SUMMARY

This summary is about the analysis done for X Education in order to get a higher lead conversion i.e to get more industrial professionals/customers to subscribe to their online courses.The basic data of 9000 data points are provided .This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

The following steps are implemented:

Importing of libraries and dataset:we imported all important and pivotal libraries like numpy,pandas,matplotlib.pyplot,seaborn and dataset which is csv file by using pandas.

Cleaning Data: Initially Prospect ID & Lead Number are dropped since these two variables are just indicative of the ID number of the Contacted People .later we dropped columns whose missing values are greater than or equals 45% and the option select is replaced with null values since it did not give us much information.

EDA: A quick EDA was done to analyze the condition of data . It was found that a lot of elements in categorical variables were irrelevant. We also found few outliers in numeric variables.

Dummy Variables: Dummy variables are created on a few categorical variables and later on dummies with not provided elements are removed.For numeric variables we implemented StandardScaler.

Train-Test split: The data was split at 70% and 30% for train data and set data respectively.

Model Building: RFE was done to obtain the top 15 relevant variables.Later the rest of the variables were removed manually depending on the VIF and P-values where variables with VIF>5 and p-value>0.05 were dropped.

Model Evaluation: A confusion matrix was made. Later on using the ROC curve we can find optimum cut off value which is used to find the accuracy,sensitivity,specificity.

Prediction: Prediction was done on the test data and with an optimum cut off as 0.30 with accuracy,sensitivity and specificity of 92%,91% and 92% respectively.

precision-Recall: This method was used to recheck the model and optimum cut off value was 0.41 with precision around 89% and recall around 91% on the test data frame.