Lead Scoring Assignment

By:

Abhishek Vamshi

Sandeep Khasnavis

Tanmay Dandekar

Problem Statement and Background

Company Name: X Education
Profiles: Sells educational courses
Target Audience: Industry Professionals.
Highlights: The firm promotes its courses on various websites and search engines such as Google.

- When these users arrive at the website, they may explore the courses, fill out a course registration form, or view some videos.
- If someone fills out a form with their email address or phone number, they are labelled as a lead. Furthermore, the corporation receives referrals from previous clients provide leads.
- Once these leads are obtained, members of the sales team begin making calls, composing emails, and so on. As a result of this procedure, some of the leads are converted but most are not. At X schooling, the average lead conversion rate is roughly 30%.

Business Goal

- X Education need assistance in identifying the most promising prospects, i.e., those most likely to convert into paying clients.
- The firm requires a model in which a lead score is assigned to each lead, with higher lead scores having a higher conversion chance and lower lead scores having a lower conversion chance.
- The CEO has stated that the objective lead conversion rate should be about 80%.

PROCESS

- Understanding of data
- 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.

FLOW

Understanding, Cleaning and Preparation

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

Feature Scaling and Splitting Train and Test Sets

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

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

Understanding, Cleaning and Preparation

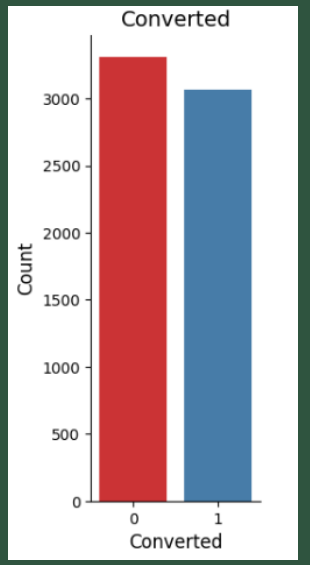
- As you can see there are a lot of column which have high number of missing values.
- Clearly, these columns are not useful. Since, there are 9000 datapoints in our data frame, let's eliminate the columns having greater than 3000 missing values as they are of no use to us.

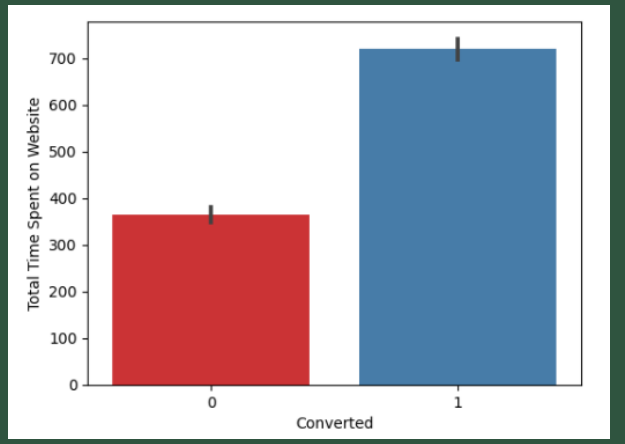
    - Thus, we dropped all the columns in which greater than 3000 missing values are present
    - As the variable City and Country won't be of any use in our analysis. So, it's best that we drop it.

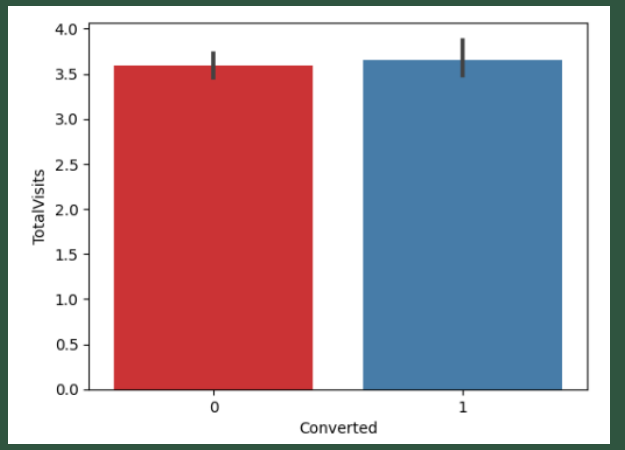
Now recall that there are a few columns in which there is a level called 'Select' which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are as good as missing values and hence we need to identify the value counts of the level 'Select' in all the columns that it is present.

- Clearly the levels Lead Profile and How did you hear about X Education have a lot of rows which have the value Select which is of no use to the analysis so it's best that we drop them.
- Similarly, dropping “What is your current occupation”
- Dropping null values rows from “Specialization‘, Total Visits’ and ‘Lead Source’

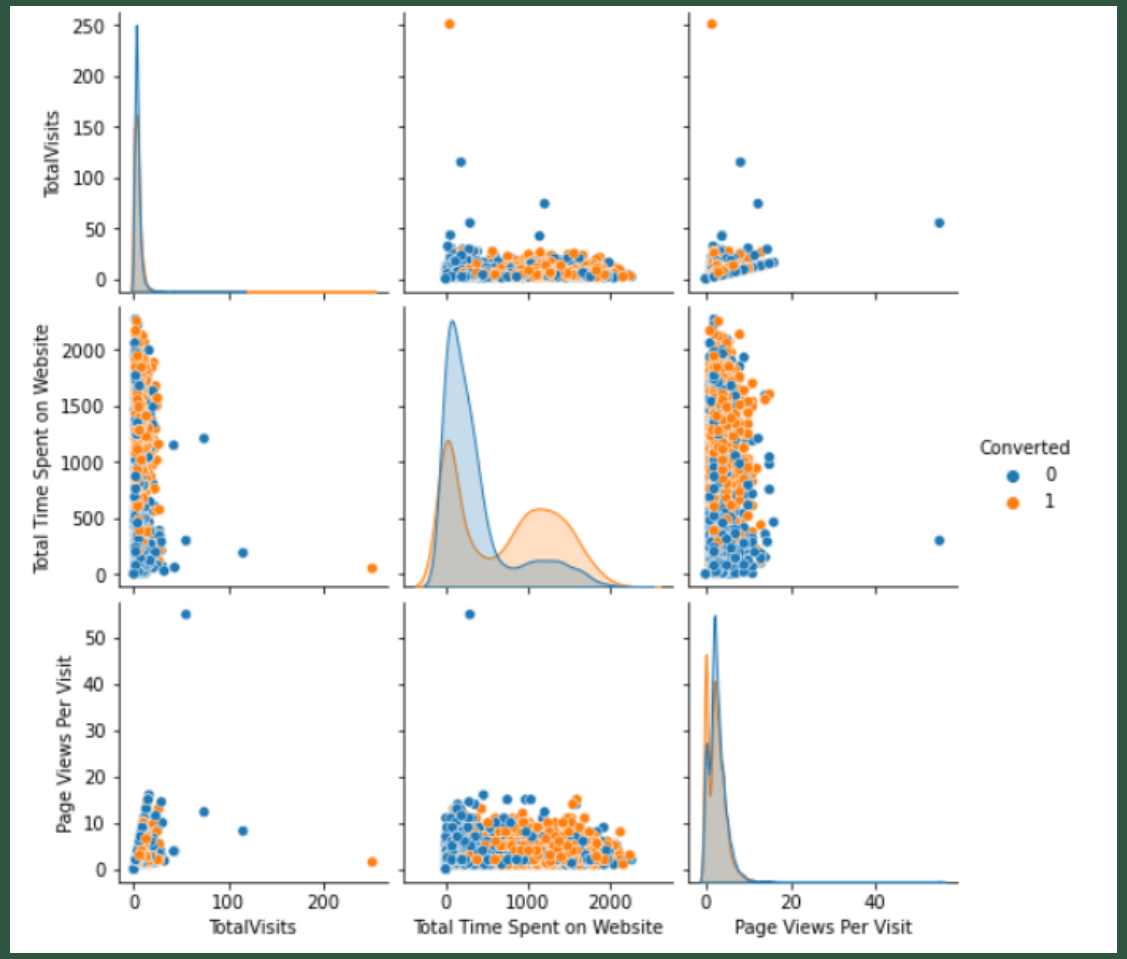
Finally, now our data doesn't have any null values. We have about 69% of data retained.

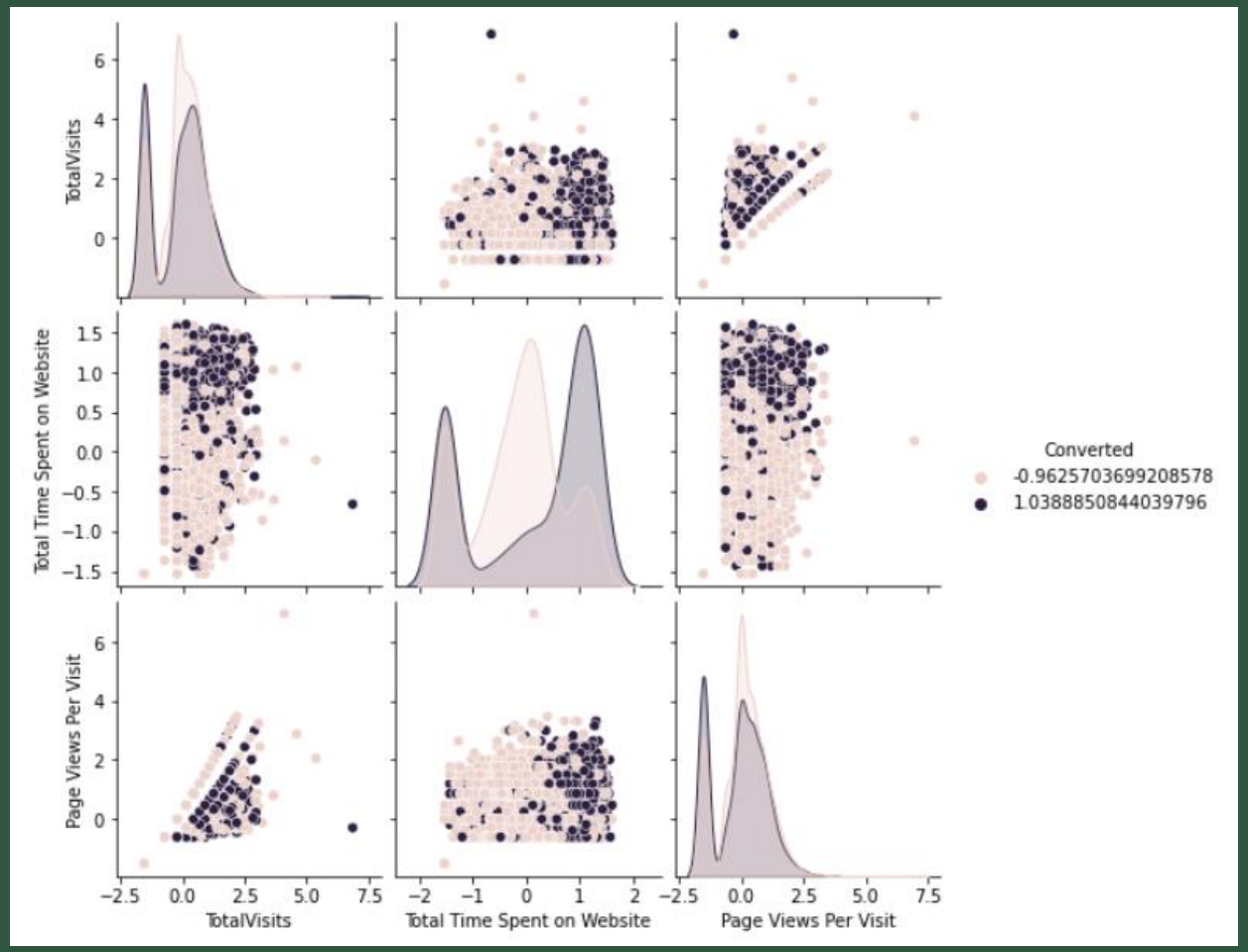






PREPARING DATA FOR MODELLING





Feature Scaling and Splitting Train and Test Sets

Dummy variable creation

The next step is to deal with the categorical variables present in the dataset. So first look at which variables are categorical variables.

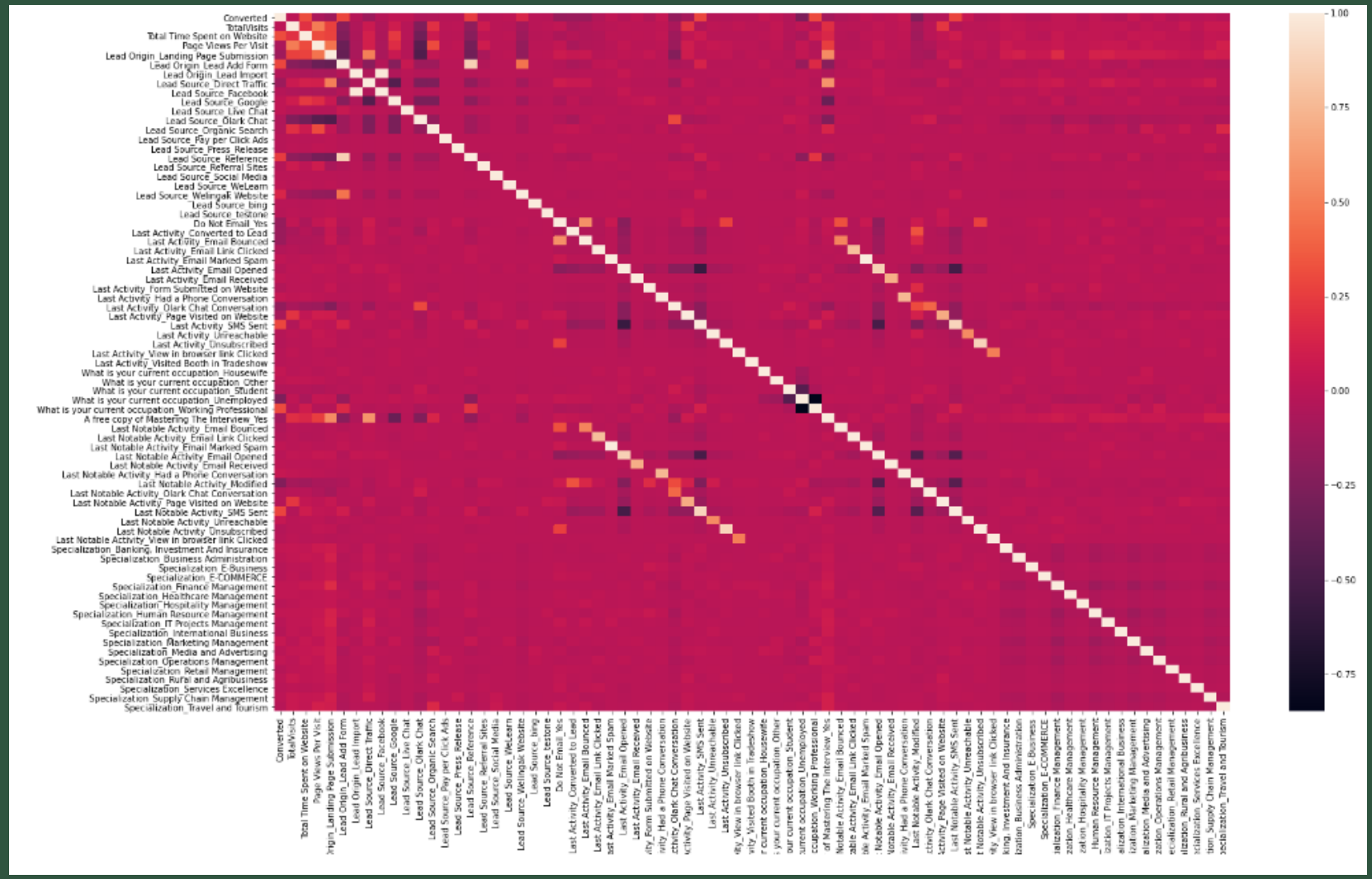
Test-Train Split

The next step is to split the dataset into training and testing sets

Scaling

Now there are a few numeric variables present in the dataset which have different scales. So, let's go ahead and scale these variables.

Let's now look at the correlations. Since the number of variables are pretty high, it's better that we look at the table instead of plotting a heatmap



Model Building

RFE for feature selection

Let's now move to model building. As you can see that there are a lot of variables present in the dataset which we cannot deal with. So, the best way to approach this is to select a small set of features from this pool of variables using RFE.

Build model

- Using all the variables selected by RFE, lets create a logistic regression model using stats models. Also, lets look at the p-values and the VIFs.

- There are quite a few variable which have a p-value greater than 0.05. We will need to take care of them.

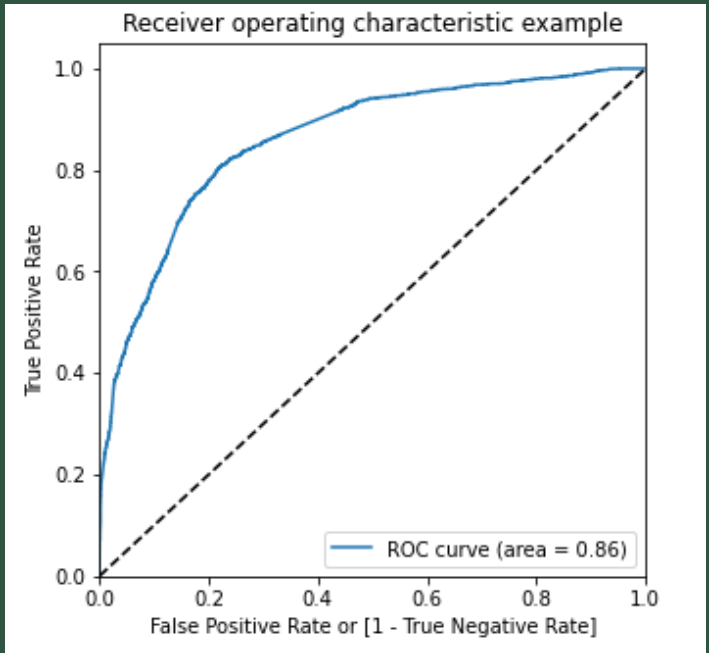
	Features
2	Lead Origin_Lead Add Form
4	Lead Source_Reference
5	Lead Source_Welingak Website
11	What is your current occupation_Unemployed
7	Last Activity_Had a Phone Conversation
13	Last Notable Activity_Had a Phone Conversation
1	Total Time Spent on Website
0	TotalVisits
8	Last Activity_SMS Sent
12	What is your current occupation_Working Profes...
3	Lead Source_Olark Chat
6	Do Not Email_Yes
10	What is your current occupation_Student
9	What is your current occupation_Housewife
14	Last Notable Activity_Unreachable

After dropping few features, the final model has both p-values and VIFs decent enough for all the variables. Here is the table of final model.

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Result: Model Evaluation

The area under the curve of the ROC is 0.86 which is quite good. So, it seems we have a good model

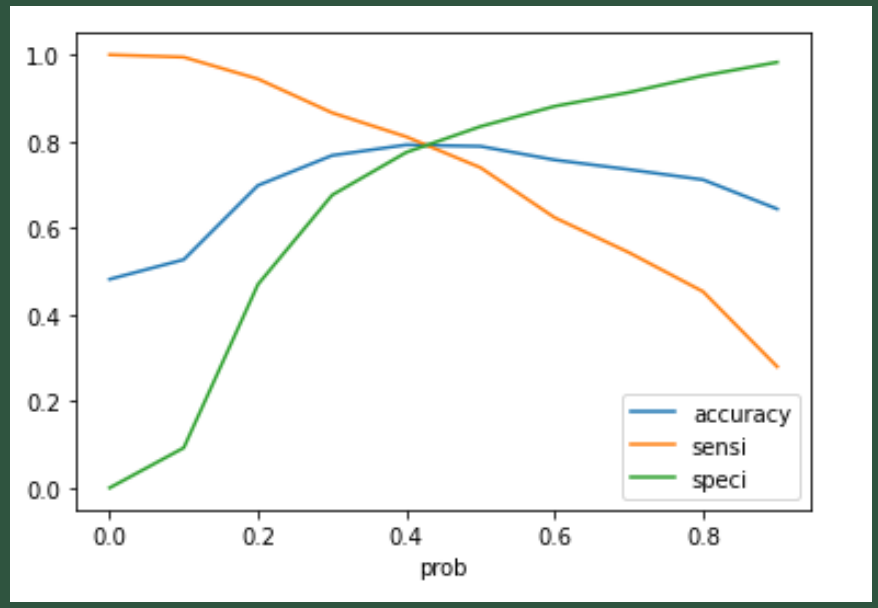


Confusion Matrix

```
[[1823,  489],
 [ 444, 1705]],
```

Accuracy – 79%
Sensitivity – 79%
Specificity – 78%

The graph depicts an optimal cut off of 0.42 based on Accuracy, Sensitivity and Specificity on trained data set,



Result: Model Evaluation

Confusion Matrix: Test Data Set

```
[[801, 195],
 [213, 703]]
```

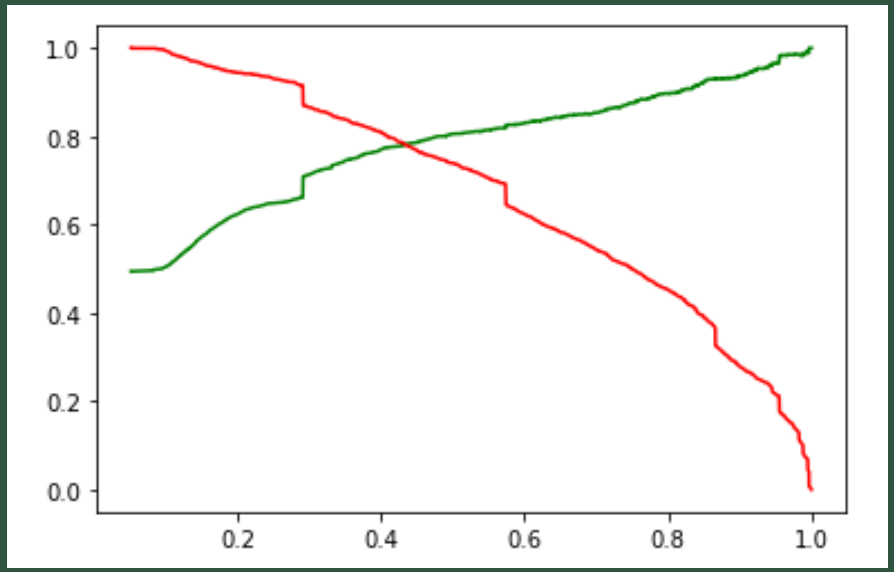
Accuracy – 78%
Precision – 78%
Recall – 76%

Confusion Matrix

```
[[1852,  460],
 [ 479, 1670]]
```

Accuracy – 78%
Precision – 78%
Recall – 77%

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



Conclusion

- Considering both Sensitivity-Specificity as well as Precision and Recall Metrics, we have taken optimal cut off
- Accuracy, Sensitivity and Specificity values of test set are around 79%, 79% and 78% which are approximately closer to the respective values calculated using trained set.
- Apart from that the lead score calculated shows the conversion rate on the final predicted model is around 79% (in train set) and 78% in test set

The top 5 variables that contribute for lead getting converted in the model are

- Total Visits
- Total Time Spent on Website
- Lead Add Form (from Lead Origin)
- Unreachable (from Last Notable Activity)
- Had a Phone Conversation (from Last Activity)

Hence overall this model seems to be good.

Business Recommendations

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom.

- Company should nurture the potential leads well (i.e., educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- The best prospects from the leads should be sorted. 'Total Visits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.
- This list can also be useful for more research and development so that company can churn out new courses, services, job offers and future higher studies. Monitor each lead carefully so that the information can be tailored before sending customers.
- Carefully provide job offerings, information or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects.
- Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them. Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.