Capstone Project - Prediction of Expected Compensation

Anna Karolin

October 30, 2025

Executive Summary

This project explores regression techniques to predict applicants' compensation expectations using a dataset of 25,000 profiles containing education, experience, and salary details. After testing and tuning multiple models, ensemble methods—Random Forest, XGBoost, and Gradient Boosting—proved most effective, achieving an R^2 of about 0.93 and RMSE of 0.18. These models accurately capture complex, non-linear relationships, with current CTC, experience, and education level emerging as the strongest predictors. The results confirm that tuned ensemble models are the most reliable choice for accurate salary forecasting and compensation analysis.

1 Introduction

In today's data-driven world, predictive modeling plays a vital role in decision-making across industries. Regression analysis, in particular, is widely used to understand relationships between variables and to predict continuous outcomes such as prices, salaries, or performance metrics. This study uses a dataset of 25,000 applicant profiles, containing features related to education, experience, and compensation, to explore the factors influencing expected salary.

The primary objective of this project is to compare the performance of various regression models, including Linear Regression, Support Vector Regression (SVR), Random Forest Regressor, and Multi-Layer Perceptron (MLP) Regressor. Each model is evaluated using standard performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) to determine which algorithm offers the best predictive accuracy and generalization capability.

Through this comparison, the project aims to identify the most effective regression approach for modeling compensation expectations, while also highlighting the strengths and limitations of different machine learning techniques in real-world predictive scenarios.

2 Data Description

The dataset consists of 25,000 applicant profiles, each containing demographic, educational, professional, and compensation-related attributes. Applicants have an average of 12.5 years of total experience, with around 6.3 years of experience in their respective applied fields, indicating a mix of early- and mid-career professionals. The graduation years range from 1986 to 2020, reflecting a diverse applicant pool in terms of seniority. About 69% of records include post-graduate details and 52% include Ph.D. qualifications, suggesting a highly educated sample. The average current annual compensation (CTC) is approximately Rs. 1.76 million, while the expected CTC averages Rs. 2.25 million, showing an anticipated salary increase among applicants. Most candidates have worked in two to five companies, possess around four publications,

and hold fewer than one certification on average. Approximately 8% of applicants have an international degree. Overall, the dataset represents a well-balanced mix of professional and academic backgrounds, suitable for analyzing factors influencing compensation expectations or career progression.

3 Exploratory Data Analysis

3.1 Data types of the features

#	Column	Non-Null Count	Dtype
0	IDX	25000 non-null	int64
1	Applicant_ID	25000 non-null	int64
2	Total_Experience	25000 non-null	int64
3	Total_Experience_in_field_applied	25000 non-null	int64
4	Department	22222 non-null	object
5	Role	24037 non-null	object
6	Industry	24092 non-null	object
7	Organization	24092 non-null	object
8	Designation	21871 non-null	object
9	Education	25000 non-null	object
10	Graduation_Specialization	18820 non-null	object
11	University_Grad	18820 non-null	object
12	Passing_Year_Of_Graduation	18820 non-null	float64
13	PG_Specialization	17308 non-null	object
14	University_PG	17308 non-null	object
15	Passing_Year_Of_PG	17308 non-null	float64
16	PHD_Specialization	13119 non-null	object
17	University_PHD	13119 non-null	object
18	Passing_Year_Of_PHD	13119 non-null	float64
19	Curent_Location	25000 non-null	object
20	Preferred_location	25000 non-null	object
21	Current_CTC	25000 non-null	int64
22	Inhand_Offer	25000 non-null	object
23	Last_Appraisal_Rating	24092 non-null	object
24	No_Of_Companies_worked	25000 non-null	int64
25	Number_of_Publications	25000 non-null	int64
26	Certifications	25000 non-null	int64
27	International_degree_any	25000 non-null	int64
28	Expected_CTC	25000 non-null	int64

Figure 1: The table shows the data type as well as the missing values in the data

3.2 Statistical Analysis

The basic description of the features of the data is as follows:

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
IDX	25000	12500.50	7217.02	1	6250.75	12500.50	18750.25	25000
Applicant_ID	25000	34993.24	14390.27	10000	22563.75	34974.50	47419.00	60000
Total_Experience	25000	12.49	7.47	0	6	12	19	25
Total_Experience_in_field_applied	25000	6.26	5.82	0	1	5	10	25
Passing_Year_Of_Graduation	18820	2002.19	8.32	1986	1996	2002	2009	2020
Passing_Year_Of_PG	17308	2005.15	9.02	1988	1997	2006	2012	2023
Passing_Year_Of_PHD	13119	2007.40	7.49	1995	2001	2007	2014	2020
Current_CTC	25000	1.76e+06	9.20e+05	0	1.03e+06	1.80e+06	2.44e+06	3.99e+06
No_Of_Companies_worked	25000	3.48	1.69	0	2	3	5	6
Number_of_Publications	25000	4.09	2.61	0	2	4	6	8
Certifications	25000	0.77	1.20	0	0	0	1	5
International_degree_any	25000	0.08	0.27	0	0	0	0	1
Expected_CTC	25000	2.25e+06	1.16e+06	2.04e+05	1.31e+06	2.25e+06	3.05e+06	5.60e+06

Table 1: Summary statistics of the applicant dataset.

3.3 Feature Distribution

The distribution of the numerical features in the data are as follows:

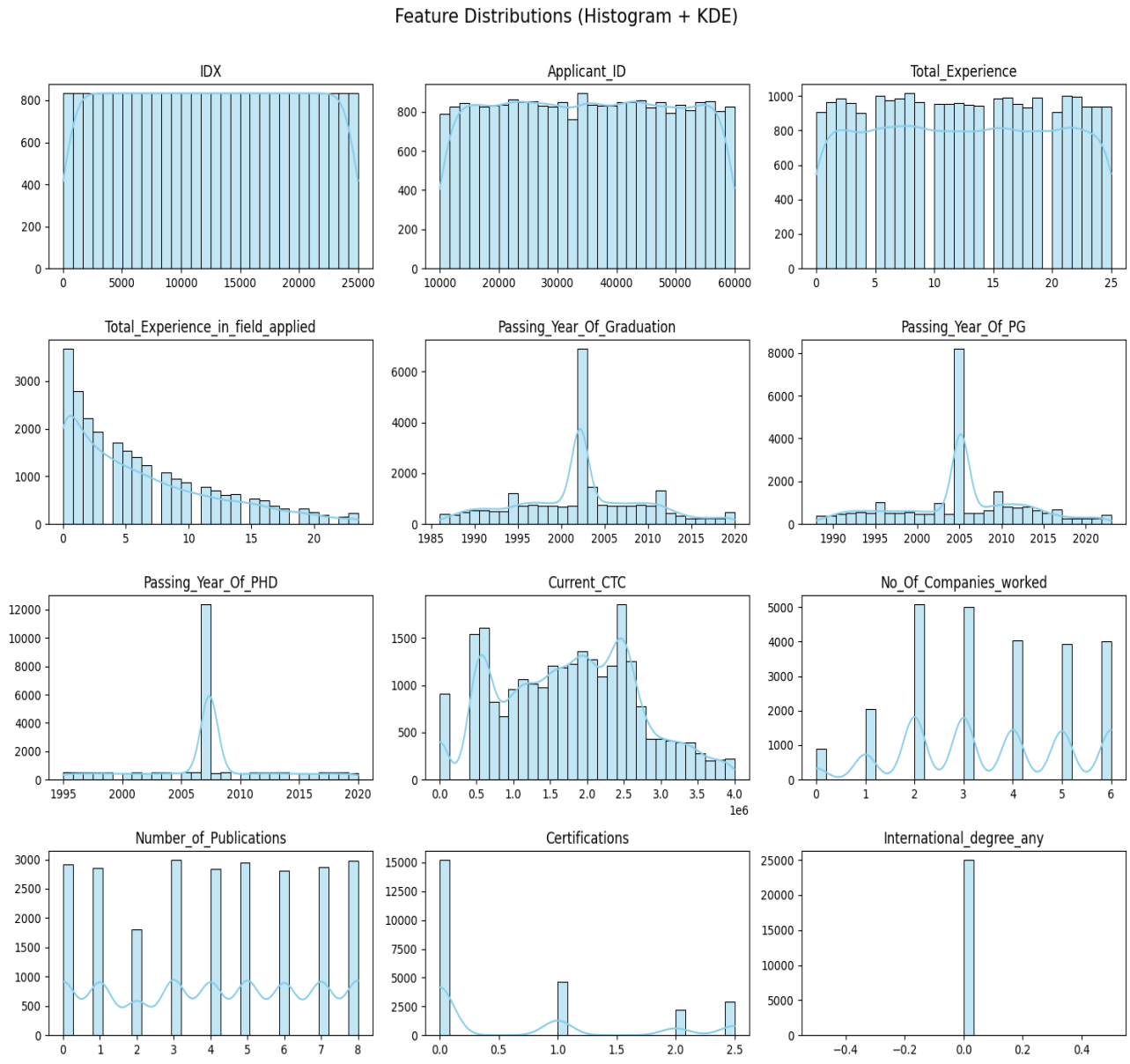

Figure 2: Distribution of the numerical features in the data

3.4 Correlations

The correlations between the features are as shown below:

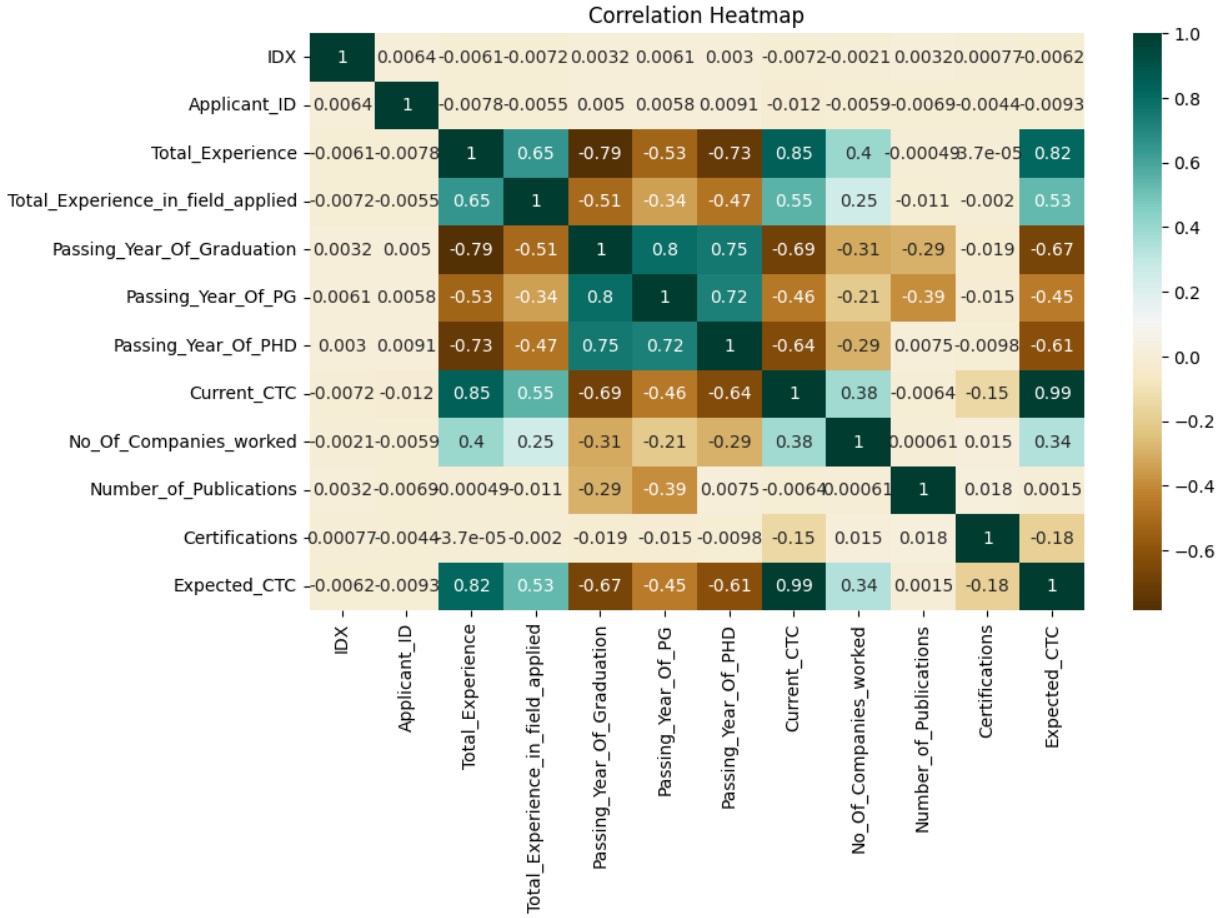

Figure 3: Correlation Matix

We can see that the features that are most correlated to Current_CTC and Total_Experience while Year_Passing_Of_Graduation and Year_Passing_OfPHD has a negative correlation with it.

3.5 Data Preprocessing

The dataset was preprocessed to ensure data quality and consistency across all algorithms. Missing values were imputed using the mean for numerical columns and the mode for categorical columns. Categorical variables such as education level and appraisal ratings were encoded using appropriate techniques (one-hot encoding) to make them compatible with machine learning models. Feature scaling was applied to all numerical features across all models to maintain uniform value ranges and prevent bias from scale differences. Additionally, outliers in compensation and experience-related features were identified using statistical methods (boxplot) and capped to minimize their impact on model performance.

4 Methodology

4.1 Models Used

The data was split into training and test set with 70% for training and 30% for test. The models used for the project are as follows:

- Linear Regression
- Support Vector Machine
- Random Forest
- AdaBoost
- Gradient Boosting
- XGBoost
- Multi Layer Perceptron

4.2 Metrics Used

Metric	Definition	Goal
MSE	Mean Squared Error	Lower = Better
RMSE	Root Mean Squared Error	Lower = Better
MAE	Mean Absolute Error	Lower = Better
R^2	Coefficient of Determination	Higher = Better

Table 2: Regression Metrics

5 Results and Analysis

Model	R^2 Train	R^2 Test	ΔR^2 (Gap)	RMSE Train	RMSE Test	Overfitting Level
Linear Regression	0.9041	0.9064	+0.0023	0.2095	0.2053	None
SVM	0.9857	0.8963	-0.0894	0.0810	0.2160	High
Random Forest	0.9894	0.9270	-0.0624	0.0696	0.1812	Moderate
Random Forest (tuned)	0.9468	0.9294	-0.0174	0.1561	0.1783	Low
AdaBoost	0.8709	0.8744	+0.0035	0.2430	0.2378	None
AdaBoost (tuned)	0.9051	0.9038	-0.0013	0.2083	0.2081	Low
Gradient Boost	0.9274	0.9256	-0.0018	0.1822	0.1830	Low
Gradient Boost (tuned)	0.9473	0.9278	-0.0195	0.1552	0.1802	Low
XGBoost	0.9779	0.9278	-0.0501	0.1006	0.1803	Moderate
XGBoost (tuned)	0.9444	0.9278	-0.0166	0.1595	0.1803	Low
MLP	0.9942	0.8741	-0.1201	0.0516	0.2381	Severe
MLP (tuned)	0.9344	0.9160	-0.0184	0.1733	0.1945	Low

Table 3: Model Performance Comparison and Overfitting Analysis

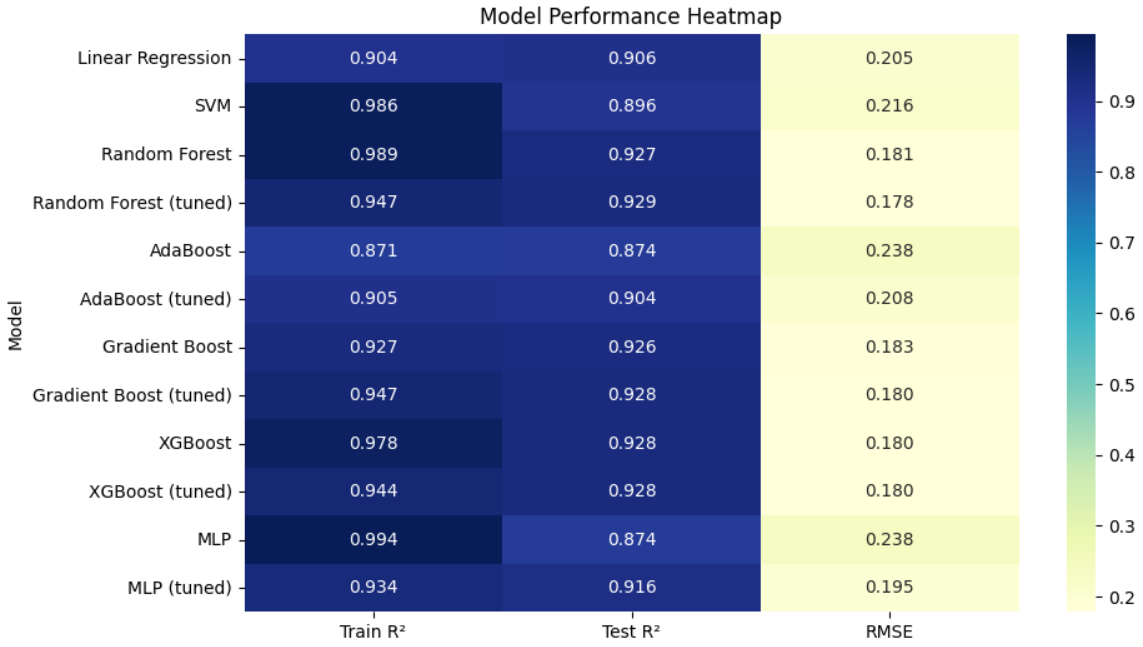

Figure 4: Heatmap of the results

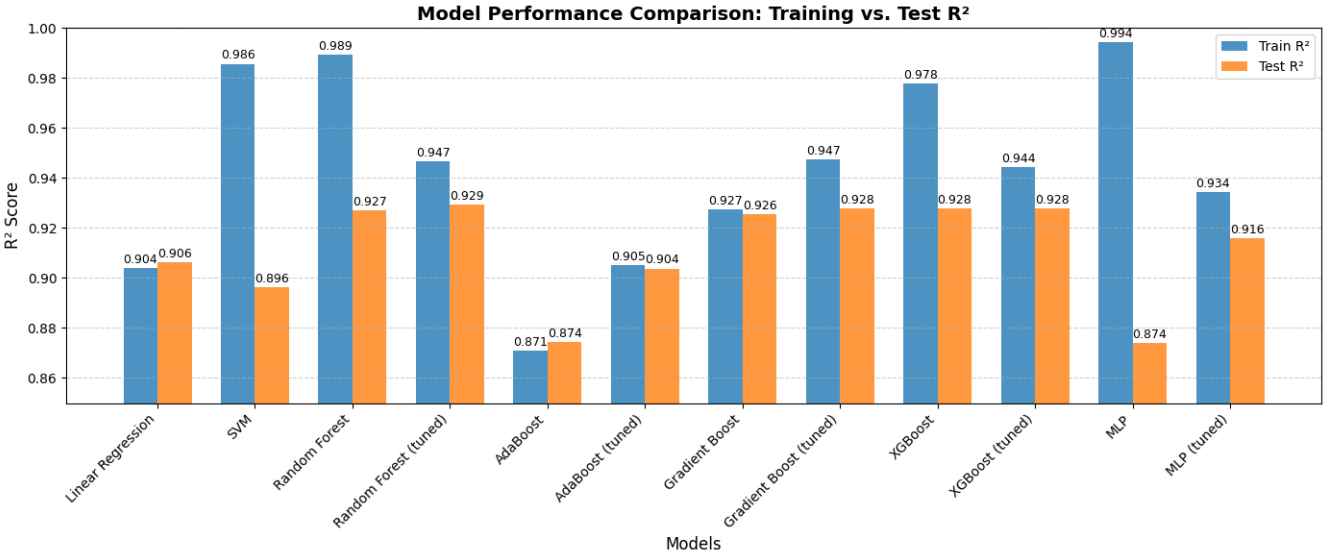

Figure 5: Training Vs Test - R^2 for all the models

5.1 Best Models

The top three models—Random Forest (tuned), XGBoost (tuned), and Gradient Boost (tuned)—all demonstrate a near-optimal balance in the bias-variance trade-off. Their performance metrics cluster tightly: an R^2 of approximately **0.93** and an RMSE of around **0.18**.

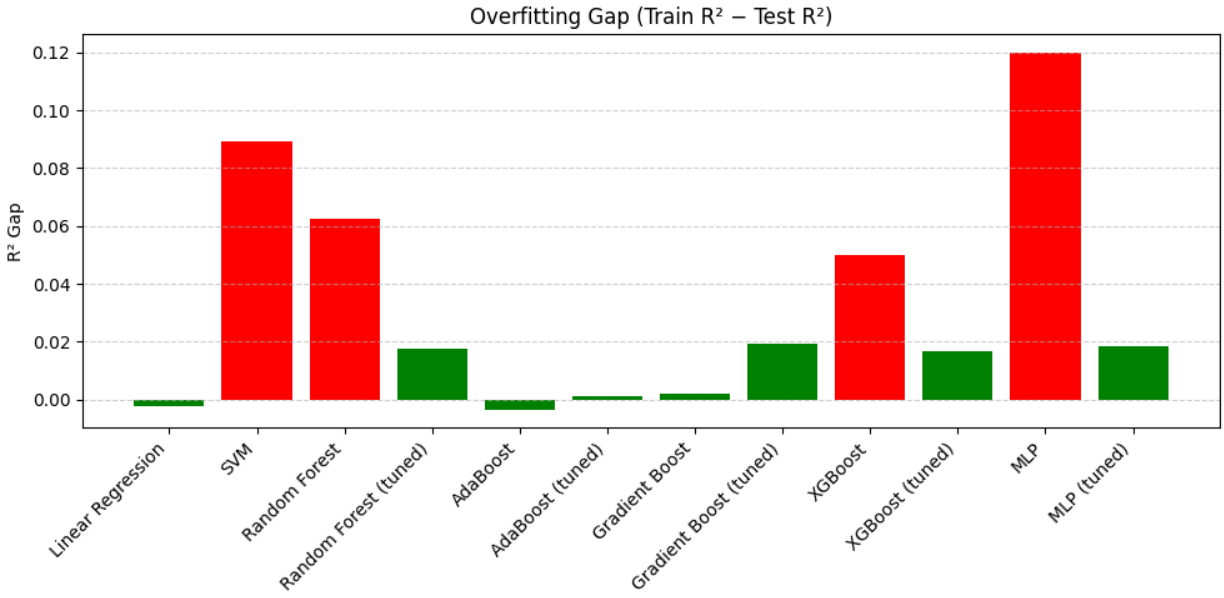

Figure 6: Summary on Overfitting by each model

5.2 Models at Risk of Overfitting (High Variance)

The models with most overfitting is MLP and SVM with the complex kernel of RBF. Although not mentioned in the results table, SVM with linear kernel also overfits the data.

5.3 Underfitting/Low Complexity (High Bias)

AdaBoost before tuning underfits the data.

5.4 Feature Importance from Ensemble Models after Fine tuning

5.4.1 Random Forest

Feature	**Importance Score**
Current_CTC	0.939993
Education_Under Grad	0.007795
Education_Grad	0.007533
Last_Appraisal_Rating_Key_Performer	0.003757
Education_Doctorate	0.003397

Table 4: Top 5 Important Features for Model Prediction

5.4.2 AdaBoost

Feature	Importance Score
Current_CTC	0.933622
Education_Doctorate	0.027894
Last_Appraisal_Rating_Key_Performer	0.021345
Total_Experience	0.007869
Last_Appraisal_Rating_C	0.004115

Table 5: Top 5 Important Features for Model Prediction

5.4.3 Gradient Boosting

Feature	Importance Score
Current_CTC	0.941561
Education_Doctorate	0.007476
Education_PG	0.005255
Last_Appraisal_Rating_D	0.004455
Last_Appraisal_Rating_C	0.004288

Table 6: Top 5 Important Features for Model Prediction

5.4.4 XGBoost

Feature	Importance Score
Current_CTC	0.666744
Education_Under Grad	0.023858
Education_PG	0.022527
Education_Grad	0.017508
Last_Appraisal_Rating_Key_Performer	0.011057

Table 7: Top 5 Important Features for Model Prediction

The column names are slightly different from the dataset given because of one code encoding. The original columns that corresponds to the ones given in the table are Current_CTC, Education, and Last_Appraisal_Rating.

6 Conclusion

The project compared several regression methods and found that tuned ensemble models—Random Forest, XGBoost, and Gradient Boosting—performed the best for predicting applicant compensation expectations. These models achieved an excellent R^2 of about 0.93 and RMSE of 0.18, showing high accuracy and stability. Their strong performance comes from effectively capturing complex, non-linear relationships between features such as current CTC, professional experience, and education level. Therefore, these ensemble models are recommended for any system designed to predict salaries or analyze the key factors influencing compensation expectations.