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 customers visit the site, the time they spend there, how they reached the site and the conversion rate. The following are the strategy used:

1. Data Cleaning and Preparation:

The dataset was mostly cleaned, but there were a couple of null values, and we decided to replace the option select with a null value since it offered minimal information. Some of the null values were converted to 'not provided' to retain more data. However, these were ultimately discarded during the process of creating dummy variables. Given that there were many entries from India and a few from other countries, the categories were modified to 'India', 'Outside India', and 'not provided'.

2. EDA:

An initial exploratory data analysis (EDA) was performed to assess the quality of our data. It revealed that many entries in the categorical variables were not applicable. The numerical values appeared satisfactory, and no outliers were detected.

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 or StandardScaler.

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 90% each.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 82%.

8. Precision – Recall:

This method was also used to recheck and a cut off of 0.41 was found with Precision around 73% and recall around 75% on the test data frame.