Summary

We have carried out the analysis for the X education company which sells the online professional courses to working professionals. The data was provided to us about the customers eg.: how they visit the website, time they spend there, how they purchase the courses like converted or not. There are lot of the other details are also mentioned. In the last we have also added some conclusion.

We have followed the below steps in this particular project:

Step 1. Cleaning data:

We have loaded the data on Jupyter notebook and performed our 1st step of process that is cleaning the data. We have gone through the data and find out that there are some null and missing values there. So we treated these columns by replacing and dropping the columns. Few values replaced by the not provided so that we can perform analysis and get better result. Also we have added some dummy variable. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

Step 2. EDA:

After completing 1st step we have carried out our second step which is exploratory data analysis. In this step we have analysed univariate and bivariate analysis of categorical and numerical variables. We have also plotted some graphs for categorical and numerical variables.

3. Dummy Variables:

The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

4. Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

5. Model Building:

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

6. Model Evaluation:

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.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.37 with

Accuracy, sensitivity and specificity of 80%.

8. Precision – Recall:

This method was also used to recheck and a cut off of 0.41 was found with Precision around 74% and recall around 77% on the test data frame.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.

2. Total number of visits.

3. When the lead source was:

a. Google

b. Direct traffic

c. Organic search

d. Welingak website

4. When the last activity was:

a. SMS

b. Olark chat conversation

5. When the lead origin is Lead add format.

6. When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy the courses.