SUMMARY:

1. **Understanding the Data and Ensuring Data Quality:**

- Preliminary exploration involved checking the dimensions and information of the data, as well as identifying duplicates and addressing missing values.

- The 'Select' option was substituted with null values, as it did not contribute meaningful information. Additionally, certain null values were replaced with 'not available' to preserve data integrity.

2. **Cleaning and Enhancing the Data:**

- The data exhibited general cleanliness, with only a few instances of null values.

- Columns with more than 45% missing values were removed to streamline the dataset.

- Immaterial and heavily skewed categorical variables were excluded from the analysis.

- Certain variables underwent consolidation to enhance comparability and facilitate accurate analysis.

3. **Exploring the Data:**

- Exploratory Data Analysis (EDA) encompassed visualizing conversion ratios in relation to pertinent variables.

- Outliers were addressed, and numeric values underwent refinement to ensure data accuracy.

4. **Identifying and Processing Categorical Variables:**

- Dummy variables were created for categorical variables, leading to the removal of the original columns.

- Numeric values underwent scaling using the MinMaxScaler to maintain consistency in the dataset.

5. Partitioning Data into Training and Testing Sets:

- Utilizing the sklearn library, the dataset was divided into training (70%) and testing (30%) sets.

6. Standardizing Numeric Variables:

- Numerical variables underwent rescaling to ensure uniformity in their magnitudes.

7. Constructing the Model:

- Recursive Feature Elimination (RFE) was applied to identify the top 15 significant variables.

- Subsequent variable removal was performed manually based on VIF values and p-values, eliminating variables with VIF > 5 and p-value > 0.05.

8. Assessing Model Performance:

- A confusion matrix was generated to evaluate the model's performance.

- Determination of the optimal cut-off value through the ROC curve facilitated the calculation of accuracy, sensitivity, and specificity, each reaching approximately 90%.

9. Making Predictions:

- Predictions were executed on the test data using the optimal cut-off value of 0.2, resulting in an accuracy, sensitivity, and specificity of 90%.

- The analysis highlighted the most influential variables in attracting potential buyers, ranked in descending order:

  1. Direct Traffic
  2. Welingak Website
  3. Last Activity - Email Bounced
  4. Last Activity - Olark Chat Conversation
  5. Tags – Busy
  6. Tags - Closed by Horizon
  7. Tags - Lost to EINS
  8. Tags - Not Specified
  9. Tags – Ringing
  10. Tags - Will revert after reading the mail
  11. Last Notable Activity - SMS Sent

- Focusing on these key variables can significantly enhance X Education's likelihood of converting potential buyers and increasing course enrolments.