Lead Scoring Assignment

By
Anuradha. R
Shruthi . J
Aarushi Vohra

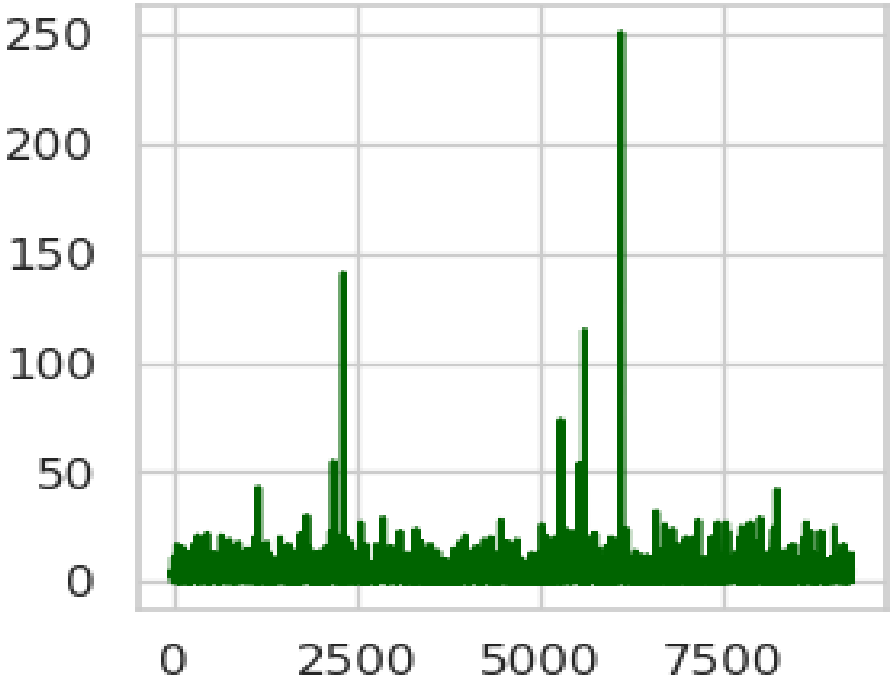
Problem Statement

- This case study involves helping X Education, an education company, improve its lead conversion rate by building a logistic regression model.
- 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.
- The company markets its courses on several websites and search engines like Google. Once these 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. Moreover, the company also gets leads through past referrals. Once these 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%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

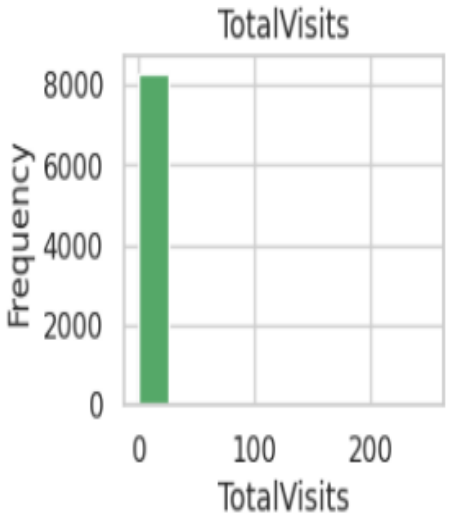
Goals of the Case Study

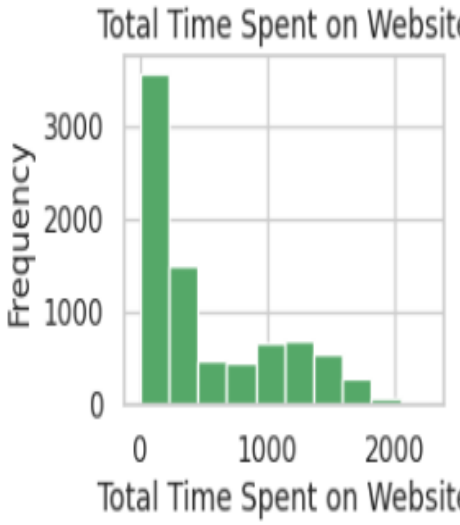
- There are quite a few goals for this case study:
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- 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.

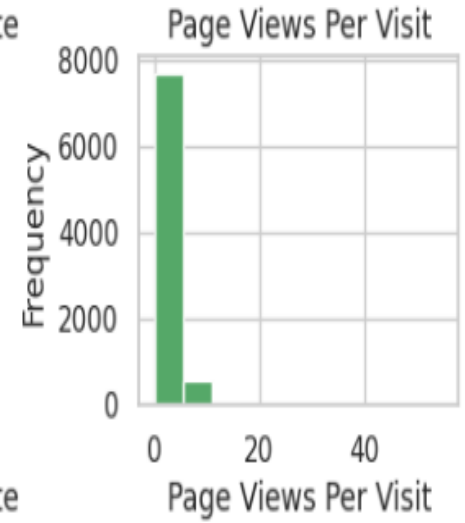
TOTAL_VISITS PLOT



EXPLORATORY DATA ANALYSIS



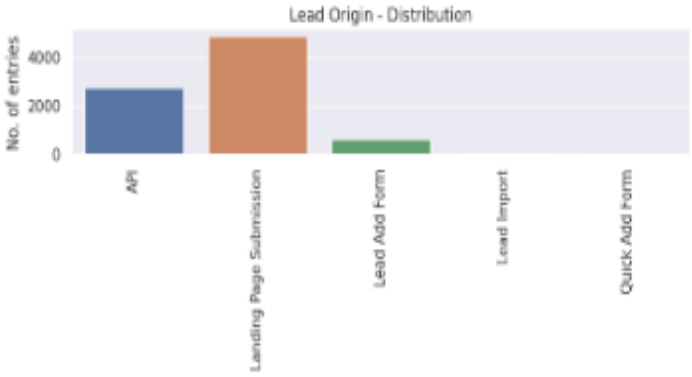


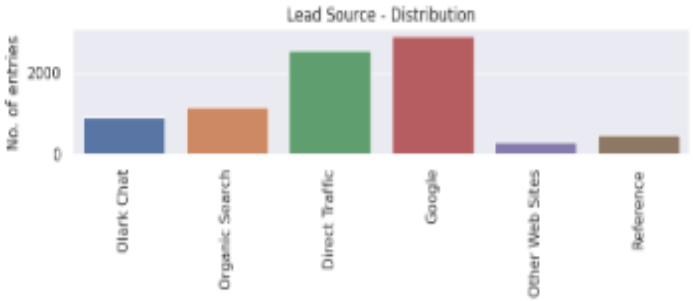


observation:

--> Outliers need to be checked as the plot is skewed and also the peak is relatively high

CATEGORICAL COLUMNS



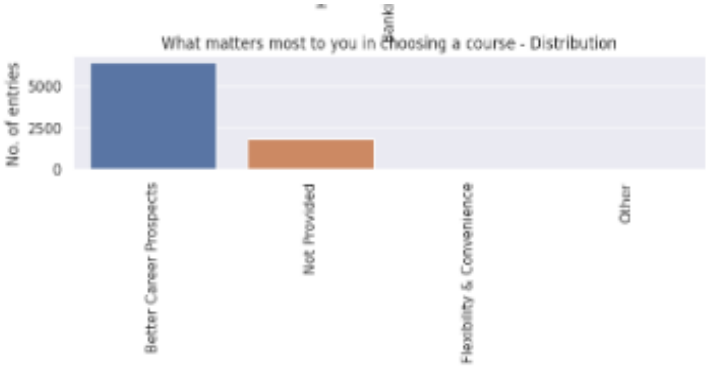


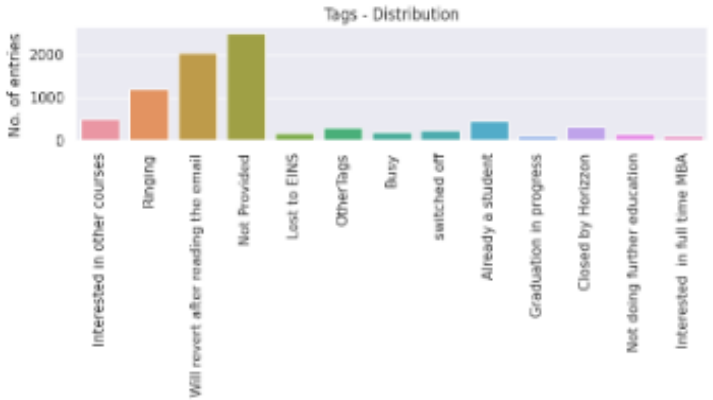
Last Activity - Distribution
No. of entries
2000
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Olark Chat Conversation
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Email Received
Email Marked Spam

Country - Distribution
No. of entries
5000
0
India
Foriegn Country

Specialization - Distribution
No. of entries
2000
1000
0
Not Provided
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business

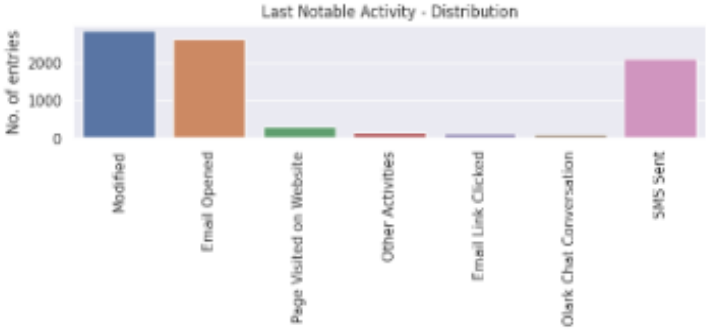
What is your current occupation - Distribution
No. of entries
5000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife


What matters most to you in choosing a course - Distribution
No. of entries
5000
2500
0
Better Career Prospects
Not Provided
Flexibility & Convenience
Other


Tags - Distribution
No. of entries
2000
1000
0
Interested in other courses
Ringing
Will revert after reading the email
Not Provided
Lost to EINS
OtherTags
Busy
switched off
Already a student
Graduation in progress
Closed by Horizzon
Not doing further education
Interested in full time MBA


City - Distribution
No. of entries
5000
2500
0
Mumbai
Other Maharashtra Cities
Non Maharashtra Cities


A free copy of Mastering The Interview - Distribution
No. of entries
4000
2000
0
No
Yes


Last Notable Activity - Distribution
No. of entries
2000
1000
0
Modified
Email Opened
Page Visited on Website
Other Activities
Email Link Clicked
Olark Chat Conversation
SMS Sent

Observations:

-->'Lead Origin' - Landing page submission & API have the highest count

-->'Lead Source' - Google, Direct Traffic have the highest count

-->'Do Not Email' - Most people choosed No, We need to check the percentage of people choosed Yes.

-->'Last Activity' - This also we need to analyse depending on Convereted. SMS Sent and Email Opened are the highest in count

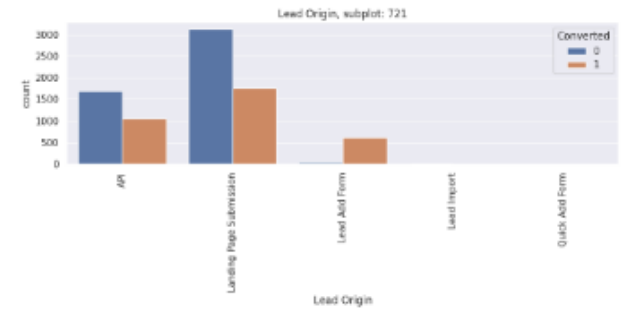
-->'Country' - Maximum customers are from India

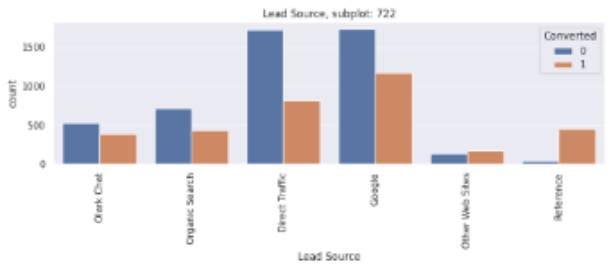
--> 'Specialization' - Among the options choosed Finance Management, Human Resource and Marketing Management are higest count but not selected by majority.

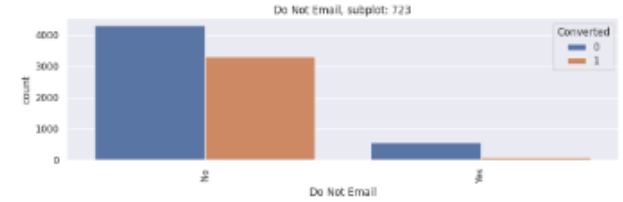
--> Also most of the users are currently unemployed.

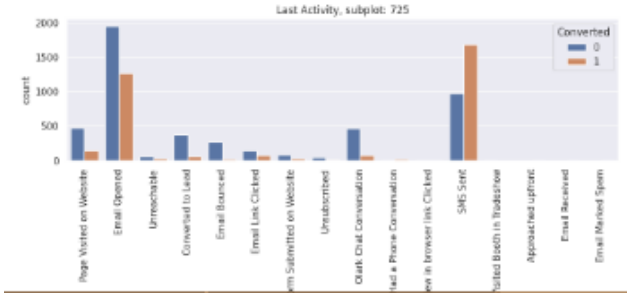
--> Most of the users are from Mumbai.

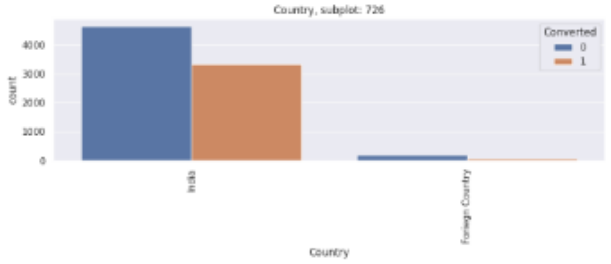
CATEGORICAL VARIABLES

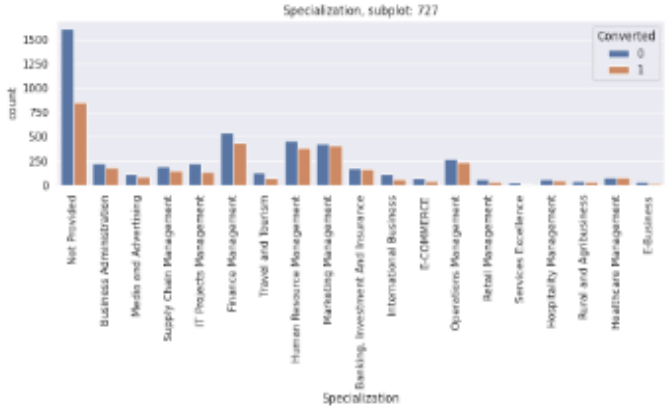


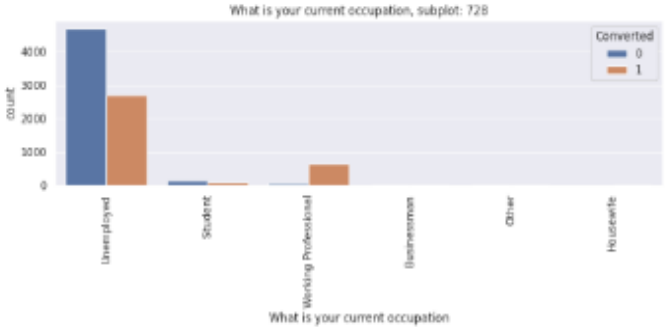


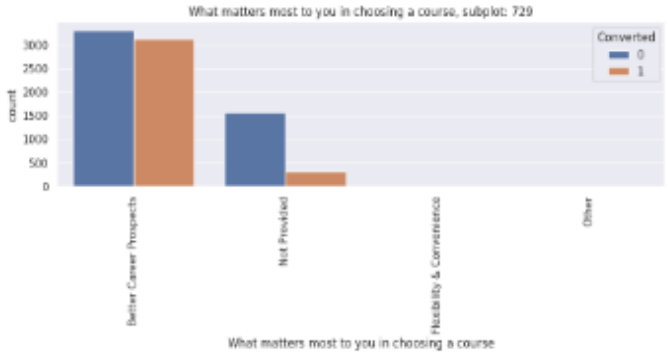


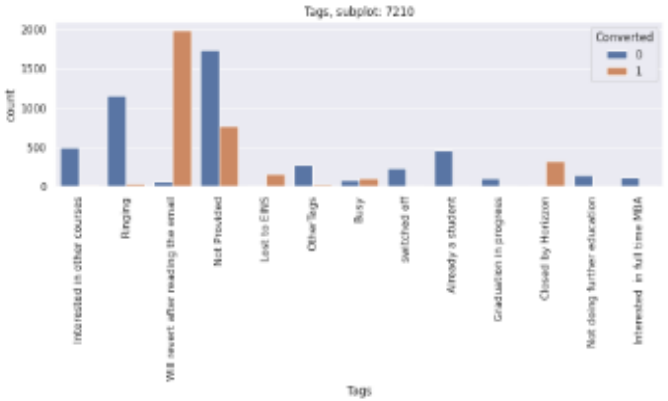





Specialization, subplot: 727
Converted
0
1
count
Not Provided
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

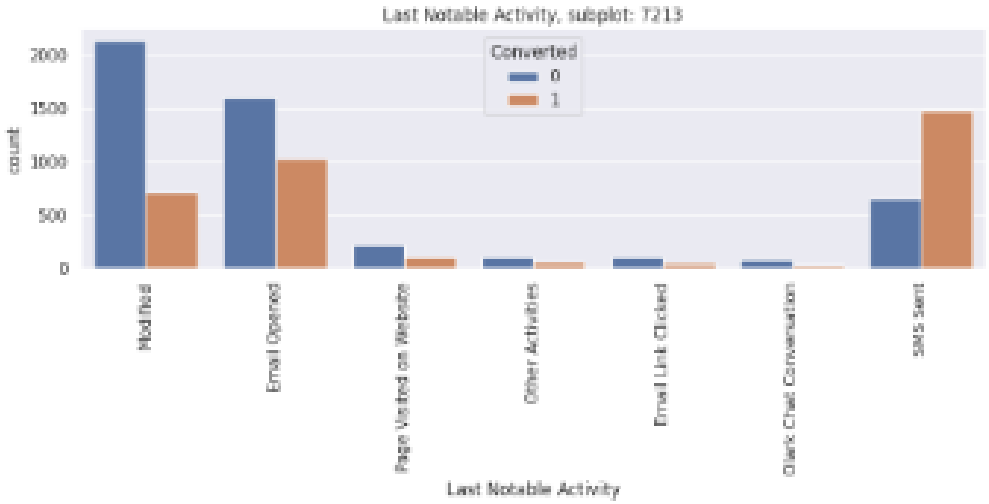

What is your current occupation, subplot: 728
Converted
0
1
count
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


What matters most to you in choosing a course, subplot: 729
Converted
0
1
count
Better Career Prospects
Not Provided
Flexibility & Convenience
Other
What matters most to you in choosing a course


Tags, subplot: 7210
Converted
0
1
count
Interested in other courses
Ringing
Will revert after reading the email
Not Provided
Lost to EINS
OtherTags
Busy
switched off
Already a student
Graduation in progress
Closed by Horizon
Not doing further education
Interested in full time MBA
Tags


City, subplot: 7211
Converted
0
1
count
Mumbai
Other Maharashtra Cities
Non Maharashtra Cities
City


A free copy of Mastering The Interview, subplot: 7212
Converted
0
1
count
No
Yes
A free copy of Mastering The Interview



Observations:

-->'lead Origin'-Leads added from orgin are mostly positively converted

-->'Lead Source'-Leads from Reference and Other websites are also mostly converted positively.

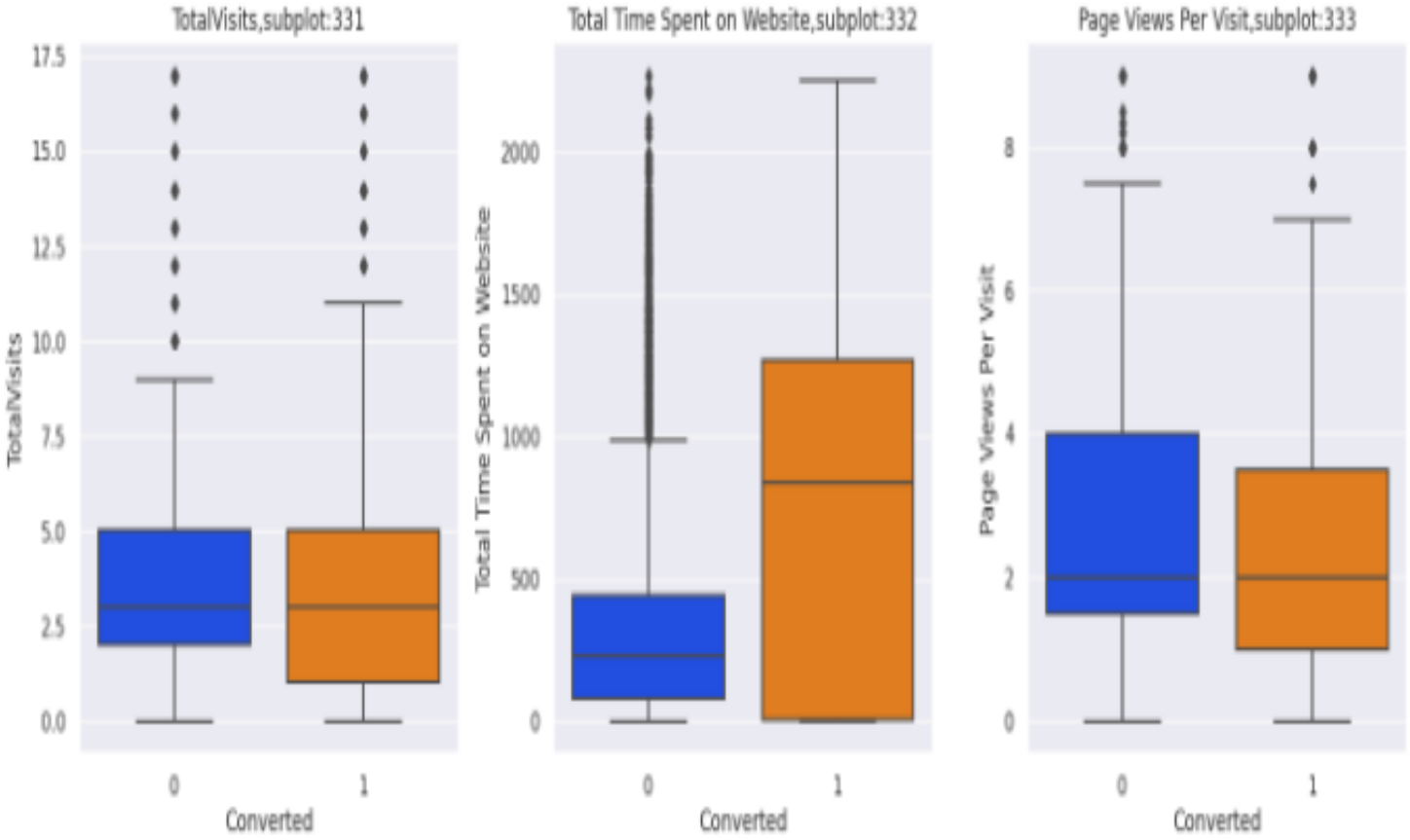
-->Working Professionals convertion ratio is very high

-->'Tags'-Will revert after reading email, Closed by Horizon are getting mostly converted

-->'Last notable Activity' column shows "SMS Sent" have high ratio of positive convertion

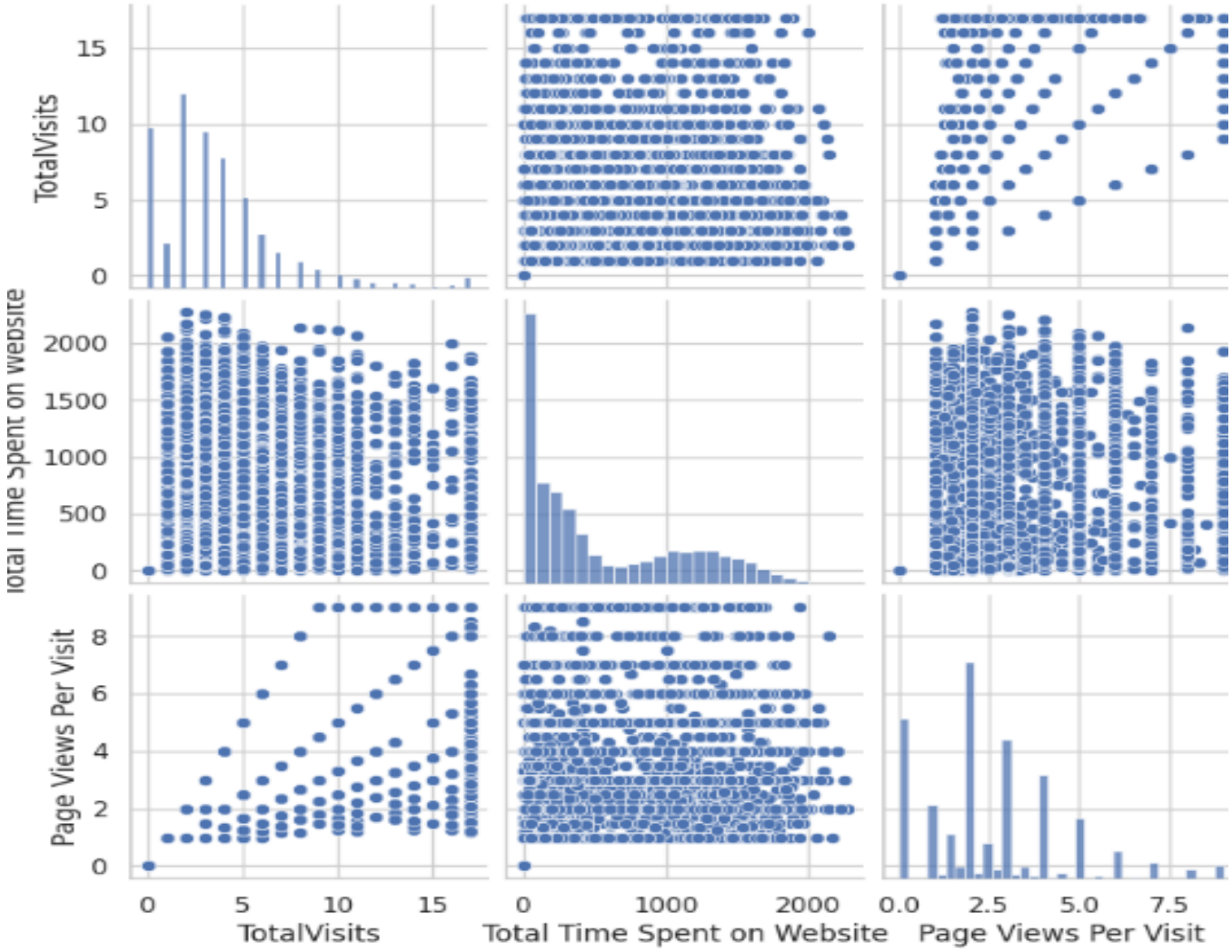
OUTLIER ANALYSIS

Observations: There are outliers in the "TotalVisits" and "Page View Per Visit" Columns. Either we can drop them or reassign outlier values to 99 percentile.

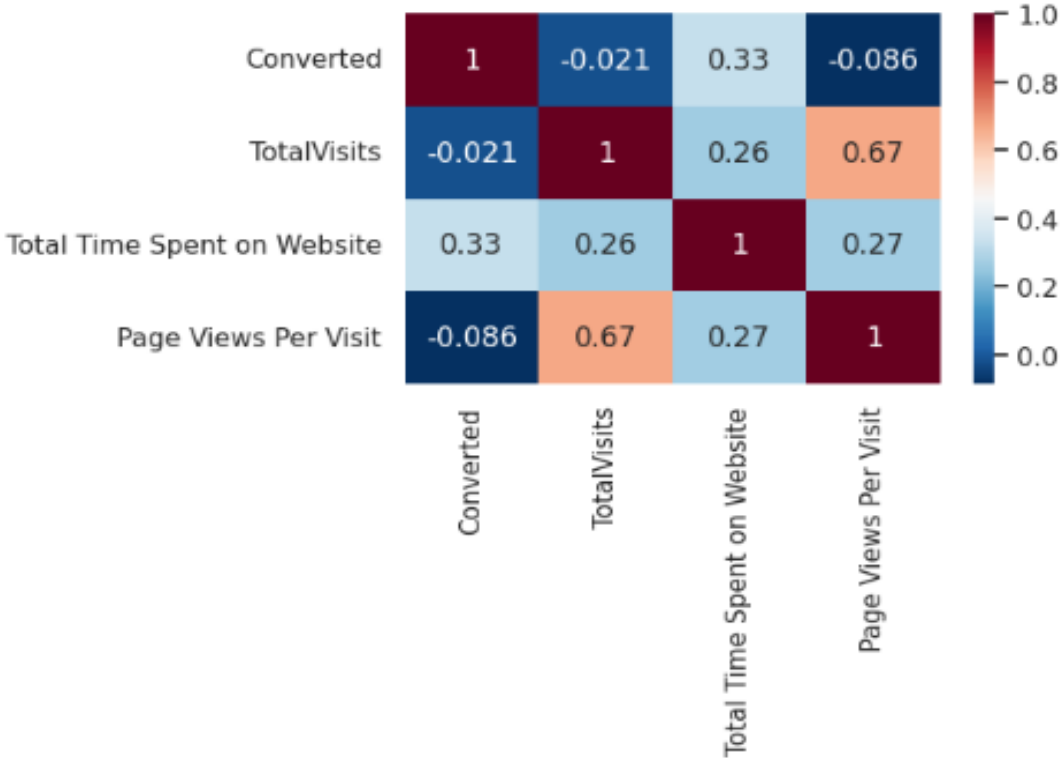


Observation : Total Time Spent on Website is very high and median value is also high for converted leads

'TotalVisits' and 'Page Views Per Visits' column are correlated.

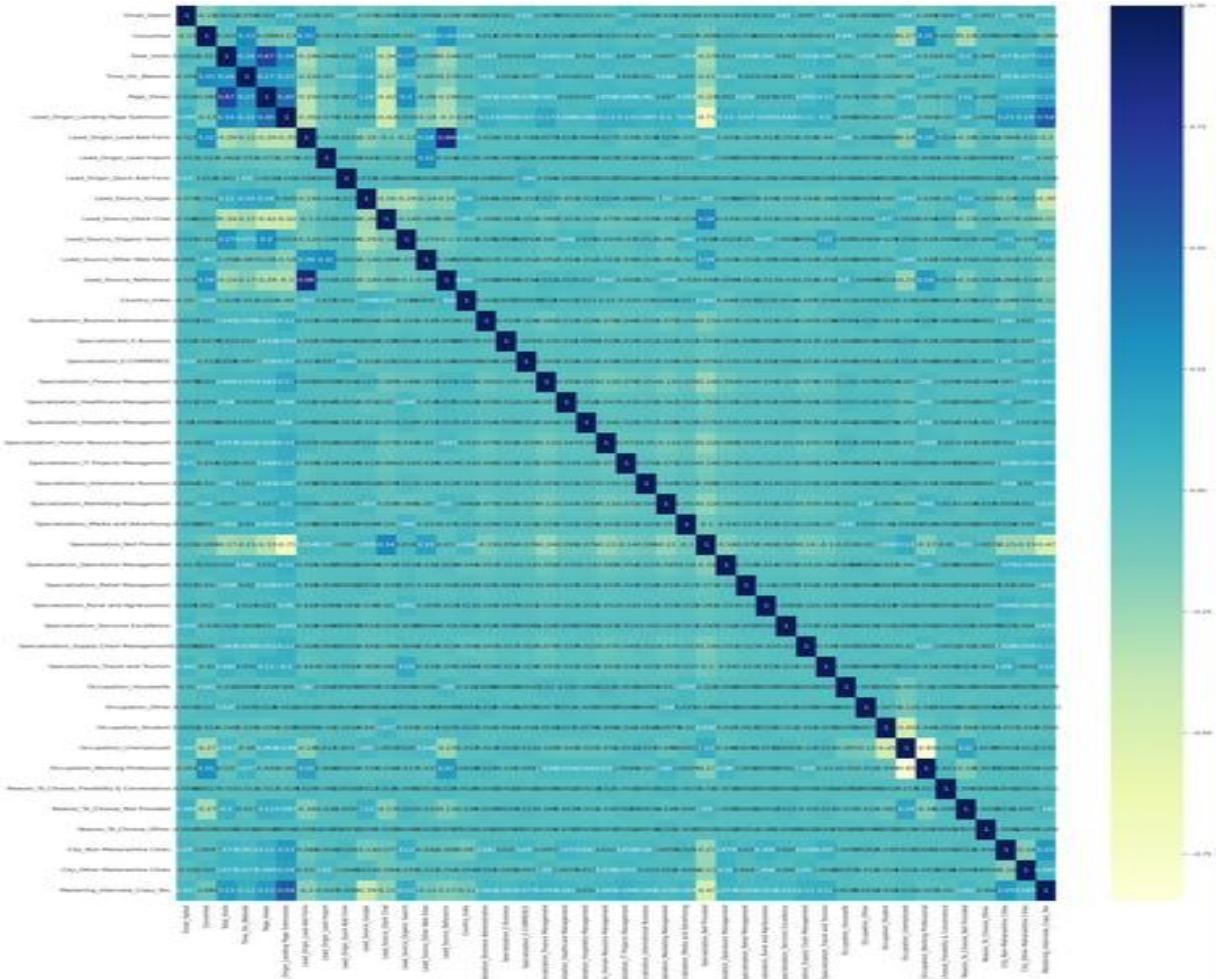


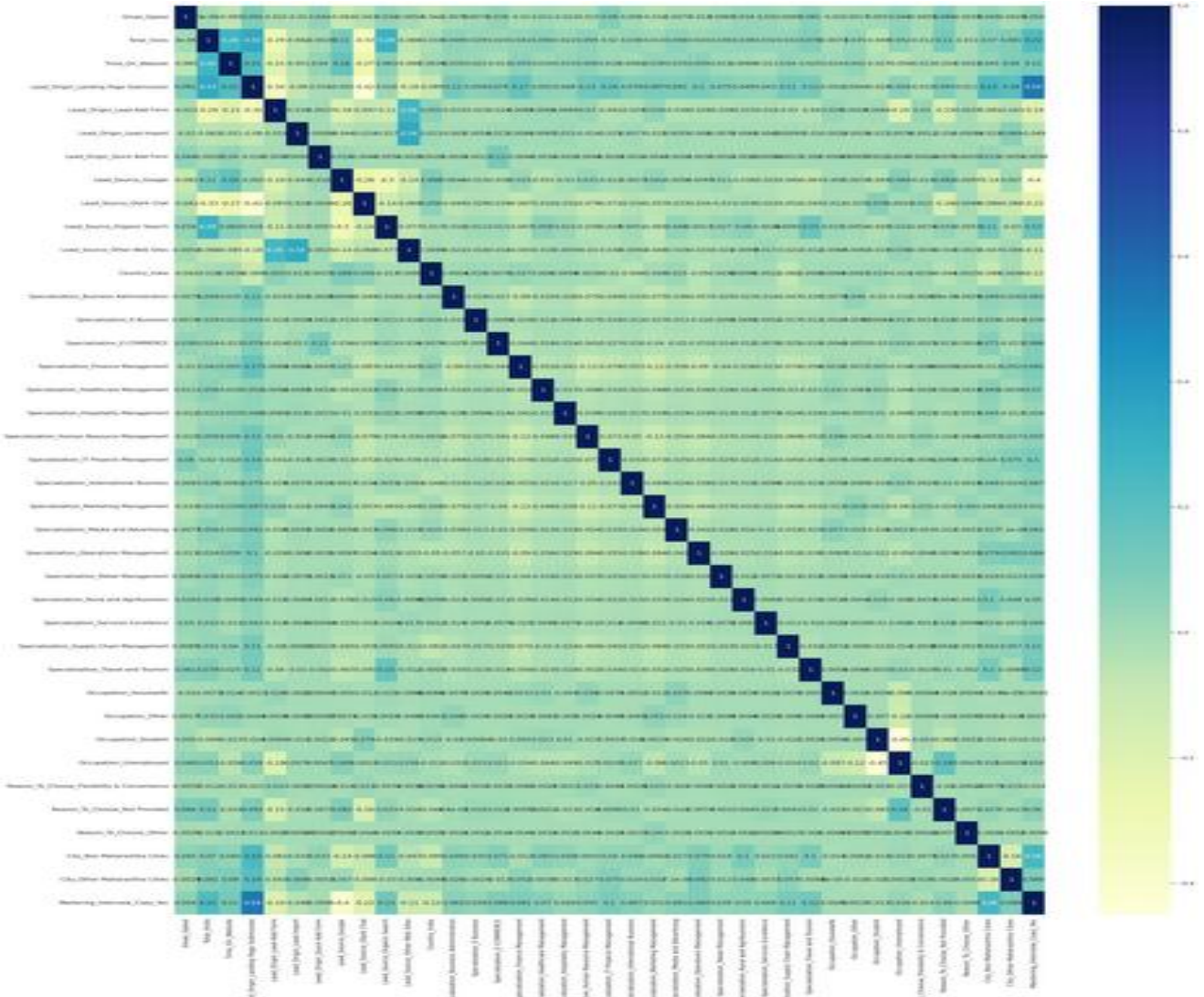
CORRELATION CHECK



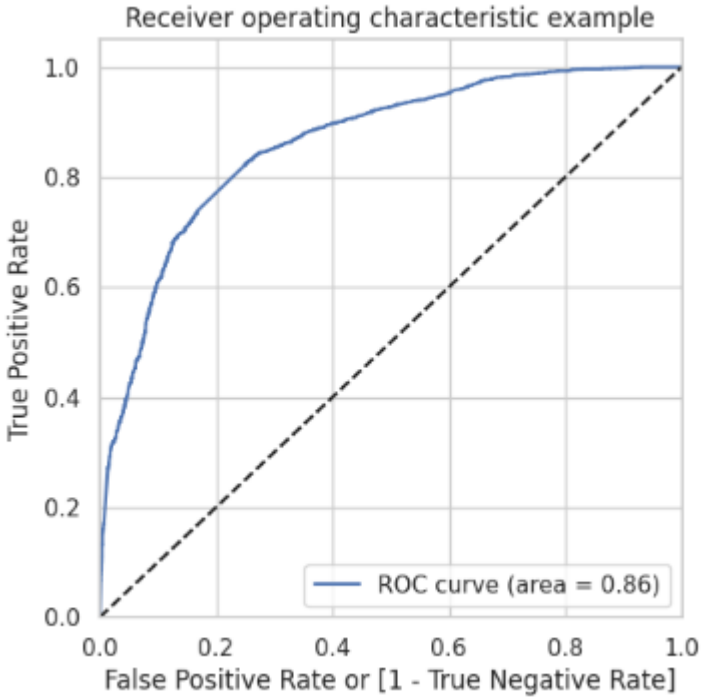
observation: 'Total Visits' & 'Page Views Per Visit' are highly correlated with each other

HEATMAP

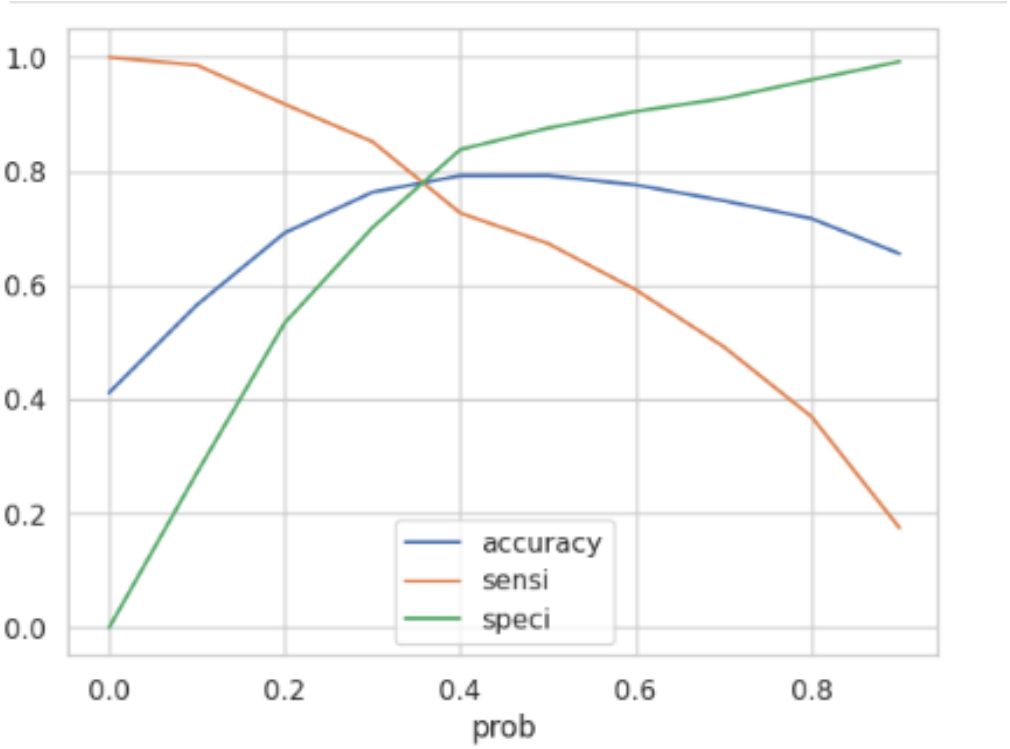




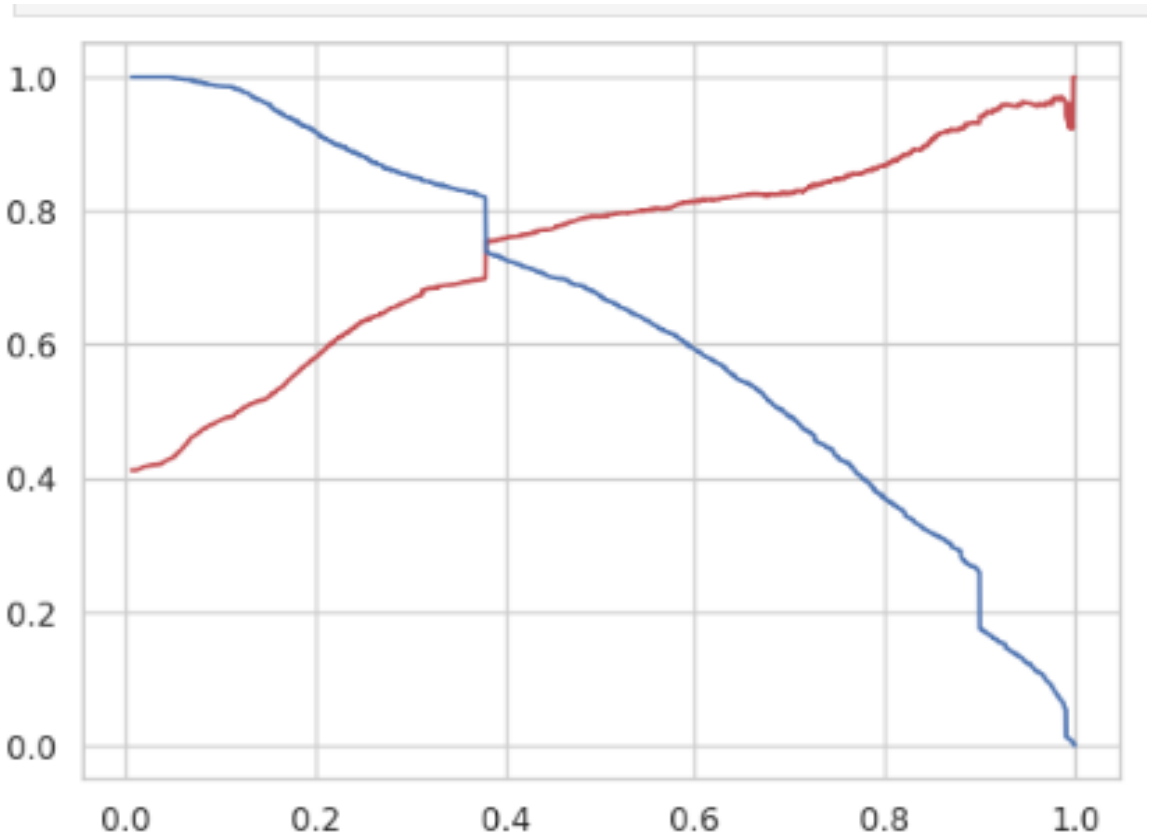
PLOTTING THE ROC CURVE



OPTIMAL CUT OFF POINT



MODEL EVALUATION



```
##Overall accuracy calculation:

metrics.accuracy_score(y_pred_final.Converted,y_pred_final.final_predicted)
```

```
0.7536057692307693
```

```
confusion2=metrics.confusion_matrix(y_pred_final.Converted,y_pred_final.final_predicted )
confusion2
```

```
array([[1018,  436],
       [ 179,  863]])
```

```
TP=confusion2[1,1]
TN=confusion2[0,0]
FP=confusion2[0,1]
FN=confusion2[1,0]
```

```
##Overall sensitivity calculation

TP / float(TP+FN)
```

```
0.8282149712092131
```

```
##Overall Specificity calculation

TN / float(TN+FP)
```

```
0.7001375515818432
```

THANK YOU