Lead Scoring Case Study Summary

This analysis is done for X Education to find out what they need to do to attract more industry professionals to their courses. Based on the Data provided we get a lot of information about the way potential customers visited the site, how long they spent on it, how they arrived at it, and the conversion rate.

The following are the steps used for Solution are:

- **Step1: Reading and Understanding Data**
  - Importing the necessary library
  - Read and analyse the data

- **Step2: Data Cleaning**
  - We replaced the 'select' option with the null value since it did not give us much information
  - The data was partially clean except for a few null values. We dropped the variables with a null value greater than 40%, and for the remaining null values we made changes in accordance with the data

- **Step3: Data Analysis (EDA)**
  - Then we started with the EDA to check the condition of our data. We performed the Categorical and Univariate analysis to get insights over dataset
  - The analysis revealed many variables that were irrelevant, so these variables were dropped

- **Step4: Creating Dummy Variables**
  - We next created dummy data for the categorical variables

- **Step5: Test Train Split**
  - In the next step, the data set was divided into train and test segments with a 70-30% ratio

- **Step6: Feature Rescaling**
  - We used the StandardScaler to scale the original numerical variables
  - Using the stats model, we created our initial model, which gave us a complete statistical view of each parameter

- **Step7: Feature selection using RFE**
  - The 15 top features were selected based on the RFE
  - Based on the VIF values and p-values (VIF < 5 and p-value < 0.05 were kept), the variables that were not significant were manually removed
  - We derived the Confusion Metrics and calculated the model's overall **Accuracy** based on the above assumptions
  - For a better understanding of the model's reliability, we also calculated the **Sensitivity** and **Specificity** matrices

- **Step8: Plotting the ROC Curve**
    - We plotted the ROC curve for the features, which showed a good area coverage of 86%, further solidifying our model

- **Step9: Finding the Optimal Cut-off Point**
    - We plotted the probability graph for Accuracy, Sensitivity, and Specificity for different probabilities and intersection was considered as optimal probability cut-off point
    - Based on a cut-off value of 0.37, we use 0.4 as the model estimate
    - We could also observe the values of the **Accuracy = 79.1%**, **Sensitivity = 75.0%**, **Specificity = 81.6%**

- **Step10: Computing the Precision and Recall metrics**
    - This method was also used to recheck and a cut off of 0.42 was found with **Precision = 71.9%** and **recall = 75.0%** on the train data frame

- **Step11: Making Predictions on Test Set**
    - We then applied these learnings to test model and found an **accuracy = 79.1%**, **Sensitivity = 74.3%** and **Specificity = 81.8%**