**STEP 1: MISSING VALUE TREATMENT**

* Missing Values in each column were treated.
* Columns with missing value > 70% were dropped straight away
* Column ‘Lead Quality’ had nearly 50% of missing values. As this column had category ‘Not Sure’, all missing values were imputed with this category
* Other columns with missing values > 40% were also dropped
* For rest of the columns, missing values were imputed with most frequently occurring value.

**STEP 2: EDA**

* Each column was analyzed with respect to the target column, i.e., ‘Converted’ column
* Importance of the column for building the logistic model was determined by analyzing each graph
* Those columns which did not have any significant effect on the target column were dropped at the end of EDA before moving on to model building

**STEP 3: DATA PREPARATION & MODEL BUILDING**

* For all categorical column, dummy columns were created, and original column was dropped.
* Dataset was split into Train and Test.
* For train dataset, numerical columns were standardized.
* Model was build using RFE, considering Top-15 columns.
* After that, each column’s p-value was checked and those column with higher p-values were dropped manually.
* Final model was arrived with 13 feature variables.
* VIF of the feature variables were also checked and found ok.

**STEP 5: FINDING OPTIMAL CUTOFF POINT**

* Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

**STEP 6: MAKING PREDICTION ON TEST SET**

* Numerical columns of Test dataset were standardized.
* Final model developed was used to predict target variable of the test dataset

|  |  |  |
| --- | --- | --- |
|  | **Sensitivity** | **Specificity** |
| **Train** | **86.05** | **93.95** |
| **Test** | **84.42** | **93.88** |