Lead Scoring Case Study

Summary

The problem at hand is to find ways to maximize the conversion of leads for the client company X Education. The data for the leads was provided by the company itself.

Data Cleaning

The data was cleaned using standard procedure. The missing values were imputed using cases specific analyses.

Visualization

Variables were visualized using bar plots mostly. These showed the spread of the variables and bifurcated the variables into converted and not converted.

Upon visualization, some of the variables showed high class imbalance. As these might impact the final results, they were dropped from further analysis to make the results of regression class neutral.

In certain cases, where the number of observations for a particular category was lesser, new umbrella categories were also created. This enhanced the overall predictive power in the analysis and refined the process.

Model Building

Visualization though yielded useful insights; the need for making a model for classification of leads was felt further in the analysis. Thus it was decided to go with a logistic regression model owing to it's simplicity in implementation and explanation.

Scaling was done to bring variables of different magnitude on comparable scales.

To select relevant features for the model, an approach with combination of Recursive Feature Elimination (RFE) and manual judgment was used.

- 1st Model: The first model gave good results, yet one variable was having a p value > 0.05 i.e. was insignificant in the analysis. Hence it was dropped building the second model.
- 2nd Model: In the second model, all variables were significant. However, two of the variables showed high VIFs. Thus they has to be dropped for variable inflation for building the third model.
- 3rd Model: In the third model, all variables were significant and all VIFs were within permissible, i.e. <5.

Hence, the third model was taken as the final model.

Values for the Train Data:

ROC Curve Value: 0.97

Accuracy : 92.17%

Sensitivity : 88.43%

Specificity : 95.00%

The ROC curve also gave a good fit. Thus the model parameters on train data gave satisfactory performance.

From plot of accuracy, sensitivity and specificity, the cutoff point was found out to be 0.3.

Thus all leads with conversion probability of >0.3 shall be taken as converted, and not converted for case otherwise.

Values for the Test Data:

Accuracy : 92.55%

Sensitivity : 92.08%

Specificity : 92.84%

Thus the test and train results show good agreement.

At the end, the parameters governing leads conversion were obtained through a logistic regression model.