Machine Learning Major Project

Salary Prediction Using Regression Models

Submitted By:

Helem Thekkumvilayil Jose

B.Tech in Electrical Engineering (2024 - 2028)
Indian Institute of Technology (Indian School of Mines), Dhanbad

May - July 2025

1 Abstract

To ensure that there is no discrimination among employees, it is imperative that the Human Resources (HR) department of Company X maintains a consistent salary range for employees with similar profiles. Apart from the existing salary, various other factors—such as experience and other assessed abilities—play a role in salary decisions. This project aims to build a predictive model that determines the salary to be offered to a selected candidate, thereby reducing human bias in the salary negotiation process.

2 Goal and Objective

The objective of this project is to build a regression model using historical hiring data to predict the expected salary of a candidate. This model aims to minimize manual judgment and potential bias, ensuring fairness and transparency in salary decisions for employees with similar profiles.

3 Exploratory Data Analysis (EDA)

The dataset contains 25,000 rows and 29 features. The `ID` and `Applicant_ID` columns were dropped as they are irrelevant for model training. Several missing values (NULLs) were found across features.

- The dataset comprises 16 categorical and 11 numerical features.
- Uni-variate analysis revealed that some features (e.g., `Total_Experience_in_Field`) were right-skewed, while others followed a near-normal distribution.
- Feature scaling was necessary due to wide variance in magnitude among numerical variables.
- Bi-variate analysis showed linear relationships between several features and the target variable `Expected_CTC`.

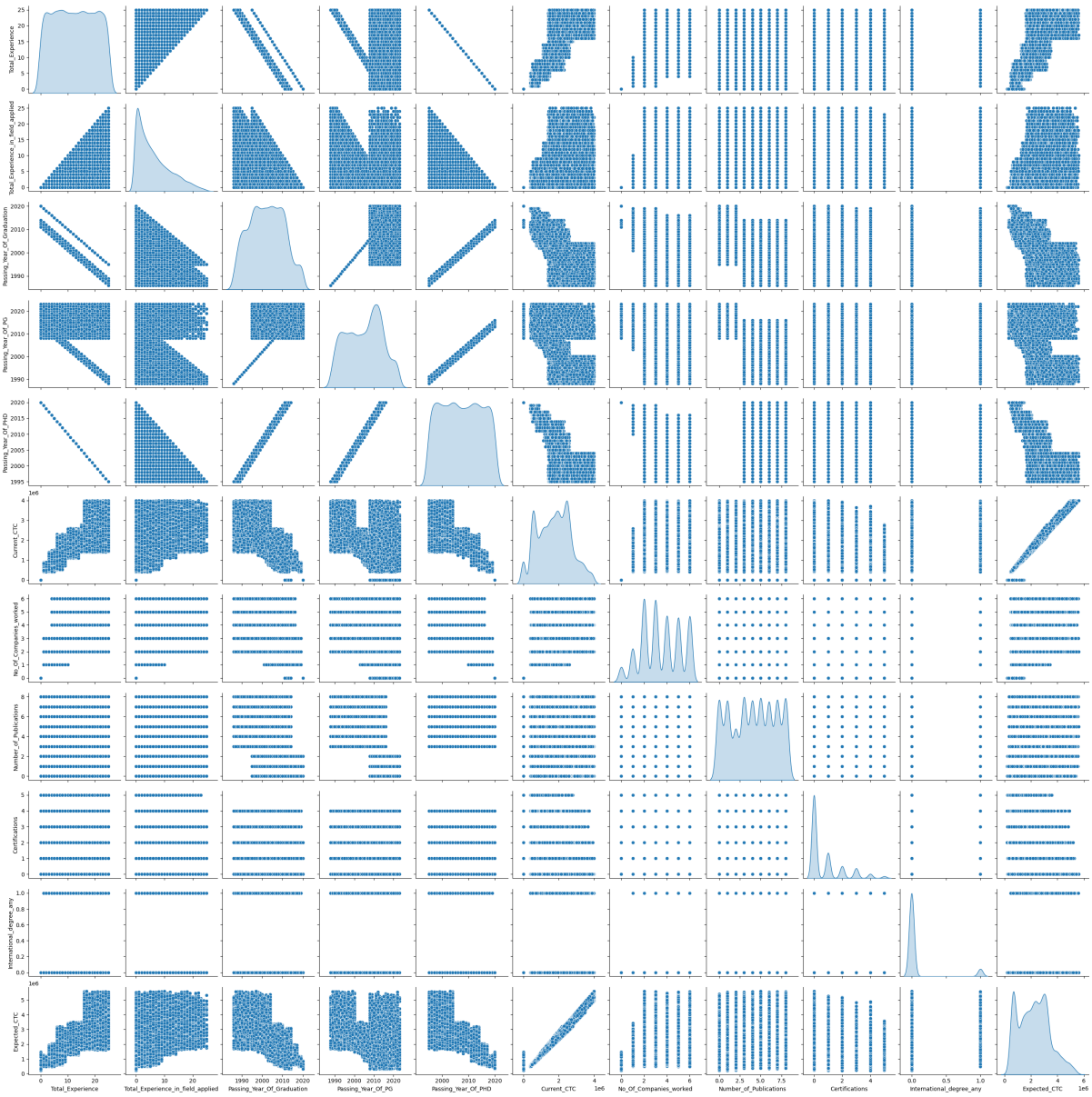

Figure 1: Pair plot showing relationships between selected features

4 Data Cleaning and Preprocessing

- Outliers were visually inspected and found to be minimal; hence, they were retained (see Figure 2).
- The correlation matrix revealed high collinearity between variables like `Passing_Year_of_PhD`, `PG`, and `Graduation`, which were therefore dropped.
- Missing values were replaced with `"NA"` for categorical variables, allowing One-Hot Encoding to handle them explicitly.
- Numerical variables were standardized using `StandardScaler()` from `scikit-learn`.

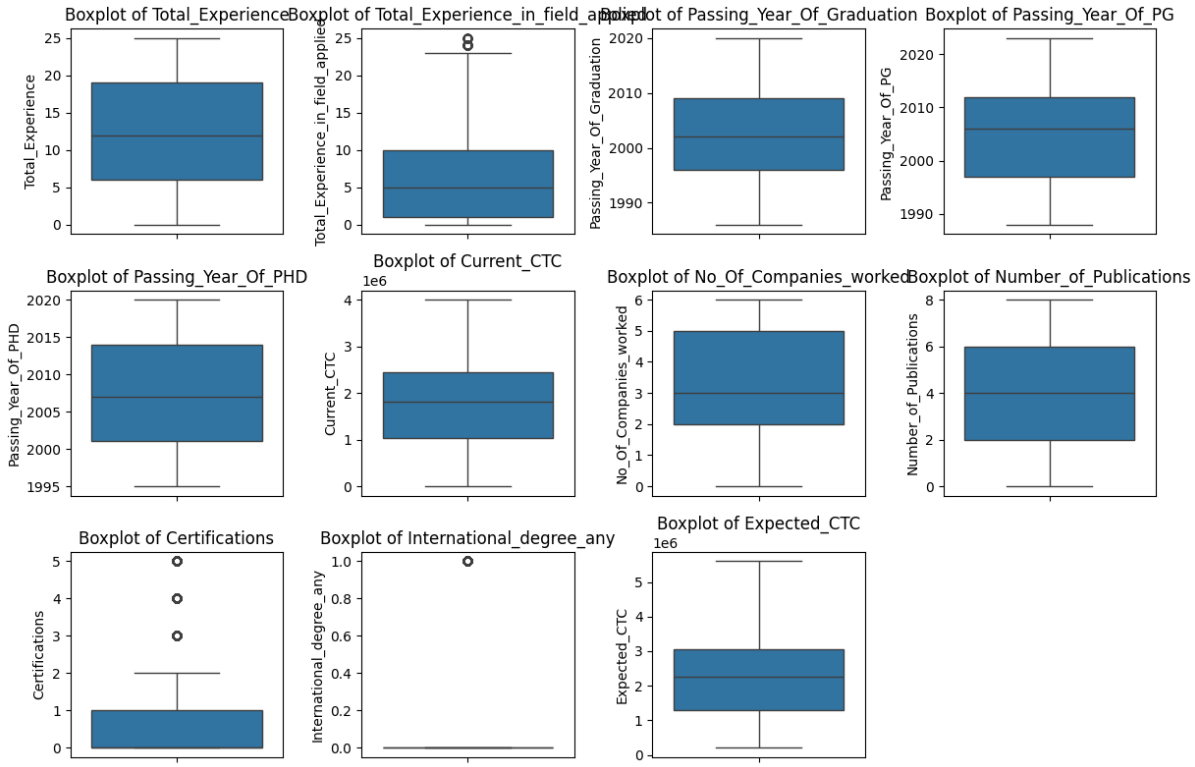

Figure 2: Box plot showing outliers in the dataset

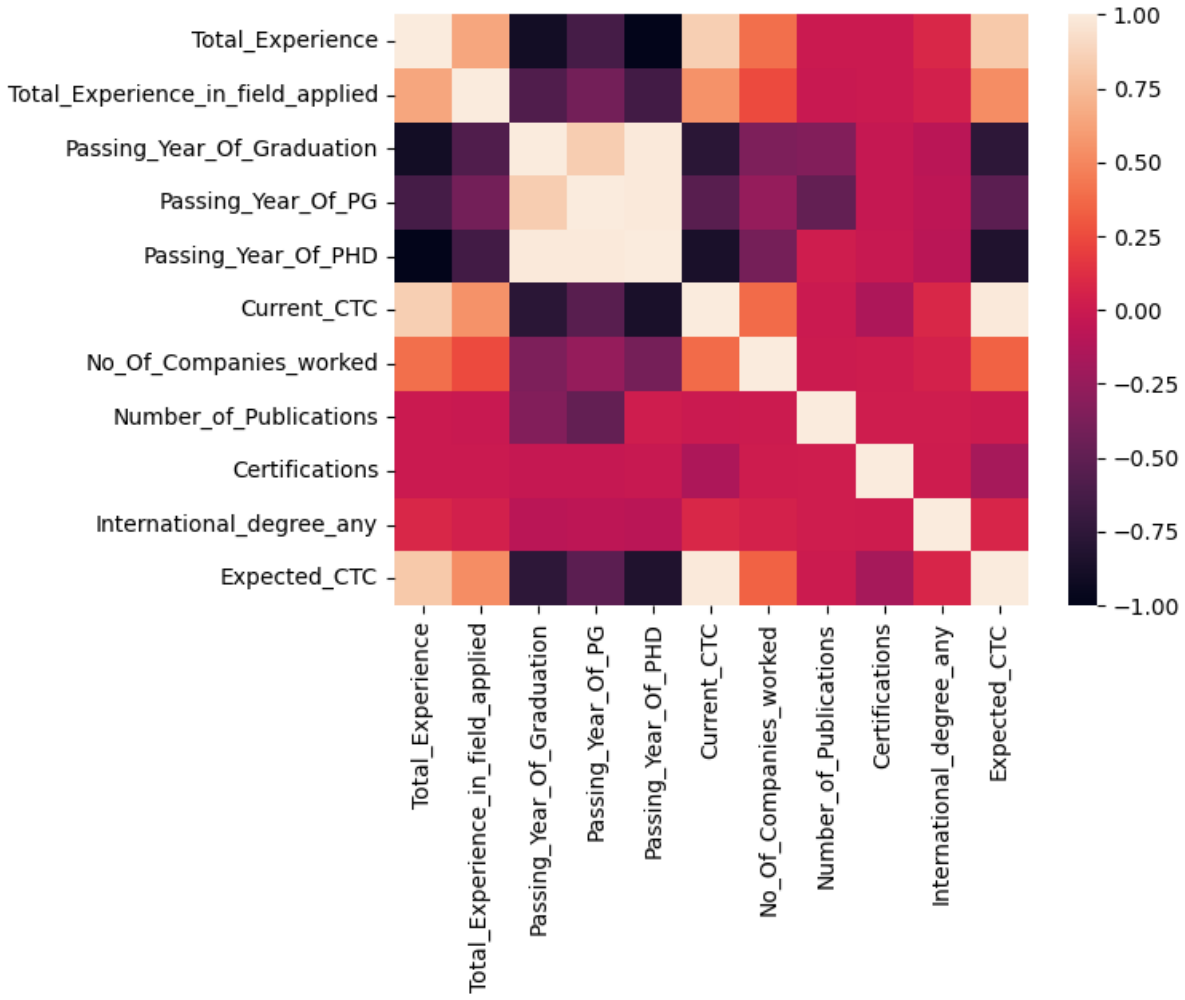

Figure 3: Heatmap of feature correlations

5 Model Selection and Training

Six regression models were trained and evaluated using cross-validation:

1. Linear Regression (Baseline)
2. Linear Regression with Principal Component Analysis (PCA)
3. Decision Tree Regressor
4. Random Forest Regressor
5. AdaBoost Regressor
6. XGBoost Regressor

Table 1: Model Performance Comparison

Model	MSE	R^2 Score
Linear Regression	5.43e+09	0.9960
Linear Regression (PCA)	5.52e+10	0.9592
Decision Tree	5.86e+08	0.9996
Random Forest	**4.31e+08**	**0.9997**
AdaBoost	2.64e+10	0.9805
XGBoost	1.19e+09	0.9991

Random Forest emerged as the best-performing model with an R^2 score of 99.97% and an average prediction error of approximately Rs. 20,757.

Model Evaluation Visualizations

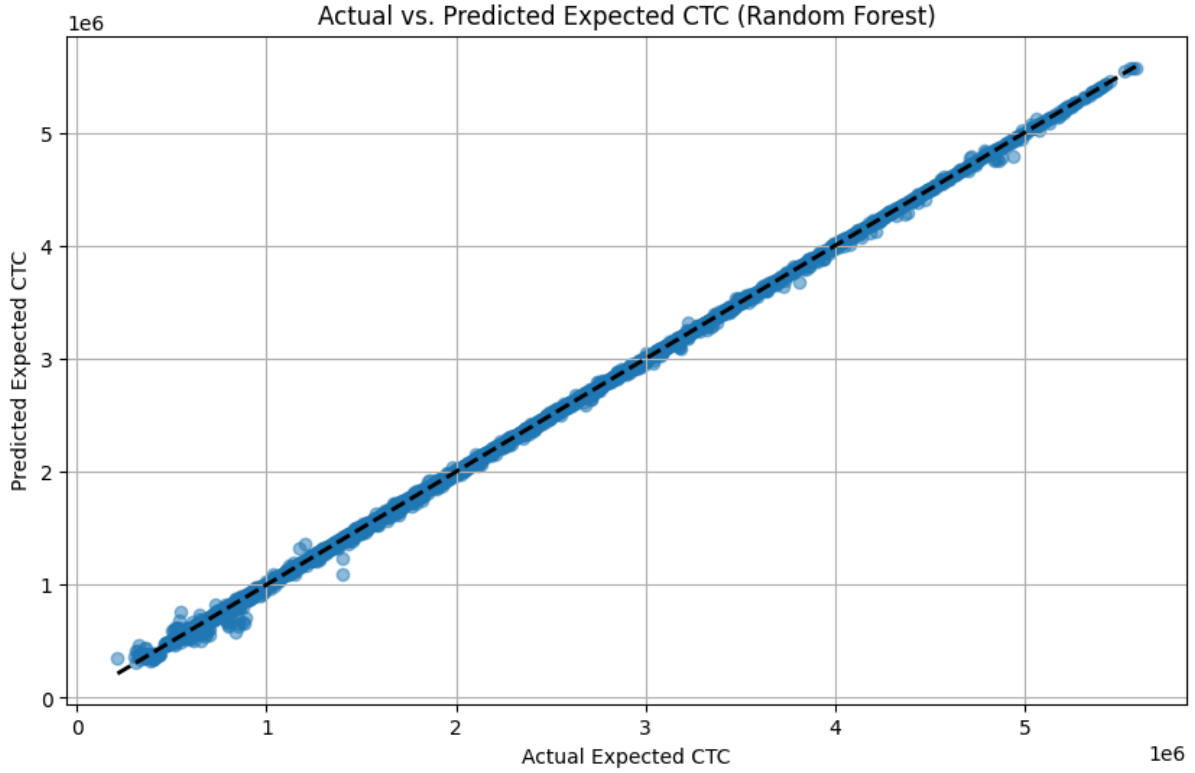

Figure 4: Random Forest Model Performance

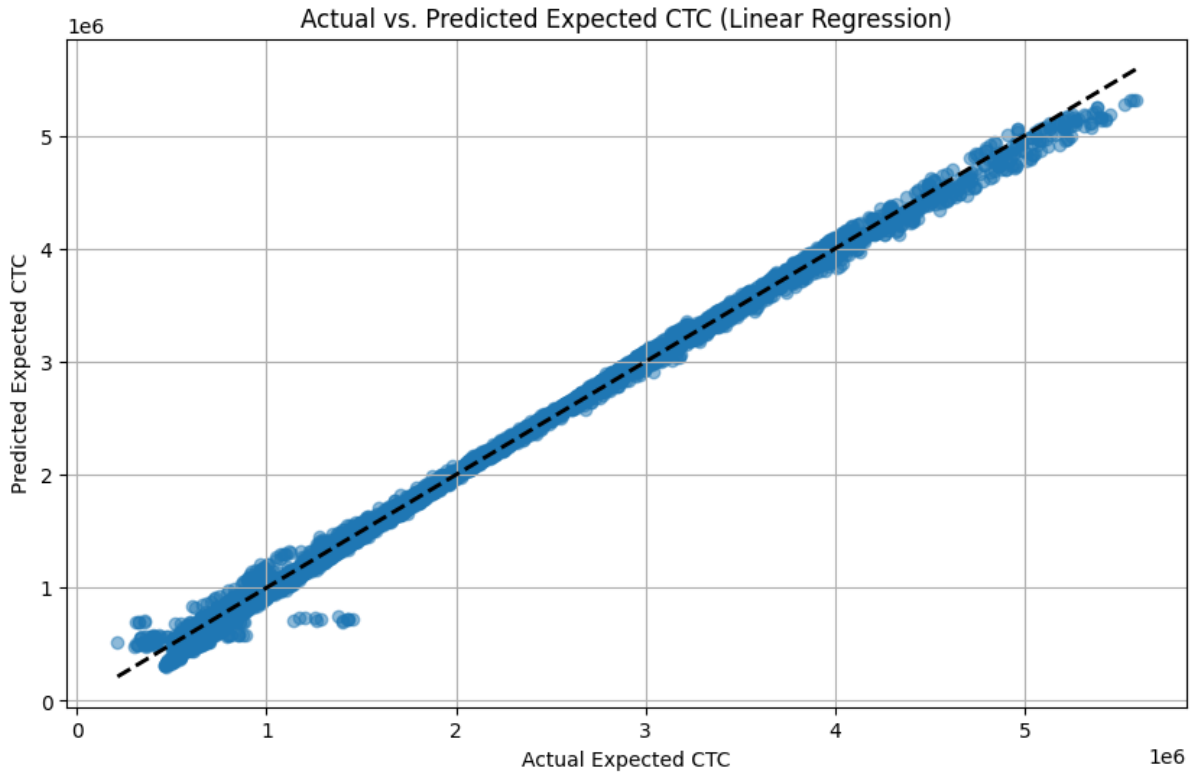

Figure 5: Linear Regression Performance

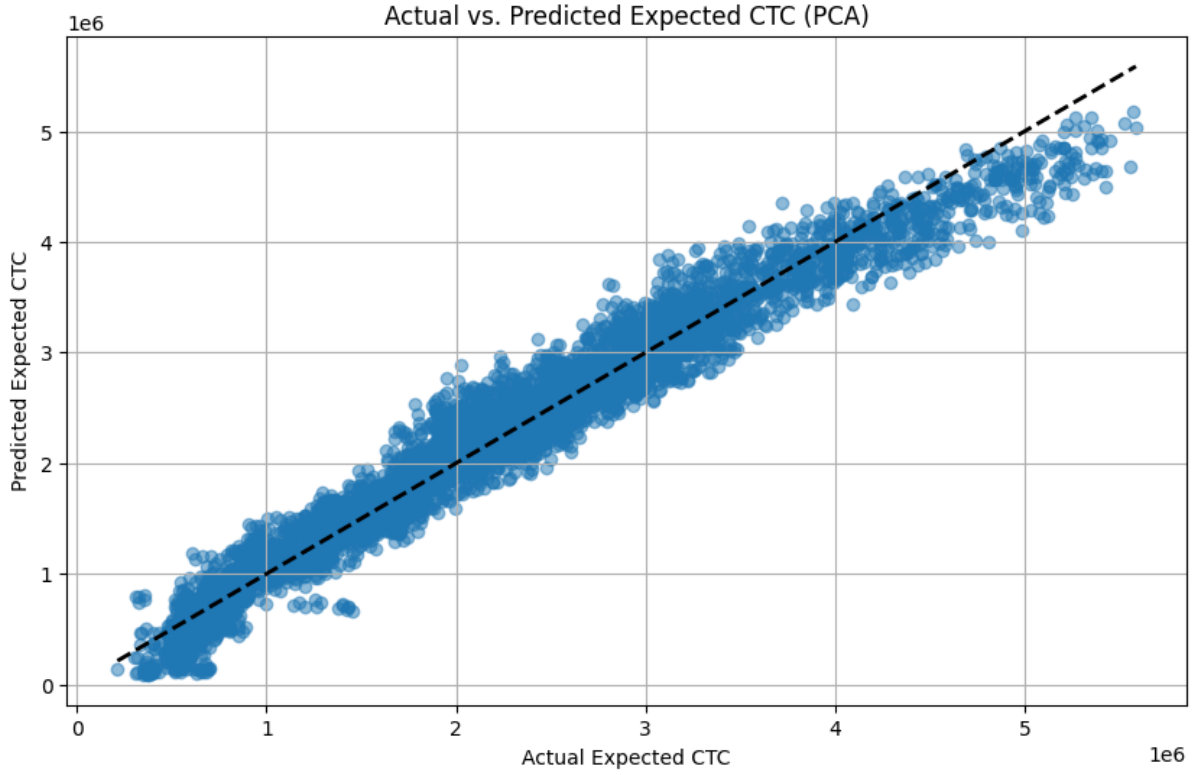

Figure 6: Performance of Linear Regression with PCA

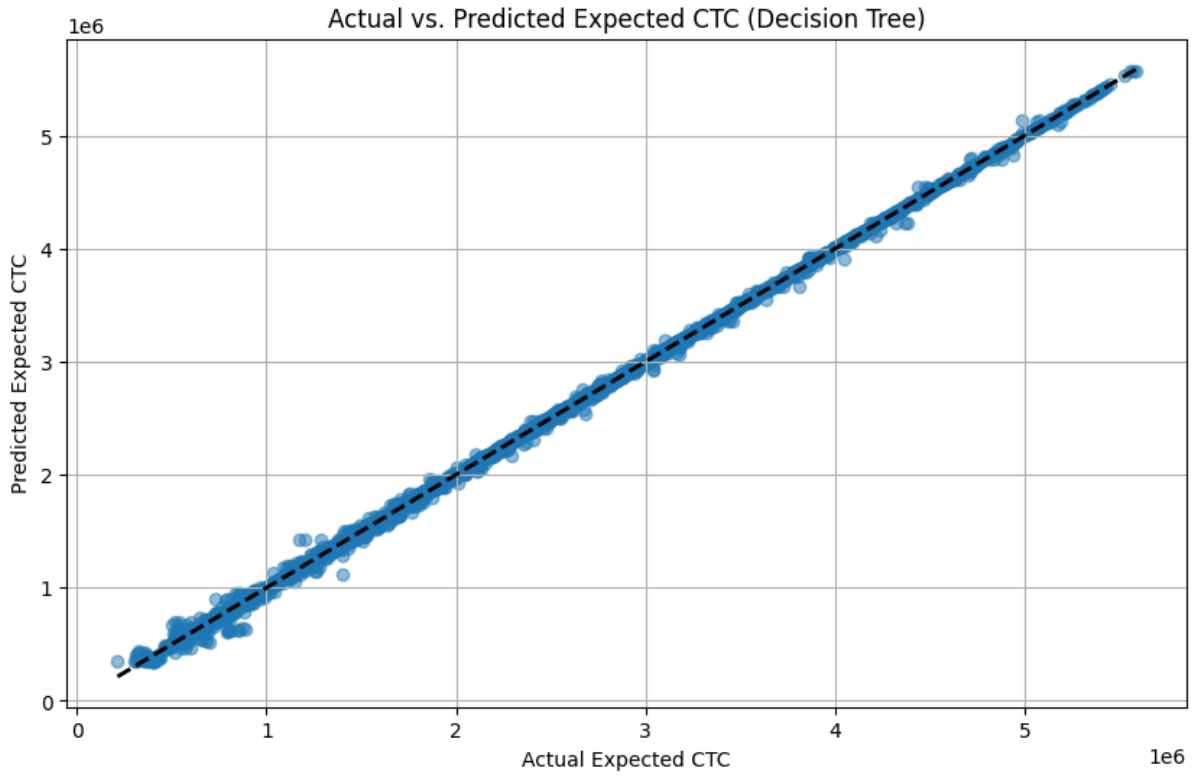

Figure 7: Decision Tree Model Performance

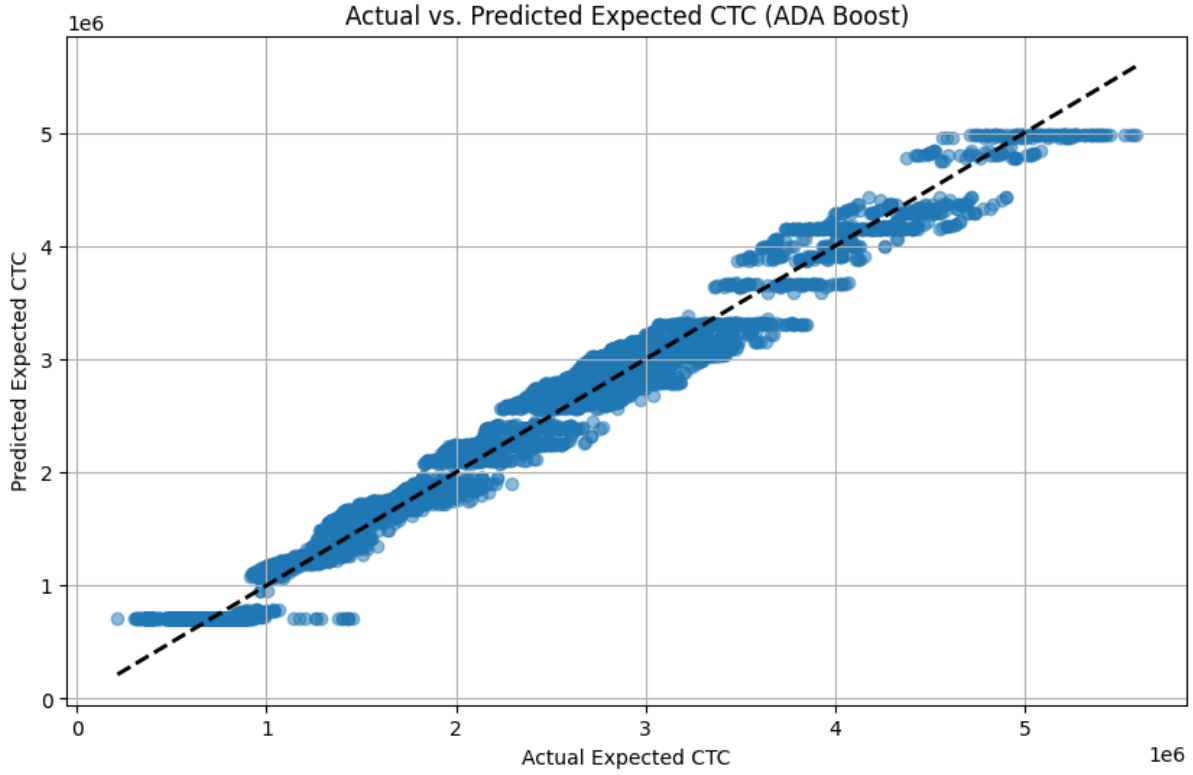

Figure 8: AdaBoost Model Performance

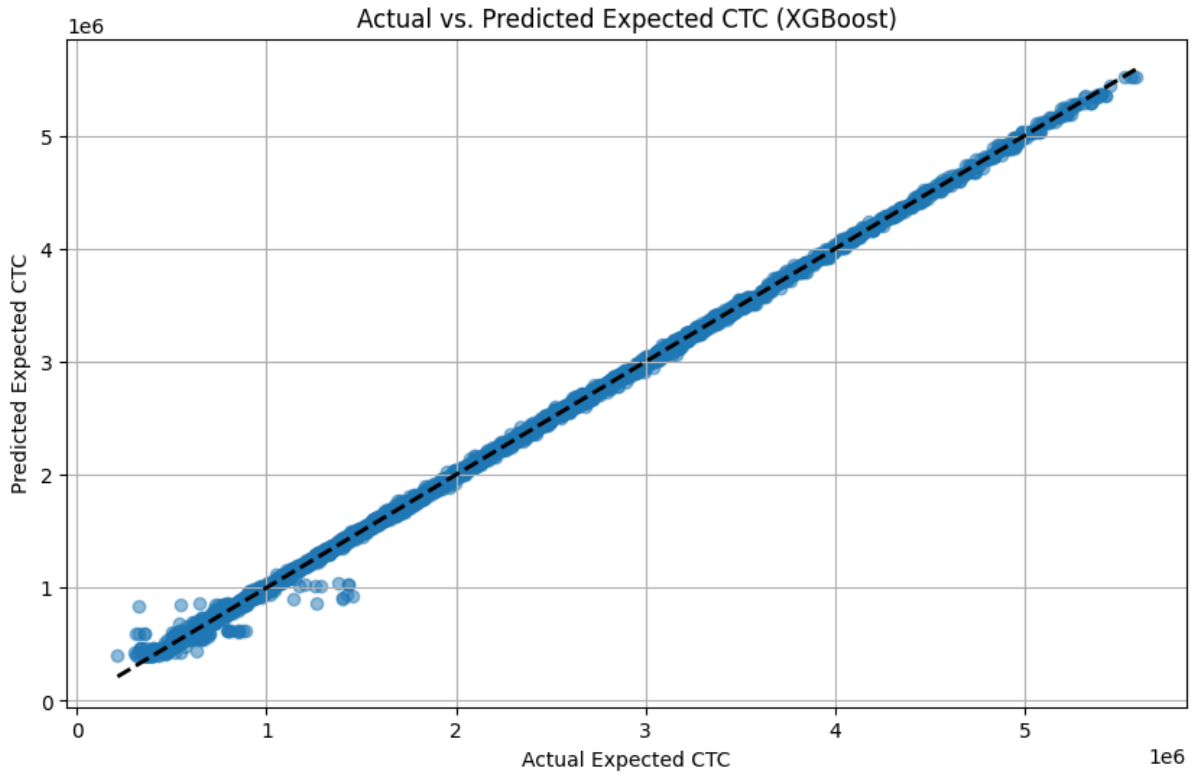

Figure 9: XGBoost Model Performance

6 Feature Importance Insights

- **Current CTC** is the most influential feature for predicting expected CTC.
- **Total experience** and **field-specific experience** significantly affect salary expectations.
- **Last appraisal rating** is critical—ratings such as "Key Performer" positively influence predicted salary, while ratings like "C" or "D" negatively affect it.
- Possession of a **Doctorate Degree** increases the predicted salary.
- Location-based features such as `Preferred Location` or `Current Location` show minimal importance, indicating no geographical bias in salary prediction.

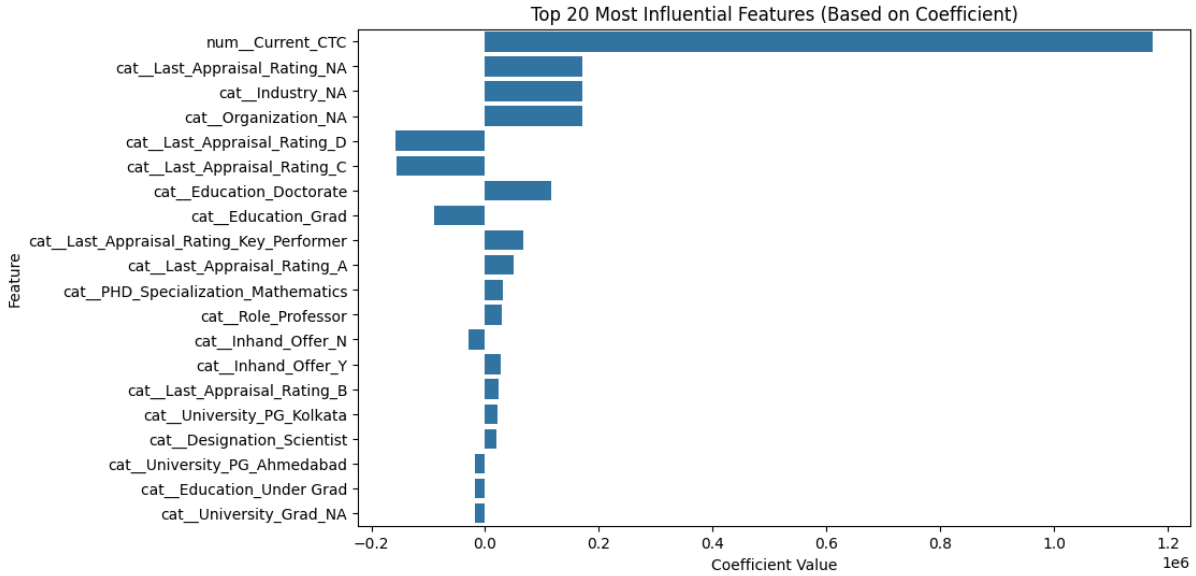

Figure 10: Top 20 Influential Features (Linear Regression Coefficients)

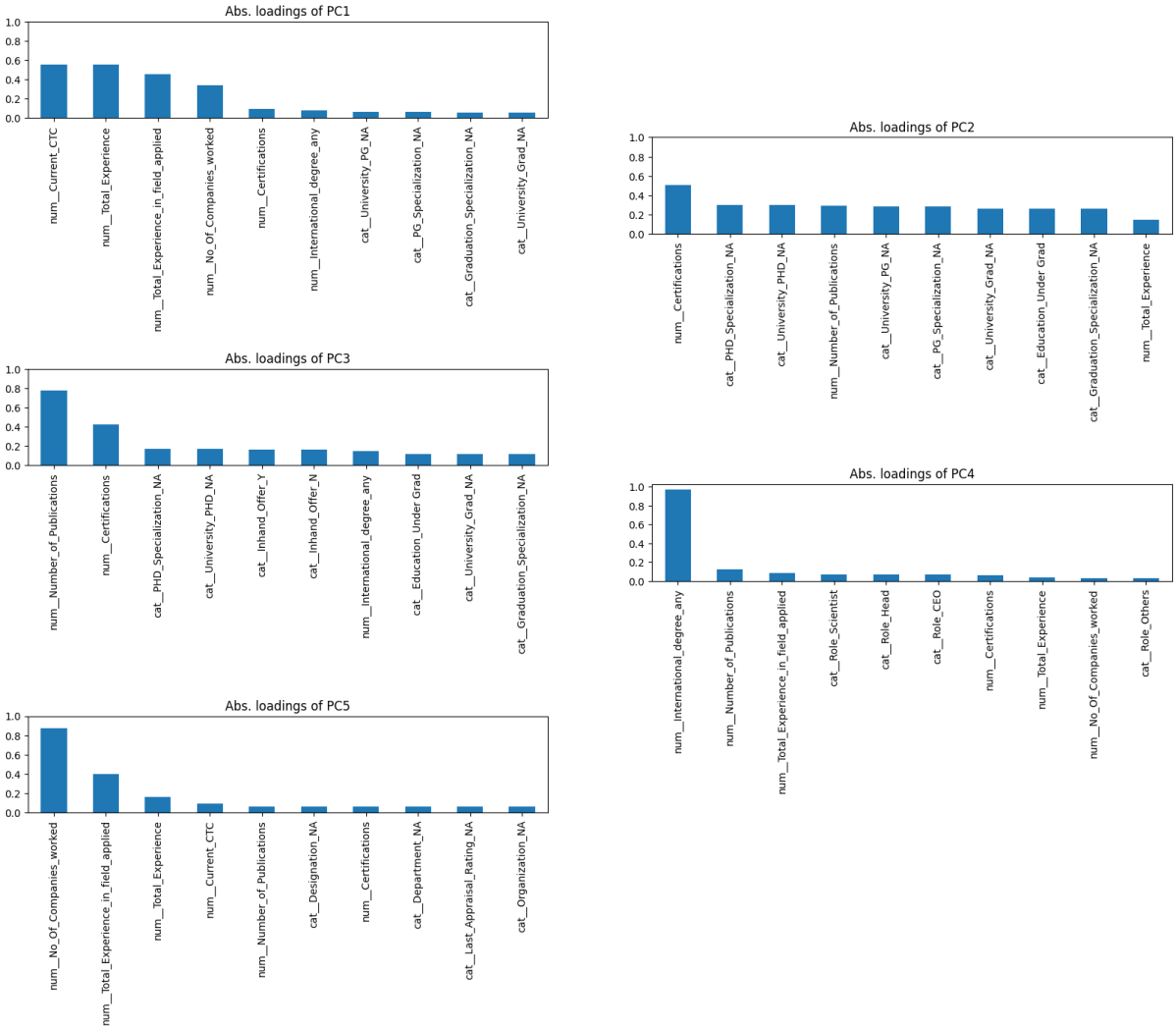

Figure 11: Top 10 Influential Features in PCA Components

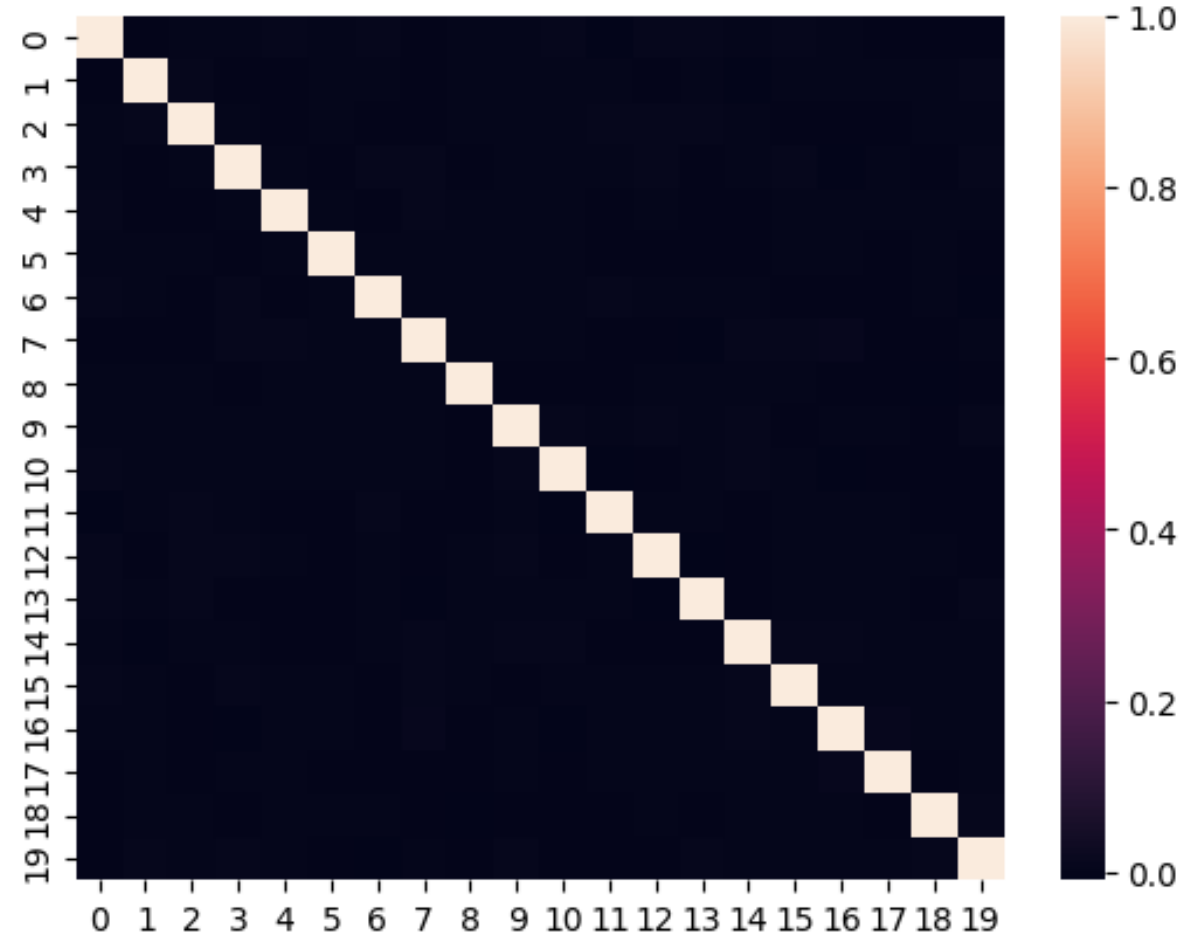

Figure 12: Correlation Among PCA Components

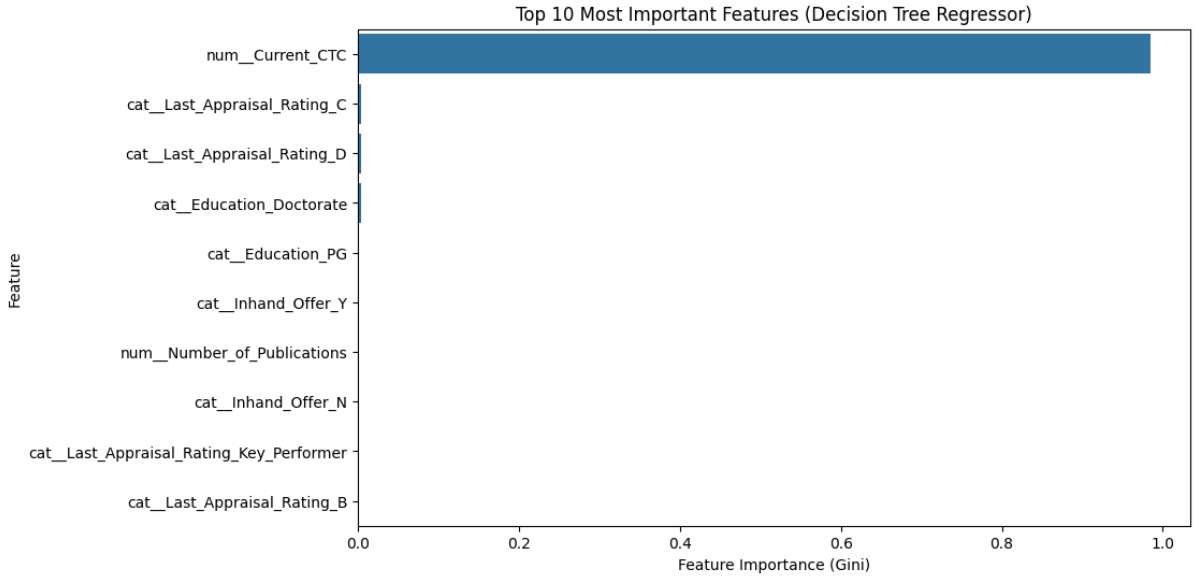

Figure 13: Top 10 Influential Features (Decision Tree - Gini Importance)

7 Conclusion and Future Work

The project successfully developed a machine learning model to predict salaries, with Random Forest achieving the best results. The insights from model explainability helped understand key drivers of salary decisions.

Future Enhancements:

- Deploy the model as an interactive web tool for HR decision-making.
- Explore fairness metrics and bias detection techniques to ensure ethical predictions.