LEAD SCORE STUDY SUMMARY

The Data Cleaning

- The data is made ready to use. We first removed all ambuiguities in records by
  - Identified Missing values
  - Verified for duplicated records
  - replaced invalid values such as ‘select’ to appropriate values
  - dropped columns which are missing lot of values or is biased
- The Categorical Data is converted to dummy variables and removed the least relevant dummy columns (e.g. ‘Others’)
- The Numerical data is checked for outliers using box plots and removed outliers
- columns are removed with high correlation with other columns

Model Training

- The data is split into train and test data, 70% for train and 30% for test
- The data is standardized after the train test split
- An Initial Logistic Regression model is trained and matrices are evaluated

Feature Evaluation and Model Analysis

- Features are dropped and refined by repeating the model building by closely observing VIF and p-value
- Once the VIF and p-value seemed to be within the acceptable range, We proceeded with final model evaluation
- We used confusion matrices to calculate accuracy sensitivity and specificity on train data
- We validated its balance using ROC curve
- Finalizing the cut-off point using accuracy, sensitivity and specificity plot
- Validated the cut off again with F1 Score