Summary

The analysis performed for X Education and find ways to get more people to subscribe courses. The data provided gave lot of information that how potential customers can visit the site, how time they spend, how they come to the site and how leads are converted.

Steps followed

1. Data Cleaning:

The data was clean except some null values and select options which has to replaced with a null value since it did not give us much information. Few 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.