Lead Score Case Study

Summary:

- A detailed case study was done to find out the potential "Leads" for X education company.
- By finding out these hot leads, the Sales team will be able to market their courses to the potential customers and need not waste time by reaching out to the cold leads who are less likely to join a course.
- The Data was analyzed in detail and data was cleansed according to the requirements by removing/ imputing the missing values, dropping the least important variables, converting the binary variables and creating dummy variables.
- Outlier Analysis was done, and the data set did not seem to have any outliers.
- The data set was split into test and train set followed by Feature scaling of Test Set to ensure that values of all the columns are in a similar scale or range.
- The Data set was also tested for imbalance and since we got a decent percentage (38%) we did not use any techniques to handle the imbalance.
- A heat map was built to find out the correlations between the variables and the most correlated variables were dropped.
- After which the first logistic regression model was built using "stats model" library.
- Then RFE technique was used to select the significant variables and the second model was built using the variables selected by RFE
- Based on the p-value, the insignificant variables were dropped.
- Further to RFE method, few insignificant variables were dropped based on the high VIF values.
- The Confusion Matrix was built and the Accuracy, Specificity and Sensitivity scores were calculated for the Train set.
- We got 88% Accuracy, 81% Specificity and 93% Sensitivity for the Model built on the Train Set.
- For the Test set we got 88% Accuracy,80% Specificity and 93% Sensitivity as close to the Train set scores.
- A ROC curve was built to understand the tradeoff between Sensitivity and Specificity and optimal cut off was obtained.
- The Area under the curve came out be 0.95 which shows that the Model built is a good model.

- The model was again evaluated for the optimal cut off which was obtained as 0.4 and the metrics did not have much difference to the values calculated for the Cutoff of 0.5
- A variable called "Lead_Score" was created and values were assigned by multiplying the respective Conversion Probability into 100
- Based on the high lead score the leads were labelled as "Hot Leads" which is the ultimate goal of the Case Study
- Sales team can now focus on contacting these Hot leads rather than investing time for other members.

Conclusion:

- A Logistic Regression Model was thus built with 35 predictor variables (Final Model) by employing a hybrid method which combines Automatic Feature Selection technique (RFE), manually selecting and neglecting variables and by choosing the variables based on p-value and VIF method.
- The Model built is pretty decent with 88% Accuracy and also the train, test sets did not have major difference in the scores.