Lead Scoring Case Study

Done By :

Shenaz Rahaman

Saqib Mohammed

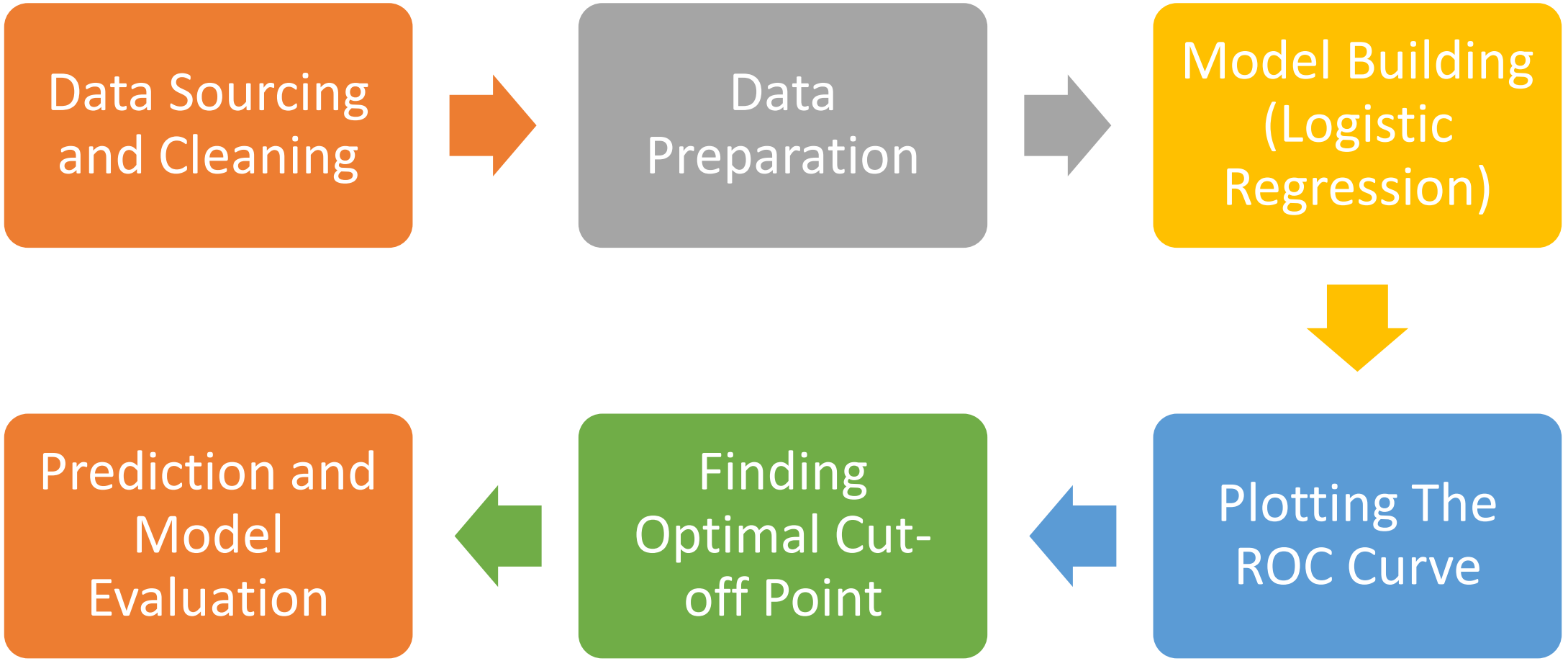
Problem Statement :

- An education company named 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%.
- 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. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

Aim:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Model should be able to adjust to if the company's requirement changes in the future.
- **The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.**

Approach


Data Sourcing and Cleaning
Data Preparation
Model Building (Logistic Regression)
Plotting The ROC Curve
Finding Optimal Cut-off Point
Prediction and Model Evaluation

Data Understanding and Manipulation

- Lead Data Set provided for analysis: 9240 rows and 37 columns.
- Converting “Select” to “Nan” (“Select” implies that user has not selected any value)
- Drop columns which have only one Unique value, OR the columns which have very less variation.
- Imputing the “NULL” across different variables, with respective values
- Converting categorical variables to dummy features (Binary variables for yes or no)
- Outlier Treatment
- Test -train split of the data
- Feature Scaling
- Data normalization, and dropping variables in scenarios showing highly co-related variables

Logistic Regression Model

- After EDA, Logistic Regression Model is built in python using **GLM()** function, under statsmodel library.
- The model contained all the variables, some of which had insignificant coefficients, Such variables are removed using
- Automated Approach: RFE (Recursive feature elimination) with number of features = 15.
- Manual approach based on VIFs and p-values.
- The final tally of variables with their respective values
  - Significant p-values near to zero
  - VIFs < 5

Final Model p_values

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.3986	0.072	5.511	0.000	0.257	0.540
Do Not Email	-1.6232	0.173	-9.371	0.000	-1.963	-1.284
Total Time Spent on Website	1.1075	0.041	27.192	0.000	1.028	1.187
LeadOrigin_Lead Add Form	3.6715	0.216	17.021	0.000	3.249	4.094
LeadSource_Olark Chat	1.3599	0.106	12.784	0.000	1.151	1.568
LeadSource_Welingak Website	2.1387	0.746	2.868	0.004	0.677	3.600
LastActivity_Not Mentioned	-1.4629	0.448	-3.264	0.001	-2.342	-0.584
LastActivity_Olark Chat Conversation	-1.1031	0.192	-5.743	0.000	-1.480	-0.727
CurrentOccupation_Not Avaiable	-1.1303	0.089	-12.740	0.000	-1.304	-0.956
CurrentOccupation_Working Professional	2.3214	0.180	12.870	0.000	1.968	2.675
LastNotableActivity_Email Link Clicked	-1.6136	0.260	-6.210	0.000	-2.123	-1.104
LastNotableActivity_Email Opened	-1.3750	0.090	-15.317	0.000	-1.551	-1.199
LastNotableActivity_Modified	-1.8315	0.099	-18.577	0.000	-2.025	-1.638
LastNotableActivity_Olark Chat Conversation	-1.5178	0.371	-4.087	0.000	-2.246	-0.790
LastNotableActivity_Page Visited on Website	-1.6540	0.213	-7.769	0.000	-2.071	-1.237

Model 3 has all variables with low p-values, making it stable.

Based on the coefficient values of the final model, the top three variables contributing towards lead conversion rate are:

1. LeadOrigin_Lead Add Form (from column Lead Origin)
2. CurrentOccupation_Working Professional (from column What is your current occupation)
3. LeadSource_Wellingak Website (from column Lead Source)

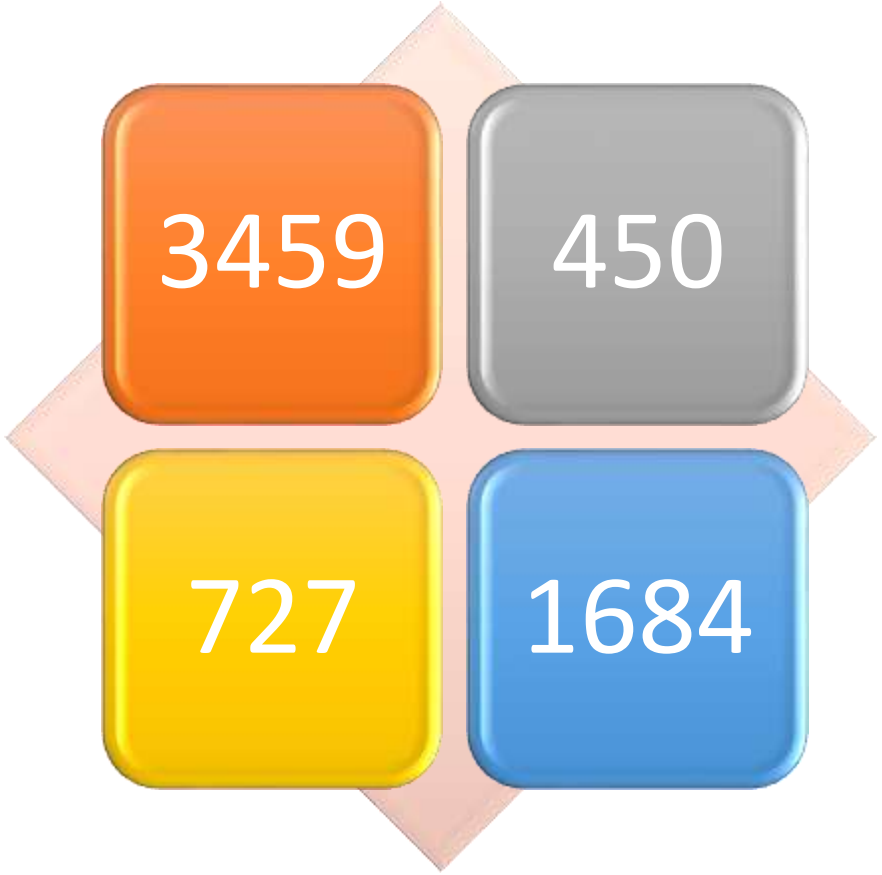
Final VIF

	Features	VIF
6	LastActivity_Olark Chat Conversation	1.96
11	LastNotableActivity_Modified	1.76
3	LeadSource_Olark Chat	1.72
2	LeadOrigin_Lead Add Form	1.61
7	CurrentOccupation_Not Avaiable	1.58
12	LastNotableActivity_Olark Chat Conversation	1.32
4	LeadSource_Welingak Website	1.29
1	Total Time Spent on Website	1.23
10	LastNotableActivity_Email Opened	1.22
5	LastActivity_Not Mentioned	1.17
8	CurrentOccupation_Working Professional	1.15
0	Do Not Email	1.11
13	LastNotableActivity_Page Visited on Website	1.04
9	LastNotableActivity_Email Link Clicked	1.03

All the variables have a low VIF value and need not be dropped.

Model Evaluation and Optimization

Confusion Matrix



Accuracy:

Here, the model Accuracy is about 0.8137658227848101 i.e **81%** which is good but is not reliable. Hence, we need to calculate other metrics also.

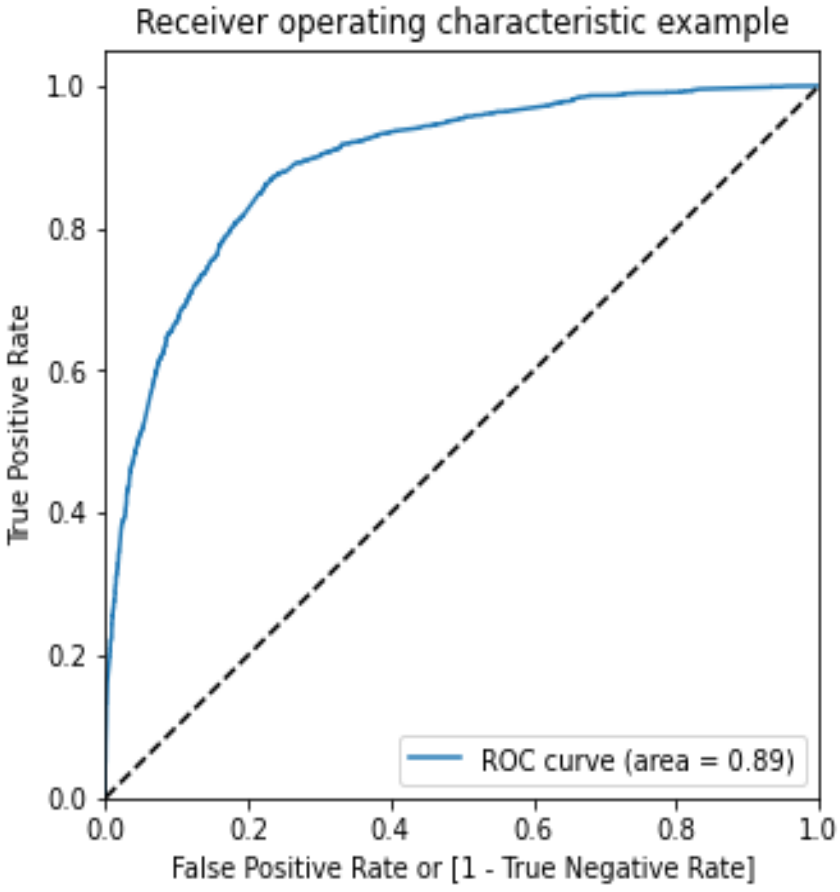
Metrics beyond accuracy

- Sensitivity : 0.6984653670676068 (70%)
- Specificity : 0.8848810437452034 (88%)
- False Positive Rate : 0.1151189562547966 2 (11%)
- Positive predictive value : 0.7891283973758201 (79%)
- Negative predictive value : 0.8263258480649786 (83%)

Plotting the ROC Curve

An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- So Our Curve Looks Fine.



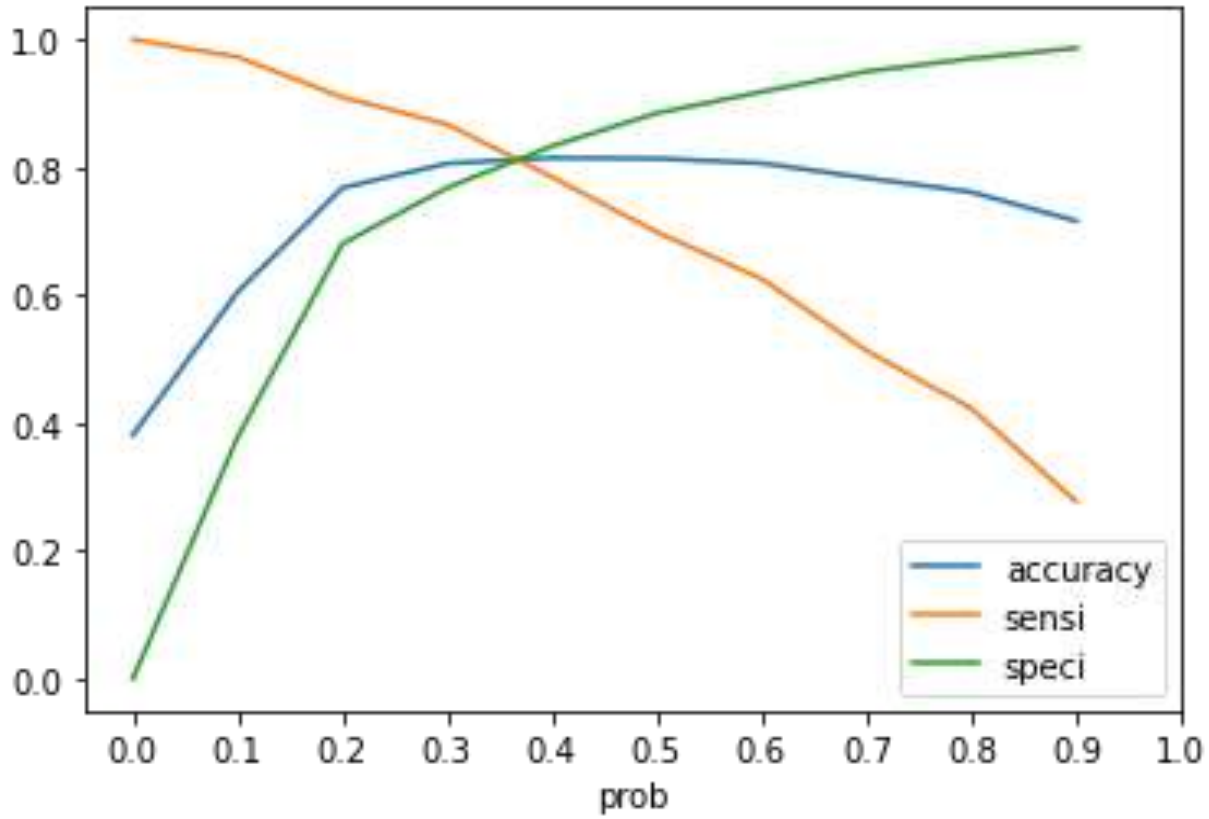
Finding Optimal Cutoff Point

- Optimal cutoff probability is that probability where we get balanced sensitivity and specificity

  - Creating columns with different probability cutoffs
  - Calculating accuracy, sensitivity and specificity for various probability cutoffs.

	prob	accuracy	sensi	speci
0.0	0.0	0.381487	1.000000	0.000000
0.1	0.1	0.605063	0.972625	0.378358
0.2	0.2	0.768038	0.909581	0.680737
0.3	0.3	0.806013	0.867275	0.768227
0.4	0.4	0.814715	0.784737	0.833205
0.5	0.5	0.813766	0.698465	0.884881
0.6	0.6	0.806487	0.625467	0.918138
0.7	0.7	0.783228	0.513065	0.949859
0.8	0.8	0.761234	0.422646	0.970069
0.9	0.9	0.715981	0.277063	0.986697

Plotting accuracy, sensitivity and specificity for various probabilities.



From the curve we can observe, 0.37 is the optimum point to take it as a cutoff probability.

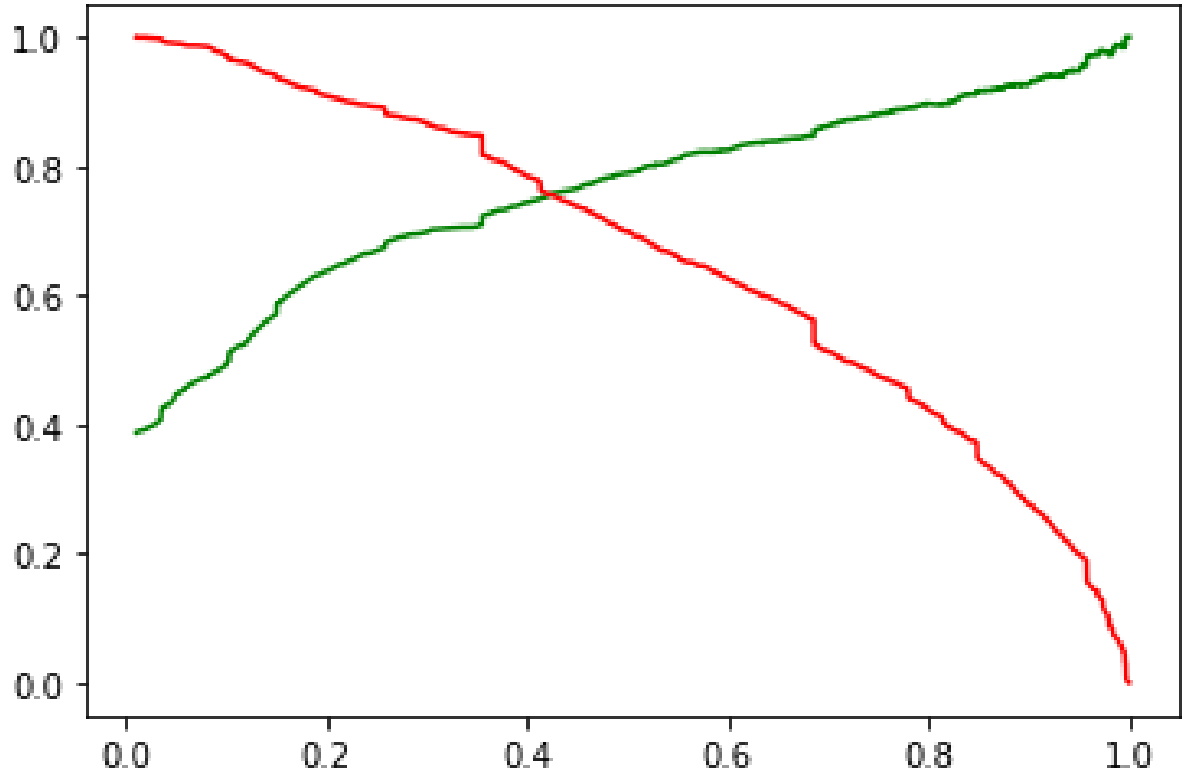
We Can Make The Final Prediction using This Cut-off

Final Model Measures (Train Set)

- Here, the final prediction of conversions is about 81% (>80% target), making it a good model.
- Accuracy : 0.8134493670886076 (81%)
- Sensitivity : 0.809207797594359 2 (81%)
- Specificity : 0.8160654898951138 (82%)
- False Positive Rate : 0.1839345101048861 5 (18%)
- Positive predictive value : 0.7307116104868914 (73%)
- Negative predictive value : 0.87397260273972 6 (87)

Precision and Recall Tradeoff (Train Set)

- Precision: 0.7891283973758201 (79%)
- Recall: 0.6984653670676068 (70%)

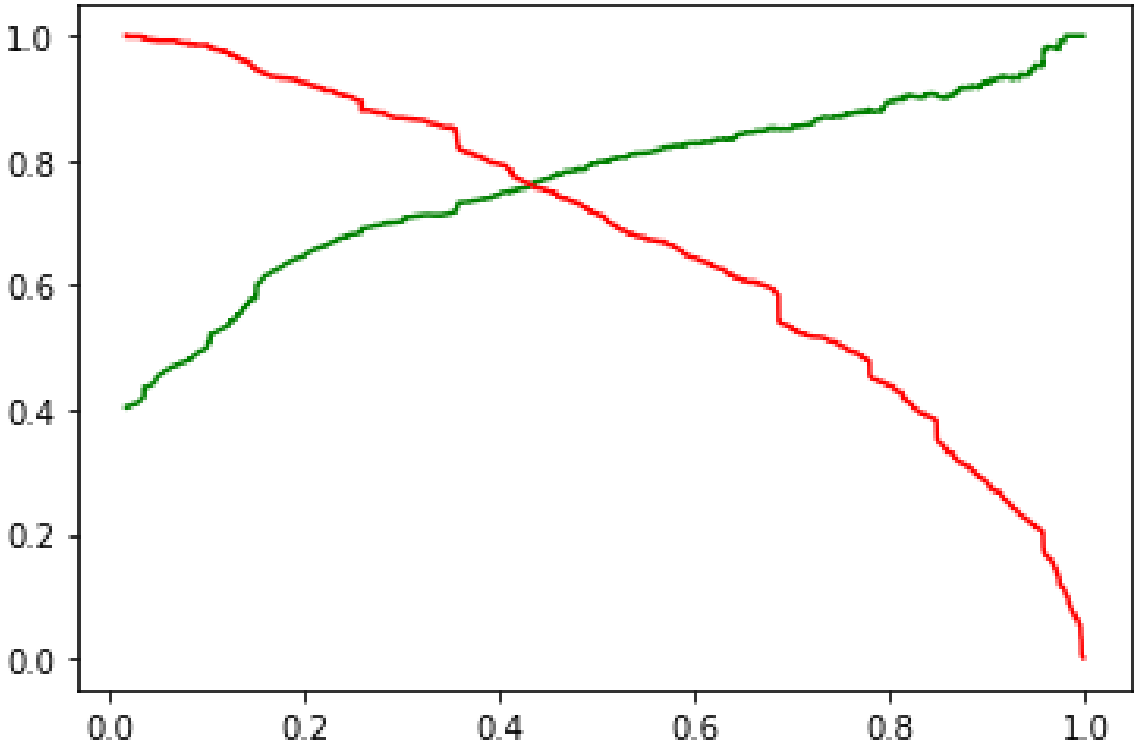


Final Model Measures (Test Set)

- Here, the final prediction of conversions is about 81% (>80% target), which is same as training data, Hence making it a good model.
- Accuracy : 0.810631229235880 4 (81%)
- Sensitivity: 0.81238273921200 75 (81%)
- Specificity: 0.80949482653682 29 (81%)
- False Positive Rate: 0.19050517346317 71 (19%)
- Positive predictive value: 0.73452078032230 7 (73%)
- Negative predictive value: 0.86928104575163 4 (87%)

Precision and Recall Tradeoff (Test Set)

- Precision: 0.734520780322307 (73%)
- Recall: 0.8123827392120075 (81%)



Conclusion

As per the final Model (logm3):

- Accuracy, Sensitivity & Specificity of the train and test set are around 81%, indicating that the model is a good predictor of the potential leads.
- The lead score calculated in the test set has a lead conversion rate of about 81%, which satisfies the target set by the CEO.
- The optimal cut-off based on Sensitivity-Specificity is 0.37 and based on Precision-Recall is 0.41, indicating that the model is stable