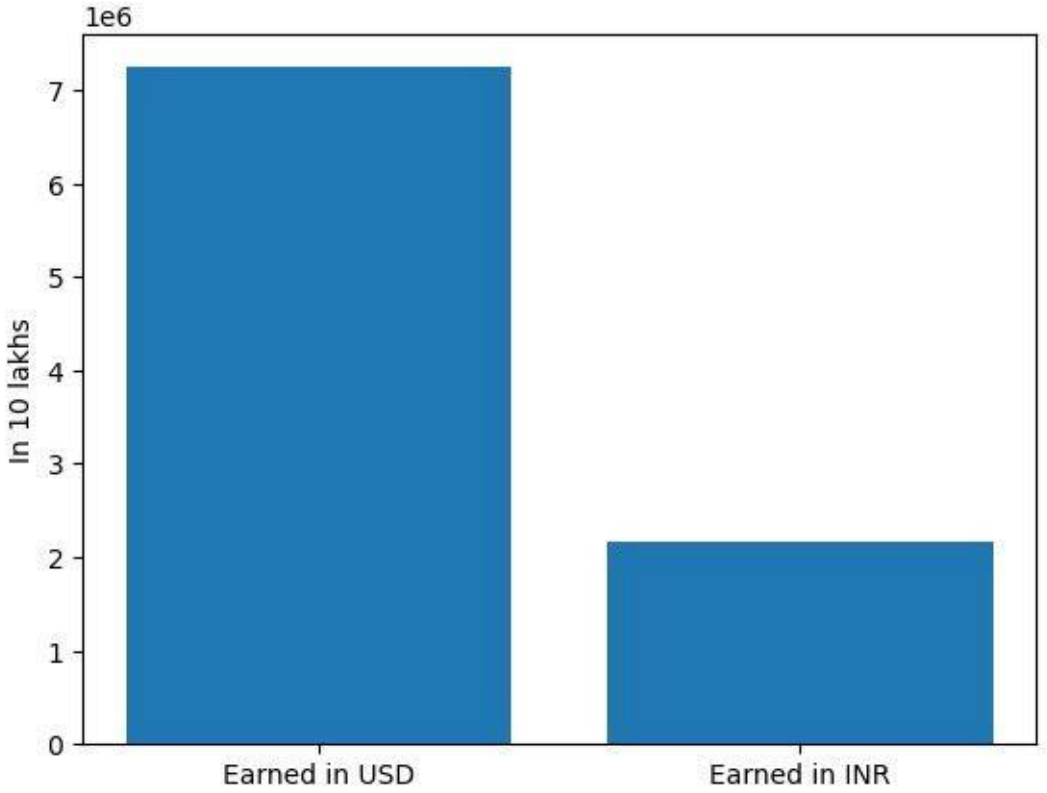
This bar plot illustrates the average salary (in INR) categorized by company size. It provides a clear comparison of salary levels across different company sizes, aiding in understanding the salary dynamics based on the scale of the employing organizations.



System Deployment Approach

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.

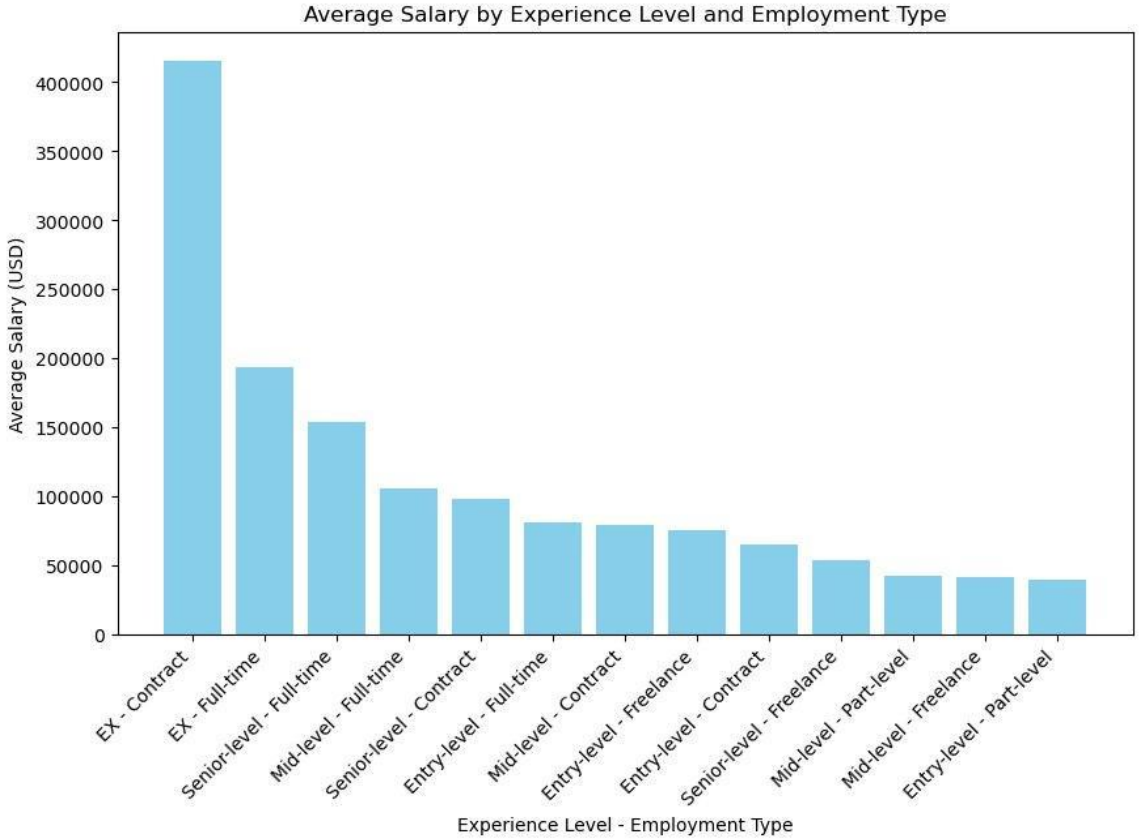


edunet foundation

System Deployment Approach

Average Salary by Experience Level and Employment Type

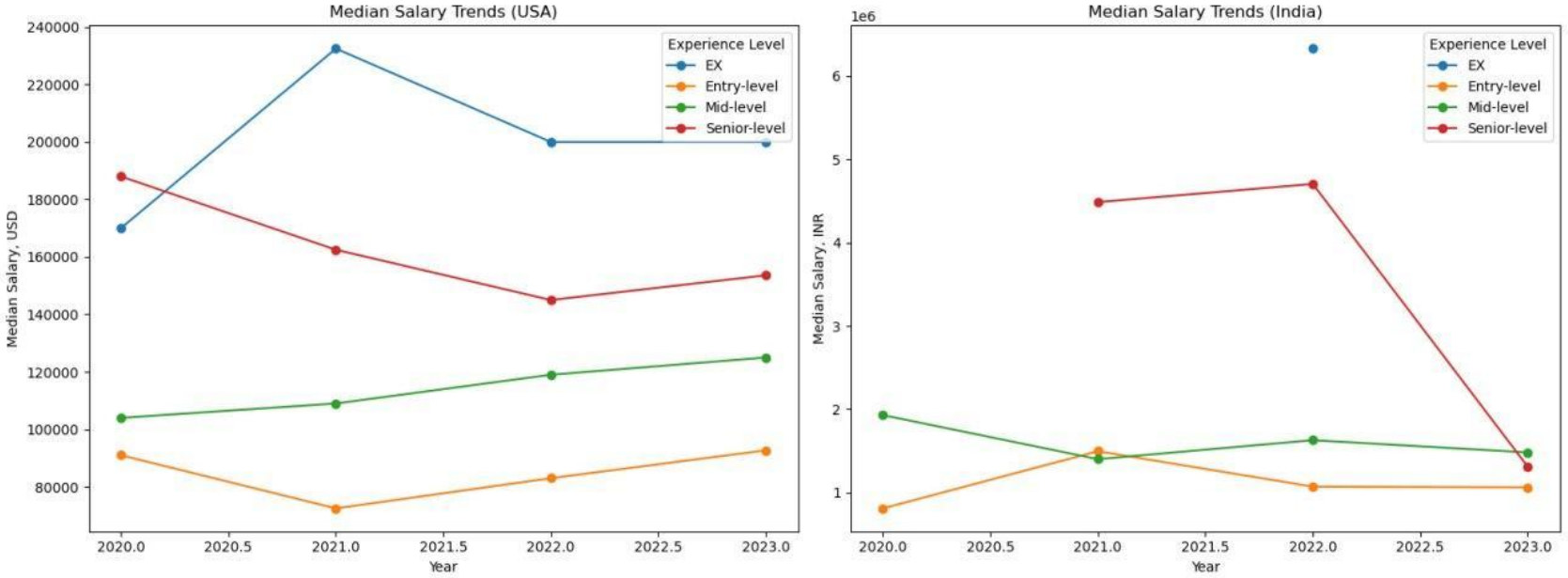
This bar plot illustrates the average salary based on different combinations of experience level and employment type. It provides a visual representation of how average salaries vary across various experience levels and types of employment.



edunet foundation

System Deployment Approach

Comparison of Salary Trends in USA and India

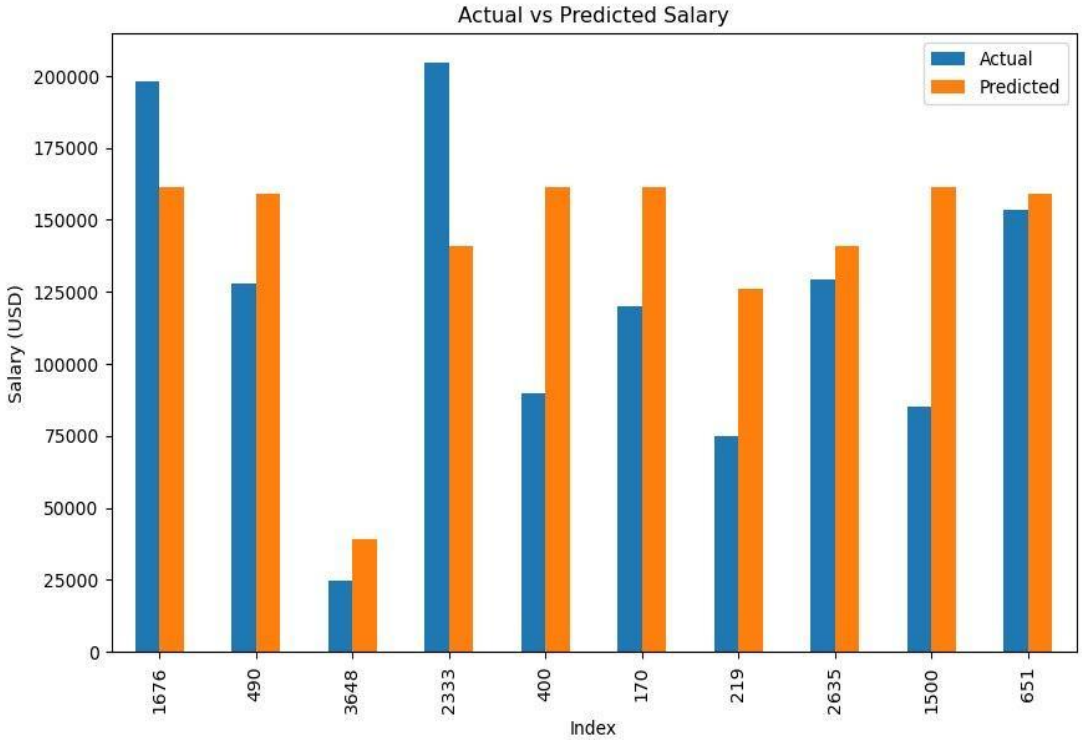


Conclusion

Result Of Model Analysis

Actual vs Predicted Salary using Random Forest

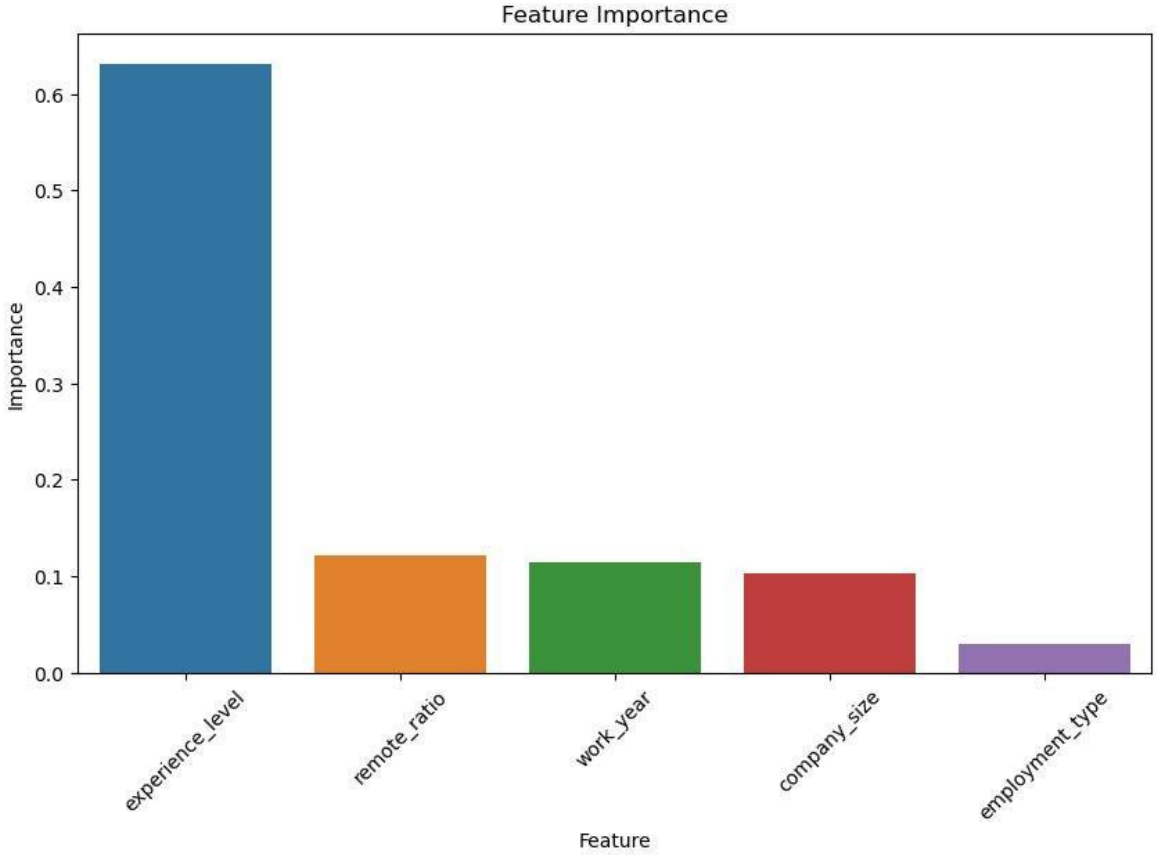
Actual vs Predicted Salary: It visualizes the comparison between actual and predicted salaries for the first 10 data points. The plot is a bar chart with the index on the x-axis and salary in USD on the y-axis.



Conclusion

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.



edunet foundation

Conclusion

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

Calculating MAE, MSE, R2 Score for model accuracy

edunet foundation

Future Scope

1. Our project on data scientist salaries suggests future analysis avenues, including factors like education, skills, industry, and location.

2. Refinement of our predictive model is essential for improved accuracy and robustness, utilizing advanced machine learning techniques.

3. Beyond data science, our methodologies can be applied to diverse professions, industries, and regions for salary analysis.

4. Our project sets the foundation for understanding and predicting data scientist salaries, offering insights and opportunities for further research.

5. Leveraging data-driven approaches, we can address challenges in the dynamic landscape of data science employment.

edunet foundation

Reference

- [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

edunet foundation

Thank you!