Lead Score Case Study

BY MANSHA CHHABRA

Contents

1.Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. Once people land on their web site, browse thru courses or watch videos. When these people fill a form providing email address and contact number they are considered as Leads. The lead conversion of X Education is very poor. Eg. On any given if they acquire 100 leads, only 30 of them are converted.

You are expected to help X Education to select most promising leads ie Hot leads that are most likely to convert to paying customers.

The company require you to build a model wherein you will assign a lead score to each of the leads such that customers with higher lead scores have higher conversion chance and the customers with lower lead score have lower conversion chance.

Lead score = 100 * Probability

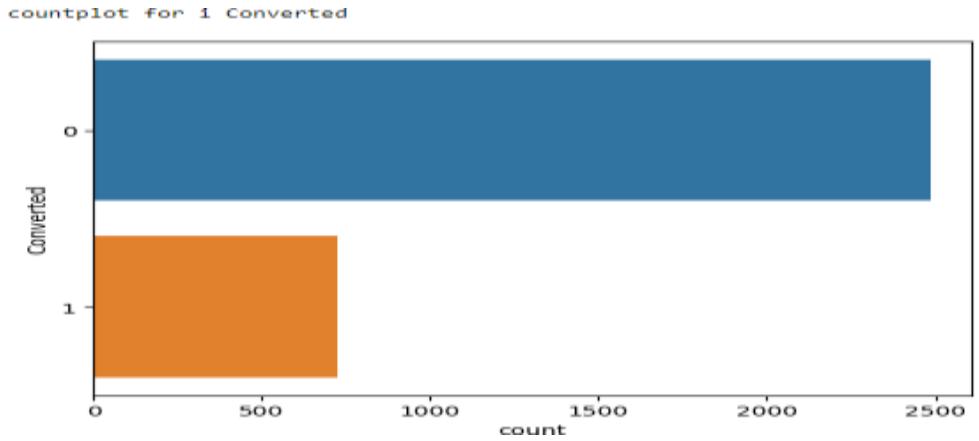
2.Approach Overall

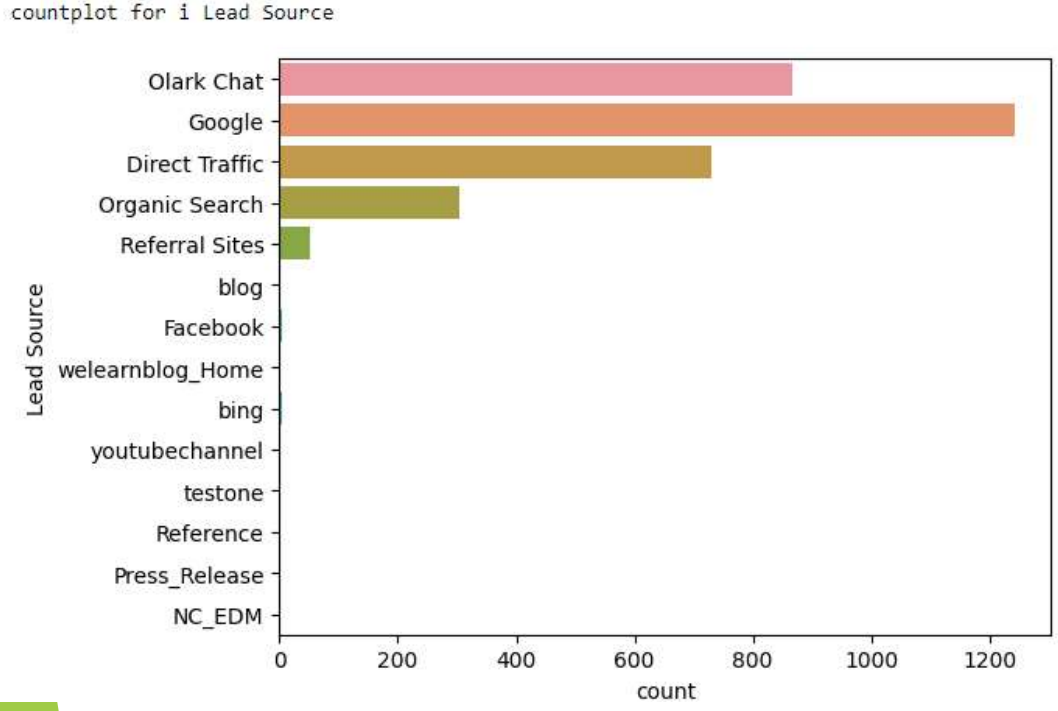
- Import leads data set ie lead.csv.
- Check the structure/metadata of the data.
- Missing value check.
- Outlier Check.
- Perform Univariate Analysis.
- Perform Bivariate Analysis.
- Segmentation done based on TARGET variable ie
- converted.
- Data Cleaning / Pre-processing
- Data preparation
- Logistics Model Building and Model Evaluation
- Predicting with Test data
- ROC Curve
- Optimal Cut off Point

Assumptions during Data Cleanup

- 'Select' value is considered as Null values for all variables.
- 30 - 40% is set as threshold for null values. ie any column with null values above this threshold will be dropped.

EDA – Univariate and Bivariate Analysis



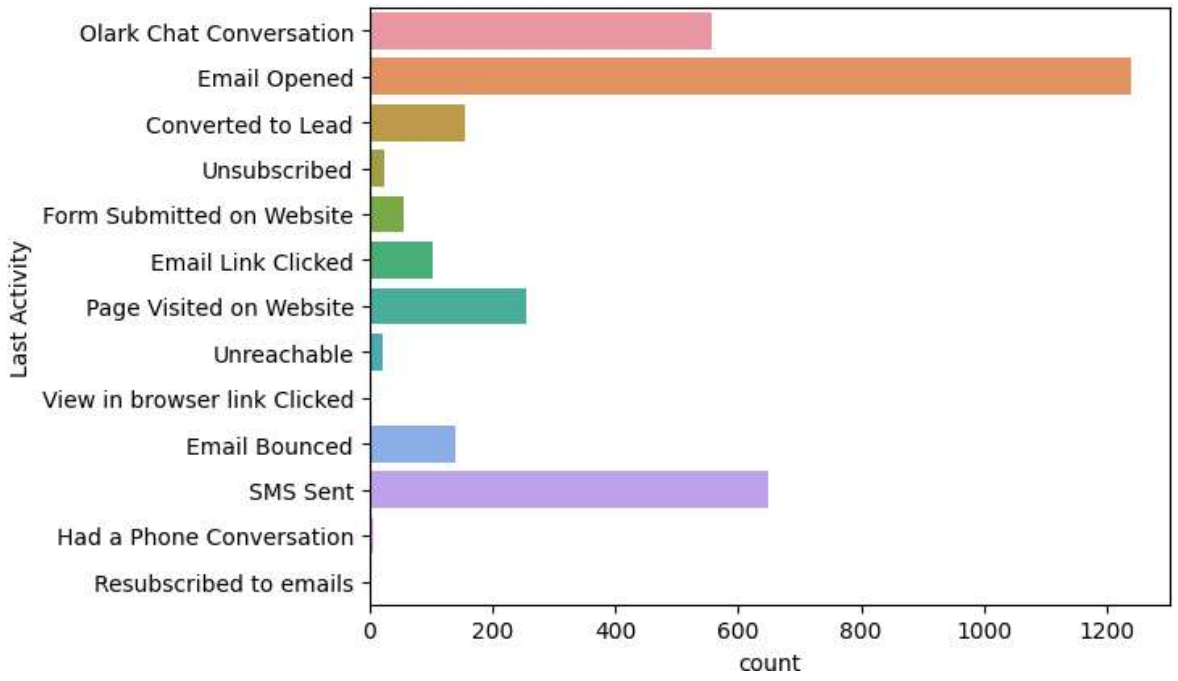


Inference : Lead Conversion is roughly 30% ie out of 2500 leads 800 are converted.

'Google' generated more inquiries. Management can spend more money on google ads compare to others Lead sources.

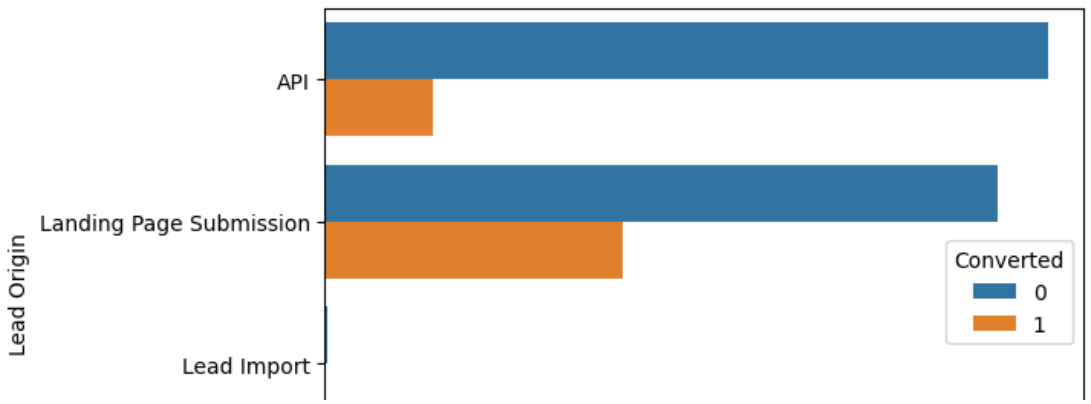
EDA – Univariate and Bivariate Analysis

countplot for i Last Activity

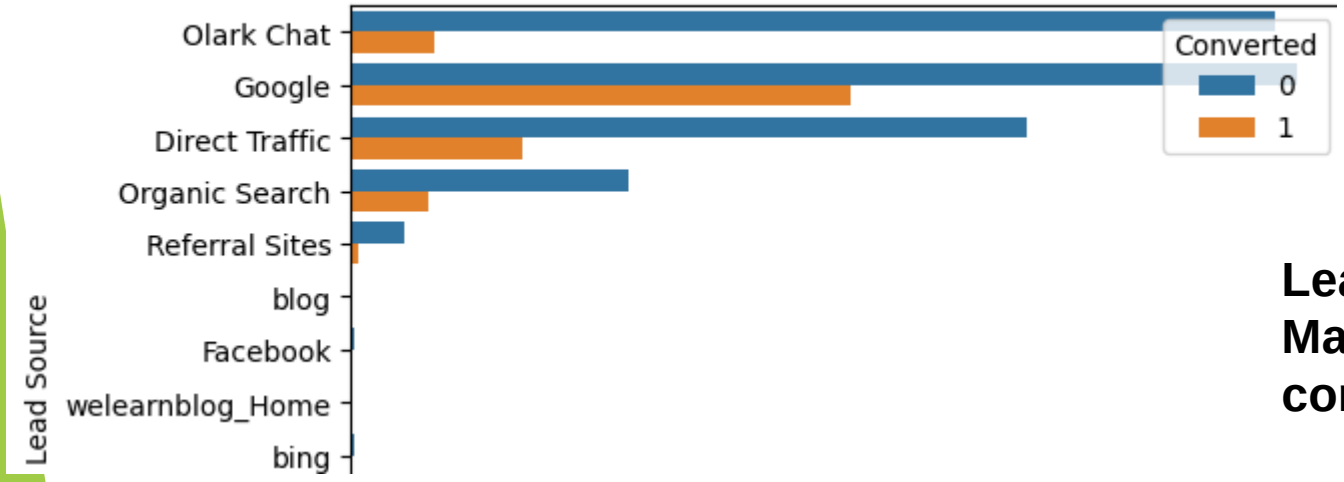


Inference : Most of Last Activity is Email Opened

EDA – Univariate and Bivariate Analysis



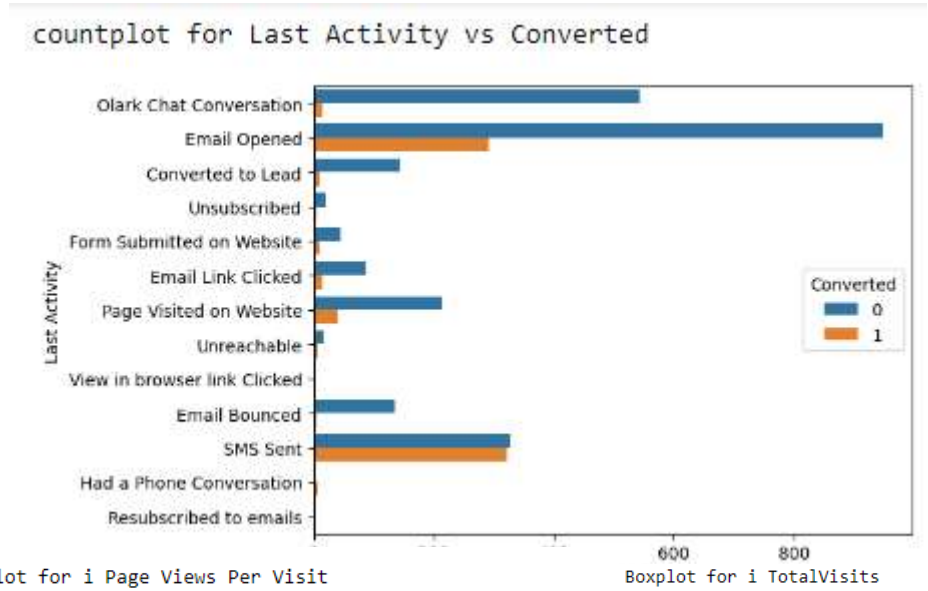
countplot for Lead Source vs Converted



Lead origin 'landing page submission ' worked well

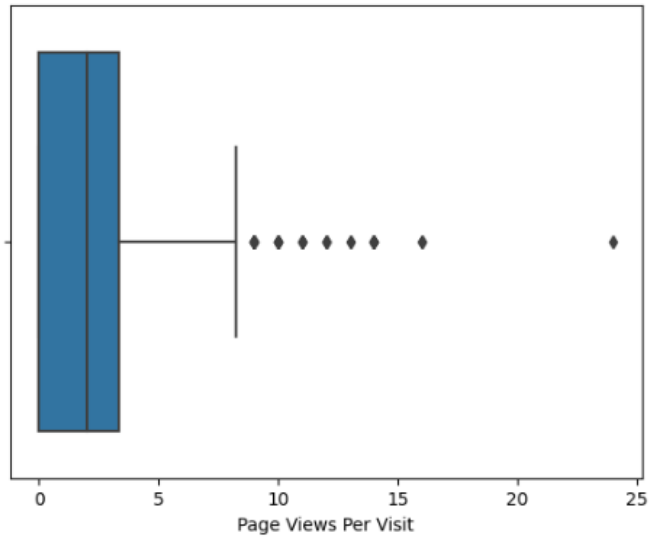
Lead source 'Google ‘ generated more conversions. Management can spend more money on google ads compare to al

EDA – Univariate and Bivariate Analysis

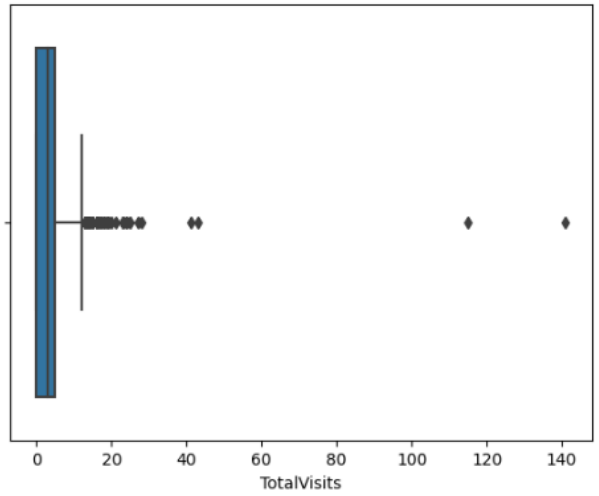


Inference : According to last activity ' Email Opened and sms sent' has been worked well

Boxplot for i Page Views Per Visit



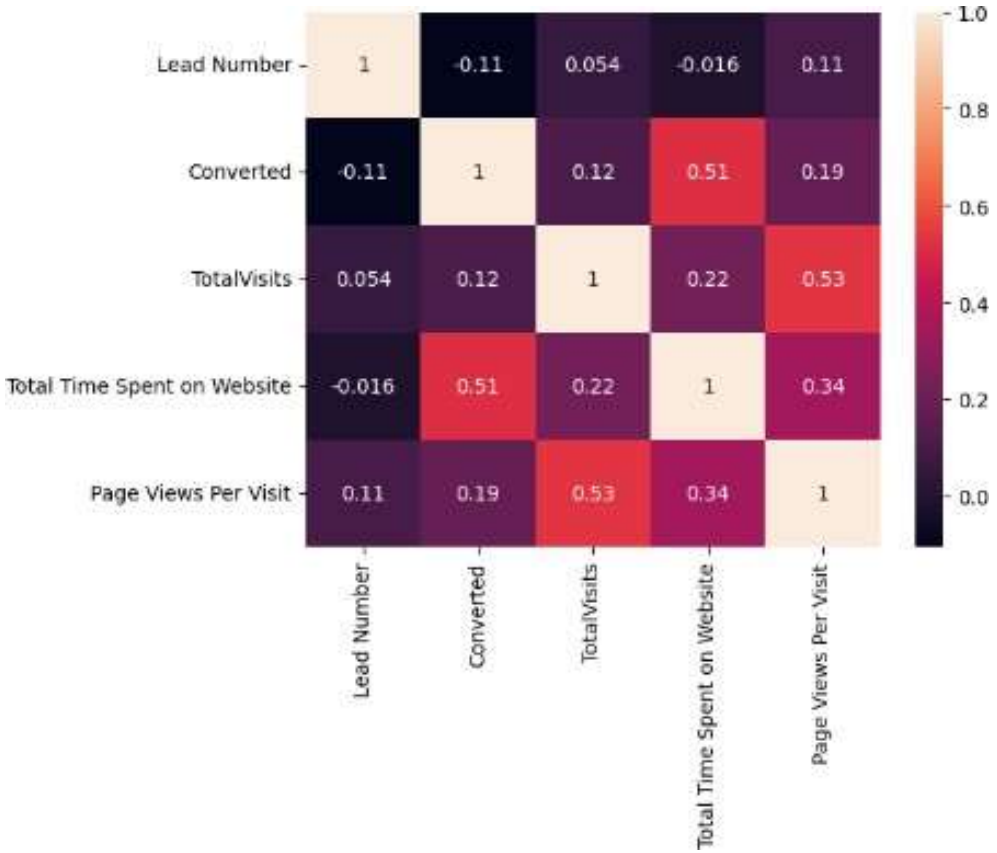
Boxplot for i TotalVisits



Inference : Box Plot for Total visits and Page views per Visit shows many Outliers.

Boxplot for i Total Time Spent on Website

EDA – Univariate and Bivariate Analysis



Inference : Heat Map shows correlation b/n

1) Total visits and Page views per visit.

2) Total time spent and Converted.

Data Cleaning / Pre-processing

1. Consider all values with '**Select**' as Null Values during missing value cleaning.
2. Drop columns **'Lead Quality','Asymmetrique Activity Index','Asymmetrique Profile Score','Asymmetrique Activity Score','Asymmetrique Profile Index','Tags'** where null values are more than 40%.
3. Drop **City** column where more than 40% is Mumbai and 28% is 'Select.
4. Drop column **Country** where more than 95% is India.
5. Similarly Drop columns **'Lead Profile' , 'How did you hear about X Education', 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque', 'What matters most to you in choosing a course'** where there data imbalance ie some unique category having maximum value.
6. Drop rows where **'What is your current occupation', 'TotalVisits', 'Page Views Per Visit', 'Last Activity'** are having null values.
7. Drop rows where **'Lead Source', 'Specialization'** are having null values.
8. Drop Prospect Id column which is not useful for Modelling.
9. Final check the percentage of rows retained in data cleaning process. **Ie 68.9%**

Data Preparation

1. Create **dummy** variables for all categorical variables with **drop_first = True**
   'Lead Origin', 'Lead Source',
   'Do Not Email', 'Last Activity',
   'What is your current occupation',
   'A free copy of Mastering The Interview',
   'Last Notable Activity'

2. Drop above **categorical** variables after creating **dummy** variables.

3. Perform **Test** and **Train** split with train 70% and test 30%.

4. Perform scaling using **MinMax** Scaler on numerical columns
   'TotalVisits',
   'Total Time Spent on Website',
   'Page Views Per Visit'

5. Split df by assigning **Converted** column to **y_train** and all other columns to **x_train**.

Data Modelling

1. Use RFE technique to perform variable selection to 15.

2. With this we got below 15 variables
   - **'TotalVisits', 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Reference', 'Lead Source_Welingak Website', 'Do Not Email_Yes', 'Last Activity_Had a Phone Conversation', 'Last Activity_SMS Sent', 'What is your current occupation_Housewife', 'What is your current occupation_Student', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional', 'Last Notable Activity_Had a Phone Conversation', 'Last Notable Activity_Unreachable'**

3. Build Regression model with sensitivity (recall).

4. Check P Value and VIF

5. Drop the columns with P Value > .05 and VIF < 5

6. Iterate the above steps (#4 and #5) till the model is stable and no further reduction in P Value and VIF Value is observed.

Final Model Features

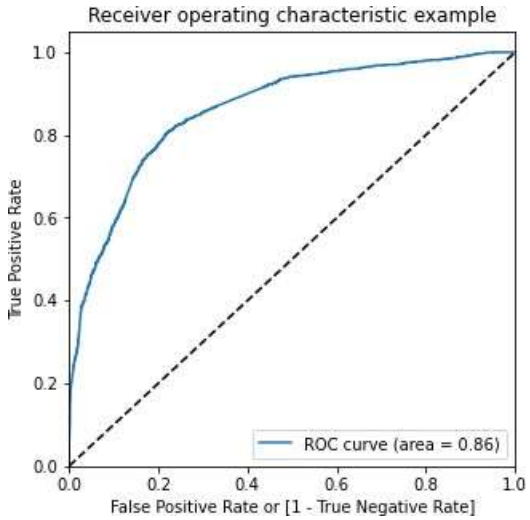
	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student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

Optimal probability cut off with ROC Curve

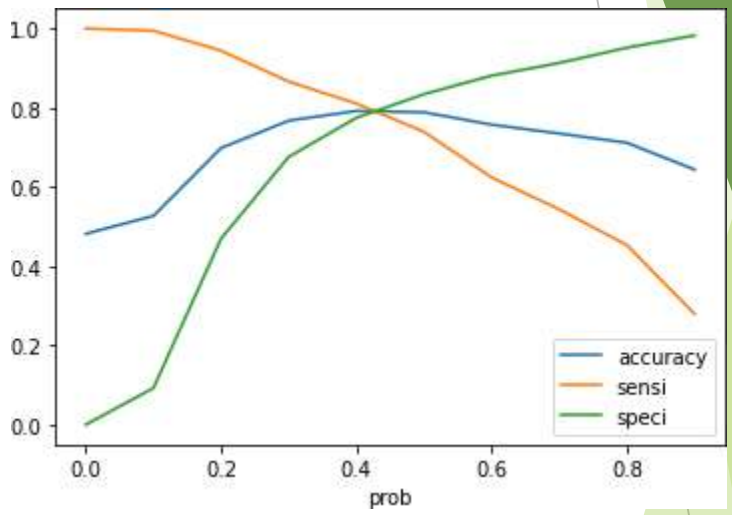
1. Calculate accuracy score
   - **78%**
2. Draw ROC Curve

3. Since we have higher **(0.86)** area under the ROC curve , therefore we can say our model is reasonably good.

4. Calculate the sensitivity
   - **73%**
5. Calculate Specificity
   - **83%**



Draw Plot for various probabilities to find optimal probability cut off -



From this plot optimum probability cut off is found to be **0.44**

Model Evaluation Test Model with test Data

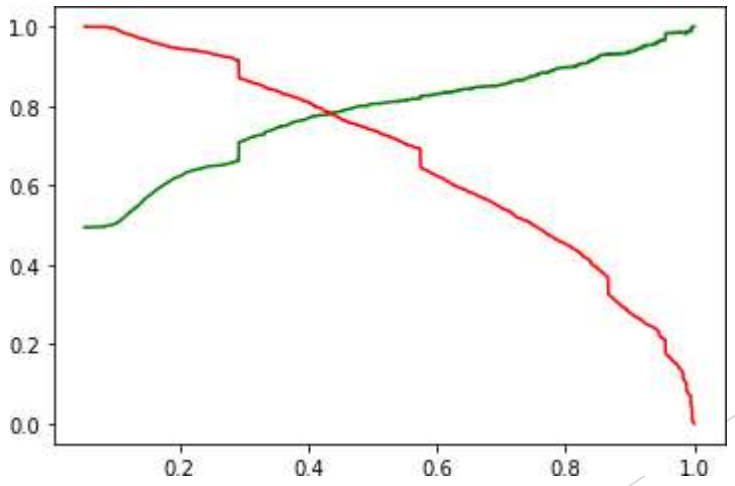
1. Calculate the overall accuracy
    - **78 %**

2. Calculate Sensitivity
    - **77 %**

3. Calculate Specificity
    - **78 %**

4. **Precision and recall trade off**

5. The graph shows trade off between the precision and recall.



Making Prediction with test set

1. Calculate overall accuracy
   - **78 %**
2. Calculate sensitivity of the model
   - **78 %**
3. Calculate specificity
   - **76 %**

Comparison values of train and test data

Train Data	**Test Data**
1. Overall accuracy • **78 %**	1. Overall accuracy • **78 %**
2. Sensitivity • **77 %**	2. Sensitivity • **77 %**
3. Specificity • **78 %**	3. Specificity • **76 %**

Hot Leads

1. Assign Lead Score to the test data.

2. We have achieved our goal of lead conversion rate around 77% . The Model seems to predict the Conversion Rate well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate.

3. Find out the leads which should be contacted.

4. The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 80%.

5. There are 458 leads which can be contacted and have a high chance of getting converted. The Prospect ID of the customers to be contacted are as below.

Final Hot_Leads

```
hot_leads
```

	Converted	Conversion_Prob	final_predicted	Lead_Score
0	1	0.996296	1	100
10	1	0.987981	1	99
14	1	0.876810	1	88
17	1	0.935454	1	94
20	1	0.979392	1	98
...	...	...	...	...
1889	1	0.875543	1	88
1904	1	0.920263	1	92
1905	1	0.973954	1	97
1906	1	0.865844	1	87
1909	0	0.799951	1	80

458 rows × 4 columns

Conclusions, Results, Final Recommendations

Important Variables in final model

TotalVisits	**11.14891**
Total Time Spent on Website	4.422291
Lead Origin_Lead Add Form	4.205123
Last Notable Activity_Unreachable	2.784594
Last Activity_Had a Phone Conversation	2.75522
Lead Source_Welingak Website	2.152559
Lead Source_Olark Chat	1.452589
Last Activity_SMS Sent	1.185594
const	0.204037
Do Not Email_Yes	-1.50368
What is your current occupation_Student	-2.35778
What is your current occupation_Unemployed	-2.54446

Recommendations based on final Model

Dos

- Call people where **Total Time Spent on Website is high**.
- Call people where **Lead Origin is Lead Add Form**
- Call people where **Last Activity is Had a Phone Conversation**
- Call people where **Lead Source is Welingak Website**
- Call people where **Lead Source is Olark Chat**
- Call people where **Last Activity is SMS Sent**

Donts

- Do not call people where **Do Not Email is Yes**
- Do not call people where **What is your current occupation is Student** unless they are really serious.
- Do not call people where **What is your current occupation is Unemployed**