LEAD SCORING ASSIGNMENT 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:

1. Reading and Understanding Data

Data was first read and analysed.

2. Data Cleaning:

We dropped the variables that had high percentage of NULL values in them. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

3. Data Analysis

We started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were around variables that were identified to have only one value in all rows. These variables were subsequently dropped.

4. Creating Dummy Variables

We proceeded with creating dummy data for the categorical variables.

5. Test Train Split:

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

6. Feature Scaling

We scaled values using standardisation. 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.

7. Feature selection using RFE

a) Using the Recursive Feature Elimination, we went ahead to rank and subsequently select the 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.

b) 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.
c) Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.
d) We also calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

8: Plotting the ROC Curve

Since the area under ROC curve is 0.88 out of 1, it indicates a good predictive model.

9. 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.

10. 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 the following:

```
Confusion Matrix
[1353  324]
[ 221  874]

************************************************

True Negative                    :  1353
True Positive                    :  874
False Negative                   :  221
False Positve                    :  324
Model Accuracy                   :  0.8034
Model Sensitivity                :  0.7982
Model Specificity                :  0.8068
Model Precision                  :  0.7295
Model Recall                     :  0.7982
Model True Positive Rate (TPR)   :  0.7982
Model False Positive Rate (FPR)  :  0.1932
```