X Education - Lead Scoring Case Study

Identification of Hot Leads to focus more on them and thus enhancing the conversion ratio for X Education

Group Members:

Vandita Tyagi

Anusha Malempati Kamma

Background

X Education Company

1. X Education , An education company named sells online courses to industry professionals
2. Many interested professionals land on their website
3. The company markets its courses on several websites 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

Background

X Education Company

4. When these people fill up a form providing their email address or phone number, they are classified to be a lead
5. 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
6. The typical lead conversion rate at X education is around 30%

Problem Statement

X Education Company's Problem

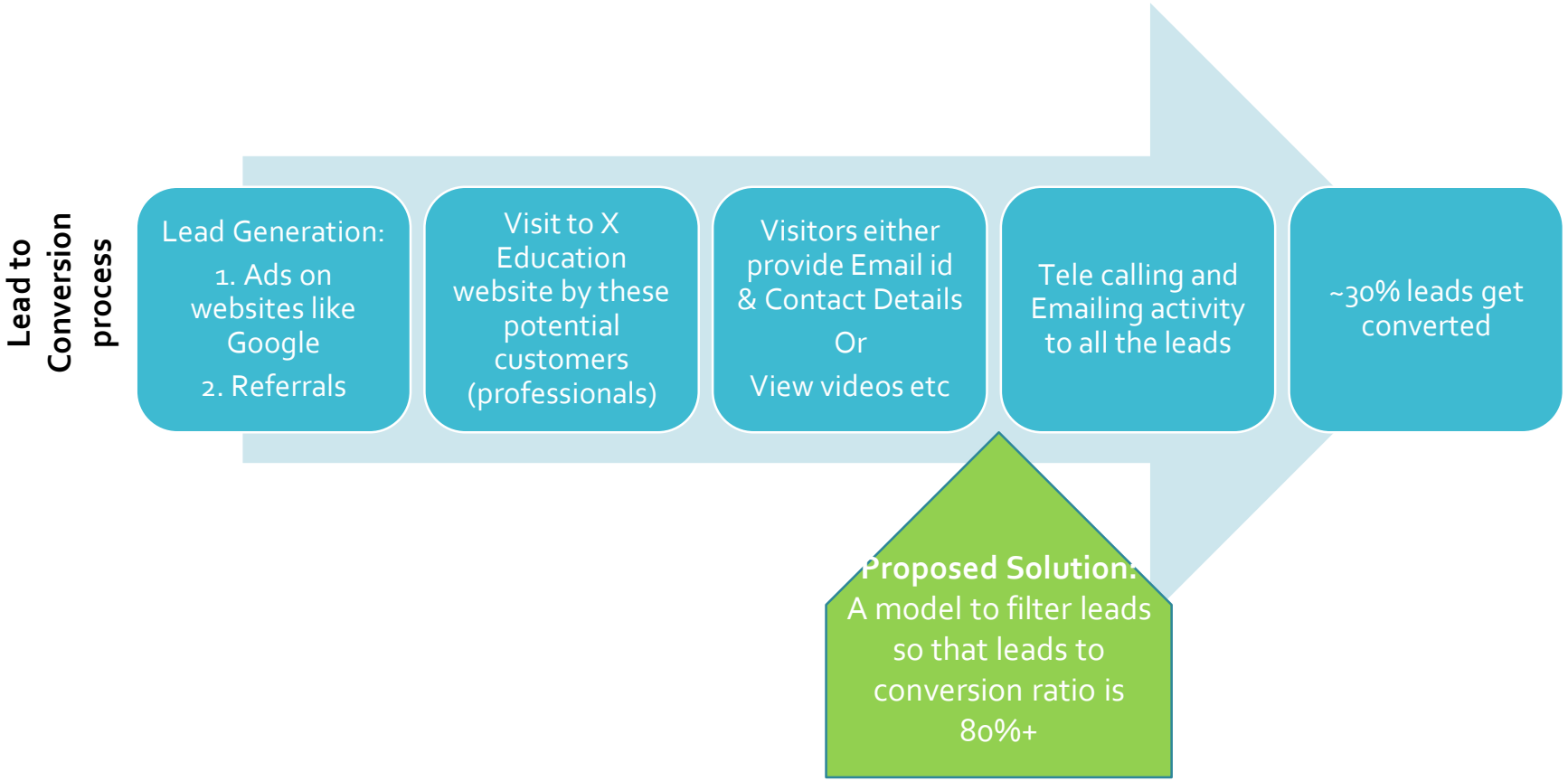
1. X Education gets a lot of leads but its lead conversion rate is very poor
2. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'
3. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

Problem Statement

X Education Company's Problem

4. We will help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
5. We are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance
6. The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

Lead – Conversion Process



Solution

Selection of Hot Leads

For our Problem Solution, the crucial part is to accurately identify hot leads.

The more accurate we obtain the hot lead, the more chance we get of higher conversion ratio.

Since we have a target of 80% conversion rate, we would want to obtain a high accuracy in obtaining hot leads.

Data Cleaning and EDA:

- Percentage of missing values was checked and the columns with more than 40% missing values were dropped.
- For columns having missing values less than 40% were replaced with others or the most common value.Eg, in the country column, since India is the most common occurrence among the non-missing values, we imputed all not provided values with India.

Data Cleaning and EDA:

1. Handled the 'Select' values - There are 'Select' values in many columns. It may be because the customer did not select any option from the list, hence it shows 'Select'. So we converted these values to null values.
2. Found the null percentages across columns
3. For some columns there are high percentage of missing values. We dropped the columns with missing values greater than 40% .
4. Found the null percentages again across columns after removing the above columns
5. Checked the columns with one unique value since it won't affect the analysis and removed

Data Cleaning and EDA:

6. Checked the values of Specialization - This column has 37% missing values
7. Checked the values of tags - 'Tags' column has 36% missing values
8. Checked the values of 'What matters most to you in choosing a course' - this column has 29% missing values

   Since 100% of values in this column is Better career prospects, so we removed this column.
9. Checked the values of 'What is your current occupation' - This column has 29% missing values

   Imputed the missing data in the 'What is your current occupation' column with 'Unemployed'
10. Column: 'Country' - This column has 27% missing values

    Imputed the missing data in the 'Country' column with 'India'

Data Cleaning and EDA:

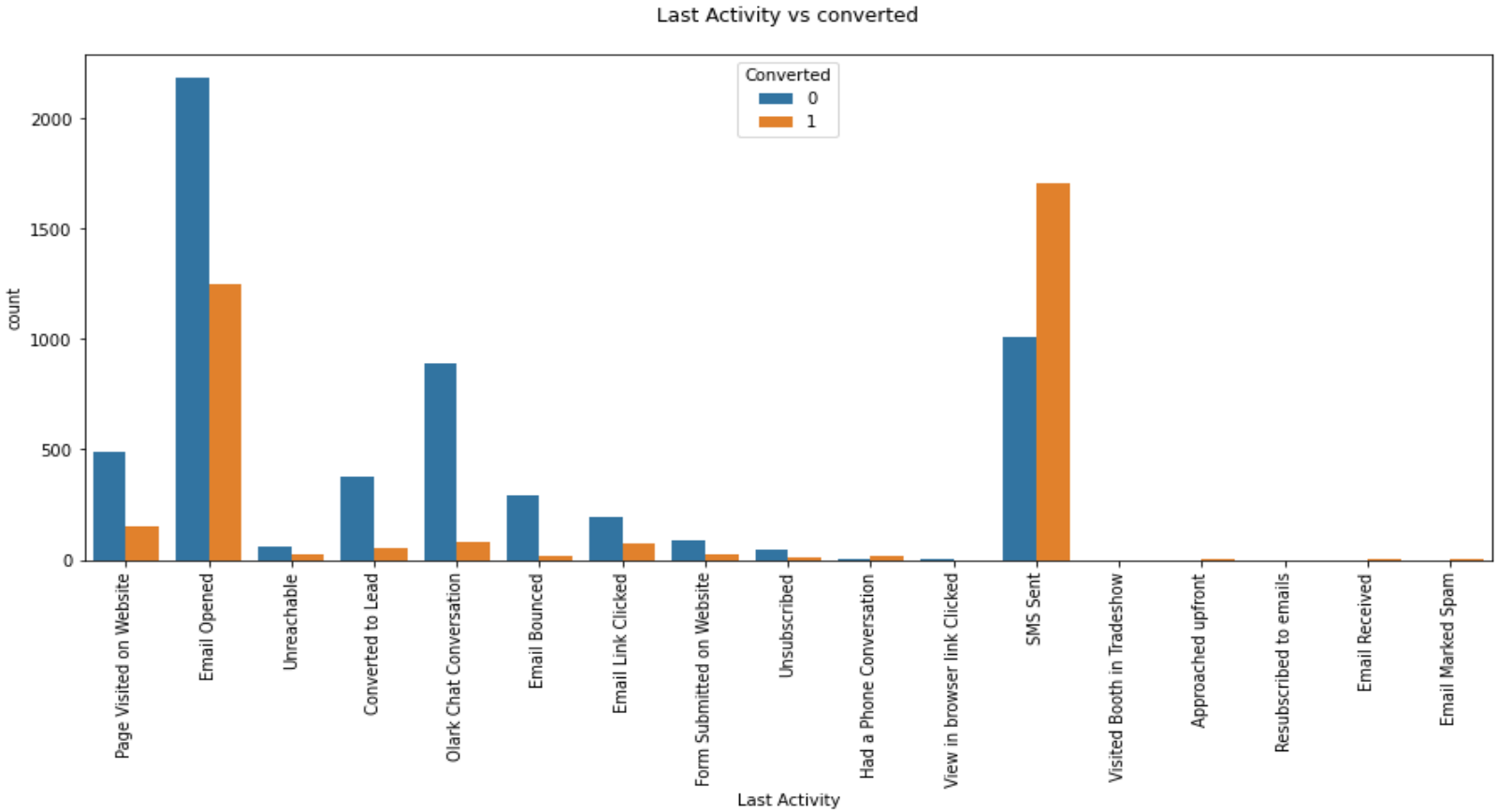
11. Checked the values of 'City' - This column has 40% missing values

    Since most values are 'Mumbai' , we can impute missing values in this column with 'mumbai'.

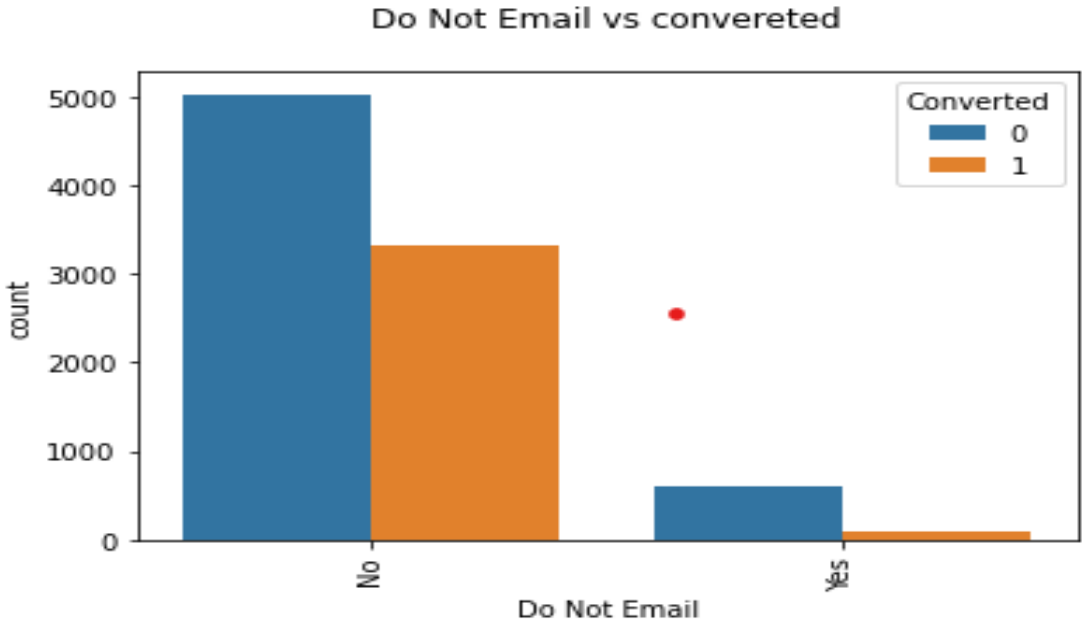
12. Found the null percentages across columns after removing the above columns
13. Rest missing values are under 2% so we Dropped these rows.
14. We have retained 98% of the rows after cleaning the data .

Univariate and bivariate analysis

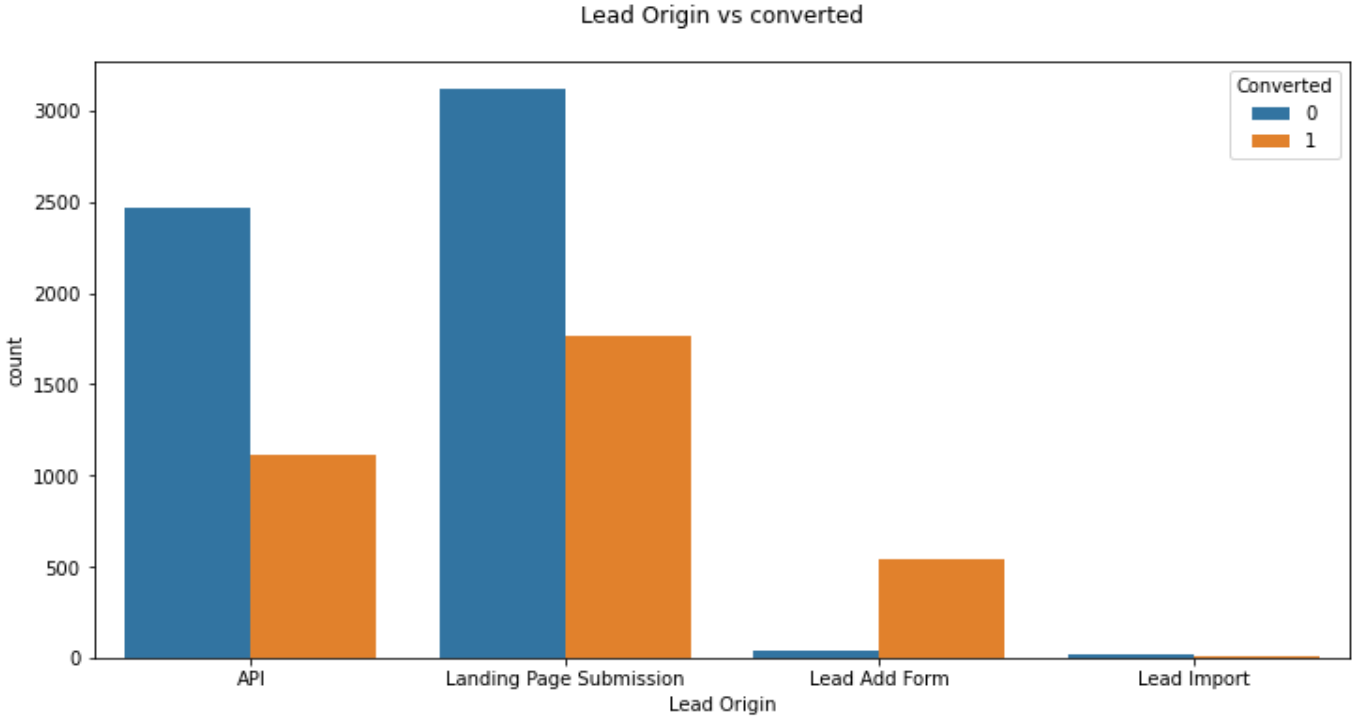
Based on the univariate analysis we have seen that many columns like 'Lead Number','Tags','Country','Search','Magazine',' Newspaper Article','X Education Forums' etc., are not adding any information to the model, hence we dropped them for further analysis



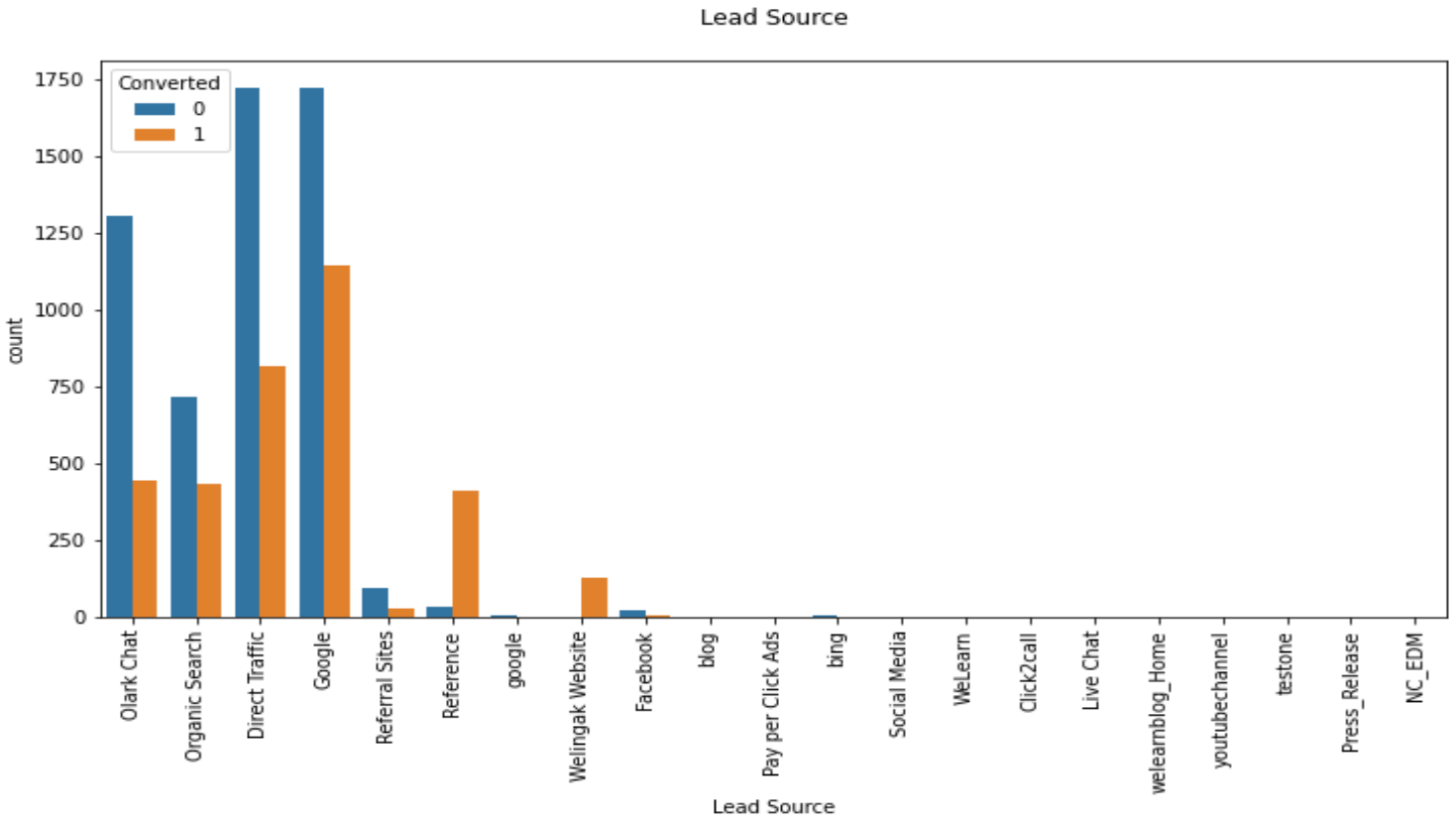
EDA plots depicting variation in categorical column (Last Activity) for those who Converted and those who didn't.



EDA plots depicting variation in categorical column (Do Not Email) for those who Converted and those who didn't.



EDA plots depicting variation in categorical column (Lead Origin) for those who Converted and those who didn't.



EDA plots depicting variation in categorical column (Lead Source) for those who Converted and those who didn't.

Data Preparation

1. Converted some binary variables (Yes/No) to 1/0
2. Created Dummy variables for the categorical features:

   'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization','What is your current occupation','City','Last Notable Activity’

   Dropped the columns for which dummies were created

Train-Test split and Scaling

1. The split was done at 70% and 30% for train and test data respectively.
2. Standard scaler was used on the variables (TotalVisits,Page Views Per Visit, Total Time Spent on Website)

Model Building

1. RFE was used for feature selection.
2. Then RFE was done to attain the top 15 relevant variables.
3. Later the rest of the variables were removed manually depending on the VIF values and p-value.
4. A confusion matrix was created, and overall accuracy was checked which came out to be 81.42%.

Model Evaluation

Sensitivity – Specificity - If we go with Sensitivity- Specificity Evaluation. We will get :

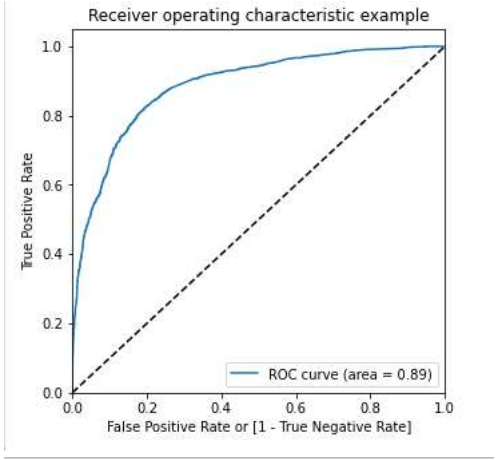
On Training Data

1. The optimum cut off value was found using ROC curve. The area under ROC curve was 0.89.
2. After Plotting we found that optimum cutoff was 0.35 which gave

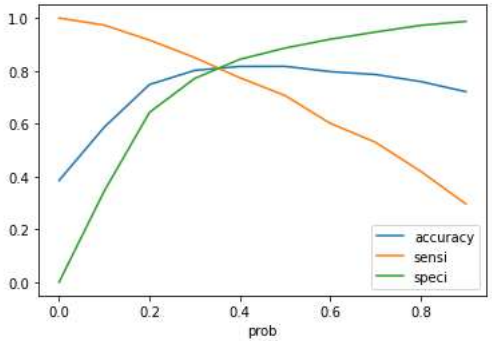
   Accuracy : 81.42%
   Sensitivity : 81.16 %
   Specificity : 81.59 %

Prediction on Test Data

Accuracy : 80.87 %
Sensitivity : 80.08 %
Specificity : 81.31 %



Since we have higher (0.89) area under the ROC curve.



From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

Precision – Recall:

On Training Data

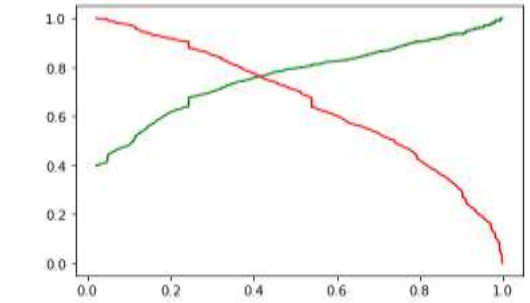
- With the cutoff of 0.35 we get the Precision and Recall of 79.57% and 70.68% respectively.
- So to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of 0.41 which gave

  Accuracy 81.72%
  Precision 76.06%
  Recall 76.66%

Prediction on Test Data

  Accuracy 81.97%
  Precision 74.17%
  Recall 77.25%

So if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be 0.35 and If we go with Precision – Recall Evaluation the optimal cut off value would be 0.41



The above graph shows the trade-off between the Precision and Recall .

Conclusion

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.

Recommendations:

- The company should make calls to the leads coming from the Lead Origin_Lead Add Form, What is your current occupation_Working Professional and Lead Source_Welingak Website, those whose last activity is sms sent, who spent “more time on the websites” as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was “Olark Chat Conversation”, leads whose lead origin is “Landing Page Submission”, leads whose Specialization was “Others”, leads who chose the option of “Do not Email” as “yes” as they are not likely to get converted.