Salary_Prediction_Project
Capstone - PPT

RAGHAVENDRA KUMAR J R

PGP – DSBA ONLINE APRIL’23 BATCH

MENTOR: MR. ABHAY PODDAR

Table of Contents

Business Problem

- To ensure there is no discrimination between employees, it is imperative for the Human Resources department of Delta Ltd. to maintain a salary range for each employee with similar profiles Apart from the existing salary, there is a considerable number of factors regarding an employee's experience and other abilities to which they get evaluated in interviews. Given the data related to individuals who applied in Delta Ltd, models can be built that can automatically determine salary which should be offered if the prospective candidate is selected in the company. This model seeks to minimize human judgment with regard to salary to be offered.
- **Goal & Objective**: The objective of this exercise is to build a model, using historical data that will determine an employee's salary to be offered, such that manual judgments on selection are minimized. It is intended to have a robust approach and eliminate any discrimination in salary among similar employee profiles.

Problem Understanding

- Delta Ltd. is creating a predictive model that standardizes salary determination in an effort to address potential biases and inconsistencies in salary offerings. The objective of the salary negotiation process is to reduce subjective judgment by employing historical data on the salaries, experience, qualifications, and other pertinent factors of employees. Through reducing the possibility of discrimination and improving hiring process transparency, this model seeks to guarantee fair, equitable, and competitive compensation packages for all employees.
- Scope:

Data Collection and Analysis: The scope of this project involves gathering historical data on employees' profiles, including their experience, skills, and existing salaries. This data will be analyzed to identify patterns and relationships between different variables and salary levels.

Model Development: The project will include the development of predictive models that utilize machine learning algorithms to predict the salary range for prospective employees based on their profiles. These models will consider various factors such as education, experience, skills, and job role.

- Objectives:

Minimize Human Bias: The primary objective is to reduce or eliminate human bias in the salary determination process by leveraging data-driven models. By relying on objective data rather than subjective judgments, the model aims to ensure fairness and equality in salary offers.

Ensure Consistency: Another objective is to establish consistency in salary offers for employees with similar profiles. The model will provide a standardized approach to salary determination, ensuring that employees with comparable qualifications and experience receive equitable compensation packages.

EDA Analysis

- The dataset contains 25,000 rows and 29 columns
- Statistical Summary - Key numerical features include:
- Total_Experience: Ranges from 0 to 25 years, with a mean of approximately 12.49 years.
- Current_CTC: The current salary ranges from 0 to approximately 3.999 million, with a mean of about 1.76 million.
- Expected_CTC: The target variable ranges from around 203,744 to 5.599 million, with a mean of approximately 2.25 million.
- Number_of_Publications and Certifications also show a wide range, indicating diverse academic and professional achievements among applicants.

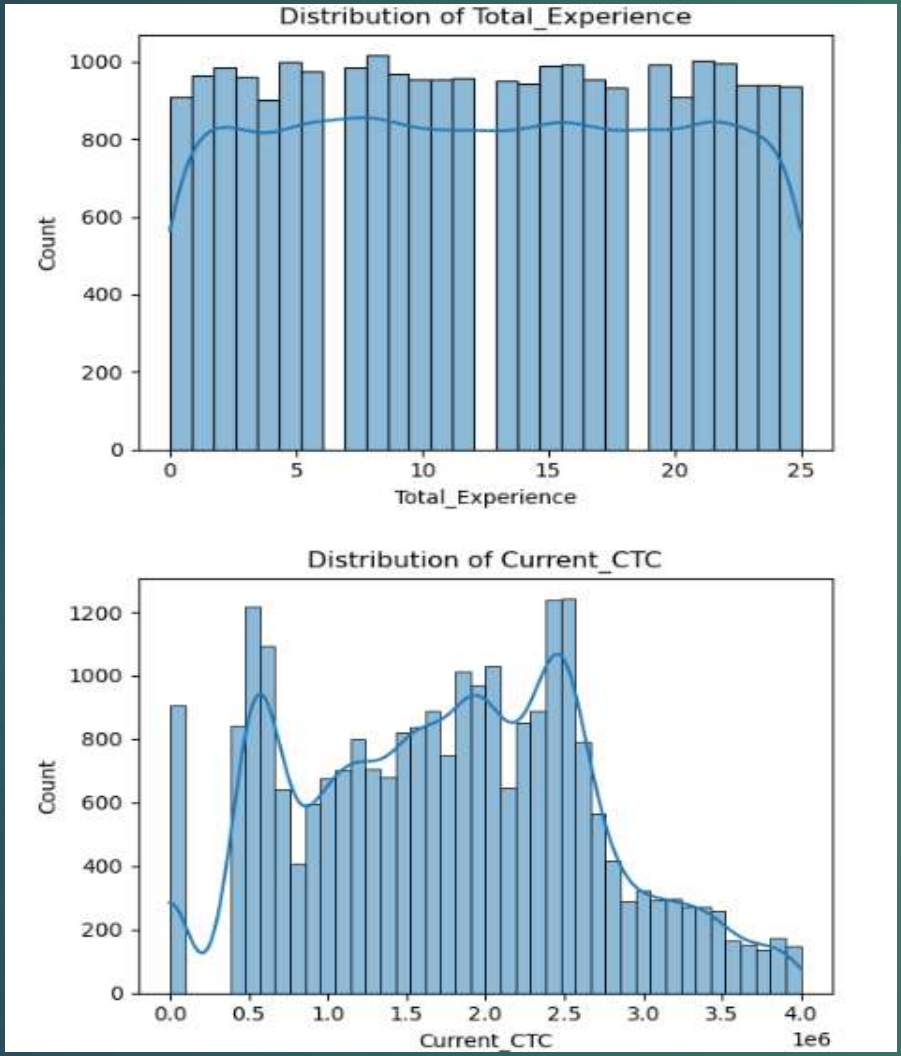
Univariate analysis and Bivariate analysis

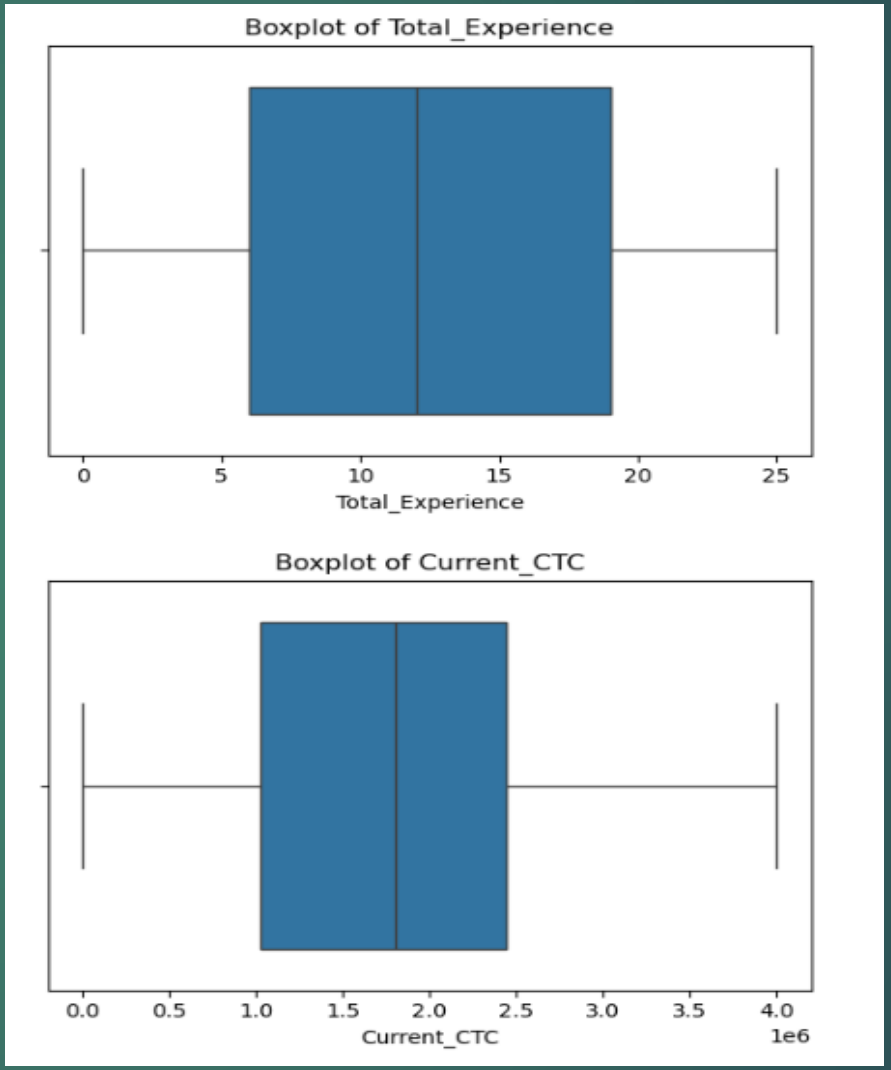
Univariate analysis:

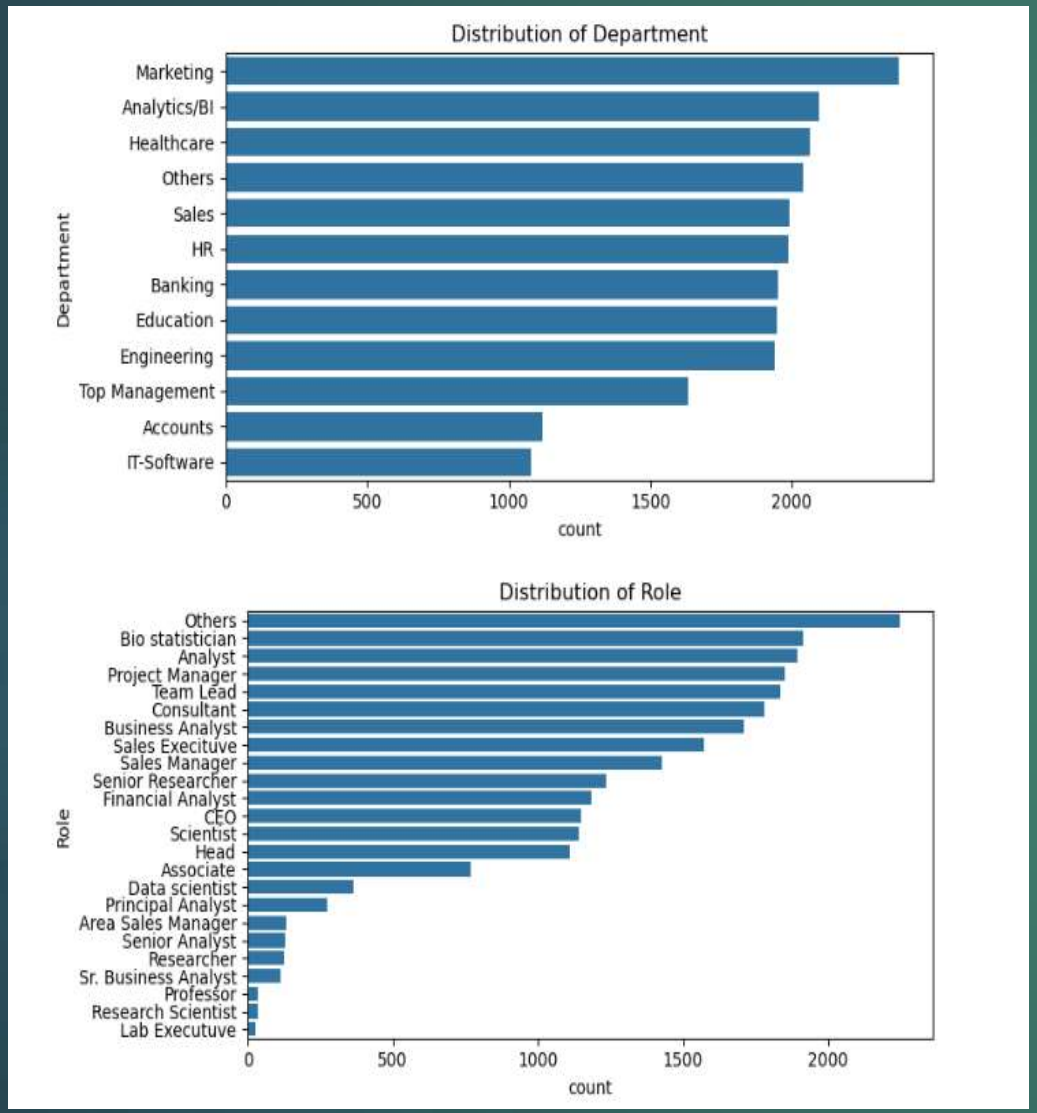
- Univariate analysis focuses on analyzing and summarizing data for a single variable at a time.
- It involves examining the distribution, central tendency, variability, and other descriptive statistics of a single variable.
- For example, if you're analyzing the ages of people in a population, univariate analysis would involve looking at the distribution of ages, calculating measures like the mean, median, and mode, and assessing the spread or variability of ages

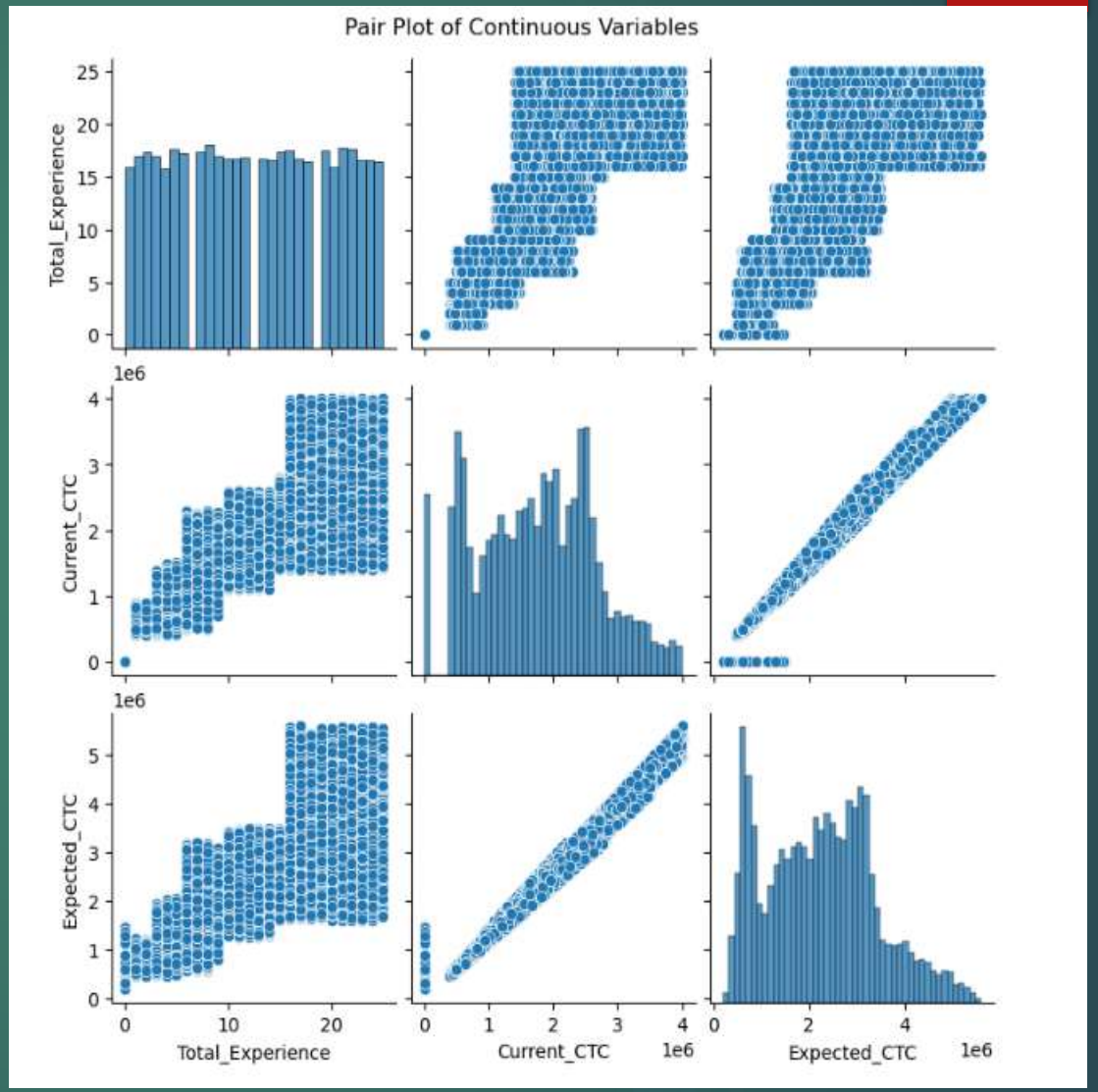
Bivariate Analysis:

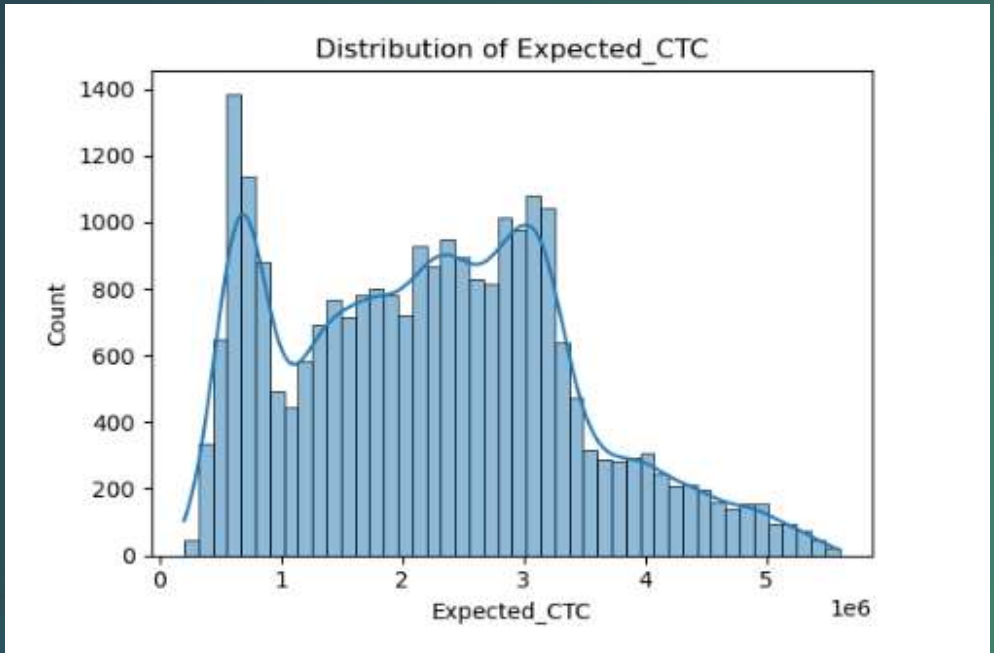
- Bivariate analysis involves analyzing the relationship between two variables simultaneously.
- It explores how changes in one variable are associated with changes in another variable.
- Common techniques in bivariate analysis include correlation analysis, scatter plots, and contingency tables.
- For example, if you're studying the relationship between temperature and ice cream sales, bivariate analysis would involve examining whether there is a correlation between higher temperatures and increased ice cream sales.

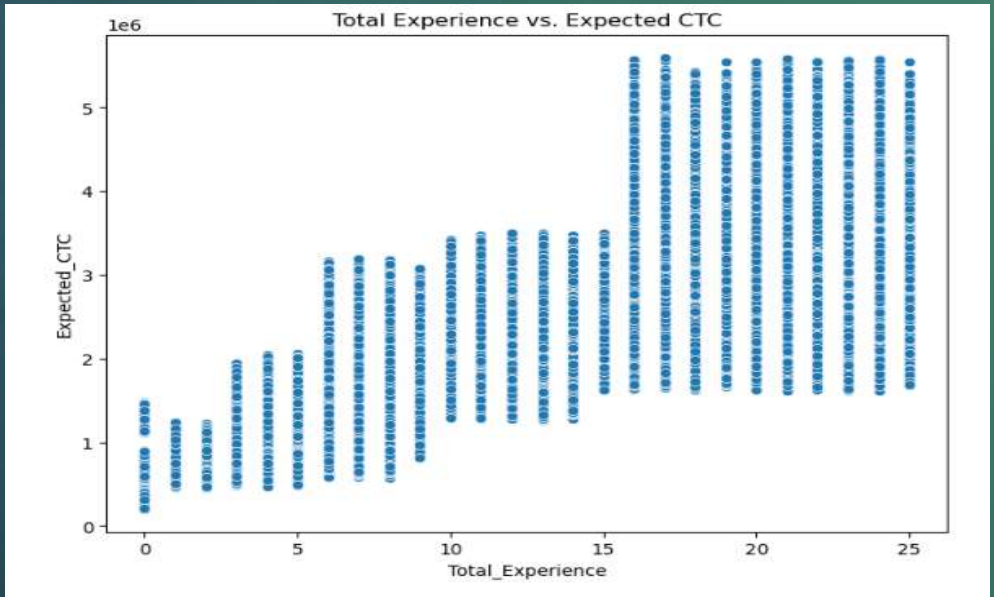

Distribution of Total_Experience
Count
Total_Experience
Distribution of Current_CTC
Count
Current_CTC
1e6


Boxplot of Total_Experience
Total_Experience
Boxplot of Current_CTC
Current_CTC
1e6


Distribution of Department
Marketing
Analytics/BI
Healthcare
Others
Sales
HR
Banking
Education
Engineering
Top Management
Accounts
IT-Software
Department
0
500
1000
1500
2000
count
Distribution of Role
Others
Bio statistician
Analyst
Project Manager
Team Lead
Consultant
Business Analyst
Sales Executive
Sales Manager
Senior Researcher
Financial Analyst
CEO
Scientist
Head
Associate
Data scientist
Principal Analyst
Area Sales Manager
Senior Analyst
Researcher
Sr. Business Analyst
Professor
Research Scientist
Lab Executuve
Role
0
500
1000
1500
2000
count


Pair Plot of Continuous Variables
Total_Experience
Current_CTC
Expected_CTC
1e6
0
5
10
15
20
25
0
1
2
3
4
0
10
20
Total_Experience
Current_CTC
Expected_CTC





- A thorough understanding of the relationships between variables is offered by these visualizations. The heatmap highlights the most strongly correlated pairs of variables and provides a brief summary of correlation coefficients. The scatter plots for variable pairs and the histograms for individual variables are visible in the pair plot, which provides a more thorough examination and is helpful in identifying non-linear relationships or variables that could benefit from transformations.
- Recall that correlation does not imply causation, and more research may be required to completely comprehend the nature of these relationships. In order to preserve the pair plot's interpretability and clarity when working with a large number of continuous variables, it may be helpful to concentrate on a small number of important variables.

Heat Map

- A heatmap is a data visualization technique that uses color-coded cells to represent the values of a matrix or a table. It is particularly useful for visualizing relationships or patterns in large datasets.
- **Color Coding**: Each cell in the heatmap is assigned a color based on its value. Typically, a gradient color scheme is used, where different shades of color represent different values.
- **Matrix Representation**: Heatmaps are commonly used to represent matrices or tables of data. The rows and columns of the matrix represent different variables or categories, and the cells contain the values of interest.
- clusters of cells with similar colors may indicate groups of variables that are correlated with each other.
- They can be used to visualize correlations, trends, distributions, and other patterns in the data.

Lets see the below models used for project:

- Linear Regression Model
- Decision Tree Regression
- Lasso Regression Model
- Ridge Regression Model
- KNN Model
- Bagging Regression Model
- Random Forest Model
- Gradient Regression Model

Linear Regression Model:

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features),The goal of linear regression is to find the best-fitting line (or plane in higher dimensions) that minimizes the difference between the observed values and the values predicted by the model.

Linear regression is widely used for several reasons:

1. **Interpretability**: The coefficients in a linear regression model represent the relationship between the independent variables and the dependent variable in a straightforward manner. This makes it easy to interpret the impact of each independent variable on the dependent variable.
2. **Simplicity**: Linear regression is a simple and computationally efficient model, making it easy to implement and understand.
3. **Predictive Power**: Despite its simplicity, linear regression can be surprisingly powerful, especially when there is a linear relationship between the independent and dependent variables.
4. **Baseline Model**: Linear regression often serves as a baseline model for more complex machine learning algorithms. It provides a simple benchmark against which the performance of more advanced models can be compared.
5. **Assumption Testing**: Linear regression allows for the testing of assumptions about the relationship between variables, such as linearity, homoscedasticity, and independence of errors.

The linear regression model demonstrates strong performance, as evidenced by the following metrics:

- **Mean Squared Error (MSE)**: The MSE, a measure of the average squared difference between the actual and predicted values, is approximately 34.48 billion (34,476,113,176.02). This value indicates that, on average, the squared difference between the predicted and actual values is relatively low, signifying good accuracy in the model's predictions.
- **Root Mean Squared Error (RMSE)**: The RMSE, calculated as the square root of the MSE, is approximately 185,677.44. This value provides an understanding of the average magnitude of the errors in the model's predictions. A lower RMSE suggests better accuracy, and the obtained value indicates relatively small errors in prediction.
- **R-squared (R^2)**: The R-squared value, a measure of how well the independent variables explain the variance in the dependent variable, is approximately 0.97 (97.43%). This high R-squared value indicates that approximately 97.43% of the variance in the dependent variable is explained by the independent variables, implying a strong fit between the observed and predicted values.

Overall, these performance metrics collectively suggest that the linear regression model is effective in making accurate predictions and explaining the relationship between the variables in the dataset.

Decision Tree Model

- A decision tree model is a predictive modeling approach that uses a tree-like structure to make decisions based on input features.
- Decision trees are used for classification and regression tasks and are known for their simplicity, interpretability, and ability to handle both numerical and categorical data.
- Decision tree models are used for several reasons:
  1. **Interpretability**: Decision trees are easy to interpret and visualize, making them useful for understanding the decision-making process.
  2. **Versatility**: Decision trees can handle both numerical and categorical data, as well as multi-class classification and regression tasks.
  3. **Non-linear Relationships**: Decision trees can capture non-linear relationships between features and the target variable without requiring feature engineering.
  4. **Automatic Feature Selection**: Decision trees perform automatic feature selection by choosing the most informative features at each split.
  5. **Robustness**: Decision trees are robust to outliers and missing values, making them suitable for noisy datasets.

The decision tree model demonstrates strong performance, as indicated by the following metrics:

- **Mean Squared Error (MSE)**: The MSE, which measures the average squared difference between the actual and predicted values, is approximately 38.64 billion (38,638,503,261.29). This value suggests that, on average, the squared difference between the predicted and actual values is relatively low, indicating good accuracy in the model's predictions.
- **Root Mean Squared Error (RMSE)**: The RMSE, calculated as the square root of the MSE, is approximately 196,566.79. This value provides an understanding of the average magnitude of the errors in the model's predictions. A lower RMSE suggests better accuracy, and the obtained value indicates relatively small errors in prediction.
- **R-squared (R^2)**: The R-squared value, a measure of how well the independent variables explain the variance in the dependent variable, is approximately 0.97 (97.12%). This high R-squared value indicates that approximately 97.12% of the variance in the dependent variable is explained by the independent variables, implying a strong fit between the observed and predicted values.

Lasso regression Model

- Lasso regression, short for Least Absolute Shrinkage and Selection Operator, is a regression technique that performs both variable selection and regularization to improve the prediction accuracy and interpretability of the model.
- It works by adding a penalty term to the traditional least squares objective function, which penalizes the absolute size of the coefficients of the regression variables. This penalty encourages simpler models with fewer features by shrinking some coefficients to zero, effectively performing feature selection.
- Lasso regression is particularly useful when dealing with datasets with a large number of features, as it helps to identify the most relevant features while reducing overfitting.
- It is commonly used in situations where there may be multicollinearity among the independent variables, as it can effectively handle this issue by shrinking less important variables' coefficients to zero.

Lasso regression is used for several reasons:

- **Feature Selection**: It automatically selects the most relevant features by setting some coefficients to zero, leading to simpler and more interpretable models.
- **Regularization**: Lasso regression applies regularization to prevent overfitting and improve generalization performance.
- **Handling Multicollinearity**: It is effective in handling multicollinearity by selecting one feature among highly correlated ones.
- **Improving Model Accuracy**: Lasso regression can improve the model's prediction accuracy by focusing on the most informative features and reducing the impact of less relevant ones.

Lasso Regression Model Performance:

- **Mean Squared Error (MSE)**: The MSE, a measure of the average squared difference between the actual and predicted values, is approximately 34.48 billion (34,476,108,782.81). This value suggests that, on average, the squared difference between the predicted and actual values is relatively low, indicating good accuracy in the model's predictions.
- **R-squared (R^2)**: The R-squared value, a measure of how well the independent variables explain the variance in the dependent variable, is approximately 0.97 (97.43%). This high R-squared value indicates that approximately 97.43% of the variance in the dependent variable is explained by the independent variables, implying a strong fit between the observed and predicted values.
- **Root Mean Squared Error (RMSE)**: The RMSE, calculated as the square root of the MSE, is approximately 185,677.43. This value provides an understanding of the average magnitude of the errors in the model's predictions. A lower RMSE suggests better accuracy, and the obtained value indicates relatively small errors in prediction.

The Lasso regression model demonstrates strong performance, as indicated by the low MSE and RMSE values, suggesting good accuracy in predicting the target variable. Additionally, the high R-squared value indicates a strong fit between the observed and predicted values, signifying that the model explains a significant portion of the variance in the data. Overall, these performance metrics suggest that the Lasso regression model effectively captures the relationship between the independent and dependent variables, providing valuable insights and accurate predictions.

Ridge Regression Model

Ridge regression is a regularization technique used to prevent overfitting by adding a penalty term to the traditional least squares objective function. It's beneficial when dealing with multicollinearity and high-dimensional datasets, as it stabilizes the model coefficients and improves prediction accuracy.

- **Ridge Regression Model Performance:**
- **Mean Squared Error (MSE)**: The MSE is approximately 34.48 billion (34,476,111,542.52), indicating good accuracy in the model's predictions.
- **R-squared (R^2)**: The high R-squared value of approximately 0.97 (97.43%) suggests a strong fit between the observed and predicted values, indicating that the model explains a significant portion of the variance in the data.
- **Root Mean Squared Error (RMSE)**: The RMSE is approximately 185,677.44, implying relatively small errors in prediction.

Model Tuning and Evaluation

KNN Model:

- K-Nearest Neighbors, is a simple and intuitive machine learning algorithm used for both classification and regression tasks.
- In KNN regression, predictions are made by averaging the target values of the K nearest data points to the query point.
- KNN is used for its simplicity and ease of implementation.
- It is effective when the data has a clear local structure and when you want to avoid making strong assumptions about the underlying distribution of the data.

KNN Regression Model Performance:

- The KNN regression model demonstrates strong performance with a low MSE of approximately 26.25 billion and a high R-squared value of around 0.98.
- The RMSE is approximately 162,031.87, indicating relatively small errors in prediction.
- Overall, the model accurately captures the relationships in the data and explains a significant portion of the variance in the target variable.

Random Forest Model:

Random Forest is like a team of decision trees working together to make predictions. Each tree is trained on a random part of the data and uses a random set of features. Then, when it's time to make a prediction, all the trees vote on the answer, and the final prediction is the average of all their votes. This teamwork helps to improve accuracy and make more reliable predictions.

- Random Forest is widely used for its high predictive accuracy, robustness to overfitting, and ability to handle large datasets with high dimensionality.
- It can capture complex relationships between features and target variables and is less prone to overfitting compared to individual decision trees.

Random Forest Model Performance:

- The Random Forest model demonstrates strong performance with a low MSE of approximately 24.22 billion and a high R-squared value of around 0.98.
- The RMSE is approximately 155,642.61, indicating relatively small errors in prediction.
- Overall, the model accurately captures the relationships in the data and explains a significant portion of the variance in the target variable.

Bagging Regression Model:

Bagging, short for Bootstrap Aggregating, is an ensemble learning technique that builds multiple models (in this case, regression models) on different subsets of the training data and then combines their predictions to improve accuracy and reduce overfitting.

Bagging regression involves training multiple regression models on random subsets of the training data (with replacement) and then averaging their predictions to make the final prediction.

We use bagging regression to increase the stability and robustness of the model, especially when the dataset is small or prone to overfitting. By combining the predictions of multiple models, bagging helps to reduce variance and produce more reliable predictions.

Bagging Regression Model Performance:

- **Mean Squared Error (MSE)**: The MSE is approximately 25.52 billion (25,519,023,789.01), indicating good accuracy in the model's predictions.
- **Root Mean Squared Error (RMSE)**: The RMSE is approximately 159,746.75, implying relatively small errors in prediction.
- **R-squared (R^2)**: The high R-squared value of approximately 0.98 (98.10%) suggests a strong fit between the observed and predicted values, indicating that the model explains a significant portion of the variance in the data.

Interpretation:

- The bagging regression model demonstrates strong performance, with relatively low MSE and RMSE values, indicating good accuracy in predicting the target variable.
- Additionally, the high R-squared value implies a strong fit between the observed and predicted values, signifying that the model effectively explains a significant portion of the variance in the data.
- Overall, the bagging regression model shows promising performance and can be considered a reliable approach for regression tasks, particularly when stability and robustness are important considerations.

Gradient Boosting regression Model

- Gradient Boosting is an ensemble learning technique that builds a strong predictive model by combining multiple weak models (typically decision trees) sequentially.
- In Gradient Boosting regression, each weak model is trained to correct the errors made by the previous models, focusing on the residuals (the differences between the actual and predicted values).
- We use Gradient Boosting regression because it often produces highly accurate predictions, even on complex datasets. It is particularly effective for tasks where high accuracy is crucial, such as in finance, healthcare, and marketing.
- Benefits of Gradient Boosting include its ability to handle heterogeneous features, its resistance to overfitting, and its flexibility in optimizing various loss functions.

Gradient Boosting Regression Model Performance:

- **Mean Squared Error (MSE)**: The MSE is approximately 20.99 billion (20,987,832,671.82), indicating excellent accuracy in the model's predictions.
- **Root Mean Squared Error (RMSE)**: The RMSE is approximately 144,871.78, implying relatively small errors in prediction.
- **R-squared (R^2)**: The high R-squared value of approximately 0.98 (98.43%) suggests an excellent fit between the observed and predicted values, indicating that the model explains a significant portion of the variance in the data.

Interpretation:

- The Gradient Boosting regression model demonstrates outstanding performance, with significantly low MSE and RMSE values, indicating high accuracy in predicting the target variable.
- Additionally, the high R-squared value implies an excellent fit between the observed and predicted values, signifying that the model effectively explains a substantial portion of the variance in the data.
- Overall, the Gradient Boosting regression model excels in accuracy and predictive power and is well-suited for tasks requiring highly accurate predictions. It offers a robust and flexible approach to regression modeling, making it a valuable tool in various domains.

Here's a comparative summary of the performance of the models:

- Gradient Boosting Machine Model:
  - MSE: 20,987,832,671.82
  - RMSE: 155,642.61
  - R-squared: 0.9843
- Random Forest Model:
  - MSE: 24,224,622,612.43
  - RMSE: 155,642.61
  - R-squared: 0.9819
- Linear Regression Model:
  - MSE: 34,476,113,176.02
  - RMSE: 185,677.44
  - R-squared: 0.9743
- Lasso Regression Model:
  - MSE: 34,476,108,782.81
  - RMSE: 185,677.43
  - R-squared: 0.9743

- Ridge Regression Model:
  - MSE: 34,476,111,542.52
  - RMSE: 185,677.44
  - R-squared: 0.9743
- Decision Tree Model:
  - MSE: 38,638,503,261.29
  - RMSE: Not provided
  - R-squared: 0.9712

Comparing these models, the Gradient Boosting Machine model stands out as the best performer with the lowest MSE, relatively low RMSE, and the highest R-squared value, indicating its superior accuracy and ability to explain variance in the data. The Random Forest model follows closely, demonstrating strong performance across all metrics. Linear Regression, Lasso Regression, and Ridge Regression models show similar performance, while the Decision Tree model trails slightly behind in terms of MSE and R-squared.

Based on the provided performance metrics, the Gradient Boosting Machine model appears to be the best-performing model among the ones listed. Here's why:

- **MSE (Mean Squared Error)**: The Gradient Boosting Machine model has the lowest MSE among all the models, indicating that it produces predictions with the smallest average squared difference from the actual values.
- **RMSE (Root Mean Squared Error)**: While the RMSE of the Gradient Boosting Machine model is slightly higher than that of the Decision Tree and Random Forest models, it is still relatively low compared to other models, indicating small errors in prediction.
- **R-squared**: The Gradient Boosting Machine model has the highest R-squared value, indicating that it explains the most variance in the data compared to other models. A higher R-squared value signifies a better fit between the observed and predicted values.
- **Overall Consistency**: The Gradient Boosting Machine model consistently performs well across all metrics, demonstrating its robustness and reliability in making accurate predictions.

Based on these observations, the Gradient Boosting Machine model is likely the best choice for making predictions on this dataset. It offers high accuracy, robustness, and a strong ability to explain variance in the data, making it a valuable tool for regression tasks.

Insights:

Data Collection Insights:

- The dataset comprises 25,000 records with 29 columns, including various attributes such as experience, current salary, and expected salary.
- Several columns have missing values, notably in Department, Designation, and Graduation_Specialization.
- Key numerical features show a wide range, indicating diversity in applicants' backgrounds and achievements.
- Exploratory Data Analysis (EDA) Insights:

Univariate Analysis:

- Continuous attributes like Total_Experience and Current_CTC exhibit diverse distributions, indicating a varied applicant pool.
- Categorical attributes such as Department and Designation highlight the distribution of applicants across different categories.

Bivariate Analysis:

- Strong correlations exist between Total_Experience and Expected_CTC, suggesting a relationship between experience and salary expectations.
- Outlier treatment is essential, particularly in Expected_CTC, to ensure accurate modelling results.

Missing Value Treatment:

- Missing values in various attributes require imputation or removal to maintain data integrity.

Insights Model Performance:

Linear Regression, Decision Tree, Lasso, and Ridge Regression Models:

- All four models demonstrate strong performance, with low mean squared error (MSE) and high R-squared values, indicating their effectiveness in predicting salary outcomes.
- "Current_CTC" emerges as a significant predictor in all models, highlighting its crucial role in determining salary predictions.

Ensemble Models (Random Forest and Gradient Boosting):

- Ensemble models outperform individual models, showcasing lower MSE and higher R-squared values.
- "Current_CTC" remains the dominant feature in both Random Forest and Gradient Boosting models, emphasizing its influence on salary predictions.

Recommendations:

Data Cleaning and Imputation:

- Address missing values in Department, Designation, and Graduation_Specialization through appropriate imputation techniques.
- Implement outlier treatment in Expected_CTC to ensure the model's accuracy and reliability.

Feature Engineering:

- Consider creating new variables that capture additional information relevant to salary determination, such as international degrees or certifications.
- Transform variables if necessary to improve model performance, such as scaling numerical features or encoding categorical variables.

Model Development:

- Develop a predictive model using historical data to estimate salary offers for prospective candidates.
- Utilize machine learning algorithms like regression or ensemble methods to capture complex relationships between attributes and salary expectations.

Validation and Evaluation:

- Validate the model's performance using appropriate metrics such as mean absolute error or R-squared.
- Ensure fairness and equity by evaluating the model's predictions across different demographic groups to identify and mitigate biases.

Continuous Monitoring:

- Regularly monitor and update the model to incorporate new data and adapt to changing trends in salary expectations.
- Conduct periodic audits to assess the model's impact on reducing discrimination and promoting fairness in salary determination.

Deployment of Ensemble Models:

- Prioritize the use of ensemble models like Random Forest and Gradient Boosting for predictive tasks due to their superior performance over individual models.

Focus on Current Salary Levels:

- Organizations should pay close attention to candidates' current salary levels when making hiring or salary adjustment decisions, as indicated by the dominant influence of "Current_CTC" in all models.

Utilization of Tuned Gradient Boosting Model:

- Deploy the tuned Gradient Boosting Machine model for critical predictive tasks, given its exceptional performance metrics. Its accuracy and reliability make it well-suited for decision-making processes that rely heavily on accurate predictions.

By implementing these recommendations, Delta Ltd. can develop a robust predictive model that minimizes manual judgments in salary determination and ensures fairness and equity across all employee profiles.

The most optimum model, the tuned Gradient Boosting Machine model, exhibits exceptional performance metrics, including a remarkably low MSE and high R-squared value. This indicates its ability to accurately capture underlying data patterns and make precise predictions.

- Improved Decision Making: The high accuracy and reliability of the tuned Gradient Boosting model enable more informed decision-making across various business domains. From resource allocation to customer segmentation, businesses can rely on the model's predictions to guide strategic initiatives effectively.
- Enhanced Operational Efficiency: By integrating the model into business operations, organizations can streamline processes and improve efficiency. Predictive insights from the model can optimize inventory management, supply chain logistics, and workforce planning, leading to cost savings and improved productivity.
- Targeted Marketing and Customer Engagement: Leveraging the model's predictive capabilities, businesses can personalize marketing campaigns and tailor customer engagement strategies based on predicted outcomes. This targeted approach can result in higher conversion rates, customer satisfaction, and long-term loyalty.

- Risk Mitigation: The accurate predictions provided by the model can help businesses identify and mitigate risks proactively. Whether its assessing credit risk, predicting market fluctuations, or identifying potential fraud, the model's insights empower businesses to make timely interventions and minimize negative impacts.
- Competitive Advantage: By harnessing the predictive power of the Gradient Boosting model, businesses gain a competitive edge in their respective industries. The ability to anticipate future trends, customer behaviours, and market dynamics allows organizations to stay ahead of the curve and adapt swiftly to changing landscapes.

Overall, the adoption of the tuned Gradient Boosting Machine model translates into tangible business benefits, ranging from improved decision-making and operational efficiency to targeted marketing efforts and risk mitigation. By leveraging the model's insights effectively, organizations can drive growth, mitigate risks, and maintain a competitive edge in today's dynamic business environment.