Summary

This analysis is done for X Education and to find ways to get more industry professionals 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 conversion rate.

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. We observe that there are 'Select' values in many columns. It may be because the customer did not select any option from the list, hence it shows 'Select'. 'Select' values are as good as NULL. So we can convert these values to null values. Dropping some missing value. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies.

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 seem good and no outliers were found.

3. Dummy Variables:

The dummy variables were created and later on the dummies with 'not provided' elements were removed. Dropping the columns for which dummies were created

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.

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 80.4 to 81.7% each.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.34 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 79% and recall around 70% on the test data frame