Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary:

Step 1: Data Reading and Understanding:
The data is analyzed by reading and understanding its contents.

Step 2: Data Cleaning:
Variables with a high percentage of NULL values are dropped. Missing values are imputed, using median values for numerical variables and creating new classification variables for categorical variables. Outliers are identified and removed.

Step 3: Data Analysis:
Exploratory Data Analysis is conducted to gain insights into the data distribution. Variables with a consistent value across all rows are identified and dropped.

Step 4: Creating Dummy Variables:
Dummy data is generated for categorical variables.

Step 5: Test-Train Split:
The dataset is divided into test and train sections with a 70-30% split.

Step 6: Feature Rescaling:
Min-Max Scaling is applied to normalize the numerical variables. The initial model is created using statistical modeling techniques to obtain a comprehensive view of the model parameters.

Step 7: Feature Selection using RFE:
Recursive Feature Elimination is used to select the top 20 important features. Based on generated statistics, insignificant values are dropped by recursively examining P-values. Finally, 15 most significant variables are identified with satisfactory VIF values.

Step 8: Plotting the ROC Curve:
The ROC curve is plotted to evaluate the model's performance, demonstrating a decent area coverage of 89% and further validating the model.

Step 9: Finding the Optimal Cutoff Point:
Probability graphs for accuracy, sensitivity, and specificity are plotted for various probability values. The intersection point of these graphs determines the optimal probability cutoff point, which is found to be 0.37. With this new value, the model accurately predicts close to 80% of the cases. The updated accuracy, sensitivity, and specificity values are approximately 81%, 79.8%, and 81.9%, respectively.

Step 10: Computing the Precision and Recall Metrics:
Precision and recall metrics are computed, resulting in values of 79% and 70.5%, respectively, on the train dataset. Considering the tradeoff between precision and recall, a cutoff value of approximately 0.42 is obtained.

Step 11: Making Predictions on Test Set:
The acquired knowledge is applied to the test model, and conversion probability is calculated based on sensitivity and specificity metrics. The accuracy value is determined to be 80.8%, with sensitivity at 78.5% and specificity at 82.2%.