**Summary**

We are currently engaged in developing a model for company X Education with the aim of identifying effective strategies to convert prospective users. Our next step is to thoroughly analyze and validate the data in order to pinpoint the appropriate target audience and enhance the conversion rate.

Firstly, we conducted a preliminary analysis of the data and dropped any columns with missing values greater than 45%.

Since some of the remaining columns contained important information, we replaced the NaN values with "not provided." Given that "India" was the most frequent value among non-missing data, we imputed all "not provided" values with "India."

However, as the majority of the data (about 97%) pertained to India, we dropped this column. We also worked on numerical variables, outliers, and created dummy variables. Train-Test Split & Scaling: The data was split into training and test sets at a ratio of 70% and 30%, respectively.

The variables "TotalVisits," "Page Views Per Visit," and "Total Time Spent on Website" were scaled using min-max scaling. Model Building: For feature selection, we used Recursive Feature Elimination (RFE) to identify the top 15 relevant variables.

We then manually removed the remaining variables based on their VIF values and p-values. A confusion matrix was created, and the overall accuracy was found to be 80.91%.

**Model Evaluation**

We evaluated the model's performance using two different methods, Sensitivity-Specificity and Precision-Recall.

For the former, the optimal cut-off value was 0.35, which gave an accuracy of 80.91%, sensitivity of 79.94%, and specificity of 81.50% on the training data. On the test data, the accuracy was 80.02%, sensitivity was 79.23%, and specificity was 80.50%. For the Precision-Recall evaluation, we found that the optimal cut-off value was 0.44. This yielded an accuracy of 81.80%, precision of 75.71%, and recall of 76.32% on the training data. On the test data, the accuracy was 80.57%, precision was 74.87%, and recall was 73.26%.

Thus, the optimal cut-off value for the model depends on the evaluation method: 0.35 for Sensitivity-Specificity and 0.44 for Precision-Recall.

**Conclusion**

In summary, the data analysis involved dropping columns with a high percentage of missing values, imputing missing values, and creating dummy variables.

The data was split into training and test sets, and the selected variables were scaled using min-max scaling. Recursive Feature Elimination (RFE) was used for feature selection, followed by manual removal of variables based on their VIF values and p-values.

The model was evaluated using two methods: Sensitivity-Specificity and Precision-Recall.

For Sensitivity-Specificity, the optimal cut-off value was 0.35, resulting in an accuracy of 80.91%, sensitivity of 79.94%, and specificity of 81.50% on the training data. On the test data, the accuracy was 80.02%, sensitivity was 79.23%, and specificity was 80.50%. For Precision-Recall, the optimal cut-off value was 0.44, resulting in an accuracy of 81.80%, precision of 75.71%, and recall of 76.32% on the training data. On the test data, the accuracy was 80.57%, precision was 74.87%, and recall was 73.26%. In conclusion, the optimal cut-off value for the model depends on the evaluation method chosen, and further analysis could be conducted to refine the model and improve its accuracy.