

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 i.e.,40% null 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 "Select" values were changed to "Nan".

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.

3 Dummy Variables:

The dummy variables were created and later on the dummies with 'other' elements were removed. For numeric values we used the MinMaxScaler.

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 $p\text{-value} < 0.05$ 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.ROC Curve & Cut-off:

Here, AUC curve came around 0.96 means, max area is under the curve.

& Optimal value as per the curve came around 0.25 so we have taken threshold as 0.25 for further calculations.

8.Prediction:

Prediction was done on the test data frame and with an optimum cut-off as 0.25 with accuracy 89%.

9.Final Model with top 3 variables:

After evaluating final model.

Top 3 variable came by computing model are as follows.

1. Tags_Closed by Horizon
2. Tags_Will revert after reading the email
3. Total Time Spent on Website