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 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.

Step3: Data Transformation
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, there were variables that were identified to have only one value in all rows. These variables were dropped. We also created dummy variables for the categorical variables for further analysis.

Step4: Test Train Split
The next step was to divide the data set into test and train sections with a proportion of 70-30%
values.

Step5: Model Building
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 and the initial conversion rate came out to be 43.88%

Step6: Model Evaluation

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 in order to select the most significant values that should be present, and dropped the insignificant values.Finally, we arrived at the 12 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 accuracy of the model.
We also calculated the '**Sensitivity**' which came out to be 89.57% and the '**Specificity**' which came out to be 95.61%

Step7: 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 98% which further solidified the of the model.

Step8: 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 0.4
Based on the new value we could observe that close to 92.95% values were rightly predicted by the model.

Step9: Computing the Precision and Recall metrics

we also found out the Precision and Recall metrics values came out to be 89.07% and 92.17% respectively on the train data set.

Step10: Conclusion

From the above observations we can conclude that our model has a high accuracy along-with high sensitivity. The values of accuracy, specificity and sensitivity are almost similar in both test and train data-set which implies our model is good enough to predict the conversion rate.