Summary of the Case Study

This analysis is conducted for X Education with the objective of attracting a greater number of industry professionals to enroll in their courses. The foundational data we have at our disposal has provided valuable insights into the visitors' behaviour on the website, including their visit duration, source of arrival, and the conversion rate.

Here are the steps employed:

- **Reading Dataset & Understanding Data**

  In the 'Reading Dataset & Understanding Data', we explore the process of loading and examining datasets. This summary provides insights into key techniques for data ingestion and initial data exploration.

- **Cleaning Data**

  The data was mostly clean, with only a few null values. However, the 'option select' field had to be replaced with a null value as it didn't provide significant information. Some of the null values were temporarily changed to 'not provided' to preserve data integrity, although they were subsequently removed during the dummy variable creation process. Given the diverse geographical origins of the respondents, the elements were categorized as 'India,' 'Outside India,' and 'not provided' to better represent the dataset.

- **EDA**

  We conducted a brief Exploratory Data Analysis (EDA) to assess the data's quality. During this analysis, we identified that several elements within the categorical variables were irrelevant. On the other hand, the numeric values appeared to be well-behaved, showing no signs of outliers. Subsequently, we generated dummy variables and eliminated those containing 'not provided' elements. To ensure consistency, we applied the MinMaxScaler to normalize the numeric values.

- **Test-Train Split**

The data was divided into training and test sets, with a split ratio of 70% for training and 30% for testing.

- **Model Building**

Initially, a Recursive Feature Elimination (RFE) process was conducted to select the top 15 relevant variables. Subsequently, the remaining variables were systematically pruned based on their Variance Inflation Factor (VIF) values and p-values. Variables meeting the criteria of VIF < 5 and p-value < 0.05 were retained.

- **Prediction**

The test data frame was used for prediction, and an optimal cutoff value of 0.38 was applied, resulting in accuracy, sensitivity, and specificity rates of 80%, 78%, and 82%, respectively.

- **Model Evaluation**

We constructed a confusion matrix and then determined the optimal cutoff value, utilizing the ROC curve. This analysis yielded accuracy, sensitivity, and specificity rates of approximately 81%, 70%, and 88%, respectively.

- **Precision-Recall**

This method was additionally employed for re-evaluation, yielding a discerned cutoff value of 0.41. The resulting performance metrics on the test data frame indicate a Precision of approximately 75% and a Recall of roughly 76%.