SUMMARY

Lead scoring is a technique used by businesses to assess the likelihood of a lead becoming a customer. By using lead scoring, companies can prioritize their sales efforts and focus on the leads that are most likely to convert. In this case study, logistic regression was used to develop a lead scoring model for a B2B software company.

As a part of the Lead Scoring case study, we have been presented with the details how the company X Education pursues customer leads from various sources and tries to convert them to potential customers. The current conversion rate is quite low at 30%. So, we have been tasked to analyze the data and come up with a model which can make predictions to the order to 80% Lead conversion. For this, we have proceeded with the basic analysis of the given data set.

These are the steps that we have followed in the case study and got following conclusions.

- Identifying the columns based on Data Dictionary.

- Elimination of invalid / redundant columns

- Removing records with > 30% missing data

- Replacing Select with NAN values in all the columns

- Identifying the potential data columns which can factor in for accurate prediction

- Identifying the relationship and distribution of column data using graphs

- Removing the outliers in numerical variables

- Plotting heat map to see the correlations

Later, we proceeded with encoding the categorical data into Dummy variables so that we can easily convert them into features which can be fed into a Model used for predictions. The Others, Unknown values that transformed into columns are dropped from the Dummy columns. The data is then split into training and test data in ratio of 70:30. The training data is scaled to avoid any disparities in magnitude of the data values impacting the model prediction. The training data is fed into a Generalized Linear Model (GLM). The ineffective variables are eliminated using RFE and VIF. We are then left with 12 variables + 1 constant which has been able to predict the training data set at more than 76% accuracy and precision 45%. Then we have found the optimal cut off using our graph. We found the optimal cutoff as 35 as the lead score. The same model has been applied to test data set after the test data has been scaled. And we have also observed more than 78.4% accuracy & precision 70% there as well.