Summary

This analysis is regarding X education firm which tries to find the most potential leads in an effective manner without spending resources and time on unturned leads. So the Logisitic regression model is employed to find the probability scores against each customer to find out the potential leads.

Below are the steps carried out :

1. Data cleaning
   It is done to understand the data and find out the missing values and filter the unwanted columns for our analysis. Few columns have select as value and we replaced it with 'User has not provided'. Null value rows are removed as they cause problem while fitting the RFE model.

2. EDA

   This is performed to visualise the data of different categorical variables using countplot and we find out India has most number of customers.

3. Dummy Variable
   Dummy variables are created for categorical columns.
4. Model Building
   Train- Test splits (70% train data set and 30% test data set) are created to build the model and after scaling has been done on the datasets and we selected top 15 features using RFE approach as we have lot of columns. After selecting the top 15 features, few more columns are deleted based on p-value and VIF values. The approach is carried out till we have decent p-values and VIF values.(p-value < 0.05 and VIF < 5)
5. Model Evaluation
   We have taken arbitrary cut-off of 0.5 to find the predicted values based on the probability of each customer.
   Confusion matrix is created and below metrics are calculated

- Accuracy
- Specificity
- Sensitivity
- Precision
- Recall

Test set metrics

I. accuracy = 77%¶
II. precision = 74%
III. recall = 80%

Train set metrics

I. accuracy = 78%
II. precision = 80%
III. recall = 73%