Summary

This analysis ids done for X Education and to find ways to get more industry professionals to join their courses . The basis 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 except for a few null values and the option select had to be replaced with null value since it did not give us much info. Few of the null values were changed to not provided somas to not lose much data. Although they were later removed while making dummies. since there were many from India and few from outside, the elements were changed India , Outside India not provided.

EDA:

A quick EDA was done to check the condition of our data . It was found that lot of elements in the

categorical variables were irrelevant. The numeric seems good and outliers were found.

Dummy Variables:

The dummy variables were created and latter on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

Train-Test-split:

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

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 and p-values (The variables with VIF< 5 and p-value <0.05 were kept).

Model Evaluation:

A confusion matrix was made .Later on the optimum cut off value(using ROC) was used to find

the accuracy, sensitivity and specificity which came to be around 80% each .

Prediction:

Prediction was done on the test data frame and the with an optimum cut off as 0.35 with the accuracy, sensitivity and specificity of 80%.

Precision -Recall:

This method was also used to recheck and cut off 0.41 was found with precision around 75% on the test data frame