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: The following steps were followed to come up with a solution for the above problem statement.

1. **Data Inspection**: Read and analyse overall data structure and variable types. Checking the null values in each column.
2. **Data Cleaning:** Dealing with the null values. Removing columns with a large percentage of null values. Handling null values with numeric data with medians and for categorical columns assigning the value which is most repeated to null values or with a generic value as 'not provided'. Replacing garbage values with meaningful values.
3. **EDA:** Finding outliers with boxplot on Numerical columns and handling the outliers. Comparing the conversion rates with the categorical variables. Dropping Categorical variables which will be of minimal use in modelling.
4. **Data Preparation:** Converting binary categorical columns to binary, Creating dummy variables for categorical columns and dropping the original categorical variables. Splitting the data into train and test sets of 70:30 ratio. Scaling the numerical columns. Feature selection with RFE
5. **Model Building:** Repeat the model building process until the columns are eliminated and the p-value and VIF values are within an acceptable range of <0.05 for p-value and <5 for VIF.
6. **Prediction & Result Analysis:** Predict the train set and do a analysis of predicted values based on 0.5 probability. Check for Accuracy, Sensitivity and Specificity. Plot an ROC Curve to determine how good our model is. Based on the ROC curve area of 0.88 we concluded the model to be good. Finding Optimal Cut-off point which came to around 0.37. Make a prediction based on optical cut off point and calculate the Accuracy, Sensitivity and Specificity which came out as below

   Train Set:

   Accuracy: 80.43

Sensitivity: 77.82
Specificity: 82.05

7. **Prediction on Test Set:** Implement the model to predict the values on the test set and do all the Calculations mentioned in the above step on it. We found the result on the test set as below:
   Test Set:
   Accuracy: 79.66
   Sensitivity: 77.55
   Specificity: 81.0

   Overall we can conclude that the model is good based on the above parameters output.