LEAD SCORE 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.

On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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%

Case Study Approach:

For this case study, where the target, resultant outcome is either the lead will join the course or not. Since the resultant outcome is categorical in nature, we will use the logistic regression model to predict the potential lead based on the given data and check its accuracy and prediction power using certain metrics.

The step-by-step procedure is given below:

Step1: **Reading and Understanding Data**: We have two data files, one containing the data of the lead and the other containing the meaning of the data, we use both of them to understand the given the nature of the leads.

Step2: **Data Cleaning**: We step includes dropped the variables that has high percentage of missing values in them. imputing the missing values as per requirement

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. Here we got some insight od data distribution and further removed the variables which were highly imbalanced by 1 particular value.

Step4: Data Preparations: In this stage we cap the variables with outliers and create the dummy variables for 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 of the given data set.

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-value and VIF value in order to select the most significant values that should be present and dropped the insignificant variables.

Finally, we arrived at the 17 most significant variables. The VIF's for these variables were also found to be good.

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 97% which further solidified the of the model.

Step9: Finding the Optimal Cutoff Point: First we created a data frame containing the cut-off from 0.1 to 1. 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 0.283. With which we achieved the accuracy of 92%.

Step10: Calculating the metrics: For the above model we got the Sensitivity = 91% and Specificity =92%, We found Precision = 88%, and Recall to be 91%.

Step11: Precision and Recall trade-off: From precision and Recall trade – off we found the cut-off to be 0.41. For the given problem statement sensitivity/Recall are important, hence we choose 0.283 cut off since with it our sensitivity/recall are maximum

Step11: **Making Predictions on Test Set** Then we implemented the learnings to the test model and calculated the co nversion probability. The metrics of the test data is as follows: Sensitivity/Recall : 92.39%, Specificity : 91.92%, precision : 87.9% and False-Positive-Rate : 8.08%.