Lead Scoring Case Study – A report

The aim of the Case Study was to arrive at a Logistic Regression Model that increases the lead conversion rate of an education company, from 30% to a target of about 80%.

A brief report of the various steps followed in building the model are given below:

Data Inspection

The dataset was analysed with the shape, describe and info methods. Also, the percentage of missing values was calculated, after replacing 'Select' with Null since this effectively meant that the user had not selected any option.

Data Cleaning

This involved dropping columns in the dataframe that had a high value of Null values and also those columns that had a heavily skewed distribution of data. In addition, a few rows that had a high percentage of missing values were also dropped. Some of the continuous variables that had missing values were imputed with Median or Mode as needed.

For columns like Country and City, where the data was heavily skewed towards one category, the remaining categories were combined under 'Other' or an appropriate category in order to ensure clean and readable data.

EDA and Outlier treatment

Where applicable, appropriate boxplots and pairplots were used to bring out visually, the presence of outliers or correlation between variables. Outlier treatment was done to make the dataset more cohesive and boxplots were charted after the treatment to ensure the same.

Data Preparation

As part of Data Preparation, the data was first split into train and test datasets. The model was built using the train dataset. Dummy variables were used to replace the categorical variables and the train data was then scaled. Scaling makes a difference to the regression model when the predictor variables have large values and are brought to a uniform scale.

Data Modeling and Evaluation

Using the Stats model, logistic regression was applied to the train dataset after which Recursive Feature Elimination was done as also manual RFE using Vif and Probability score.

The overall accuracy of the model was found to be 76%. An ROC curve was plotted to show the variance of sensitivity with specificity. Also, the area under the curve was found to be 0.78. A heatmap was done again and here, there were no highly correlated variables in the model. After checking values of accuracy, sensitivity etc for various probabilities, 0.3 was found to be the optimal probability cutoff.

The model was then tested on the test dataset where an accuracy of ~71% was obtained.

Conclusion

The Logistic Regression model arrived at has no multi collinearity among the variables and the overall model accuracy is at 77% and the probability threshold is at 0.3 which gives us a dependable model that very nearly approaches the target accuracy given, of 80%.

It has been found that the main predictor variables are:

1. Welingak Website
2. Reference
3. Had a Phone Conversation