Summary

Problem Statement:

An X Education requires assistance in identifying the most promising leads, i.e. those most likely to become paying customers. The company wants us to create a model in which each lead is given a lead score, with higher lead scores having a higher conversion chance and lower lead scores having a lower conversion chance.

The target lead conversion rate, according to the CEO, is roughly 80%.

Solution:

Step1: **Load and Inspect Data**.

Read, analyze and try do basic cleaning in
the data.

Step2: **Data Cleaning**:

The variables with a high percentage of NULL values were removed. In the case of numerical variables, this stage also included imputing missing values when needed with median values, and in the case of categorical categories, creating additional categorization variables. Outliers were found and eliminated.

Step3: **EDA**

Then we did an exploratory data analysis of the data set to obtain a sense of how it was organized. Around three variables were discovered in this step as having only one value in all rows. These variables were removed from the equation.

Step4: **Dummy Variables**

For the categorical variables, we continued to create dummy data.

Step5: **Test Train Split**:

The following stage was to divide the data set into test and train halves, with a 70-30 percent split.

Step6: Feature Rescaling

To scale the original numerical variables, we used Min Max Scaling. Then, using the stats model, we built our first model, which would provide us with a complete statistical perspective of all of our model's parameters.

Step7: **RFE**

We went ahead and picked the 20 most important characteristics using Recursive Feature Elimination.

We iteratively looked at the P-values using the statistics generated in order to choose the most significant values that should be present and eliminate the unimportant ones.

Finally, we came up with a list of the 15 most important variables.

These variables' VIFs were likewise judged to be satisfactory. We then generated a data frame with the converted probability values, with the premise that a probability value greater than 0.5 indicates 1 and less than 0.5 means 0. We computed the overall Accuracy of the model using the Confusion

Metrics based on the above assumption. To determine how trustworthy the model is, we calculated the 'Sensitivity' and 'Specificity' matrices.

Step8: ROC Curve

We then plotted the ROC curve for the features, which turned out to be quite good, with an area coverage of 88%, confirming the model's accuracy.

Step9: Optimal Cutoff

Then, for different probability values, we displayed the probability graph for 'Accuracy,' 'Sensitivity,' and 'Specificity.' The appropriate probability cutoff point was determined by crossing the graphs. 0.35 was discovered to be the cutoff point. Based on the new number, we can see that the model correctly predicted over 80% of the values.

We could also see the new 'accuracy=80%','sensitivity=80.4%', and 'specificity=80.25%' figures.

Also, the lead score was determined, and the final projected variables yielded an approximate target lead prediction of 81%.

Step10: Precision and Recall metrics

On the train data set, we discovered that the Precision and Recall measures were 75 percent and 76 percent, respectively. We arrived at a cut off value of 0.41 based on the Precision and Recall tradeoff.

Step11: **Making Predictions on Test Set**

Then, using the Sensitivity and Specificity metrics, we estimated the conversion probability for the test model and discovered that the accuracy value was 81.5 percent; Sensitivity=78.5 percent; Specificity= 82.2 percent.