Summary Report

Step- 1 : Reading and understanding the data

1. The given excel sheet is read using pandas and converted into the data frame.
2. The total number of rows present in the data frame is checked first.
3. The feature variable and the categorical variables are identified.
4. The nature of the values present in each column (null, Select) is identified and the way of treating them is analyzed.
5. The columns which has no role in analysis are identified.

Step-2 : Data Cleaning and Transformation:

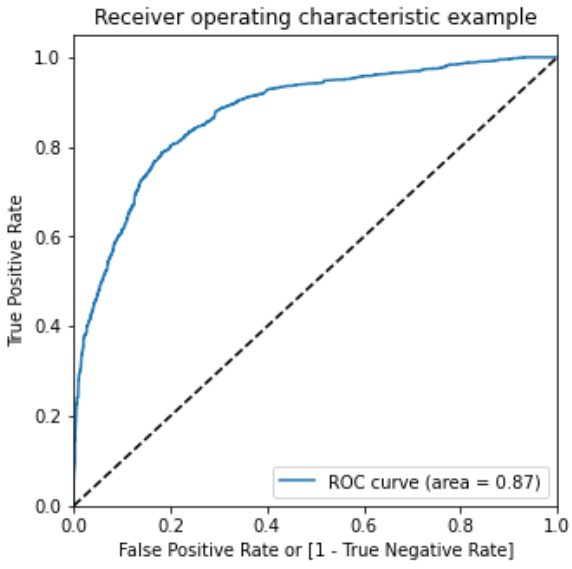
1. At first the categorical variables which might play a major role in the analysis is identified.
2. The variables are as follows,
    - Total Visits,
    - Total Time spent on website,
    - Page Views per visit

3. The 'Select' entries present in the each column is handled appropriately in the columns.
4. The columns which doesn't contribute to our data analysis are removed.
5. The null values and improper entries are analyzed and then the data is removed.
6. Then the data is divided into Converted and Non-Converted values for better visualization and understanding.
7. The relationship between the Lead Source, Current Occupation, Specialization, User Activity on the website is analyzed and visualized for better understanding.
8. The columns which has a null value more than 35% is removed.
9. The dummy values are created for the categorical variables using the pandas and the redundant columns are removed.
10. The rows having the null values are checked and removed from the data frame.

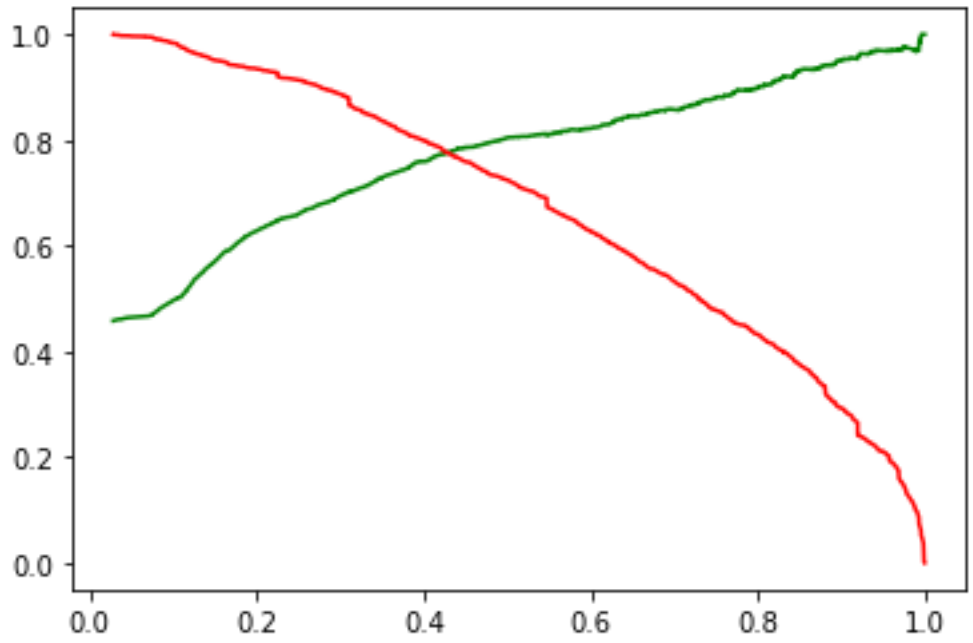
Step-3: Data Preparation and Model building:

1. The data is split into train-test model.
2. Using RFE, the top 15 variables are selected from the train data.
3. Then the columns which has higher p-values are removed from the train model and the new model is generated.
4. The VIF analysis has been performed on the model and the VIF values are very low (<2) which means the correlation between the variables are very low.
5. The predicted values has been obtained from the train set and the conversion probabilities are identified.
6. Then using confusion matrix, accuracy, sensitivity, specificity has been calculated as 0.8, 0.72, 0.86.
7. The ROC curve has been plotted and the area under the ROC curve has been found to be 0.87, which is a good ROC value.
8. Then the cut-off value is plotted for accuracy, sensitivity and specificity and the value is found to be 0.4.
9. Using it as the optimum point, the separate column is added and populated with a boolean variable to denote whether an applicant can be converted to lead or not.
10. The confusion matrix has been generated for the final predicted values and precision, specificity, sensitivity, accuracy and recall values are calculated as 0.76,0.82,0.77,0.80 and 0.77.

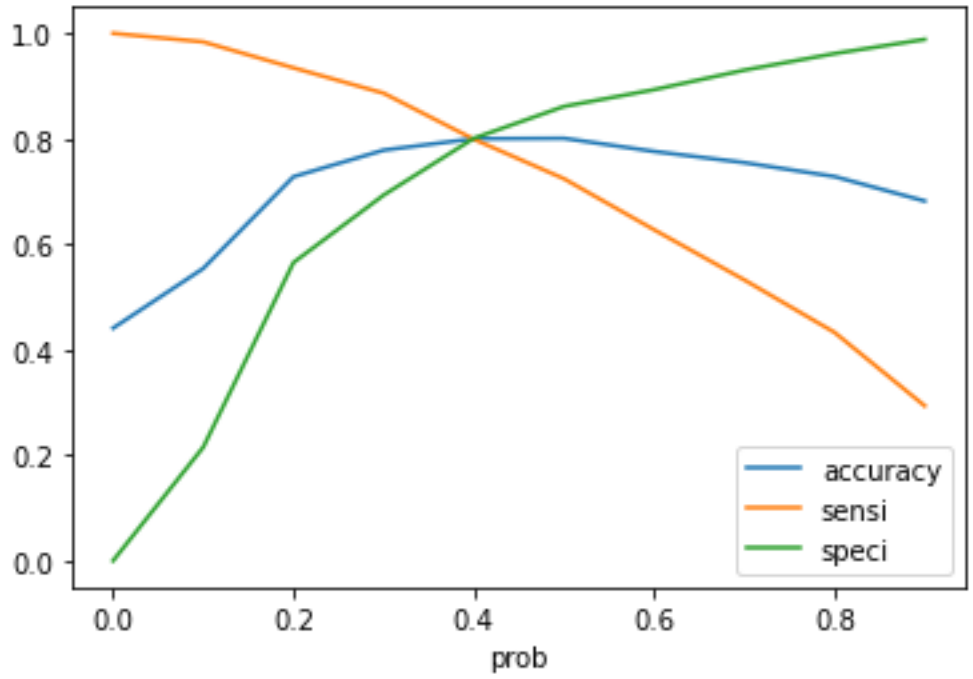
ROC Curve:



Precision-Recall Curve:



Probability Cut-off curve:



Step-4: Conclusion:

1. The model has been built with an RFC count of 15.
2. The predictions has been performed on the test set and the values of accuracy, sensitivity, specificity and precision are found to be in the acceptable range.
    - The major categorical/dummy variables that should be focused are, Total Time Spent on Website
    - Lead Origin_Landing Page Submission (Direct traffic)
    - Page Views per visit