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:

Step1: **Reading and Understanding Data**.
Read and analyze the data.

Step2: **Data Cleaning**:
We dropped the variables that had 30% or more of null values in them. Then we drop the variables having uniquevalues. Then, there were few columns with value 'Select' which means the leads did not choose any given option. We changed those values to Null values. Then again, we dropped the columns having null values greater than 30%. And dropped some columns having skewed data. Handled the outliers in numerical columns

Step3: **Data Analysis**
Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step we drop country column as its huge portion is showing India which is no use for our analysis

Step4: **Creating Dummy Variables**
we went on with creating dummy data for the categorical variables.

Step5: **Test Train Split**:
The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling
We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step7: **Feature selection using RFE**:
Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values. Then by using VIF we eliminate high VIF variable

Finally, we arrived at the 16 most significant variables. The VIF’s for these variables were also found to be good. We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0. Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracyof the model. We also calculated the ‘**Sensitivity**’ and the ‘**Specificity**’ matrices to understand how reliable the modelis.

Step8: Plotting the ROC Curve

We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 88% .

Step9: Finding the Optimal Cutoff Point

Then we plotted the probability graph for the ‘Accuracy’, ‘Sensitivity’, and ‘Specificity’ for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.3 Based on the new value We could observe the new values of the ‘accuracy=79.83%, ‘sensitivity=82.80%’, ’specificity=78.01%’.

Step10: Computing the Precision and Recall metrics

we also found out the Precision and Recall metrics values came out to be 69.74% and 82.80% respectively onthe train data set.

Step11: **Making Predictions on Test Set**

Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.22%; Sensitivity=82.20%; Specificity= 78.39%.