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 has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary:

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

Step2: **Data Cleaning**: We dropped the variables that had a high percentage of NULL values in them. This step also included imputing the missing values. The outliers were identified and removed.

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.

- For the column country, about 95% values are India. This will not help in our analysis as it does not bring in any variance into the dataset.
- For the column city, the majority of values are Mumbai and about 28% data is 'Select' which represents missing data.
- For the Leads Profile, how did you hear about X education we can see a lot of values which are equal to "select" and are about 63% and 71% respectively.
- For the column what matters most to you in choosing the course , about 99% values are Better Career Prospects, since most of the values are the same, this column does not bring in any variance necessary for our analysis.

hence, we should drop these columns, as they are of no use

Initially we had about 9240 records , whereas after processing of the null values we are left with 6373 records .It means we have lost about 30% of the data during our initial processing and we are only left with 70% of data for our further 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 selected the 15 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.

Finally, we arrived at the 11 most significant variables. The VIF's for these variables were also found to be good. Then we 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 Accuracy of the model. We also calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

Step8: **Plotting the ROC Curve**: Then we tried plotting the ROC curve for the features and the curve came out to be decent with an area coverage of 86% which further solidified the model.

Step9: Finding the Optimal Cut-off/ Threshold Point: Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different cut off probability values. The intersecting point of the graphs was considered as the optimal probability cut-off point . The cut-off point was found out to be around 0.42. We could also observe the new values of the 'accuracy=79%, 'sensitivity=79.3%', 'specificity=78.8%'.

Step10: **Making Predictions on Test Set:** We implemented the learnings to test the model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 78.4%; Sensitivity=77.9%; Specificity= 78.9%.