SUMMARY

This analysis is done for X Education and to extract more information about professionals joining courses. The data which is provided gives a lot of information about how customers visit the site and how much time they spend there.

The steps that are used:

1. Data Pre-processing:

All the null values are handled by dropping some and by changing some columns to unknown so as to not lose much data. However they are later removed by label encoder (by creating objects).

2. EDA:

In order to check the condition of our data EDA was done. The elements in the categorical variables were found to be irrelevant and the numeric values were found to be good and hence no outliers were found.

3. Label Encoder:

Instead of creating dummy variables, label encoder was used for the elements to be removed. MinMaxScaler is used for numeric values.

4. Train-Test Split:

The split was done at 70% and 30% for train and test data respectively.

5. Model Building:

RFE was applied to 15 relevant variables which are most important. Later on the variables which are removed manually are based on the VIF values and p-value.

6. Model Evaluation:

A Confusion matrix was made. Later on ROC curve was used to find the accuracy, sensitivity and specificity which was 80%.

7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.30 with accuracy, sensitivity and specificity of 75%.

8. Precision – Recall:

Again to recheck the model, this method was used and a cut off of 0.40 was found with Precision around 68% and recall around 71% on the test data frame.