Case Study Assignment

Lead conversion rate in X eduation

Aravind Babu Chandrika

18/12/23

Problem Statement

From the given data set containing lead data, identify the driving factors that are indicators of a lead getting converted to an actual paying customer.

Analysis Approach

After data preparation, selected features based on RFE before building a prediction model on these variables to identify the probability of lead conversion.

Data Preparation

- After loading the data from csv files into data frame the first step was to identify the unnecessary columns in the data which are not required for analysis prima facie.
- Next step was to identify the columns with null values and based on null percentage calculation of the data set, decision was made to remove columns with more than 30% of null values as they would not aid in providing an accurate analysis.
- The rows with minimal null values were also removed based on the percentage of null in a particular column.

Data Preparation

- Few columns had “select” as a value which was converted to null values for data standardization.
- For categorical variables with multiple levels dummy variables were created.
- Binary variables with Yes/No values were converted to 0/1.
- Outliers in continuous variables were identified using percentile calculation and rows were omitted based on the results.

Train-Test Split and Feature scaling

- Next the data set was split in to train and test data set.
- Fit transform was run on the continous variables before calculating lead conversion rate of the data set which came out to be 48%.

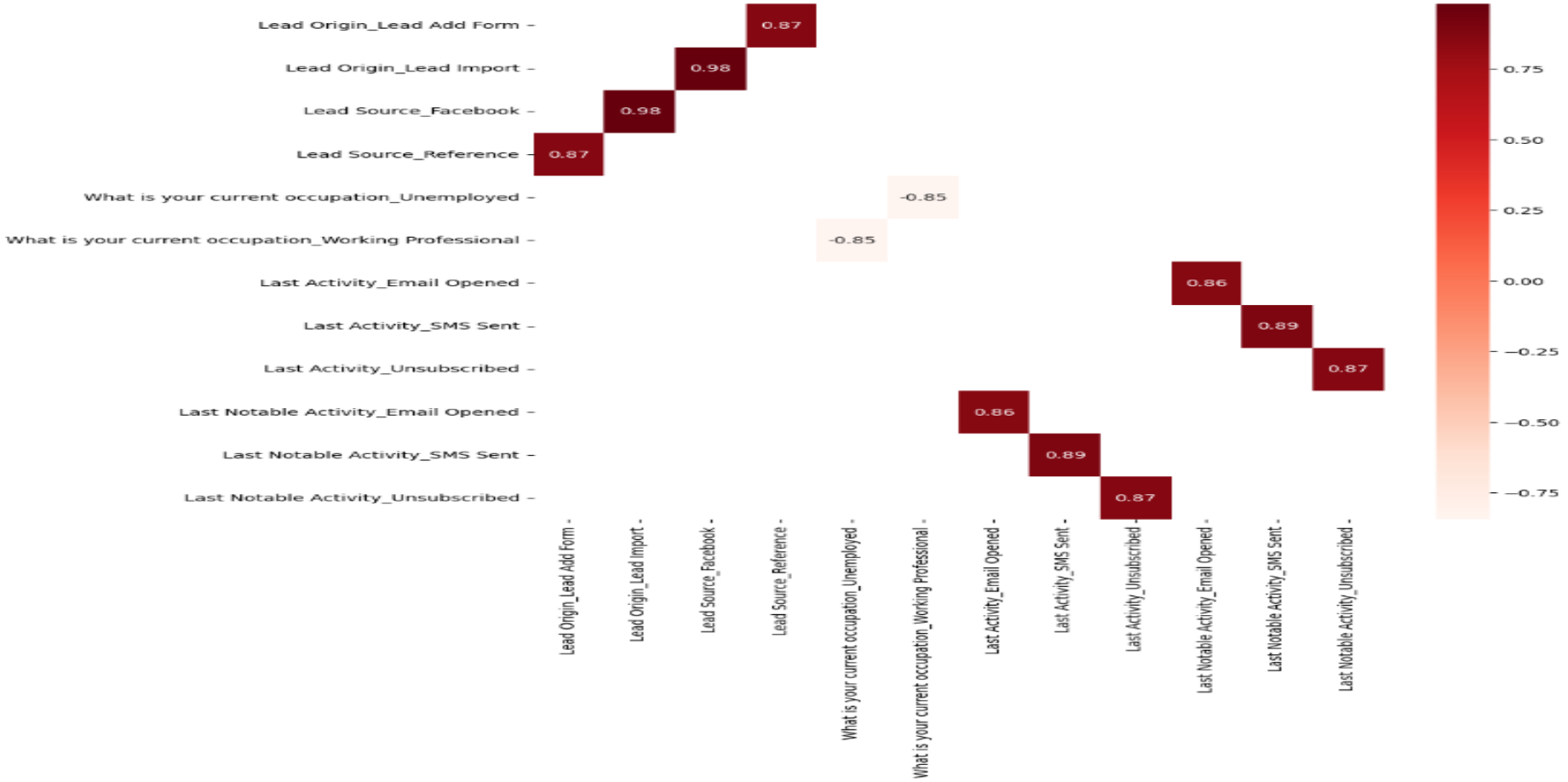
```
In [944]:
### Checking the Leads
Convert = (sum(Leads['Converted'])/len(Leads['Converted'].index))*100
Convert

Out[944]: 48.02810154877854
```

We have almost 48% Leads rate

Correlation Analysis

- Next step is to identify correlations between independent variables.
- For better clarity, the heatmap was restricted to 0.8 and greater correlation. The identified highly correlated variables were removed from the dataset.



Model Building

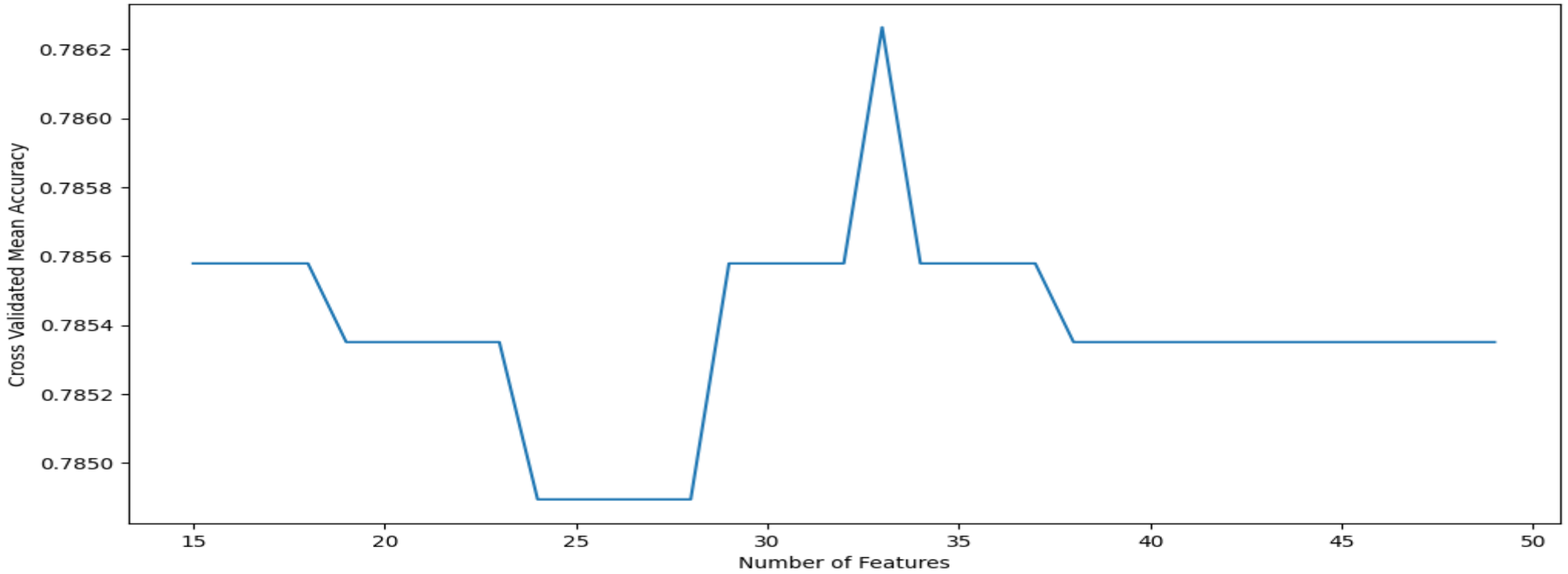
- A Generalized linear model (GLM) was run on the training data set.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4384
Model:	GLM	**Df Residuals:**	4315
Model Family:	Binomial	**Df Model:**	68
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Mon, 18 Dec 2023	**Deviance:**	1.1577e+05
Time:	21:48:08	**Pearson chi2:**	5.66e+18
No. Iterations:	100	**Pseudo R-squ. (CS):**	nan
Covariance Type:	nonrobust		

Feature selection using RFE

- Next step is to identify the optimum number of features using RFE and decided to use 33 as the optimal number of features based on the highest accuracy.




```
Optimal number of features to use is 33.0 which gives 0.786263361501174 accuracy.
```

Feature selection using RFE

- We used 33 as the step value in our RFE function to get the list of variables.

```
Index(['Total Time Spent on Website', 'Lead Origin_Landing Page Submission',
       'Lead Source_Direct Traffic', 'Lead Source_Google',
       'Lead Source_Olark Chat', 'Lead Source_Organic Search',
       'Lead Source_Referral Sites', 'Lead Source_Welingak Website',
       'Specialization_Banking, Investment And Insurance',
       'Specialization_Business Administration',
       'Specialization_Finance Management',
       'Specialization_Healthcare Management',
       'Specialization_Human Resource Management',
       'Specialization_IT Projects Management',
       'Specialization_Marketing Management',
       'Specialization_Operations Management',
       'Specialization_Rural and Agribusiness',
       'Specialization_Supply Chain Management',
       'How did you hear about X Education_Email',
       'What is your current occupation_Housewife',
       'What is your current occupation_Student',
       'Lead Profile_Dual Specialization Student',
       'Lead Profile_Lateral Student', 'Lead Profile_Potential Lead',
       'Lead Profile_Student of SomeSchool', 'Last Activity_Converted to Lead',
       'Last Activity_Email Bounced', 'Last Activity_Email Received',
       'Last Activity_Form Submitted on Website',
       'Last Activity_Had a Phone Conversation',
       'Last Activity_Olark Chat Conversation',
       'Last Activity_Page Visited on Website',
       'Last Activity_View in browser link Clicked',
       'Last Notable Activity_Email Link Clicked',
       'Last Notable Activity_Email Received',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Modified', 'Last Notable Activity_Unreachable'],
      dtype='object')
```

Assessing the model

- Next step is to get the GLM regression model summary.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4384
Model:	GLM	**Df Residuals:**	4345
Model Family:	Binomial	**Df Model:**	38
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1955.7
Date:	Mon, 18 Dec 2023	**Deviance:**	3911.5
Time:	22:00:26	**Pearson chi2:**	4.98e+03
No. Iterations:	24	**Pseudo R-squ. (CS):**	0.3888
Covariance Type:	nonrobust		

Assessing the model

- The confusion matrix was generated, and the overall accuracy was also analyzed.
- We need to perform few more adjustments as the model still has class imbalances.

[726]:
```
# Confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Leads, y_train_pred_final.predicted )
print(confusion)
```

```
[[1893  392]
 [ 535 1564]]
```

[727]:
```
# Let's check the overall accuracy.
print(metrics.accuracy_score(y_train_pred_final.Leads, y_train_pred_final.predicted))
```

```
0.7885492700729927
```

Variance Inflation Factors

- Next step is to generate VIFs and eliminate features with high VIF's which ensuring the accuracy doesn't drop much.

[730]:

	Features	VIF
1	Lead Origin_Landing Page Submission	5.61
2	Lead Source_Direct Traffic	3.72
3	Lead Source_Google	2.51
34	Last Notable Activity_Had a Phone Conversation	2.46
29	Last Activity_Had a Phone Conversation	2.46
35	Last Notable Activity_Modified	2.08
33	Last Notable Activity_Email Received	2.01
27	Last Activity_Email Received	2.01
5	Lead Source_Organic Search	1.51

```
0.7833029197080292
```

So overall the accuracy hasn't dropped much.

	Features	VIF
15	Last Notable Activity_Modified	2.02
2	Lead Source_Google	1.38
12	Last Activity_Olark Chat Conversation	1.38
1	Lead Source_Direct Traffic	1.33
3	Lead Source_Olark Chat	1.31
9	Last Activity_Converted to Lead	1.29
7	Lead Profile_Potential Lead	1.24
0	Total Time Spent on Website	1.19
13	Last Activity_Page Visited on Website	1.14
4	Lead Source_Organic Search	1.13
6	Specialization_Marketing Management	1.13
10	Last Activity_Email Bounced	1.10
8	Lead Profile_Student of SomeSchool	1.09
14	Last Notable Activity_Email Link Clicked	1.03
5	Lead Source_Referral Sites	1.02
11	Last Activity_Had a Phone Conversation	1.01

Confusion matrix & ROC Curve

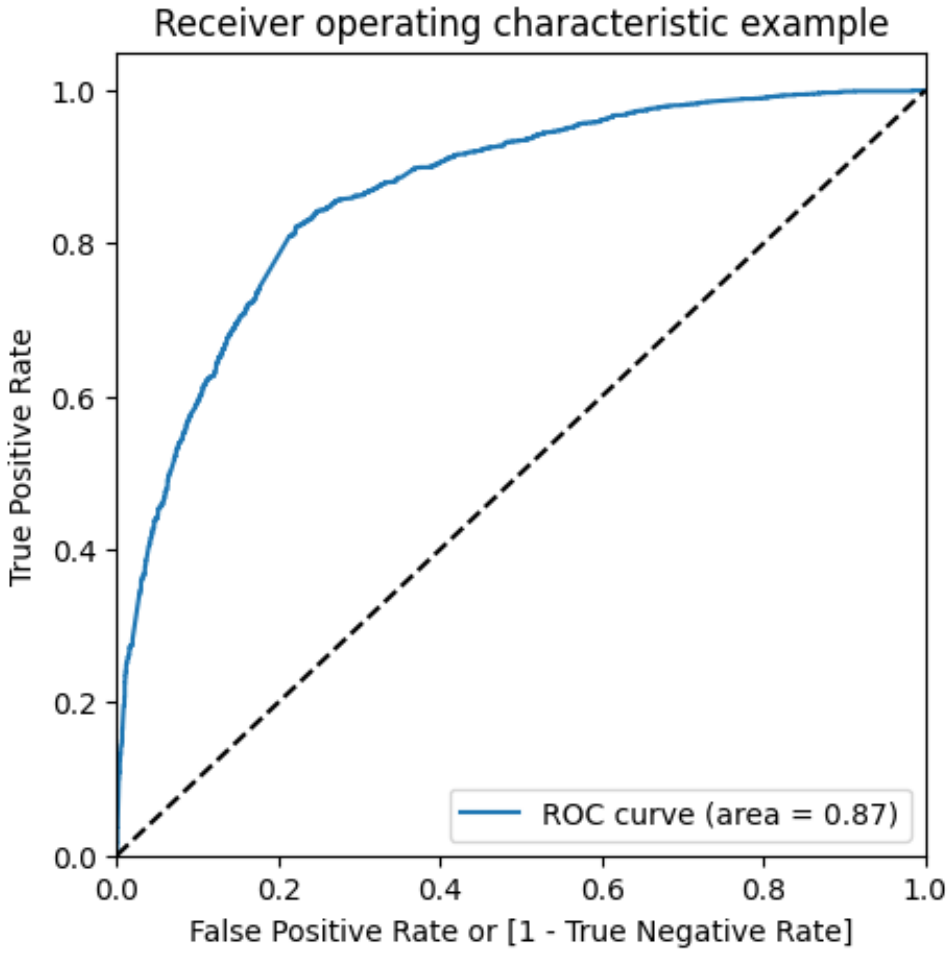
- Next the sensitivity and specificity is calculated along with the ROC curve.
- All values appear to be in acceptable limits.

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.7389232968080038
```

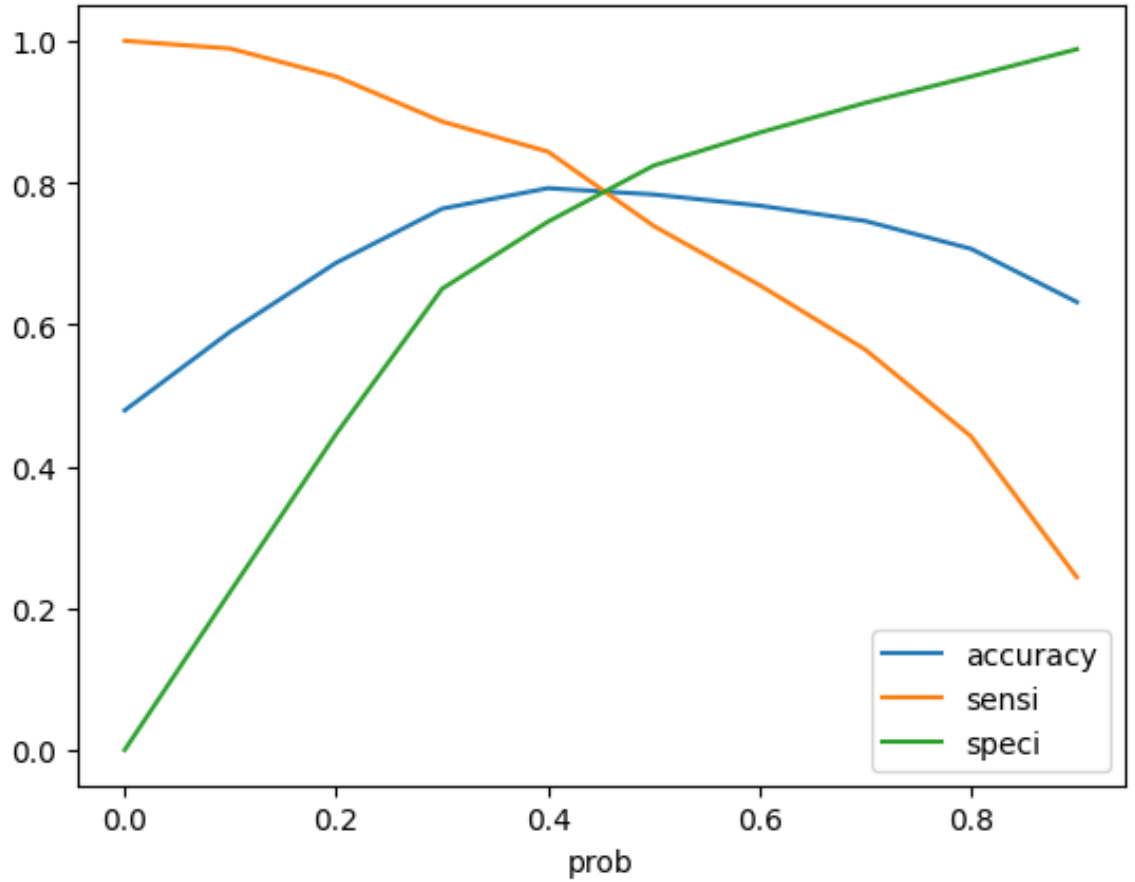
```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8240700218818381
```



Optimal Cut-off point

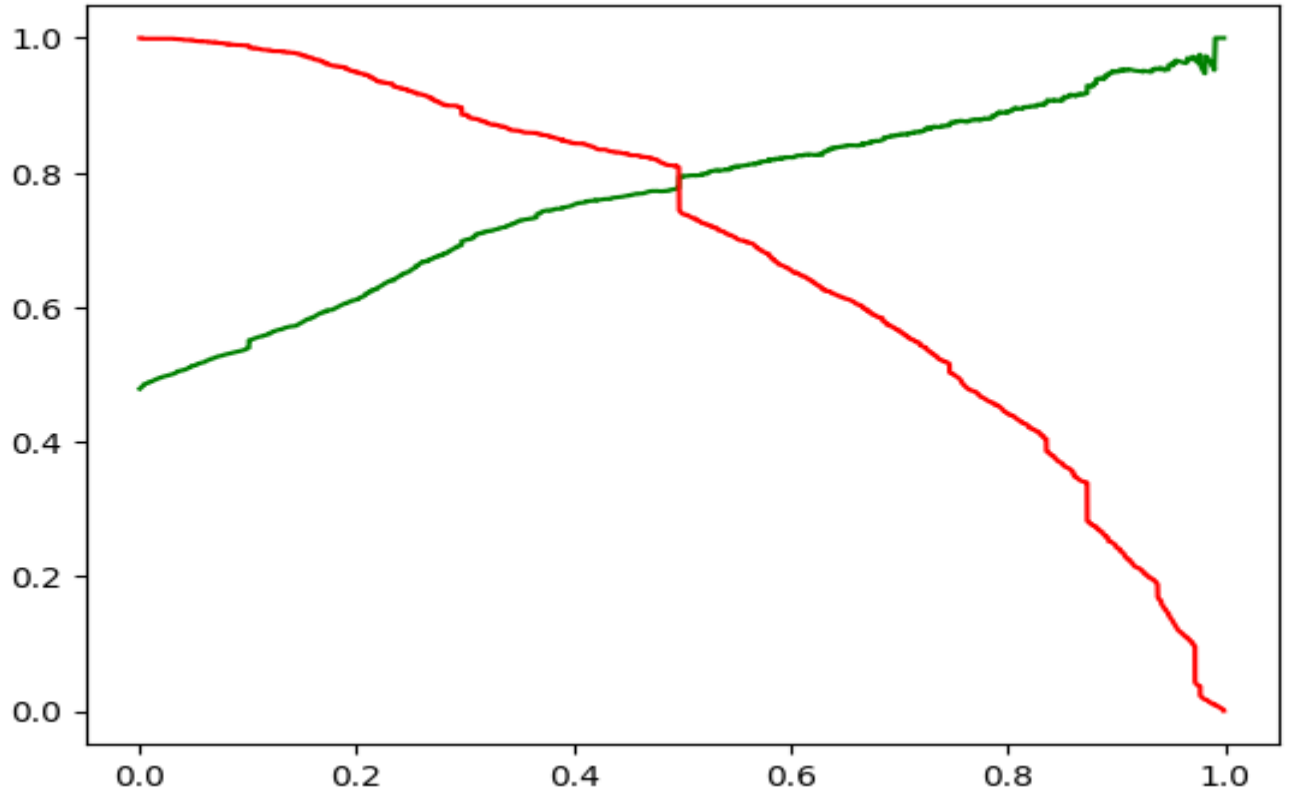
- Next step is identifying the optimal cut-off point for the lead probability which is 0.5 in our case.




```
Optimum cut-off value is: 0.5
```

Precision – Recall tradeoff

- Next step is to analyze the precision recall curve to identify any anomalies which turned out to be okay.



Predictions on test set

- Finally, we identify the accuracy, specificity and sensitivity of the test set data.

```
# Let's check the overall accuracy.
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.768493879723257
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8010309278350516
```

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.7337733773377337
```