Lead scoring case study

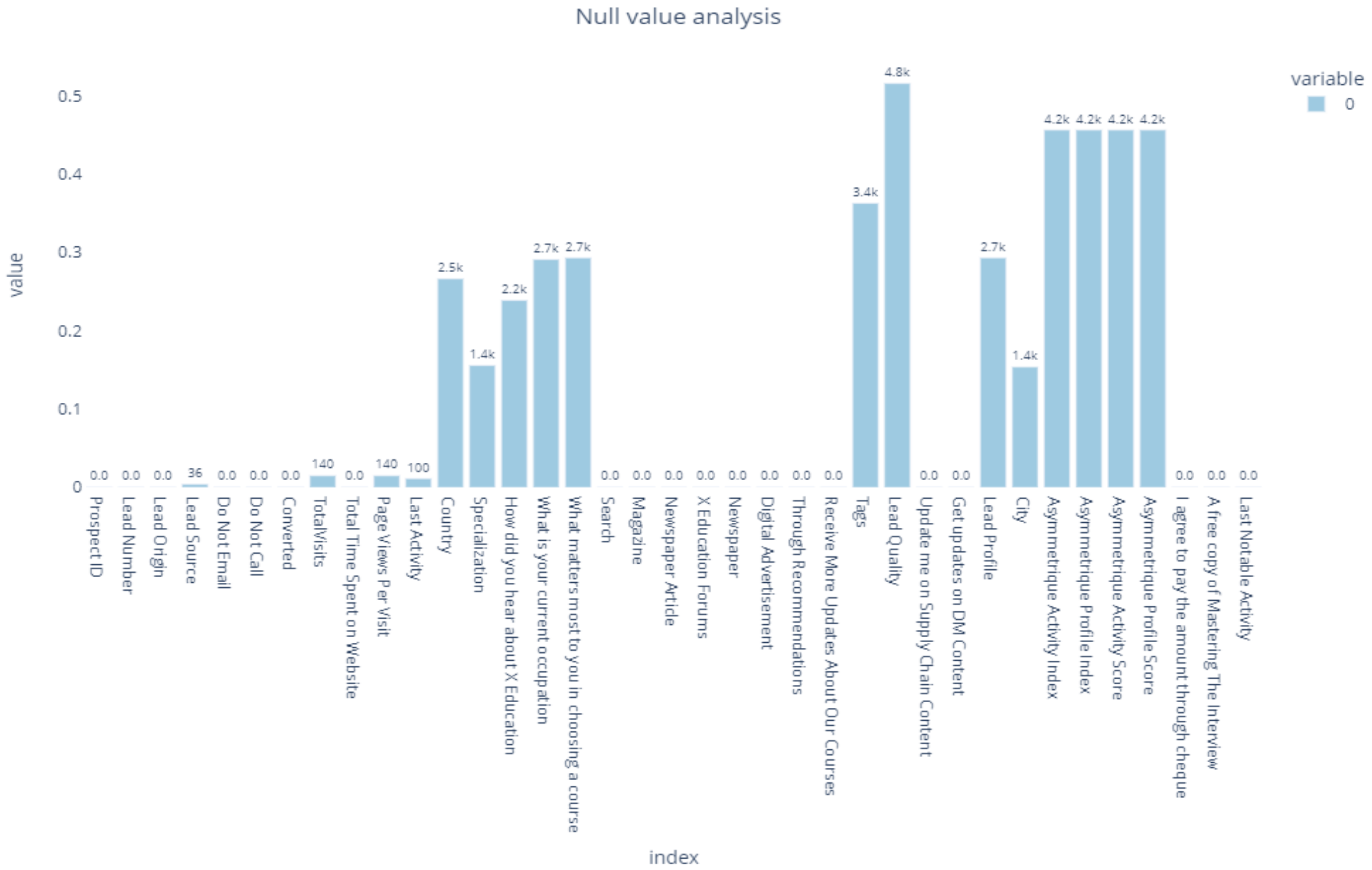
Problem Statement

An schooling business enterprise named X Education sells on-line publications to enterprise professionals. On any given day, many experts who are fascinated in the publications land on their internet site and browse for courses.

The corporation markets its guides on a number of web sites and search engines like Google. Once these humans land on the website, they may browse the publications or fill up a structure for the path or watch some videos. When these humans fill up a shape offering their electronic mail tackle or cellphone number, they are categorized to be a lead. Moreover, the agency additionally receives leads via previous referrals. Once these leads are acquired, personnel from the income group begin making calls, writing emails, etc. Through this process, some of the leads get transformed whilst most do not. The regular lead conversion charge at X schooling is round 30%.

Now, though X Education receives a lot of leads, its lead conversion charge is very poor. For example, if, say, they gather a hundred leads in a day, solely about 30 of them are converted. To make this procedure greater efficient, the enterprise desires to discover the most manageable leads, additionally recognized as ‘Hot Leads’. If they efficiently perceive this set of leads, the lead conversion charge need to go up as the income group will now be focusing greater on speaking with the achievable leads as a substitute than making calls to absolutely everyone

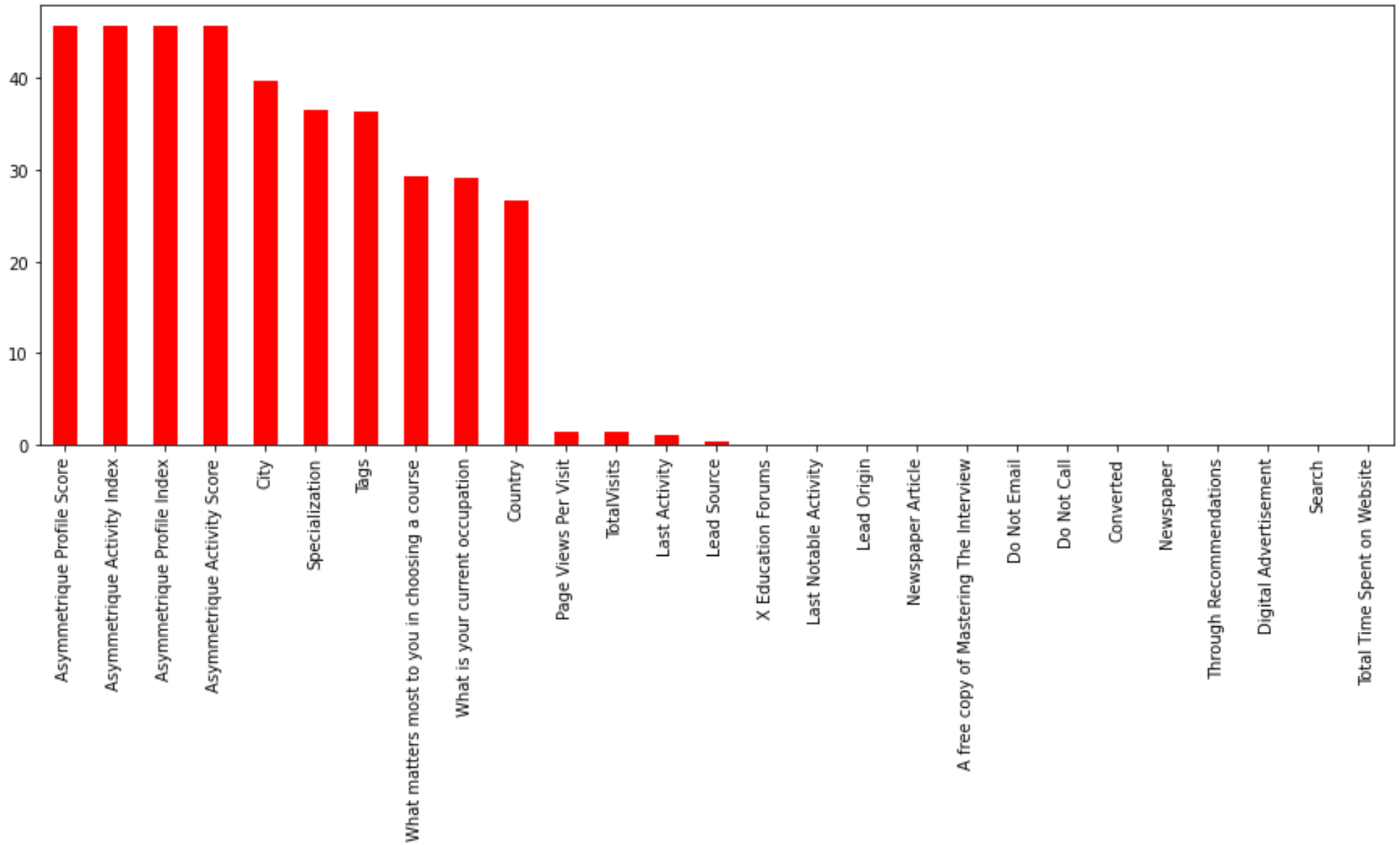
Missing values and Data cleaning



We dropped the missing values which having the percentage greater than 50 And we successfully dealt with the missing values having percentage less than our threshold

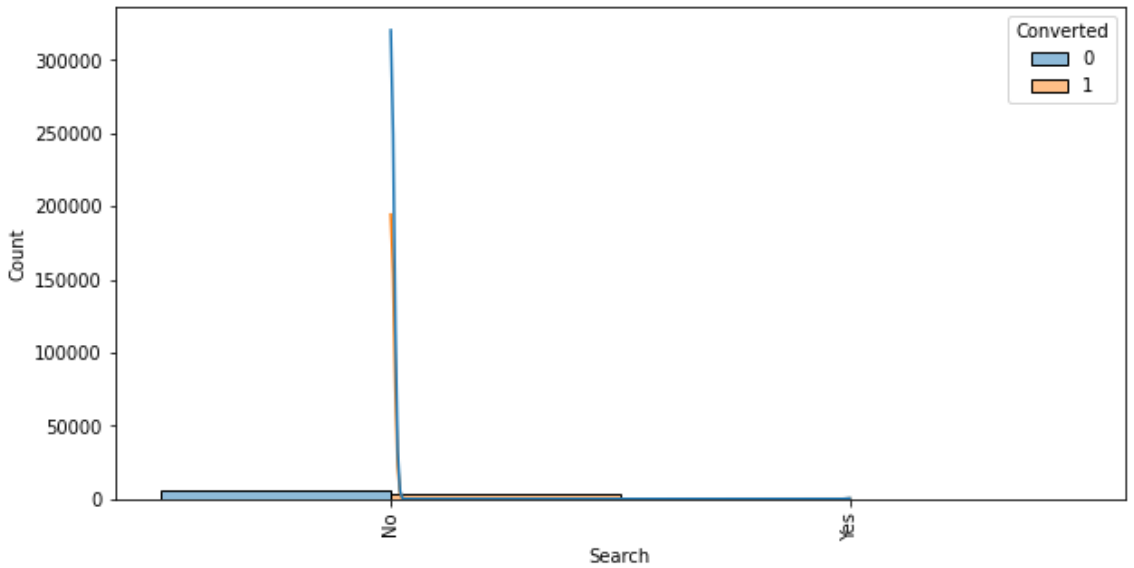
Majority of the customers has not seen the add at X education forum , Newspaper Article , Not searched , Not seen digital advertisement, Through Recommendation, or Newspaper Customers who wants updates via call, Email are more likely to be a lead

This are the features remained after dropping the missing values

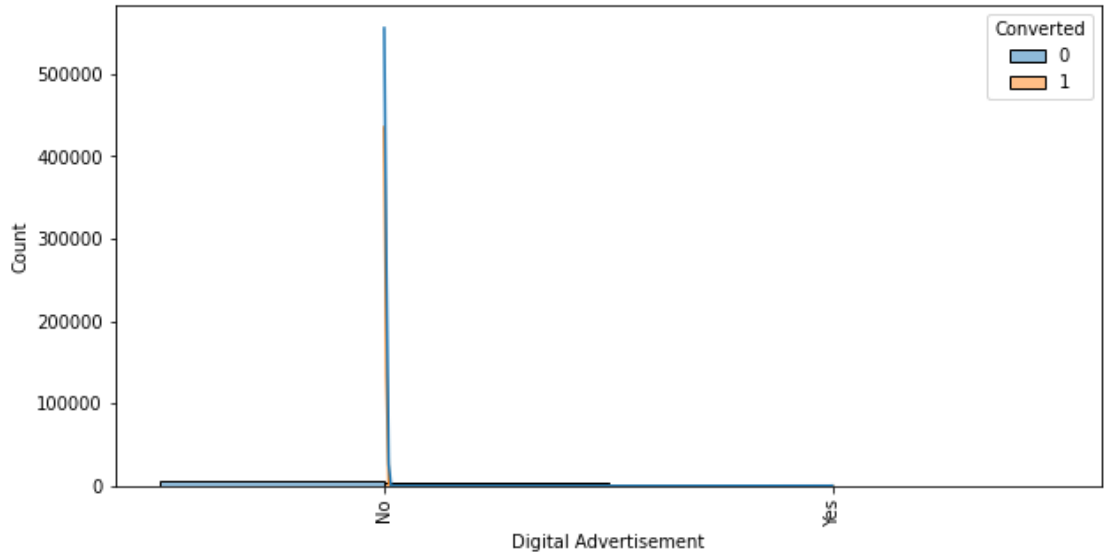


EDA

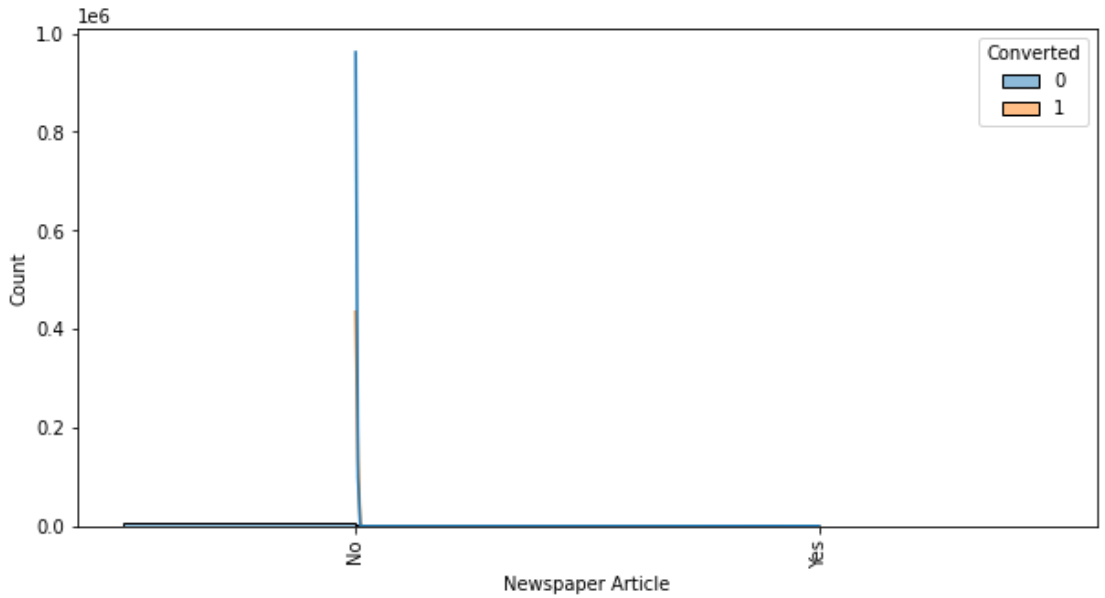
Search



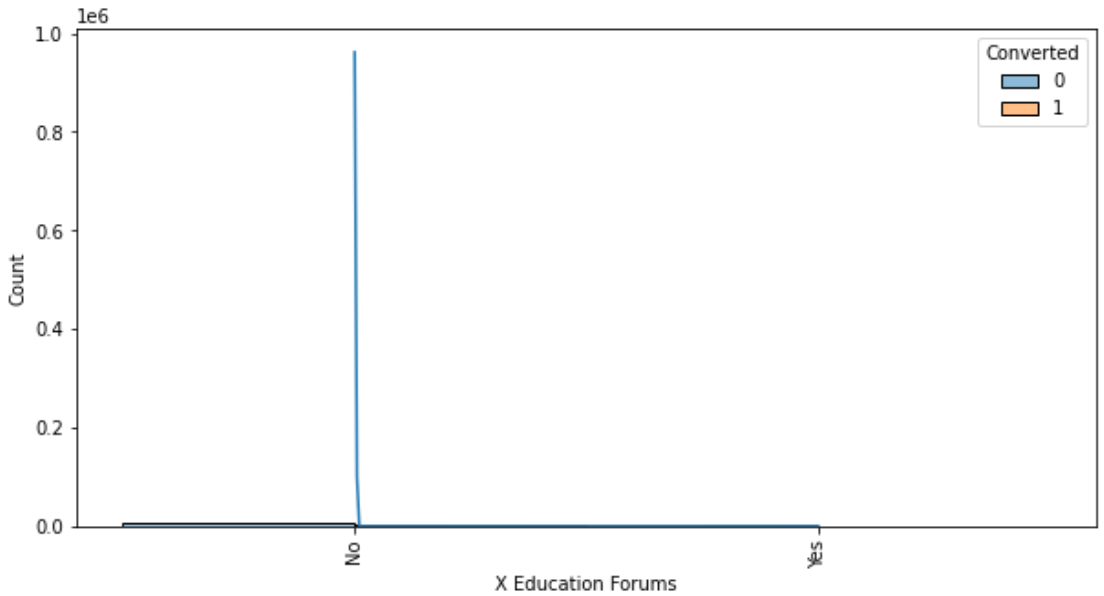
Digital Advertisement

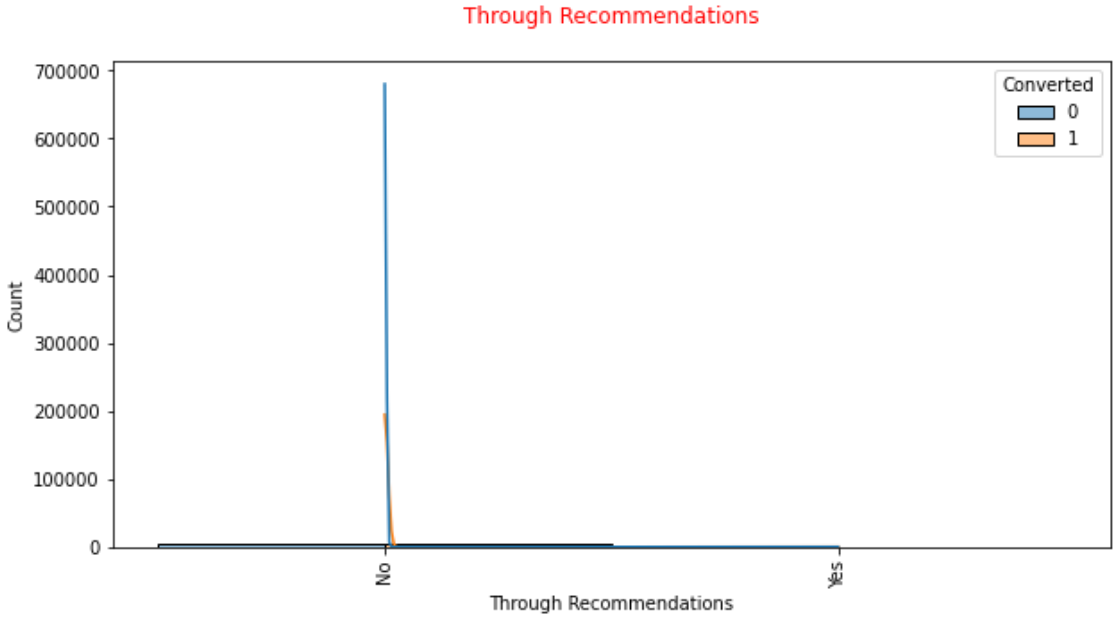


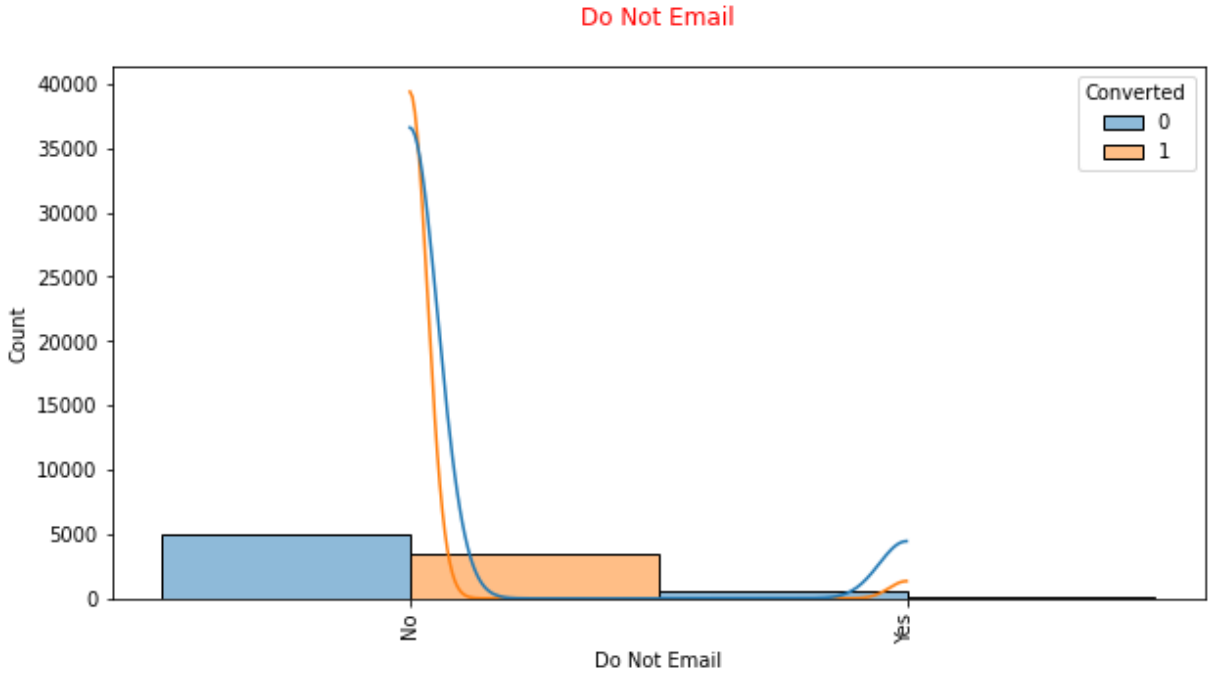
Newspaper Article

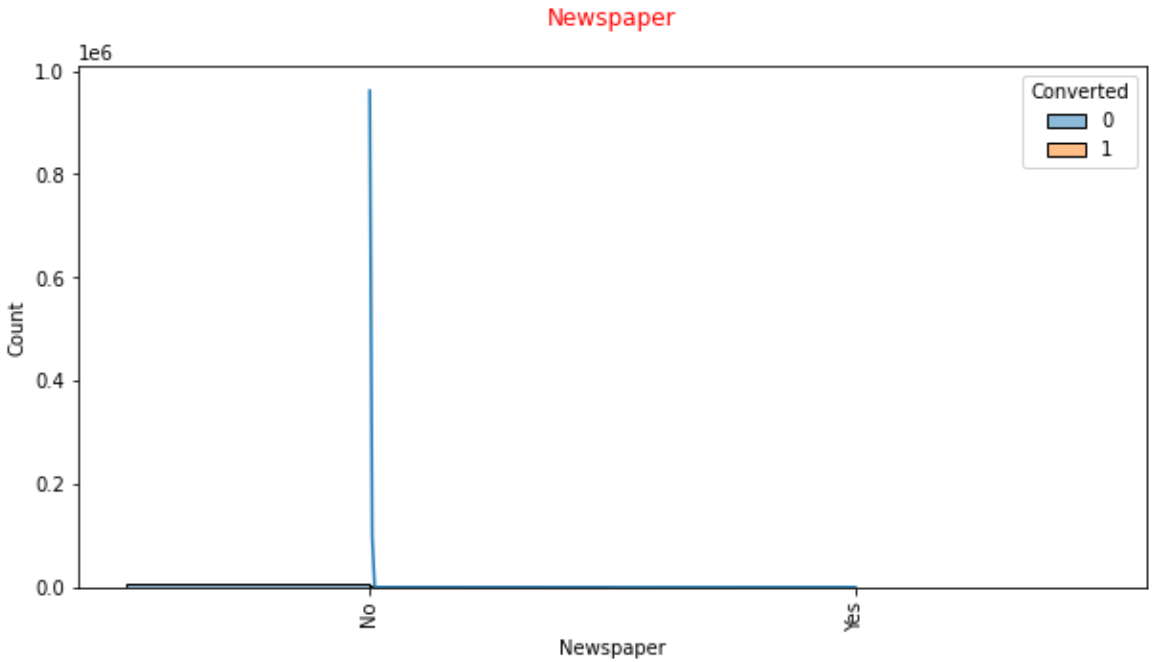


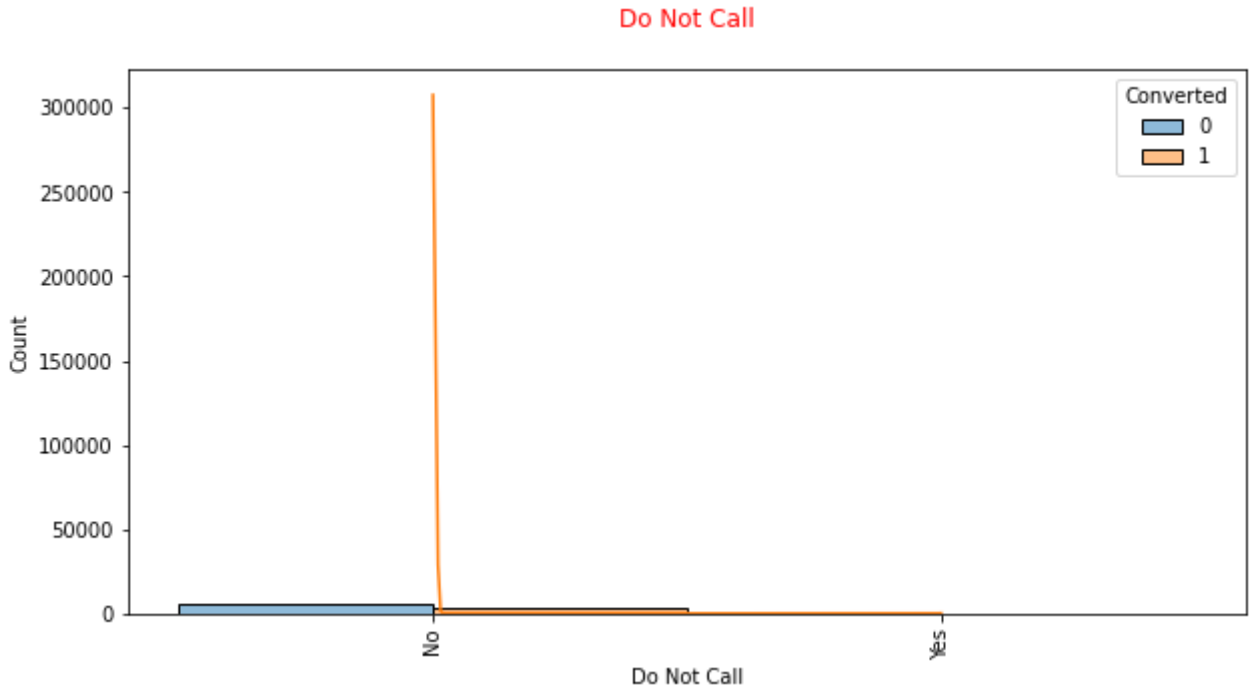
X Education Forums



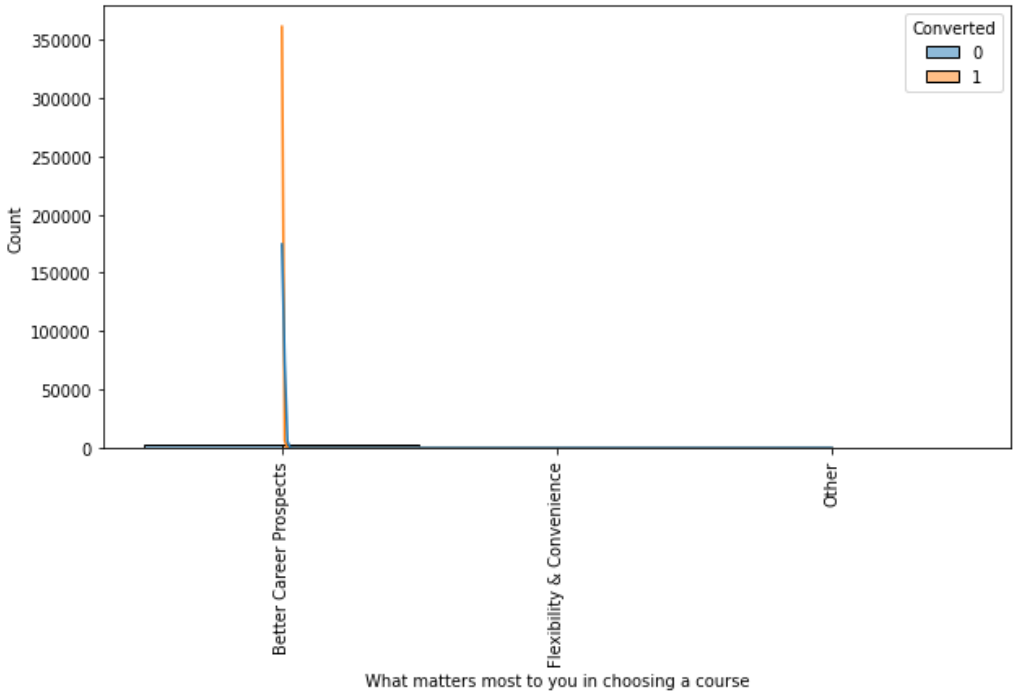

Through Recommendations
Converted
0
1
Count
700000
600000
500000
400000
300000
200000
100000
0
No
Yes
Through Recommendations


Do Not Email
Converted
0
1
Count
40000
35000
30000
25000
20000
15000
10000
5000
0
No
Yes
Do Not Email

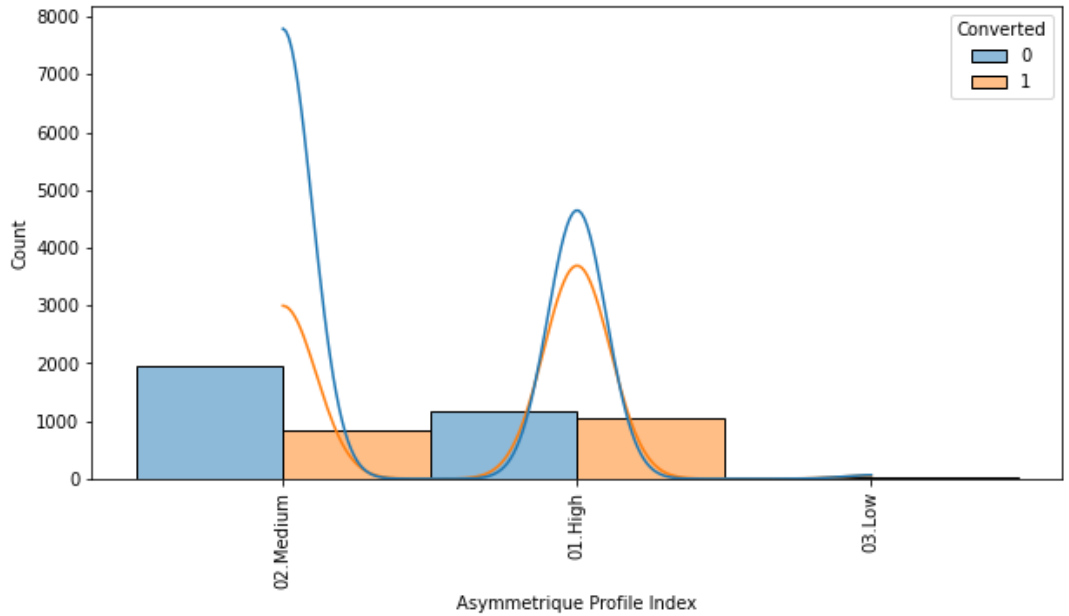

Newspaper
1e6
Converted
0
1
Count
1.0
0.8
0.6
0.4
0.2
0.0
No
Yes
Newspaper


Do Not Call
Converted
0
1
Count
300000
250000
200000
150000
100000
50000
0
No
Yes
Do Not Call

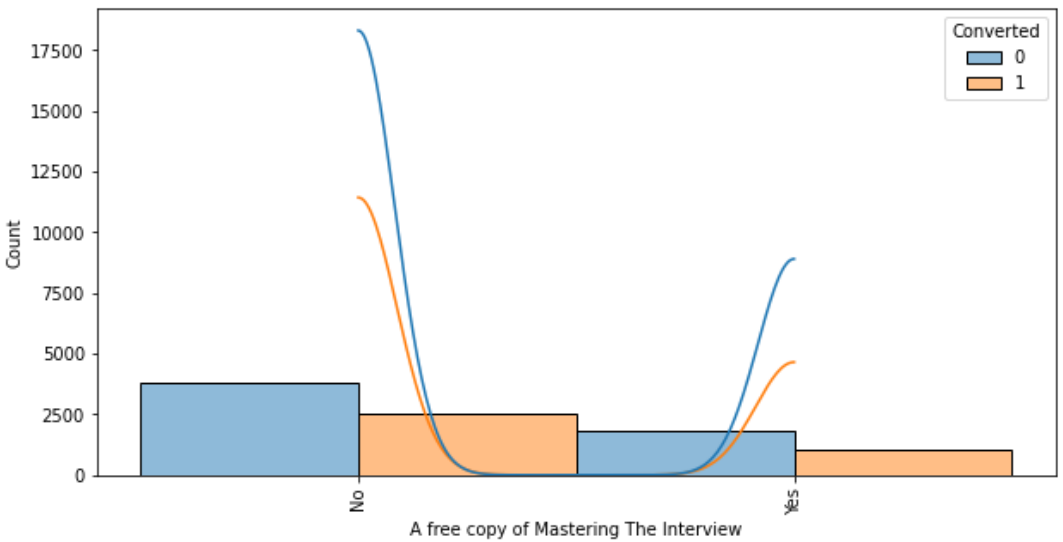
What matters most to you in choosing a course



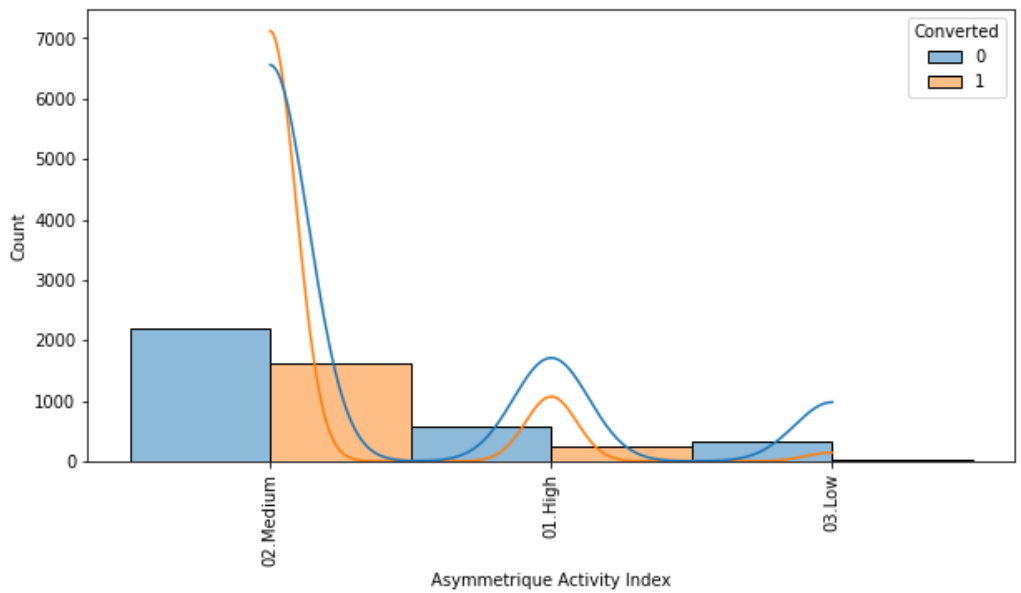
Asymmetrique Profile Index



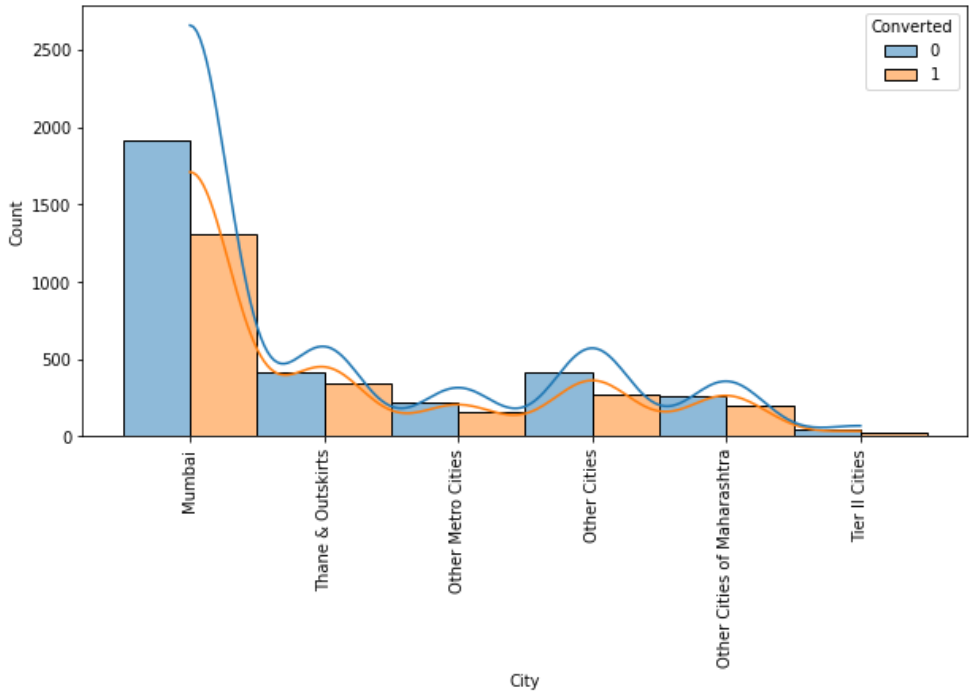
A free copy of Mastering The Interview



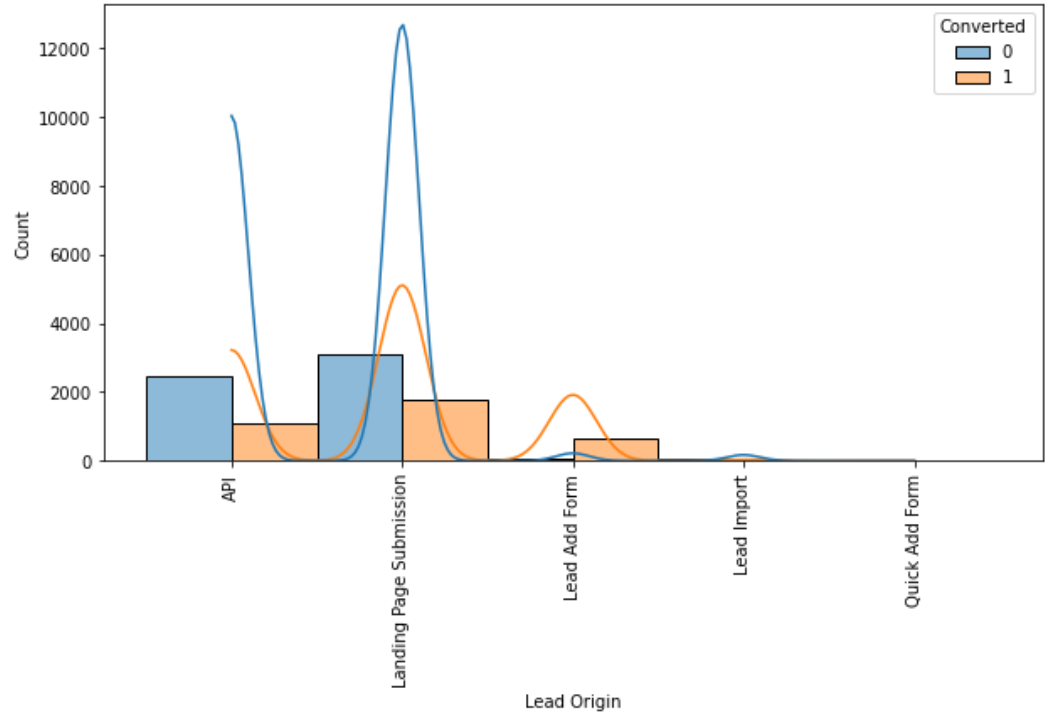
Asymmetrique Activity Index



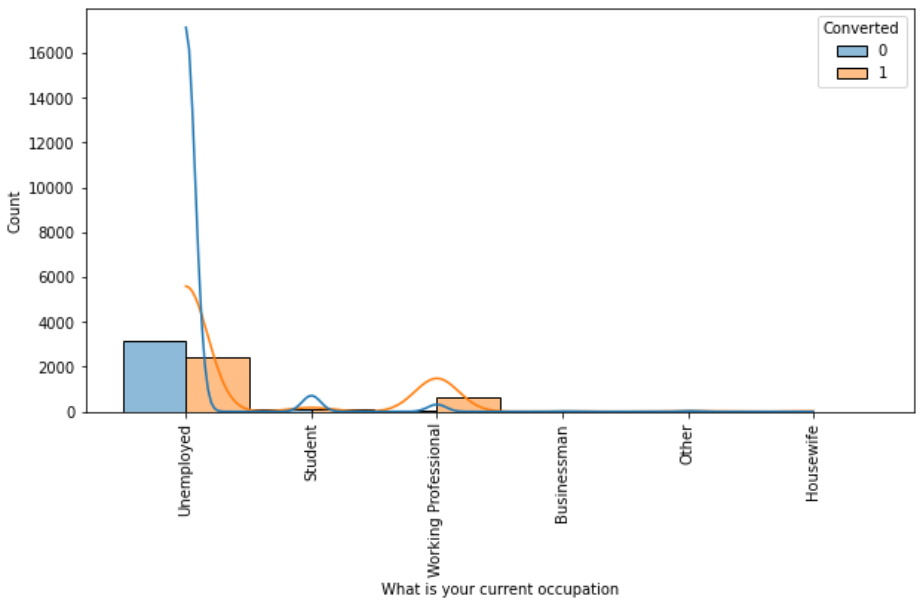
City



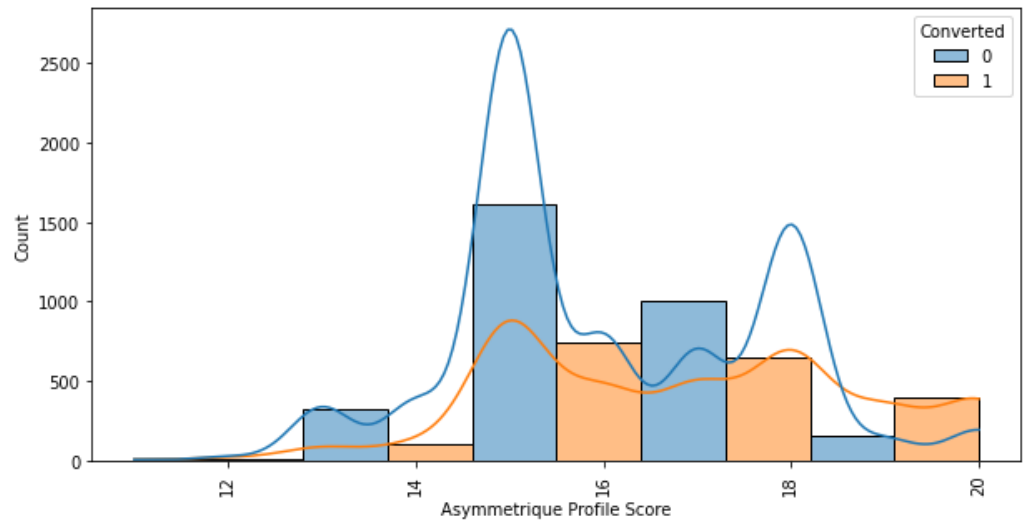
Lead Origin



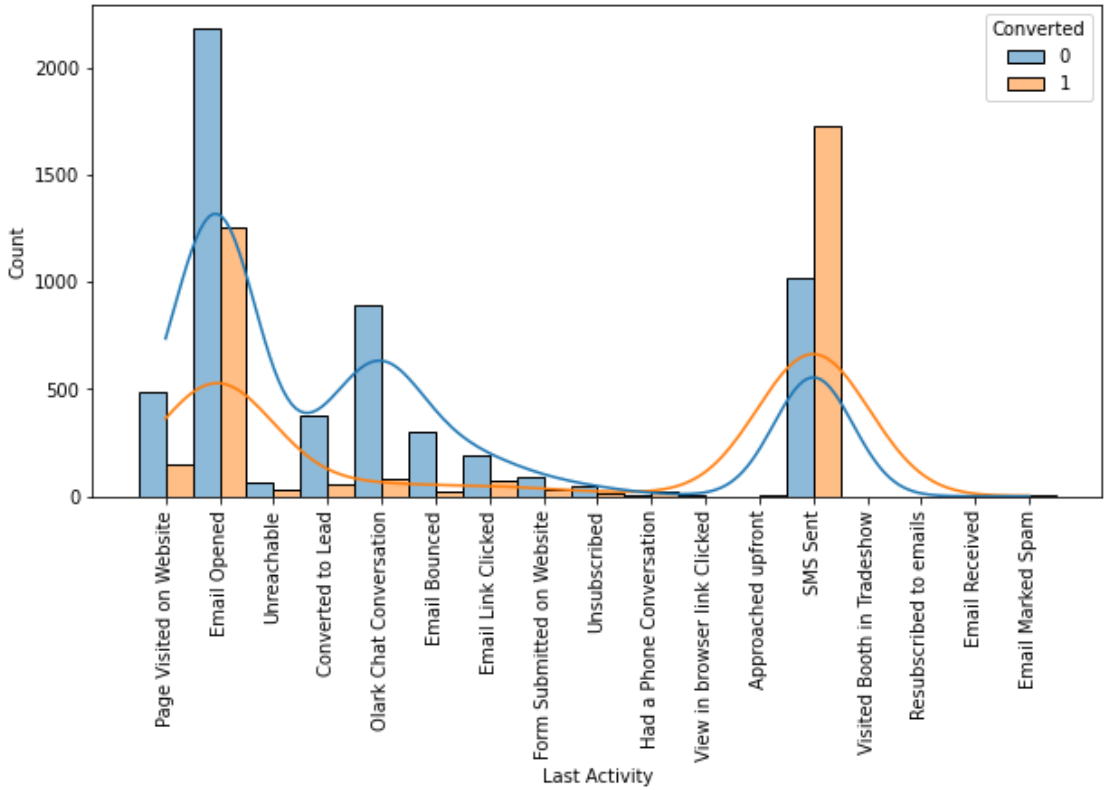
What is your current occupation



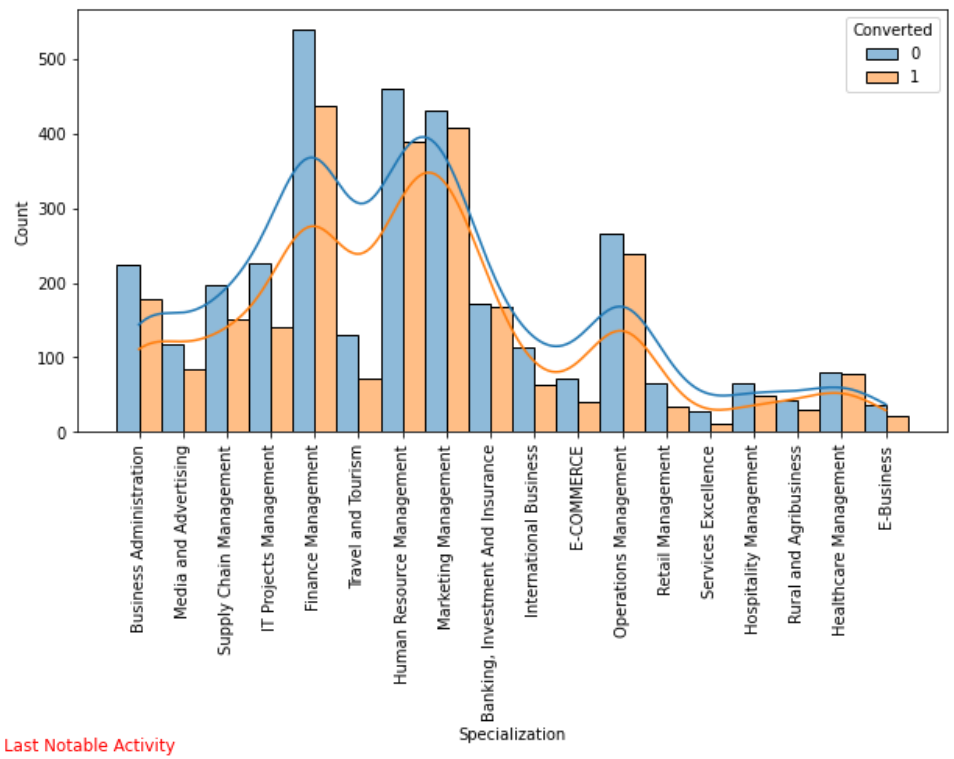
Asymmetrique Profile Score



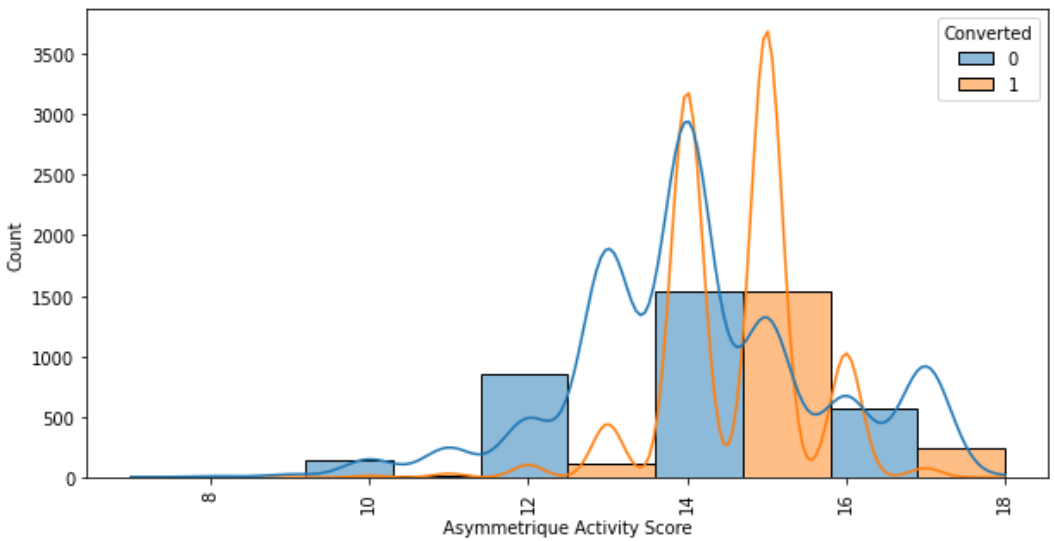
Last Activity



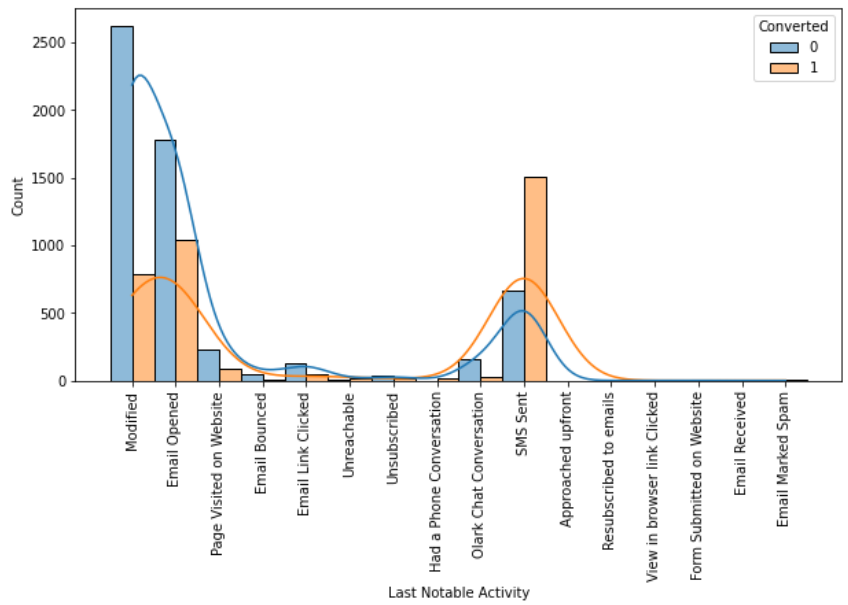
Specialization



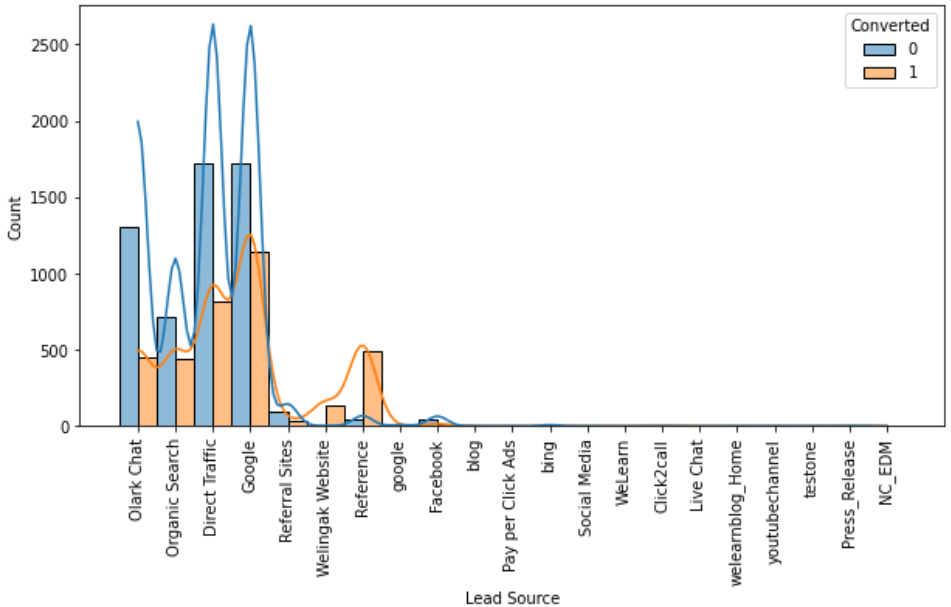
Asymmetrique Activity Score



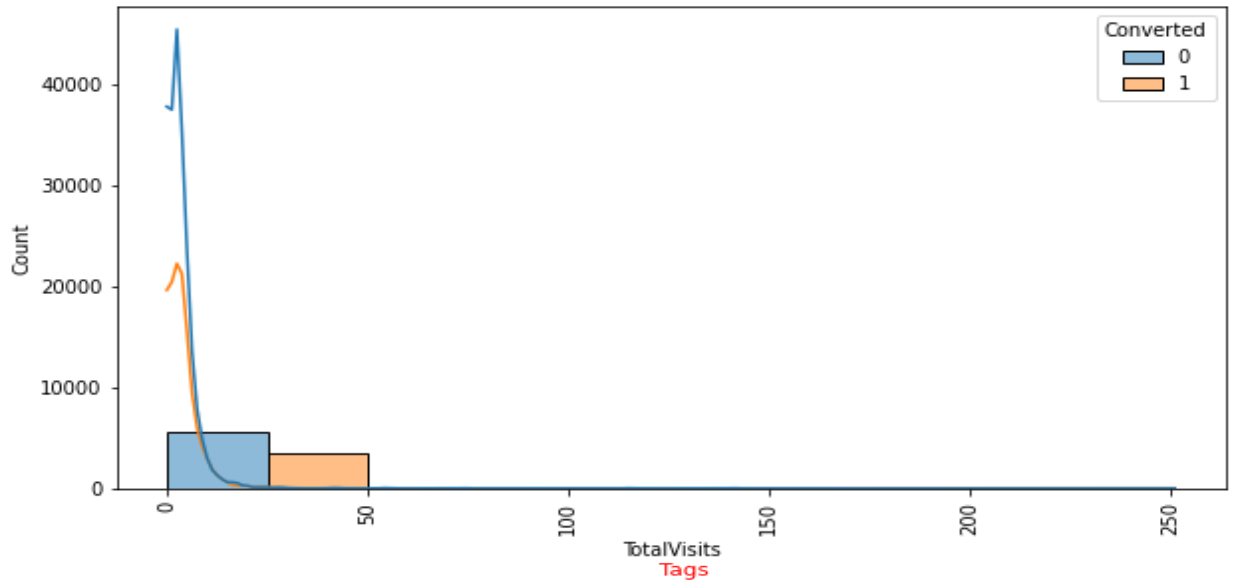
Last Notable Activity



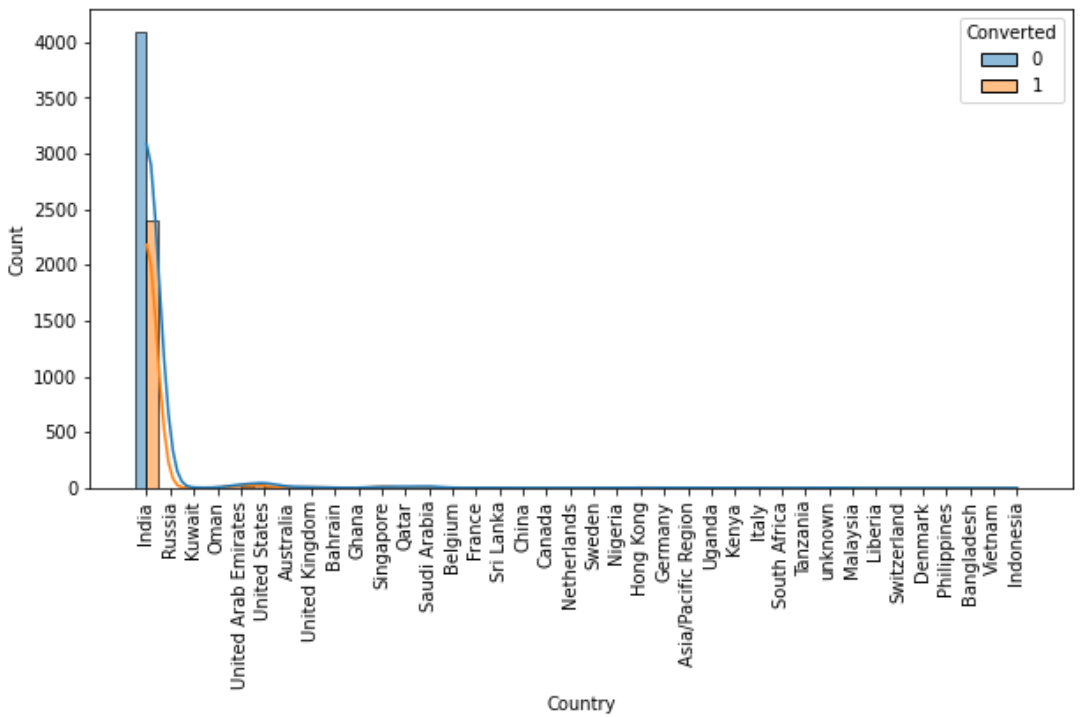
Lead Source



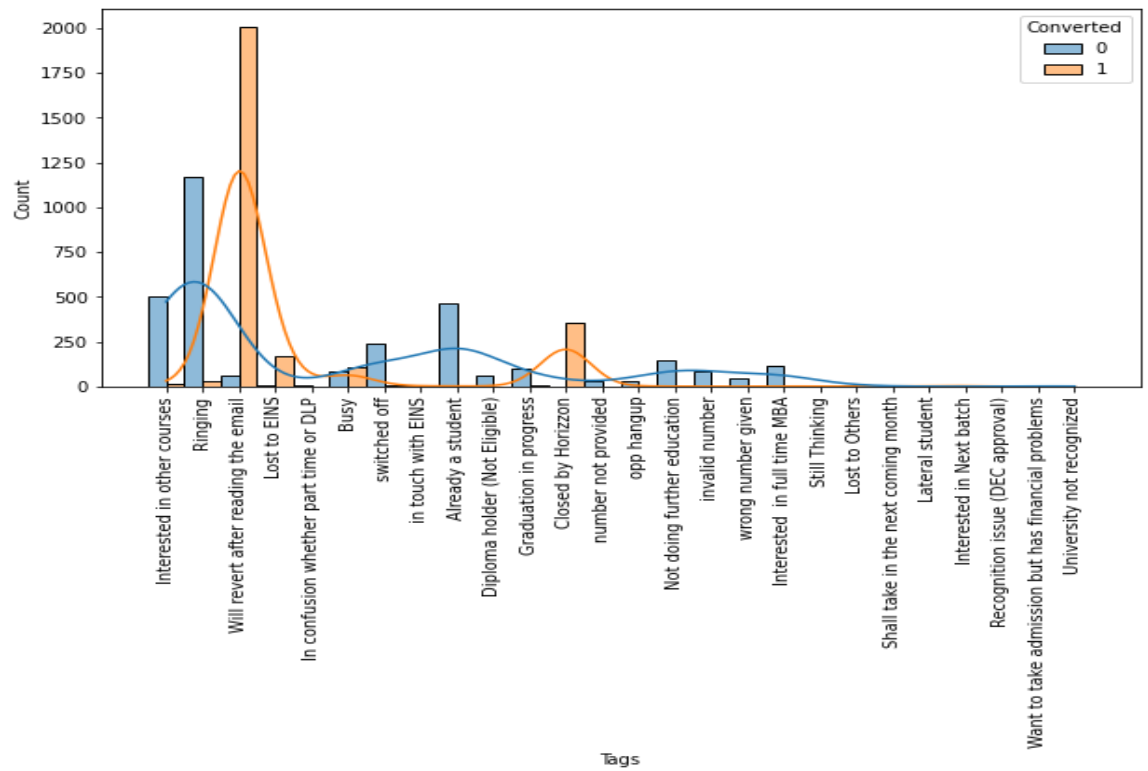
TotalVisits



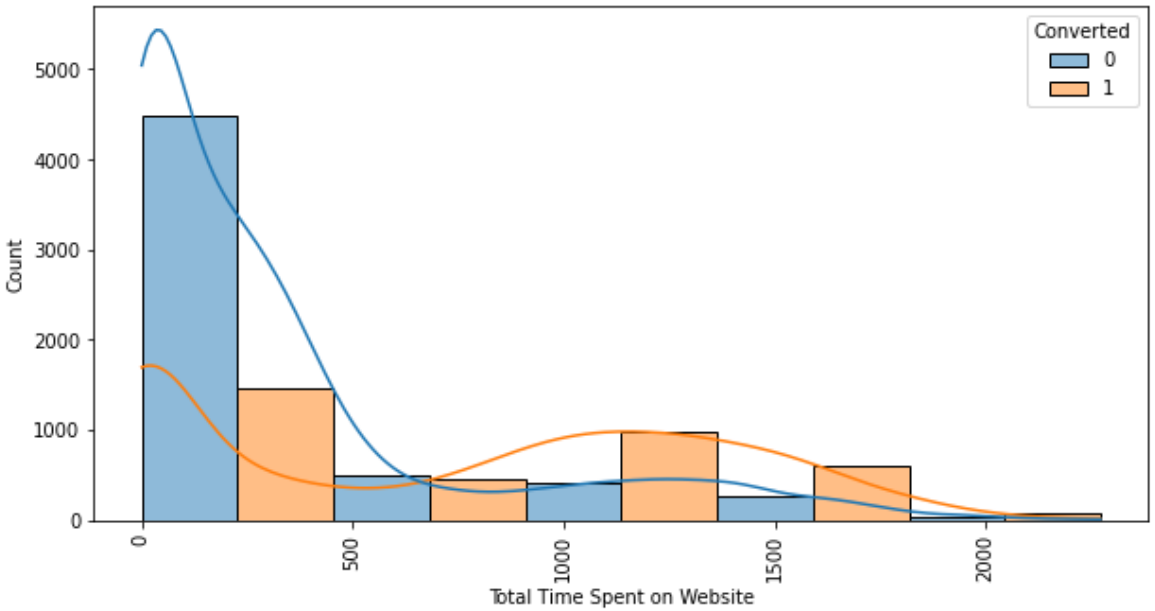
Country



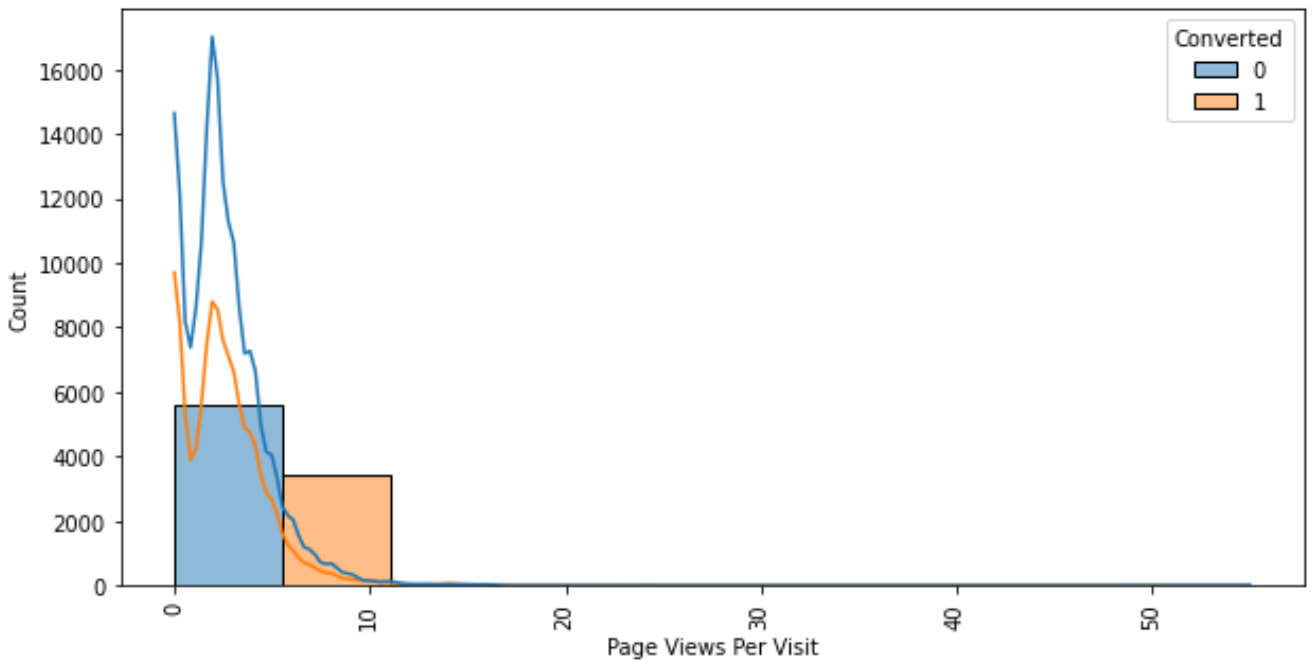
Tags



Total Time Spent on Website



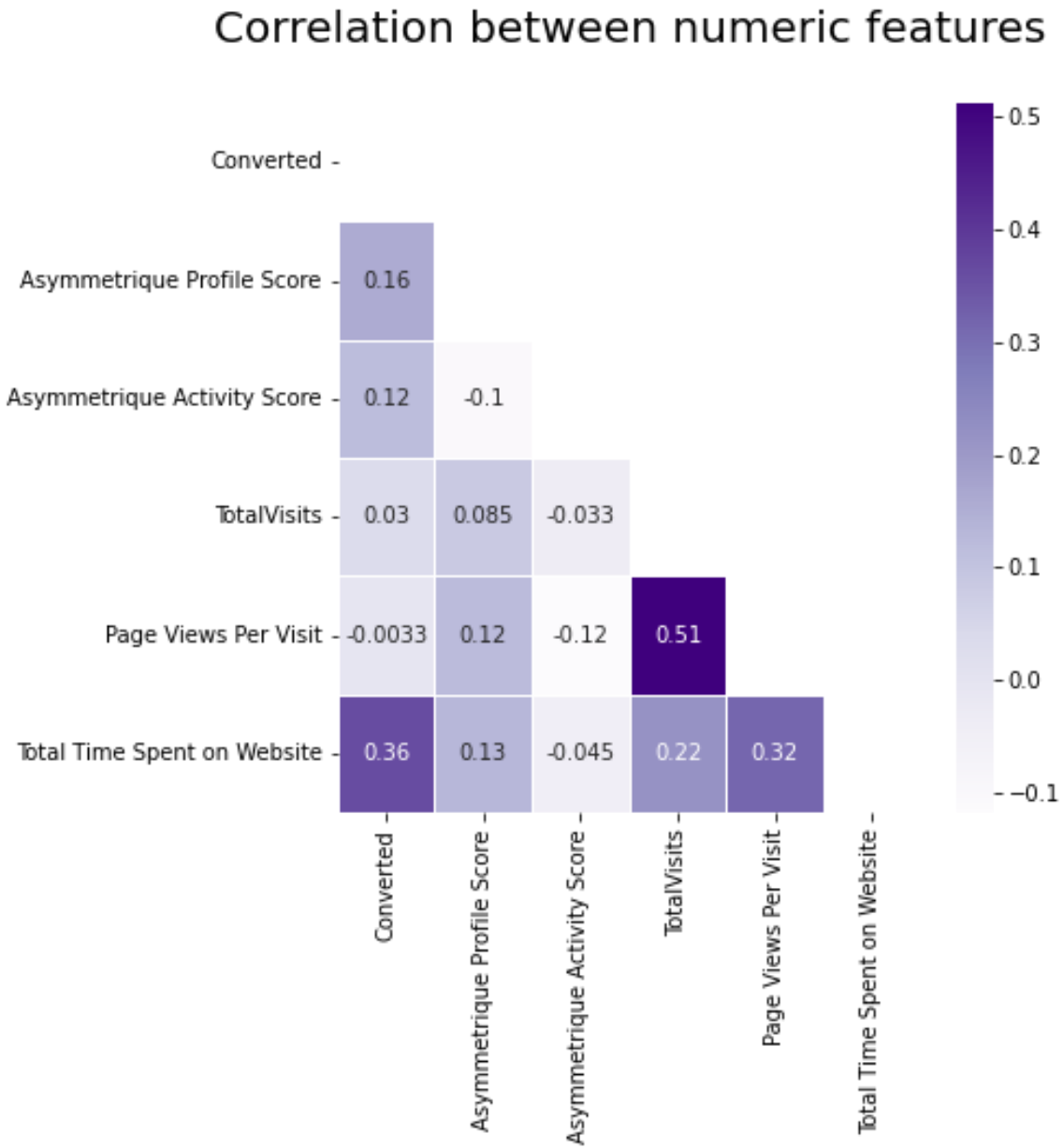
Page Views Per Visit



leaders show following behavior :

- They do not want a free copy of mastering the interview
- Medium asymmetries activity index
- high asymmetries profile index
- They want better career prospectus
- They might have lead quality and Landing page submission
- They are unemployed and from Mumbai and thane and belong to India
- They heard from online search and students from same school
- have asymmetries profile score ranging from 15 -18,Asymmetrique activity score ranging from 14 - 16
- they recently did the modification , opened their email and sent the SMS and probably did the same as a last activity
- majority of them are specialized in projects management ,finance management ,human resource management ,marketing management.
- they comes from sectors like banking , investment and insurance and operation management
- majority lead sources from google ,direct traffic and organic search
- they do revert after reading the email that shows their interest
- they visit the website at least 50 times they visit the page at-least 10 times and spend at least 500 s on the website

This Heatmap provides us the inter relation between the variables



Class imbalance dealing

We used SMOTETomek to deal with the class imbalance and we successfully dealt with that.

Before

```
0    5679
1    3561
Name: Converted, dtype: int64
```

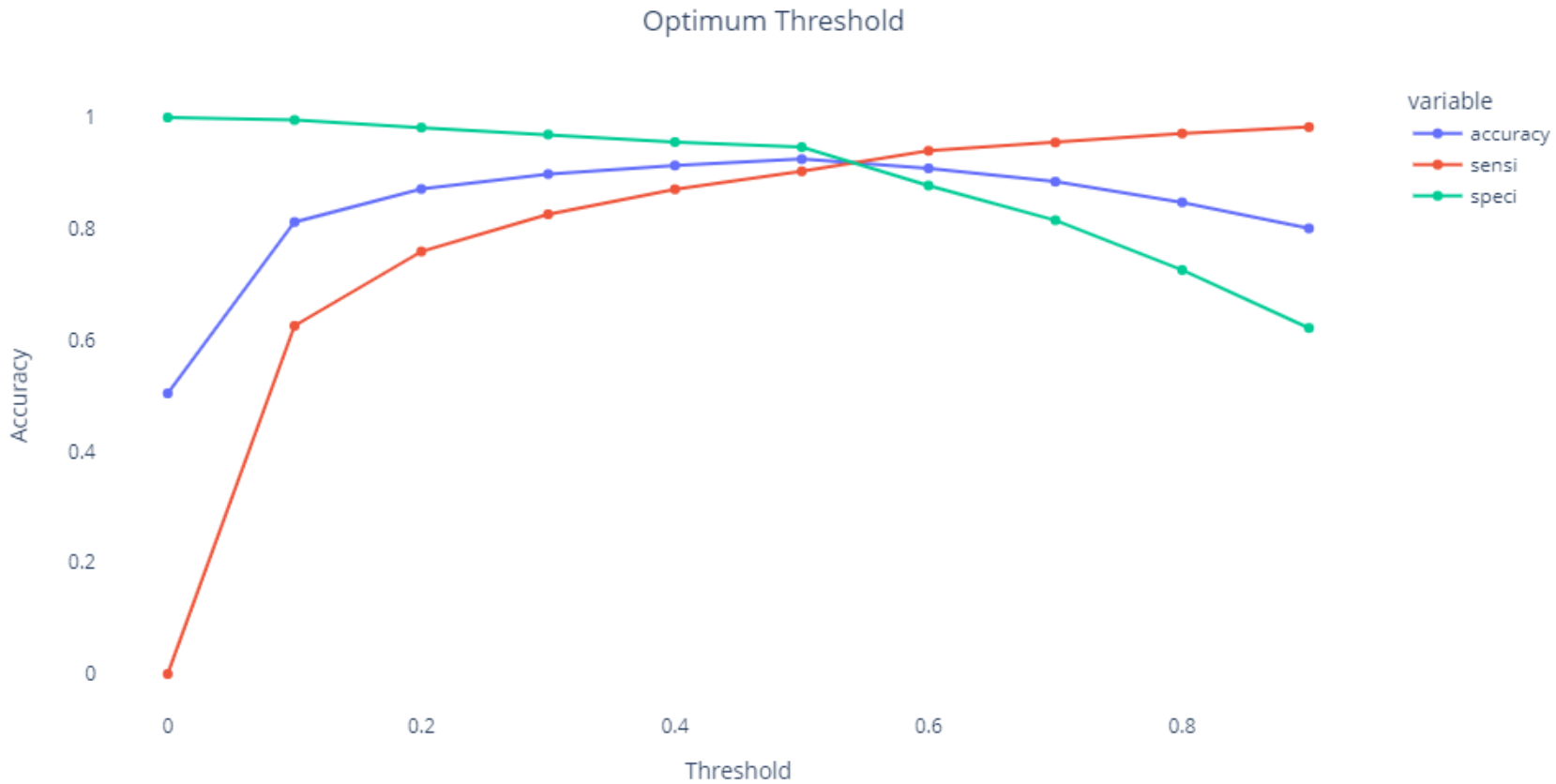
After

```
0    5470
1    5470
Name: Converted, dtype: int64
```

Model building and prediction

- We scaled the model using min max scaling and then performed the split
- For the splitting we split the original data frame into train and validation data frame and after that we even split the tarin data into test data as of 80,20 split
- We used RFE for selecting the best possible features and after from that we dropped the features one by one who had the p value greater than 0.05
- We used GLM from statsmodels library for the model building

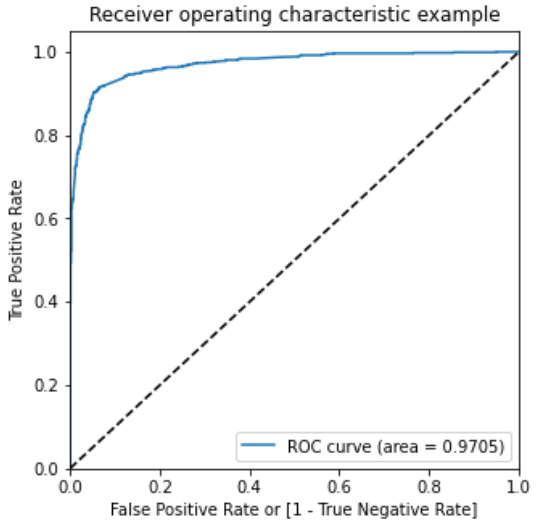
thresh	accuracy	sensi	speci
0.5	0.925186	0.903226	0.946772
0.4	0.913764	0.870968	0.955832
0.6	0.908624	0.940092	0.877690
0.3	0.898344	0.826037	0.969422
0.7	0.884637	0.955069	0.815402
0.2	0.871502	0.759217	0.981880
0.8	0.847516	0.971198	0.725934
0.1	0.812107	0.625576	0.995470
0.9	0.800685	0.982719	0.621744
0.0	0.504283	0.000000	1.000000



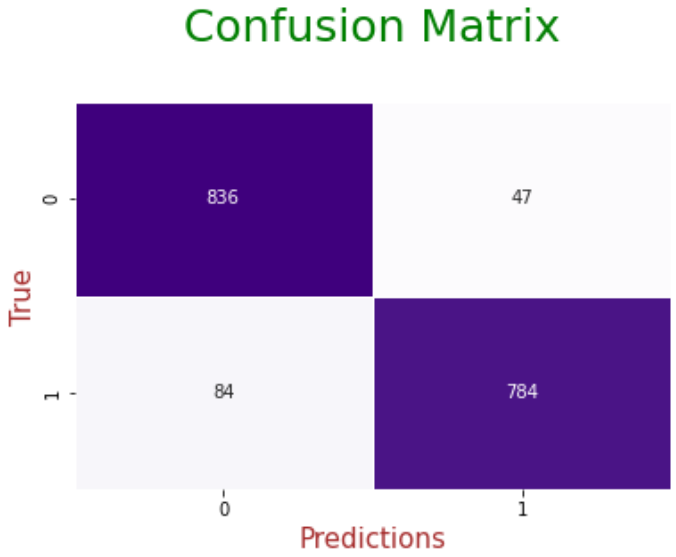
We selected best possible threshold with the help of above graph to divide the predictions into two parts

Model evaluation

- On test data we got accuracy of 92.52 percentage.
- Further we did model decomposition with the help of PCA and got accuracy of around 97.05 percentage on the test data
- On the validation data we got the accuracy of 89.58 percent and after applying the PCA we got the accuracy of around 95.61 percentage.



	precision	recall	f1-score	support
0.0	0.9087	0.9468	0.9273	883
1.0	0.9434	0.9032	0.9229	868
accuracy			0.9252	1751
macro avg	0.9261	0.9250	0.9251	1751
weighted avg	0.9259	0.9252	0.9251	1751



Conclusion and Important features

this are the top three positively correlated features :

	Features
Total Time Spent on Website	0.373628
Tags_Will revert after reading the email	0.310597
Lead Origin_Lead Add Form	0.280926

this are the top three negatively correlated features :

	Features
Last Activity_Olark Chat Conversation	-0.260607
Last Notable Activity_Modified	-0.310238
Tags_Ringing	-0.322744

This are the features that are important for the prediction of the model

Overall we have got a good model