Lead Score Case Study

Logistic Regression

Submitted by:
Varshini Srinivasan
Srilathaa Vasu

Problem Statement:

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

When people visit their website, they browse the courses or fill up a form for the course or watch some videos. Among all, the people who fill up a form providing their email address or phone number, are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal:

They want to select most promising leads that can be converted into paying customers.

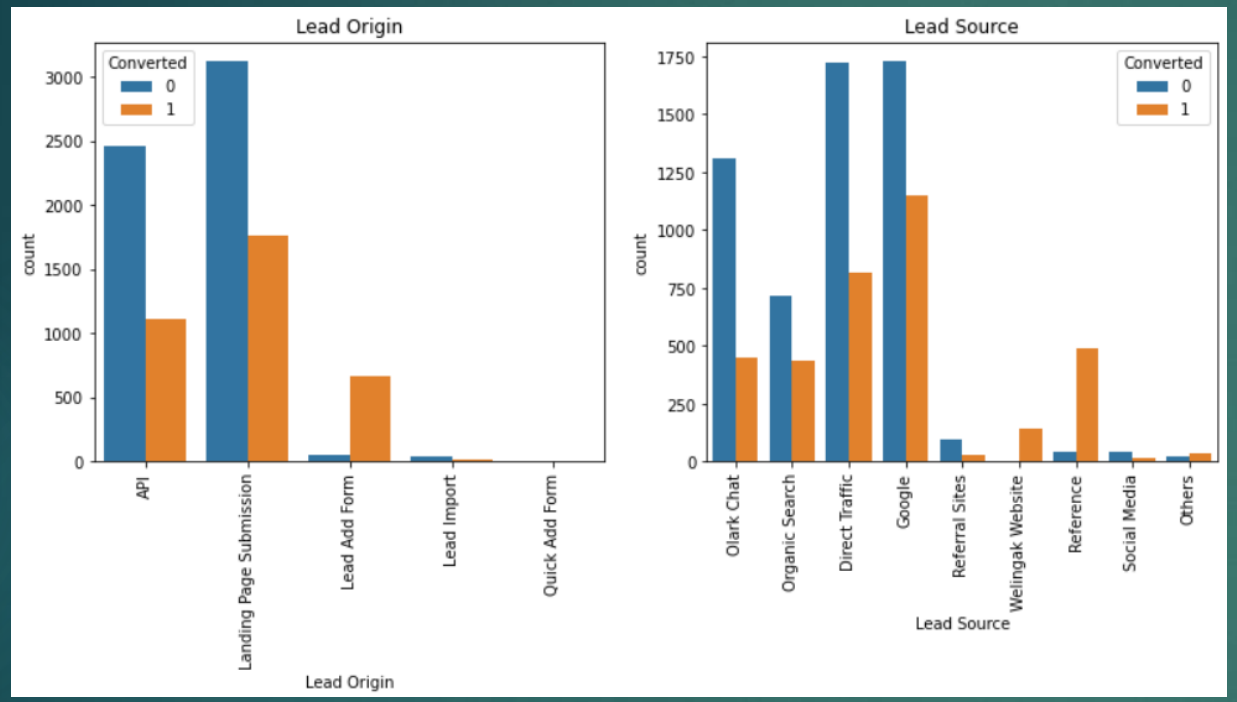
Lead score should be assigned to each lead, so that they determine how promising the lead could be. Higher the lead score, higher the probability of the lead to get converted into promising customer.

The model is expected to have a target lead conversion rate to be 80%.

Strategy

- Understanding the problem
- Data Collection, which includes importing data from different sources
- Cleaning and Preparing the data for analysis
- Exploratory Data Analysis
- Feature Scaling
- Preparing the data for model building
- Splitting the data into train and test dataset
- Build a logistic regression model and assign lead scores
- Model Evaluation using metrics such as- Specificity, Recall and Sensitivity or Precision
- Applying the best model on Test data based on Specificity and Sensitivity metrics
- Measure the accuracy of the model along with other metrics for overall evaluation

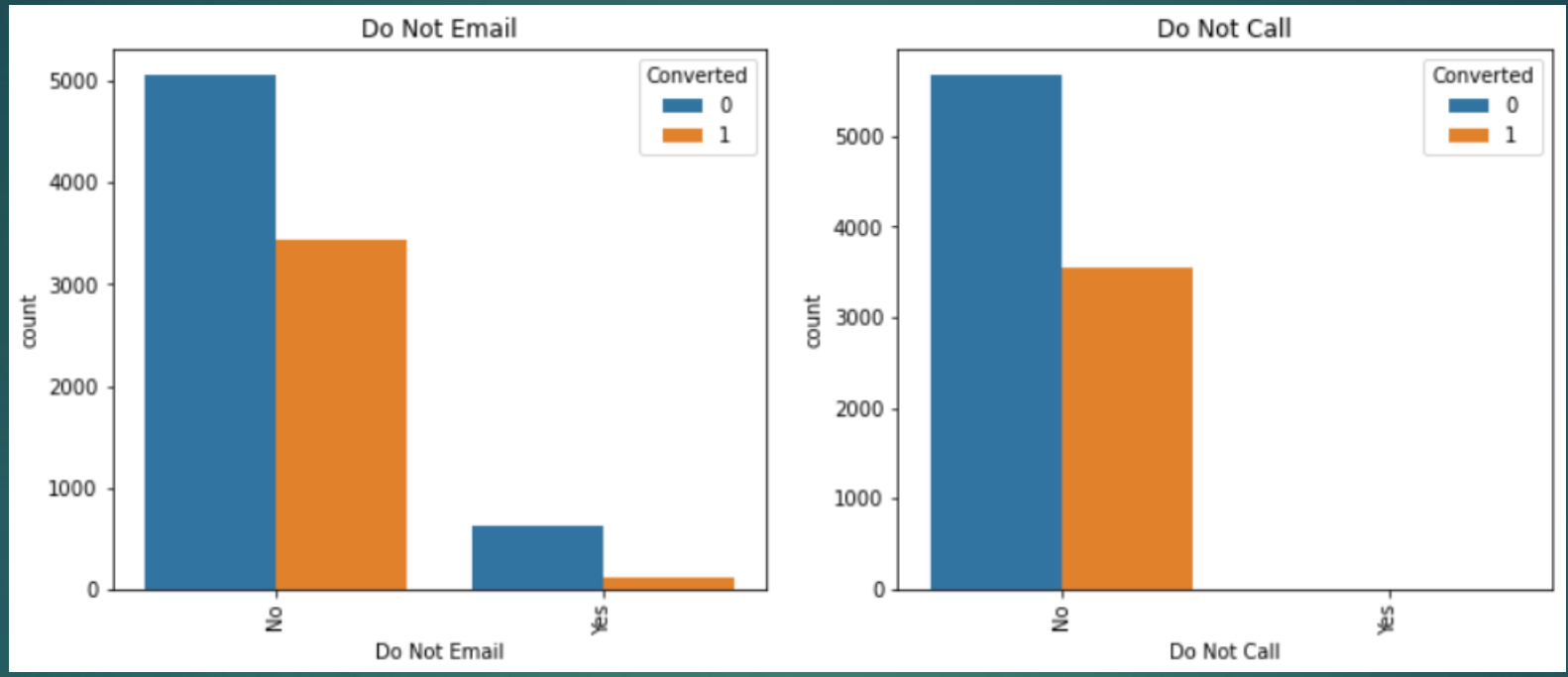
Exploratory Data Analysis



- Lead Origin Vs Converted
  - Customers from "Landing Page submission" have highest conversion rate as compared to others.

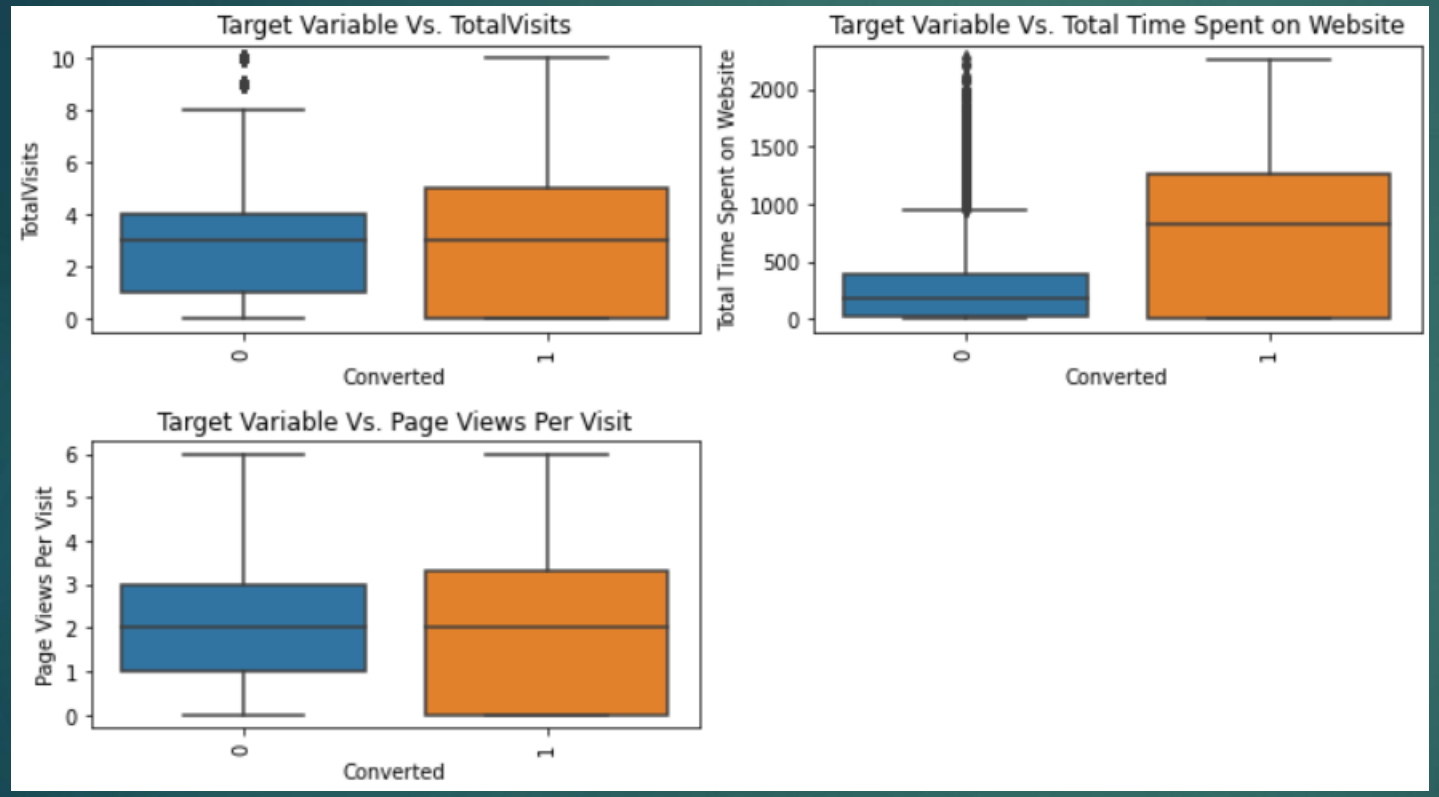
- Lead Source Vs Converted
  - Google searches have highest conversion probability.

Exploratory Data Analysis



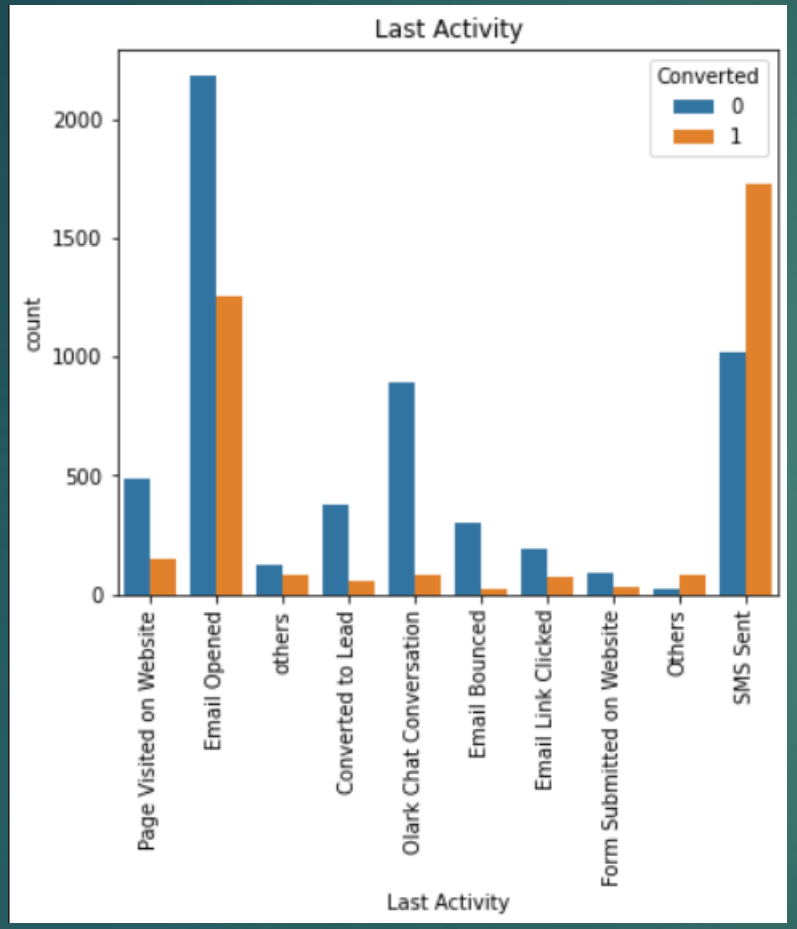
- Customers preferring not to mail and not to call are maximum in number.
- Customers who do not opt for Do Not call have Higher conversion rate which is around 38%.These constitute the majority of the leads.

Exploratory Data Analysis



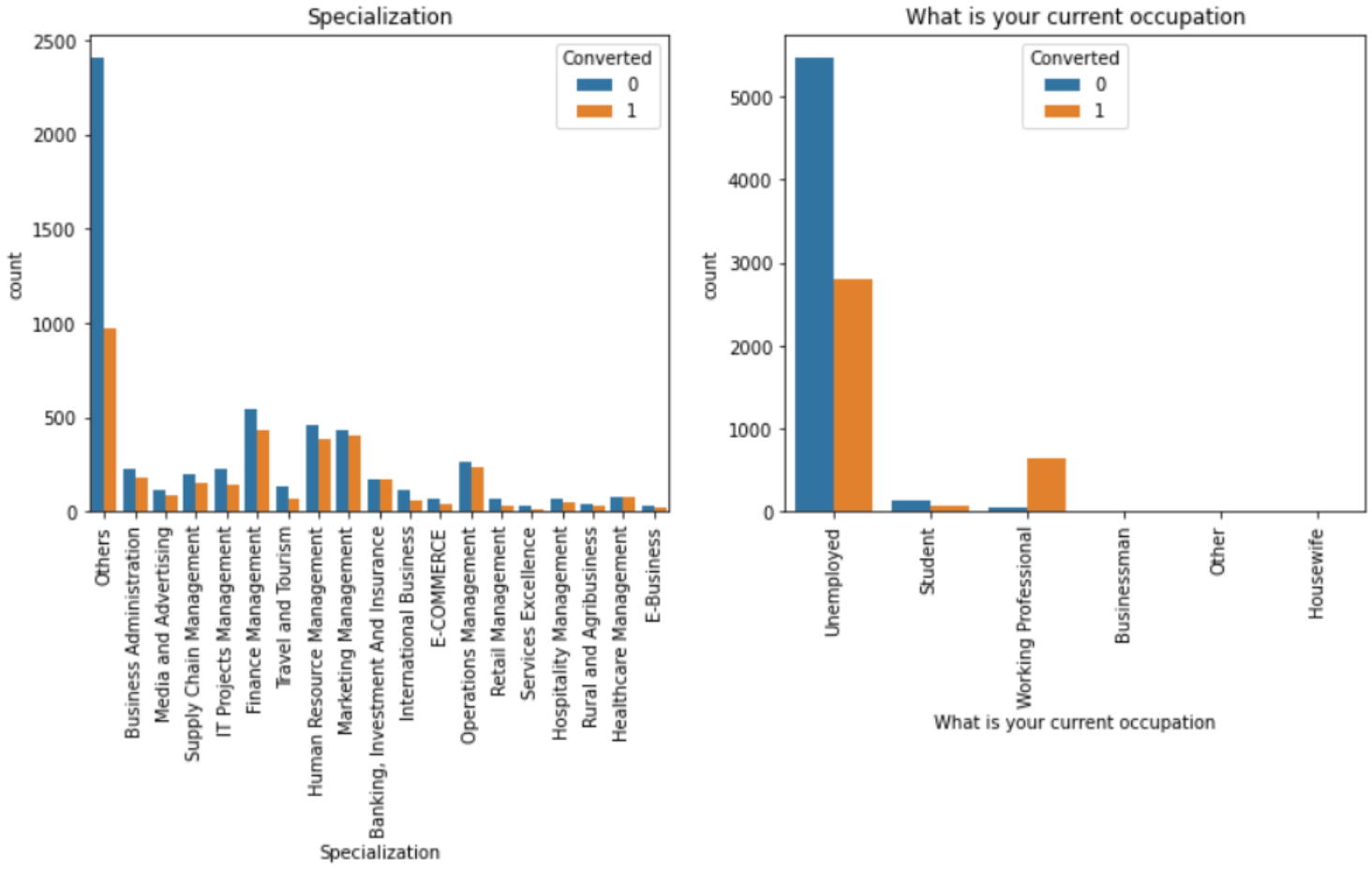
- Total Visits & Pages Views Per Visit Vs Converted:
  - Median for both types of Leads : converted and non converted are similar.
- Total Time Spend on Website Vs Converted:
  - Leads who spends most of the time on website have higher chances for conversion.

Exploratory Data Analysis



- **Last Activity Vs Converted**
- Customers who last activity was SMS Sent have higher conversion rate.
- Customers who last activity was Email Opened constitute majority of the customers.

Exploratory Data Analysis



- **Specialization Vs Converted**
  - Leads with specialization in Management & Others have maximum conversion rate, while Rural and Agribusiness leads have minimum.

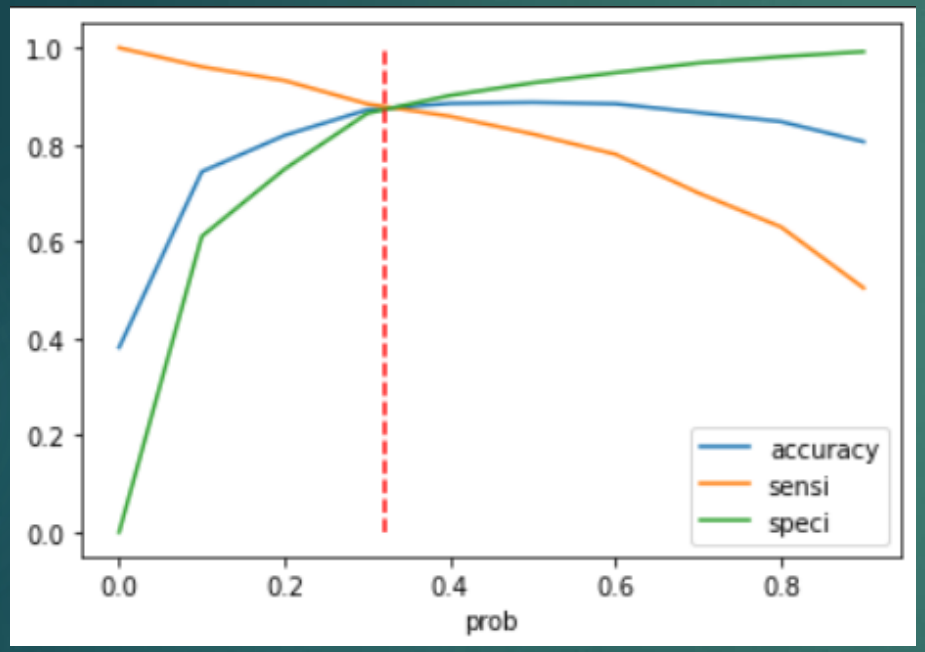
- **What is your current occupation Vs Converted**
  - People who are unemployed seemed to be converted the most. While housewife seems to have least conversion rate.

Variables Impacting the Conversion Rate

- Lead Number
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Reference
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- Specialization_Others
- Tags_Closed by Horizzon
- Tags_Interested in other courses
- Tags_Ringing
- Tags_Will revert after reading the email
- Lead Quality_Low in Relevance
- Lead Quality_Might be
- Lead Quality_Worst
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional
- Last Notable Activity_Modified
- Last Notable Activity_SMS Sent

Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.32 based on Accuracy, Sensitivity and Specificity



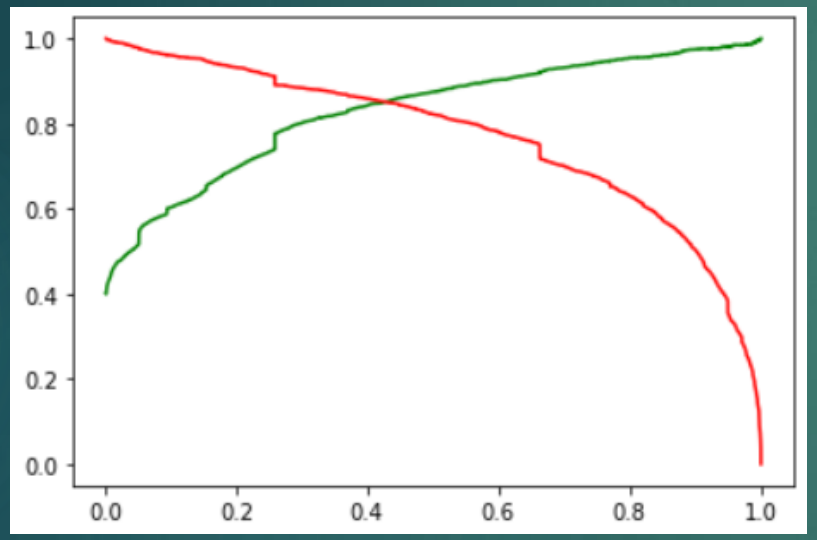
Confusion Matrix

3496	506
295	2171

- Accuracy - 87%
- Sensitivity - 88 %
- Specificity - 87 %
- False Positive Rate - 12 %
- Positive Predictive Value - 81 %
- Negative Predictive Value – 92%

Model Evaluation- Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.42 based on Precision and Confusion Matrix Recall



Confusion Matrix

3712	290
439	2027

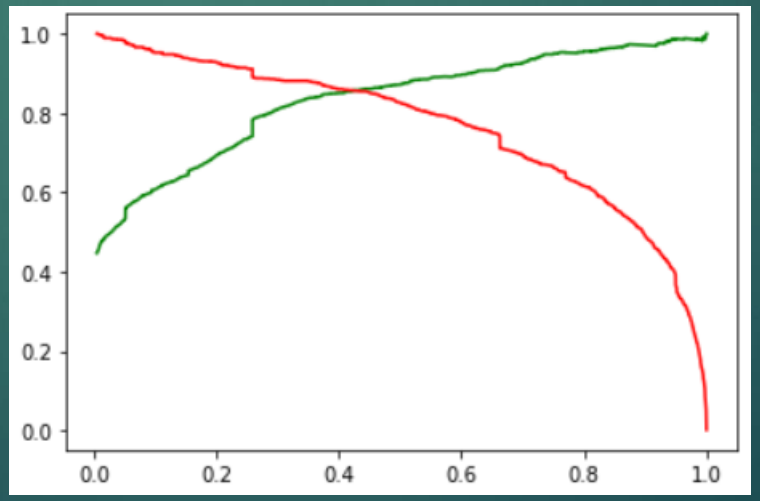
- Precision - 87%
- Recall - 82 %

Model Evaluation – Sensitivity and Specificity on Test Dataset

Confusion Matrix

1465	212
130	965

- Accuracy - 87%
- Sensitivity - 88 %
- Specificity - 87 %



- Precision - 81%
- Recall - 88 %
- F1 - 84 %

CONCLUSION

- Model has an accuracy, sensitivity and specificity of 87% , 88 % and 87 % when calculated using train dataset.

- The threshold was selected on the basis of Accuracy, Sensitivity, Specificity, precision and recall curves.

- When model was run in Test Data, we obtained the following results:
  - Accuracy: 87.68%
  - Sensitivity: 88.12%
  - Specificity: 87.35%

- The top three variables contributing to convert a lead are:
  - Tags_Closed by Horizzon
  - Total time spent on Website
  - Lead Origin_Lead Add Form

- The top three variables that seems to need improvement are:
  - Tags_Ringing
  - Lead Quality_Worst
  - Tags_Interested in other courses

Overall model seems good.