Summary

This analysis is done for X Education and to track down ways of getting more industry experts to join their courses. The given data provided us with a lot of info about how the potential clients visit the site and their conversion rates, the time they spend there, how they arrived at the website etc.

The existing data had a few missing values which were handled by filling them with appropriate values. A few of the null values were changed to ‘not provided’ so as to not lose much data. Since there were many from India and few from outside, the elements were changed to ‘India’, ‘Outside India’ and ‘not provided’.

Visualization techniques were used and a quick EDA was performed to check the condition of the data. It was observed that a large number of components in the categorical variables were unimportant. The numeric qualities appear to be great and no anomalies were found. Dummy variables were created and later on the dummies with ‘not provided’ elements were removed.

Then the data split was done at 70% and 30% for train and test data respectively. Firstly, RFE was done to attain the top 15 relevant variables. Later the other factors were taken out manually relying upon the VIF values and p-value. A confusion matrix was made. Later on the optimum cut-off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

The prediction was done on the test data frame and with an optimum cut-off as 0.35 with accuracy, sensitivity and specificity of 80%. This technique was likewise used to review and a cut-off of 0.41 was found with a Precision of around 73% and recall of around 75% on the test information outline.