Process of the Logistic Regression

- We start the regression by, reading and checking the data.

- Next, we check for the Null values in the data frame. We observe the data frame and see that there are multiple cells which have been filled as ‘Select’ from the problem statement, we understand that, these are the unfilled cells.

- So, technically these are also Null Values and we replace the Select with Null Values.

- Next, we drop the columns with most amount of Null Values and we have selected the threshold to be 40%.

- Next, we check the rows with most missing values with intent of removing rows with more than 70% Null Values.

- Next Stage will be the missing Value imputation. The columns with less number of missing values have to be imputed. The following process is followed for imputing
  - For Categorical Variables, missing values are imputed using **Mode**.
  - For Numerical Variables, missing values are imputed using **Mean** if the distribution of data is normal.
  - For Numerical Variables, missing values are imputed using **Median** if the distribution of data is **not** normal.

- In many variables, the total values are very high like there are some 15 to 20 Variables also. We can reduce these based on the percentage of the value counts. These with minimal percentage are bunched into a single category of Others.

- Also, we checked the value counts on all the categorical variables and checked for Skewed Variables, which means the variable which has most of the variables in a single category. These variables are not much of a use for the model. So, we drop all these columns.

- Next, we go for the Univariate and Bivariate analysis for performing Exploratory Data Analysis on the data we have to check the variables which are most dependant and correlated and drop if there are most correlated variables. But, there are not highly correlated variables in the data frame.

- After that, we drop the variables that were added by the Sales team, since, as per Problem Statement, we need to Analyse data from the internet only, so we drop the columns such as Tags etc which were added by Sales Team.

- Then, we perform the binary mapping of the variables and also create dummy Variables.

- After the above step, we go for train and test split and Scaling of the variables using Standard Scaler.

- Then, we start with the ranking of the variables using RFE.

- Next, we go for modelling of the data, create the predicted variables and make a Confusion Matrix and check the accuracy of the Model and check the VIF of the variables.

- Based on the P Value and VIF, we can start eliminating features which are not much helpful for the Model. We repeat this process till there is no much change in the accuracy of the model.

- Next we check Sensitivity and Specificity of the Model and plot the ROC Curve and find optimal cut-off point.

- Next, we check the accuracy, sensitivity and specificity of the model at the cut off point.

- Next, we check the Precision and recall values.

- Next, we move on to the test set and get the predictions of the test data set.

- After that, we concatenate the prediction values from train and test set and merge with the original data frame and create a score column by multiplying the predictions with 100.

- Hence, we also get the score on which lead is most likely to get converted.