Lead Scoring Case Study Summary

Problem Statement:

X Education requires assistance in developing a system that can find the most qualified leads for their online courses. Each lead should be given a score by the system, indicating how probable it is that they will convert to paying clients. The likelihood of the lead converting increases with the score. This system will assist X Education in meeting the CEO's target of an 80% conversion rate by concentrating their efforts on the leads that are most likely to convert.

Solution Summary:

Step1: **Reading and Understanding Data**.

Read and analyze the data.

Step2: **Data Cleaning**:
The variables with a high percentage of NULL values were removed. In addition, when necessary, missing values were imputed using median values for numerical variables and new categorization variables were created for categorical data. The outliers were found and eliminated.

Step3: **Data Analysis**
Once we had a sense of the data's orientation, we began the data set's exploratory data analysis. At this phase, approximately 3 variables were found to have the same value across all rows. These factors were removed.

Step4: **Creating Dummy Variables**

For the categorical variables, we kept on creating dummy data.

Step5: **Test Train Split**:
The data set was then split between test and train sections, with a 70–30% split between each.

Step6: Feature Rescaling
The original numerical variables were scaled using the Min Max Scaling method. We then developed our initial model utilising the statistics model, which provided us with a comprehensive statistical perspective of all the model's parameters.

Step7: **Feature selection using RFE**:
We chose the top 20 features by employing the Recursive Feature Elimination method. We recursively attempted examining the P-values using the statistics produced in order to choose the most significant values that should be present and eliminate the unimportant ones.

We finally reached the 15 most important variables. These variables' VIFs were likewise discovered to be good.

The transformed probability values were then placed in a data frame, and we made the initial assumption that if the probability value was greater than 0.5, it meant that the value should be 1, otherwise.

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

To assess how accurate the model is, we also computed the "Sensitivity" and "Specificity" matrices.

Step8: Plotting the ROC Curve
When we tried to draw the ROC curve for the features, it looked very good and had an area coverage of 89%, which further supported the model.

Step9: Finding the Optimal Cutoff Point
The probability graph for "Accuracy," "Sensitivity," and "Specificity" was then displayed for various probabilities. The ideal probability cutoff point was thought to be where the graphs intersected. The threshold value was determined to be 0.37.

Based on the updated value, we could see that the model correctly predicted values for close to 80% of the values. Also, we could see the updated accuracy, sensitivity, and specificity scores of 81%, 79.8%, and 81.9%.

Also, the lead score was calculated, and it was determined that the final projected variables provided a target lead prediction of 80%.

Step10: Computing the Precision and Recall metrics
For the train data set, we also discovered that
the Precision and Recall metrics values were
79% and 70.5%, respectively.
We obtained a cut off value of roughly 0.42 based on the Precision and Recall tradeoff.

Step11: **Making Predictions on Test Set**
The conversion probability was then determined based on the Sensitivity and Specificity metrics, and we discovered that the accuracy value was 80.8%; Sensitivity was 78.5%; and Specificity was 82.2%. Next, we applied the learnings to the test model.