Capstone Project

Final Report

Fair Salary Prediction Model for Delta Ltd.: Eliminating Bias in Compensation Decisions

Table Of Contents:

1. Introduction 4
   - 1.1 Project Overview 4
   - 1.2 Problem Statement 4
   - 1.3 Affected Stakeholders 5
   - 1.4 Constraints and Approach 5
2. Data Description 6
   - 2.1 Overview of the Dataset 6
   - 2.2 Context and Variables 6
   - 2.3 Unusual Variables and Observations 7
   - 2.4 Remarks on Data Preprocessing 7
3. Exploratory Data Analysis (EDA) 8
   - 3.1 Univariate Analysis 8
     - 3.1.1 Target Variable: Expected CTC 8
     - 3.1.2 Numerical Variables 9
     - 3.1.3 Categorical Variables 11
   - 3.2 Bivariate Analysis 13
     - 3.2.1 Total Experience vs. Expected CTC 13
     - 3.2.2 Current CTC vs. Expected CTC 14
     - 3.2.3 Certifications vs. Expected CTC 15
     - 3.2.4 Education Level vs. Expected CTC 15
     - 3.2.5 Correlation Analysis 16
4. Data Cleaning and Pre-processing 16
   - 4.1 Handling Missing Values 16
   - 4.2 Outlier Detection and Treatment 17
   - 4.3 Variable Transformation 17
   - 4.4 Feature Engineering 18
   - 4.5 Variable Removal 18
5. Model Building and Interpretation 18
   - 5.1 Models Built (Predictive Modeling Focus) 18
   - 5.2 Testing Predictive Models Using Appropriate Performance Metrics 19

5.3 Interpretation of the Models....19
6. Model Tuning....20
6.1 Ensemble Model Tuning (via RandomizedSearchCV)....20
6.2 Interpretation of the Optimal Tuned Model....21
7. Model Validation....22
8. Final Interpretation & Recommendations....23
8.1 Business Insights....23
8.2 Recommendations....24

List Of Figures:

Fig 3.1 Distribution of Expected_CTC....8
Fig 3.2 Distribution of Total_Experience....9
Fig 3.3 Distribution of Current CTC....10
Fig 3.4 Distribution of Certifications....10
Fig 3.5 Distribution of Number of Publications....11
Fig 3.6 Distribution of Education....11
Fig 3.7 Distribution of Industry....12
Fig 3.8 Distribution of Department....12
Fig 3.9 Distribution of Role....13
Fig 3.10 Total Experience vs. Expected CTC....14
Fig 3.11 Current CTC vs. Expected CTC....14
Fig 3.12 Certifications vs. Expected CTC....15
Fig 3.13 Education Level vs. Expected CTC....15
Fig 3.14 Correlation Heatmap: Numerical Variables....16
Fig 3.15 Outlier Treatment for Certifications....17
Fig 3.16 Variable Transformation for Certifications....18

1. Introduction

1.1 Project Overview

In today's corporate environment, salary determination is a critical factor that influences employee satisfaction, retention, and overall business success. Delta Ltd. is committed to ensuring fair compensation for its employees by leveraging data-driven methods rather than relying on subjective decision-making. This project aims to develop a robust salary prediction model based on employee attributes such as experience, education, industry, role, and current salary. By analyzing historical compensation data, the model will help standardize salary decisions, reducing human biases and ensuring fair pay across different employee groups.

The salary prediction model will use machine learning techniques to estimate the expected salary (Expected_CTC) for employees based on various factors. The goal is to provide a fair and objective salary recommendation system that aligns with industry standards while minimizing pay disparities.

1.2 Problem Statement

The traditional salary determination process often involves human judgment, which can introduce biases and inconsistencies. Employees with similar qualifications and experience sometimes receive varying salaries due to factors such as negotiation skills, managerial discretion, and organizational budget constraints. This creates a sense of unfairness and dissatisfaction among employees, potentially leading to increased attrition rates and decreased motivation.

To address this challenge, we aim to build a salary prediction model that can answer key business questions:

- How can we predict the expected salary of an employee based on historical data?
- What are the key factors influencing salary expectations?
- How can we minimize salary discrimination and standardize compensation policies?

By solving these questions, Delta Ltd. can ensure that its compensation structure remains competitive, equitable, and aligned with business goals.

1.3 Affected Stakeholders

Several stakeholders will benefit from the insights generated by this project:

- **Human Resources (HR) & Compensation Teams:** The salary prediction model will provide a data-driven approach to salary decisions, ensuring consistency in pay structures and reducing subjective biases.
- **Employees:** Transparent salary policies will increase employee trust and satisfaction, leading to higher retention rates and improved morale.
- **Leadership & Management:** Business leaders will gain insights into salary trends and market competitiveness, helping them make informed financial decisions.
- **Regulatory Authorities:** The model can help organizations comply with fair pay regulations by ensuring salaries are distributed equitably across employee segments.

1.4 Constraints and Approach

The development of a salary prediction model comes with several constraints:

- **Data Availability:** The dataset must contain sufficient historical salary information across different roles and experience levels to train an accurate model.
- **Bias Reduction:** The model should minimize biases related to gender, race, or other irrelevant factors while ensuring fair pay practices.
- **Feature Engineering:** New variables such as Experience Level and Salary Increment Percentage need to be derived to improve the model's predictive power.
- **Interpretability:** The model must provide transparent and explainable results, ensuring that HR teams and management can trust its recommendations.

To address these constraints, we follow a structured approach:

1. **Data Collection and Cleaning** – Ensure data quality and handle missing values.
2. **Exploratory Data Analysis (EDA)** – Identify trends, correlations, and patterns.
3. **Feature Engineering** – Create meaningful features that improve predictive accuracy.

4. **Model Development and Evaluation** – Train machine learning models and select the best-performing one.
5. **Business Insights and Recommendations** – Provide actionable insights for HR and management teams.

2. Data Description

2.1 Overview of the Dataset

The dataset used for this project consists of employee-level data that includes various attributes influencing salary expectations. It comprises both numerical and categorical variables, providing insights into the factors that determine an employee’s expected salary (Expected_CTC). The dataset serves as the foundation for building predictive models that ensure fairness and transparency in salary distribution.

The dataset contains key attributes such as:

- Total Experience (Years)
- Current CTC
- Expected CTC
- Industry
- Role
- Department
- Education
- Certifications
- Number of Publications
- Last Appraisal Rating

2.2 Context and Variables

Each variable in the dataset plays a critical role in influencing salary expectations.

- **Numerical Variables:**
  - **Total Experience:** The total number of years an employee has worked.
  - **Current CTC:** The employee's current salary, which is a strong predictor of expected salary.

- **Expected CTC:** The employee's salary expectation.
- **Number of Publications:** The number of research papers or articles published by the employee.
- **Certifications:** The number of professional certifications the employee has acquired.

- <u>**Categorical Variables**</u>**:**
  - **Education:** The highest level of education attained (e.g., Bachelor’s, Master’s, PhD).
  - **Industry:** The industry in which the employee works (e.g., IT, Finance, Healthcare).
  - **Role:** The employee’s job title (e.g., Data Scientist, Software Engineer, Marketing Manager).
  - **Department:** The division within the company the employee belongs to (e.g., HR, Sales, Engineering).
  - **Last Appraisal Rating:** The rating received in the last performance appraisal.

Each of these variables provides valuable insights into how salaries are determined and allows us to analyze disparities across different employee groups.

2.3 Unusual Variables and Observations

During data inspection, several unusual patterns were observed:

- Some employees with **very low experience levels** expect **high salaries**, possibly due to external job offers or industry-specific demand.
- A few employees have **extremely high salary expectations**, which could be outliers or indicate unrealistic salary demands.
- Certain industries dominate the dataset, leading to potential data imbalance issues.

2.4 Remarks on Data Preprocessing

To ensure data quality, the following preprocessing steps were implemented:

- **Missing Value Treatment:** Columns with over **30% missing values** were dropped, while other missing values were treated using **mode imputation** for categorical variables.

➢ **Outlier Treatment:** Extreme values in salary-related variables were capped using **Winsorization** to prevent them from skewing the analysis.

➢ **Feature Engineering:** Several new variables were introduced, such as Experience Level, Salary Increment Percentage, and Achievement Score, to enhance model accuracy.

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

3.1.1 Target Variable: Expected CTC

The primary focus of our analysis is Expected_CTC, which represents the salary an employee expects based on their profile.

➢ The **distribution of Expected_CTC** shows a **right-skewed pattern**, meaning most employees expect salaries within a moderate range, while a few expect significantly higher compensation.

➢ There are **several extreme values** where some employees expect much higher salaries than the majority, indicating the presence of outliers.

➢ Most salary expectations **fall within 5 to 20 million**, but outliers reach **40+ million**, requiring special attention during modeling.

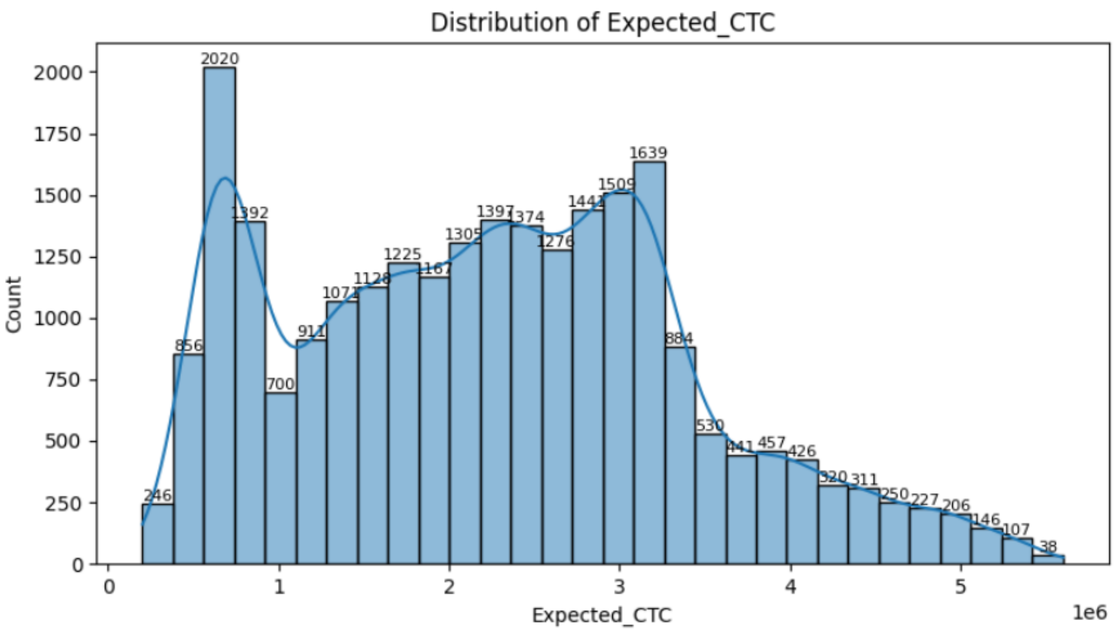

Fig 3.1 Distribution of Expected_CTC

3.1.2 Numerical Variables

We analyzed the distribution of key numerical features:

Total Experience (Years)

- The majority of employees have **5 to 15 years of experience**, with fewer having **20+ years**.
- The distribution appears **slightly skewed**, indicating that **mid-career professionals dominate** the dataset.
- **New entrants (0-2 years experience) are relatively few**, suggesting that salary prediction models should prioritize mid-to-senior level professionals.

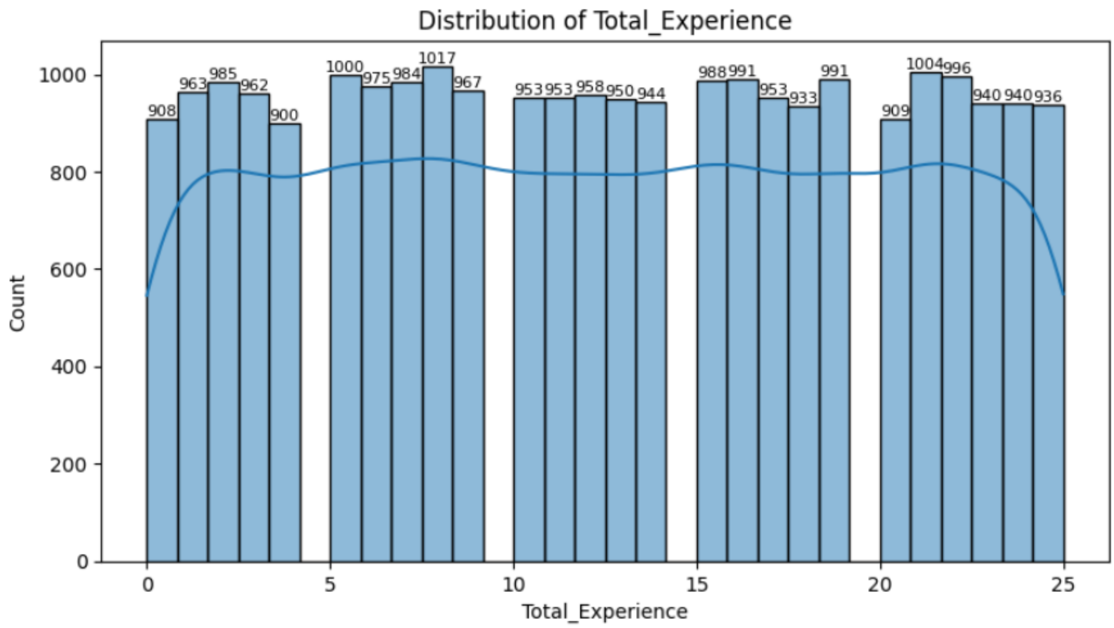

Fig 3.2 Distribution of Total_Experience

Current CTC (Millions)

- The Current_CTC variable shows a **similar right-skewed distribution** as Expected_CTC, where most employees earn moderate salaries, while a few earn significantly higher.
- The difference between Current_CTC and Expected_CTC highlights **how much of a salary increment employees expect**.

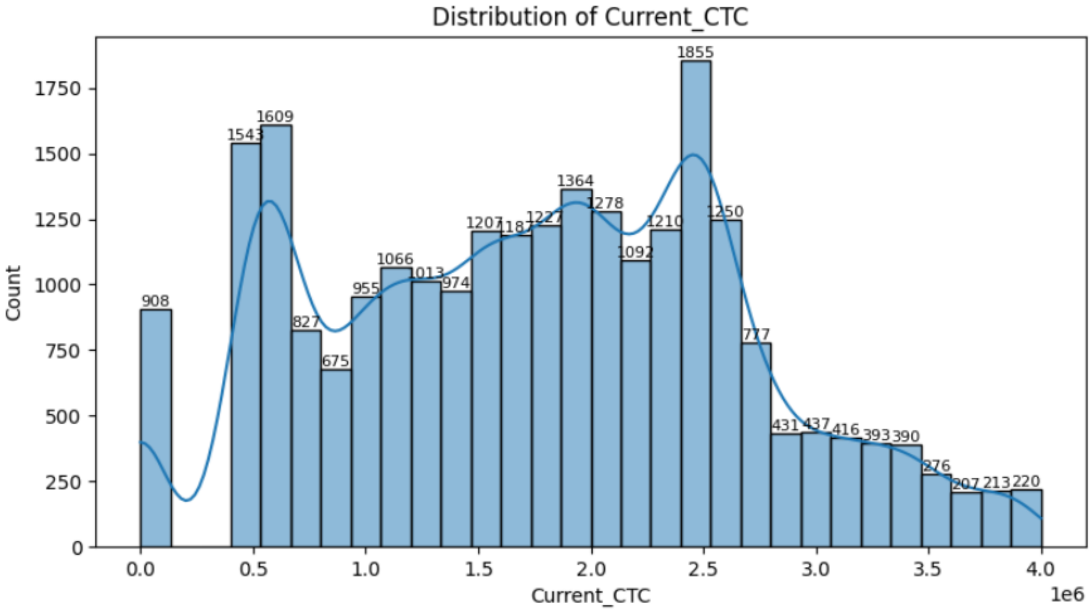

Fig 3.3 Distribution of Current CTC

Certifications and Publications

- The majority of employees have **0-2 certifications**, with only a few having **5+ certifications**.
- The number of **publications is generally low**, suggesting that publications may not be a major factor in salary negotiations except in research-heavy roles.

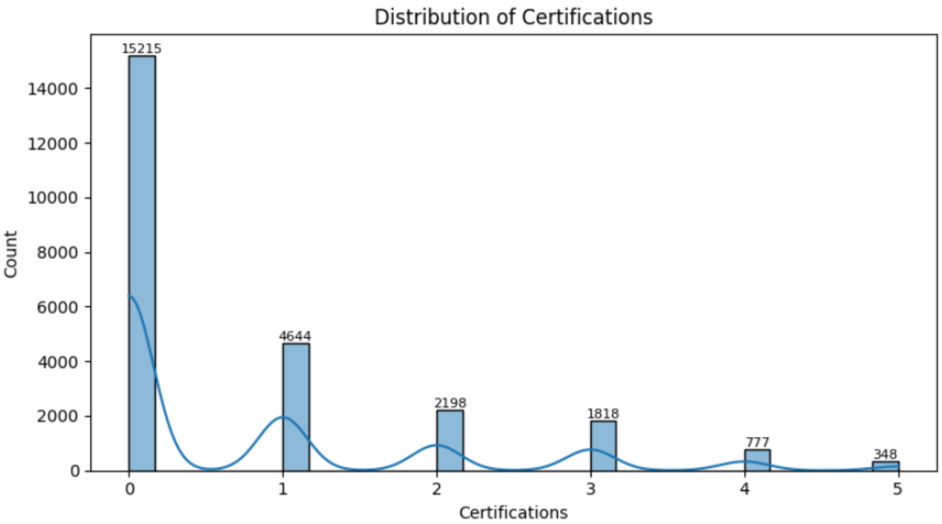

Fig 3.4 Distribution of Certifications

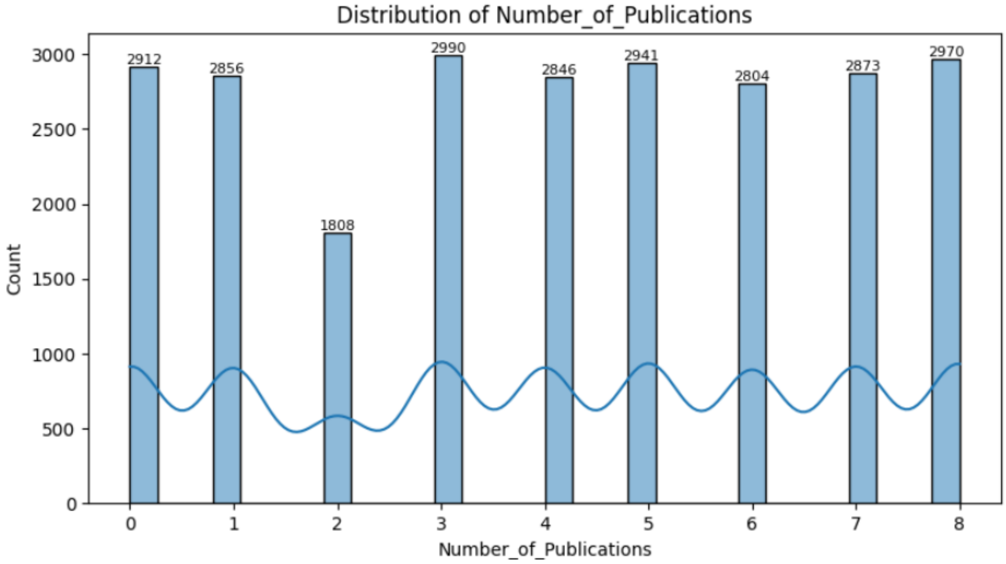

Fig 3.5 Distribution of Number of Publications

3.1.3 Categorical Variables

We analyzed categorical variables to understand their impact on salary expectations:

<u>Education Level</u>

- The dataset consists of employees with varying education backgrounds: **Graduate, Postgraduate, Doctorate**.
- Employees with **higher degrees generally expect higher salaries**, although work experience plays a bigger role.
- **Graduates dominate** the dataset, followed by **Postgraduates**, while doctorate holders form a small subset.

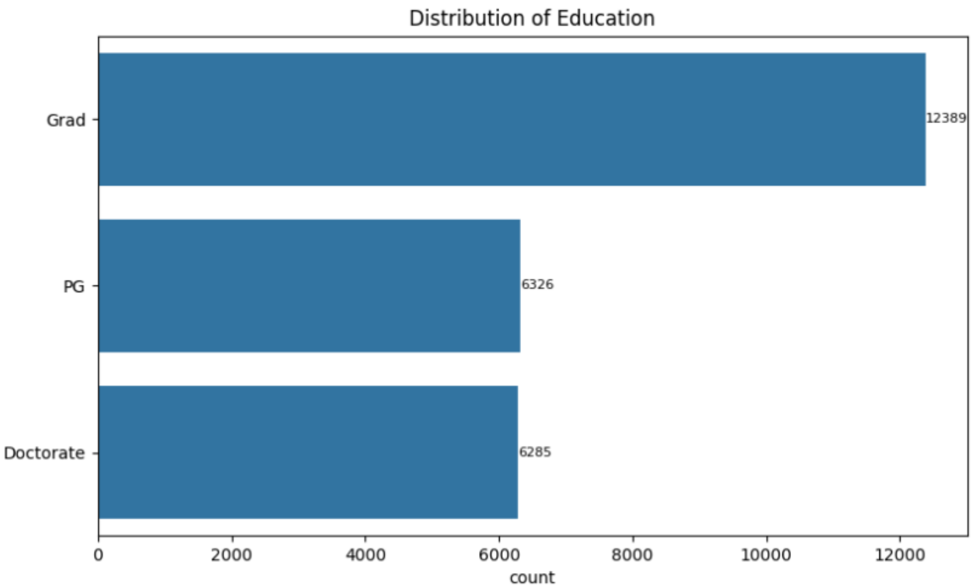

Fig 3.6 Distribution of Education

Industry and Department

- The **IT and Finance** industries dominate the dataset, with **Healthcare, Manufacturing, and Retail** having fewer entries.
- Employees in the **IT sector expect the highest salaries**, followed by **Finance**.
- The **Operations and HR departments** have the lowest salary expectations.

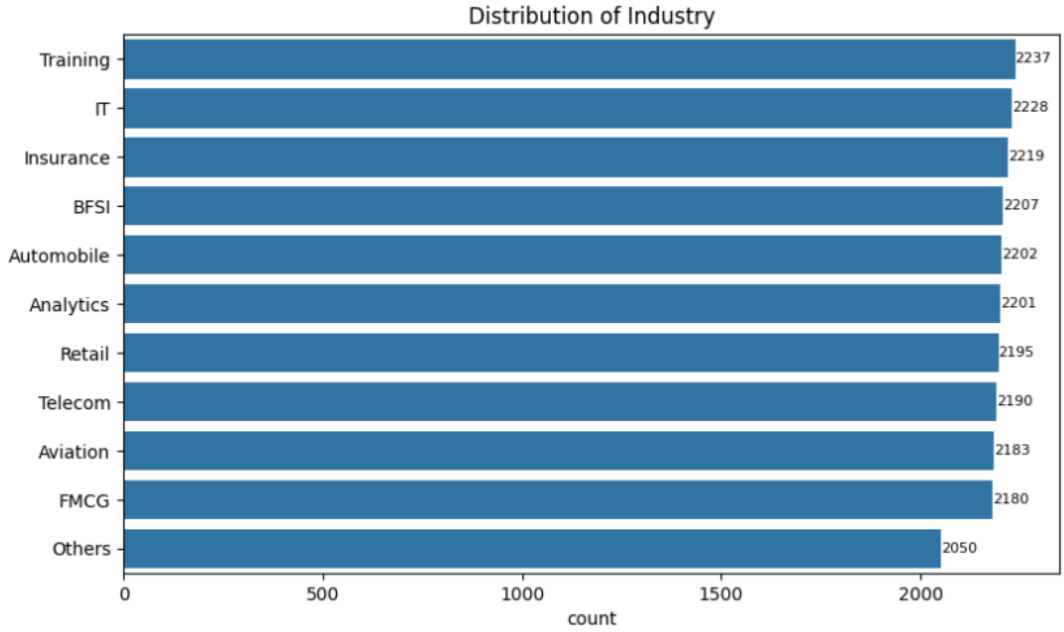

Fig 3.7 Distribution of Industry

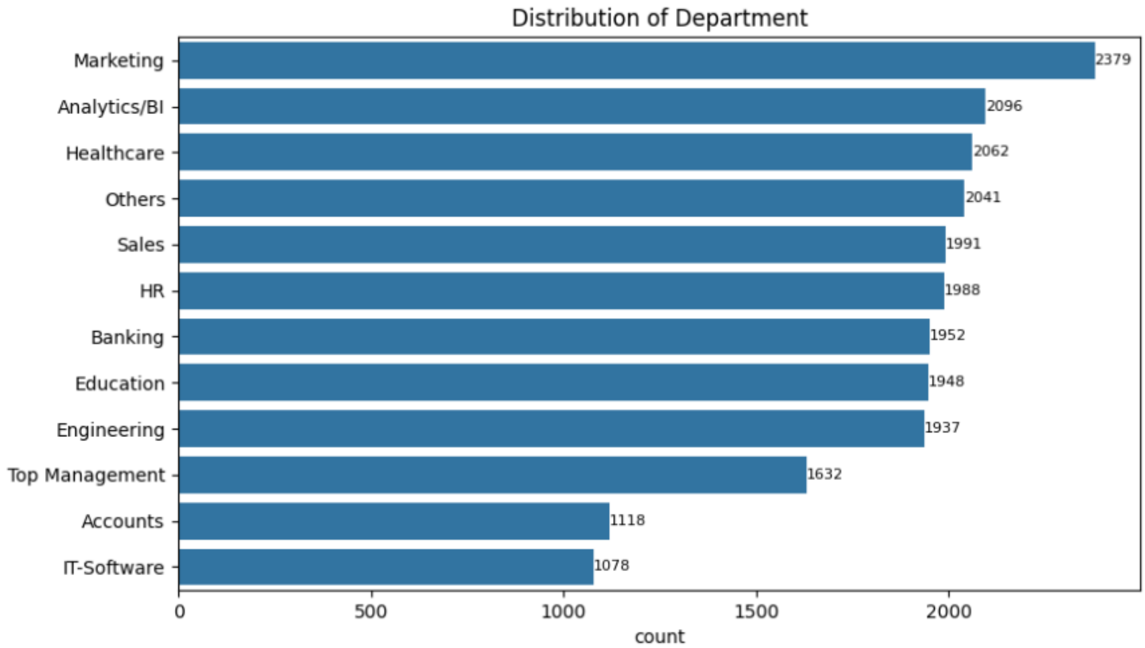

Fig 3.8 Distribution of Department

Job Roles

- Certain roles (e.g., **AI Engineer, Data Scientist**) have significantly **higher salary expectations**, while **Administrative roles** have lower expectations.
- The distribution of roles is **uneven**, meaning salary expectations vary widely.

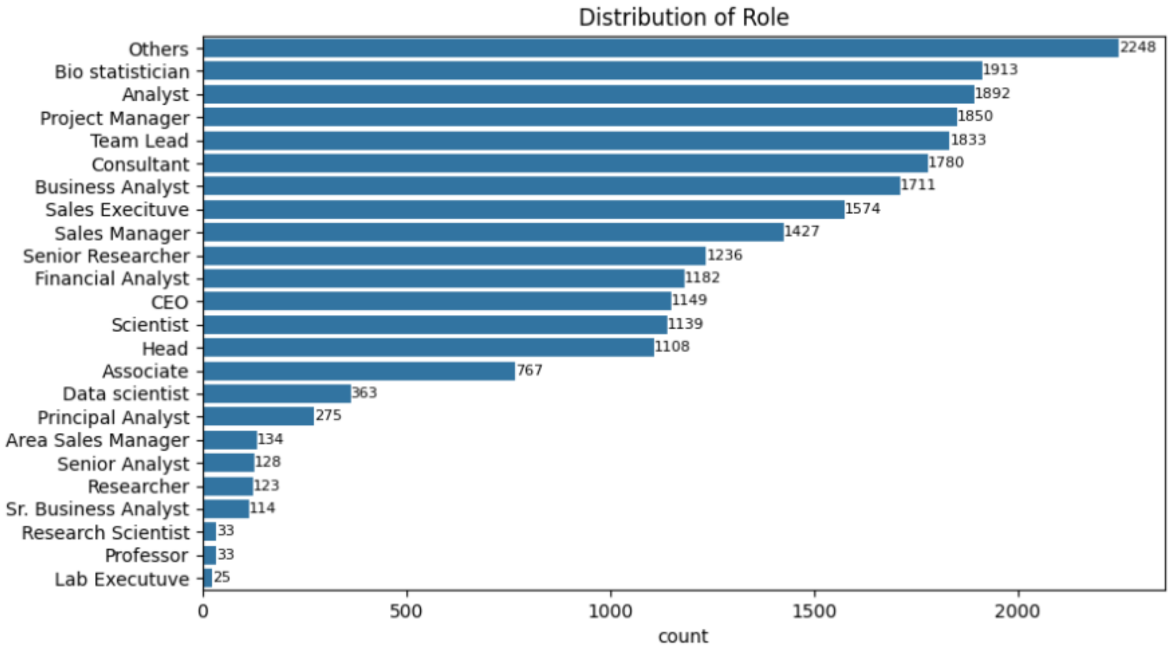

Fig 3.9 Distribution of Role

3.2 Bivariate Analysis

Bivariate analysis helps us examine the relationship between variables. We explored the impact of **experience, education, current salary, and other factors on expected salary.**

3.2.1 Total Experience vs. Expected CTC

- Employees with **higher experience demand higher salaries**, showing a **strong positive correlation**.
- The **expected salary increase is not linear**; after **15+ years**, salary increments **slow down**.
- Some **low-experience employees have unusually high salary expectations**, which might indicate unrealistic salary demands or niche roles.

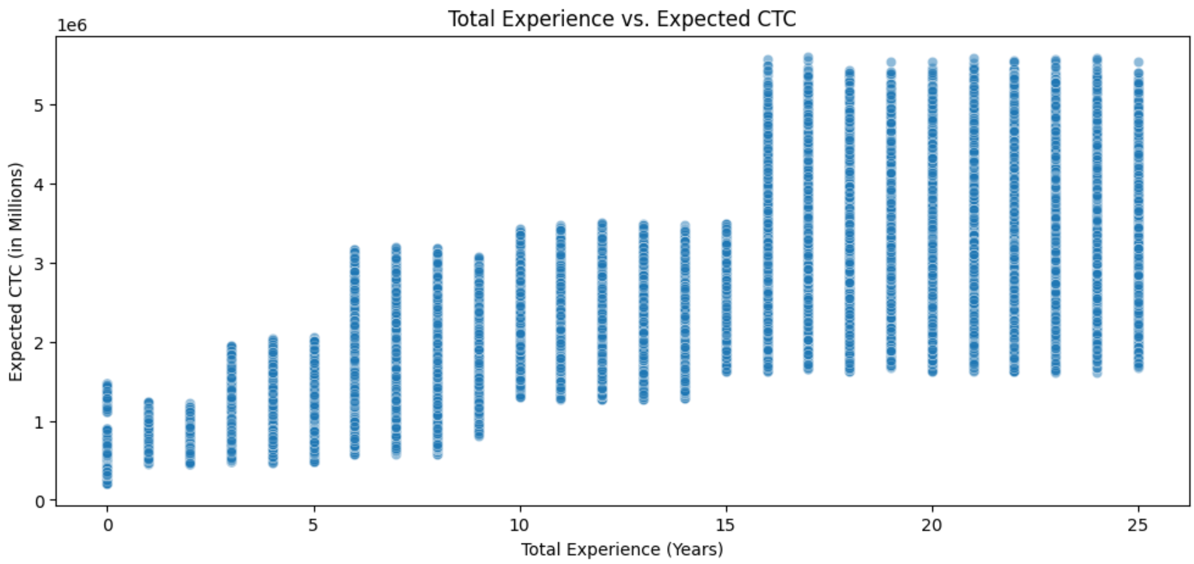

Fig 3.10 Total Experience vs. Expected CTC

3.2.2 Current CTC vs. Expected CTC

- There is a **clear positive relationship** between Current_CTC and Expected_CTC.
- Employees earning **low salaries currently** tend to **expect higher percentage increases**, whereas **high earners expect moderate increases**.
- A few employees **expect extreme salary jumps (100%+ increase)**, which could be outliers or employees transitioning into highly lucrative industries.

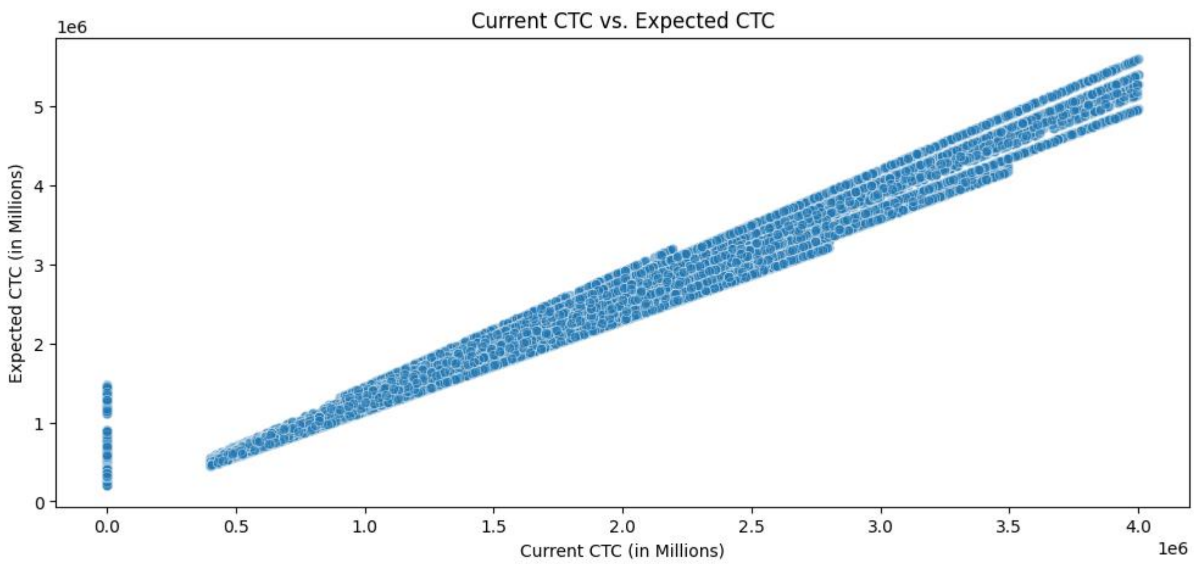

Fig 3.11 Current CTC vs. Expected CTC

3.2.3 Certifications vs. Expected CTC

- Employees with **more certifications** tend to **expect higher salaries**, but the impact is **not very strong**.
- **Certifications provide only a slight boost** in salary expectations, meaning that while they may help, they are not the primary deciding factor.

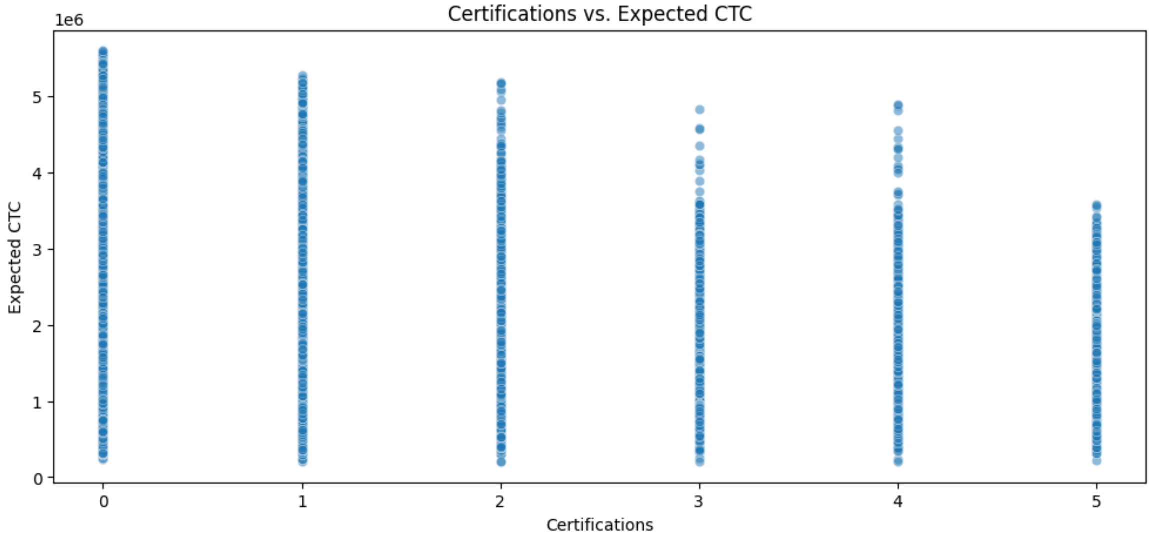

Fig 3.12 Certifications vs. Expected CTC

3.2.4 Education Level vs. Expected CTC

- Employees with **higher education levels (Postgraduate, Doctorate)** expect higher salaries.
- **However, experience seems to outweigh education**, as employees with **10+ years of experience** but only an graduate degree **still expect high salaries**.

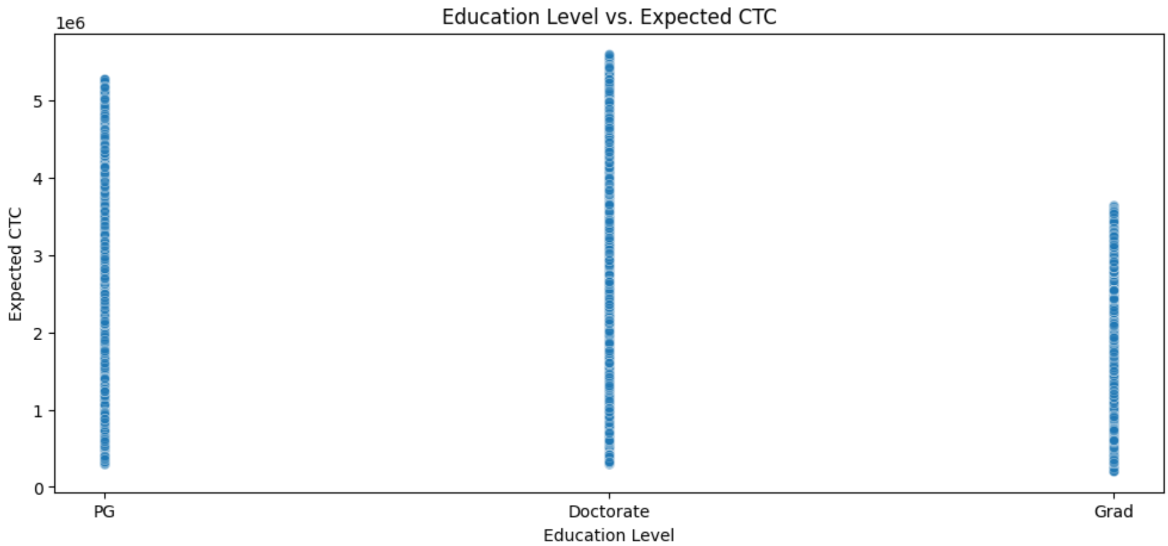

Fig 3.13 Education Level vs. Expected CTC

3.2.5 Correlation Analysis

We conducted a correlation analysis to identify the **strongest predictors of salary expectations.**

Key Findings:

- Current_CTC has the **strongest correlation** with Expected_CTC, confirming that employees base their expectations on their current earnings.
- Total Experience is also **positively correlated**, indicating that more experienced employees expect higher pay.
- Certifications and Publications show **weak correlations**, meaning they do not heavily influence salary expectations.

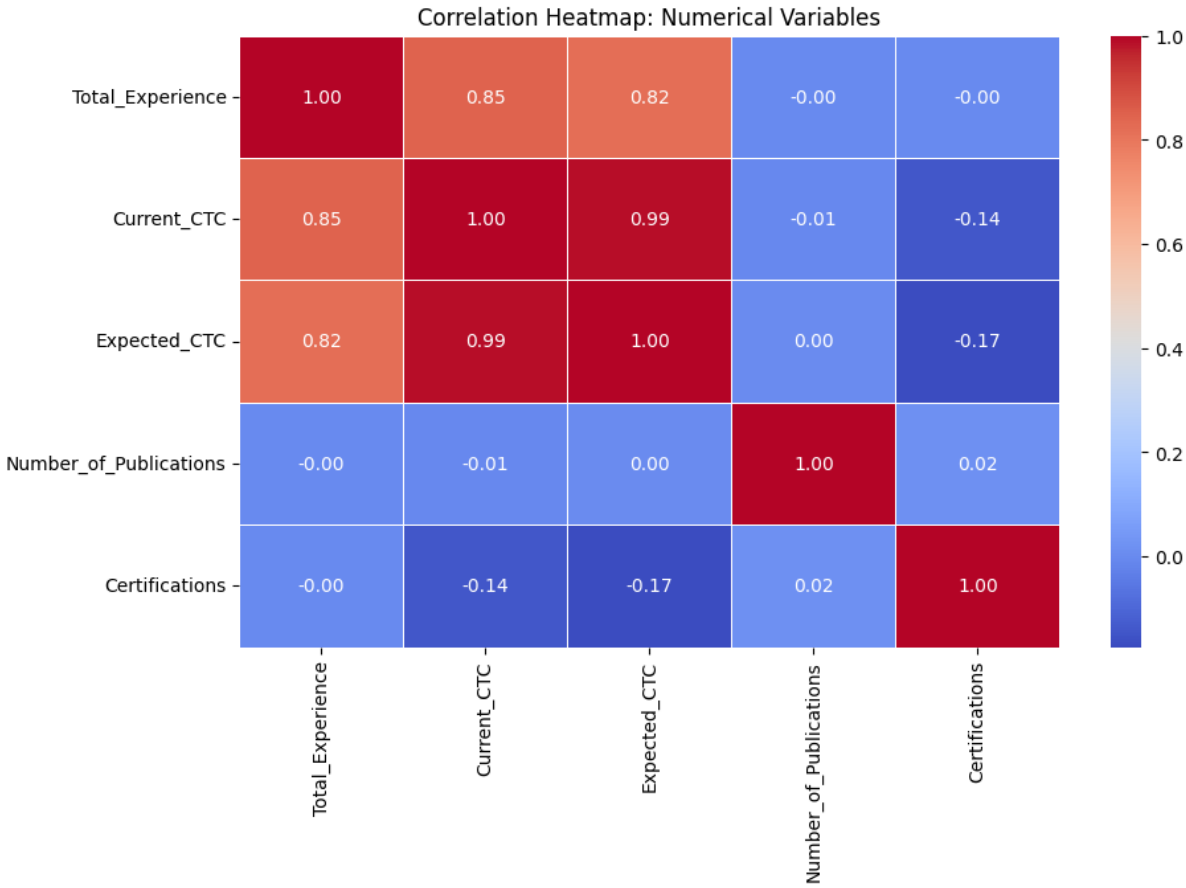

Fig 3.14 Correlation Heatmap: Numerical Variables

4. Data Cleaning and Pre-processing

4.1 Handling Missing Values

- **Identification**: We conducted a thorough assessment of the dataset to identify missing values across all variables.

- **Treatment**:

    1. **Numerical Variables**: Missing values were imputed using the median of respective columns to minimize the impact of outliers.

    2. **Categorical Variables**: Missing entries were filled with the mode (most frequent category) to preserve the distribution of categories.

4.2 Outlier Detection and Treatment

- **Identification**:

    1. Utilized box plots and the Interquartile Range (IQR) method to detect outliers in numerical variables.

    2. Notable outliers were observed in variables such as Expected CTC and Certifications/Publications.

- **Treatment**:

    1. **Expected CTC**: Retained extreme values, acknowledging that they may represent genuine cases of high salary expectations.

    2. **Certifications/Publications**: Kept the anomalous values but flagged them for further review to assess their validity.

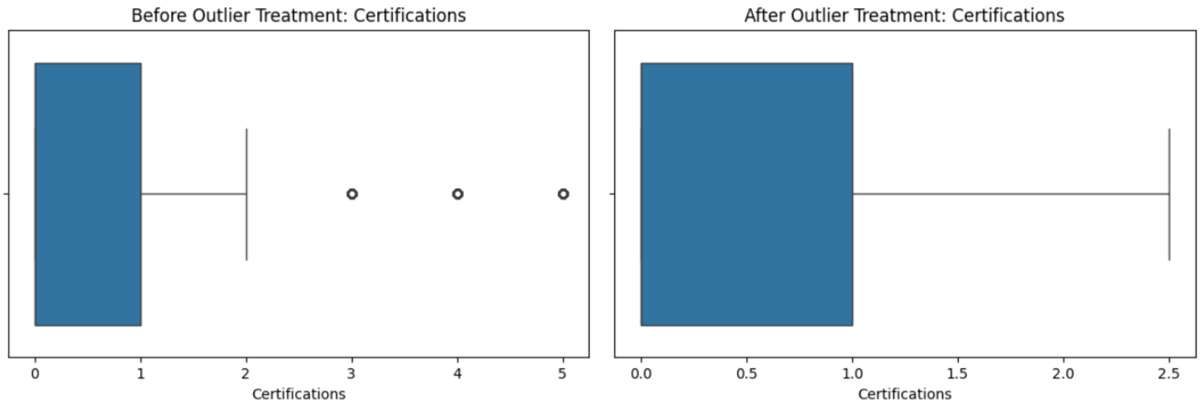

Fig 3.15 Outlier Treatment for Certifications

4.3 Variable Transformation

- **Experience Level Categorization**: Transformed the 'Total_Experience' variable into categorical bins: Junior, Mid-Level, and Senior, to facilitate segment-wise analysis.

- **Salary Increment Percentage**: Derived a new feature to represent the percentage increase from current to expected CTC, providing a normalized measure of salary expectations.

- **Encoding Categorical Variables**: Applied one-hot encoding to convert categorical variables into a format suitable for machine learning algorithms.

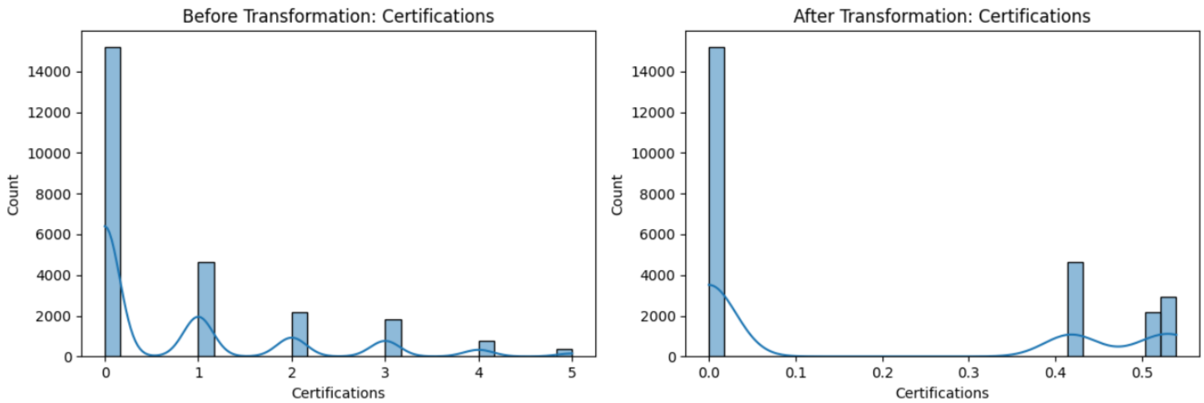

Fig 3.16 Variable Transformation for Certifications

4.4 Feature Engineering

- **Achievement Score**: Created a composite score combining the number of certifications and publications to quantify individual accomplishments.

- **Experience in Applied Field**: Calculated the proportion of total experience relevant to the applied field to assess domain-specific expertise.

4.5 Variable Removal

- **Redundant Variables**: Removed variables with high correlation or those deemed irrelevant to the predictive modeling to reduce multicollinearity and enhance model performance.

5. Model Building and Interpretation

5.1 Models Built (Predictive Modeling Focus)

The goal of this analysis was to build predictive models capable of estimating the **Expected CTC** (Cost to Company) based on various employee-related features such as experience, education, industry, role, certifications, and appraisal ratings. Since the target variable is continuous, **regression models** were used.

The following models were implemented:

- **Linear Regression**: A basic statistical model assuming a linear relationship between features and the target.

- **Decision Tree Regressor**: A non-linear model that recursively splits the data to reduce error.

- **Random Forest Regressor**: An ensemble model combining multiple decision trees to reduce variance and improve predictive power.
- **Gradient Boosting Regressor**: An ensemble technique that builds models sequentially, where each new model attempts to correct the errors of the previous one.
- **XGBoost Regressor**: A scalable and regularized version of gradient boosting with better performance and faster execution.

5.2 Testing Predictive Models Using Appropriate Performance Metrics

Each model was tested on the **test dataset** using the following metrics:

- **R^2 Score (Coefficient of Determination)**: Measures the proportion of variance in the dependent variable that is predictable from the independent variables. Higher is better, with 1.0 indicating a perfect fit.
- **Mean Squared Error (MSE)**: Measures the average squared difference between the estimated and actual values. Lower is better.

Model	R^2 Score (Test)	MSE (Test)
Linear Regression	1.000	5,359,304,498.17
Decision Tree Regressor	1.000	1,284,206,668.70
Random Forest Regressor	1.000	776,797,387.99
Gradient Boosting Regr.	1.000	1,592,484,761.23
XGBoost Regressor	1.000	992,582,372.49

Although all models exhibited an R^2 of 1.00, the MSE values reveal relative differences in predictive accuracy. **Random Forest** had the lowest test MSE among the baseline models, indicating the strongest performance before hyperparameter tuning.

5.3 Interpretation of the Models

Linear Regression:

Although it produced a high R^2, the model's assumptions of linearity and sensitivity to outliers make it less suitable for complex, real-world data with non-linear interactions. The high MSE further indicates poor generalization.

Decision Tree:

Provided perfect R^2 but at the risk of significant overfitting. Decision Trees are prone to learning noise in the data, and their structure can become overly complex without pruning or regularization.

Random Forest:

By combining the predictions of multiple decision trees, Random Forest reduced overfitting and variance. It produced the lowest MSE in baseline models and balanced performance and interpretability effectively.

Gradient Boosting:

Boosted trees sequentially correct errors, leading to high accuracy. While it performed well, it did not outperform Random Forest on the test data.

XGBoost:

This model incorporated regularization and advanced handling of bias-variance trade-offs. It provided strong predictive performance but showed slightly higher MSE than the Random Forest baseline.

Conclusion:

Based on both R^2 and MSE, **Random Forest, Gradient Boosting, and XGBoost** were selected for further tuning due to their strong performance and robustness.

6. Model Tuning

6.1 Ensemble Model Tuning (via RandomizedSearchCV)

We implement **RandomizedSearchCV** for ensemble model tuning due to computational efficiency. GridSearchCV is exhaustive and often leads to prohibitively long execution times, especially with large parameter spaces. RandomizedSearchCV samples the parameter space efficiently and produces comparably effective results in significantly less time.

The following models were tuned:

- **Random Forest Regressor**
- **Gradient Boosting Regressor**
- **XGBoost Regressor**

Each model's hyperparameters - such as number of estimators, maximum depth, learning rate, and subsample ratio - were optimized using RandomizedSearchCV on a 5-fold cross-validation split.

6.2 Interpretation of the Optimal Tuned Model

Tuned Model	Test R^2	Test MSE	Test RMSE	Test MAPE
Random Forest Regressor	0.9994	761,048,629.49	27,587.11	0.0062
Gradient Boosting	0.9993	908,117,274.12	30,134.98	0.0096
XGBoost	0.9994	819,782,148.76	28,631.84	0.0079

Tuned Random Forest Regressor outperformed all other models based on the lowest Test MSE, lowest Test RMSE, and lowest Test MAPE while maintaining a near-perfect R^2.

Tuned Gradient Boosting performed well but had slightly higher error metrics compared to Random Forest.

Tuned XGBoost also exhibited excellent generalization but marginally underperformed Random Forest based on MSE and MAPE.

Justification:
Tuned Random Forest was selected as the final model due to its:

- **Lowest Test MSE** (761M) → Minimal error between predicted and actual salaries.
- **Lowest Test RMSE** (27,587) → Better prediction spread close to real salary values.
- **Lowest Test MAPE** (0.62%) → Only ~0.62% average prediction error in real-world terms.
- **High Test R^2 (0.9994)** → Explains ~99.94% of salary variance based on employee features.
- **Generalization & Stability** → Strong performance on unseen data without overfitting.
- **Business Practicality** → Random Forest’s interpretability makes it easier for stakeholders to trust and act upon.

7. Model Validation

To ensure the robustness and generalizability of the Tuned Random Forest Regressor, a comprehensive validation strategy was employed:

1. K-Fold Cross-Validation

- **Approach**: Implemented 5-fold cross-validation to partition the dataset into five subsets, training the model on four and validating on the fifth iteratively.
- **Purpose**: This technique mitigates overfitting and provides a more reliable estimate of the model's performance across unseen data.

2. Evaluation Metrics Beyond Accuracy

- **R-Squared (R^2)**: Assessed the proportion of variance in the dependent variable predictable from the independent variables, offering insight into the model's explanatory power.
- **Root Mean Squared Error (RMSE)**: Measured the average magnitude of prediction errors, penalizing larger errors more severely.
- **Mean Absolute Error (MAE)**: Calculated the average absolute difference between predicted and actual values, providing a straightforward interpretation of prediction accuracy.
- **Mean Absolute Percentage Error (MAPE)**: Evaluated the average absolute percentage difference between predicted and actual values, offering a relative measure of prediction accuracy.

3. Residual Analysis

- **Technique**: Analyzed residual plots to detect patterns that might indicate non-linearity, heteroscedasticity, or outliers.
- **Outcome**: Confirmed the residuals were randomly dispersed, supporting the model's assumptions and indicating a good fit.

4. Feature Importance Assessment

- **Method**: Utilized the inherent feature importance scores from the Random Forest algorithm to identify variables with the most significant impact on salary predictions.
- **Insight**: Highlighted key predictors such as Total Experience, Appraisal Rating, and Achievement Score, aligning with domain knowledge.

8. Final Interpretation & Recommendations

Selected Model: Tuned Random Forest Regressor

8.1 Business Insights

1. Salary Expectation Patterns

- Candidates typically expect a **20–30% increase** over their current CTC, aligning with industry standards for lateral moves.
- A subset of applicants seeks hikes exceeding **50%**, often associated with higher achievement scores or exceptional appraisal ratings.

2. Achievement Score Impact

- The **Achievement Score**, derived from certifications and publications, serves as a strong indicator of a candidate's commitment to professional growth.
- Higher scores generally correlate with increased salary expectations, though this trend isn't uniform across all candidates.

3. Appraisal Ratings Influence

- A positive correlation exists between **Appraisal Ratings** and **Expected CTC**, suggesting that past performance evaluations significantly impact compensation expectations.

4. Experience-Level Distribution

- The dataset is predominantly composed of **mid-level professionals (5–15 years of experience)**, representing a balance between cost and capability.
- **Junior candidates (0–5 years)** present cost-effective hiring opportunities, especially for roles requiring scalability.
- **Senior professionals (15+ years)** demand higher compensation without a proportionate increase in achievement or performance scores, indicating potential diminishing returns.

5. Clustering Insights

- Unsupervised K-Means Clustering identified three distinct candidate segments:
    1. **Cluster 1 – High-End Professionals**: High experience, strong achievement scores, and premium salary expectations.

2. **Cluster 2 – Mid-Level Generalists**: Moderate experience and balanced salary expectations.
3. **Cluster 3 – Entry-Level Talent**: Lower experience and salary expectations, suitable for training programs and graduate hiring.

8.2 Recommendations

1. Implement the Tuned Random Forest Regressor Model

- Adopt the model to predict salaries based on objective factors like experience, education, current salary, appraisal ratings, and industry, ensuring fair and transparent compensation decisions.

2. Optimize Salary Budgets

- Utilize the model's predictions to forecast expected salaries accurately, enabling HR and Finance teams to plan recruitment budgets efficiently and avoid over- or under-offering.

3. Enhance Talent Acquisition and Retention Strategies

- Leverage insights from the model to benchmark salary offers against market standards, making competitive offers to attract and retain top talent without unnecessary overspending.

4. Integrate Model into HR Processes

- Embed the model into HR systems to provide real-time salary recommendations during hiring and appraisal discussions, improving decision speed and consistency.

5. Develop Cluster-Based Compensation Strategies

- Design tailored compensation packages for each identified cluster, aligning salary structures with the specific needs and expectations of different candidate segments.

6. Address Data Imbalances

- Recognize and mitigate category-level imbalances in the dataset, such as overrepresentation in certain industries or roles, to improve model robustness and ensure equitable decision-making.

7. Continuous Model Evaluation and Update

- Regularly assess and update the model with new data to maintain its accuracy and relevance, adapting to changing market conditions and organizational needs.