Lead Scoring Case study

Group members :

Meghanshu Bhatia

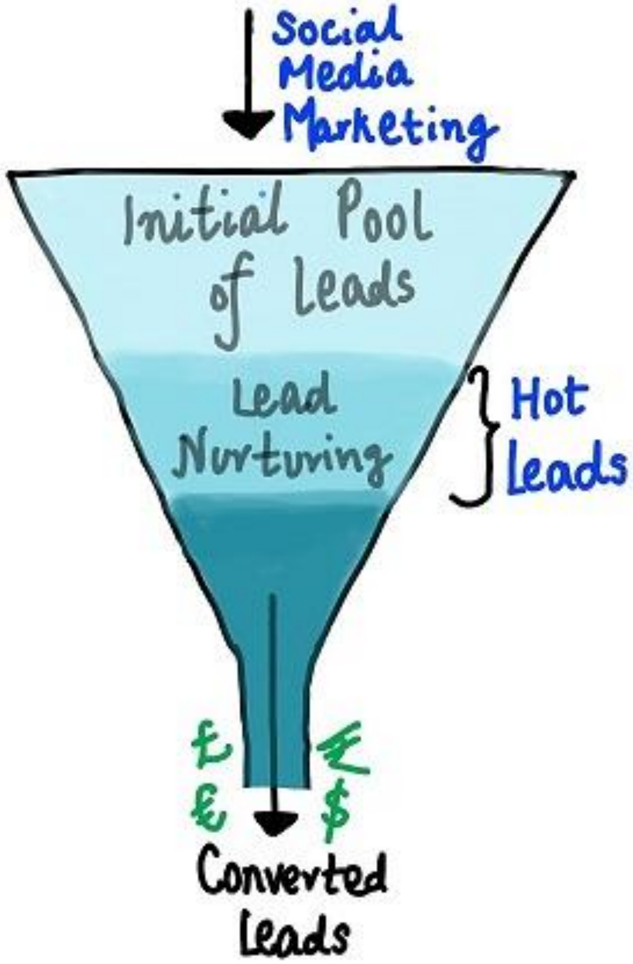
Megha Pankhuri

Introduction

- Educational company named Education X sells educational online courses to professionals.
- Company advertises on several websites and search engines.
- Interested people land on the company’s website and give their information by filling some forms.
- Individuals providing such information are classified as “**Leads**” by Education X.
- Company then contacts such individuals through calls and emails.
- This contact converts some leads to paying customers but this conversion rate is very poor ~ 30%

Problem statement

- Education X wants a model which can efficiently identify the potential leads by assigning conversion score to them.
- These leads which are likely to convert are known as “Hot leads” and the sales team will avidly communicate with these Hot leads only, instead of everyone.
- Target leads conversion score expected from Education X ≈ 80%



Approach towards solution

1) Checking data sanity
2) Data preparation and cleaning
3) Data visualization through EDA
4) Dummy creation
5) Scaling of variables
6) Feature selection using RFE
7) Model optimization.
8) Detecting specificity, sensitivity and accuracy of the model.
9) Finding optimal probability cut off point
10) Testing the model

Data Preparation and cleaning

a) Checking the rows and columns data type using ‘Describe’ and ‘info’ function.

b) Several columns having entries named ‘Select’ were replaced by ‘Null’.

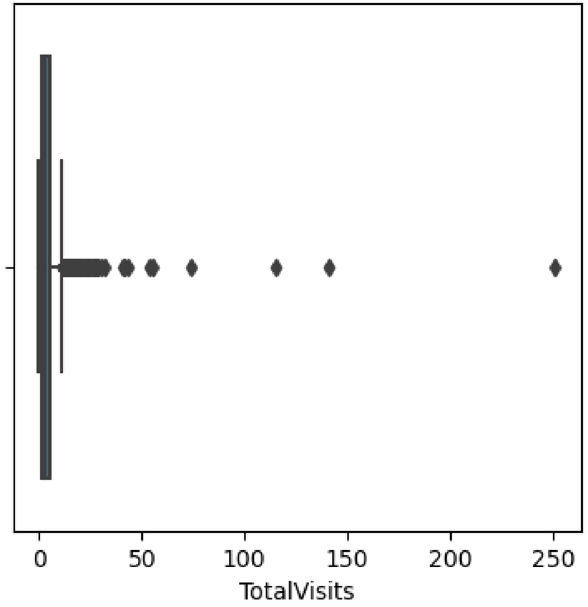
c) Missing values treatment:

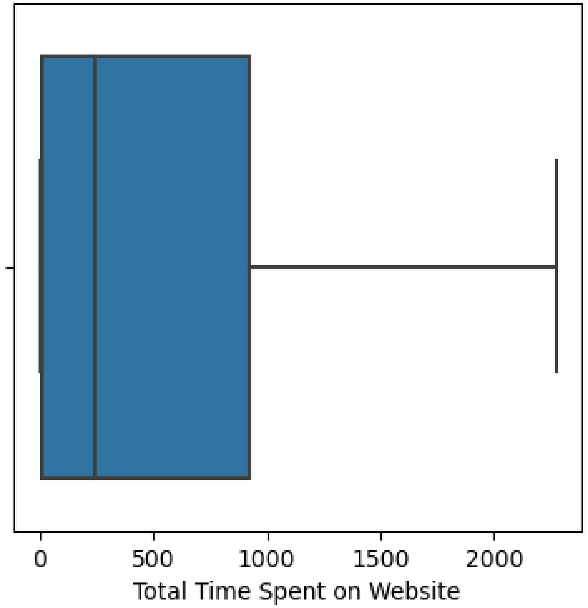
   - In columns: Columns having missing values more than 25% will be dropped for better model.
   - In rows: Rows having “Total visits” and “Lead score” column as null will be dropped since they don’t hold any importance for the model. A total of 1.8% of rows were dropped.

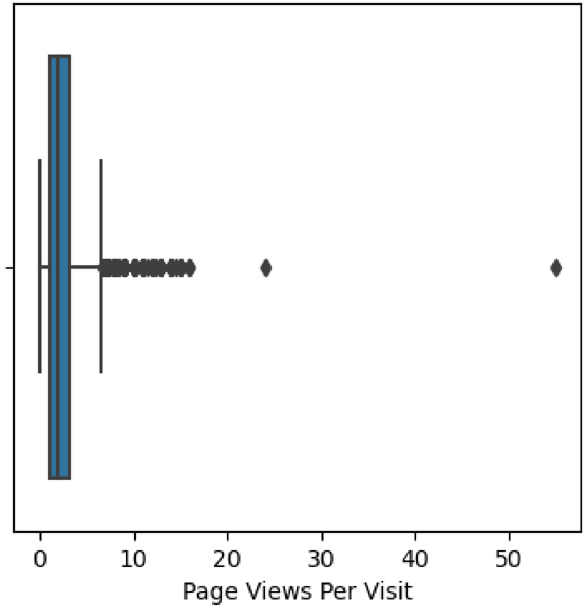
d) Removal of columns that have a very low count of conversion as ‘yes’. These won’t help in training the model as per our requirement.

Outlier treatment

Box plot was used to detect outliers in customers visits of site







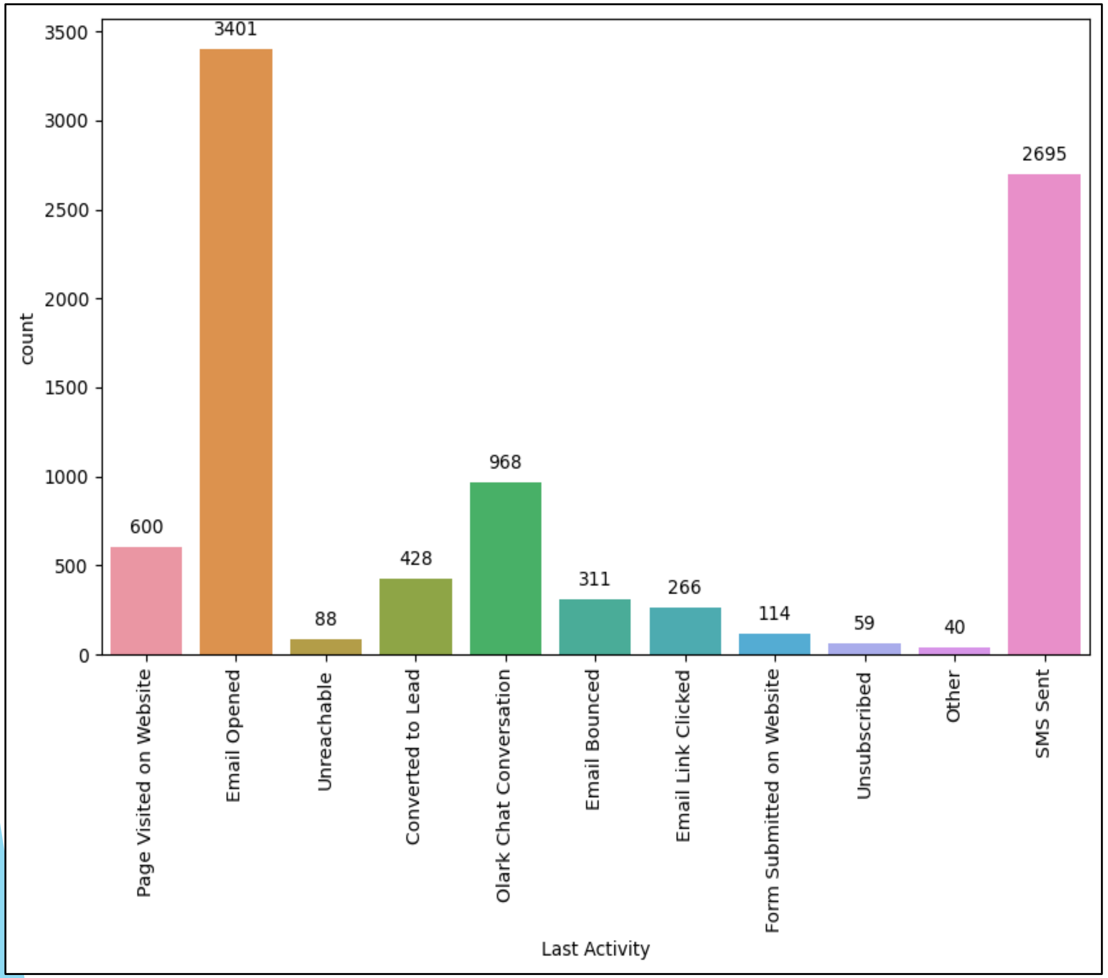
Few outliers in Total Visit column

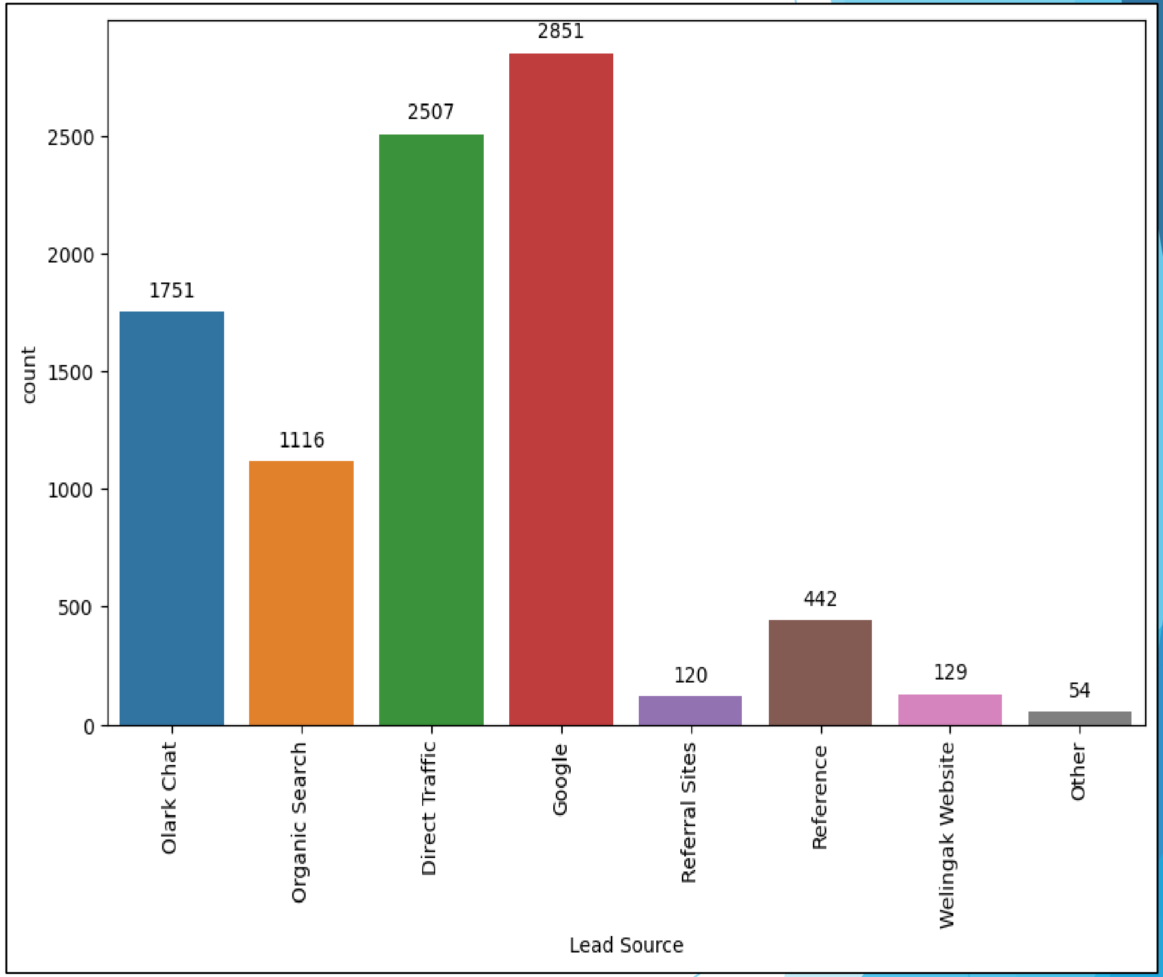
No outliers in Total time spend on website column

Few outliers in Page views per Visit column

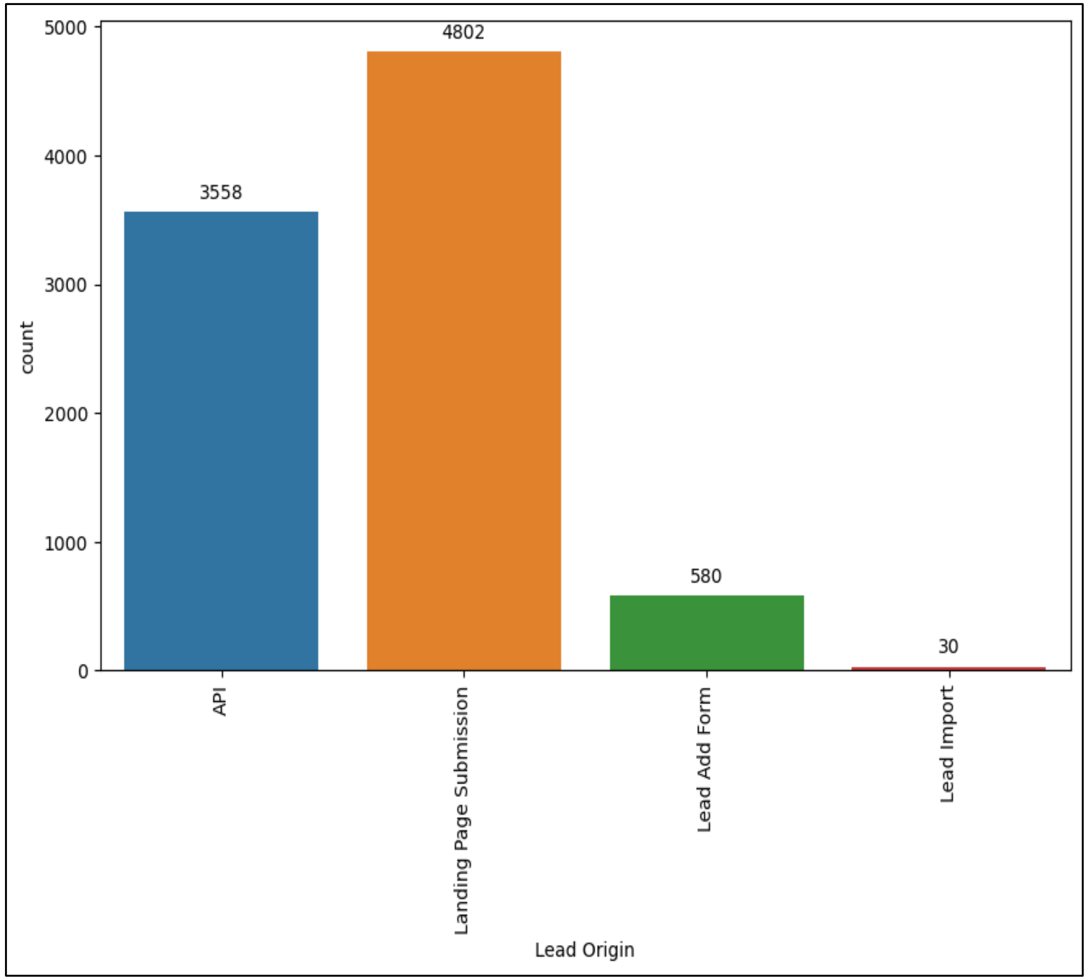
Outlier treatment was done for 10th and 90th percentile

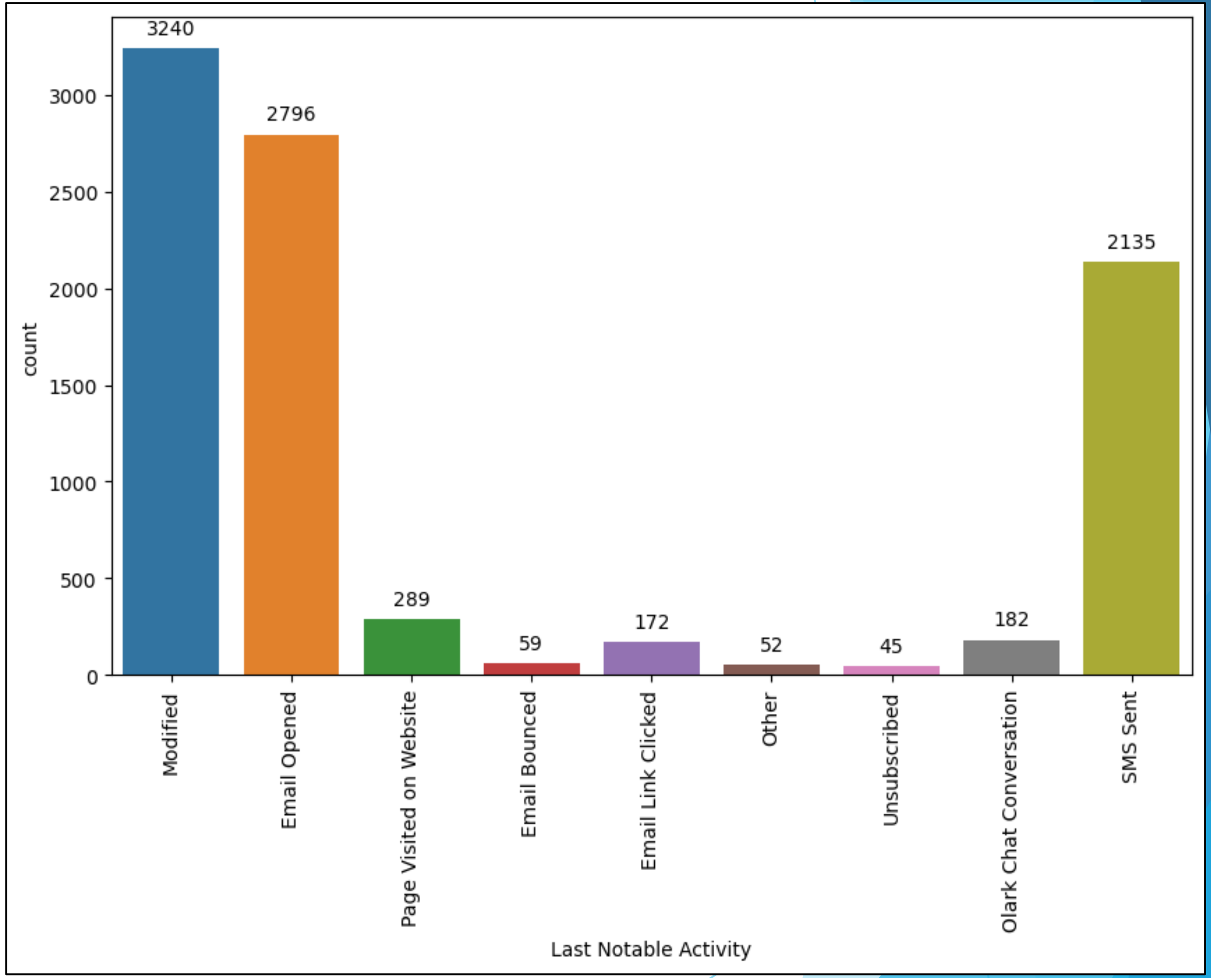
Data Visualization – Lead details



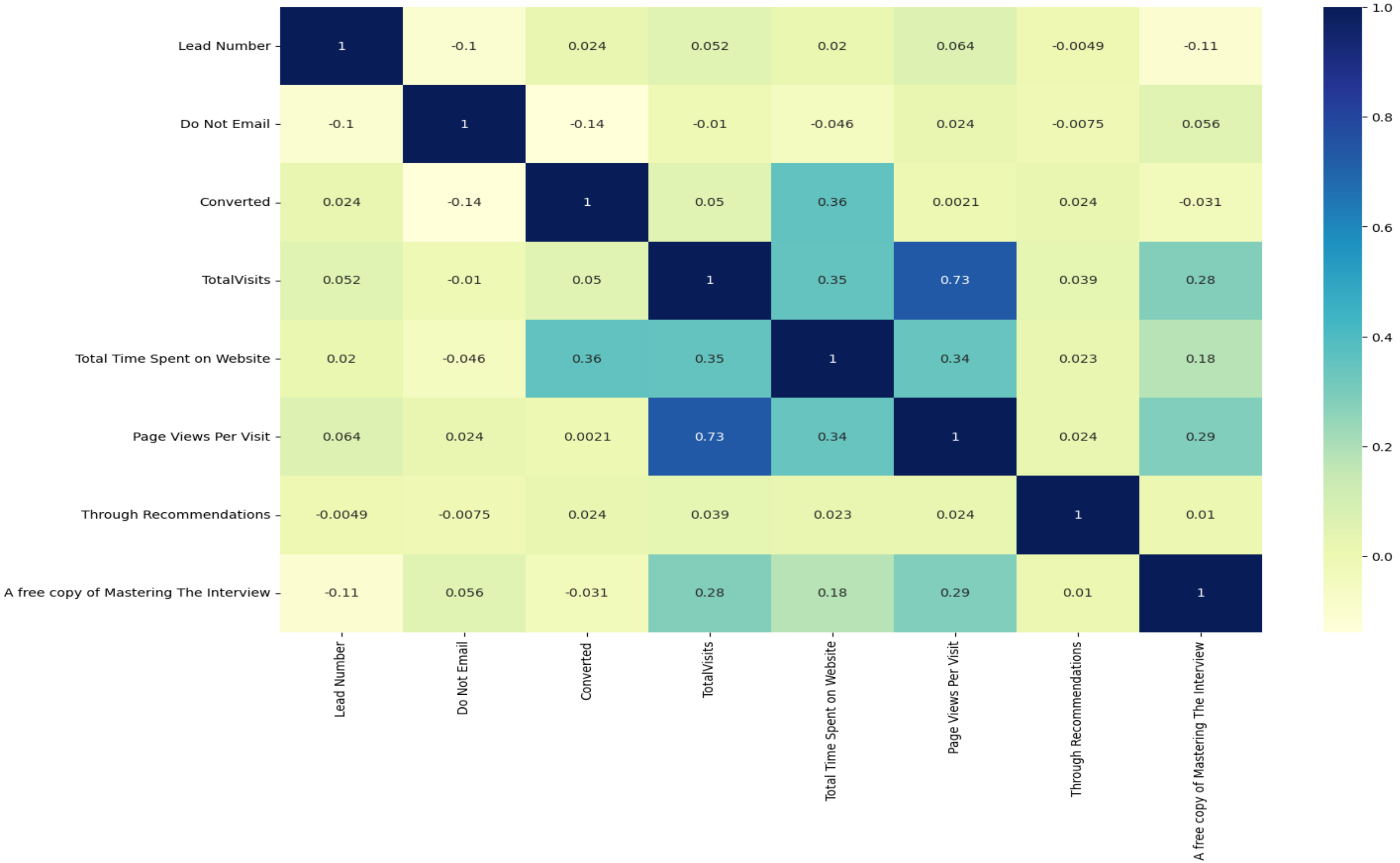


Data Visualization - Lead details contd..





Data Visualization – Correlation



	Lead Number	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Through Recommendations	A free copy of Mastering The Interview
Lead Number	1	-0.1	0.024	0.052	0.02	0.064	-0.0049	-0.11
Do Not Email	-0.1	1	-0.14	-0.01	-0.046	0.024	-0.0075	0.056
Converted	0.024	-0.14	1	0.05	0.36	0.0021	0.024	-0.031
TotalVisits	0.052	-0.01	0.05	1	0.35	0.73	0.039	0.28
Total Time Spent on Website	0.02	-0.046	0.36	0.35	1	0.34	0.023	0.18
Page Views Per Visit	0.064	0.024	0.0021	0.73	0.34	1	0.024	0.29
Through Recommendations	-0.0049	-0.0075	0.024	0.039	0.023	0.024	1	0.01
A free copy of Mastering The Interview	-0.11	0.056	-0.031	0.28	0.18	0.29	0.01	1

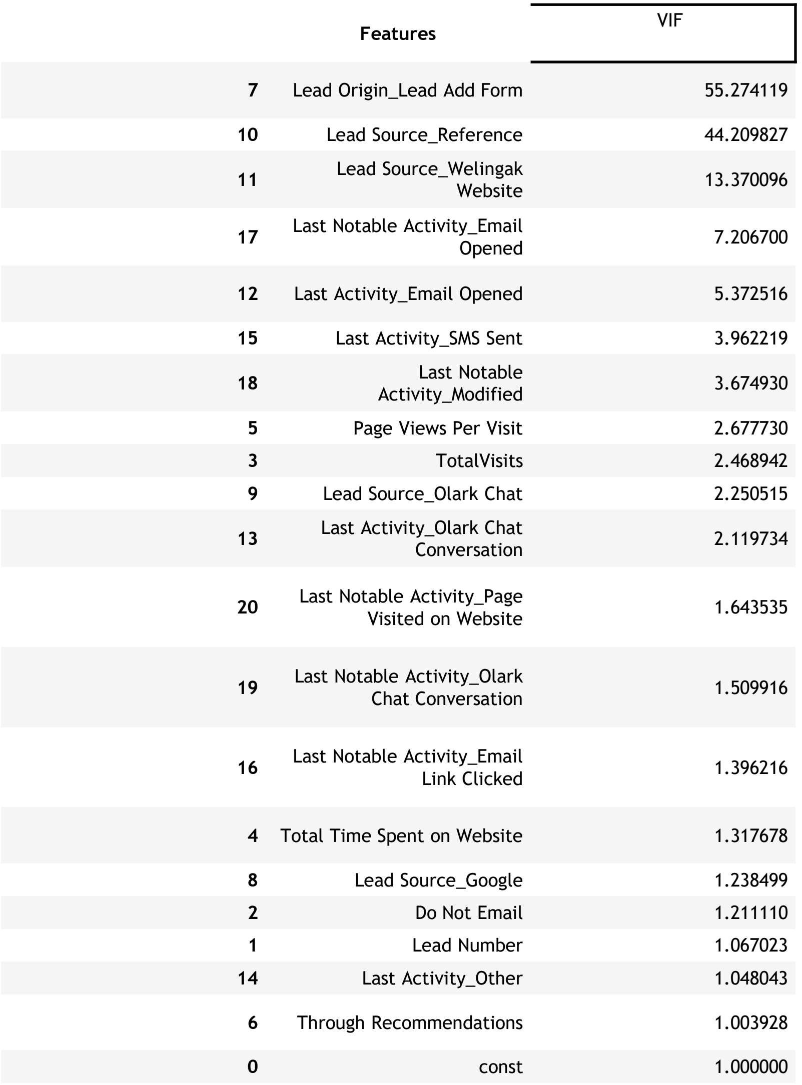
Pre-processing of data for Model building

This includes following steps for building efficient model:

1. ***Dummy creation*** for all the categorical variables, having >2 unique values.
2. Concatenation of dummy columns in the main dataframe.
3. Checking the ***correlation*** between variables through heatmap.
4. Scaling of variables using Standard scaler.
5. ***Selection of Feature*** through RFE to find and keep only the features relevant for training and testing sets.
6. Further refining of features on the basis of P-values and VIF.

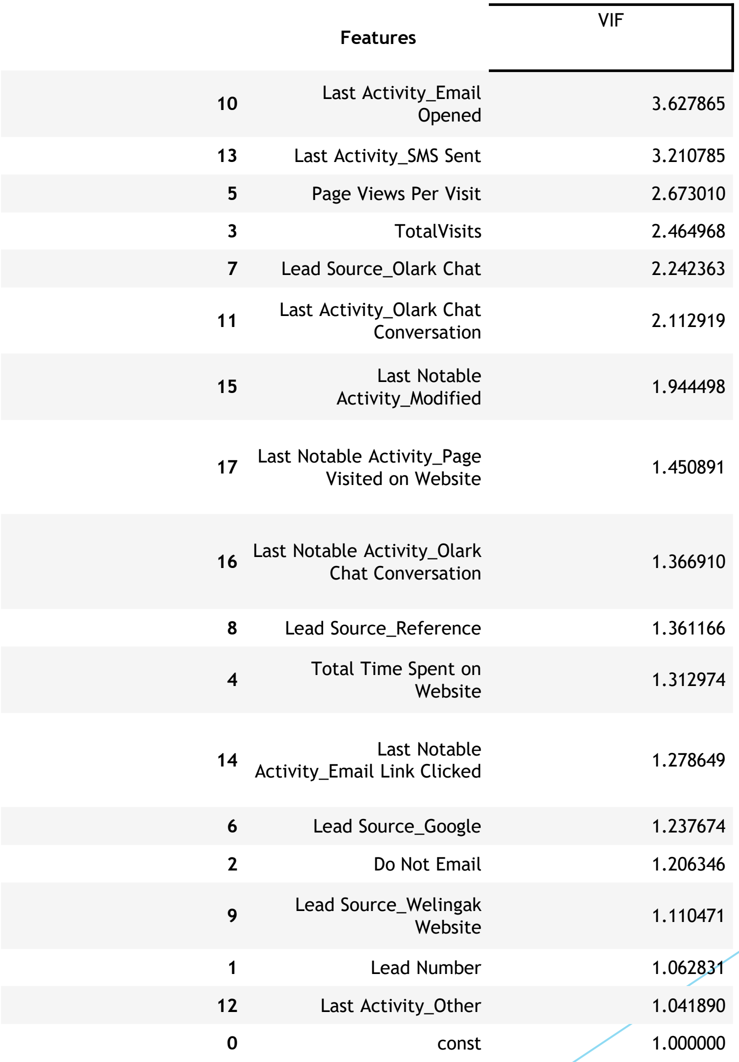
Model building and optimization

VIF of 1st model



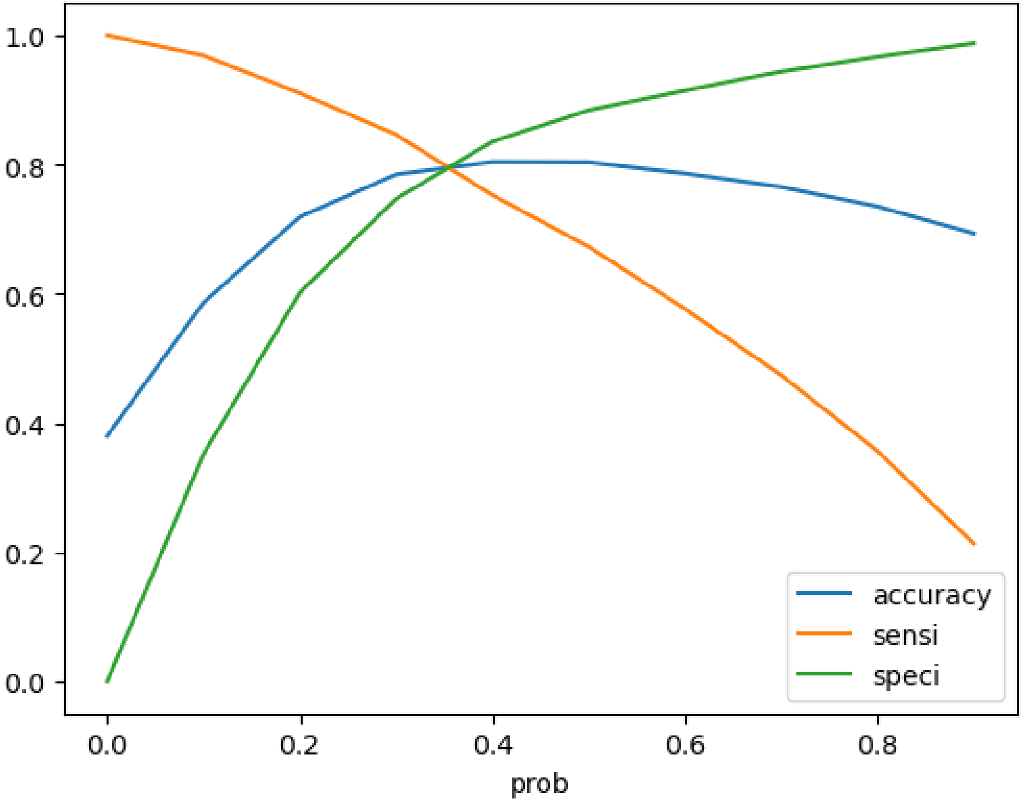
	Features	VIF
7	Lead Origin_Lead Add Form	55.274119
10	Lead Source_Reference	44.209827
11	Lead Source_Welingak Website	13.370096
17	Last Notable Activity_Email Opened	7.206700
12	Last Activity_Email Opened	5.372516
15	Last Activity_SMS Sent	3.962219
18	Last Notable Activity_Modified	3.674930
5	Page Views Per Visit	2.677730
3	TotalVisits	2.468942
9	Lead Source_Olark Chat	2.250515
13	Last Activity_Olark Chat Conversation	2.119734
20	Last Notable Activity_Page Visited on Website	1.643535
19	Last Notable Activity_Olark Chat Conversation	1.509916
16	Last Notable Activity_Email Link Clicked	1.396216
4	Total Time Spent on Website	1.317678
8	Lead Source_Google	1.238499
2	Do Not Email	1.211110
1	Lead Number	1.067023
14	Last Activity_Other	1.048043
6	Through Recommendations	1.003928
0	const	1.000000

VIF of final model after manual feature elimination



	Features	VIF
10	Last Activity_Email Opened	3.627865
13	Last Activity_SMS Sent	3.210785
5	Page Views Per Visit	2.673010
3	TotalVisits	2.464968
7	Lead Source_Olark Chat	2.242363
11	Last Activity_Olark Chat Conversation	2.112919
15	Last Notable Activity_Modified	1.944498
17	Last Notable Activity_Page Visited on Website	1.450891
16	Last Notable Activity_Olark Chat Conversation	1.366910
8	Lead Source_Reference	1.361166
4	Total Time Spent on Website	1.312974
14	Last Notable Activity_Email Link Clicked	1.278649
6	Lead Source_Google	1.237674
2	Do Not Email	1.206346
9	Lead Source_Welingak Website	1.110471
1	Lead Number	1.062831
12	Last Activity_Other	1.041890
0	const	1.000000

Optimal probability cut- off point



Analysis of the above curve:

Accuracy - Becomes stable after 0.35

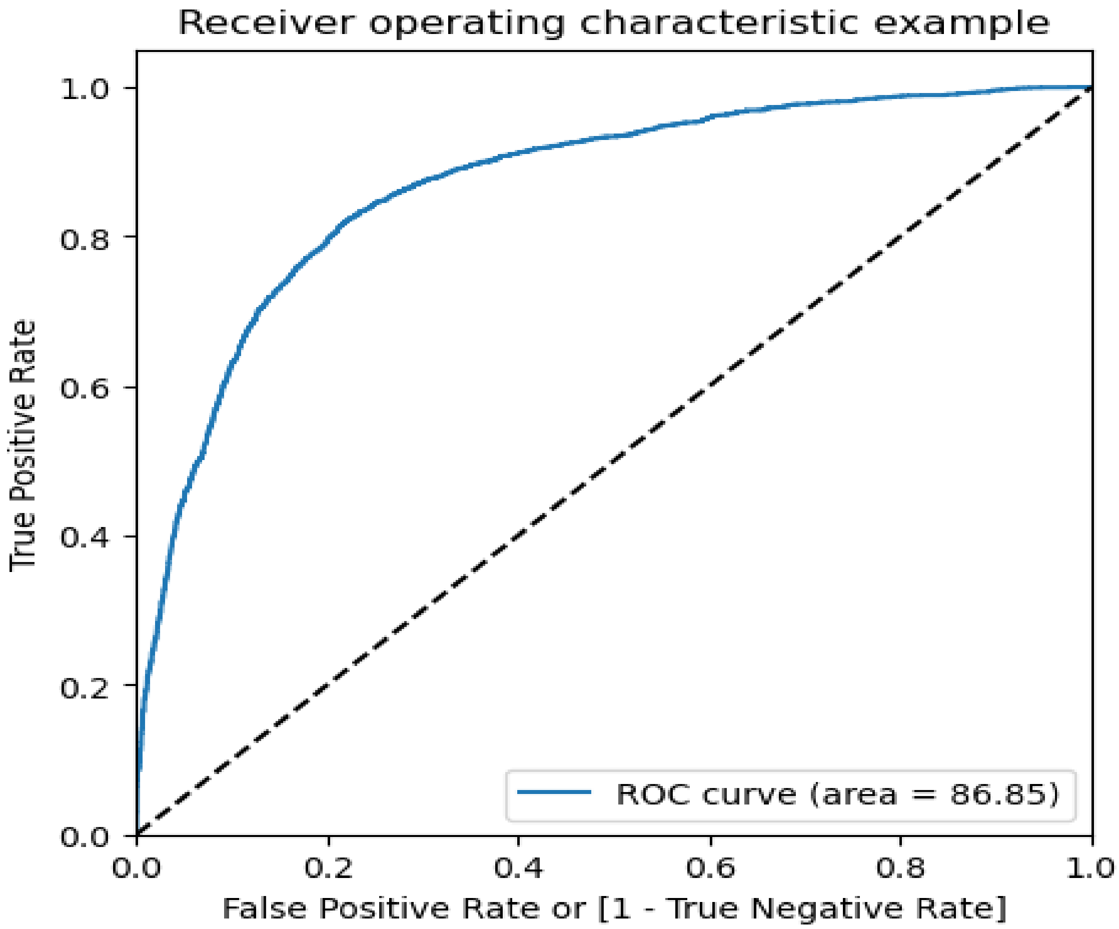
Sensitivity - Decreases with the increased probability.

Specificity - Increases with the increasing probability.

At point 0.35 where three parameters cut each other, we can see that there is a balance between sensitivity and specificity with a good accuracy.

So, 0.35 may be the optimal cutoff point.

ROC curve analysis



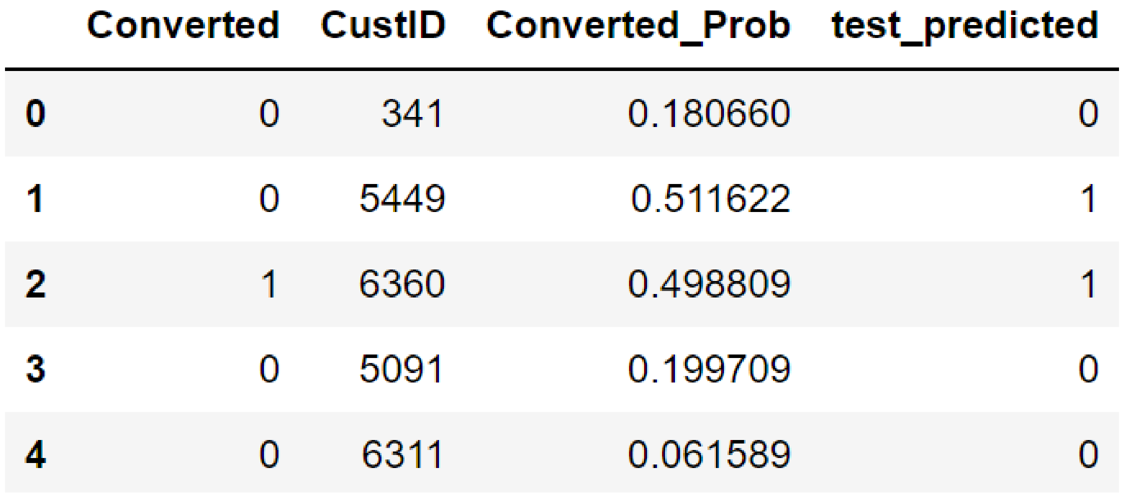
Gini of the model

We can see from the ROC curve, that the area of the curve is 86.85 %, which is the Gini of the model.

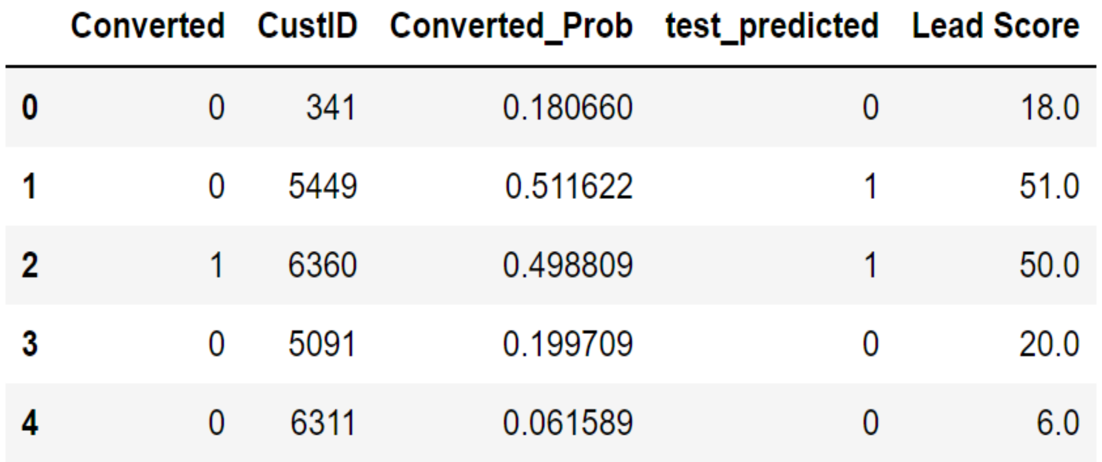
The curve is hugging the true positive rate axis.

Model Evaluation & Conclusion

In the test set, using 0.35 probability cut-off



	Converted	CustID	Converted_Prob	test_predicted
0	0	341	0.180660	0
1	0	5449	0.511622	1
2	1	6360	0.498809	1
3	0	5091	0.199709	0
4	0	6311	0.061589	0



	Converted	CustID	Converted_Prob	test_predicted	Lead Score
0	0	341	0.180660	0	18.0
1	0	5449	0.511622	1	51.0
2	1	6360	0.498809	1	50.0
3	0	5091	0.199709	0	20.0
4	0	6311	0.061589	0	6.0

Assigning lead score

- **Conclusion: Train vs test**
- Train set
  - Accuracy = 79.77 %
  - Sensitivity = 79.31 %
  - Specificity = 80.04 %
- Test set
  - Accuracy = 79.60 %
  - Sensitivity = 80.90 %
  - Specificity = 78.83 %