Lead Scoring Case Study Summary

Problem Statement:

Industry professionals may purchase online courses from X Education. X Education need assistance in identifying the most promising prospects—that is, the leads with the highest likelihood of becoming paying clients.

The business requires a model in which each lead is given a lead score, giving consumers with higher lead scores a better probability of converting, and customers with lower lead scores a lesser chance.

Specifically, the CEO has said that an approximate 80% lead conversion rate is the goal.

Summary:

Step 1: **Reading and Understanding data**.

Read and analyze the data.

Step 2: **Cleaning Data**

The variables with a high percentage of NULL values were eliminated. In this stage, missing values were also imputed when needed using median values for numerical variables, and new categorization variables were created for categorical variables. We located and eliminated the outliers.

Step 3: **Data Analysis**

After that, in order to acquire a sense of the data's orientation, we began the data set's exploratory data analysis. Approximately three variables were found to have a single value in every row in this stage. These were removed as variables.

Step 4: **Data Preparation**

We proceeded to generate pseudo-information for the classification variables.

Step 5: **Test Train**

Subdividing the data set into test and train segments using a 70–30% value ratio was the next step

Step 6: Rescaling with Min-Max

The original numerical variables were scaled using the Min Max Scaling method. We then built our first model utilizing the statistics model, which provided us with a comprehensive statistical perspective of all the model's parameters.

Step 7 and 8: **Model Build and RFE**:

We went ahead and chose the top 20 significant features using the Recursive Feature Elimination method. We attempted iteratively to examine the P-values using the produced statistics in order to exclude the less significant values and choose the most significant ones that ought to be there.

We finally identified the fifteen most important factors. It was also discovered that the VIFs for these variables were good.

Next, we constructed the data frame containing the transformed probability values. Initially, we assumed that if a probability value was more than 0.5, it would indicate either 1 or 0.

We determined the Confusion Metrics and the overall Accuracy of the model based on the aforementioned assumption.

Additionally, in order to determine the model's level of reliability, we computed the "Specificity" and "Sensitivity" matrices.

Step 9: ROC Curve

The model was further validated when we attempted to display the ROC curve for the characteristics. The curve showed a respectable area coverage of 89%.

Step 10: Finding Cutoff Point

Next, we produced the probability graph for various probability values for the variables "Accuracy," "Sensitivity," and "Specificity." The best probability cutoff point was thought to be where the graphs intersected. It was discovered that the cutoff threshold was 0.37.

We could see from the new number that the model had over 80% of the values correctly predicted. Additionally, we could see the updated "accuracy=81%, sensitivity=79.8%, and specificity=81.9%" numbers.

Additionally, the lead score was computed, and it was determined that the final projected variables roughly yielded an 80% target lead prediction.

Step11: Computing the Precision and Recall metrics

Additionally, we discovered that on the train data set, the Precision and Recall metrics values were, respectively, 79% and 70.5%.

We obtained a cut off value of around 0.42 based on the Precision vs. Recall tradeoff.