Lead Scoring Case Study

Building predictive models for filtering out leads most likely to convert from data sets.

Presented By

Vatsal Gohel & Sai Charan

Problem Statement:

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads but its lead conversion rate is very poor. For example, if they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Goal:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Methodology:

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis (EDA)
- Dummy Variables
- Splitting the data into Test and Train dataset
- Feature Scaling
- Building a logistic Regression model
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall
- Applying the best model in Test data
- Conclusions

Data Cleaning and Analysis:

- We reviewed and cleaned the data for the analysis. To gain a more refined visualization, it is important to clean the data to remove any unnecessary data sets.
- We replaced the 'select' option with the null value since it did not give us much information.
- We dropped the variables with a null value greater than 40%, and for the remaining null values we made changes in accordance with the data.
- We checked if there is any duplicate values in the data frame and also dropped the columns which are not that much important for the analysis.

```
# checking the null values in the dataframe

round(100*(leads.isnull().sum()/len(leads.index)), 2)
```

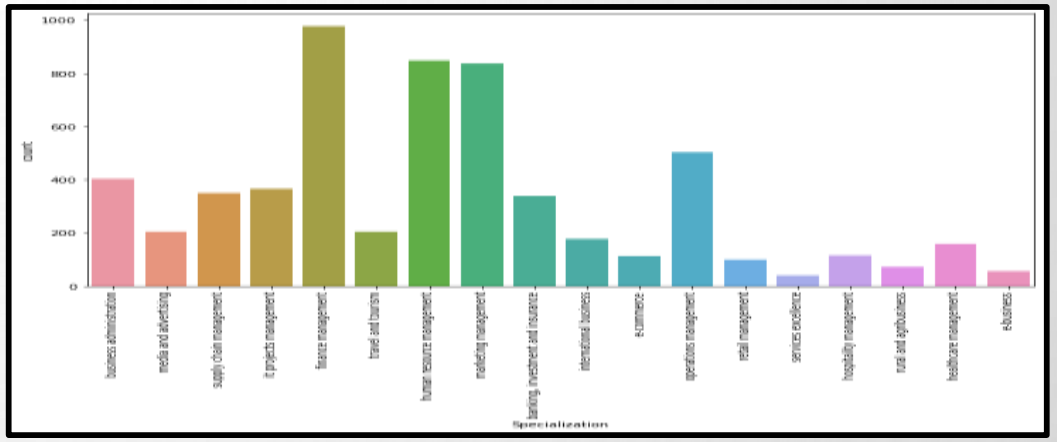
```
Prospect ID                                      0.00
Lead Number                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.39
Do Not Email                                     0.00
Do Not Call                                      0.00
Converted                                        0.00
TotalVisits                                      1.48
Total Time Spent on Website                      0.00
Page Views Per Visit                             1.48
Last Activity                                    1.11
Country                                         26.63
Specialization                                  15.56
How did you hear about X Education              23.89
What is your current occupation                 29.11
What matters most to you in choosing a course   29.32
Search                                           0.00
Magazine                                         0.00
Newspaper Article                                0.00
X Education Forums                               0.00
Newspaper                                        0.00
Digital Advertisement                            0.00
Through Recommendations                          0.00
Receive More Updates About Our Courses           0.00
Tags                                            36.29
Lead Quality                                    51.59
Update me on Supply Chain Content                0.00
Get updates on DM Content                        0.00
Lead Profile                                    29.32
City                                            15.37
Asymmetrique Activity Index                     45.65
Asymmetrique Profile Index                      45.65
Asymmetrique Activity Score                     45.65
Asymmetrique Profile Score                      45.65
I agree to pay the amount through cheque         0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

Exploratory Data Analysis (EDA):

❖ Categorical Analysis:

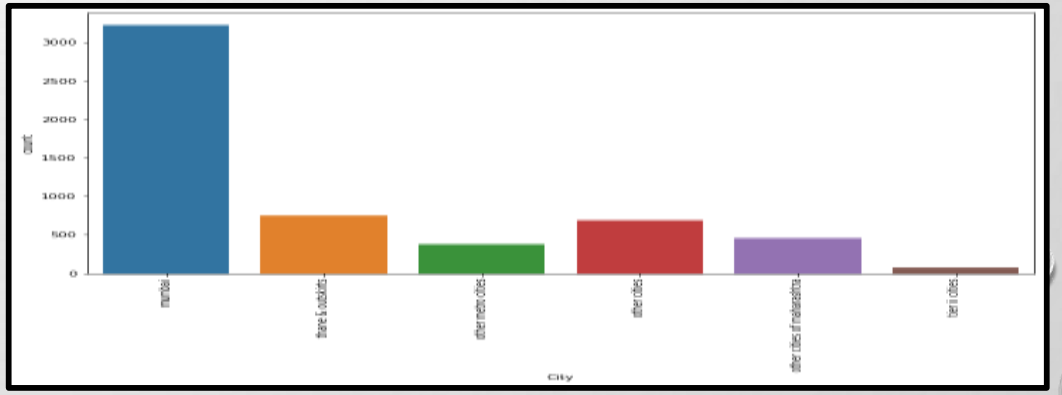
➢ ‘Specialization’ column

- We can see that the whole **management field** has converted the most. Whereas, **‘services excellence’**, **‘e-business’**, **‘agribusiness and rural’** has the lowest conversion rate.



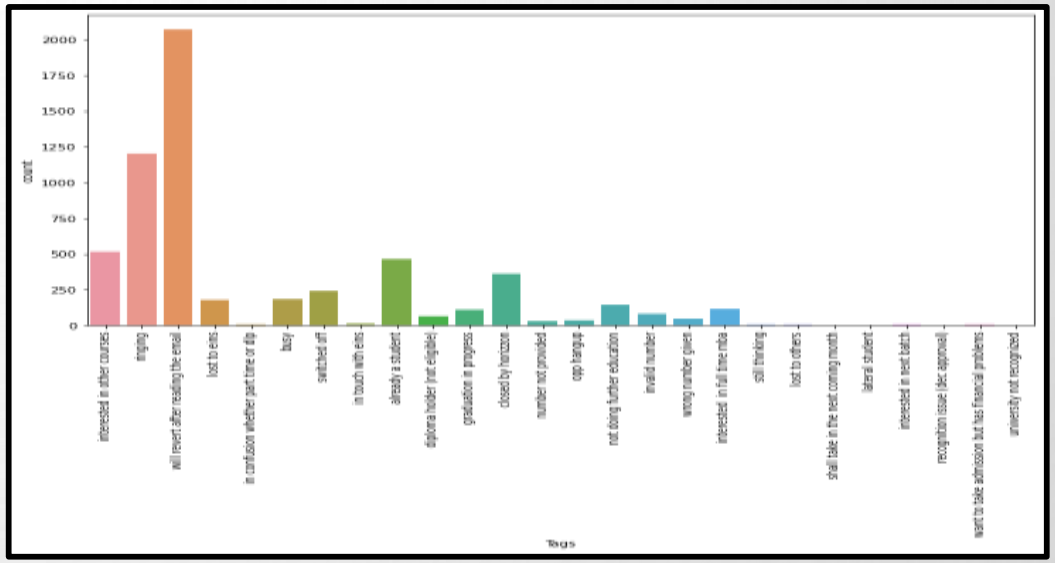
➢ ‘City’ column

- **‘Mumbai’** is the highest in terms of conversion rate, whereas, **‘Tier 2 cities’** have the minimum conversion rate followed by **‘other metro cities’**.



Exploratory Data Analysis (EDA):

- **‘Tags’ column**
  - Here we can observe that **‘Will revert after reading the mail’** has converted the most, followed by **‘ringing’**.
  - This column contains more null values so after analyzing we dropped the column.

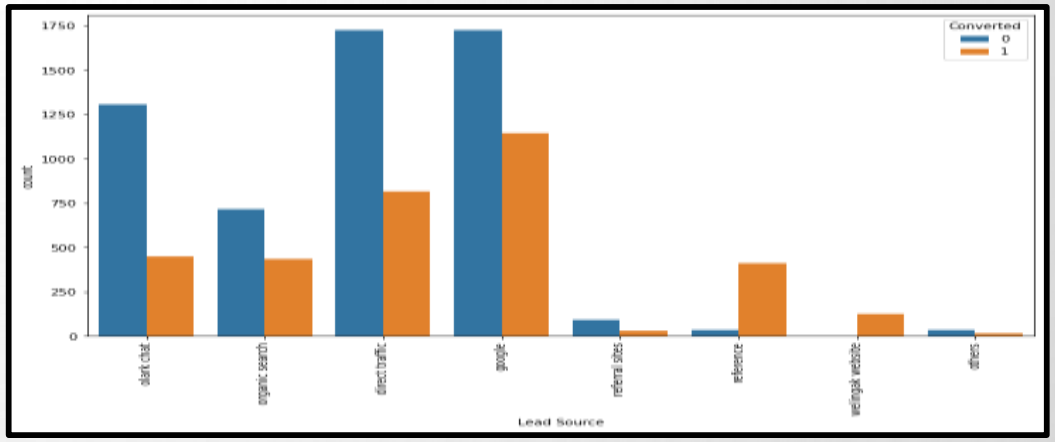


Exploratory Data Analysis (EDA):

❖ Univariate Analysis:

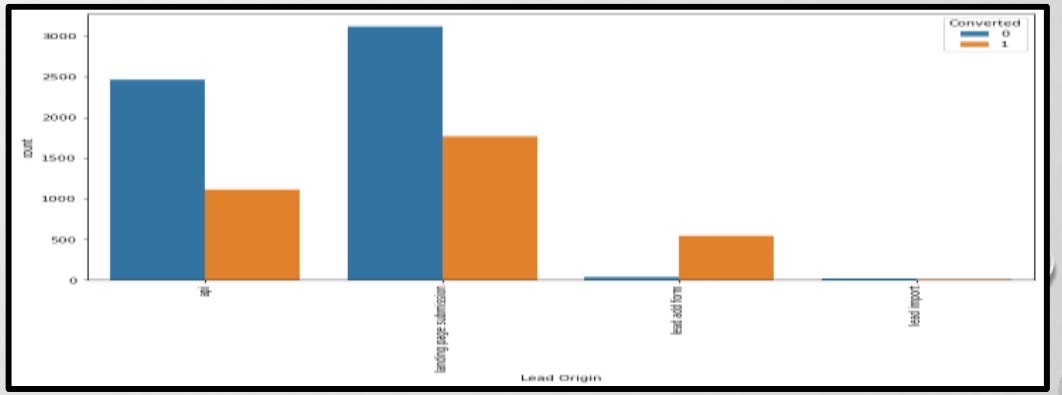
➢ ‘Lead Source’ column

- we can say that **'direct traffic'** and **'Google'** convert the maximum number of leads.
- The rate of conversion is high for the **'welingak website'** and **'reference'**.



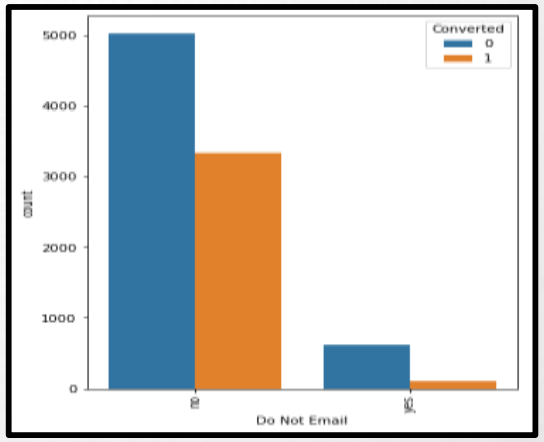
➢ ‘Lead Origin’ column

- we can see that **'lead import'** count is minimum.
- The most conversions came from the **'landing page submission'**, followed by the **'api'**.

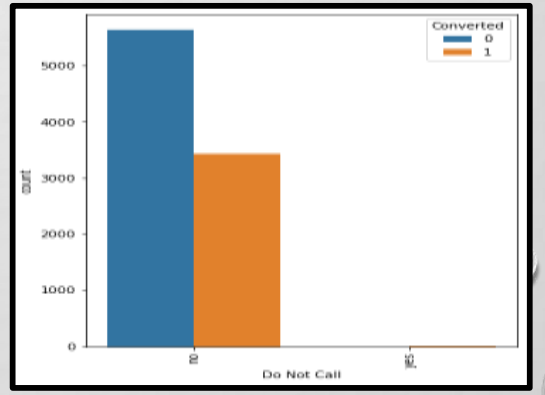


Exploratory Data Analysis (EDA):

- **‘Do Not Email’ column**
  - The overall count of **'yes'** is very low and among them majority of them are not converted.
  - The leads who choose **'no'** is very high compared to the overall count of **'yes'**.

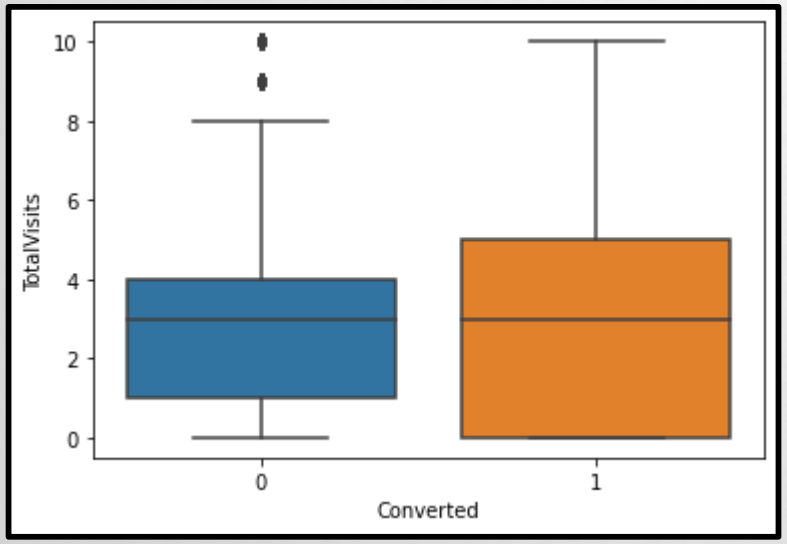


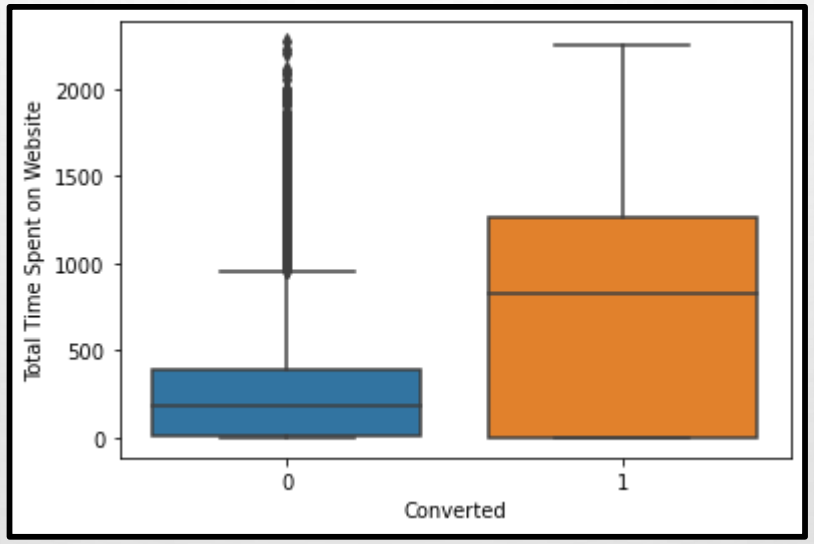
- **‘Do Not Call’ column**
  - There are no leads who has selected **'yes'** as a option.
  - Votes are significantly high for **‘no'**, with most of the counts not being converted.

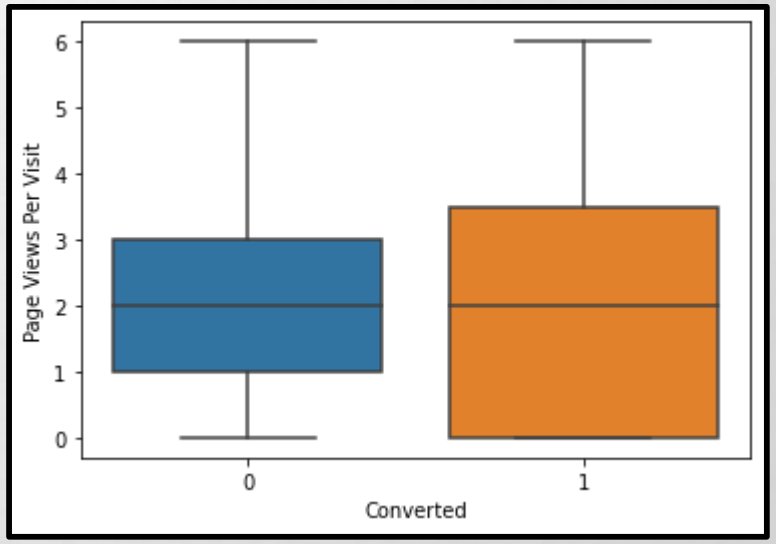


Exploratory Data Analysis (EDA):

- **‘Total Visits’, ‘Total Time Spent on Website’ and ‘Page Views Per Visit’ column**
  - The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit.

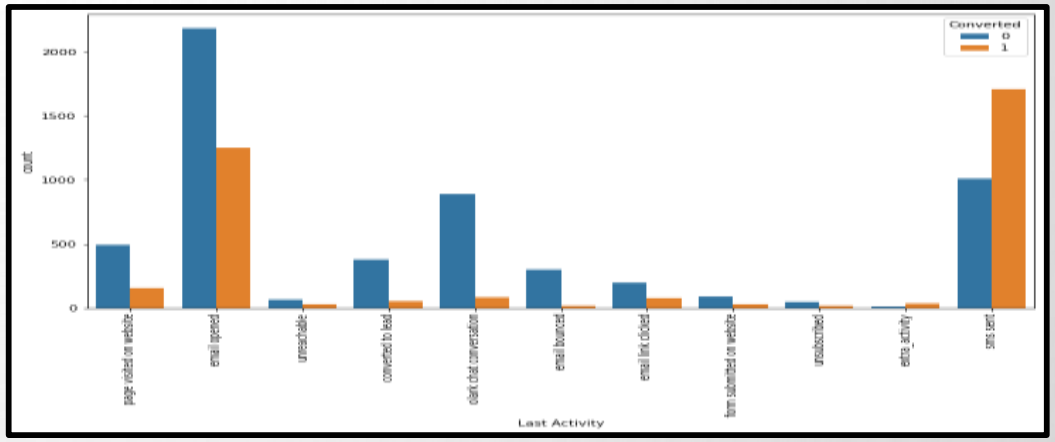




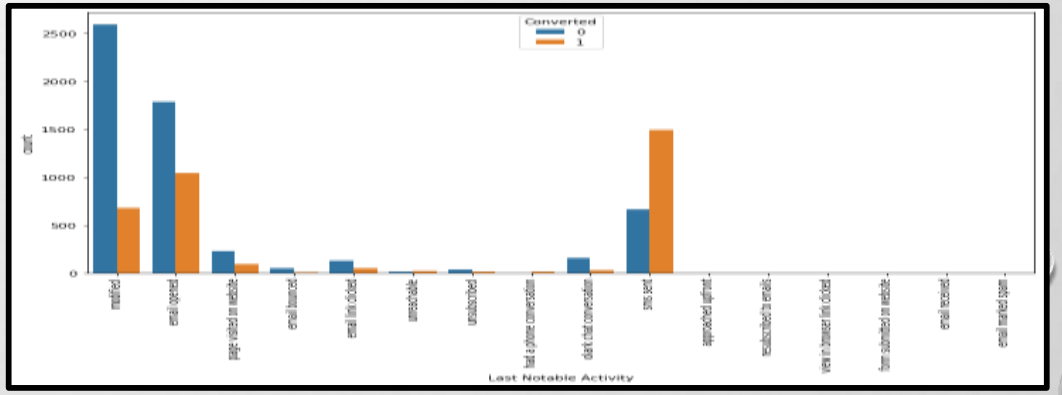


Exploratory Data Analysis (EDA):

- **'Last Activity' column**
  - we can say that Leads are most likely to have opened their email as their last activity.
  - Highest conversion rate is seen in **'sms sent'**.



- **'Last Notable Activity' column**
  - The number of non-converted leads from **'sms sent'** is considerably lower as compared to **'e-mail Opened'** and **'modified'**.
  - Most converted leads are from the **'sms sent'** followed by **'e-mail opened'** and then **'modified'**.



Dummy Variables:

- First we will create a dummy variable for the categorical variables.
- Later, after creating a dummies we will merge those variables with the original data frame.
- Dropping the unnecessary columns as we have already created dummy variables out of it.

Splitting the data into Test and Train dataset:

- We will split the data set into train and test data with a ratio of 70-30%.
- After splitting, we will build a model on train data set and later test it on the test data set.

Feature Scaling:

- We used the StandardScaler to scale the original numerical variables.
- Using the stats model, we created our initial model, which gave us a complete statistical view of each parameter.

Building a logistic Regression model:

- The next step is to create X and y train dataset for building a model.
- We will use the Logistic Regression function from the Scikit learn for making it suitable with Recursive Feature Elimination (RFE).
- The 15 top features were selected based on the RFE.
- Based on the VIF values and p-values (VIF < 5 and p-value < 0.05 were kept), the variables that were not significant were manually removed.

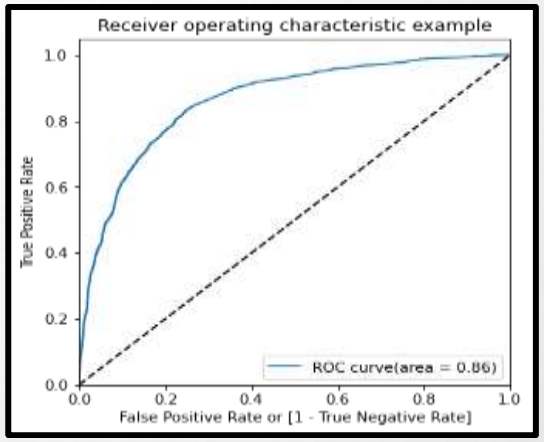
Evaluating the model by using different metrics:

- Using Final Model, we will create the column **'Predicated'** with probability cutoff of 0.5.
- We derived the Confusion Metrics and calculated the model's overall Accuracy based on the above assumptions.
- We could observe the values of the **Accuracy = 79.3%**, **Sensitivity = 67.1%**, **Specificity = 87.0%**.

Evaluating the model by using different metrics:

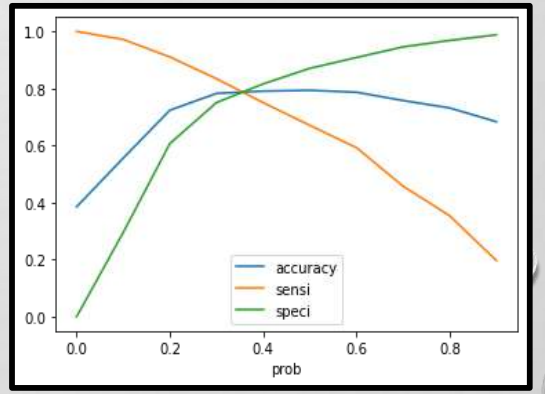
❖ Roc Curve:

- We plotted the ROC curve for the features, which showed a good area coverage of 86%, further solidifying our model.



❖ Optimal Cut-off Point:

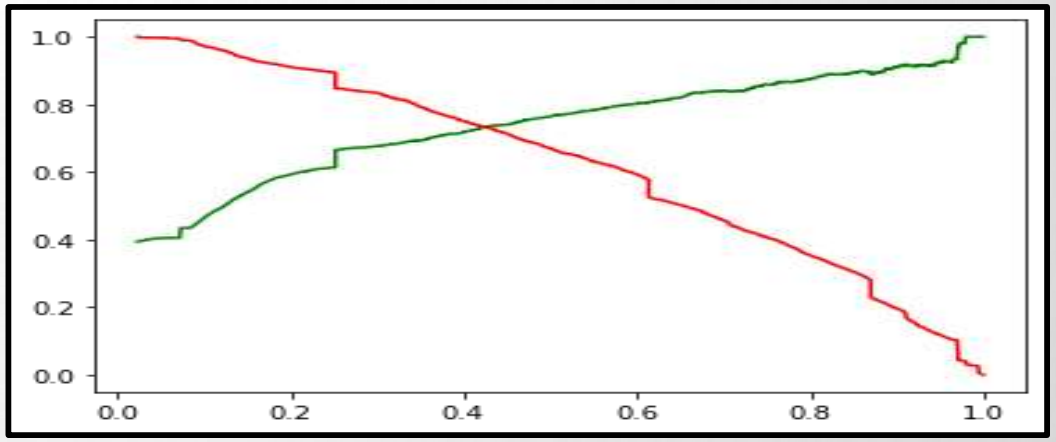
- We plotted the probability graph for **Accuracy**, **Sensitivity**, and **Specificity** for different probabilities and intersection was considered as **optimal probability cut-off point**.
- Based on a cut-off value of 0.37, we use 0.4 as the model estimate.
- We could also observe the values of the **Accuracy = 79.1%**, **Sensitivity = 75.0%**, **Specificity = 81.6%**.



Evaluating the model by using different metrics:

❖ Precision and Recall score:

- The graph depicts an optimal cut off of 0.42 based on Precision and Recall.
- We found **Precision = 71.9%** and **Recall = 75.0%** on the train data frame.
- ➢ **When plotting the ROC curve and the graph based on Precision and Recall metrics, we get almost the same intersection value.**



Applying the best model in Test data:

➢ We then applied these learnings to test model and found an **Accuracy = 79.1%**, **Sensitivity = 74.3%** and **Specificity = 81.8%**.

Conclusions:

- we have considered the optimal cut-off of 0.4 for calculating the final prediction after checking both Sensitivity-Specificity as well as Precision and Recall metrics.
- Accuracy, Sensitivity and Specificity values of test set are around 79.1%, 74.3% and 81.8% which are approximately closer to the respective values calculated using trained set.
- The top 3 variables that contribute for lead getting converted in the model are:
  - Lead Source_olark chat
  - Last Activity_olark chat conversation
  - Total Time Spent on Website
- This model seems to be very accurate in predicting conversion rates and seems stable when running on a **TRAIN SET** as well as a **TEST SET**.
- With this model, we should be able to give the CEO confidence in making good decisions.

THANK YOU!!!