Summary Report:

The following document encompasses the approach to the Lead Generation Case Study.

The company required a model wherein a lead score was desired such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

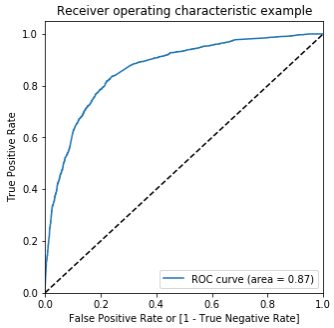
Data Cleaning:
A sense check of the data was done, after which an analysis on the columns to understand the shape, data types and content & quality of the data. The 'Select' values were converted to NaN as it indicated that the field was not selected. The columns that only had one value were deemed irrelevant as the column had no variance and would not contribute to the final model.

Missing Value & Outlier Analysis:
The missing values were analyzed and, on average, the columns that had over 25% missing values were removed as imputing these columns with statistic methods would introduce a bias in the model. Outlier analysis was performed on the numerical features.

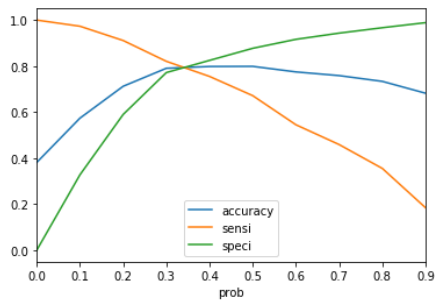
Binary Mapping, One-hot encoding and Heatmap:
The columns that had "Yes" and "No" were converted to 1 and 0 respectively. One-hot encoding was performed on the categorical features. Post this, the features that had multicollinearity were removed as they would not contribute to the model that is to be created.

Model building:
The data was then split into train and test data, post which the numerical features (not dummy variables) were scaled using the Standard Scaler. Recursive Feature elimination was used to select the top features. An additional column was also added as the statsmodels package does not account for constant. Post this, the logistic model object was instantiated, and the model was run on the train data. The p-values and Variable Inflation Factor was analyzed, and the irrelevant columns were excluded one by one and the model was re-trained while keeping the evaluation metrics in purview.

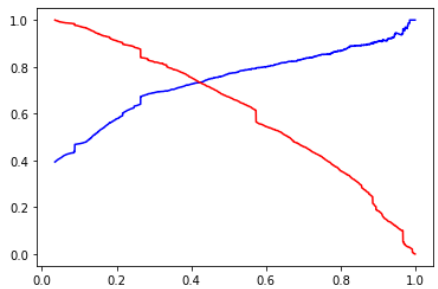
Model Evaluation:
The ROC curve was plotted to analyze the performance of the model. The Specificity-Sensitivity trade-off was also analyzed, where the Sensitivity, Precision and Accuracy was analyzed. Post this, the model was finally tested on the test data and we were able to achieve an accuracy of about 80%.



ROC Curve



Sensitivity-Specificity Plot


Precision-Recall plot