SUMMARY

Our analysis is done for X Education and aims to find ways to attract more industry professionals to join their courses. The basic data provided gave us a lot of information about how potential customers visit the site, the time they spend there, how they reached the site, and the conversion rate.

The following are the steps we 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. The column 'Country' had more than 96% data as Null or India, so we dropped this column.

2.EDA (Exploratory Data Analysis):

A quick EDA was done to check the condition of our data. It was found that many elements in the categorical variables were irrelevant. The numeric values seemed good, but a few outliers were found in the 'Total Visits' and 'Page views per visit' columns, so we performed outlier treatment on them.

3.Dummy Variables:

Dummy variables were created for columns with more than 2 categories, and we dropped columns with binary categories which had very high data imbalance.

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 dataframe, and with an optimum cut-off as 0.3, the accuracy, sensitivity, and specificity were around 80%.

8.Precision – Recall:

This method was also used to recheck, and a cut-off of 0.39 was found with precision and recall of around 71% each on the test dataframe.