LEAD SCORING CASE STUDY

BY

Manu Saxena

Batch DS C45

PROBLEM STATEMENT & GOALS

Problem Statement

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

There are quite a few goals for this case study:

1.) 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.

2.)There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

ANALYSIS APPROACH AND SOLUTION

1. Data understanding, preparation and EDA - All data quality checks are performed and issues were addressed. Dummy variables were also created for all applicable. The data is converted to a clean format suitable for analysis in Python.

2. Model building - Logistic regression models are made using appropriate variables and the best one is chosen based on key performance metrics post model parameters tuning.

3. Model evaluation – The selected model is evaluated using both evaluation metrics viz sensitivity-specificity view, and precision-recall view.

Data Cleaning

Here it is observed that below columns are having high number of missing values i.e. > 40%

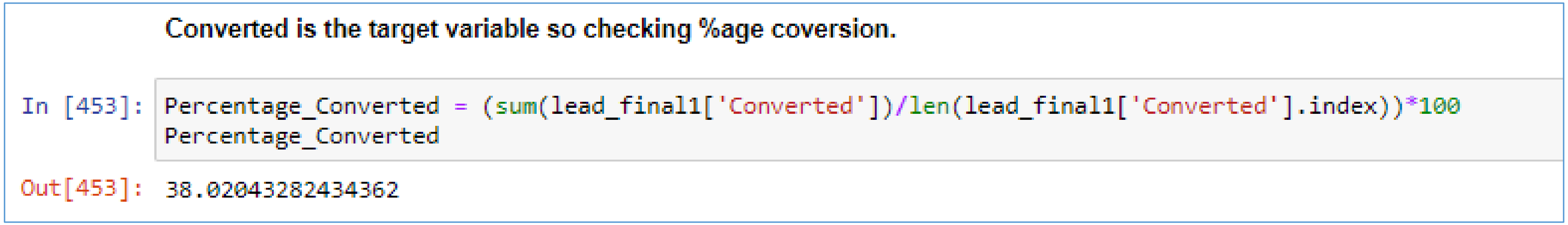
- Lead Quality.
- Asymmetrique Activity Index.
- Asymmetrique Profile Index.
- Asymmetrique Activity Score.
- Asymmetrique Profile Score.

Apart from the above tags also have a very high percentage of missing value. Removed the same.

Removed Lead profile column as we need to identify the potential leads and it has not so many details.

After deleting the high value missing columns we imputed missing values with “Not Available in data” for below columns:
1.) Country.
2.) Specialization.
3.) How did you hear about X Education.
4.) What is your current occupation.
5.) What matters most to you in choosing a course.
6.) City.

Conversion rate of Lead



Converted is the target variable so checking %age coversion.

```
In [453]: Percentage_Converted = (sum(lead_final1['Converted'])/len(lead_final1['Converted'].index))*100
          Percentage_Converted

Out[453]: 38.02043282434362
```

So over-all lead conversion rate is ~38%.

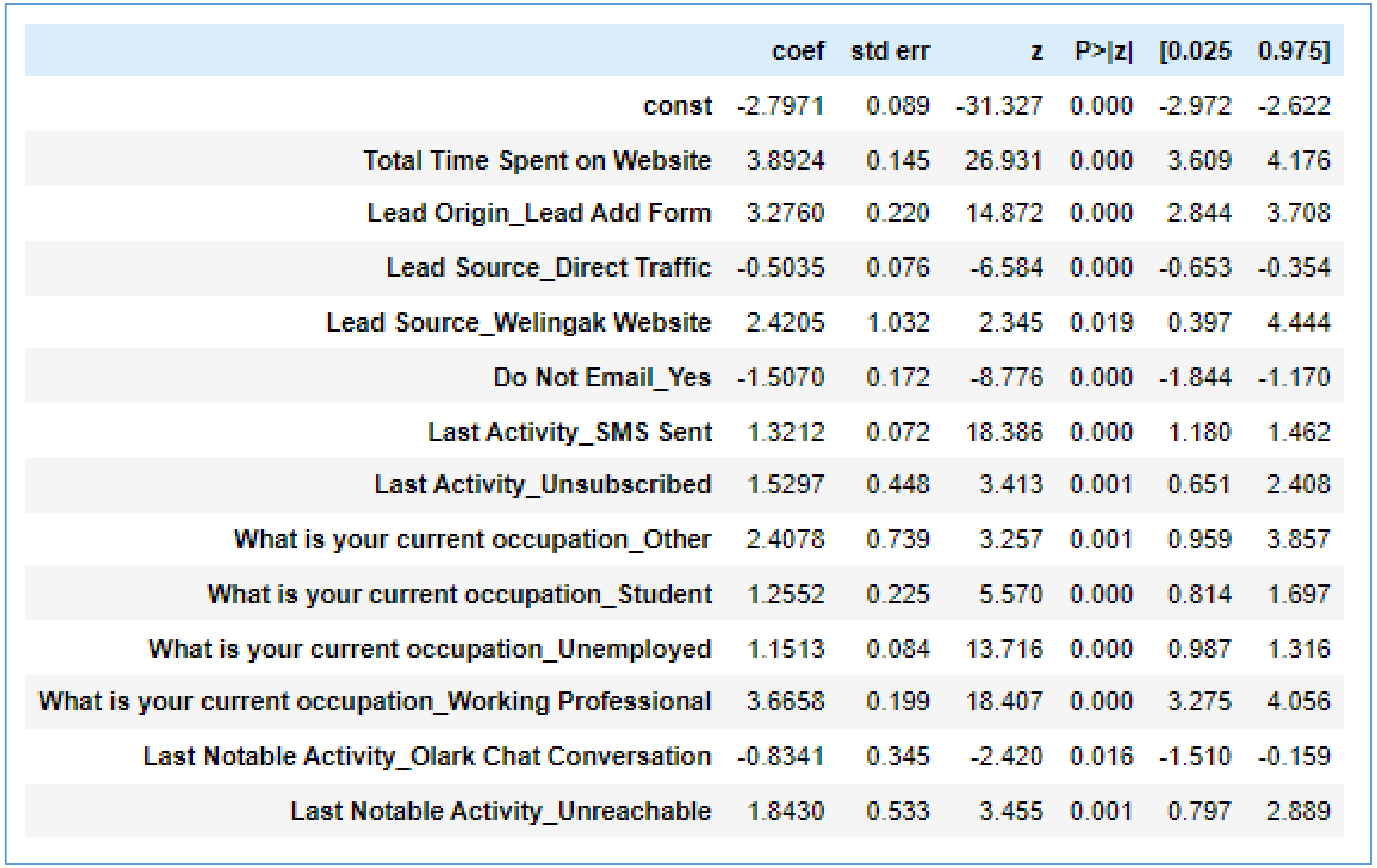
Model Building

Logistic regression models are made with the finalized variables in such a way that P-values of the model is less than 0.05 and VIF is less than 5.

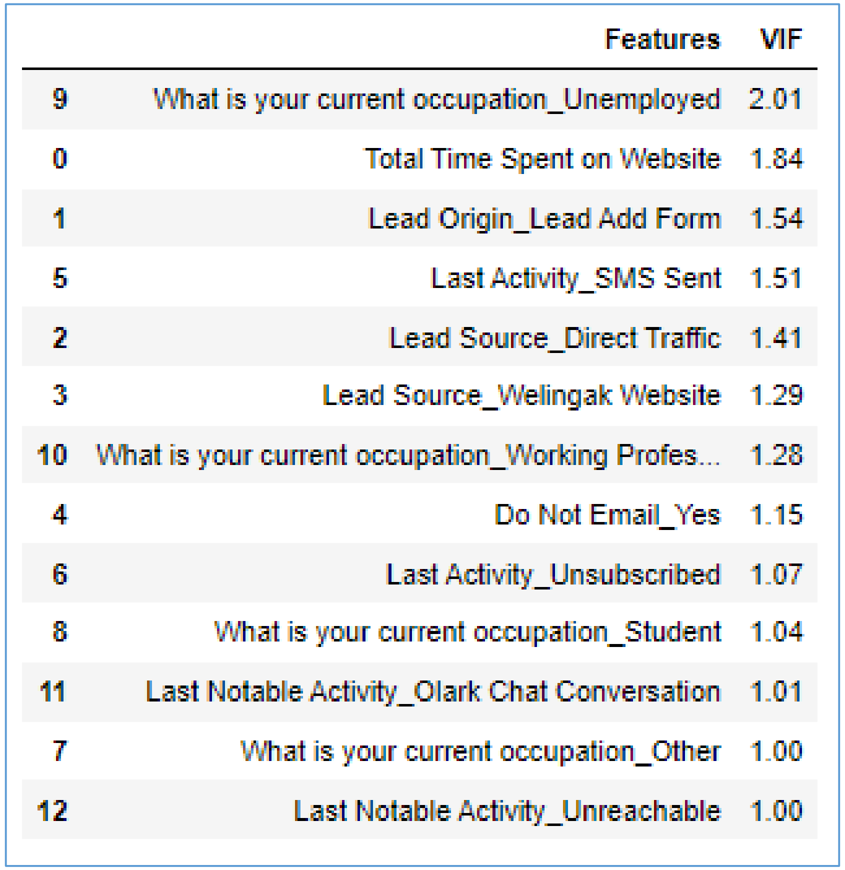
Below model was best fitted in the equation.

ln[P/(1-P)] = -2.79 + 3.8924*(Total Time Spent on Website) + 3.2760*(Lead Origin_Lead Add Form) – 0.5035*(Lead Source_Direct Traffic) + 2.4205*(Lead Source_Welingak Website) - 1.5070*(Do Not Email_Yes) + 1.3212*(Last Activity_SMS Sent) + 1.5297*(Last Activity_SMS Sent) + 2.4078*(What is your current occupation_Other) + 1.2552*(What is your current occupation_Student) + 1.1513*(What is your current occupation_Unemployed) + 3.6658*(What is your current occupation_Working Professional) – 0.8341*(Last Notable Activity_Olark Chat Conversation) + 1.8430*(Last Notable Activity_Unreachable).

Final Model



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7971	0.089	-31.327	0.000	-2.972	-2.622
Total Time Spent on Website	3.8924	0.145	26.931	0.000	3.609	4.176
Lead Origin_Lead Add Form	3.2760	0.220	14.872	0.000	2.844	3.708
Lead Source_Direct Traffic	-0.5035	0.076	-6.584	0.000	-0.653	-0.354
Lead Source_Welingak Website	2.4205	1.032	2.345	0.019	0.397	4.444
Do Not Email_Yes	-1.5070	0.172	-8.776	0.000	-1.844	-1.170
Last Activity_SMS Sent	1.3212	0.072	18.386	0.000	1.180	1.462
Last Activity_Unsubscribed	1.5297	0.448	3.413	0.001	0.651	2.408
What is your current occupation_Other	2.4078	0.739	3.257	0.001	0.959	3.857
What is your current occupation_Student	1.2552	0.225	5.570	0.000	0.814	1.697
What is your current occupation_Unemployed	1.1513	0.084	13.716	0.000	0.987	1.316
What is your current occupation_Working Professional	3.6658	0.199	18.407	0.000	3.275	4.056
Last Notable Activity_Olark Chat Conversation	-0.8341	0.345	-2.420	0.016	-1.510	-0.159
Last Notable Activity_Unreachable	1.8430	0.533	3.455	0.001	0.797	2.889

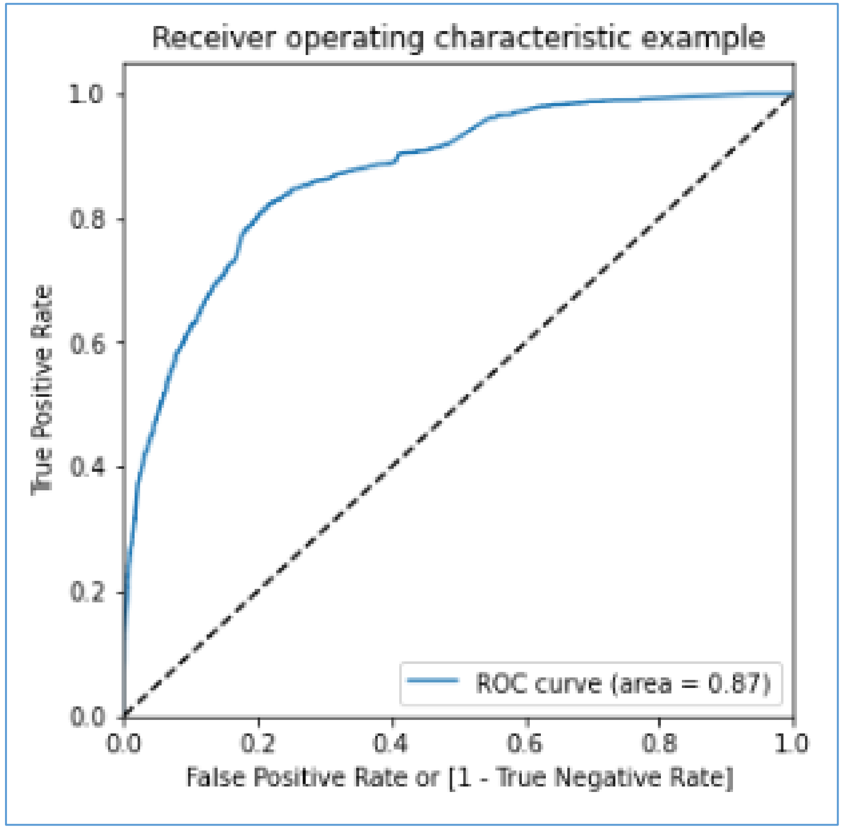


	Features	VIF
9	What is your current occupation_Unemployed	2.01
0	Total Time Spent on Website	1.84
1	Lead Origin_Lead Add Form	1.54
5	Last Activity_SMS Sent	1.51
2	Lead Source_Direct Traffic	1.41
3	Lead Source_Welingak Website	1.29
10	What is your current occupation_Working Profes...	1.28
4	Do Not Email_Yes	1.15
6	Last Activity_Unsubscribed	1.07
8	What is your current occupation_Student	1.04
11	Last Notable Activity_Olark Chat Conversation	1.01
7	What is your current occupation_Other	1.00
12	Last Notable Activity_Unreachable	1.00

Model Evaluation

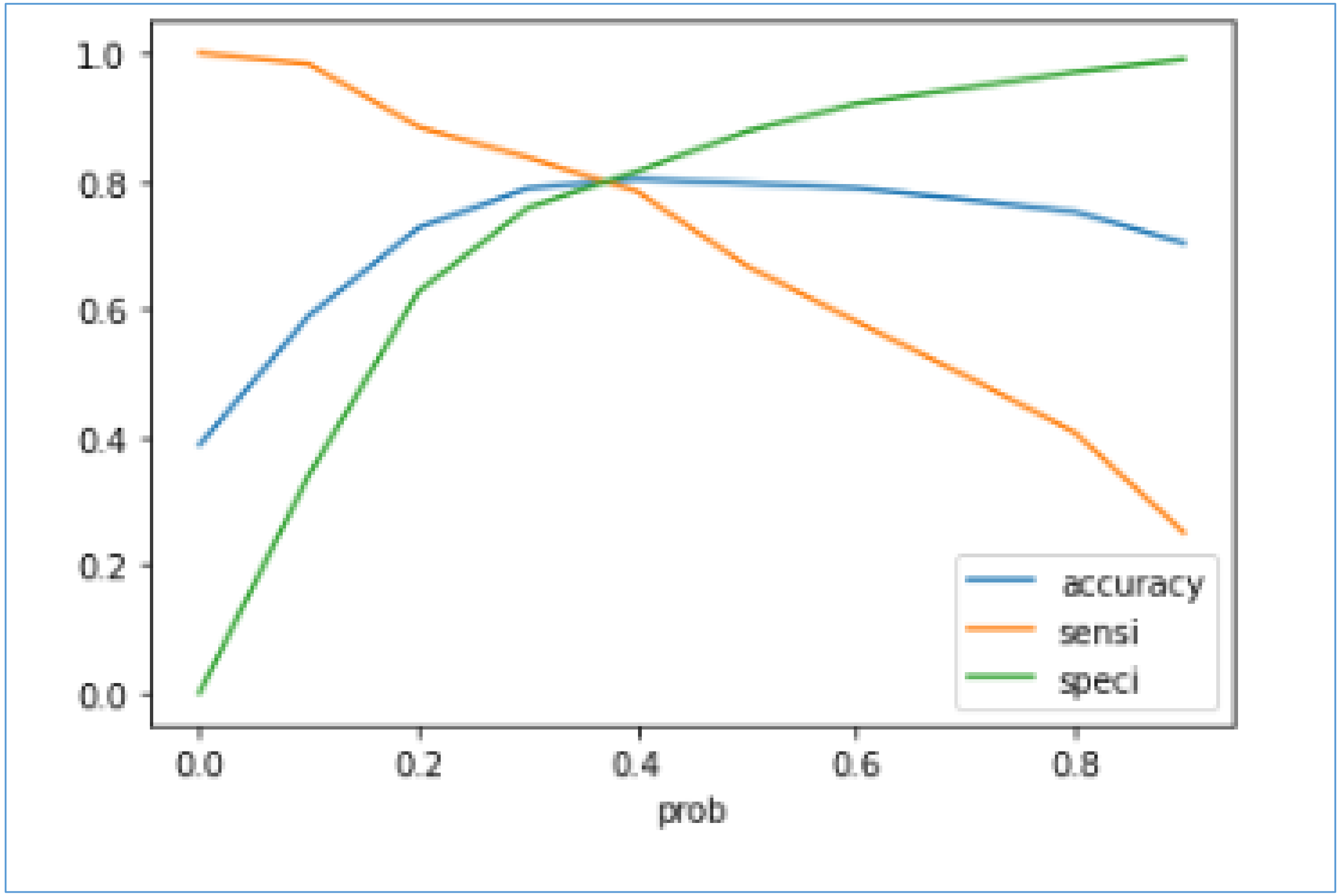
The accuracy obtained on the train set is 0.7962, sensitivity = 0.6677 and specificity = 0.8777 with optimal cut off = 0.40.

Area under ROC curve is 0.87



Model Evaluation

From accuracy, sensitivity & specificity graph, optimal value of probability cut off is coming as 0.4



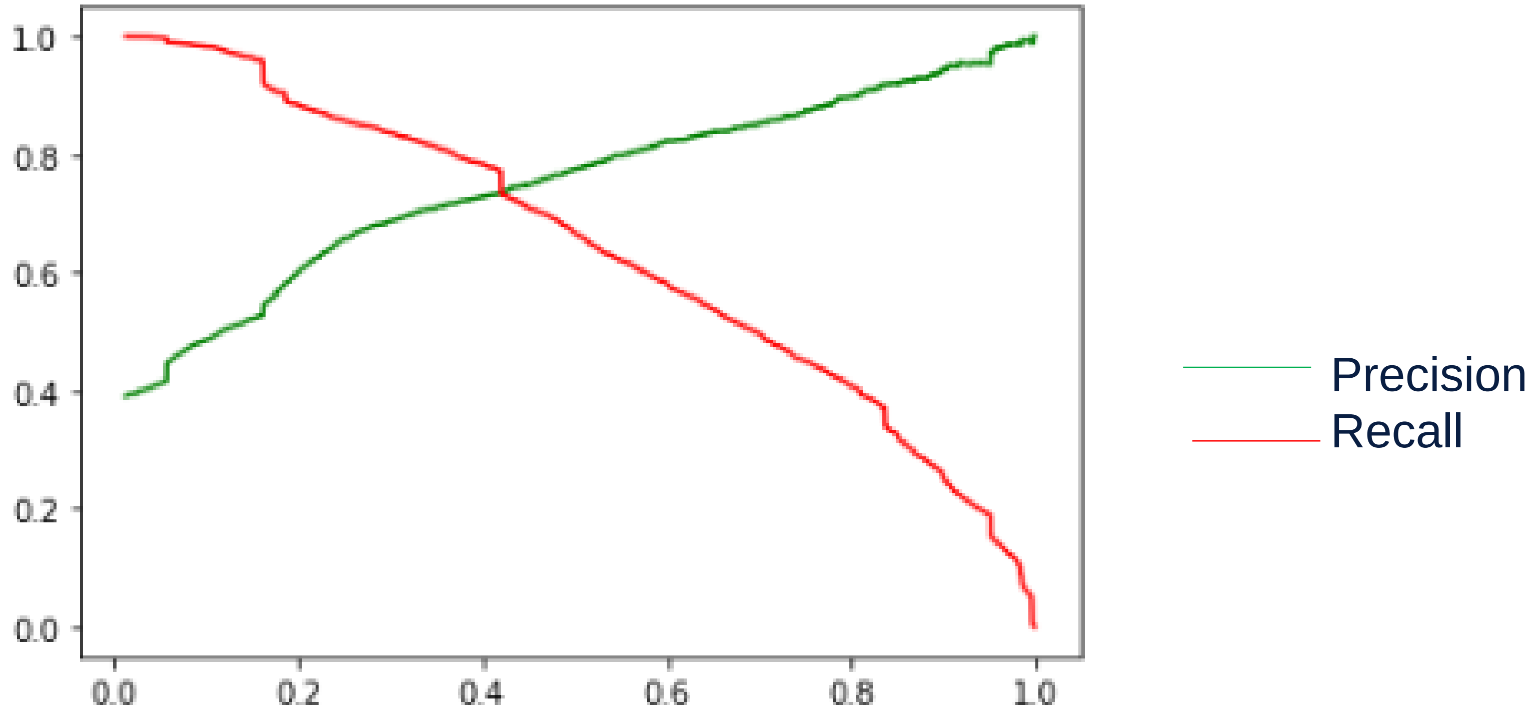
Model Evaluation

- The accuracy on the train set is 0.8032, sensitivity = 0.7838 and specificity = 0.8154 with probability cut off = 0.40

- The accuracy on the test set is 0.8125, sensitivity = 0.8101 and specificity = 0.8139 with probability cut off = 0.40

Model Evaluation: Precision and Recall

- The accuracy on the train set is 0.7942, precision = 0.7353 and recall = 0.7333 with probability cut off = 0.42
- From Precision & Recall graph, optimal value of probability cut off is coming as 0.40.



Model Evaluation: Precision and Recall

- The accuracy on the train set is 0.8032, sensitivity = 0.7838 and specificity = 0.8154 with probability cut off = 0.40

- The accuracy on the test set is 0.8033, precision = 0.7318 and recall = 0.7222 with probability cut off = 0.42

Model Observation

The model is evaluated with both Sensitivity-Specificity view & Precision-Recall view

- The area under ROC curve is 0.87, hence the model is very good.
- Probability cut off for Sensitivity-Specificity view is 0.40
- Probability cut off for Precision-Recall view is 0.42

Below is equation for log odds.

ln[P/(1-P)] = -2.79 + 3.8924*(Total Time Spent on Website) + 3.2760*(Lead Origin_Lead Add Form) – 0.5035*(Lead Source_Direct Traffic) + 2.4205*(Lead Source_Welingak Website) - 1.5070*(Do Not Email_Yes) + 1.3212*(Last Activity_SMS Sent) + 1.5297*(Last Activity_SMS Sent) + 2.4078*(What is your current occupation_Other) + 1.2552*(What is your current occupation_Student) + 1.1513*(What is your current occupation_Unemployed) + 3.6658*(What is your current occupation_Working Professional) – 0.8341*(Last Notable Activity_Olark Chat Conversation) + 1.8430*(Last Notable Activity_Unreachable).

Model Conclusion

Below are the top 3 variables which contribute most towards the probability of a lead getting converted.

Total Time Spent on Website –Higher the value, higher is the probability of lead converting successfully. It has highest coefficient viz 3.8924.
What is your current occupation – Working professional looking to upgrade the skills have higher probability of lead converting successfully. Its coefficient is 3.6658
Lead Origin – If “Lead origin” = “Lead Add Form”, probability of lead converting successfully is high. Its coefficient is 3.2760