SUMMARY:

1. Data Import and Inspection:

- The required libraries were imported, and the CSV file was read to inspect the dataframe. Initial checks included examining the first few rows, number of rows and columns, and reviewing column-wise information and numeric summaries.

- Key Insights: This initial examination provided a clear understanding of the dataset and guided subsequent steps.

2. Data Preparation:

- The dataset was mostly clean, but some columns contained null values. The 'Select' option was replaced with null, and columns with more than 30% null values were dropped. Null values below 30% and above 2% were handled by imputing appropriate values (median, mean, or mode).

- Encoding: Categorical data was converted using one-hot encoding to prepare it for modeling.

3. Exploratory Data Analysis (EDA):

- Univariate analyses were conducted to understand the data distribution and relationships.

- Outlier detection and handling were performed, including capping outliers in numerical columns.

- Data imbalance was checked to understand if any class was underrepresented.

4. Dummy Variable Creation:

- Dummy variables were created for categorical columns.

5. Train-Test Split:

- The dataset was split into training (70%) and testing (30%) sets.

6. Feature Scaling:

- Standard scaling was applied to the numerical features to normalize the data.

7. Correlation Analysis:

- Correlations between variables were analyzed, and highly correlated variables were removed to avoid multicollinearity.

8. Model Building:

- Recursive Feature Elimination (RFE) was used to select the top 15 relevant variables. Additional variables were removed based on VIF values and p-values (variables with VIF < 5 and p-value < 0.05 were kept).

- The logistic regression model was chosen to predict the probability of conversion.

- Adjusted R-squared, VIF, and p-values were used to further refine the model.

9. Model Evaluation:

- Confusion matrix and ROC curve analysis were used for evaluation.

- The optimum cutoff (0.35) was determined using the ROC curve to maximize accuracy, sensitivity, and specificity (81%, 83%, and 80%, respectively).

10. Conclusion:

- The logistic regression model effectively predicts the probability of customer conversion.

- Key features influencing the model include 'Do Not Email', 'Lead Origin_Lead Add Form', 'Last Activity Had a Phone Conversation', 'Last Activity_SMS Sent', 'Last Activity_Unsubscribed', 'What is your current occupation_Unemployed', 'Tags_Busy', 'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Last Notable Activity_Modified', and 'Last Notable Activity_Olark Chat Conversation'.
- The top three features in the final model are 'Tags_Closed by Horizzon', 'Tags_Lost to EINS', and 'Tags_Will revert after reading the email'.
- The model achieved a sensitivity of 83%
indicating its ability to correctly identify 72.77% of converted customers.

This summary provides a comprehensive overview of the analysis conducted to optimize customer conversion for X Education.