LEAD SCORING ASSIGNMENT

PRESENTED BY

ABHINAYA B

RAHUL BHARGAVA V

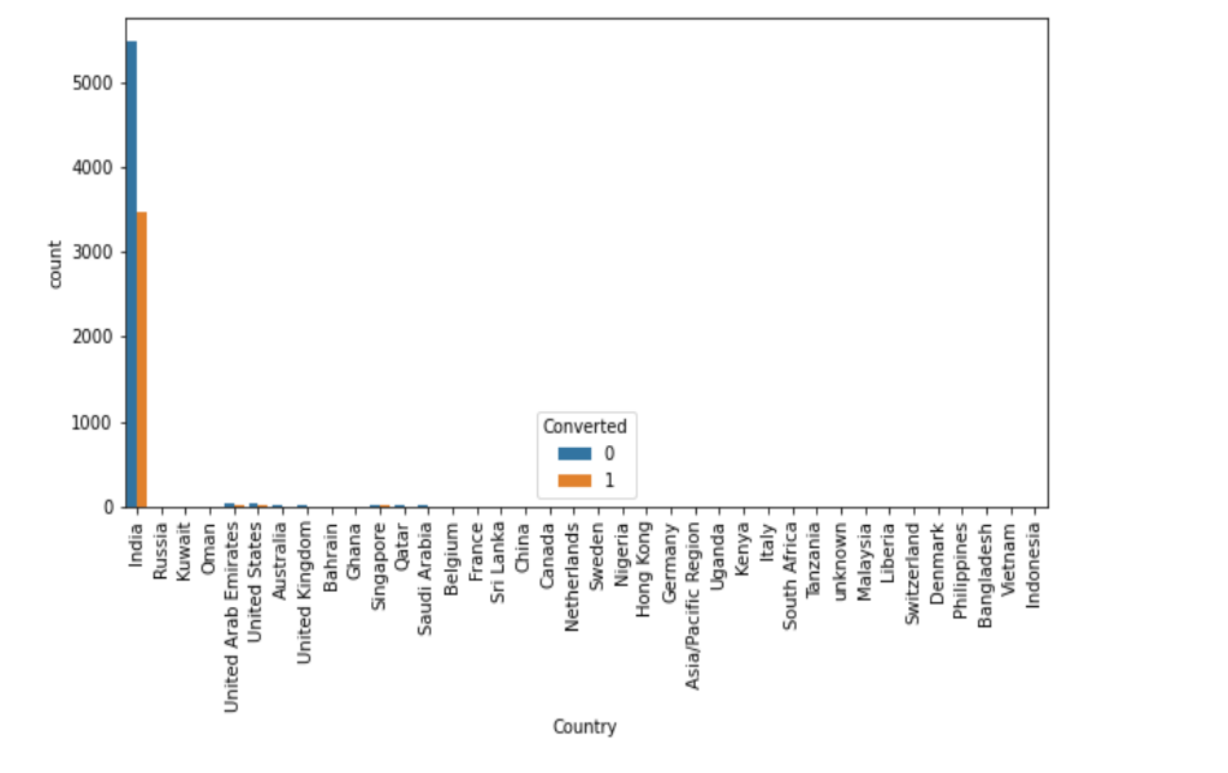
CATEGORICAL VARIABLE ANALYSIS

CHECKING COUNT_VALUES ON COUNTRY COLUMN

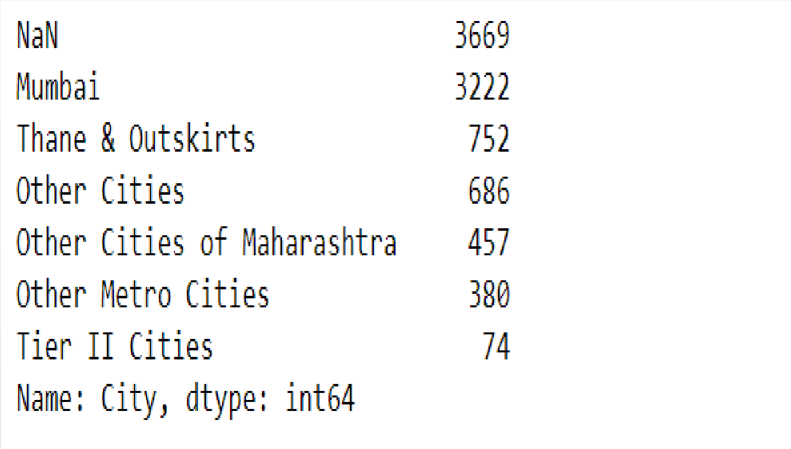
INSIGHTS

- As we can see the Number of Values for India are quite high with nearly 97% of the Data

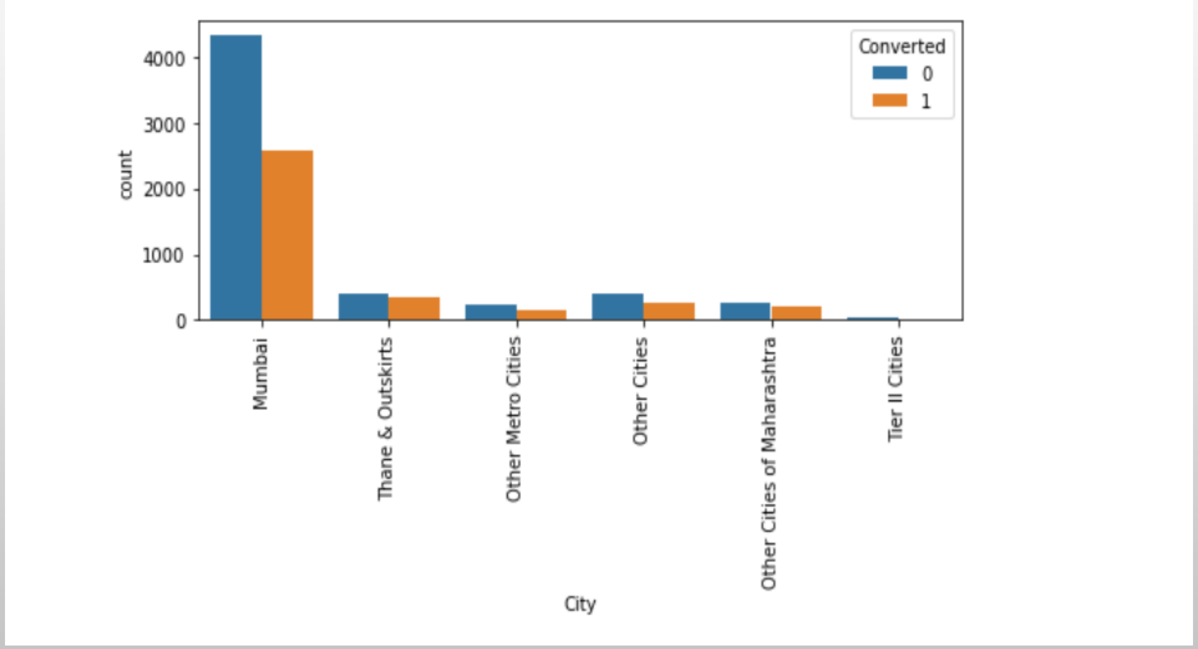
- It wont help analysis the data fully, this column can be dropped.



CHECKING COUNT_VALUES IN CITY COLUMN



```
NaN                            3669
Mumbai                         3222
Thane & Outskirts               752
Other Cities                    686
Other Cities of Maharashtra     457
Other Metro Cities              380
Tier II Cities                   74
Name: City, dtype: int64
```

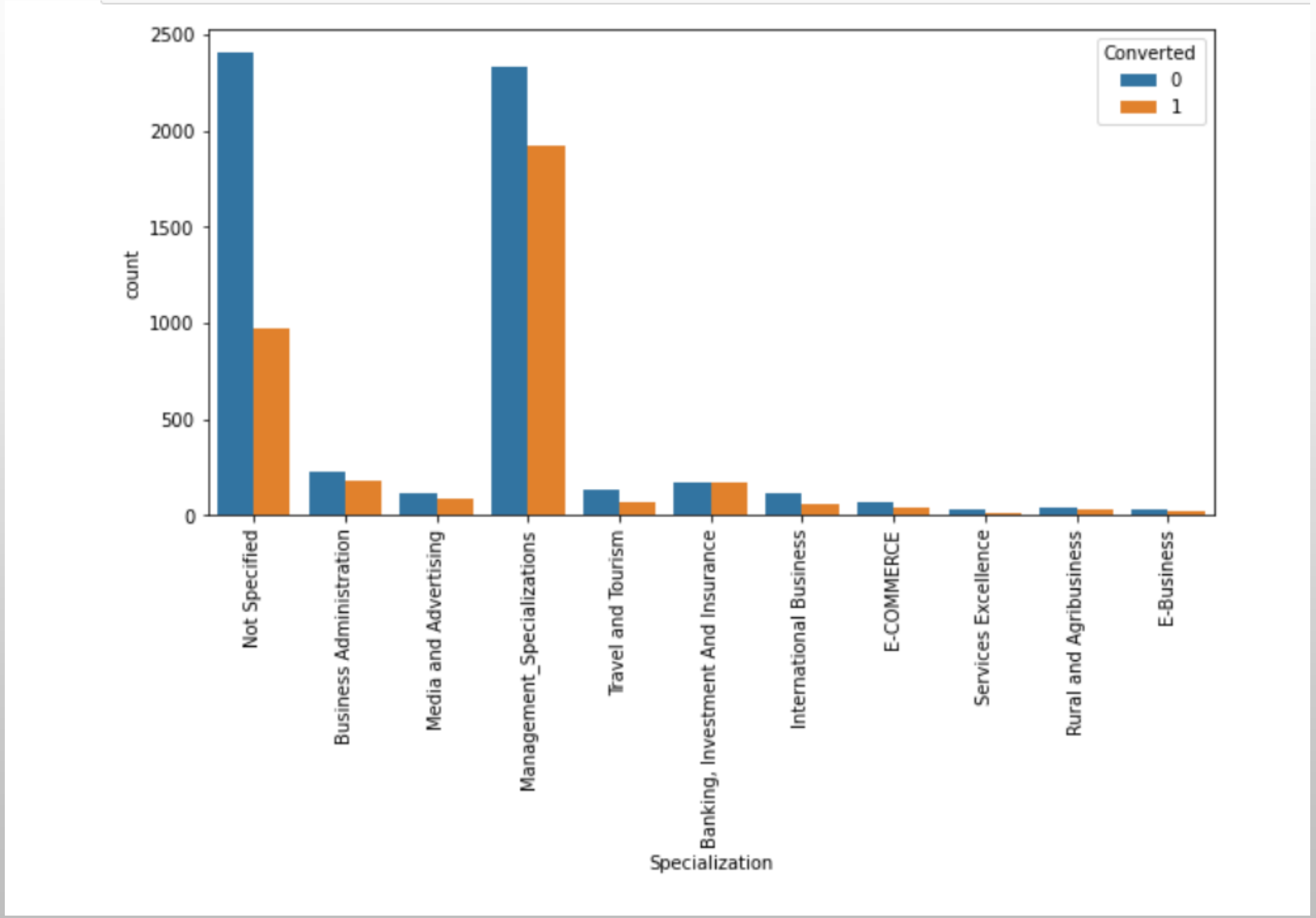


CHECKING COUNT_VALUES IN SPECIALIZATION COLUMN

INSIGHTS :

- As we can see that **Management** specialization has Higher conversion rate than others.

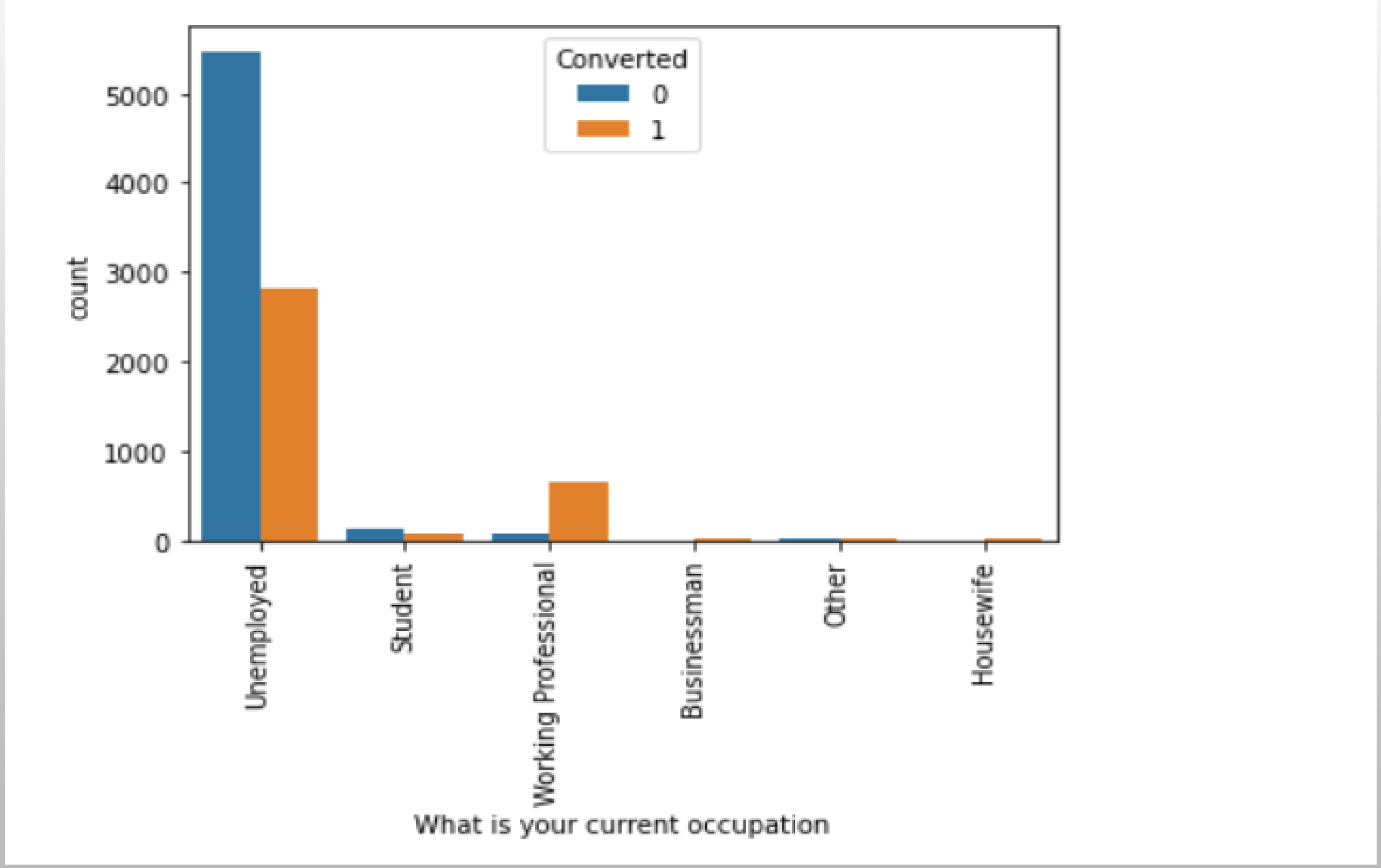
- So this is definitely a significant variable and should not be dropped.



CHECKING COUNT_VALUES IN THE WHAT IS YOUR CURRENT OCCUPATION COLUMN

INSIGHTS

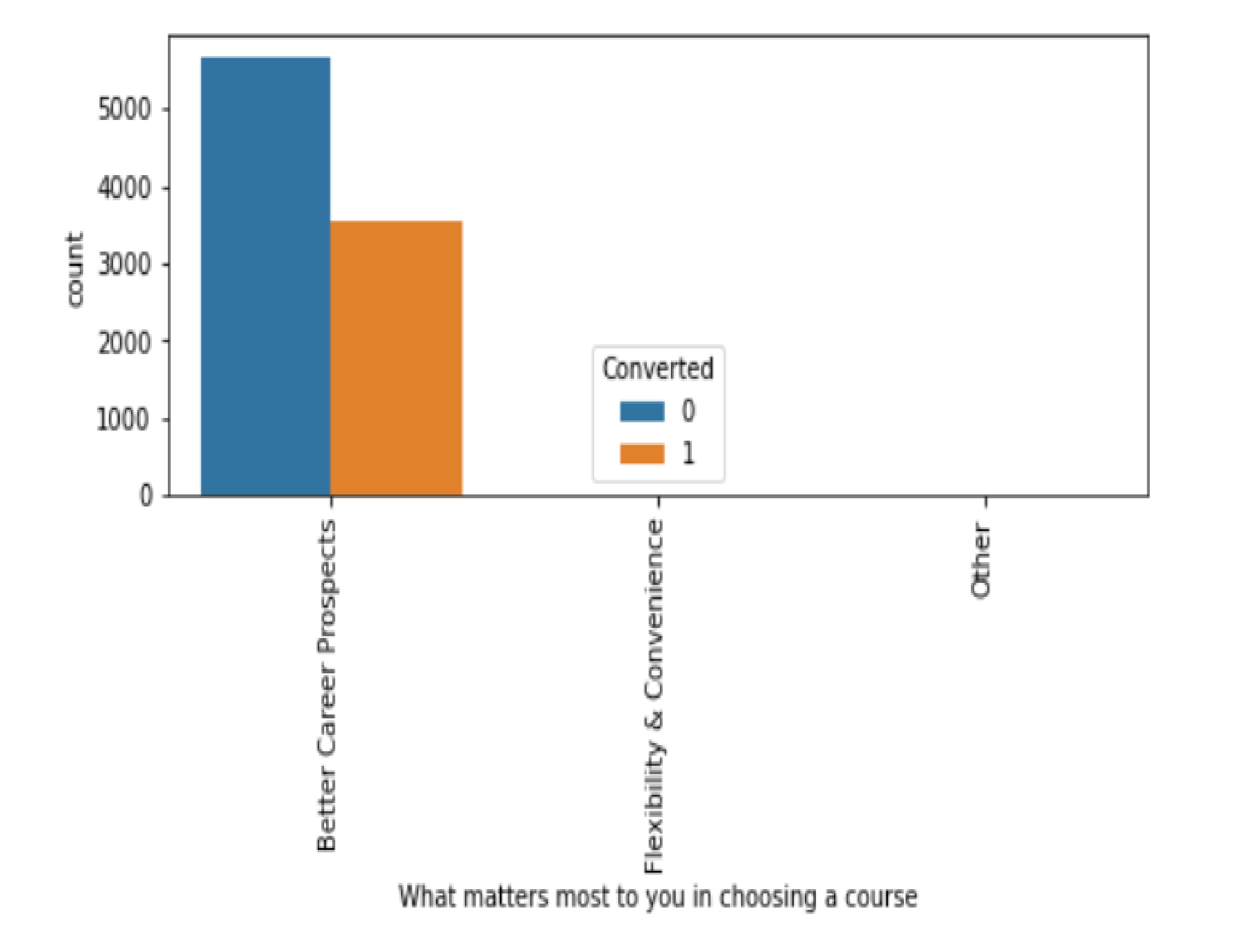
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.



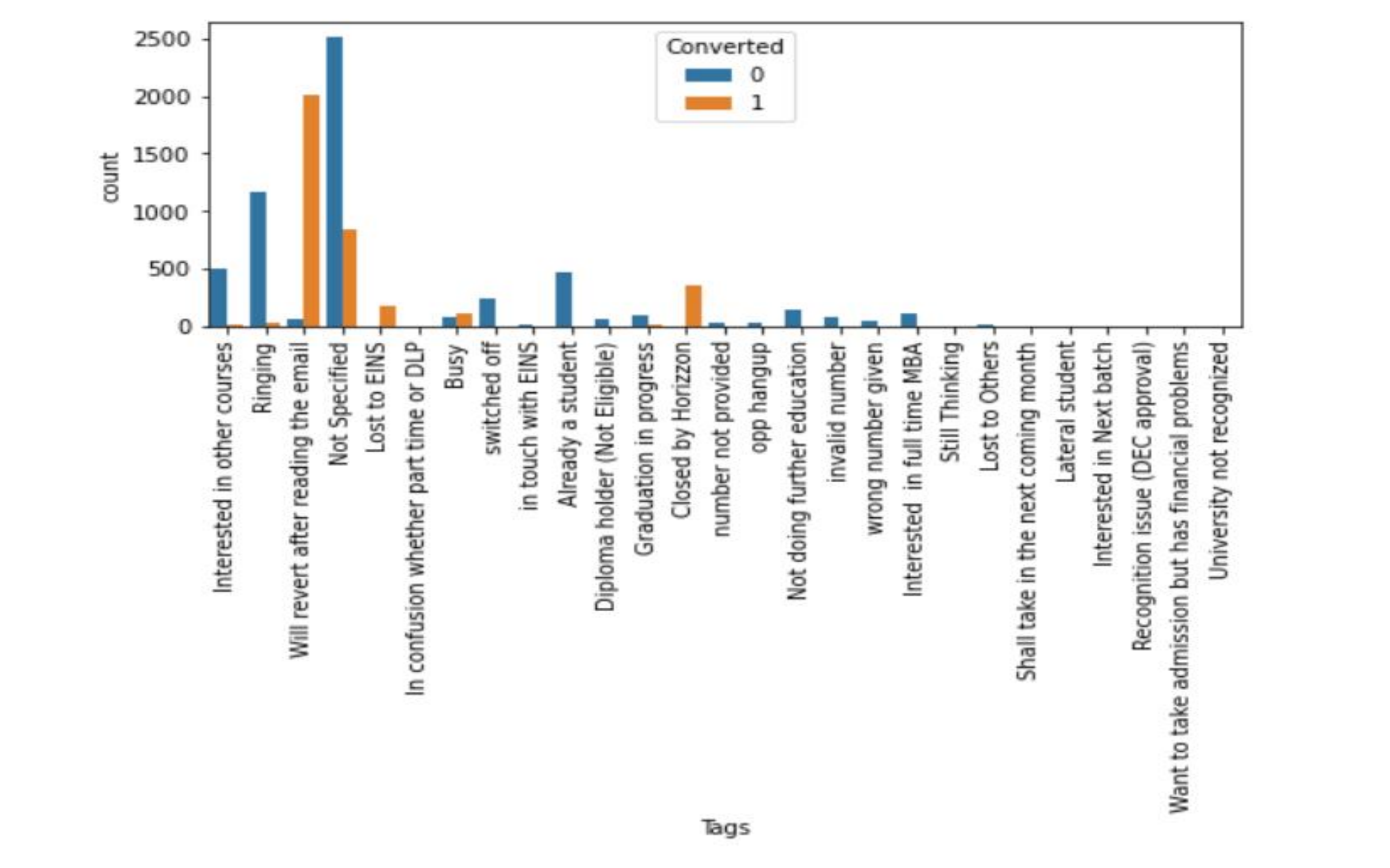
CHECKING COUNT_VALUES IN THE WHAT MATTERS MOST TO YOU IN CHOOSING THIS COURSE COLUMN

INSIGHTS :

- As we can see that motive of choosing a course is for better career prospects mostly.



CHECKING COUNT_VALUES IN THE TAGS COLUMN



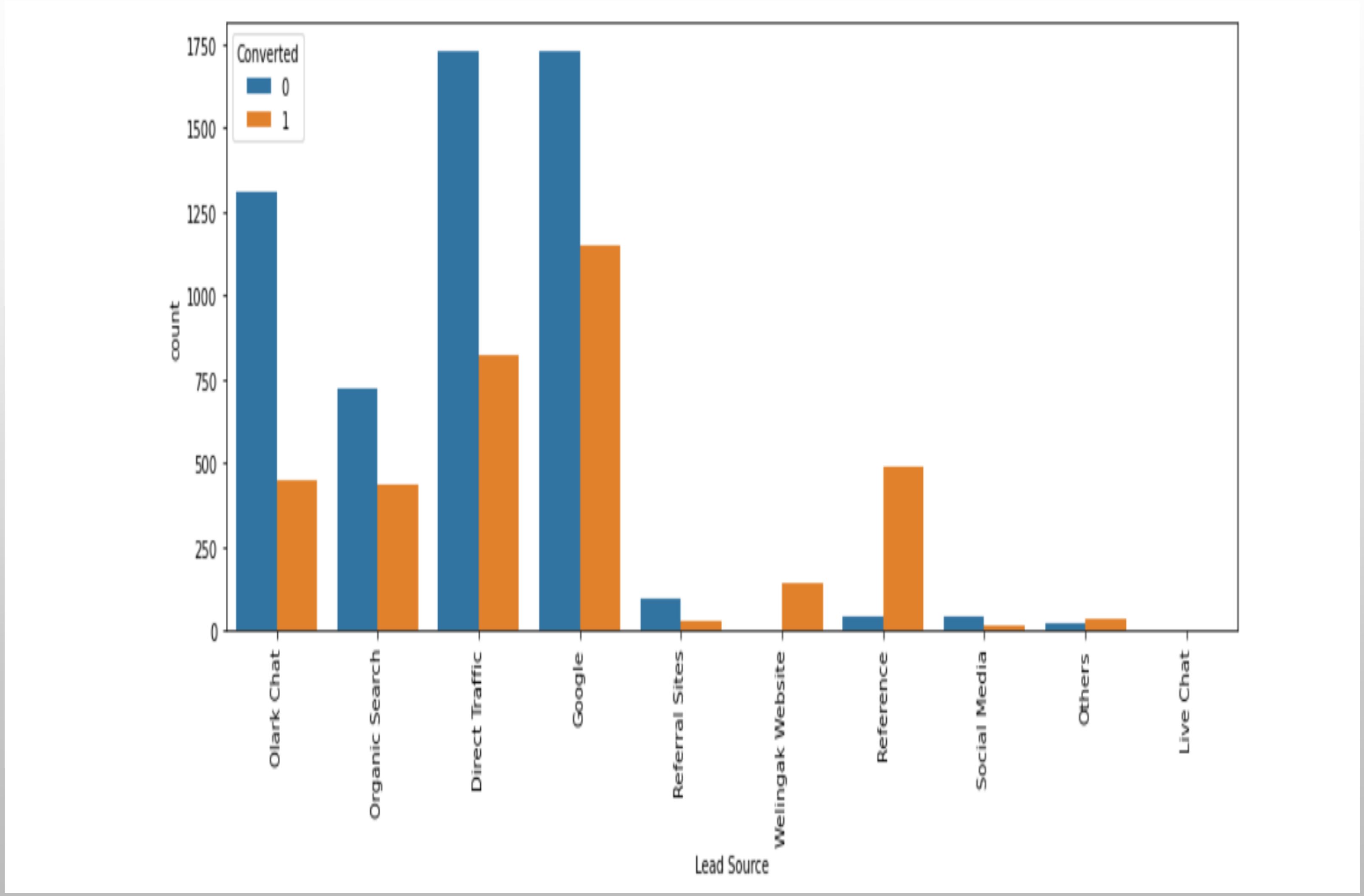
INSIGHTS :

As we can see the current status of the lead is Mostly they 'will revert after reading the email'.

CHECKING COUNT_VALUES IN THE LEAD SOURCE COLUMN

INSIGHTS :

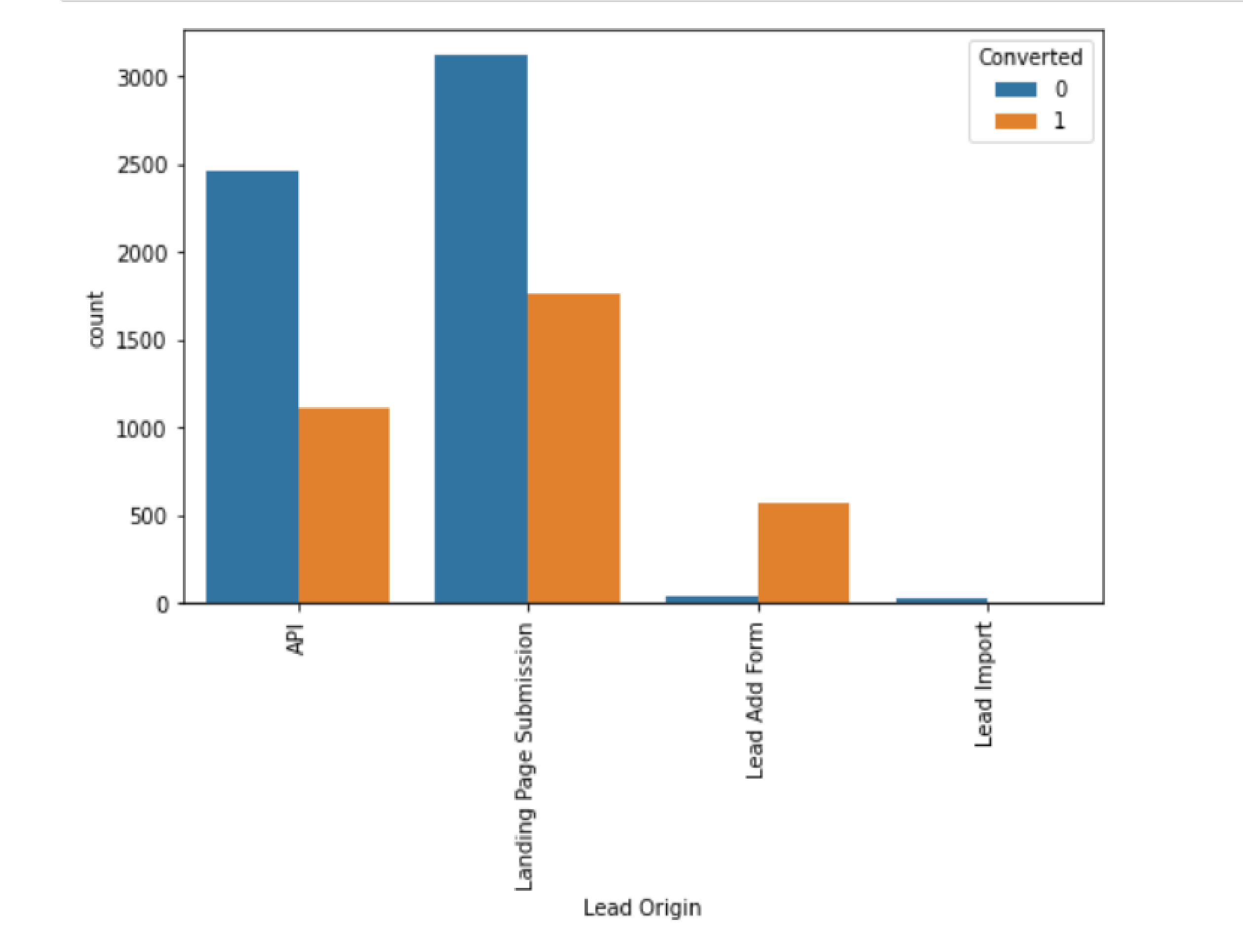
- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.



CHECKING COUNT_VALUES IN THE LEAD ORIGIN COLUMN

INSIGHTS :

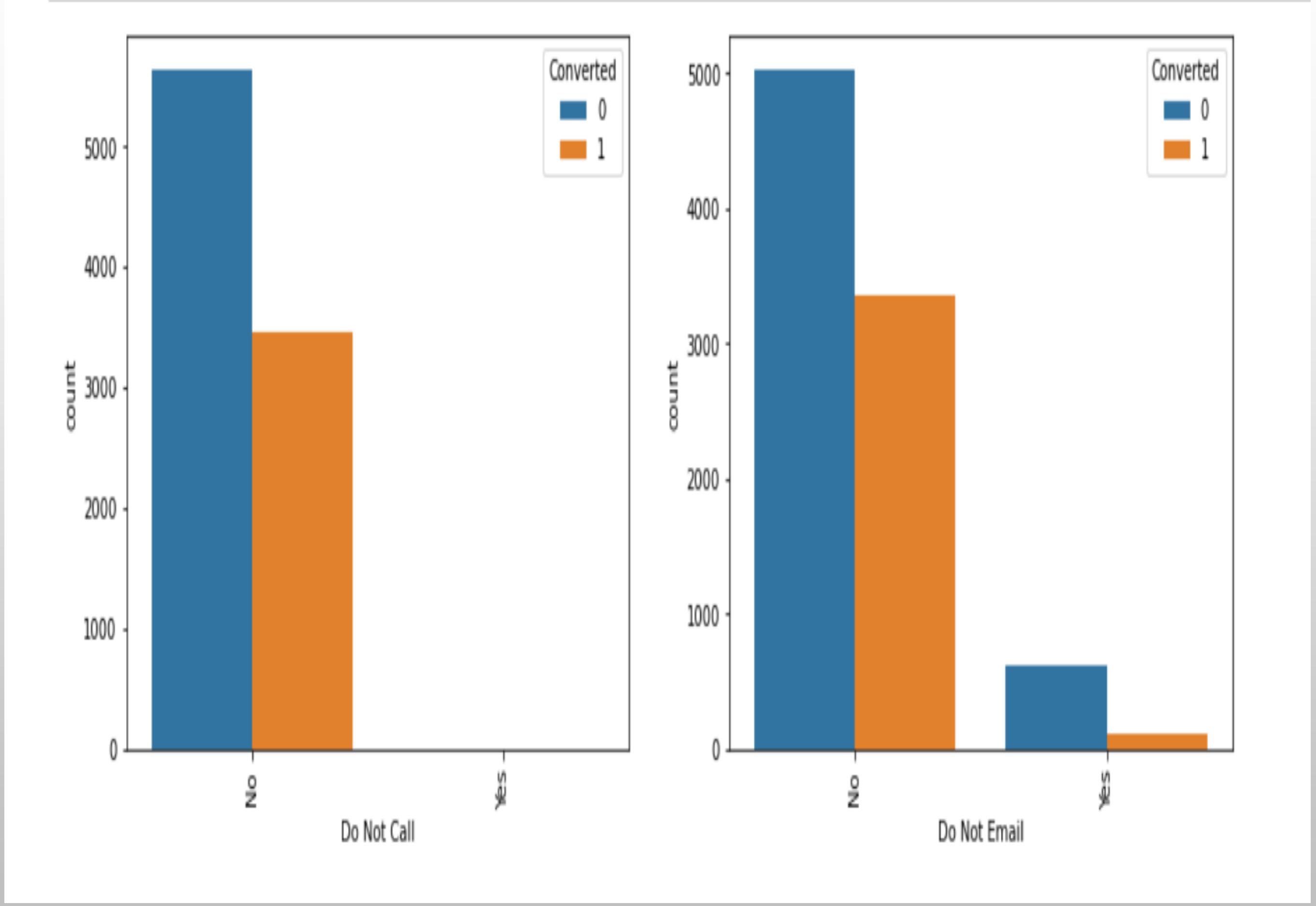
- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



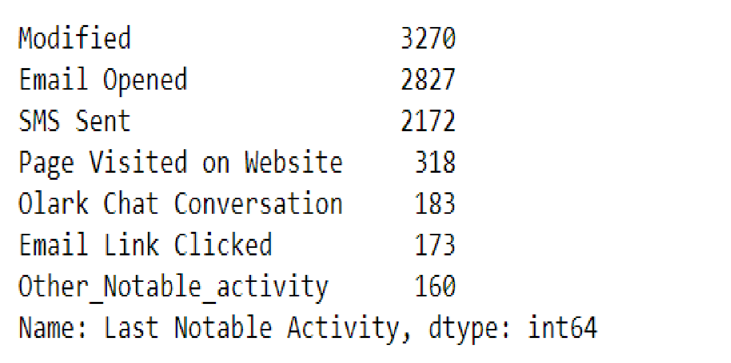
CHECKING COUNT_VALUES IN THE COLUMN

INSIGHTS :

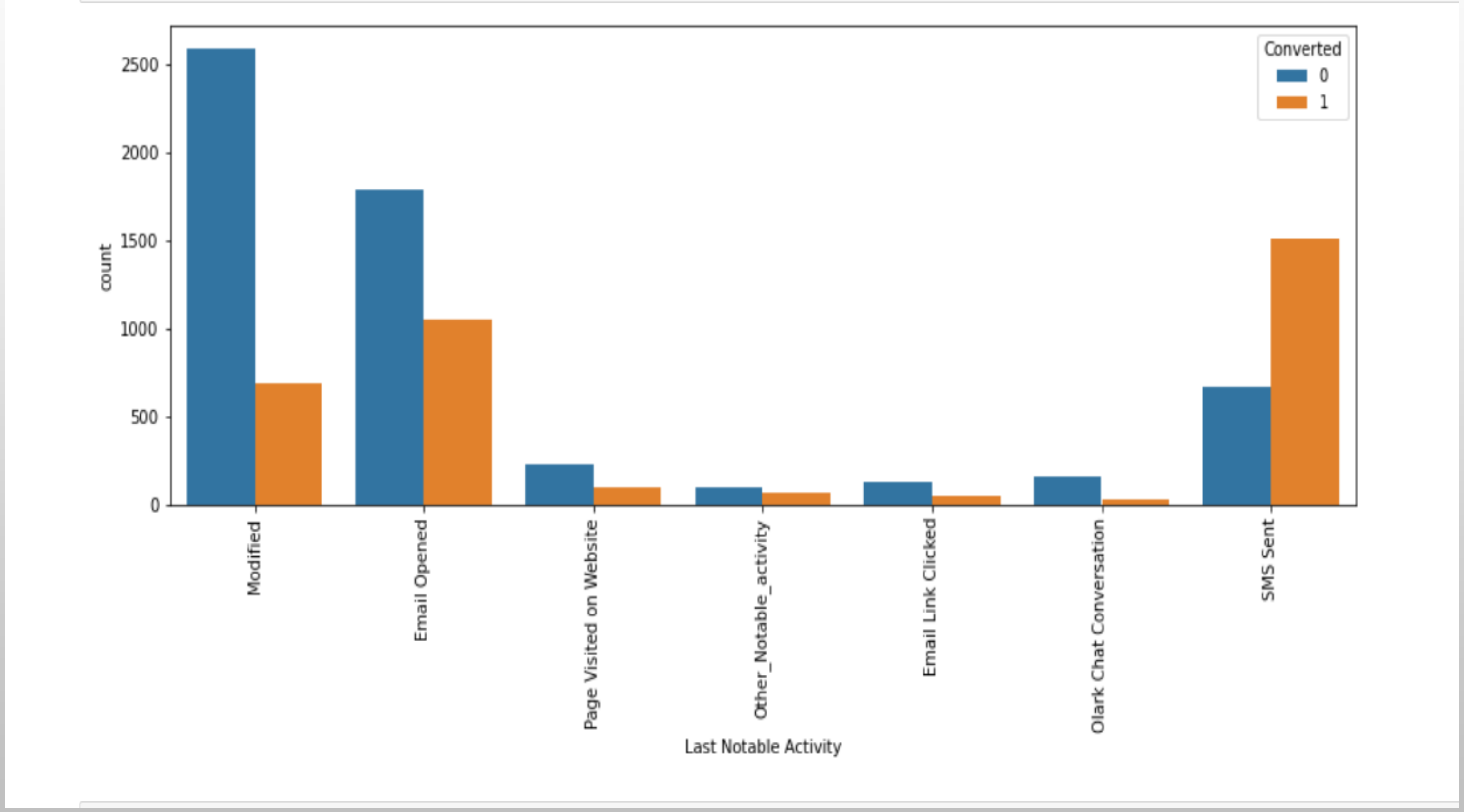
- We Can append the Do Not Call Column to the list of Columns to be Dropped since > 90% is of only one Value



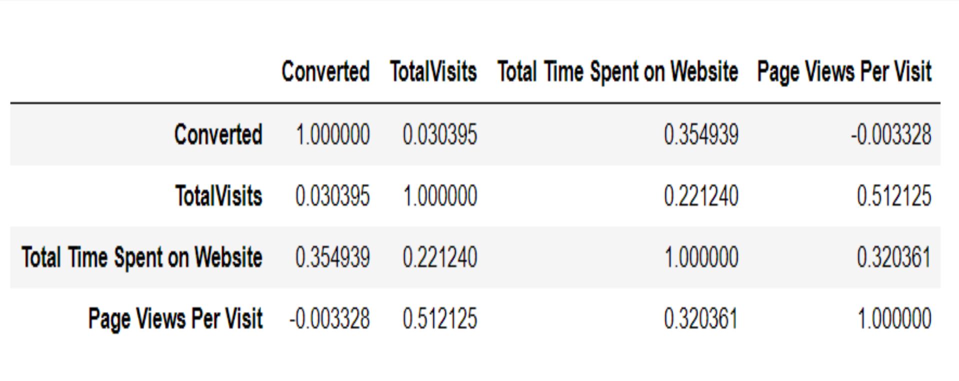
CHECKING COUNT_VALUES IN THE LAST NOTABLE ACTIVITY COLUMN



```
Modified                   3270
Email Opened               2827
SMS Sent                   2172
Page Visited on Website     318
Olark Chat Conversation     183
Email Link Clicked          173
Other_Notable_activity      160
Name: Last Notable Activity, dtype: int64
```

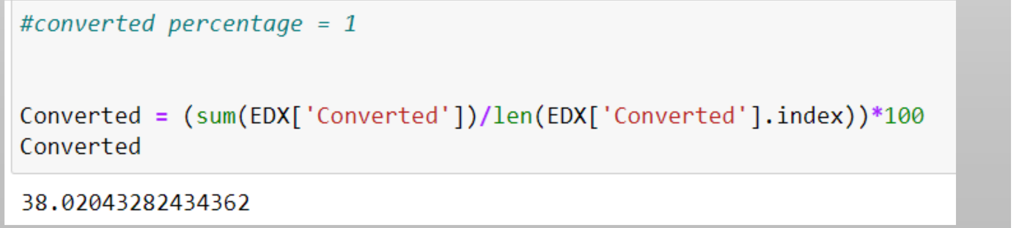


NUMERICAL VARIABLE ANALYSIS



	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1.000000	0.030395	0.354939	-0.003328
TotalVisits	0.030395	1.000000	0.221240	0.512125
Total Time Spent on Website	0.354939	0.221240	1.000000	0.320361
Page Views Per Visit	-0.003328	0.512125	0.320361	1.000000

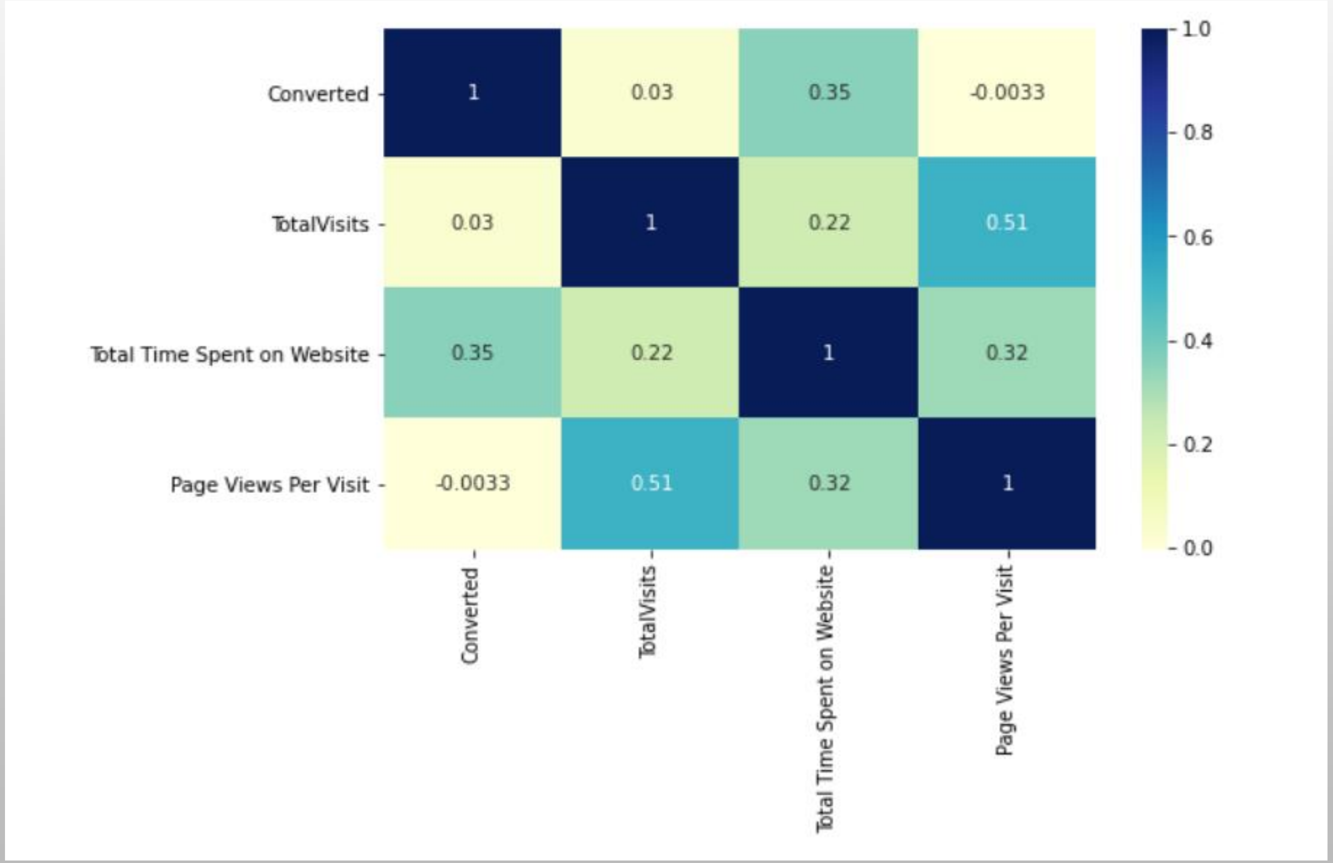
PERCENTAGE OF CONVERTED



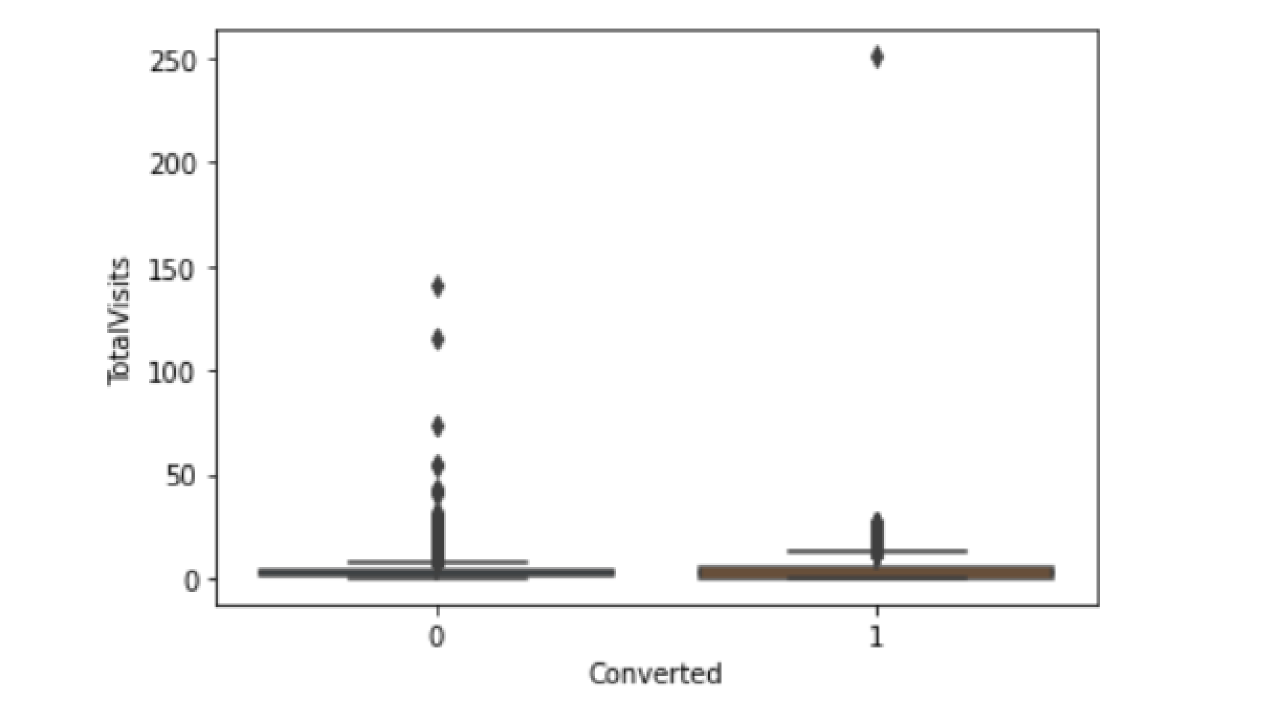
```
#converted percentage = 1


Converted = (sum(EDX['Converted'])/len(EDX['Converted'].index))*100
Converted
```

```
38.02043282434362
```

CHECKING COUNT_VALUES IN THE TARGETED VARIABLE



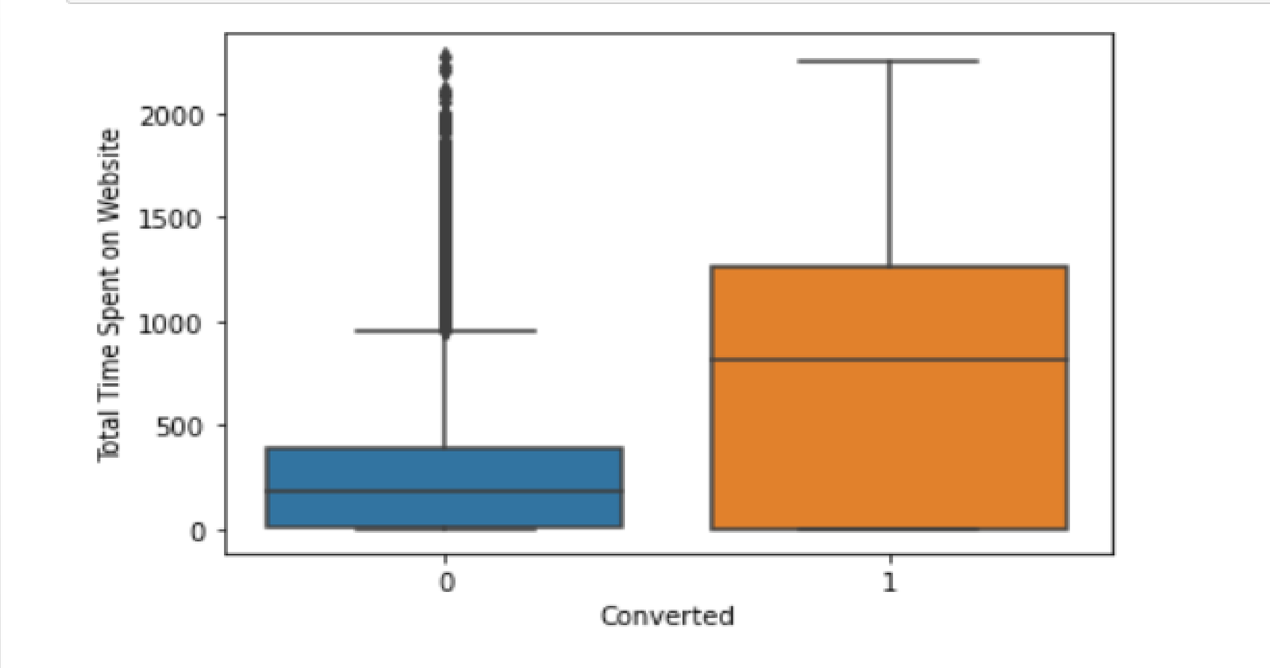
TOTAL VISITS   VS CONVERTED VARIABLE



INSIGHTS :

- Nothing can be conclude on the basis of Total Visits

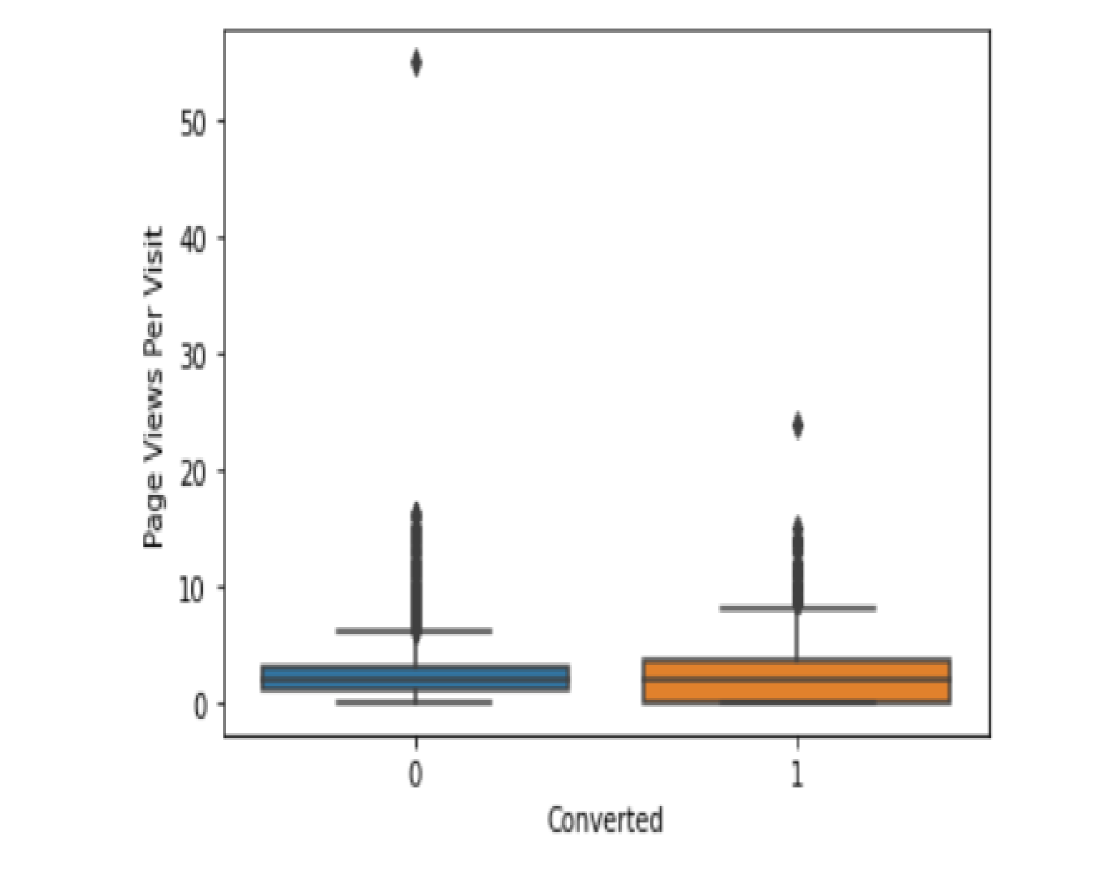
TOTAL TIME SPENT ON WEBSITE VS CONVERTED



INSIGHTS :

- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.

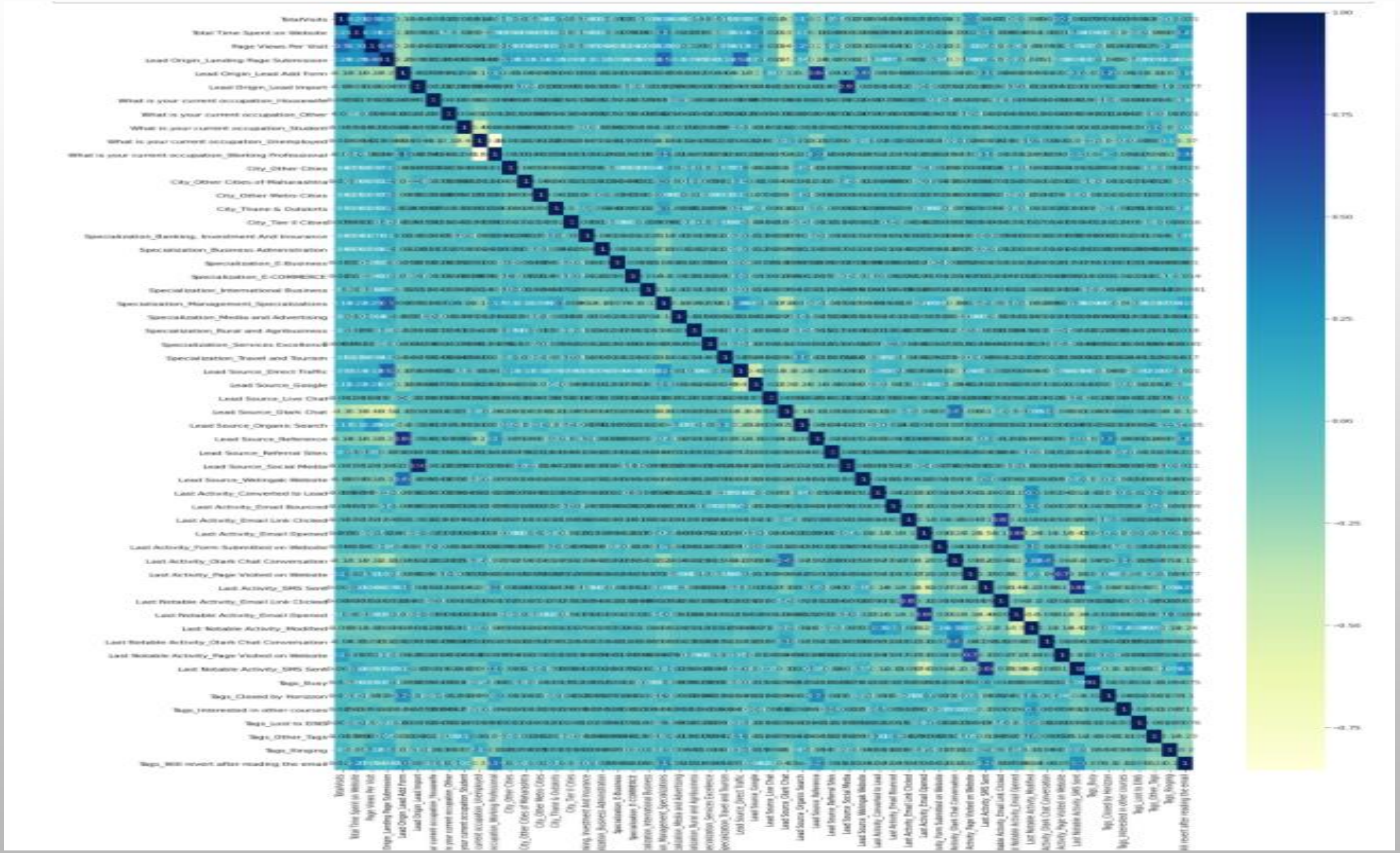
PAGE VIEWS PER VISITS VS CONVERTED



INSIGHTS :

- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit.

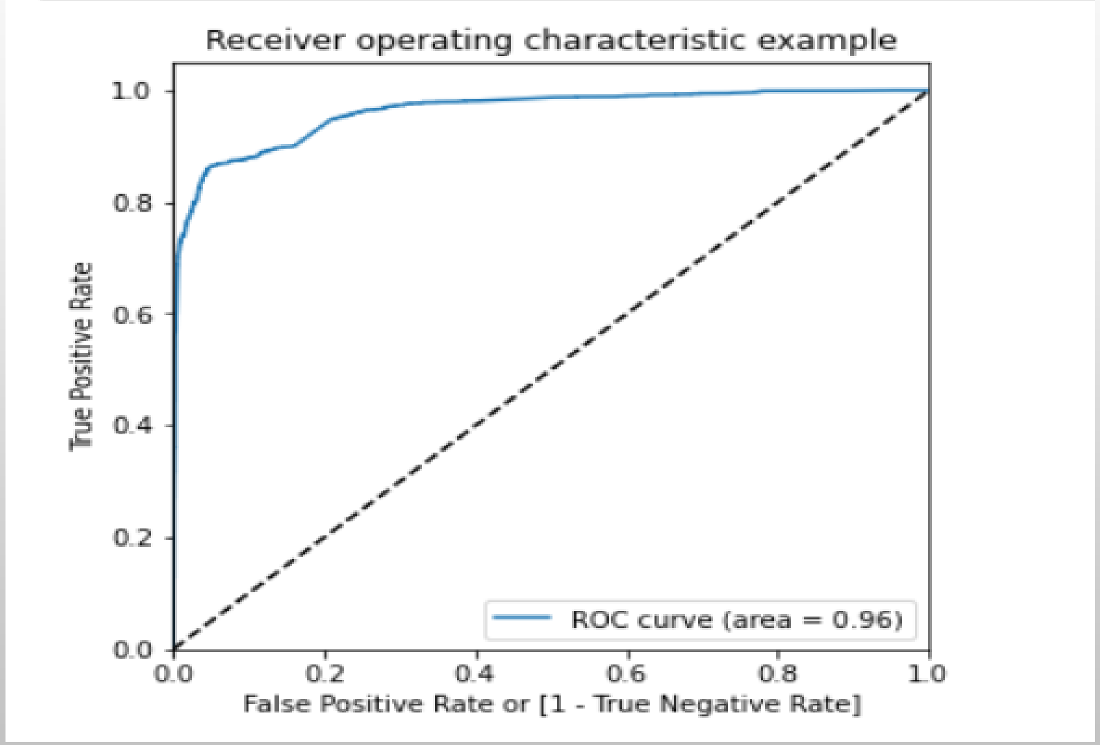
CORRELATION HEAT MAP ON TRAIN DATA_SET



OPTIMAL CUT-OFF PROBABILITY

The ROC Curve should be a value close to 1.

We are getting a good value of 0.96 indicating a good predictive model.

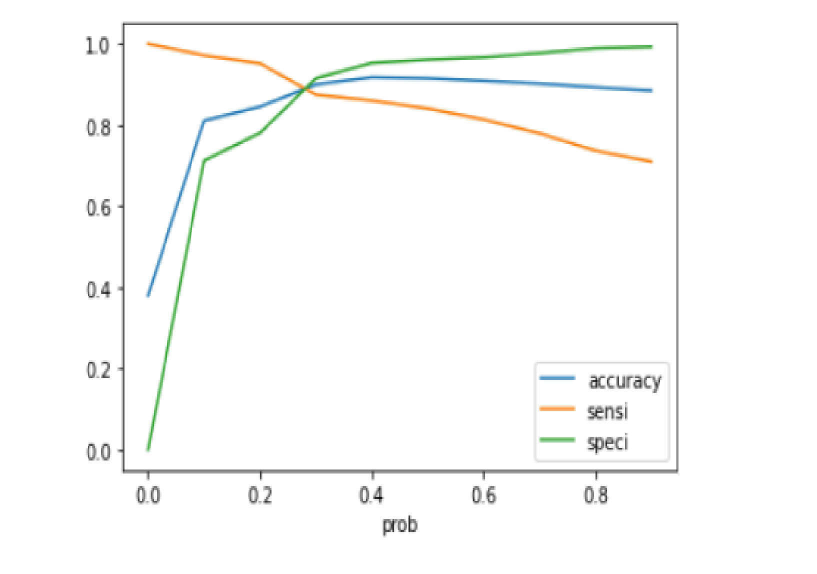


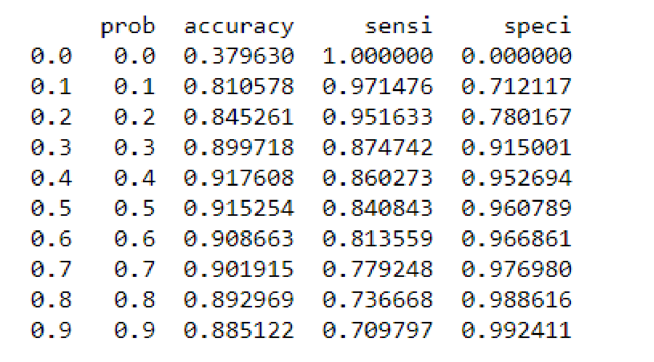
FINDING FINAL OPTIMAL CUT-OFF

INSIGHTS :

So as we can see above the model seems to be performing well. The ROC curve has a value of 0.96, which is a good value. We have the following values for the Train Data:

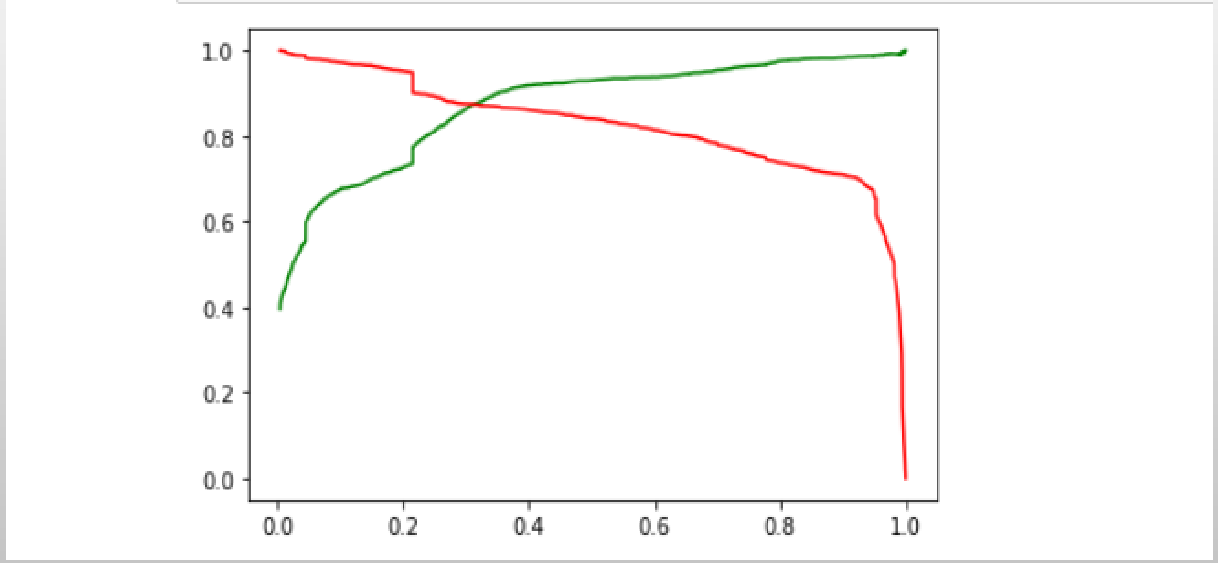
- **Accuracy :** 89.9%
- **Sensitivity :** 87%
- **Specificity :** 91.5%



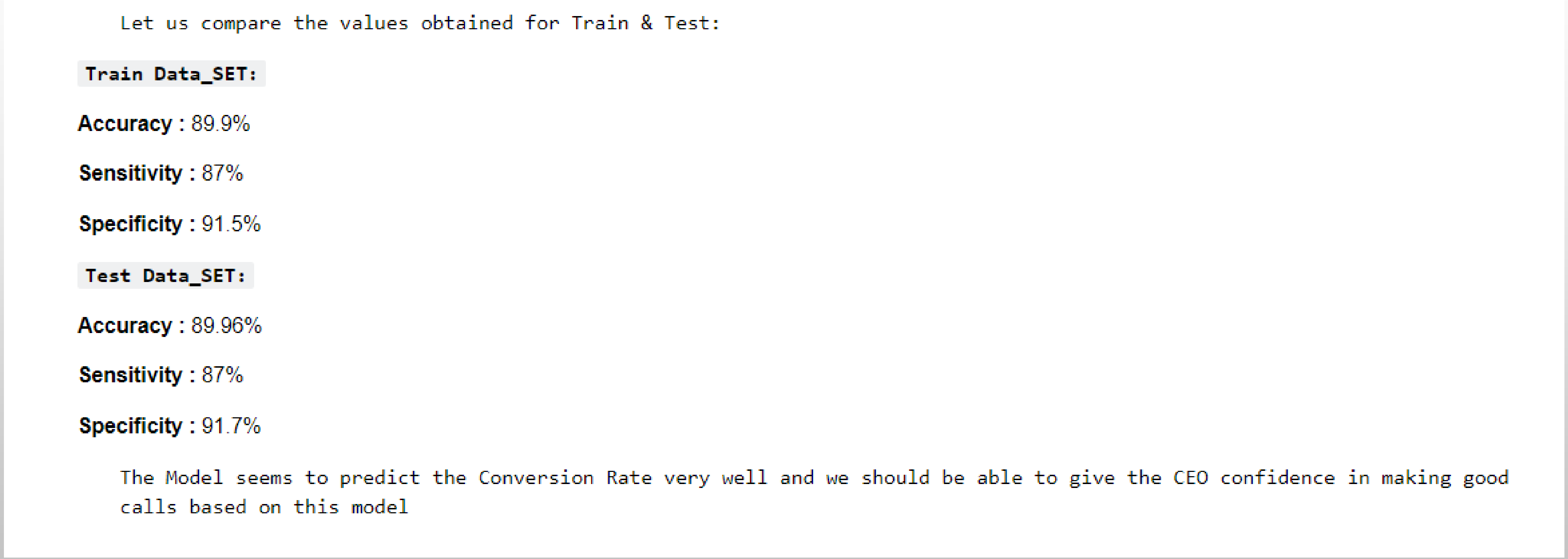


```
     prob  accuracy     sensi     speci
0.0   0.0  0.379630  1.000000  0.000000
0.1   0.1  0.810578  0.971476  0.712117
0.2   0.2  0.845261  0.951633  0.780167
0.3   0.3  0.899718  0.874742  0.915001
0.4   0.4  0.917608  0.860273  0.952694
0.5   0.5  0.915254  0.840843  0.960789
0.6   0.6  0.908663  0.813559  0.966861
0.7   0.7  0.901915  0.779248  0.976980
0.8   0.8  0.892969  0.736668  0.988616
0.9   0.9  0.885122  0.709797  0.992411
```

PRECISION RECALL CURVE



FINAL OBSERVATION



Let us compare the values obtained for Train & Test:

Train Data_SET:

Accuracy : 89.9%

Sensitivity : 87%

Specificity : 91.5%

Test Data_SET:

Accuracy : 89.96%

Sensitivity : 87%

Specificity : 91.7%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model