upGrad

Lead Scoring Case Study

ISHAAN BHATTACHARYA &
HRISHITA SINGH

Problem statement

- Create a model in such a way that the customers with high lead score have higher conversion chance and low lead score have lower conversion chance. The ballpark of the target lead conversion rate is around 80%.
- Also the model should be able to adjust if the company's requirement changes in near future.

Analysis Workflow

1. Reading The Dataset
2. Missing Value Treatment
3. Outlier Analysis
4. Exploratory Data Analysis
5. Data Preparation
6. Model Building
7. Making Predictions On The Test Set
8. Scoring Each Lead
9. Conclusion

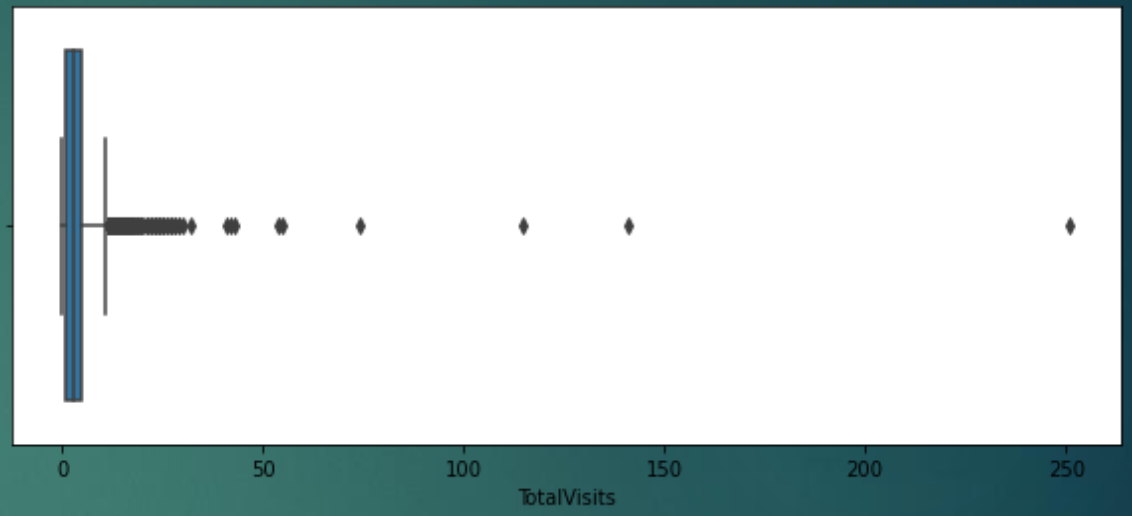
Missing Values

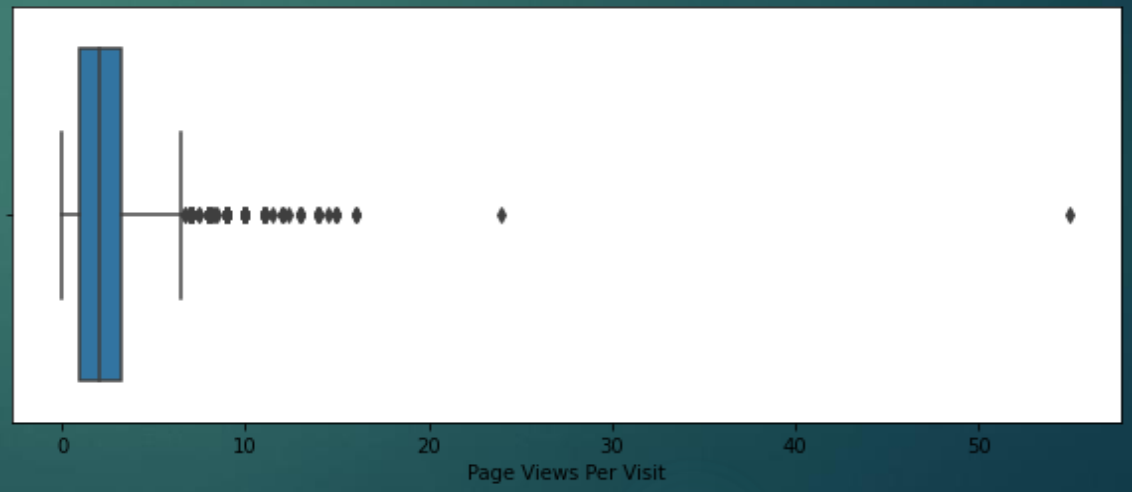
- There were a number of columns with missing values.
- Columns with 30% or more missing values were dropped.
- Categorical Columns which were heavily skewed and had missing values were dropped.
- For the columns having low percentage of missing values, the rows having missing values were dropped.

Outlier Analysis

As we can see in the images on the right, the columns 'TotalVisits' & 'Page Views Per Visit' have a good number of outliers.

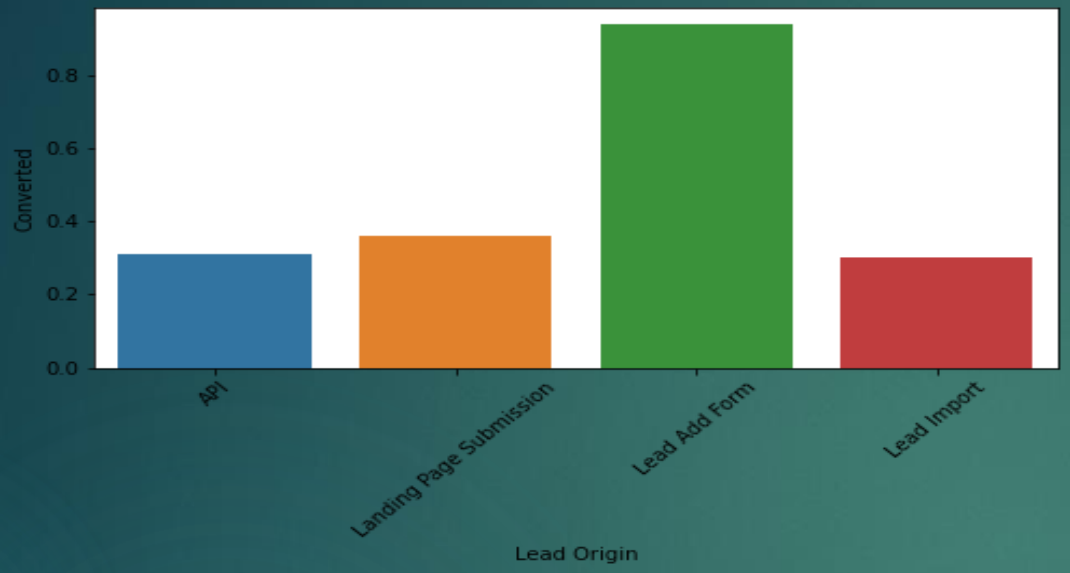
For this analysis we are not removing the outliers.



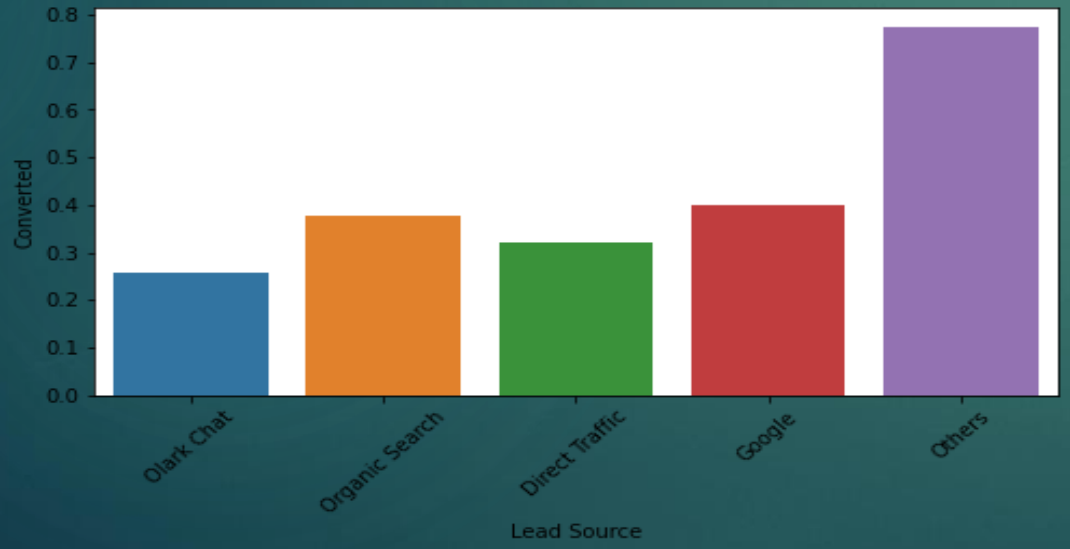


Exploratory Data Analysis

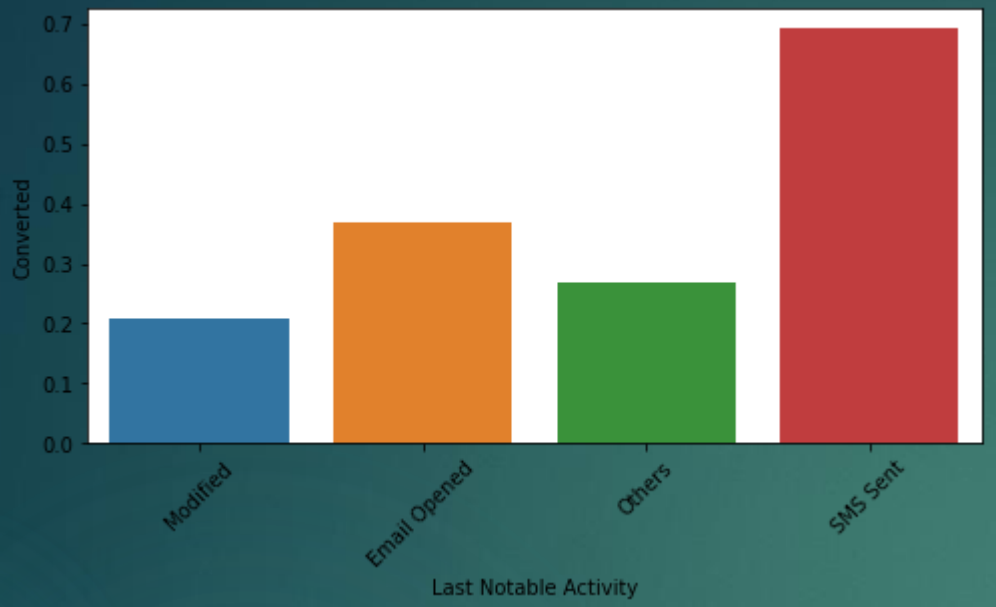
- Over the next few slides we will present the different plots and the insights derived from them.
- The features having significant differences between different categories are important for our analysis.



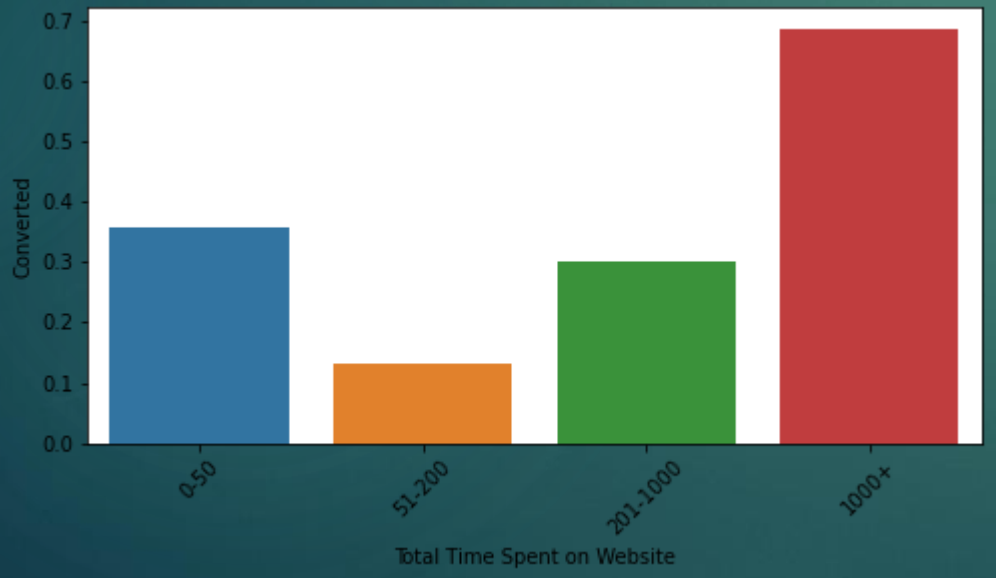
Those who have Lead Add Form in Lead Origin have a 80% + chance of getting converted.



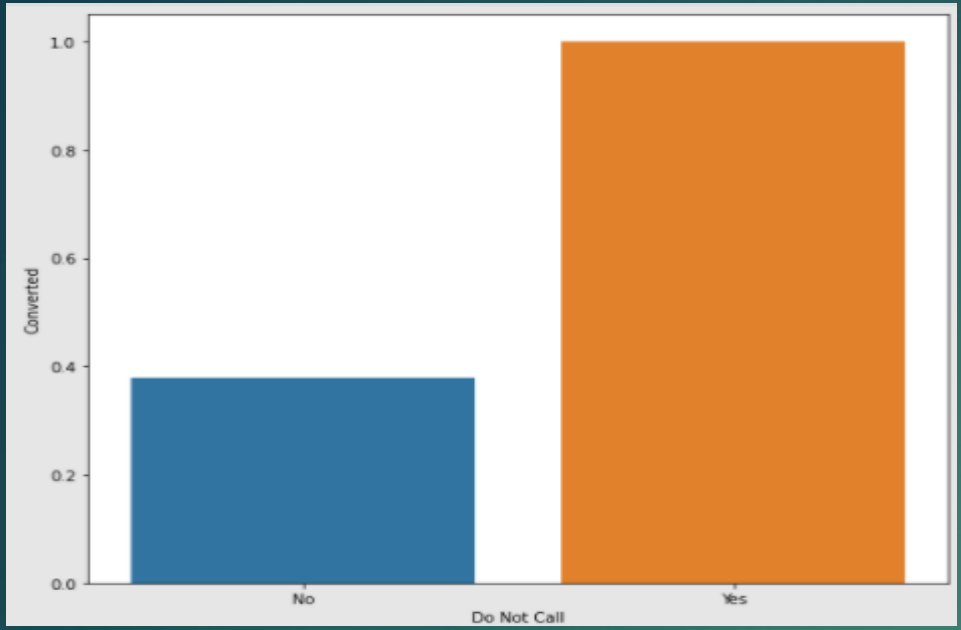
Those who have ‘Others’ in ‘Lead Source’ have almost 80% chance of getting converted.



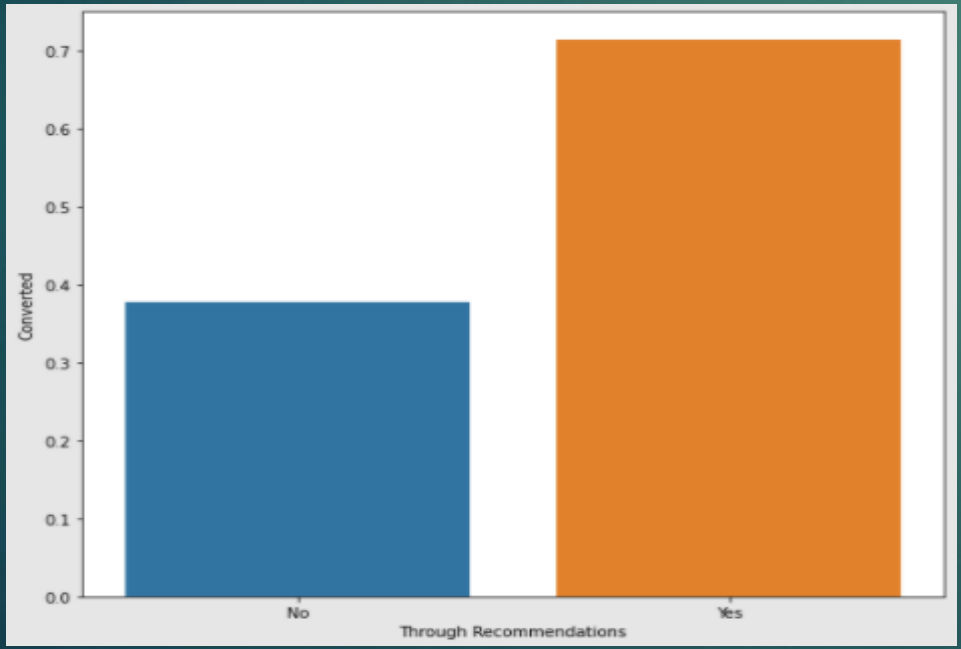
Those who have ‘SMS Sent’ in ‘Last Notable Activity’ have almost 70% chance of getting converted.



Those who have 1000+ in ‘Total Time Spent on Website’ have almost 70% chance of getting converted.

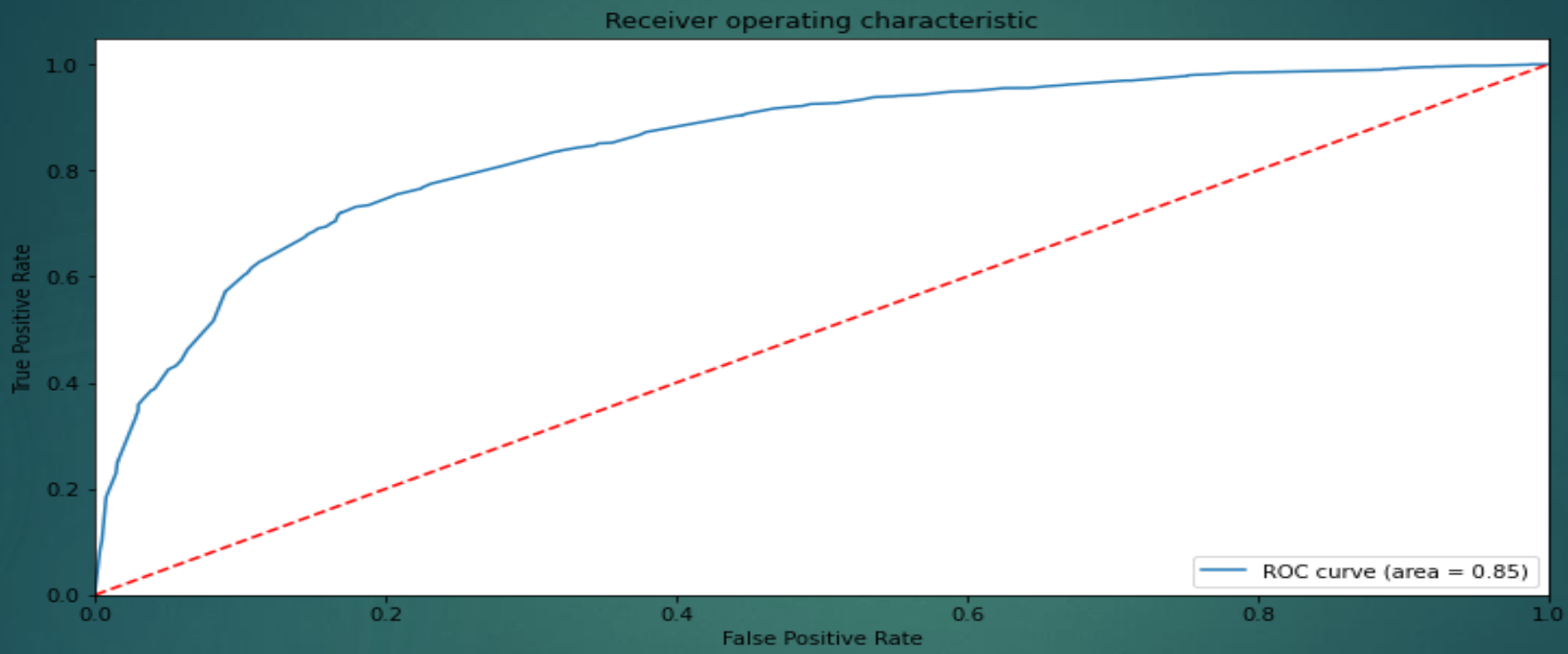


Those who have ‘Yes’ in Do Not Call have almost 100% chance of getting converted.

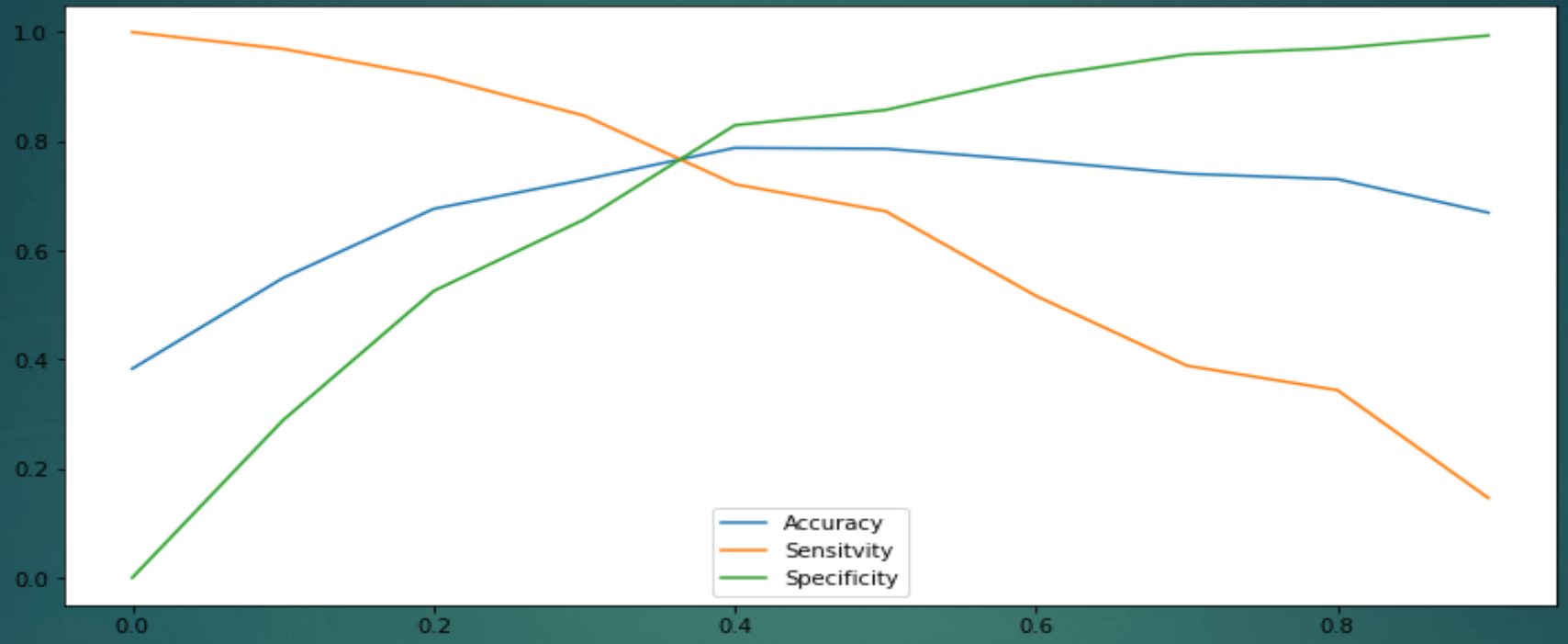


Those who have ‘Yes’ in ‘Through Recommendations’ have 70% chance of getting converted.

Final Model



From the ROC Curve we can see that the Area Under Curve is 0.85 which is good.



We can see that at approximately 0.4 value of probability Accuracy, Sensitivity, Specificity converge. So we have set the probability cut off to be 0.4.

Conclusion

- For the test set the values of sensitivity and specificity are in acceptable range.
- In business terms, our model is having stability in predicting whether a lead will be converted or not.
- Also it can adjust with company's changing requirement in the coming future.