Lead Scoring Case Study Summary

Problem Statement: -

A company name 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 requires you to build a model wherein you need to assign a lead score 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: -

Step1: Reading and Understanding Data.

Read and analysing the data.

Step2: Data Cleaning.

We dropped the variables more than 40% of NULL values in them. Whereas the variables which have less than 13% of NULL values where we 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.

Step3: Exploratory Data Analysis (EDA).

In EDA we can visualizing, summarizing and interpreting the information that is hidden in rows and column format. After completing the EDA. we know how the data is oriented.

Step4: Creating Dummies Variables.

We creating dummy variable for categorical data.

Step5: Test Train Split

In this step we divide the data set into test and train sections with a proportion of 70% and 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 15 top important features. Using the statistics generated, we recursively tried looking at the P-values and VIF in order to select the most significant values that should be present and dropped the insignificant values.

Finally, we arrived at the 10 most significant variables. The VIF's for these variables were also found to be good and in permissible range. 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 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

We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 84%. ROC curve shows the trade-off between sensitivity and specificity.

Step9: Finding the Optimal Cut-off 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 cut-off point. The cut-off point was found out to be 0.3. Based on the new value we could observe that close to 79% values were rightly predicted by the model. We could also observe the new values of the 'accuracy=79.5%, 'sensitivity=78.01%', 'specificity=80.45%'.

Step10: Computing the Precision and Recall metrics

Then we also found out the Precision and Recall metrics values came out to be 76% and 64.5% respectively on the train data set. Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.4.

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 =79.6%, Sensitivity=78.8%, Specificity= 80.2%.