Lead Score Case Study

BY SAI YASWANTH S

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 in search engines like Google.
- Once the 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 and the company also gets leads through past referrals.
- After the 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%.

Business Goal

- 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.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

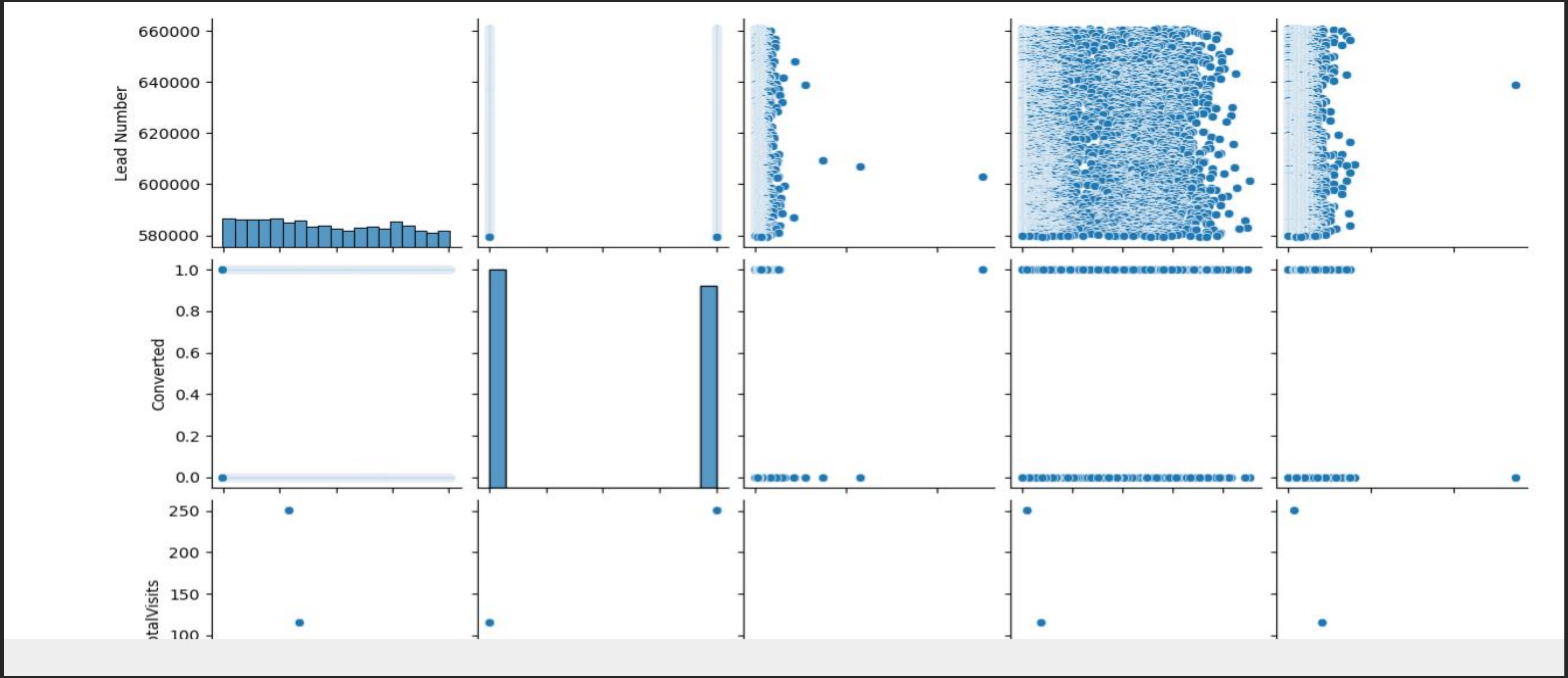
Steps

- Reading and Understanding of Data
- Data Cleaning
- Exploratory Data Analysis
- Feature Scaling
- Splitting the Train and Test Data
- Build a Model
- Evaluating the Model – Sensitivity, Specificity, Precision, Recall.
- Predictions on Test Data

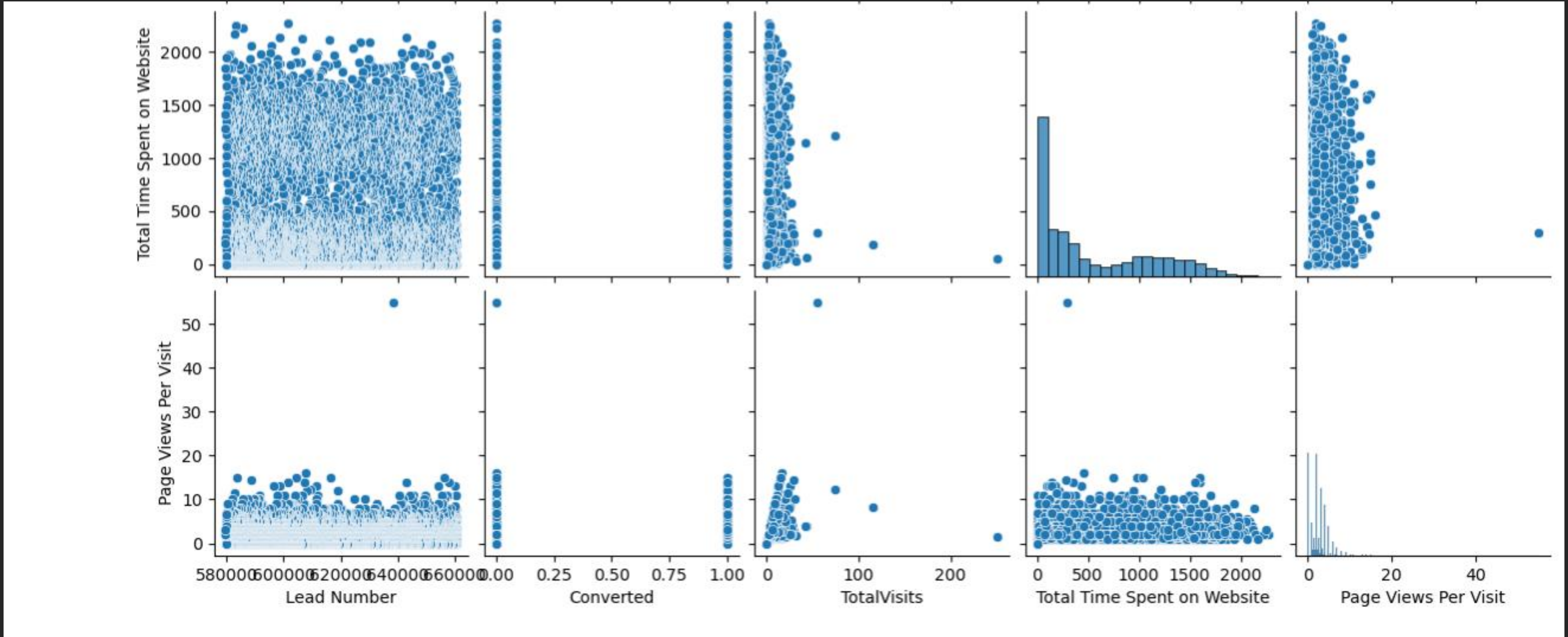
Reading, Understand and Cleaning the Data

- Read the data using Pandas and Numpy libraries.
- User shape, info and describe functions to understand.
- We got 9240 rows and 37 columns of data.
- All the data is in Object/float/int data types.
- Using isnull().sum() found the total number of null values in each column.
- Dropped the columns with more than 3000 null values.
- There are columns with 'Select' as a level – This might be the student had not filled anything in those columns.
- After handling null values, we got ~69% of the data to deal with.

Visualization of Data



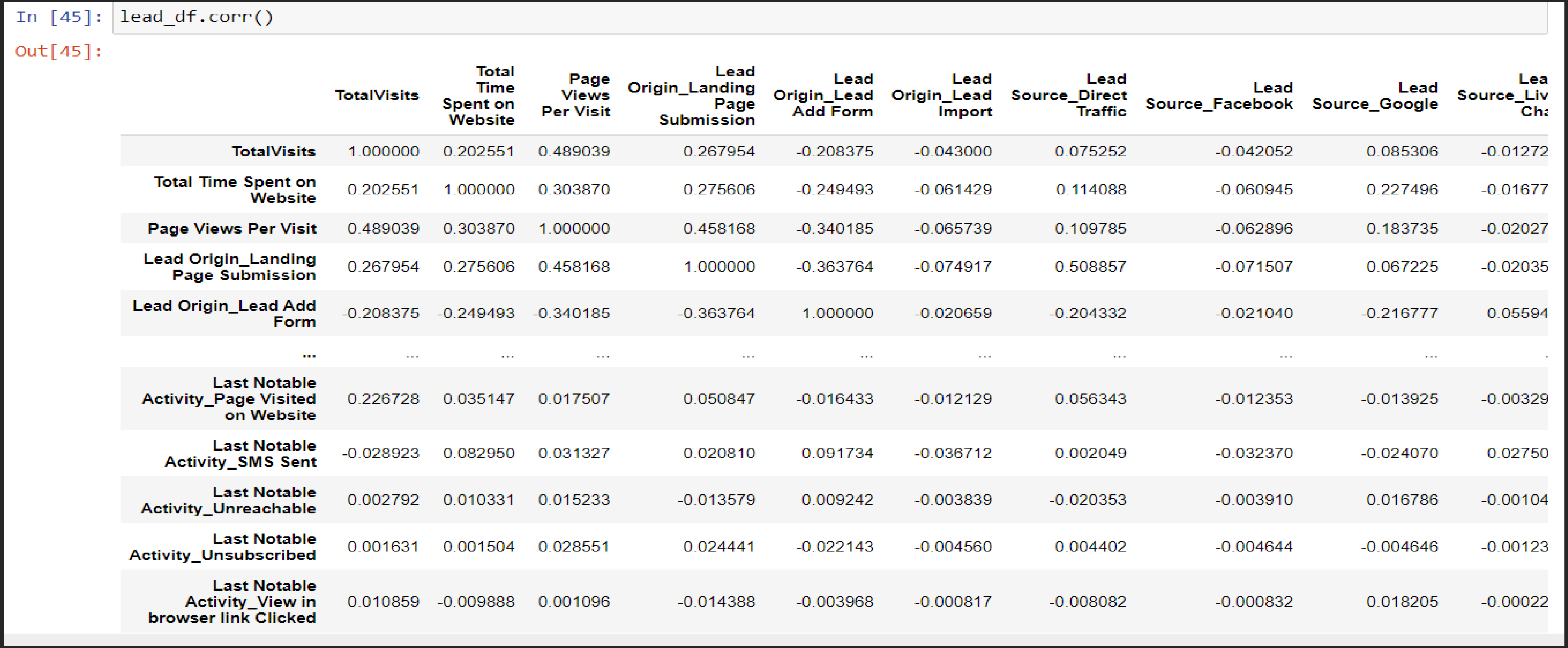
Visualization of Data Cont...



Splitting Train and Test Sets and Scaling

- Splitting the data in Train and Test data sets using
- Assign Converted as Target variable.
- Assign all independent variables to train data.
- Scale the numerical data using MinMaxScaler().
- Check the correlation of the data after scaling.

Correlation



In [45]:
```
lead_df.corr()
```

Out[45]:

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lea Source_Liv Cha
TotalVisits	1.000000	0.202551	0.489039	0.267954	-0.208375	-0.043000	0.075252	-0.042052	0.085306	-0.01272
Total Time Spent on Website	0.202551	1.000000	0.303870	0.275606	-0.249493	-0.061429	0.114088	-0.060945	0.227496	-0.01677
Page Views Per Visit	0.489039	0.303870	1.000000	0.458168	-0.340185	-0.065739	0.109785	-0.062896	0.183735	-0.02027
Lead Origin_Landing Page Submission	0.267954	0.275606	0.458168	1.000000	-0.363764	-0.074917	0.508857	-0.071507	0.067225	-0.02035
Lead Origin_Lead Add Form	-0.208375	-0.249493	-0.340185	-0.363764	1.000000	-0.020659	-0.204332	-0.021040	-0.216777	0.05594
...	...	...	...	...	...	...	...	...	...	.
Last Notable Activity_Page Visited on Website	0.226728	0.035147	0.017507	0.050847	-0.016433	-0.012129	0.056343	-0.012353	-0.013925	-0.00329
Last Notable Activity_SMS Sent	-0.028923	0.082950	0.031327	0.020810	0.091734	-0.036712	0.002049	-0.032370	-0.024070	0.02750
Last Notable Activity_Unreachable	0.002792	0.010331	0.015233	-0.013579	0.009242	-0.003839	-0.020353	-0.003910	0.016786	-0.00104
Last Notable Activity_Unsubscribed	0.001631	0.001504	0.028551	0.024441	-0.022143	-0.004560	0.004402	-0.004644	-0.004646	-0.00123
Last Notable Activity_View in browser link Clicked	0.010859	-0.009888	0.001096	-0.014388	-0.003968	-0.000817	-0.008082	-0.000832	0.018205	-0.00022

Feature Variables

- Total Visit
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Welingak Website
- Do not Email_Yes
- Last Activity_Had a Phone Conversation
- Last Activity_SMS Sent
- What is your Current Occupation_Student
- What is your Current Occupation_Unemployed
- Last Notable Activity_Unreachable

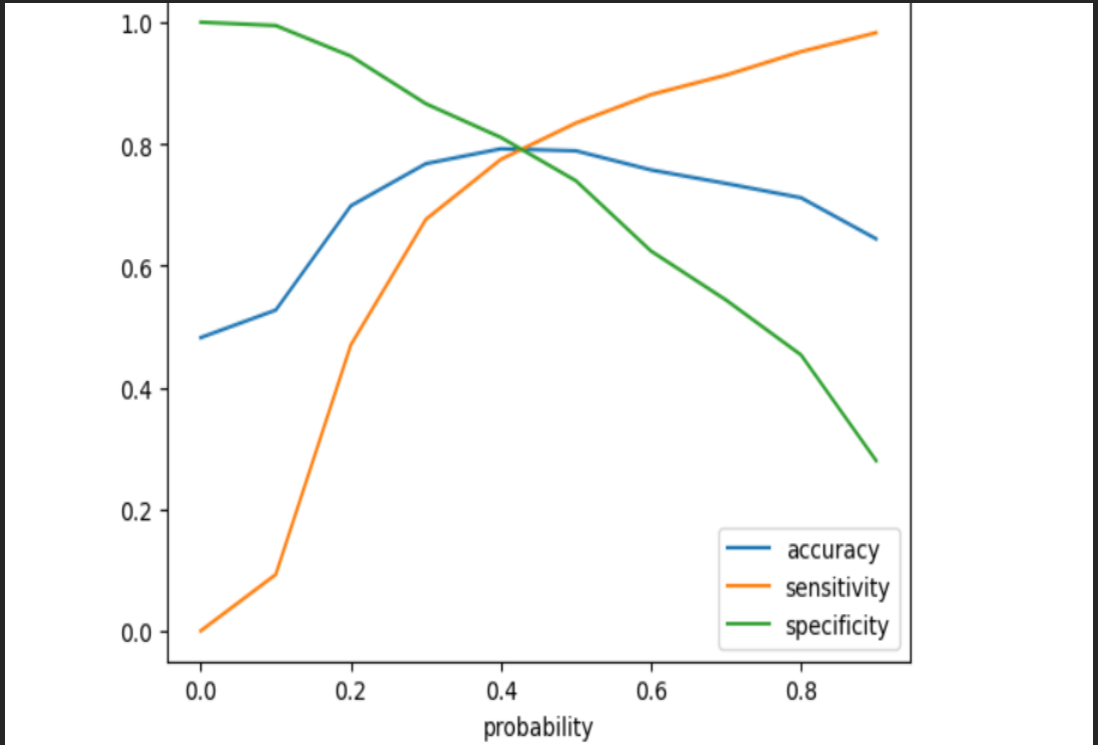
Model Evaluation

- Confusion Matrix :

  [[1815 497]

  [431 1718]]

  - Accuracy: 79%
  - Sensitivity: 80%
  - Specificity: 79%
  - Precision: 77%
  - Recall: 79%



Evaluation On Test Data

- Confusion Matrix:

[[1852, 460],

[479, 1670]]

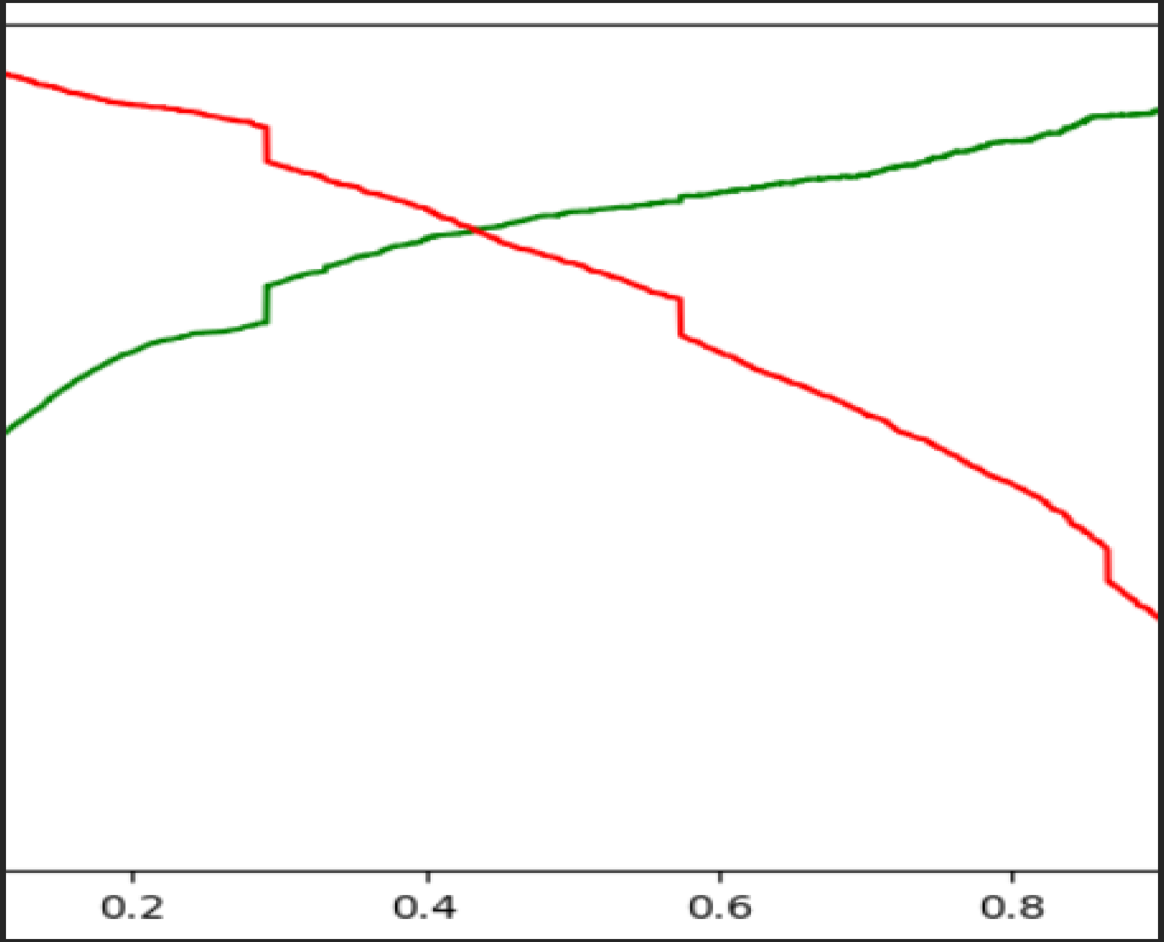
Accuracy: 79%

Sensitivity: 78%

Specificity: 80%

Precision: 78%

Recall: 77%



Conclusion

- We have calculated the Accuracy, Sensitivity and Specificity, Precision and Recall for the Train and Test Data.
- The above metrics Accuracy (79%), Sensitivity(80%) and Specificity(79%) for Train data are almost equal to those metrics on Test data.
- The following are top three variables:

  TotalVisits

  Total Time Spent on Website

  Lead Origin_Lead Add Form
- The overall model is good.

Thank You