Summary of Lead Score Case study

The aim of X education is to find potential leads and convert them into a hot lead. We proceeded with data import and data cleaning. We performed univariate analysis on most of the variables and eliminated the ones, which are not helping in the prediction.

We also eliminated the columns with null values or treated them with a suitable substitution of values from the respective columns. While doing so, we found out that most of the requests came from Mumbai. Many of the customers who were interested in the course were from Financial Management. In addition, people were looking for Better Career Prospects while choosing a course. The majority of the people who were interested were Unemployed.

After data cleaning, we performed univariate analysis to understand how each variable behaved with the Converted column. We created dummy variables for the categorical variables. We split the data into train and test data. In feature scaling, we found out the churn rate to be 37 percent. After this, we did the Feature selection using RFE. Here, we assessed the model with statsmodel. We dropped unnecessary column, and got the predicted values on the train set.

We created a new column Predicted, which predicts the churn probability as 1 if it is greater than 0.5 else 0. We created a confusion matrix and predicted the overall accuracy. We also checked for the VIFs, and calculated the true positive, true negative, false positive and false negative. From this, we calculated the sensitivity, specificity, false positive rate, positive predicted value, negative predicted value. We also plotted the Receiver Operating characteristic (ROC) curve.

We found the optimal cutoff point where accuracy, sensitivity and specificity met, to be around 0.275. Lead Score was assigned varying from 0 to 100, where 100 means the highest probability of a lead getting converted into a hot lead. In addition, precision and recall was also calculated.

Using the same model, predictions were made on the test model. The accuracy was checked for the test model, which turned out to be around 0.897, which is a pretty good accuracy for the model.