Summary

The analysis is done for X Education and to find ways to get more people to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the leads that are converted.

The following are the steps used:

1. Cleaning data:

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' and 'not mentioned' so as to not lose much data. Since there were many from India and few from outside, the elements were changed to 'India', 'Other Countries' and 'Not mentioned'.

2. EDA:

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and a few outliers were found.

3. Dummy Variables:

The dummy variables were created and later on one of the variable was dropped with drop_first = True condition. For numeric values we used the MinMaxScaler for scaling the data sets.

4. Train-Test split:

The split was done at 70% and 30% for train and test data respectively with a randomness as 100.

5. Model Building:

Firstly, the model was developed and then RFE was performed to attain the top 20 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:

First the cut-off was assumed to be 0.5 and then predictions were made. Later a confusion matrix was made. Later on the optimum cut off value (using ROC curve) was found to be 0.33 and then used to find the accuracy, sensitivity and specificity which came to be around 75% to 80% each.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.33 with accuracy, sensitivity and specificity of around 80%.

8. Precision – Recall:

This method was also used to recheck and a cut off of 0.4 was found with Precision around 73% and recall around 75% on the test data frame.

It was found that the variables that mattered the most in the potential buyers are:

1. The total time spent on the Website.

2. Total number of visits.

3. When the lead source was:

    a. Google

    b. Olark chat

    c. Organic search

4. When the lead origin is Lead add format.

5. 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 their courses.