Lead Scoring Assignment

Team Members
- Manohar Datta Gundu
- Prathish C Suvarna
- Devanshi Gupta

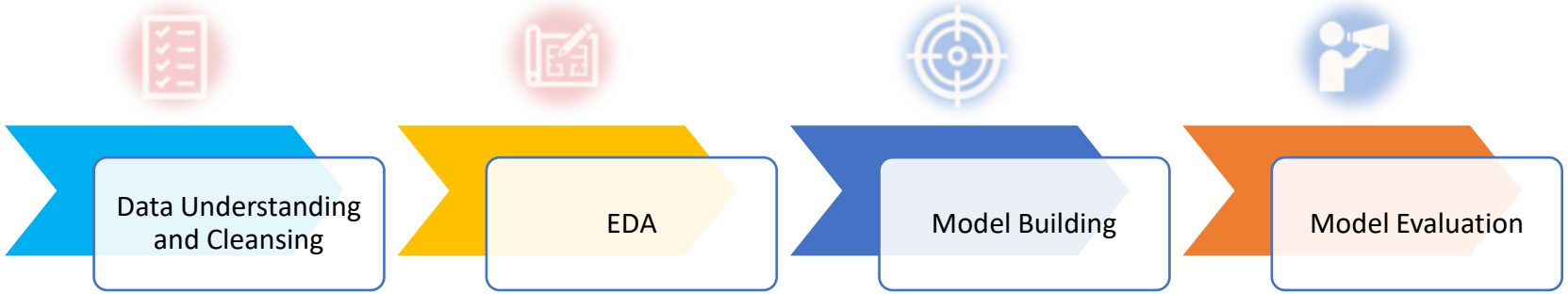
PROBLEM STATEMENT

- X Education is an online course provider targeting industry professionals. Prospective clients discover their courses through various channels like Google. Upon visiting the website, individuals may browse courses, fill out forms, or watch videos.
- Leads are generated when users provide contact information, either through website forms or referrals. The sales team then contacts these leads via calls and emails.
- The average lead conversion rate for X Education is approximately 30%.

OBJECTIVE

- X Education has tasked with creating a lead scoring model to boost their lead conversion rates.
- The company aims to identify high-potential leads, termed 'Hot Leads,' with a focus on achieving an ambitious 80% lead conversion rate.
- 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.

FLOWCHART OF SOLUTION



Data Understanding and Cleansing

Total Rows: 9240

Total Columns: 37.

There are few columns with value selected, replaced it with null

cialization	How did you hear about X Education	What is your current occupation	matters most to you in choosing a course	Search	Magazine	Newspaper Article	X Education Forums	Newspaper	Adve
Select	Select	Unemployed	Better Career Prospects	No	No	No	No	No	
Select	Select	Unemployed	Better Career Prospects	No	No	No	No	No	
Business dministration	Select	Student	Better Career Prospects	No	No	No	No	No	
Media and Advertising	Word Of Mouth	Unemployed	Better Career Prospects	No	No	No	No	No	
Select	Other	Unemployed	Better Career Prospects	No	No	No	No	No	

```
#Checking for percentage of missing values
leads_data_df.isna().mean().sort_values(ascending=False)*100
```

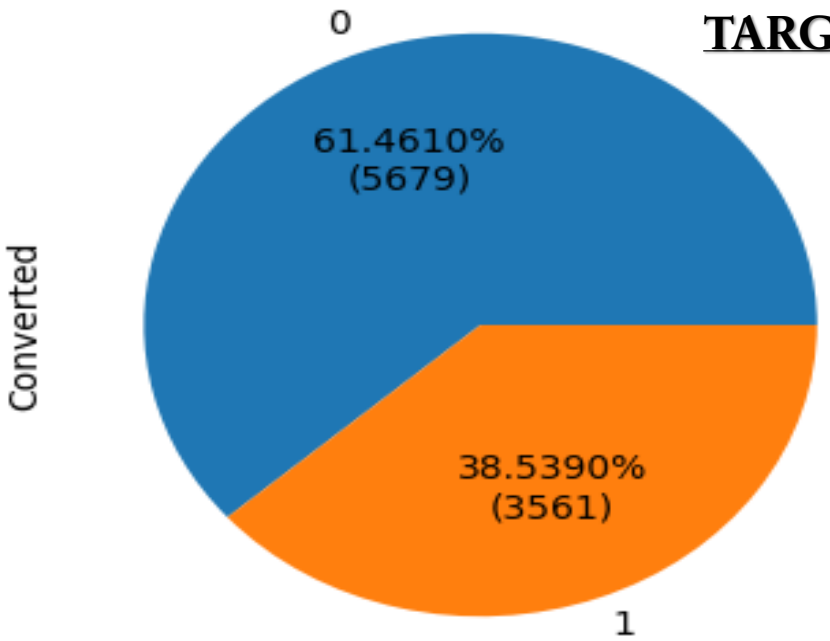
```
How did you hear about X Education      78.463203
Lead Profile                            74.188312
Lead Quality                            51.590909
Asymmetrique Profile Score              45.649351
Asymmetrique Activity Score             45.649351
Asymmetrique Activity Index             45.649351
Asymmetrique Profile Index              45.649351
City                                    39.707792
Specialization                          36.580087
```

Handling Nulls

- Dropped columns with missing values percentage>40%
- Imputed the nulls in categorical columns with Mode.
- Imputed the nulls in continuous columns with median

EDA

TARGET DATA IMBALANCE RATIO

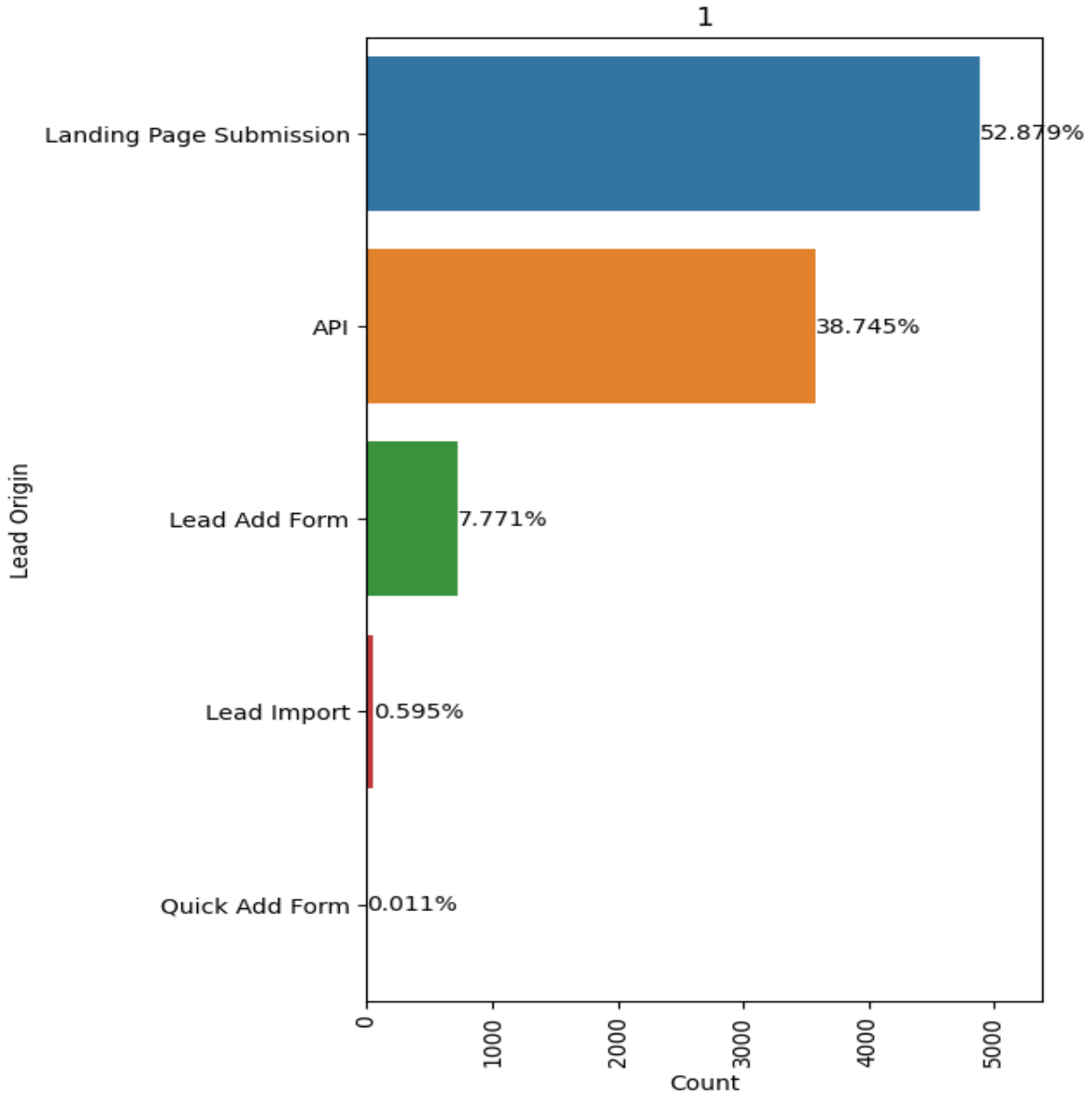


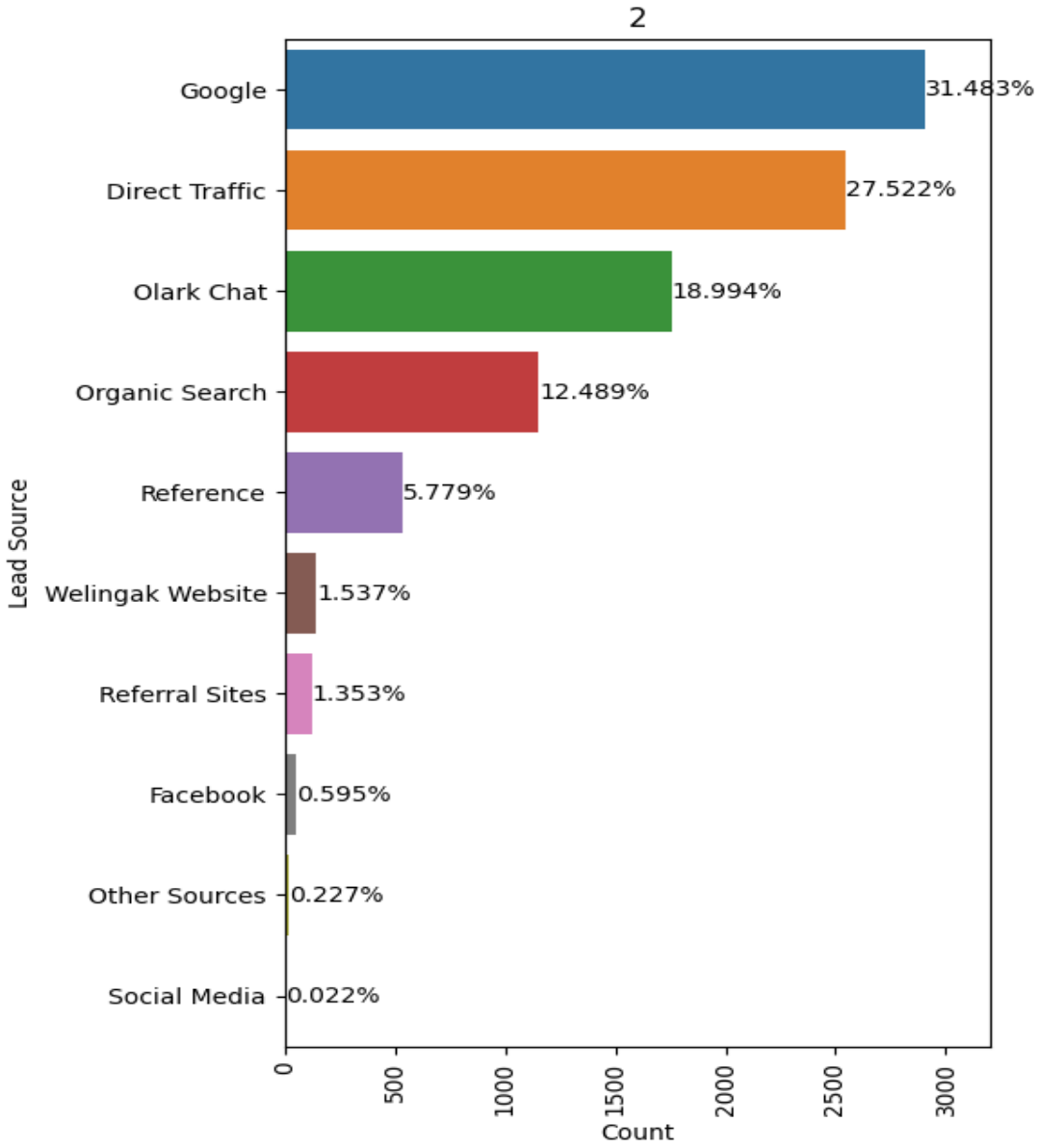
- Total Converted: 3561
- Total non-Converted: 5679
- Imbalance_Ratio: 0.6270470153195985

CATEGORICAL UNIVARIATE ANALYSIS

- Lead Source
- Lead Origin
- Do not Email
- Do not Call
- Country
- City
- Last Activity
- Specialization
- What is your current occupation
- What matters most to you in choosing a course
- Tags
- A free copy of mastering the interview
- Last Notable Activity

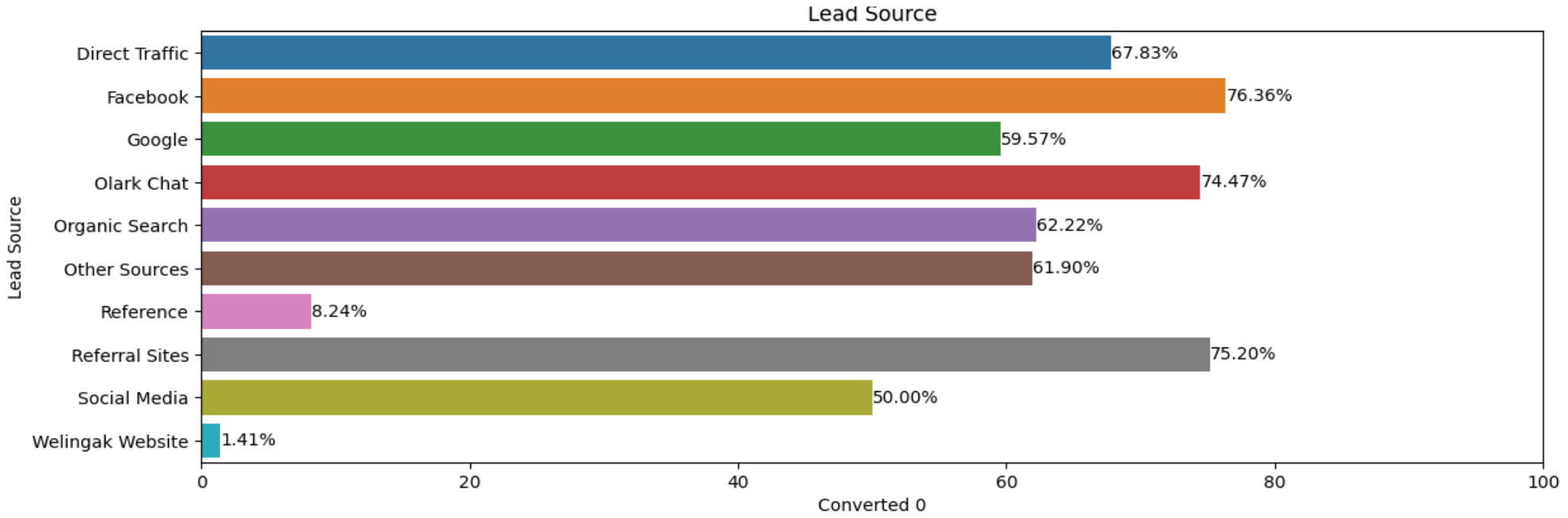
Lead Origin and Lead Source

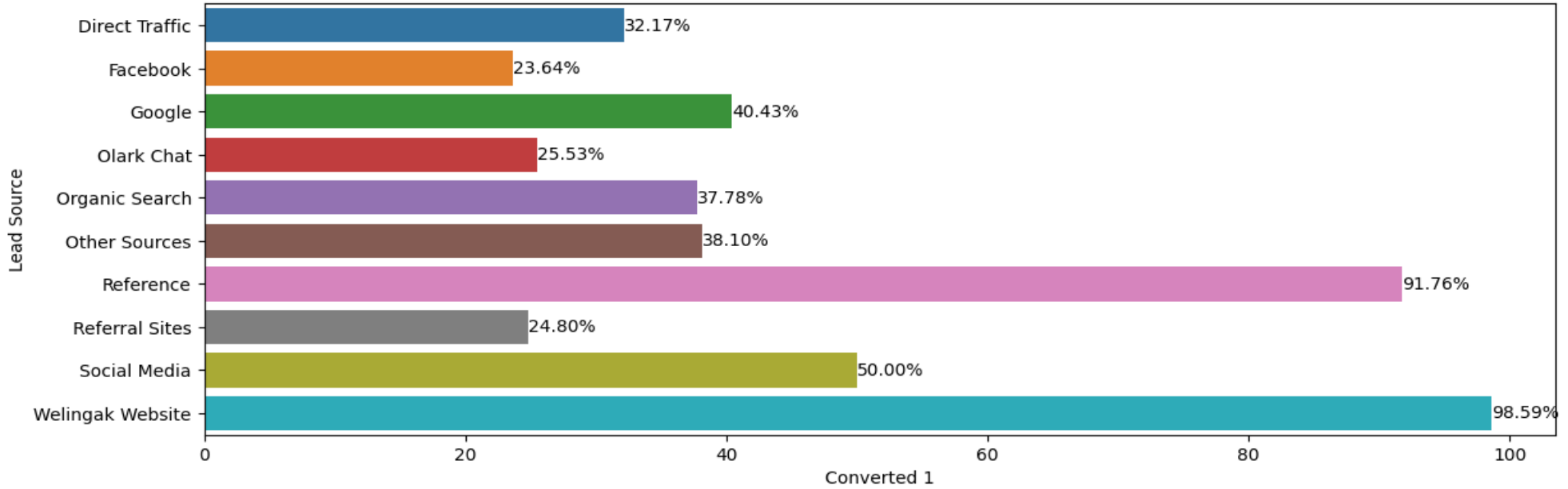




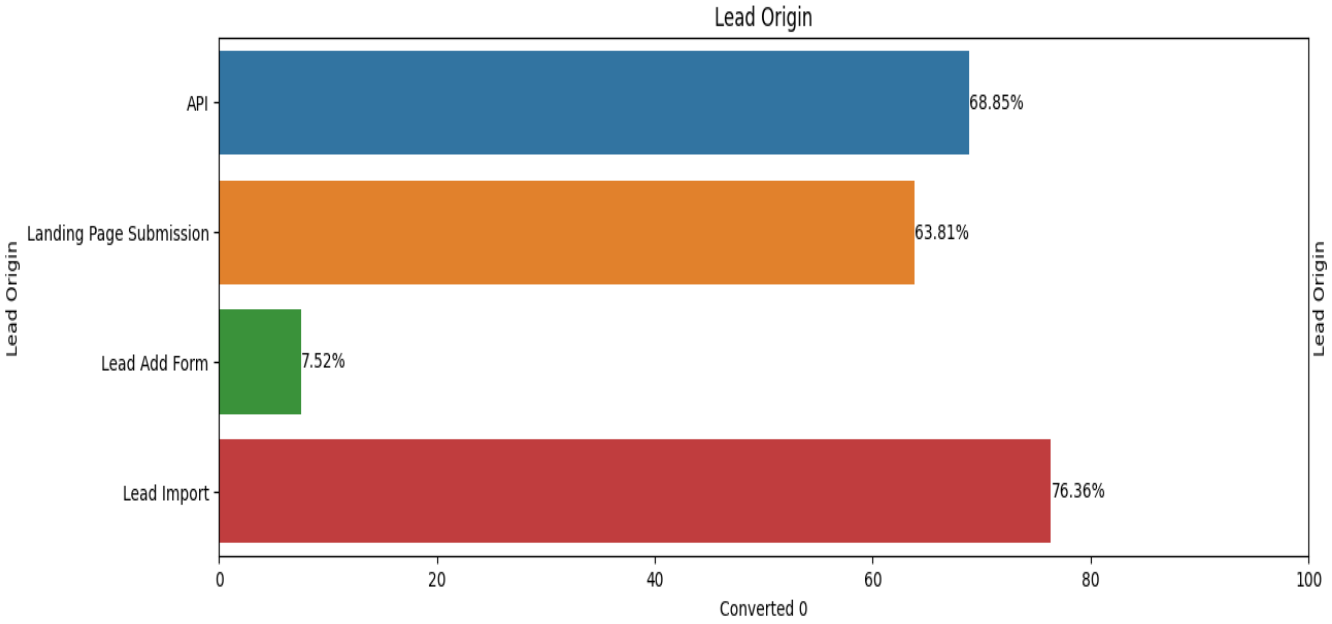
- Note: its seems that most if the leads from Lead Source are having very small percentage of counts and hence can be combined together for better readability and analysis and are named as Other Sources.

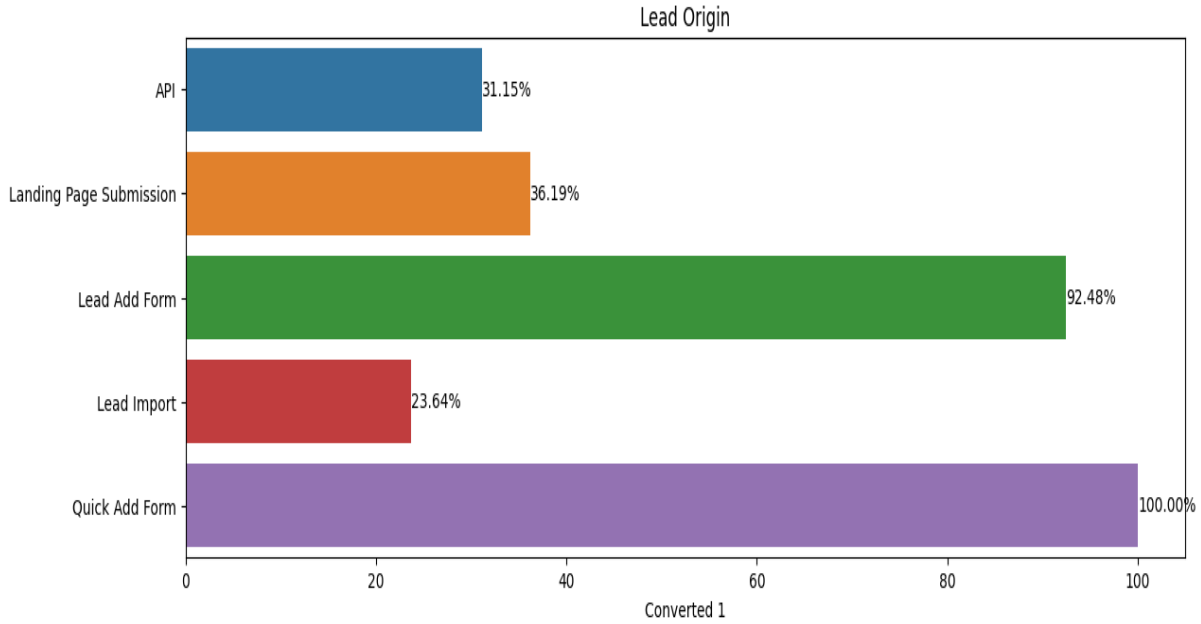
Lead Source wrt Converted





Lead Origin wrt Converted





Highlights

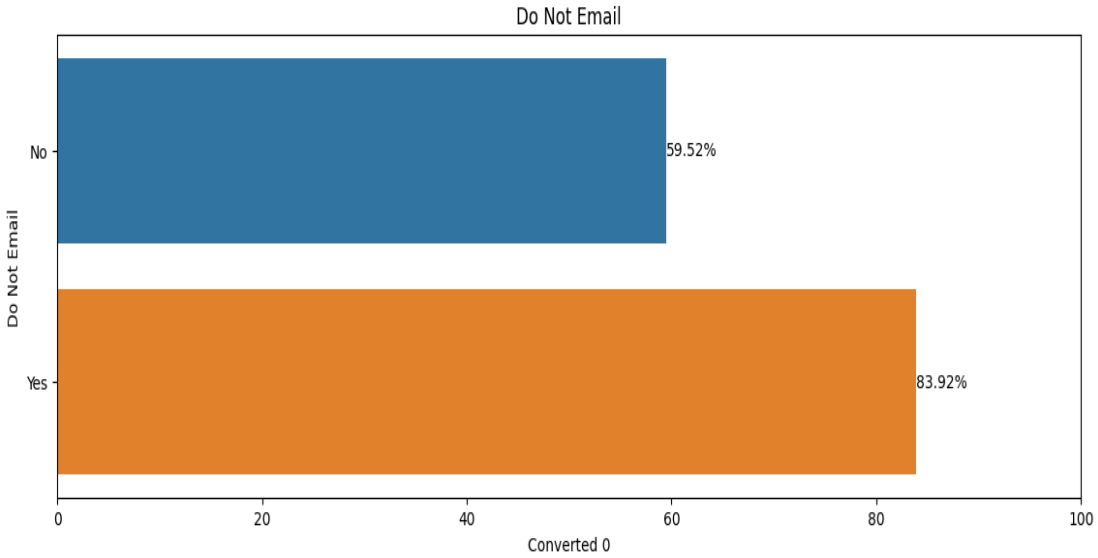
1.The Lead Origin "Lead Add Form" has very few leads but the conversion rate is very good(nearly 93%)

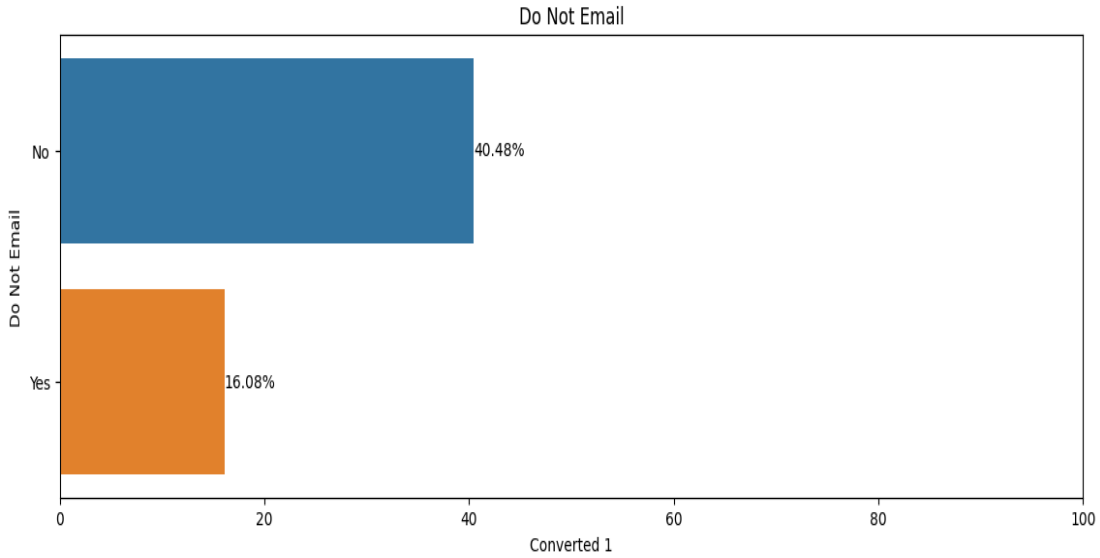
2.The Lead Origins "Landing Page Submission" and "API" have many leads but the conversion is low (about 36% for "Landing Page Submission" about 31% for "API")

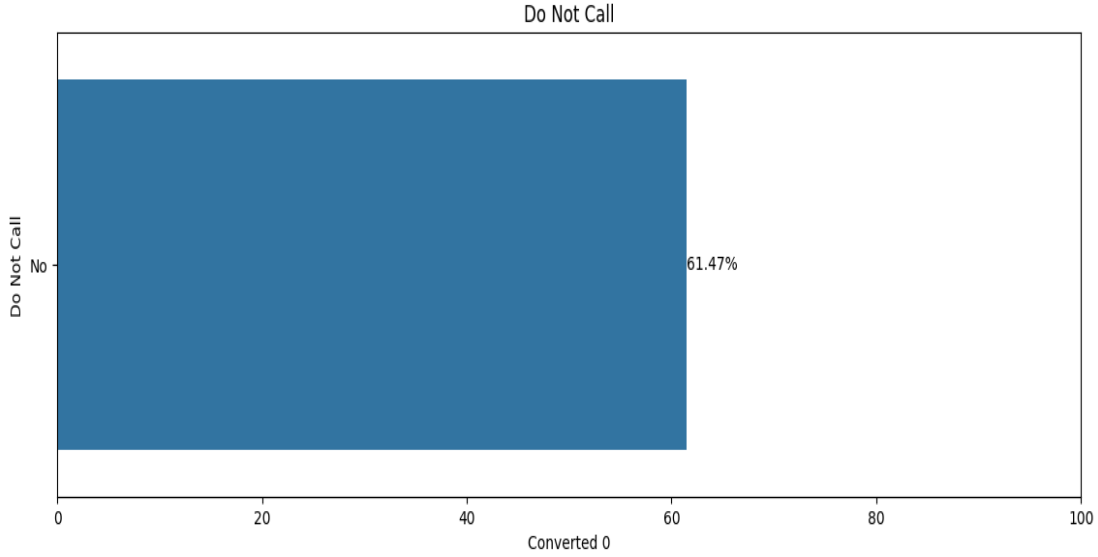
3.Of all the Lead Sources "Reference" has medium leads but a good conversion rate with 92%.Similarly Wellingak Websiite also has better lead conversion rate(98%).

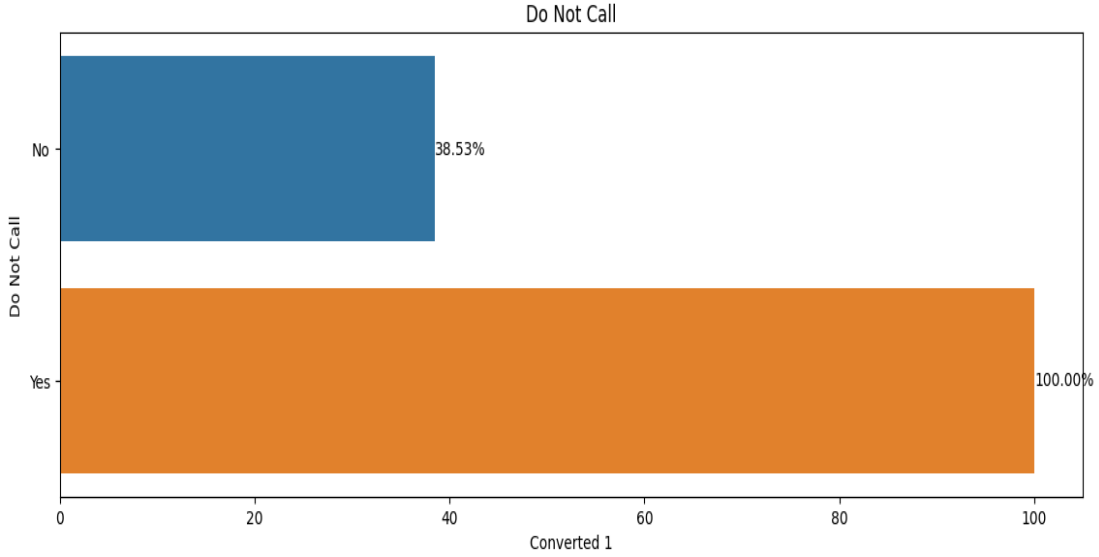
4.Leads from "Google" are good but conversion rate is just below average(40%).

Do not Email and Do not Call wrt Converted







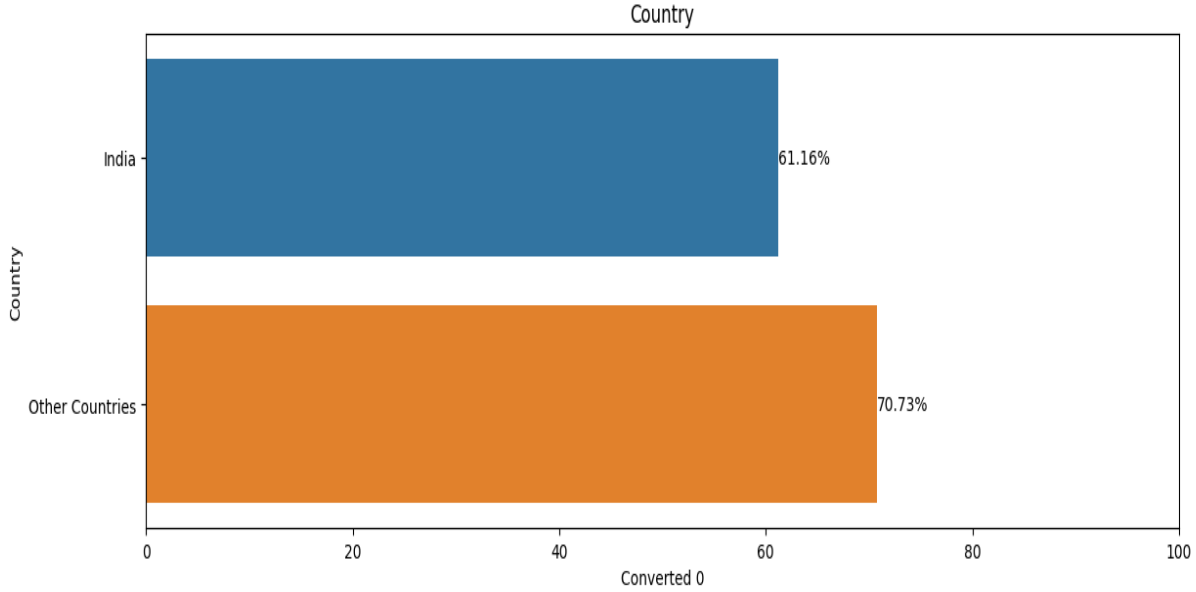


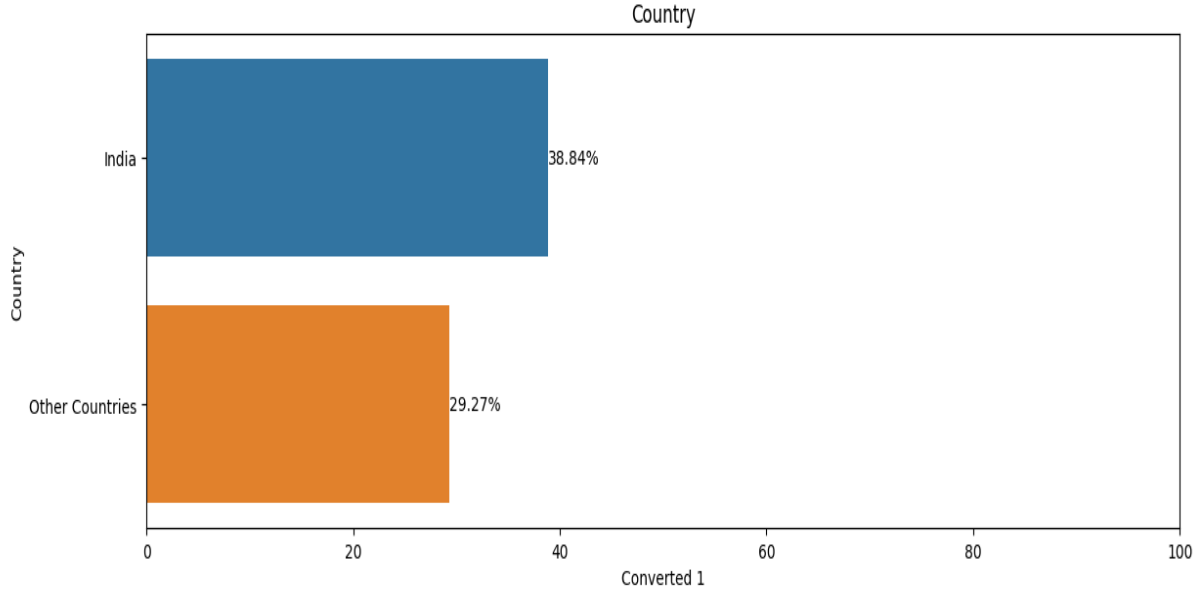
Highlights

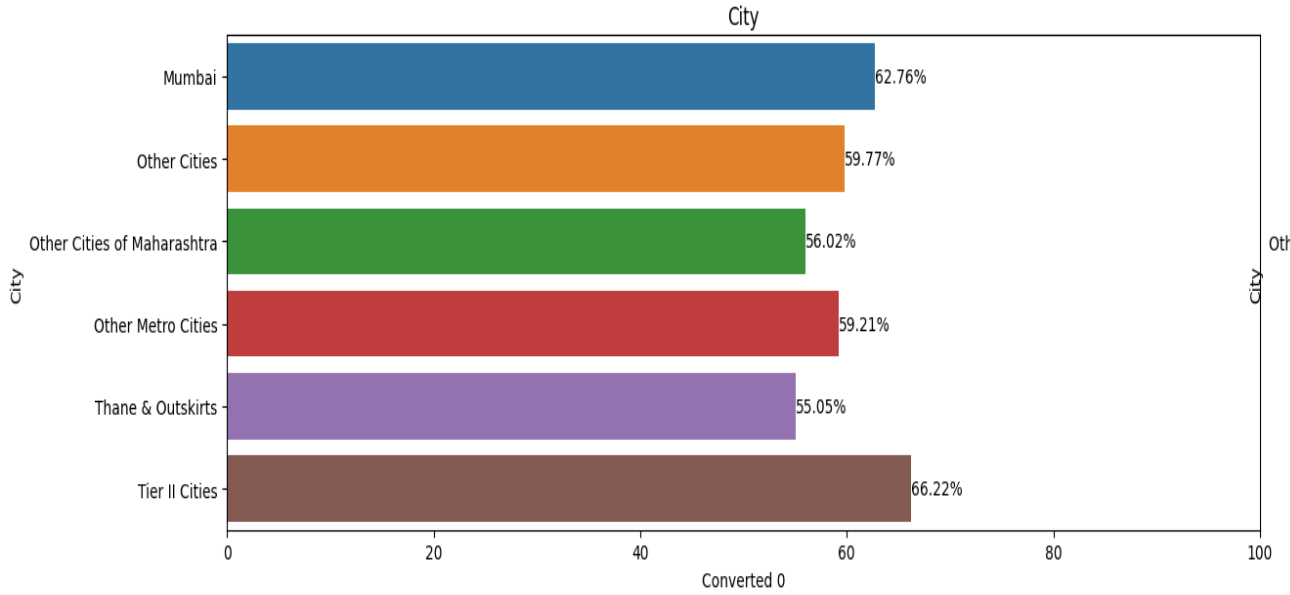
1.Nothing much can be deduced from "Do not call" column as most of the leads(99%) selected "No".This column can actually be dropped.

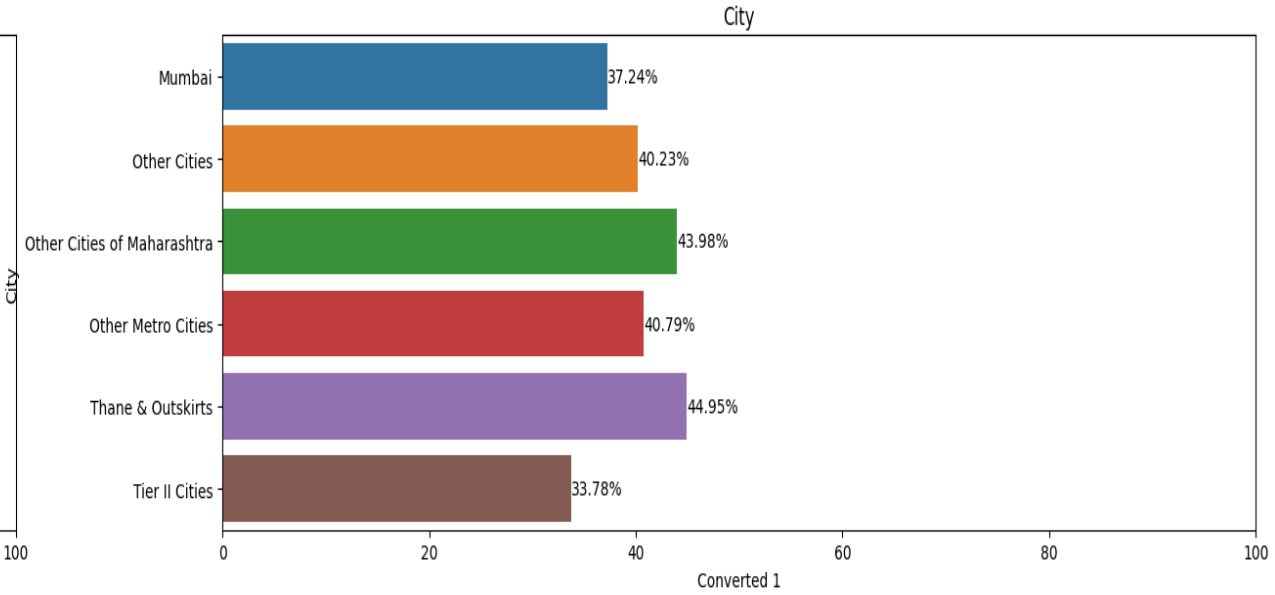
2.From "Do not email" it can be seen that leads who opted "Do not call" as "No" has better conversion rate than who opted for "Yes".

Country And City wrt Converted

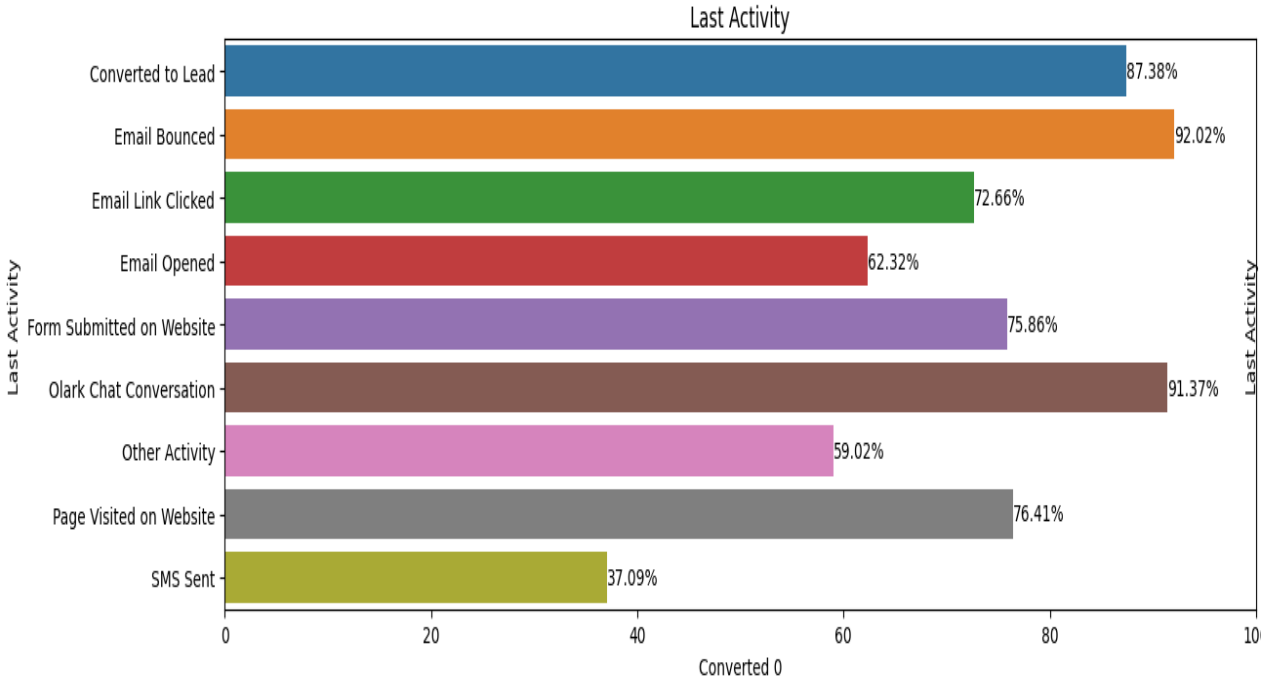


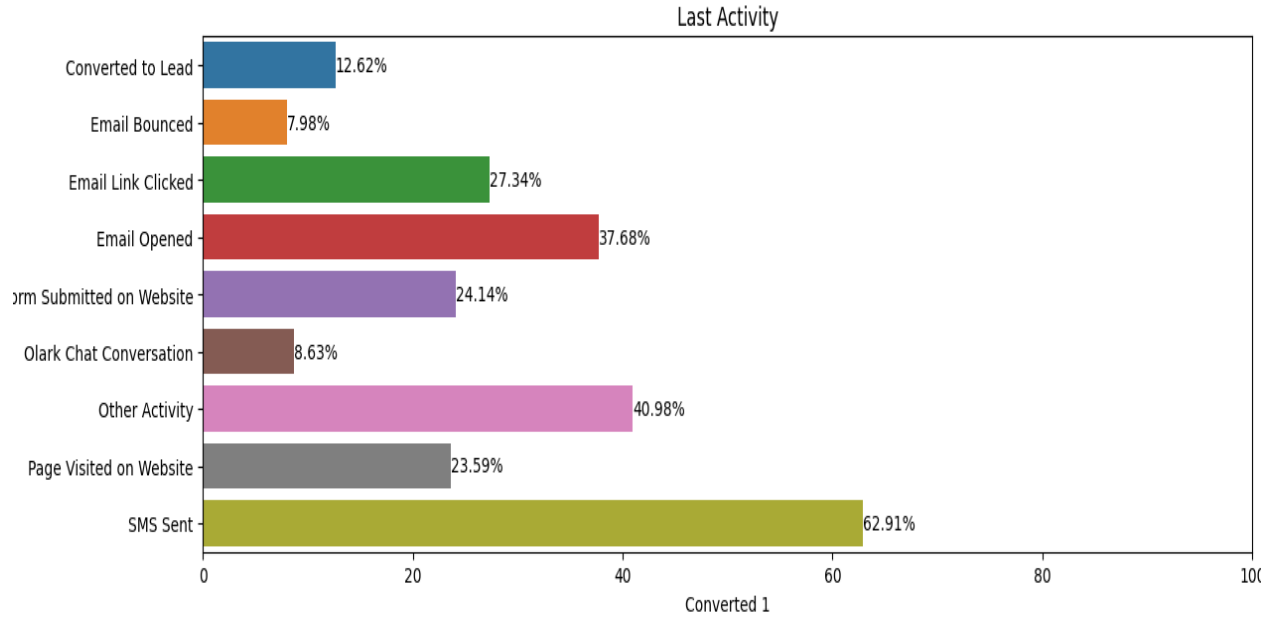






Last Activity wrt Converted





Highlights

1.The Country column seems to be too imbalanced and can be dropped from our analysis.

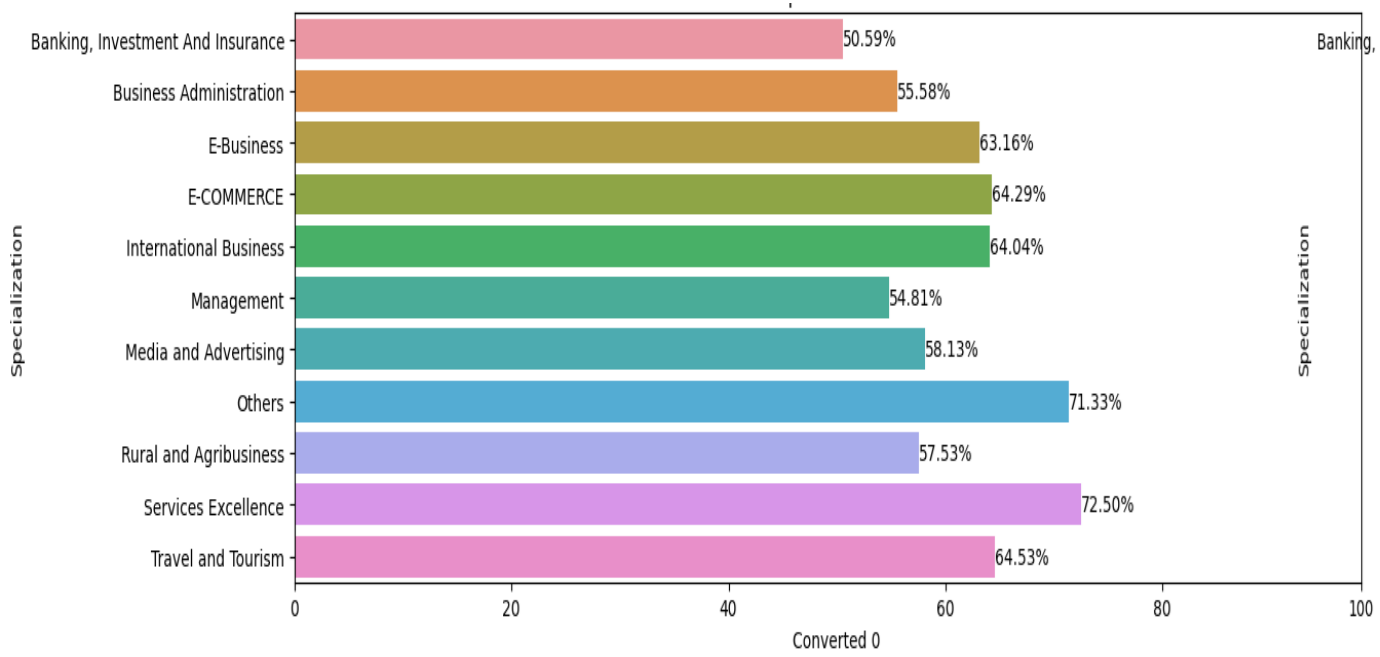
2.The "Last Activity" "SMS Sent" have a good lead counts and good conversion rate as well.

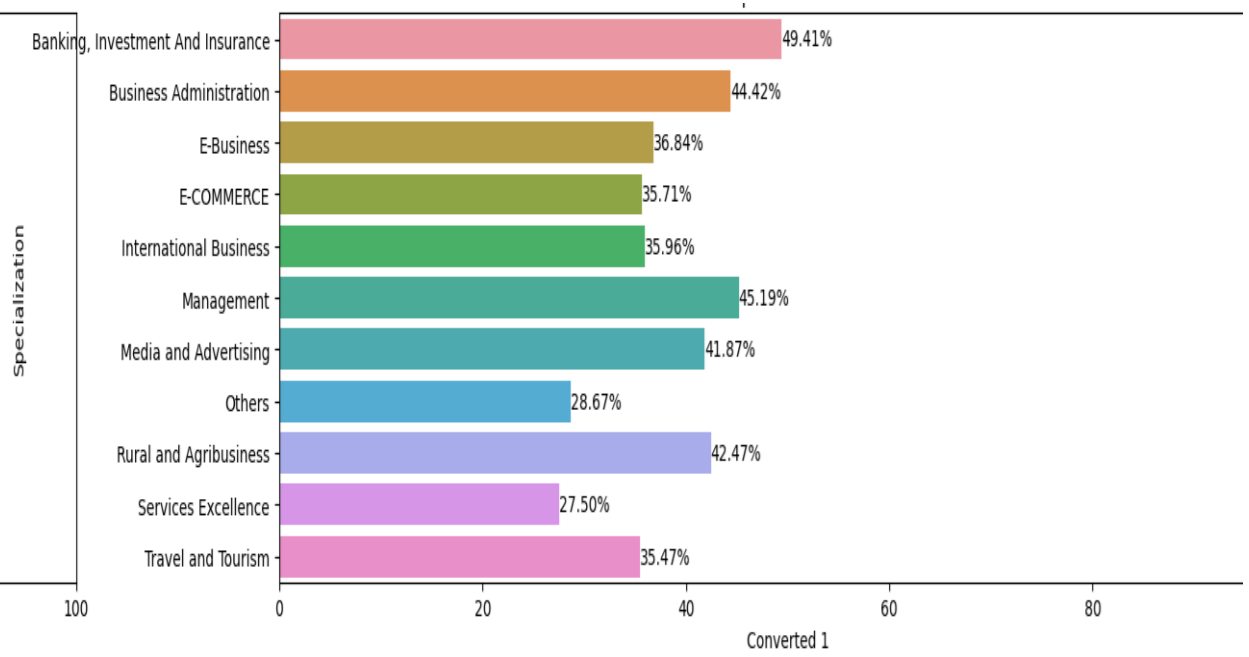
3.Though Leads with "Last Activity" "Email Opened" are high their conversion rate is not that good(only 37%)

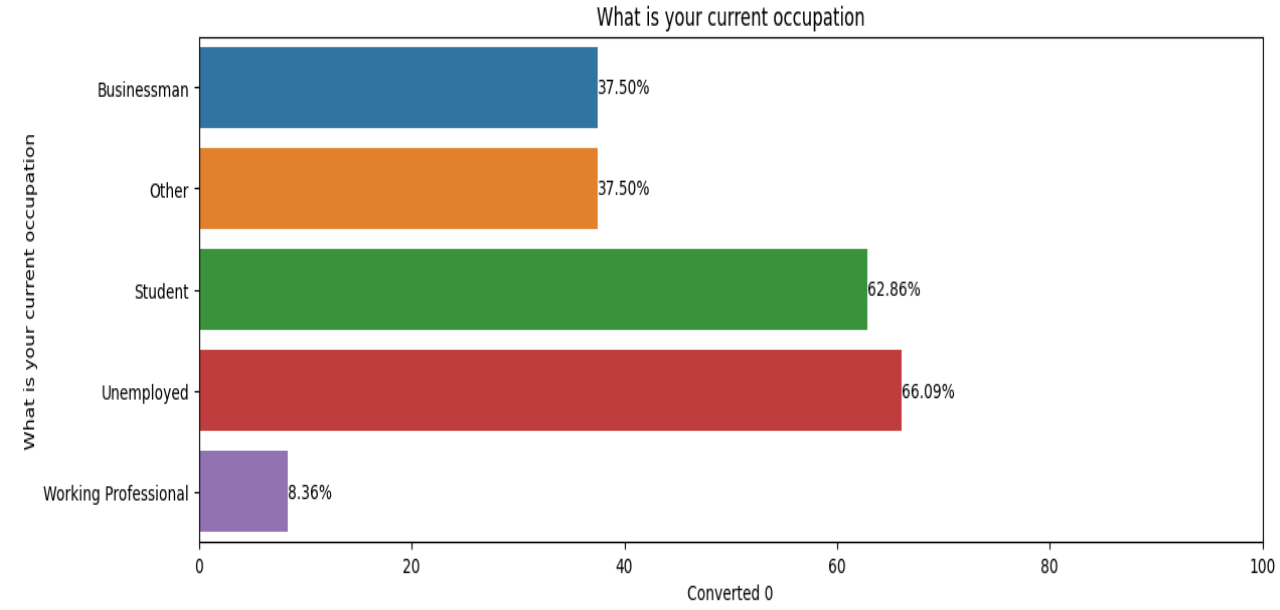
4.Also, Leads with "Last Activity" "Email Opened" are good but their conversion rate is just 8%.

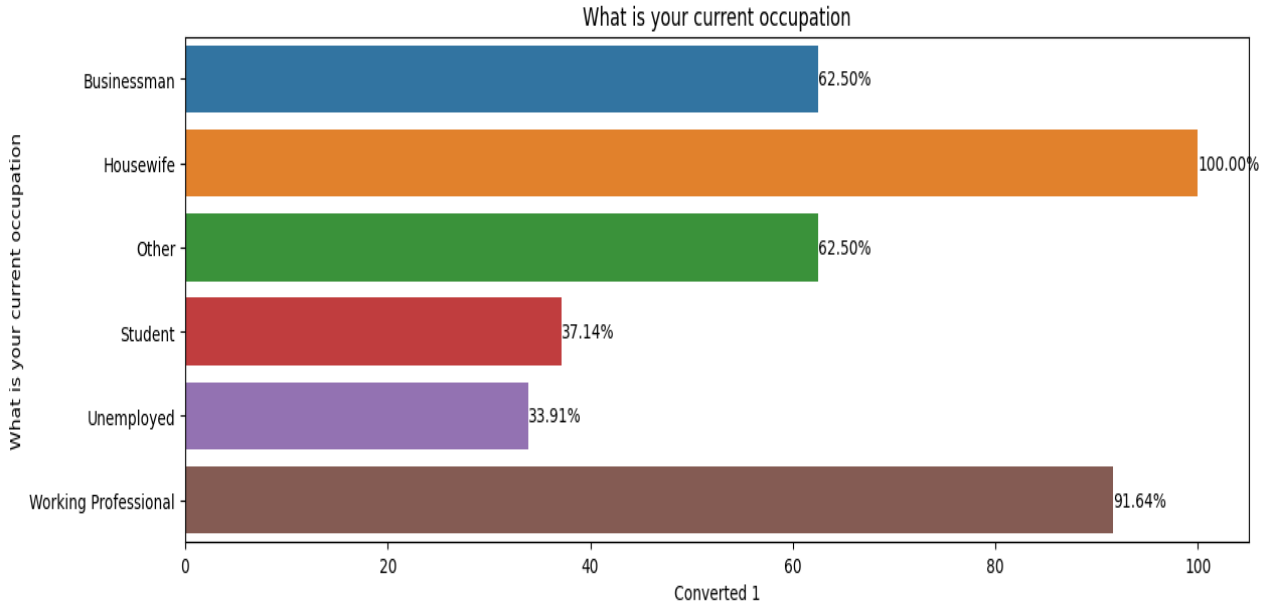
5.Leads count from "Mumbai" is very much higher than other cities , however the conversion rate is similar in all cities , ranging between 33% to 45%.So nothing much can be inferred from this column and it can be dropped.

Specialization, What is your current occupation wrt Converted

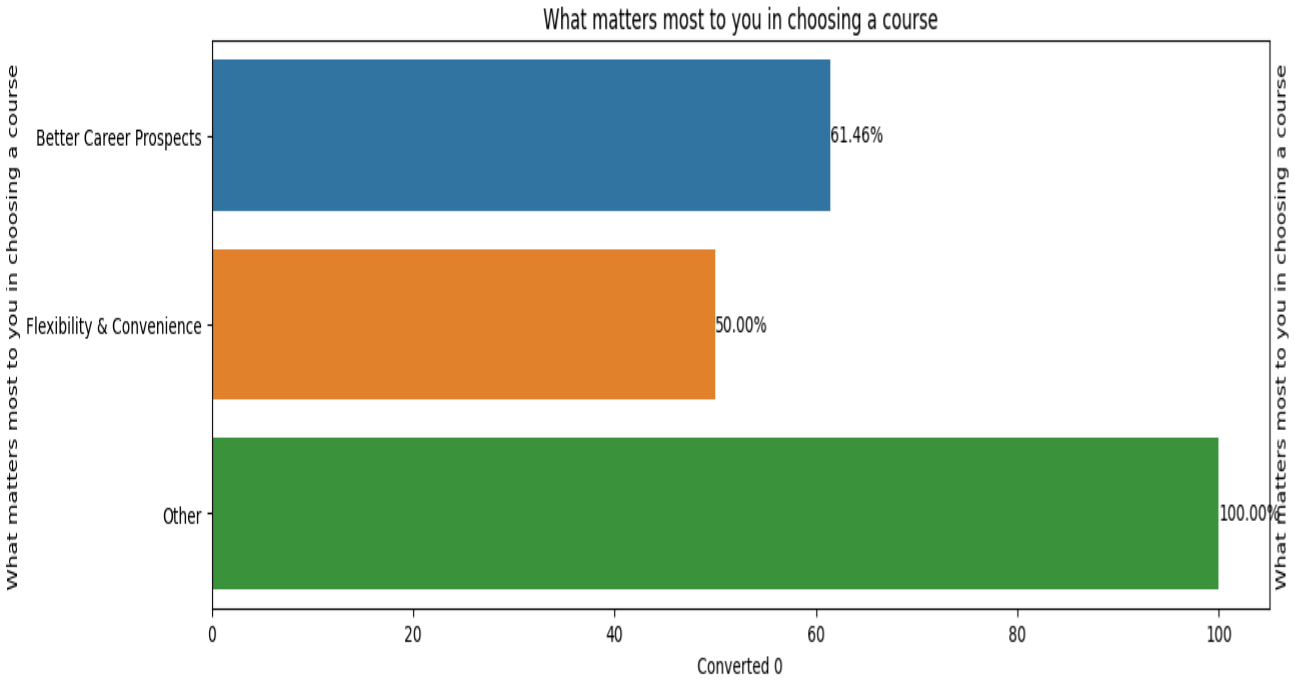


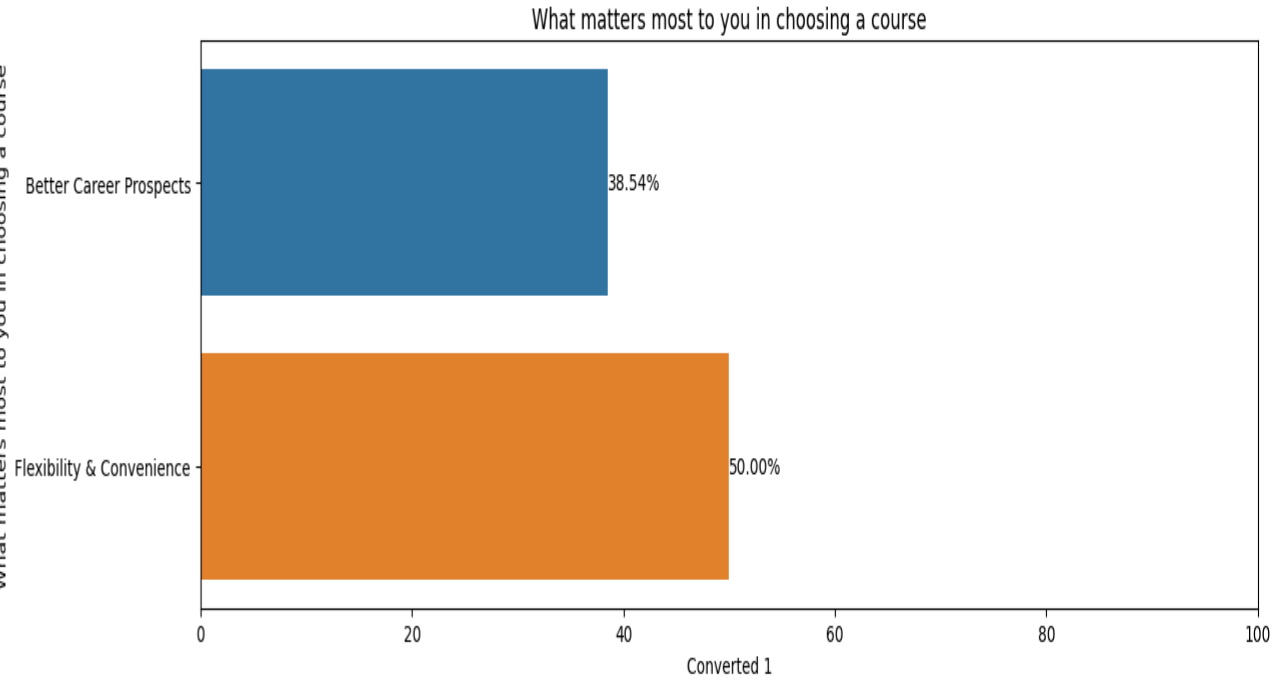






What matters most to you in choosing a course wrt Converted





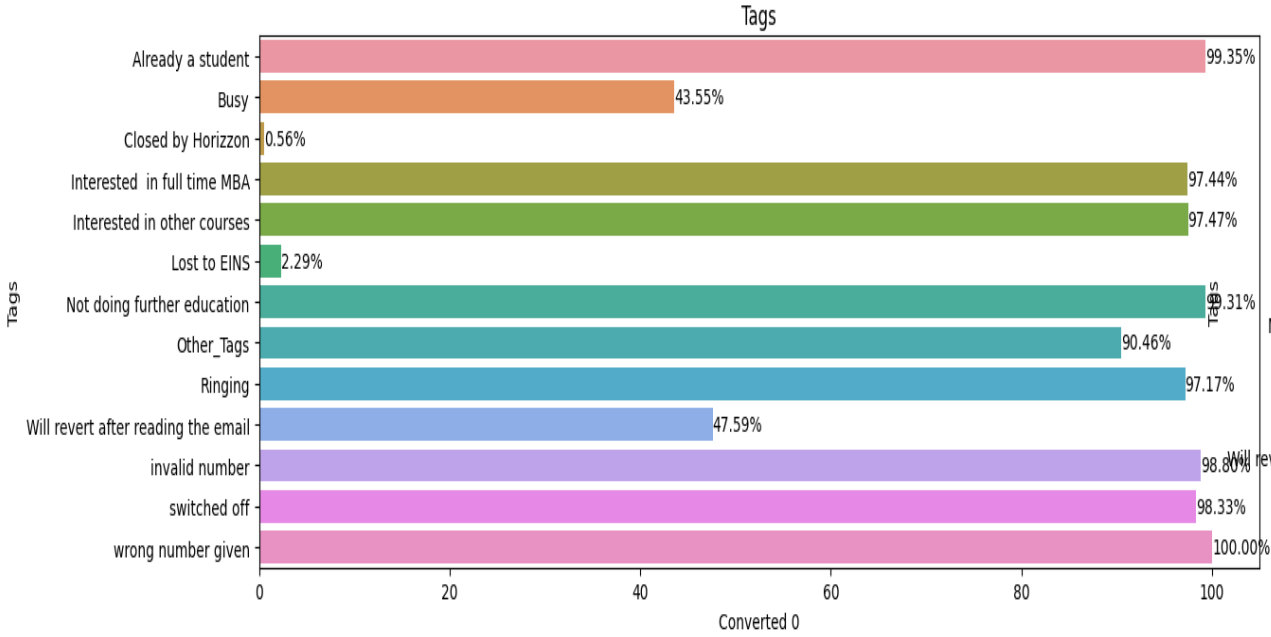
Highlights

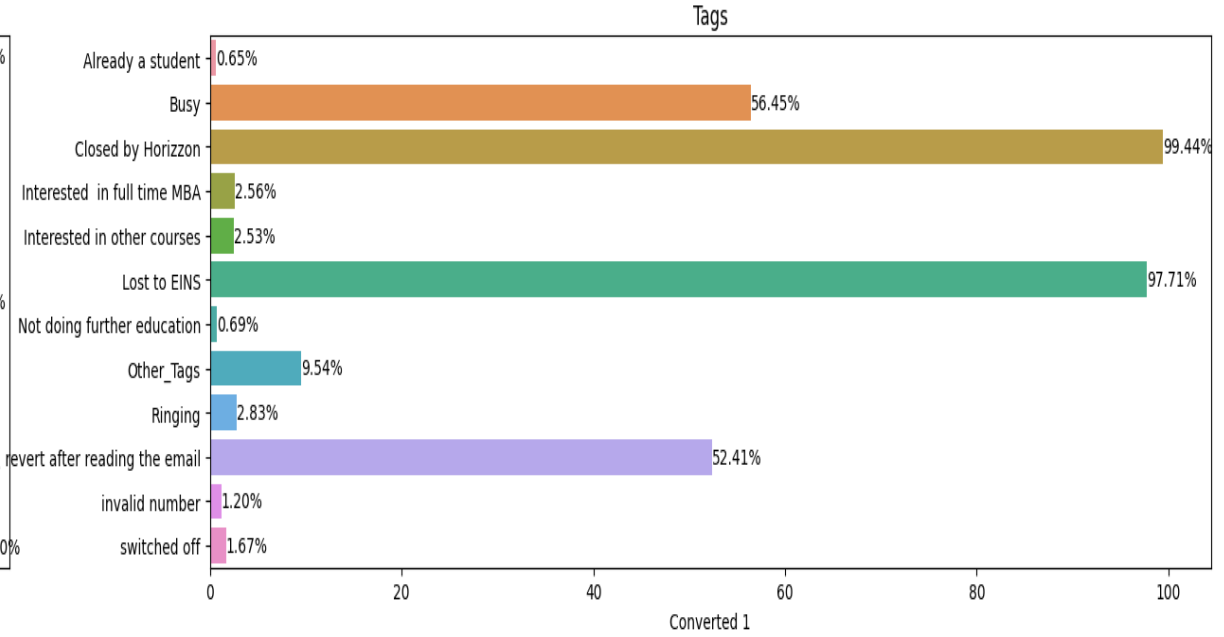
1.The Column "What matters most to you in choosing a course" can be dropped as it is showing almost zero variance.

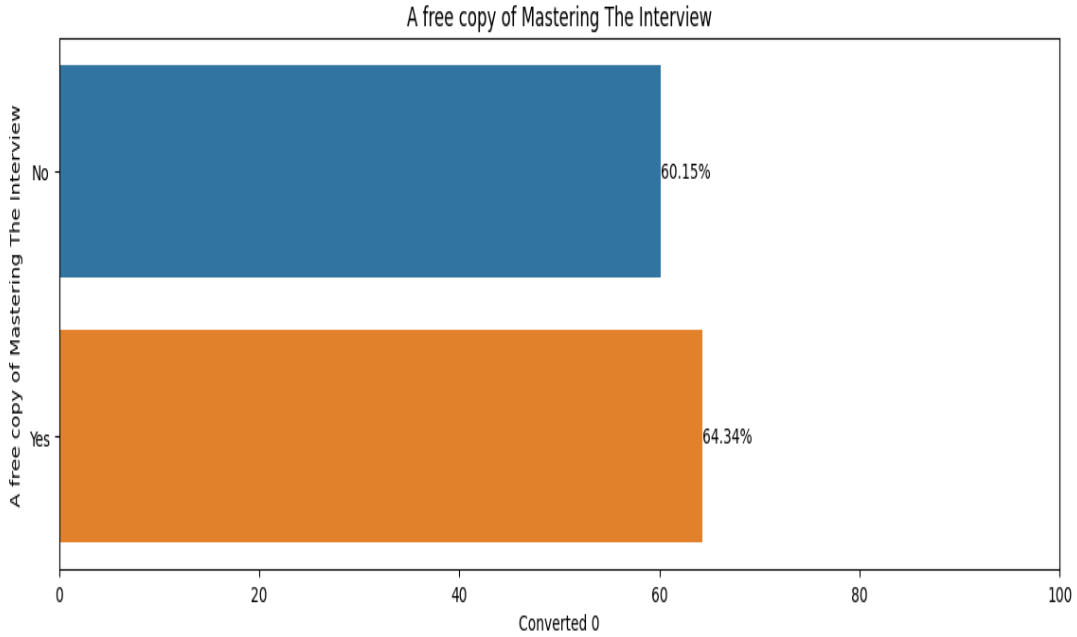
2.From the Column "'What is your current occupation", it can be seen that the conversion rate of "working professionals" is better than others.

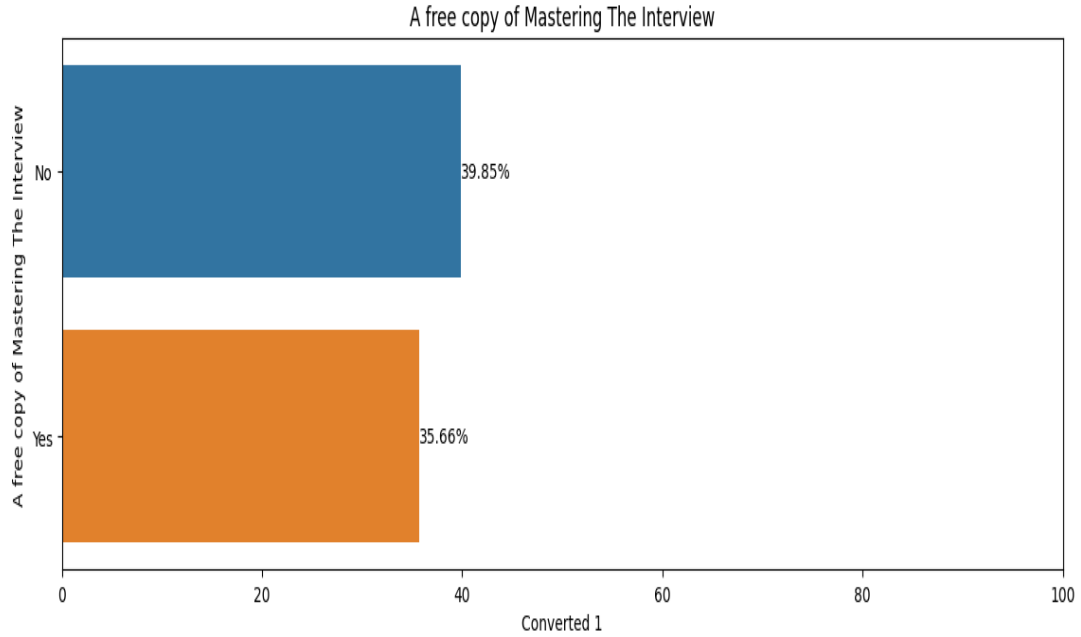
3.The conversion rates of specialization is similar for all categories with leads from "Banking, Investment and Insurance" having a better edge over others.

Tag, A free copy of Mastering The Interview wrt Converted

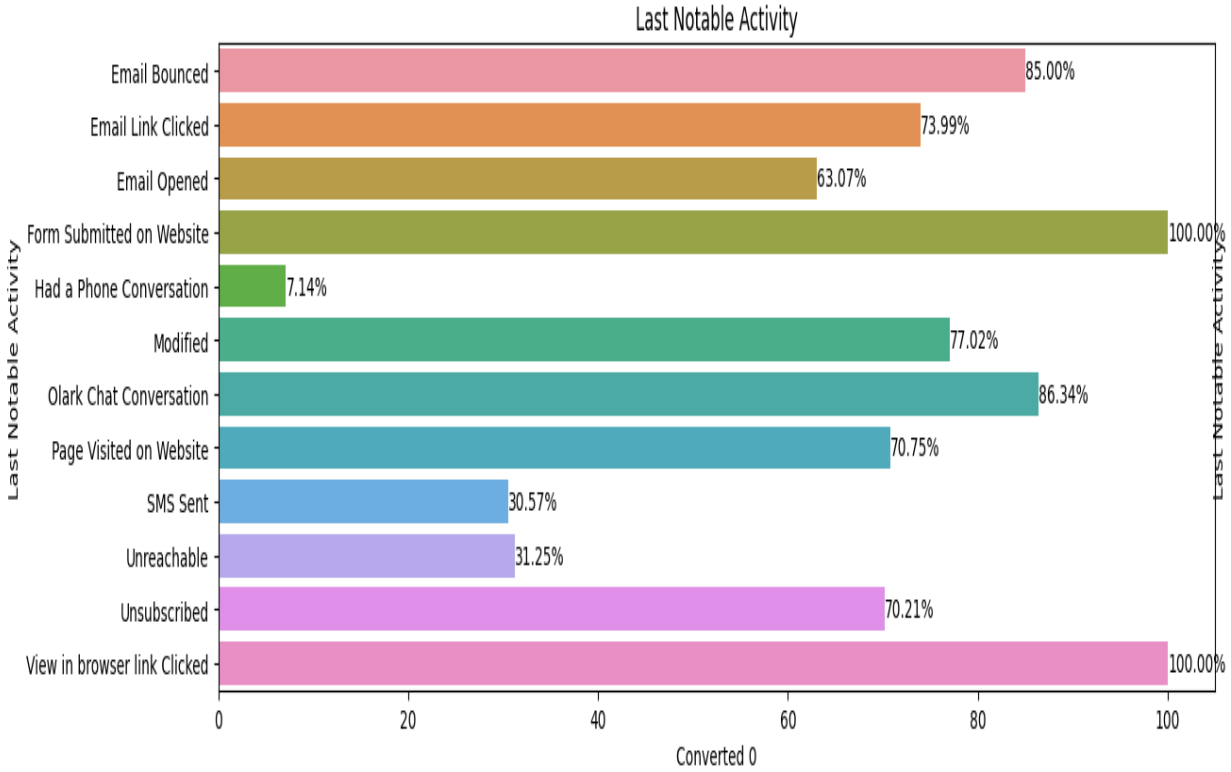


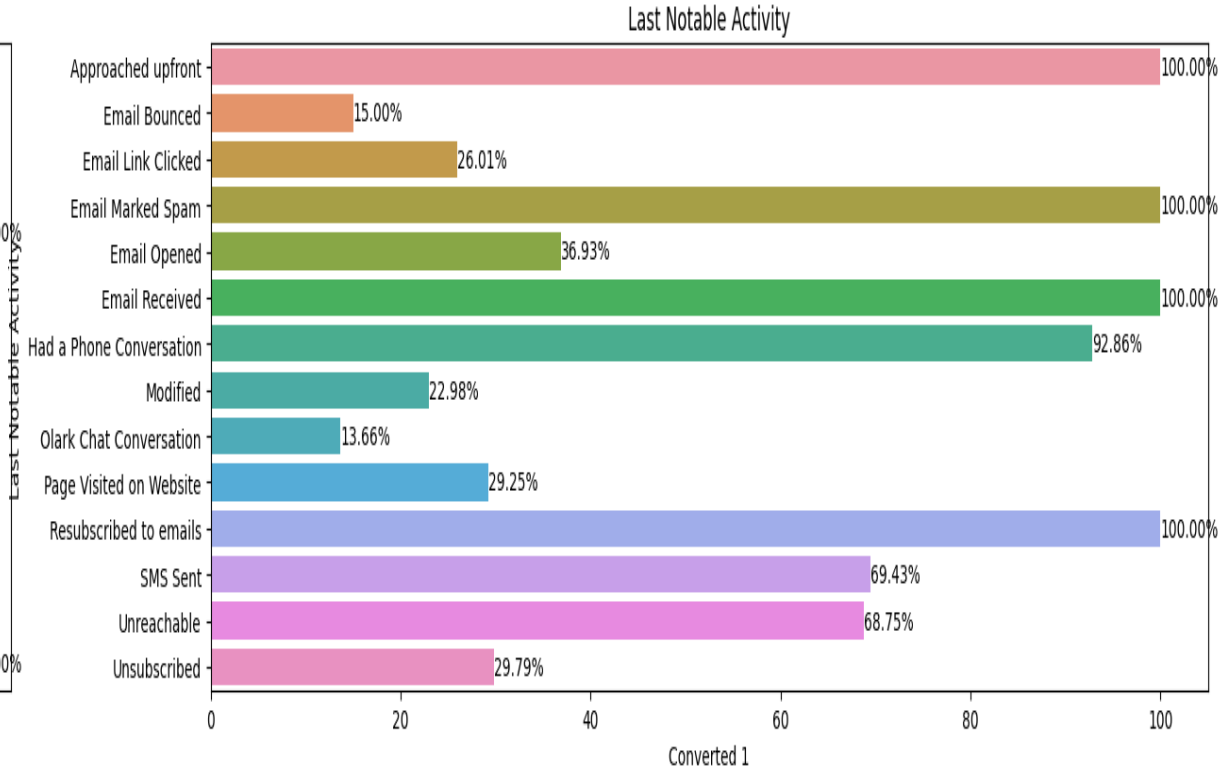






Last Notable Activity wrt Converted





Highlights:

1.Leads with last notable activity as "Modified" are high however their conversion rate is just 22%.

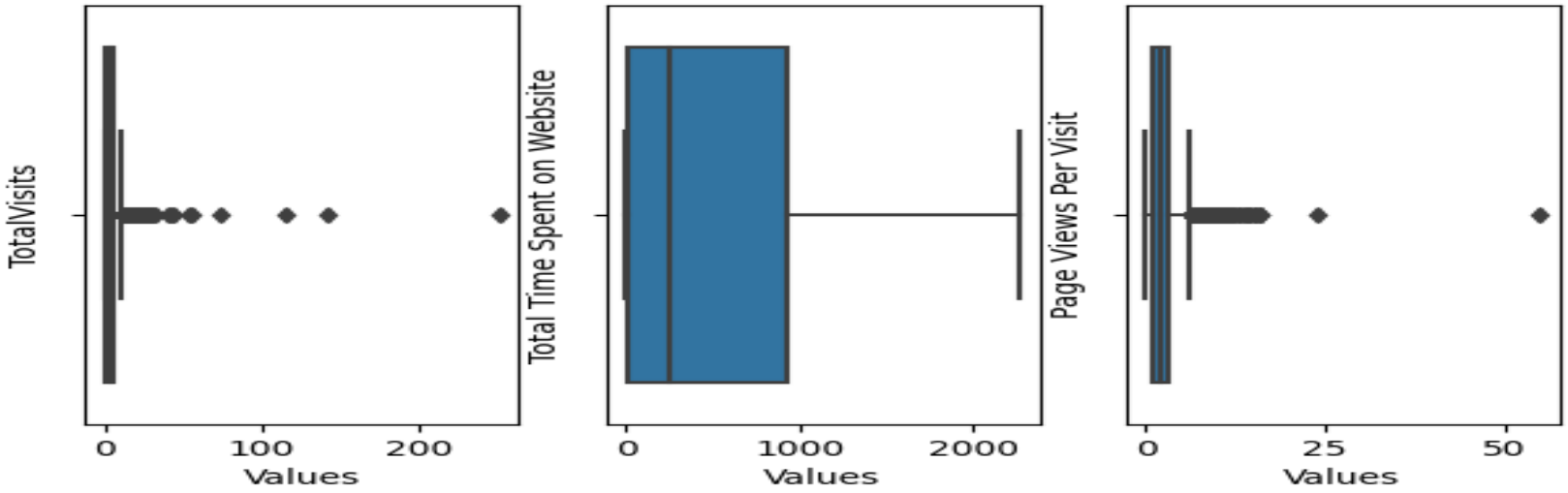
2.Leads with last notable activity as "SMS sent" are having good converson rate of 69%

3.Leads with Tag "Will revert after reading the mail" are better leads with good conversion rate

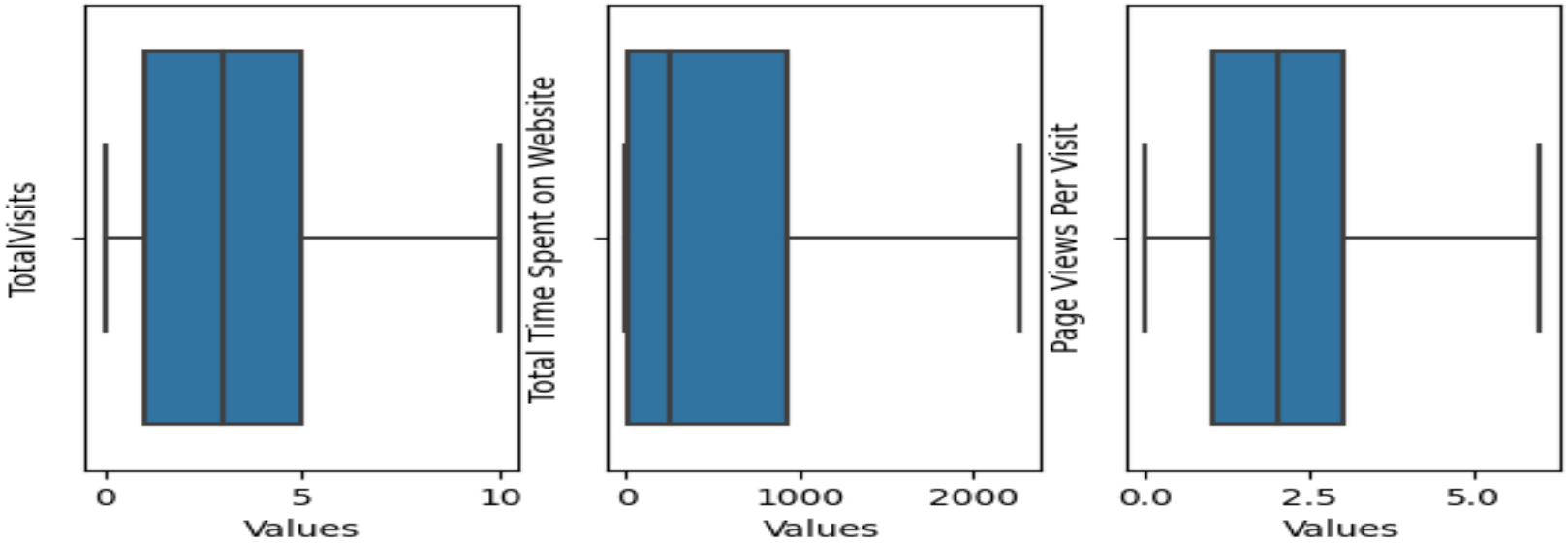
NUMERICAL UNIVARIATE ANALYSIS

- Total Visits
- Total Time Spent on Website
- Page views per Visit

TotalVisits, Total Time Spent on Website and Page Views Per Visit



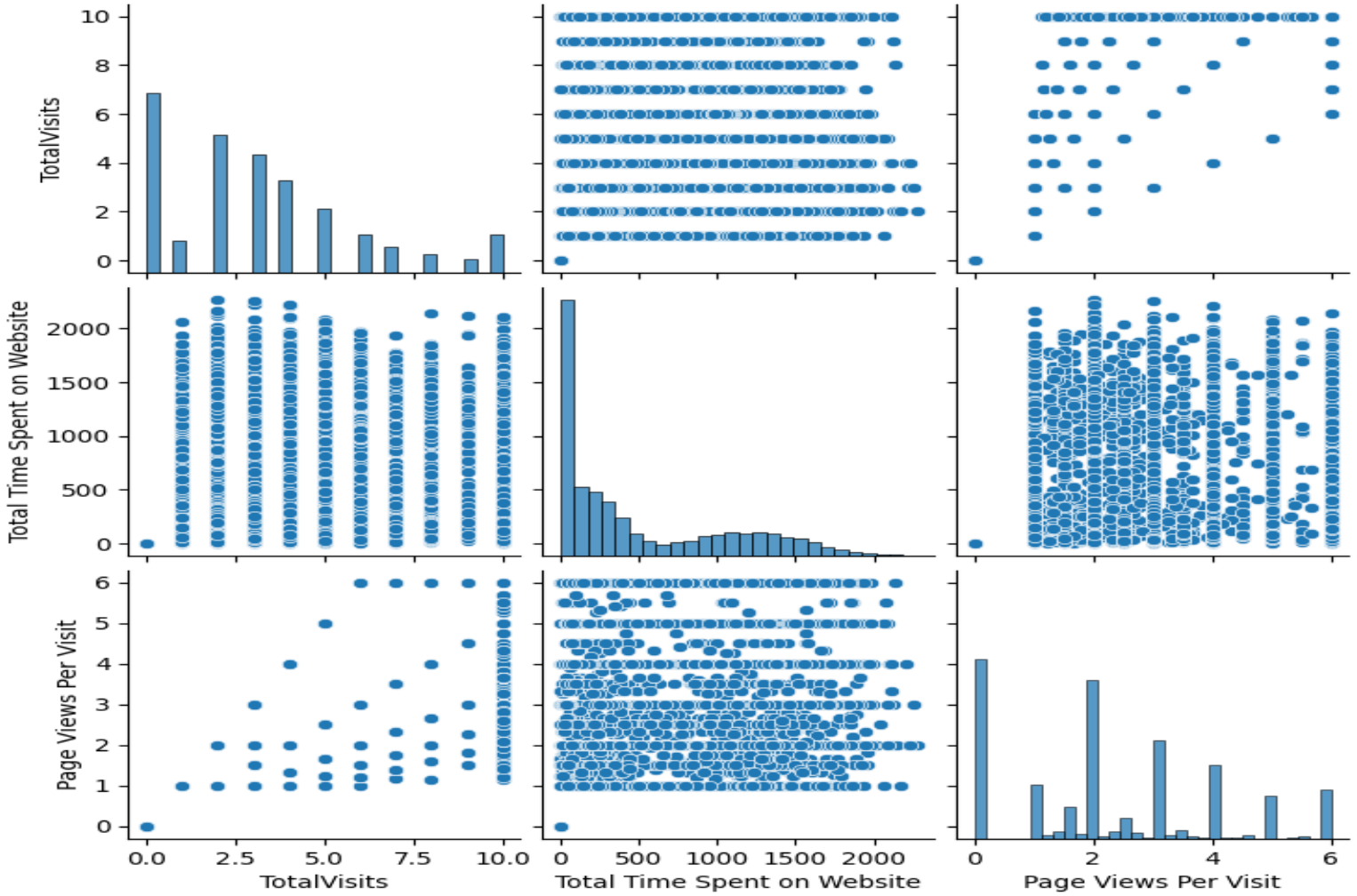
- There are outliers in Total Visits, Page views per visit. After Treating them,

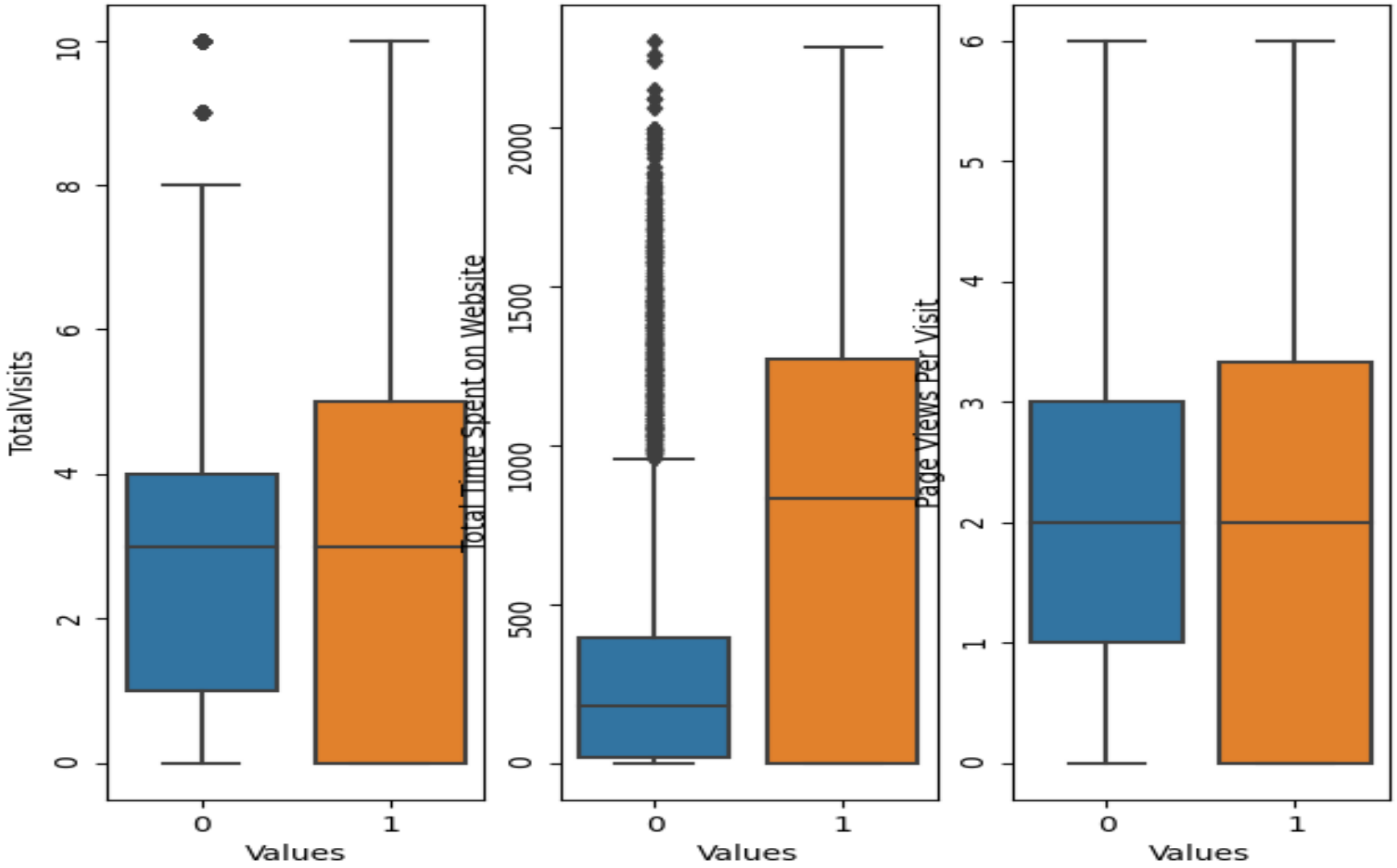


NUMERICAL BIVARIATE ANALYSIS

- Total Visits
- Total Time Spent on Website
- Page views per Visit

TotalVisits, Total Time Spent on Website and Page Views Per Visit



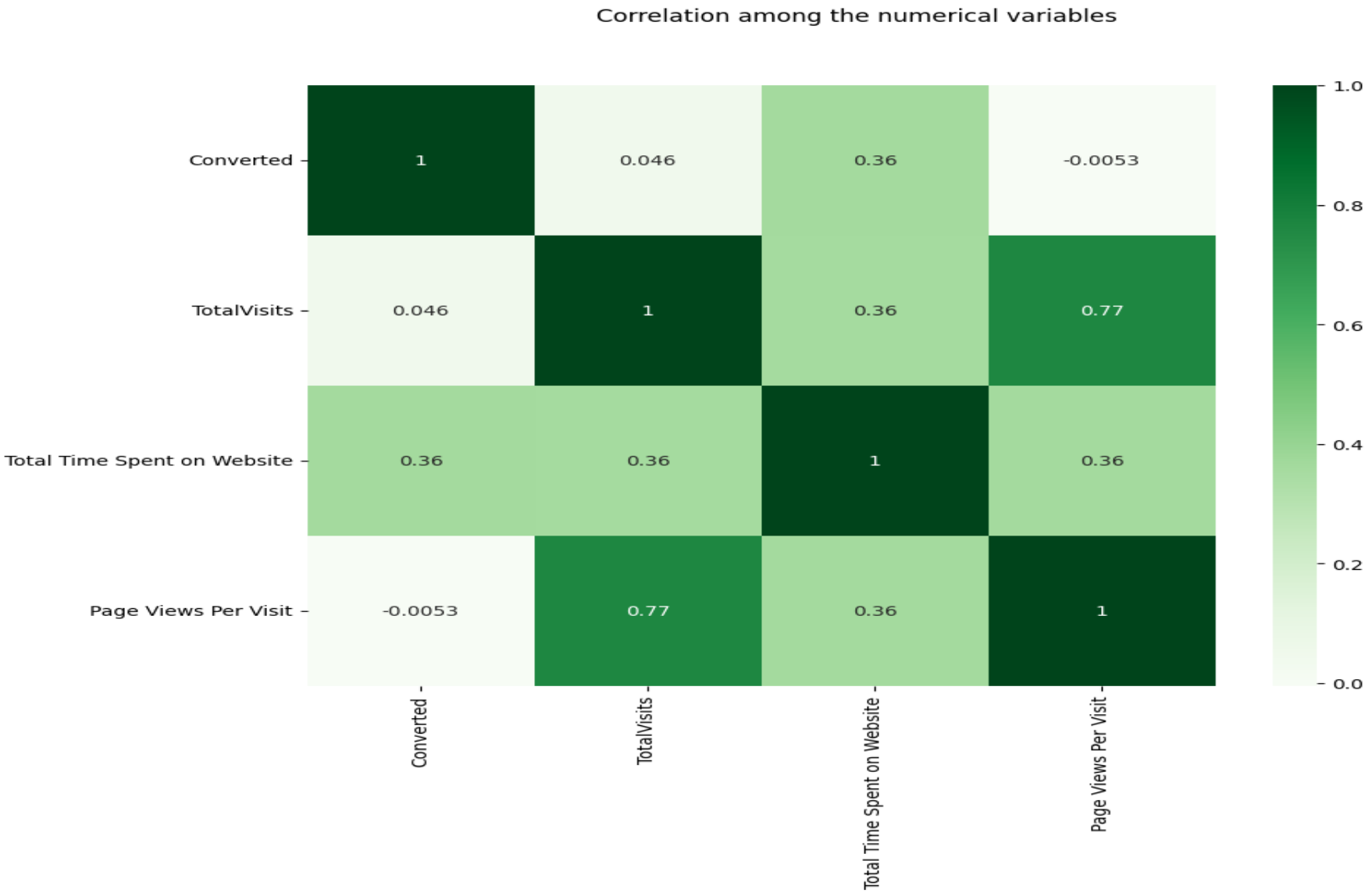


Highlights:

1.The columns "TotalVisits" and "Page Views per visit" have positive correlation and nothing much can be inferred from them as of now

2.The column "Total Time spent on each visit seems to have positive correlation with target column "Converted“**.**

TotalVisits, Total Time Spent on Website and Page Views Per Visit And Converted



Model Building

Feature Selection and Statistics of Final Model

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	7392
Model:	GLM	Df Residuals:	7378
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3045.7
Date:	Sun, 19 Nov 2023	Deviance:	6091.5
Time:	04:02:12	Pearson chi2:	8.11e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3975
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5205	0.119	-4.362	0.000	-0.754	-0.287
TotalVisits	0.7974	0.170	4.678	0.000	0.463	1.131
Total Time Spent on Website	4.2229	0.147	28.777	0.000	3.935	4.511
Page Views Per Visit	-1.3034	0.177	-7.355	0.000	-1.651	-0.956
Lead Origin_Landing Page Submission	-1.2853	0.114	-11.233	0.000	-1.510	-1.061
Lead Origin_Lead Add Form	2.7116	0.185	14.653	0.000	2.349	3.074
Lead Source_Welingak Website	2.5897	0.742	3.490	0.000	1.135	4.044
Do Not Email_Yes	-1.4253	0.158	-9.008	0.000	-1.735	-1.115
Last Activity_Other Activity	1.1222	0.219	5.115	0.000	0.692	1.552
Last Activity_SMS Sent	1.3797	0.070	19.618	0.000	1.242	1.518
What is your current occupation_Working Professional	2.6226	0.180	14.599	0.000	2.271	2.975
Last Notable Activity_Modified	-1.0192	0.074	-13.818	0.000	-1.164	-0.875
Last Notable Activity_Olark Chat Conversation	-1.1317	0.308	-3.677	0.000	-1.735	-0.528
Specialization_Others	-1.1637	0.114	-10.185	0.000	-1.388	-0.940

	Features	VIF
0	const	14.66
4	Lead Origin_Landing Page Submission	3.20
13	Specialization_Others	2.93
3	Page Views Per Visit	2.86
1	TotalVisits	2.51
5	Lead Origin_Lead Add Form	1.63
6	Lead Source_Welingak Website	1.27
2	Total Time Spent on Website	1.21
10	What is your current occupation_Working Profes...	1.15
11	Last Notable Activity_Modified	1.13
9	Last Activity_SMS Sent	1.12
7	Do Not Email_Yes	1.05
12	Last Notable Activity_Olark Chat Conversation	1.05
8	Last Activity_Other Activity	1.04

Correlation Among Features of Final Model

Model Evaluation on Train Dataset

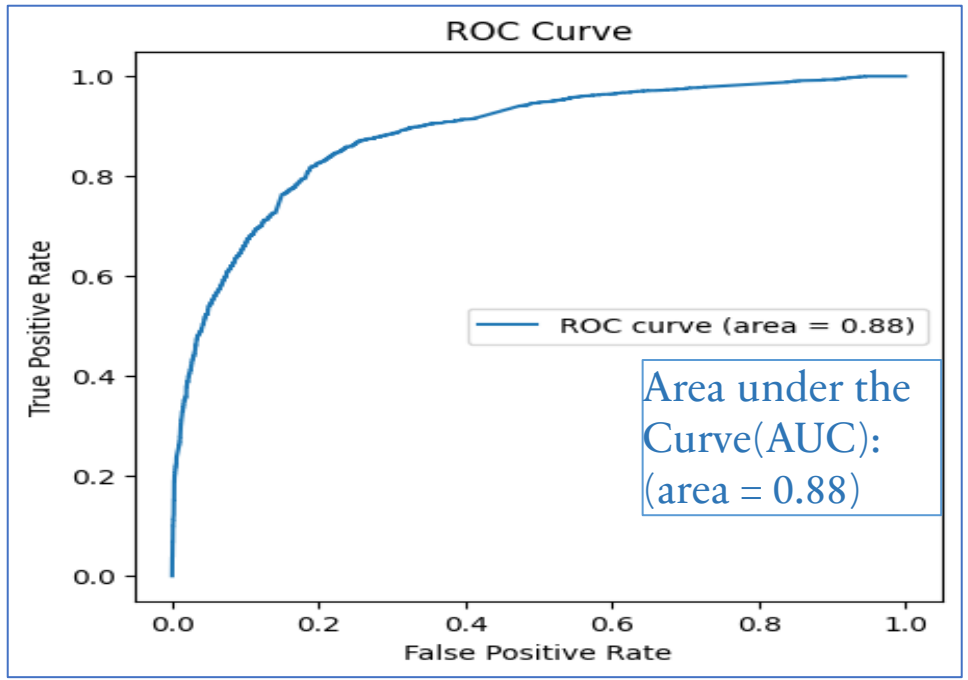
Predicting the Conversion Probability and Predicted Column

	Converted	Conversion_Prob
0	1	0.062777
1	0	0.390658
2	1	0.364888
3	0	0.876405
4	0	0.062777

Creating a dataframe with the actual Converted flag and the Conversion Probabilities.

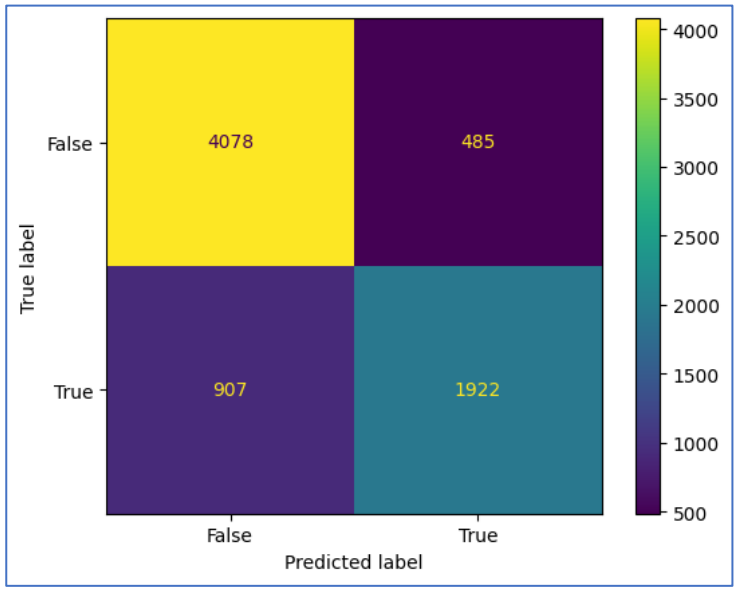
	Converted	Conversion_Prob	Predicted
0	1	0.062777	0
1	0	0.390658	0
2	1	0.364888	0
3	0	0.876405	1
4	0	0.062777	0

Creating a new column 'Predicted' with 1 if Conversion_Prob>0.5 else 0.



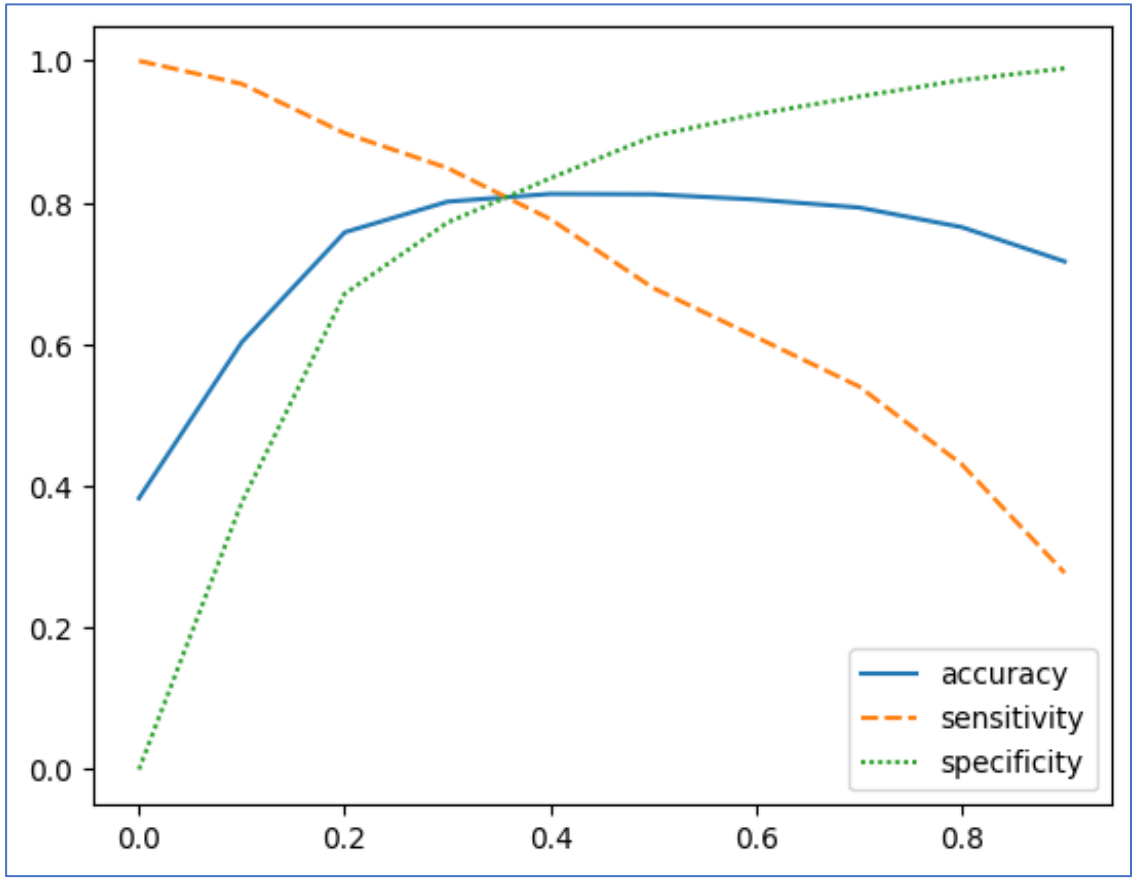
Area under the Curve(AUC): (area = 0.88)

- Confusion Matrix after considering cutoff value of 0.5



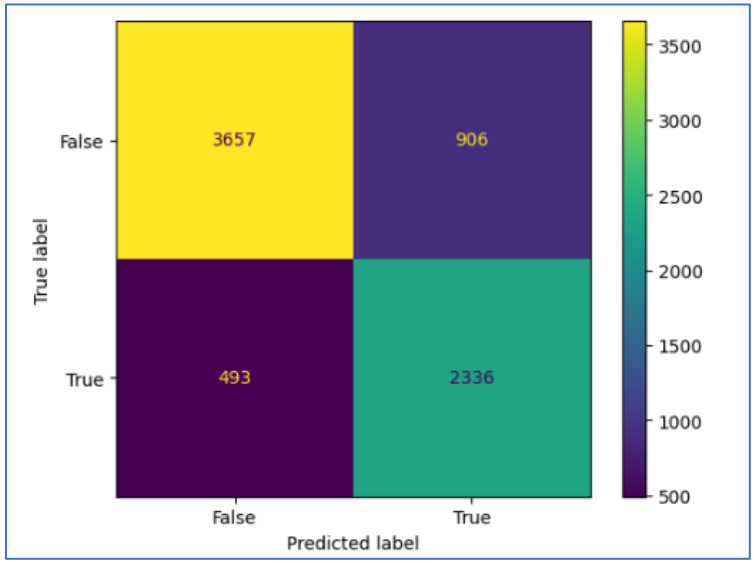
- Accuracy : 0.81
- Specificity :0.89
- Sensitivity: 0.68

Optimal Probability Threshold



From the curve above, 0.35 is found to be the optimum point for cutoff probability.

- Confusion Matrix after considering cutoff value of 0.35



- Accuracy : 0.81
- Specificity : 0.80
- Sensitivity: 0.82

Model Evaluation with optimum cut-off

Predicting the Conversion Probability and Predicted Column

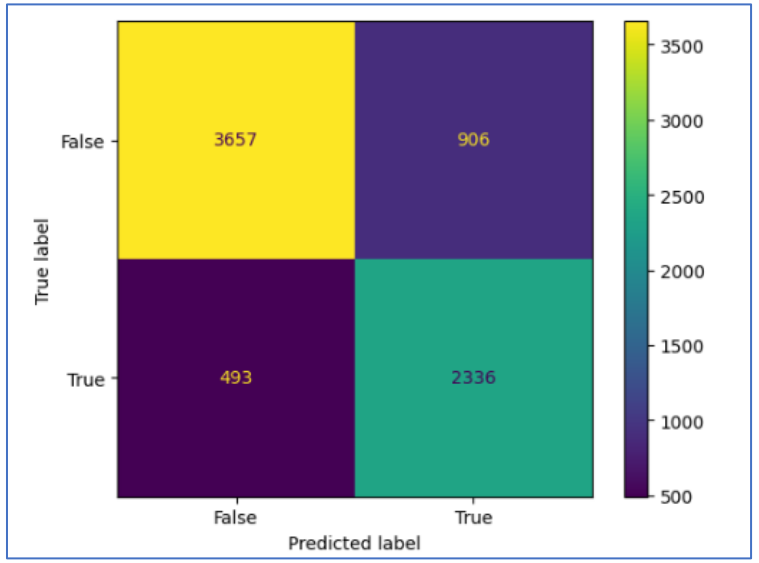
	Converted	Conversion_Prob
0	1	0.062777
1	0	0.390658
2	1	0.364888
3	0	0.876405
4	0	0.062777

Creating a dataframe with the actual Converted flag and the Conversion Probabilities.

	Converted	Conversion_Prob	Predicted
0	1	0.062777	0
1	0	0.390658	0
2	1	0.364888	0
3	0	0.876405	1
4	0	0.062777	0

Creating a new column 'Predicted' with 1 if Conversion_Prob>0.35 else 0.

- Confusion Matrix after considering cutoff value of 0.35



- Accuracy : 0.81
- Specificity : 0.80
- Sensitivity: 0.82

Model Evaluation on Test Dataset

Predicting the Conversion Probability and Predicted Column

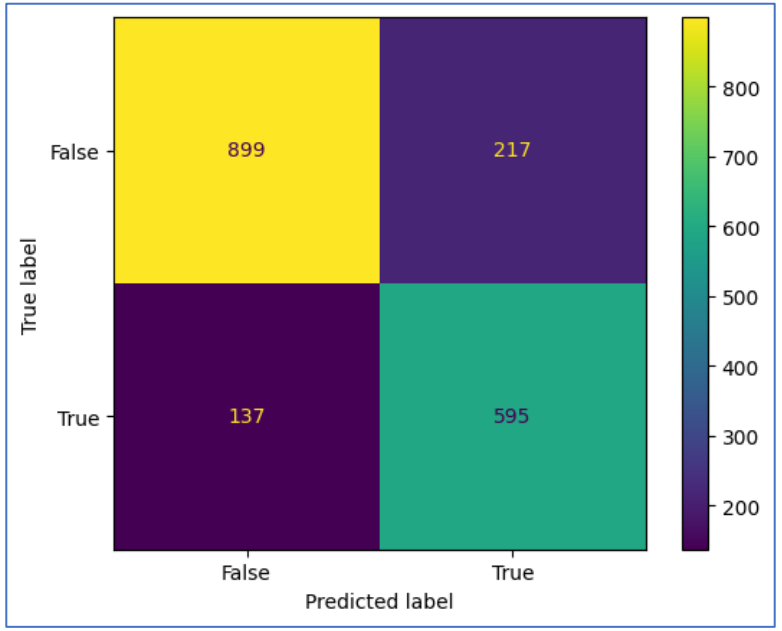
	Converted	Conversion_Prob
0	1	0.712835
1	1	0.917371
2	1	0.933794
3	0	0.062777
4	1	0.899449

Creating a dataframe with the actual Converted flag and the Conversion Probabilities.

	Converted	Conversion_Prob	final_predicted
0	1	0.712835	1
1	1	0.917371	1
2	1	0.933794	1
3	0	0.062777	0
4	1	0.899449	1

Creating a new column 'Predicted' with 1 if Conversion_Prob > 0.35 else 0.

- Confusion Matrix with cutoff value of 0.35:



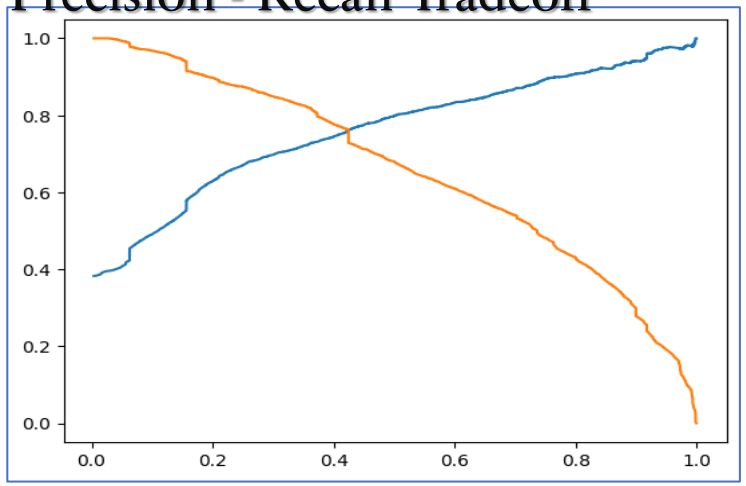
- Accuracy : 0.81
- Specificity :0.80
- Sensitivity: 0.81

Model Evaluation on Train Dataset based on Precision Recall Curve

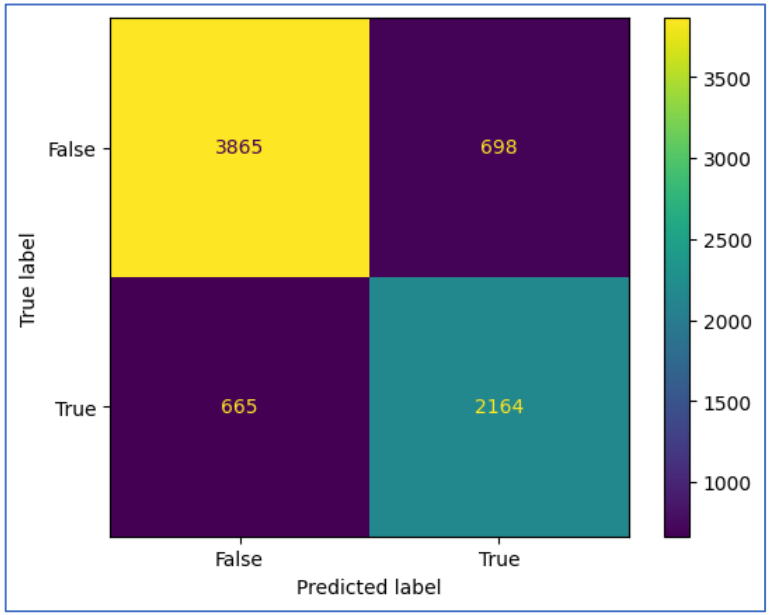
Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead_Score
1	0.062777	0	1	0	0	0	0	0	0	0	0	0	0	6
0	0.390658	0	1	1	1	1	0	0	0	0	0	0	1	39
1	0.364888	0	1	1	1	1	0	0	0	0	0	0	1	36
0	0.876405	1	1	1	1	1	1	1	1	1	1	0	1	88
0	0.062777	0	1	0	0	0	0	0	0	0	0	0	0	6

Creating a new column 'Lead_Score'

Precision - Recall Tradeoff



Confusion Matrix



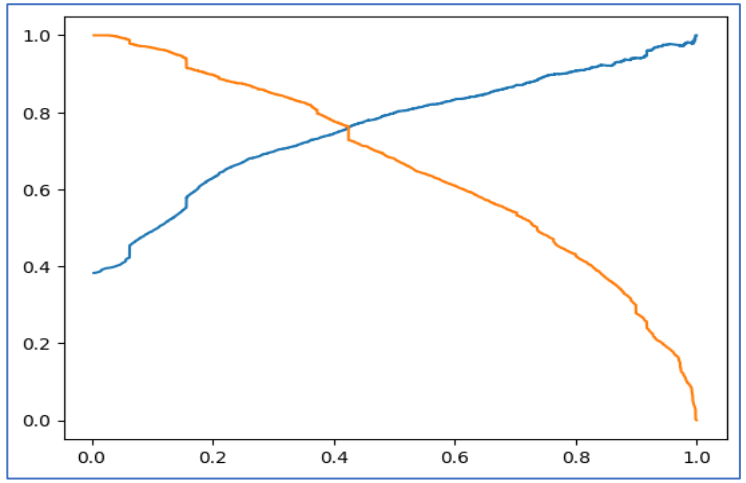
- Accuracy : 0.81
- Specificity : 0.80
- Sensitivity: 0.82
- Precision : 0.72
- Recall : 0.83

Model Evaluation on Test Dataset based on Precision Recall Curve

	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead_Score
0	1	0.062777	0	1	0	0	0	0	0	0	0	0	0	0	6
1	0	0.390658	0	1	1	1	1	0	0	0	0	0	0	1	39
2	1	0.364888	0	1	1	1	1	0	0	0	0	0	0	1	36
3	0	0.876405	1	1	1	1	1	1	1	1	1	1	0	1	88
4	0	0.062777	0	1	0	0	0	0	0	0	0	0	0	0	6

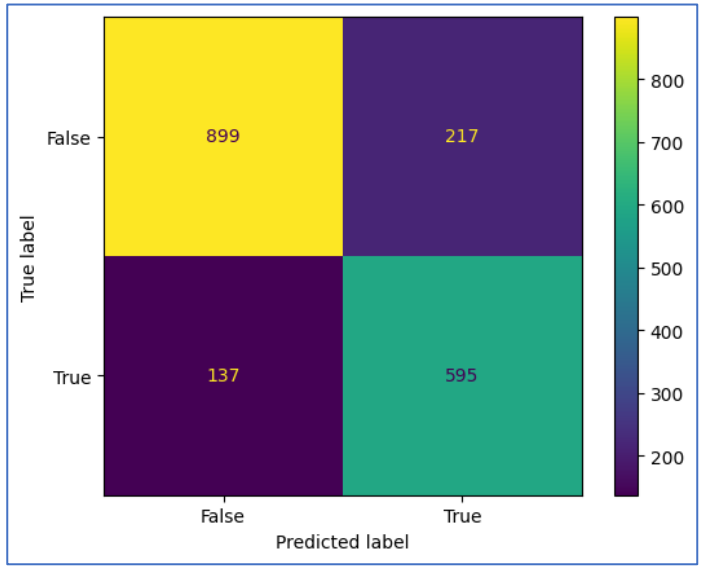
Creating a new column 'Lead_Score'

Precision - Recall Tradeoff



0.42 is found to be the optimum point for cutoff probability.

Confusion Matrix



- Accuracy : 0.81
- Specificity : 0.80
- Sensitivity: 0.81
- Precision : 0.73
- Recall : 0.82

Hot Leads and Cold Leads

Hot Leads Sample:

	Converted	Conversion_Prob	final_predicted	Lead_Score
1	1	0.917371	1	92
2	1	0.933794	1	93
4	1	0.899449	1	90
13	1	0.943718	1	94
15	1	0.899449	1	90
33	1	0.993286	1	99
40	1	0.984659	1	98
49	1	0.989899	1	99
88	1	0.987033	1	99
92	1	0.922298	1	92
93	1	0.972636	1	97
94	1	0.991681	1	99
96	1	0.993286	1	99
140	1	0.917371	1	92
157	1	0.981615	1	98
165	1	0.940554	1	94
187	1	0.930731	1	93
193	1	0.935046	1	94
196	1	0.997961	1	100
212	1	0.900691	1	90

Cold Leads Sample:

	Converted	Conversion_Prob	final_predicted	Lead_Score
0	1	0.712835	1	71
3	0	0.062777	0	6
5	1	0.471656	1	47
6	1	0.351978	1	35
7	1	0.218806	0	22
8	0	0.069266	0	7
9	1	0.367921	1	37
10	0	0.349425	0	35
11	0	0.157772	0	16
12	1	0.889274	1	89
14	1	0.184554	0	18
16	1	0.163973	0	16
17	0	0.020950	0	2
18	0	0.056473	0	6
19	1	0.760920	1	76
20	0	0.307081	0	31
21	1	0.536428	1	54
22	0	0.299150	0	30
23	0	0.067225	0	7
24	1	0.442929	1	44

Recommendations

Company should focus on leads with

a. Lead Origin_Landing Page Submission

b. Lead Origin_Lead Add Form

c. Lead Source_Welingak Website

d. High Total_Visits

e. High Time spent on the website

f. High count on page views per visit

g. current occupation as "Working Professionals“

Company should ignore leads

a. who opted for "Don't Email","Don't Call"

b. who are unemployed

c. with last notable activity "Olark Chat Conversation“

The Company can decrease the Probability cut-off inorder to increase the lead conversion aggressively else they can increase the cut-off inorder to reduce the rate of hiring

Conclusion

With the help of the final model an accuracy of 81%, a sensitivity/recall of 81%, a precision of 73% and a specificity of 80% were achieved. The model has achieved a sensitivity of 81% which is at par with the desired 80% conversion rate of the X Education Company.