***Summary***

This analysis is done for X education to target potential leads.

Input data gives the information about customer activities like how they reached to the website, how many times they have visited and how much time they spent on it, occupation and lead conversion rate etc.

The following steps were followed:

1. **Reading Data**

Data is read from CSV file and stored in pandas’ data frame.

1. **Cleaning Data**

Data was cleaned except for some null values. ‘Select’ is changed to null value in some columns. For some columns null values are imputed with median. Grouping is also followed for some columns

1. **EDA**

From EDA, it is observed that there is no specific relation in categorical variables. Also some outliers are found in case of numeric variables, Such outliers are capped to 90th percentile.

1. **Creation of Dummy Variables**

Dummy variables are created for Categorical Variables and MinMaxSaler is used for numeric ones.

1. **Creation of Train and Test Dataset**

Entire dataset is divided in train (70%) and Test (30%) datasets.

1. **Model Building**

RFE was done to attain top 15 relevant variables. Later variable having VIF >5 and p-value > 0.05 were removed manually.

1. **Model Evaluation**

A confusion matrix is created, optimum cut off value (using ROC curve) was used to find the accuracy (72.18), sensitivity (85.44) and specificity (64.01) .

1. **Prediction**

Prediction was done on test data set and with optimum cut off as with accuracy (72.47), sensitivity (86.48) and specificity (63.32).

1. **Precision: Recall**

Used to recheck and a cut off of 0.4 was found.

It is also observed that variables mattered the most in finding potential buyers are:

1. Total time spent on website
2. Total number of visits
3. When a lead source was,
4. Google
5. Direct Traffic
6. Organic Search
7. When the last activity was SMS or Olark chat conversation
8. When current occupation is working professional