Lead Scoring Case Study
Summary

Problem Statement

X Education sells on-line courses to industry professionals.

They need 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

After reading and analysing the data we started to clean the data. We dropped the variables that had high percentage of NULL values in them. For cleaning the NULL values we also imputed the missing values when required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. Outliers were removed or adjusted then the Exploratory Data Analysis of the data set started to understand the data.

In this step, we identified 3 variable to have only one value, they were dropped eventually. We started creating dummy variables for categorical values. The next step was to divide the data set into test and train sections with a proportion of 70-30% values. We also scaled the numerical values for smoother evaluation.

Feature selection using RFE was done next. We went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values. We also checked the VIF to see any incorrect values and removed them as well. 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 found the cutoff point to be 0.4 We observed that our prediction accuracy was 79% values were rightly predicted by the model. Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 79%

Finally we made predictions on test set. When we calculated the accuracy if that model prediction it came around to 78%

Overall this was a good result and the model was able to predicted as expected.