SUMMARY REPORT

Introduction:

We were provided with the leads dataset with shape 9341*37. Our goal is to make a logistic regression model which will help the sales team to find out the most potential leads or the hot leads on the basis of the lead score so that they can focus on these leads with full resources and effort.

Approach:

Initially, we imported all the basic libraries and Data.csv file required for Exploratory data analysis. The following steps are used:

- **Reading and understanding** the data set with the help of data dictionary, Problem statement.
- **Importing Libraries and data:** Imported necessary libraries and data set then added further ibraries as and when required.

- **Data cleaning**: After importing the data it is properly analyzed for various data types, missing values, data imbalance etc., and necessary actions were taken by eliminating or imputing technique. Checked the uniqueness of each column individually and necessary actions were taken.

- **EDA**: Various feature behaviors w.r.t the target column 'Converted' is visualized using the plots and based on which decisions were taken to deal with certain features. Looking into skewness of some columns they were eliminated and few categorical columns having more options but less counts for each variety are then grouped to a common category.

- **Creation of Dummy Creation:** For the categorical variables dummies were created by dropping first column to convert them into numerical data type for the model.

- **Creating Train-Test data sets:** X variable is created by taking all the columns except 'Converted' and y variable includes the target column 'Converted'. Train test data are then separated to 70% & 30% respectively.

- **Scaling**: The numerical columns are then scaled using a standard scaler.

- **Feature selection**: Using RFE 15 features are selected out of a total of 65 columns.

- **Assessing the model**: Added a constant to the train data, data is fit to the logistic regression model. Various results are then checked to determine the wellness of the model. By checking the p-value and the VIF value feature are further eliminated to get the best Model.

- **Model evaluation**: Metrics – Accuracy, Sensitivity/Recall, Precision, F-1 Score, ROC AUC Score are checked to evaluate the model. A cutoff of 0.5,0.4 is first randomly provided and checked for all the metrics then changed to 0.28 cutoff as we can see in the accuracy, sensitivity and specificity intersection point 0.28 is the optimum cutoff for the model.

- **ROC (Receiver Operating Characteristics Curve):** ROC curve of True positive rate vs False positive rate is plotted to check the trend and observed that it follows a normal ROC curve shape.

- **Lead score column**: A lead score column is added to the table for the predicted probability of each customer to help the salesperson determine the priority.

- **Testing**: The model is then tested on the test data and it is then observed that test model metrics are having very close values to the train data models, thus model can be used for conversion problem.