SUMMARY

The X Education wants to join more industry professionals to their course. To achieve the agenda of company they provided basic in information in the form of data about how customers visit the sites, how much time they are spending on the particular website, how customer reached the site also what is the conversion rate .

We used the following steps :
1.Data cleaning :

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’ so as to not loose much data.
Although they were later removed while making dummies.
Since there were many from India and few from outside, the elements were changed to ‘India’, ‘Outside India’ and ‘not

2. EDA(Exploratory Data Analysis):

EDA was done to check the condition of our data in terms of different graphs for better understanding of insights of data .After the analysis ,It was found that a lot of elements in the categorical variables were irrelevant.
The numeric values looks 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.
For numeric values we used the MinMaxScaler.

4.Train-Test split:

The first basic step for regression is performing a train-test split, we have chosen 80:20 ratio.

5.Model Building:

RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value i. e the variables with VIF < 4 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 0.40 with accuracy, sensitivity and specificity of above 80%.

8. Precision – Recall:

This method was also used to recheck and a cut off of 0.48 was found with Precision around 75% and recall around 80% on the test data frame.