Summary:

The model building and prediction efforts are undertaken for Company X Education with the aim of identifying strategies to convert potential users. The process involves a comprehensive exploration and validation of the data to make informed decisions on targeting specific groups and enhancing the conversion rate. The following steps outline our approach:

EDA:

- A swift assessment of null values led to the removal of columns with over 45% missing values.
- Handling of null values involved replacing them with 'not provided' to preserve important columns.
- Imputation of 'not provided' values with 'India,' the most prevalent non-missing value.
- Considering the overwhelming prevalence of 'India' (nearly 97% of the data), the column was dropped.
- Treatment of numerical variables, addressing outliers, and creating dummy variables were performed.

Train-Test Split & Scaling:

- Data was split into 70% for training and 30% for testing.
- Min-max scaling was applied to the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'].

Model Building:

- Recursive Feature Elimination (RFE) was utilized for feature selection.
- Top 15 relevant variables were identified through RFE.
- Manual removal of remaining variables based on VIF values and p-values was performed.
- A confusion matrix was generated, and the overall accuracy checked, yielding 80.91%.

Model Evaluation:

- Sensitivity-Specificity Evaluation:
- On Training Data, an optimum cut-off value of 0.35 was determined using the ROC curve with an area under the curve of 0.88.
- For this cut-off, accuracy was 80.91%, sensitivity was 79.94%, and specificity was 81.50%.
- On Test Data, accuracy was 80.02%, sensitivity was 79.23%, and specificity was 80.50%.
- Precision-Recall Evaluation:
- On Training Data, with a cut-off of 0.35, precision and recall were 79.29% and 70.22%, respectively.

- After finding the optimum cut-off of 0.44, accuracy increased to 81.80%, precision to 75.71%, and recall to 76.32%.
- On Test Data, accuracy was 80.57%, precision was 74.87%, and recall was 73.26%.

Conclusion:

- The optimal cut-off value for Sensitivity-Specificity Evaluation is 0.35, while for Precision-Recall Evaluation, it is 0.44.