SUMMARY REPORT

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. So, to make this process more efficient, the company wishes to identify the most potential leads. So as a data analyst we have been provided with a leads dataset from the past. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Brief Summary: -

1. Data cleaning and Preparation.
2. Test -train Split and Scaling.
3. Model Building
    a. Feature elimination based on correlations.
    b. Coarse tuning using RFE.
    c. Manual fine tuning using p-Values and VIF's.
4. Model Evaluation
    a) Accuracy

b) Sensitivity and Specificity

c) Threshold determination using ROC.

d) Precision and recall.

5. Prediction on Test data.

Now based on the above business requirement we will proceed with the steps in creating Model which will predict whether the lead will get converted or not.

1. The first step we followed was reading the data:

   After reading the dataset we found, it is having 9240 rows and 37 columns. Rows over here signifies the number of leads while the columns are the attributes based on which we need to predict if the leads will opt for the course or not.

2. Data Preparation: -

   As Model need clean data with no missing values we need to do sanity checks and missing value treatments on data before proceeding with the model building. Also, data should not have any outliers as it will impact the model and will give incorrect result. Categorical data should also be converted to numerical using binary mapping for Yes and no while for categorical variables having more than two values need to be converted to dummy variables.

3. Test-Train Split: -

   Before model building features need to be split as train and test data wherein train data will be used to give training to the model and test data will be used for testing different metrics of the model. Over here we are using 70% data as a training data and 30% as test data.

4. Feature Scaling: -

   All the columns should be scaled before model building using StandardScaler function of sklearn.

5. Model Building: -

   I. Using Stats Model of Logistic regression model have created the model.

   II. Feature Selection Using RFE.

   III. After course tuning using RFE method will do fine tuning using vif and P value method for final section of the features.

   IV. After final selection of the features will proceed with the checking of different metric of the Model like Accuracy, sensitivity, specificity.

   V. Finally, by Plotting ROC curve to get the overview of the model and Area under the curve.

   VI. Now using different Metrics find the optimal cutoff point for the model and then again check for the metrics to get the final Model.

6. So, after model building we will be making predictions on the test set and will check all the metrics on the test data set to get if our model is stable and giving accurate result or not.