Lead Score Case Study

Submitted by :

Biswajeet Singh

Maninder Singh

Pushpa Monica

Problem Statement :

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they 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:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Steps Followed to achieve the Goal :

- Load the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling.
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Problem solving Methodology :

1.Data sourcing and Cleaning

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization

2.Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

3.Model Building

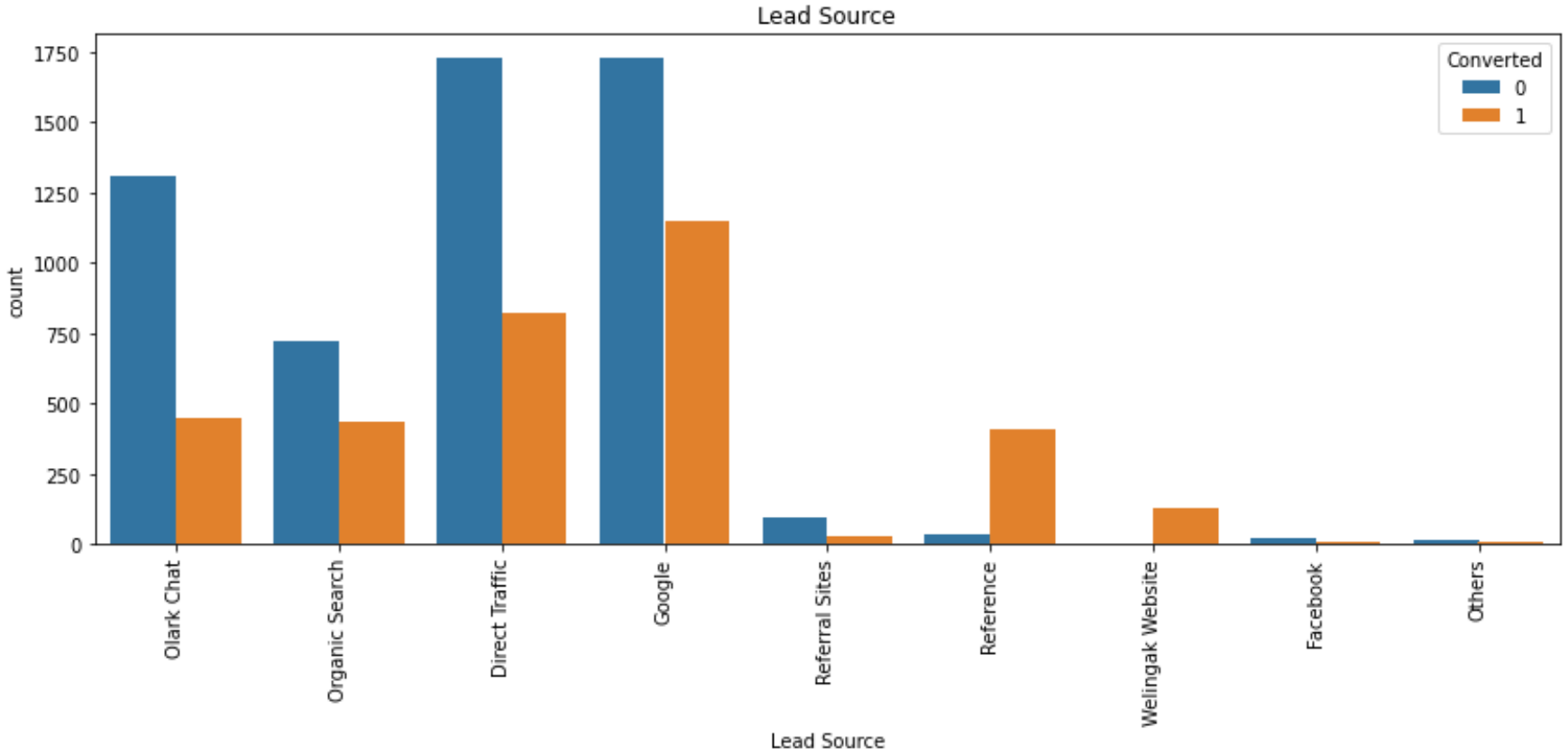
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

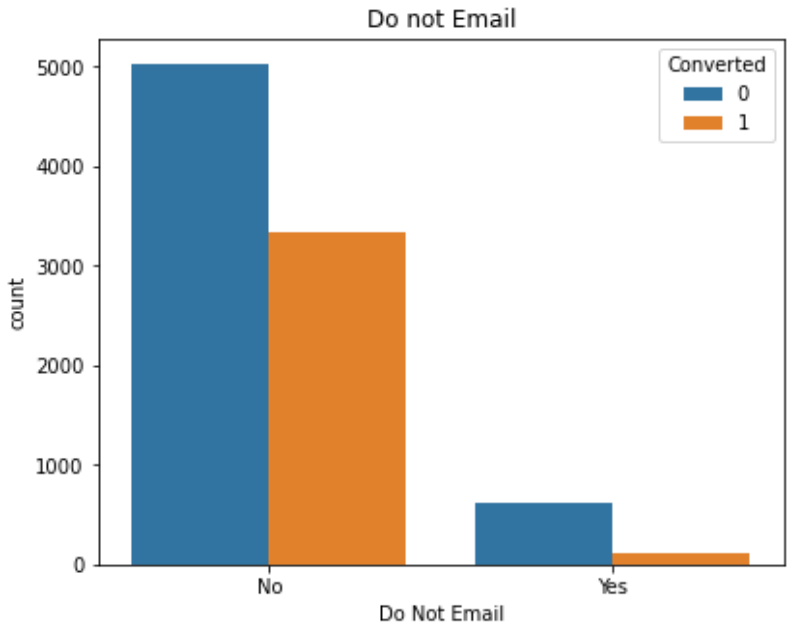
EDA :

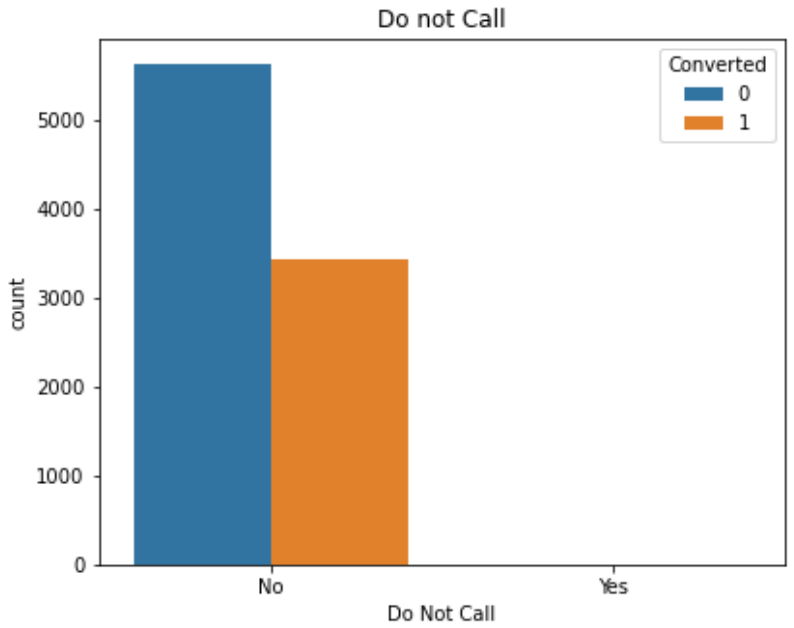
- Google and Direct Traffic generates maximum number of leads
- Conversion rate of reference and Welingak website is high
- To Improve overall lead conversion rate, we should focus on improving lead conversion of Olark chat, Organic Search, Direct Traffic, Google and we need to generate more leads from reference and Welingak website



EDA :

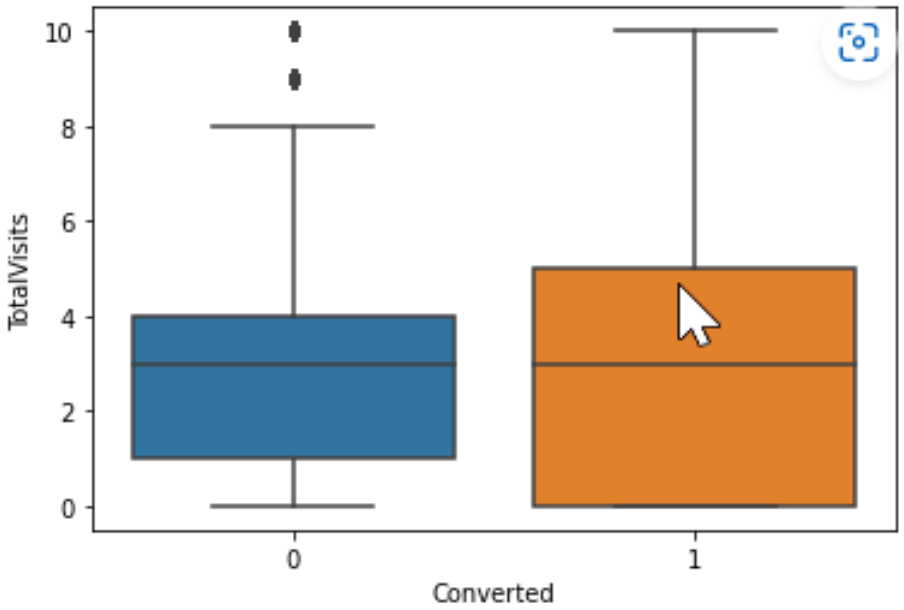
- Most for both the enteries are 'NO'. No inference can be drawn with this parameter





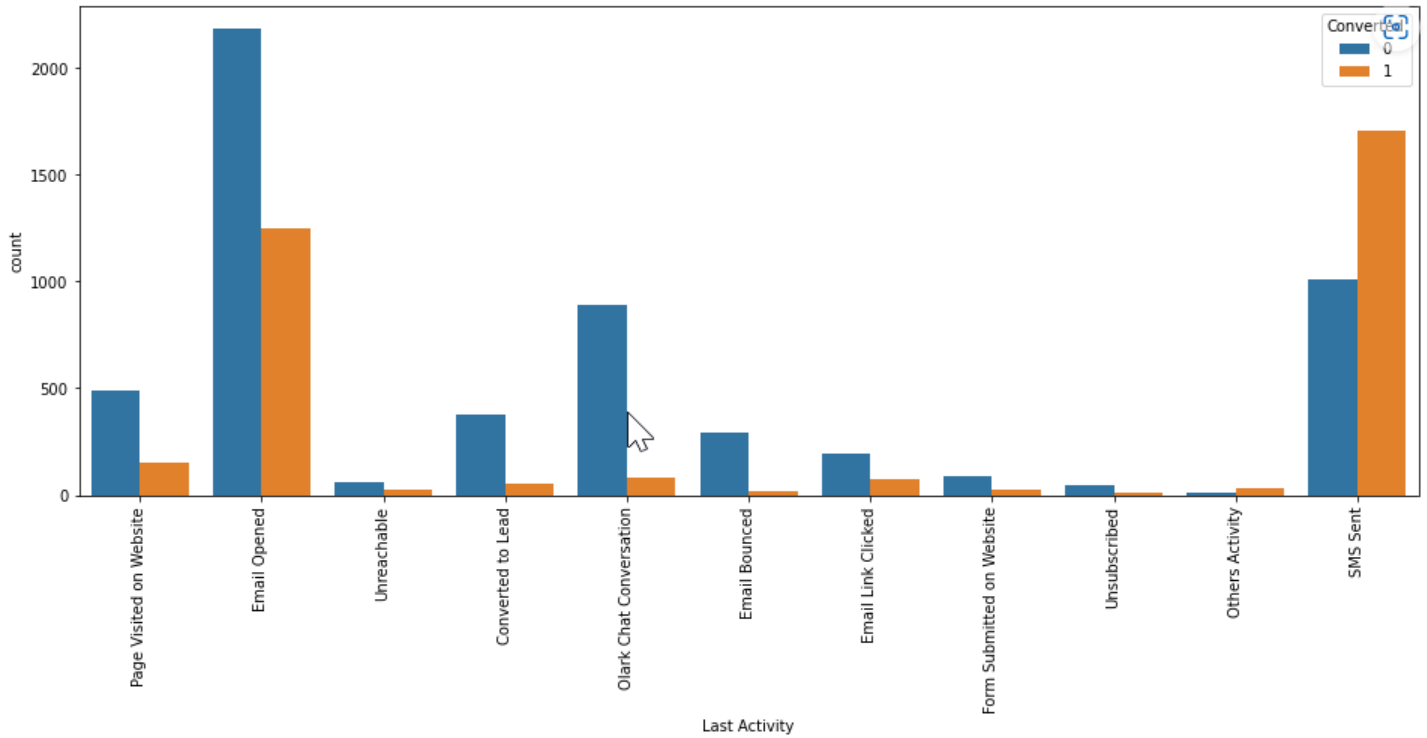
EDA :

- Median for converted and not converted leads are the same
- Nothing can be Concluded on the basis of Total Visits



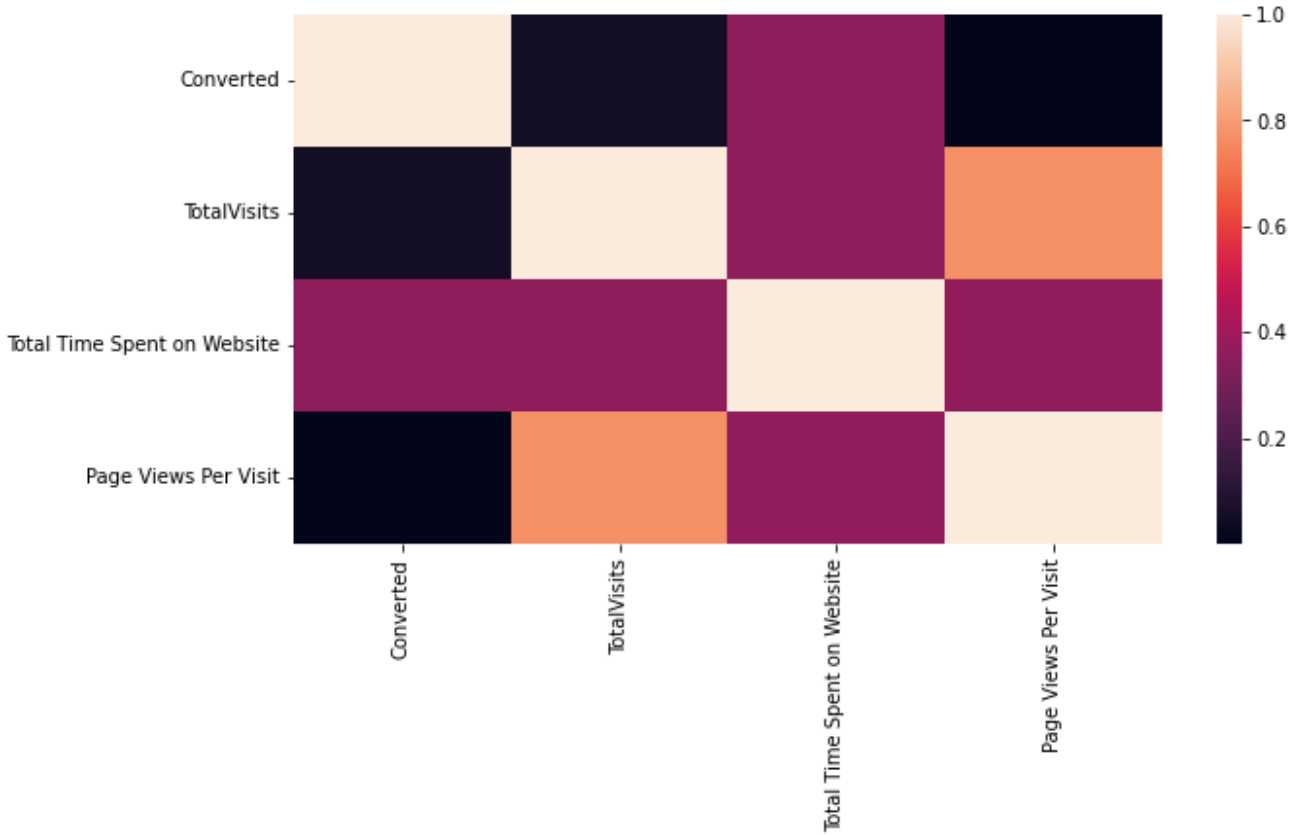
EDA :

- Most of the lead have their Email Opened as their last activity
- Conversion rate for leads with last activity as SAMS sent is almost 60%.



Correlation Matrix :

It is understandable from the above EDA that there are many elements that have very little data and so will be of less relevance to our analysis.

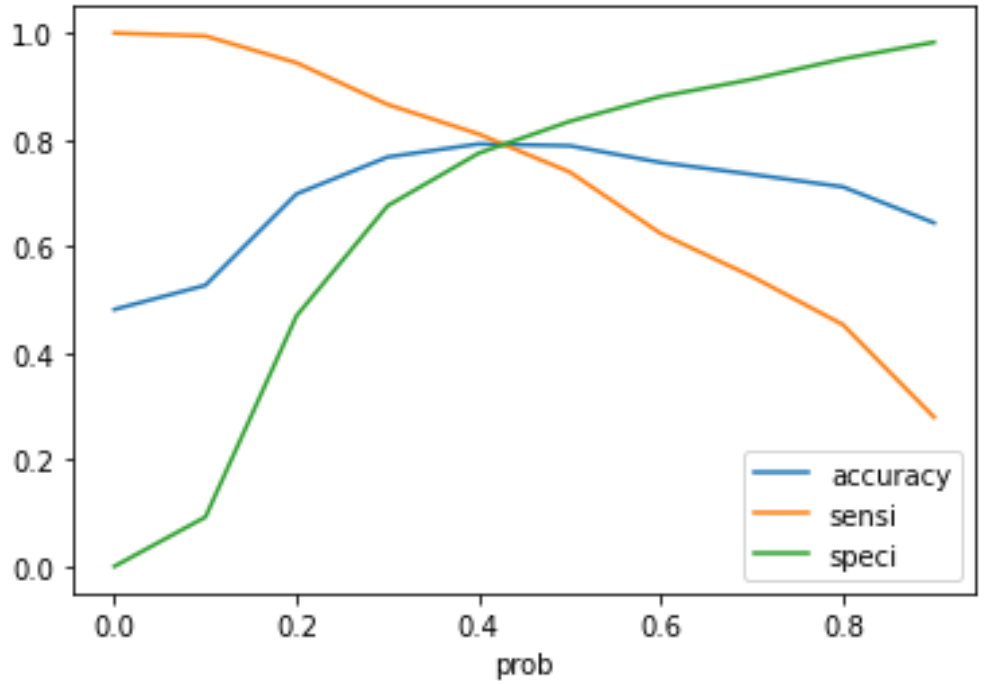


Variables Impacting the Conversion Rate :

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity – SMS Sent
- Current Occupation – No Information
- Current Occupation – Working Professional
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity - Unreachable

Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.34 based on Accuracy, Sensitivity and Specificity



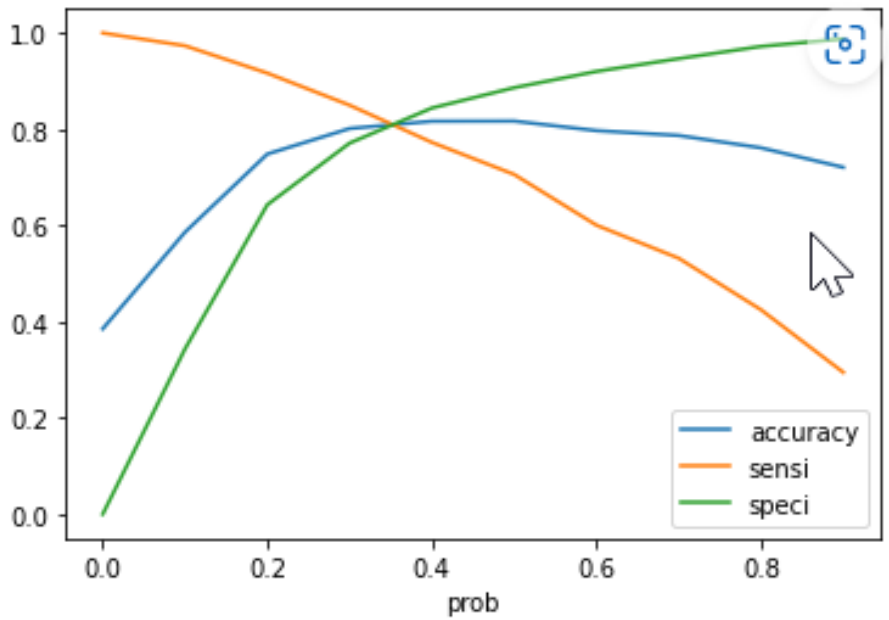
Confusion metrics :

```
array([[3151,  754],
       [ 447, 1999]], dtype=int64)
```

- Accuracy - 81%
- Sensitivity – 81.7 %
- Specificity – 80.6%

Model Evaluation - Sensitivity and Specificity on Test Data Set

The graph depicts an optimal cut off of 0.34 based on Accuracy, Sensitivity and Specificity



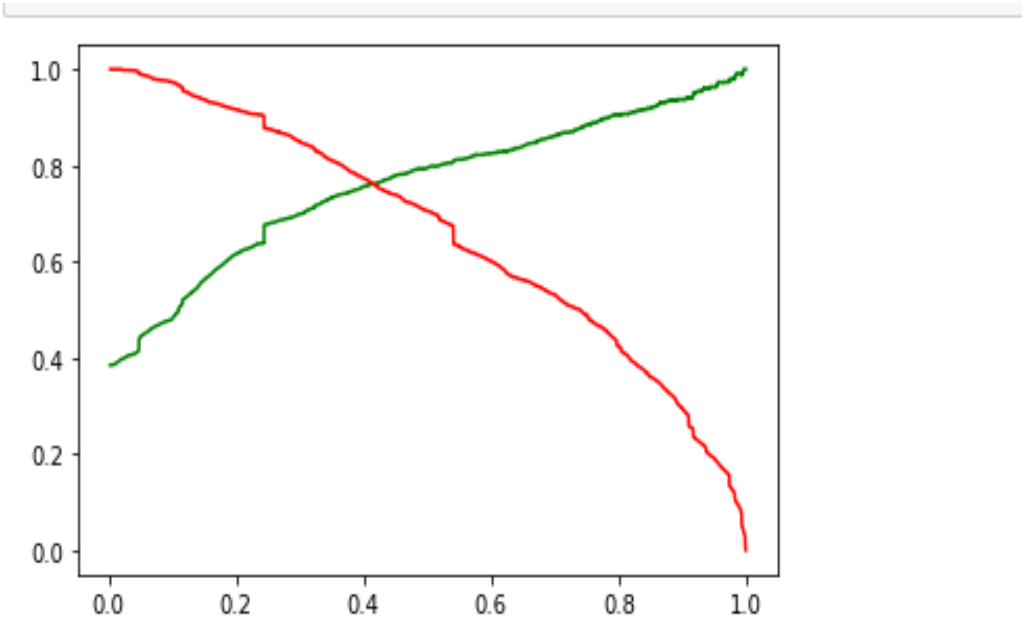
Confusion metrics :

```
array([[1396,  338],
       [ 193,  796]], dtype=int64)
```

- Accuracy - 80.4%
- Sensitivity – 80.4 %
- Specificity – 80.5%

Model Evaluation - Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.34 based on Precision and Recall



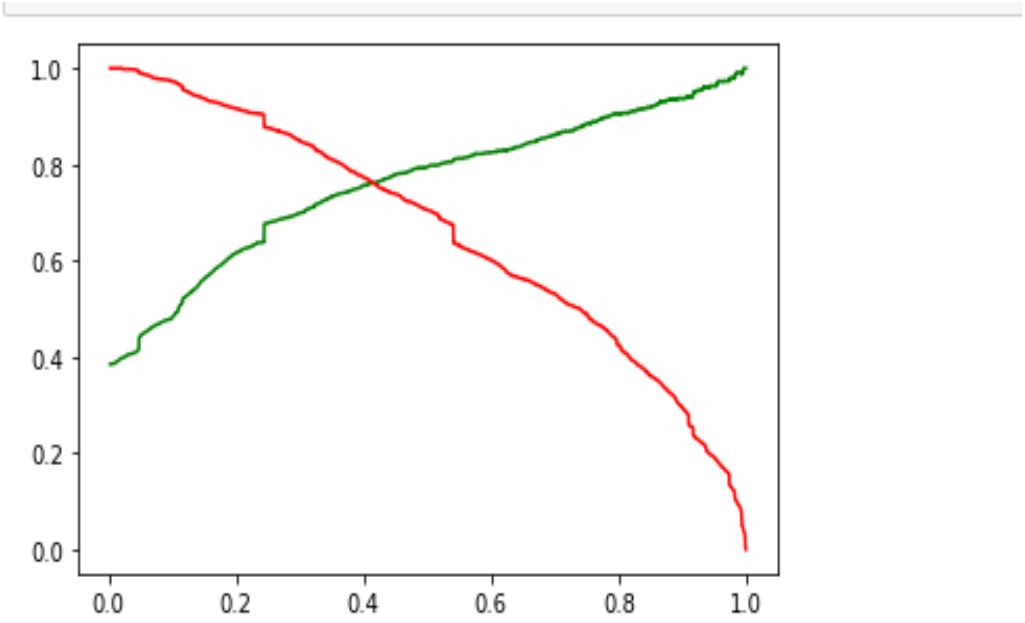
Confusion metrics :

```
array([[3151,  754],
       [ 447, 1999]], dtype=int64)
```

- Precision – 72.6 %
- Recall – 81.7%

Model Evaluation - Precision and Recall on Test Dataset

The graph depicts an optimal cut off of 0.34 based on Precision and Recall



Confusion metrics :

```
array([[1396,  338],
       [ 193,  796]], dtype=int64)
```

- Precision – 70.1 %
- Recall – 80.4%

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 81.7% and 80.6% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 70.1% (in train set) and 80.4% in test set
- It was found that the variables that mattered the most is the potential buyers are:
  1.the total time spend on website
  2.total number of visits.
  3.the working professionals as current occupations
  4.the last Activity of SMS sent, olark chat conversation