

Lead Scoring Case Study


by- Yash Kumar Roy
Yash Pandey



Problem Statement

X Education sell online course to industry professionals. 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.

The conversion rate is around 30% which is not satisfactory for the company so they want to increase that conversion rate which helps company to grow more and do more profit.



Dataset

You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

Which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out are the levels present in the categorical variables.

Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

This is the overview of the Dataset

#	Column	Non-Null Count	Dtype
0	Prospect ID	9240 non-null	object
1	Lead Number	9240 non-null	int64
2	Lead Origin	9240 non-null	object
3	Lead Source	9204 non-null	object
4	Do Not Email	9240 non-null	object
5	Do Not Call	9240 non-null	object
6	Converted	9240 non-null	int64
7	TotalVisits	9103 non-null	float64
8	Total Time Spent on Website	9240 non-null	int64
9	Page Views Per Visit	9103 non-null	float64
10	Last Activity	9137 non-null	object
11	Country	6779 non-null	object
12	Specialization	7802 non-null	object
13	How did you hear about X Education	7033 non-null	object
14	What is your current occupation	6550 non-null	object
15	What matters most to you in choosing a course	6531 non-null	object
16	Search	9240 non-null	object
17	Magazine	9240 non-null	object
18	Newspaper Article	9240 non-null	object
19	X Education Forums	9240 non-null	object
20	Newspaper	9240 non-null	object
21	Digital Advertisement	9240 non-null	object
22	Through Recommendations	9240 non-null	object
23	Receive More Updates About Our Courses	9240 non-null	object
24	Tags	5887 non-null	object
25	Lead Quality	4473 non-null	object
26	Update me on Supply Chain Content	9240 non-null	object
27	Get updates on DM Content	9240 non-null	object
28	Lead Profile	6531 non-null	object
29	City	7820 non-null	object
30	Asymmetrique Activity Index	5022 non-null	object
31	Asymmetrique Profile Index	5022 non-null	object
32	Asymmetrique Activity Score	5022 non-null	float64
33	Asymmetrique Profile Score	5022 non-null	float64
34	I agree to pay the amount through cheque	9240 non-null	object
35	A free copy of Mastering The Interview	9240 non-null	object
36	Last Notable Activity	9240 non-null	object

EDA

(Exploratory Data Analysis)



df[object_columns].isna().sum()	
Prospect ID	0
Lead Origin	0
Lead Source	36
Do Not Email	0
Do Not Call	0
Last Activity	103
Country	2461
Specialization	3380
How did you hear about X Education	7250
What is your current occupation	2690
What matters most to you in choosing a course	2709
Search	0
Magazine	0
Newspaper Article	0
X Education Forums	0
Newspaper	0
Digital Advertisement	0
Through Recommendations	0
Receive More Updates About Our Courses	0
Tags	3353
Lead Quality	4767
Update me on Supply Chain Content	0
Get updates on DM Content	0
Lead Profile	6855
City	3669
Asymmetrique Activity Index	4218
Asymmetrique Profile Index	4218
I agree to pay the amount through cheque	0
A free copy of Mastering The Interview	0
Last Notable Activity	0

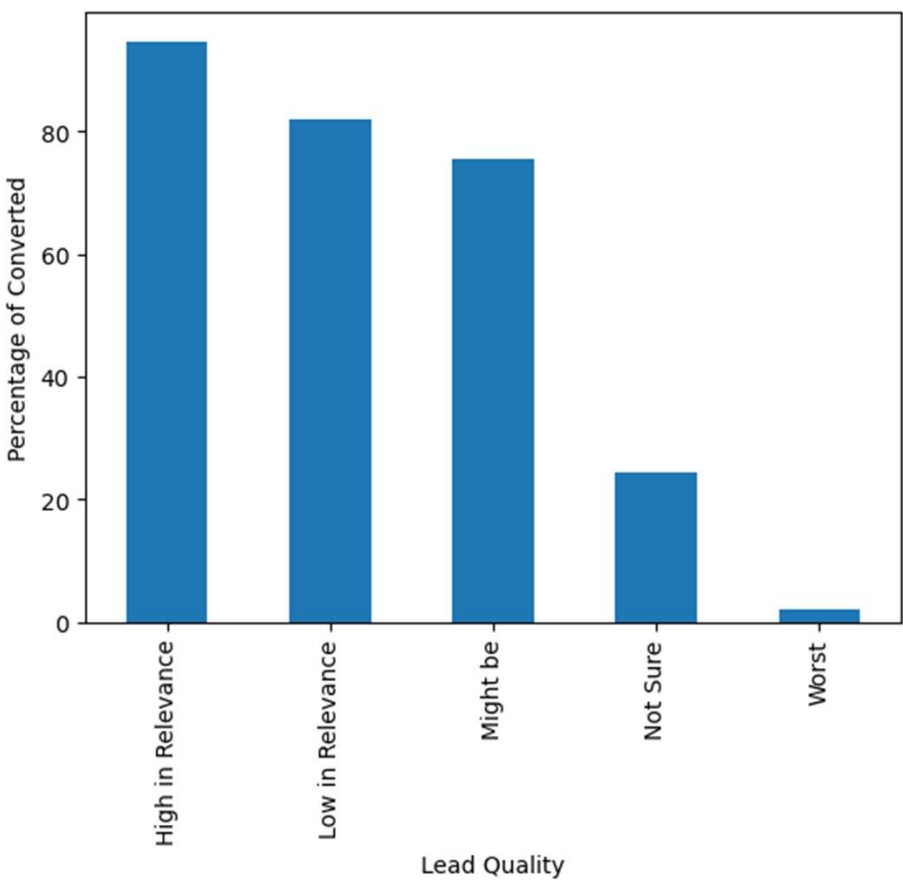
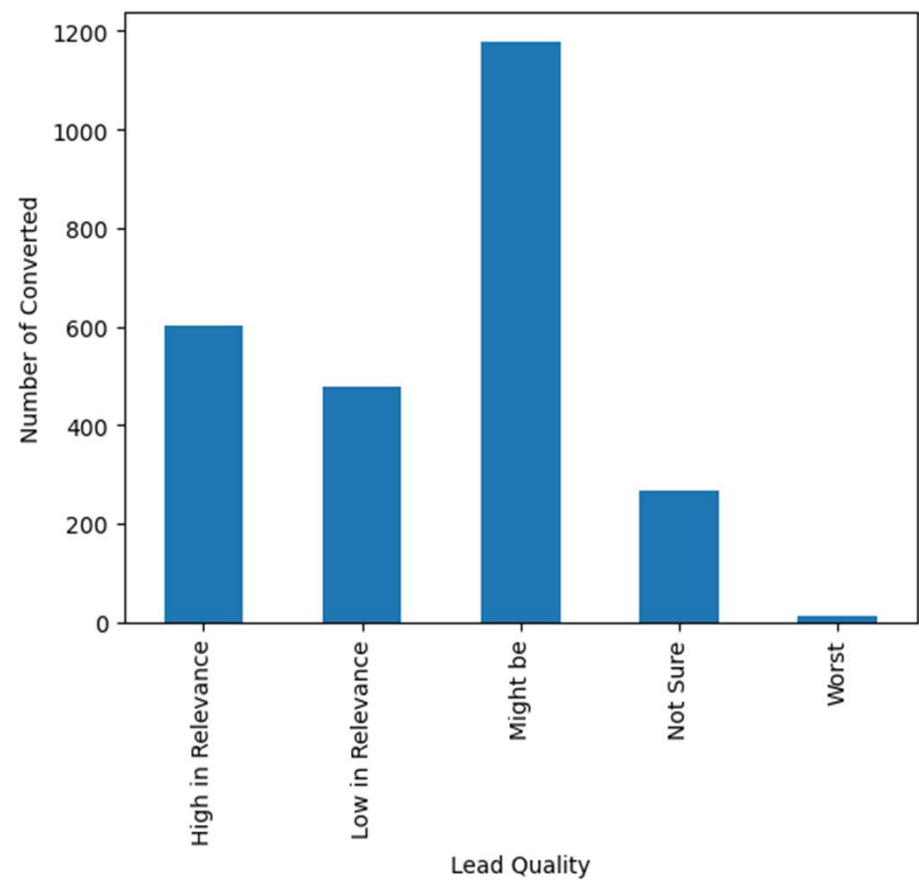
← Is showing the number of null Values in the columns

Is showing the percentage of null values in the columns →

round(df.isna().sum()/df.shape[0]*100,2)	
Prospect ID	0.00
Lead Number	0.00
Lead Origin	0.00
Lead Source	0.39
Do Not Email	0.00
Do Not Call	0.00
Converted	0.00
TotalVisits	1.48
Total Time Spent on Website	0.00
Page Views Per Visit	1.48
Last Activity	1.11
Country	26.63
Specialization	36.58
How did you hear about X Education	78.46
What is your current occupation	29.11
What matters most to you in choosing a course	29.32
Search	0.00
Magazine	0.00
Newspaper Article	0.00
X Education Forums	0.00
Newspaper	0.00
Digital Advertisement	0.00
Through Recommendations	0.00
Receive More Updates About Our Courses	0.00
Tags	36.29
Lead Quality	51.59
Update me on Supply Chain Content	0.00
Get updates on DM Content	0.00
Lead Profile	74.19
City	39.71
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65
I agree to pay the amount through cheque	0.00
A free copy of Mastering The Interview	0.00
Last Notable Activity	0.00

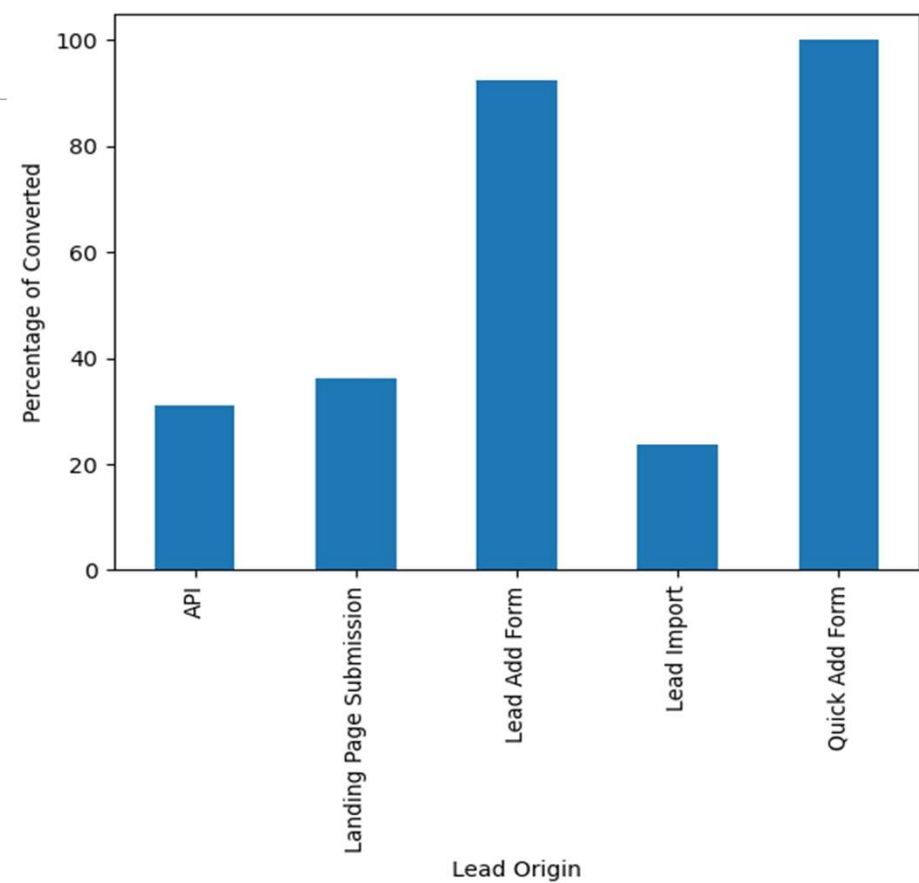
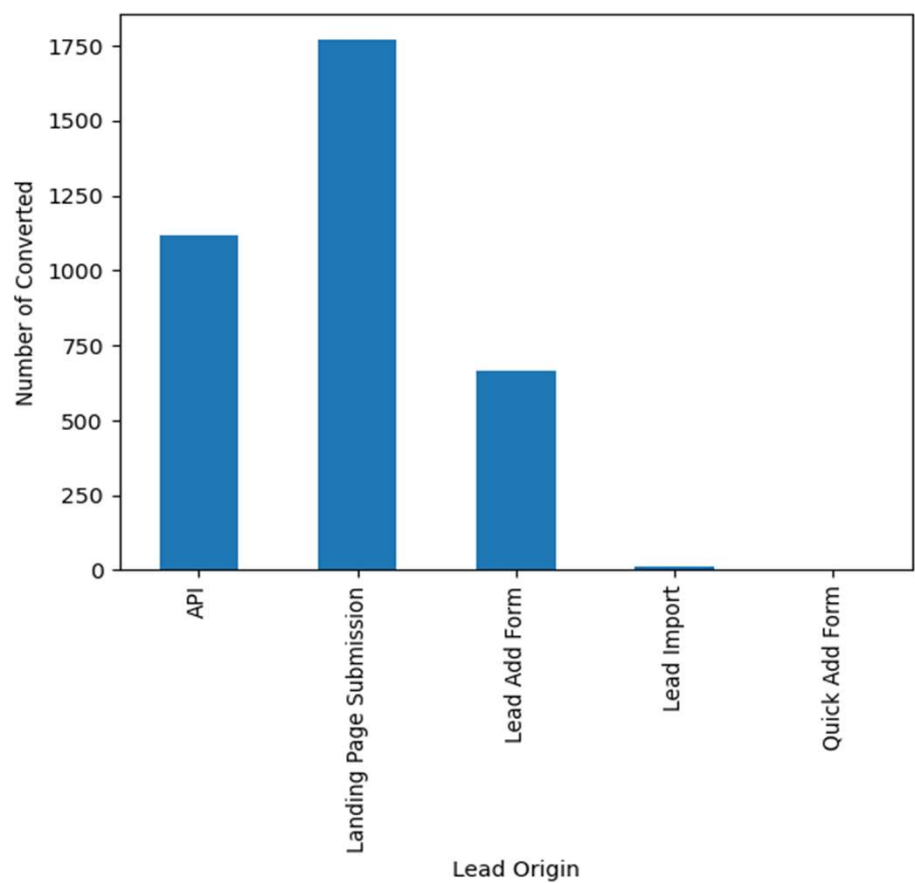
Distribution of Lead Quality Vs Converted

We can see that Lead Quality with Might be are the highest in number but conversion rate of High in Relevance and for the Low in Relevance and Might be are having also good conversion rate.



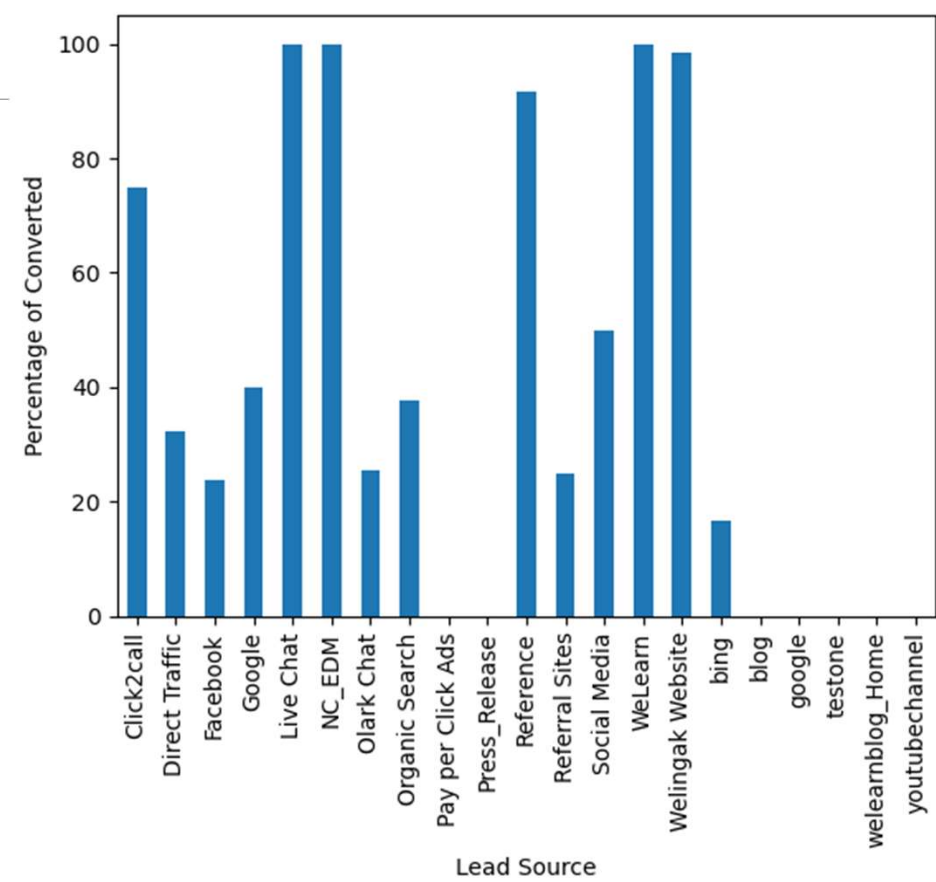
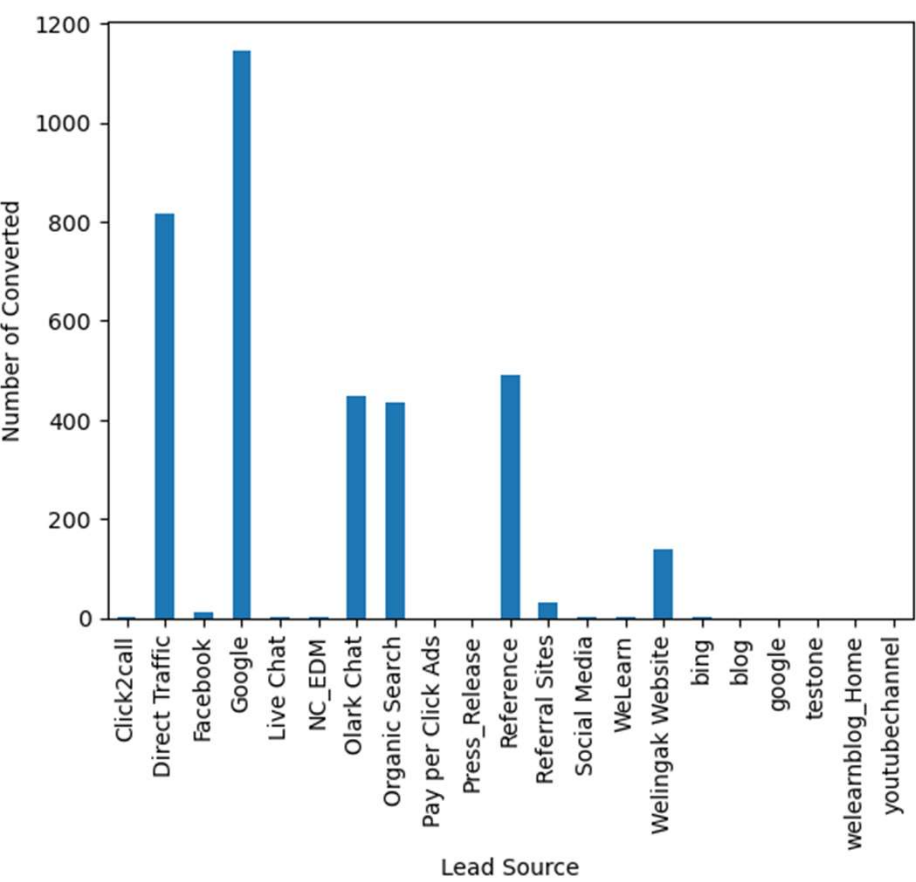
Distribution of Lead Origin Vs Converted

We can see that Lead Origin who are Landing Page Submission have the highest number of customers but very low in conversion rate. Lead Add From are having few number but high conversion Rate.



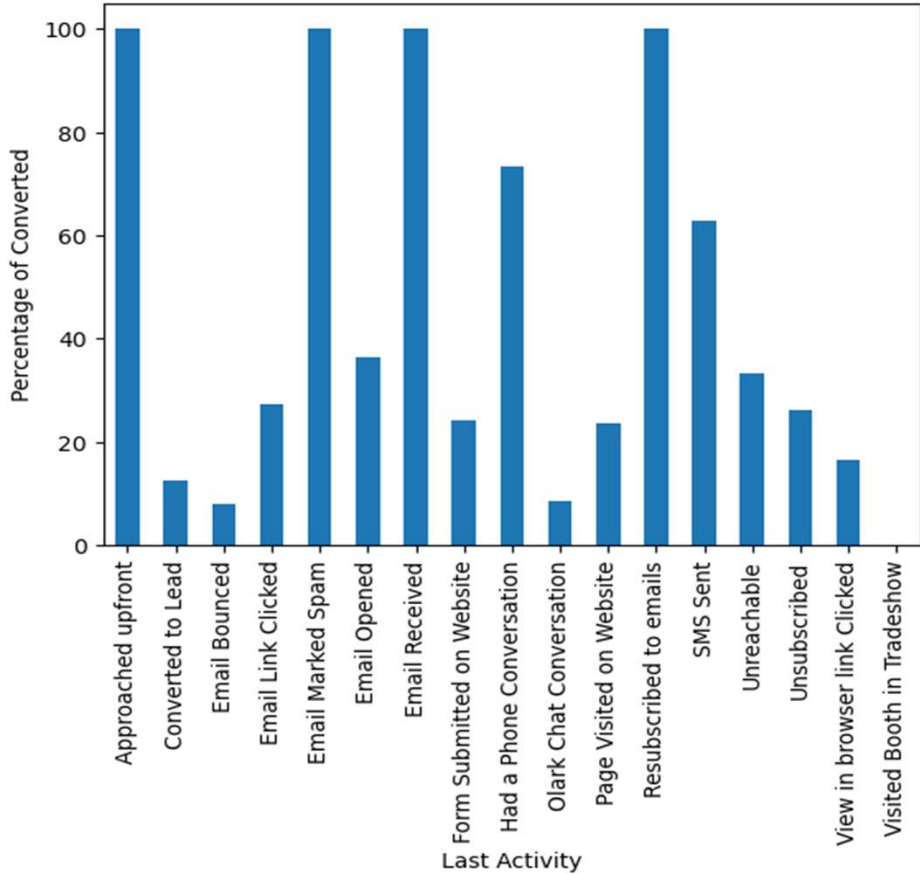
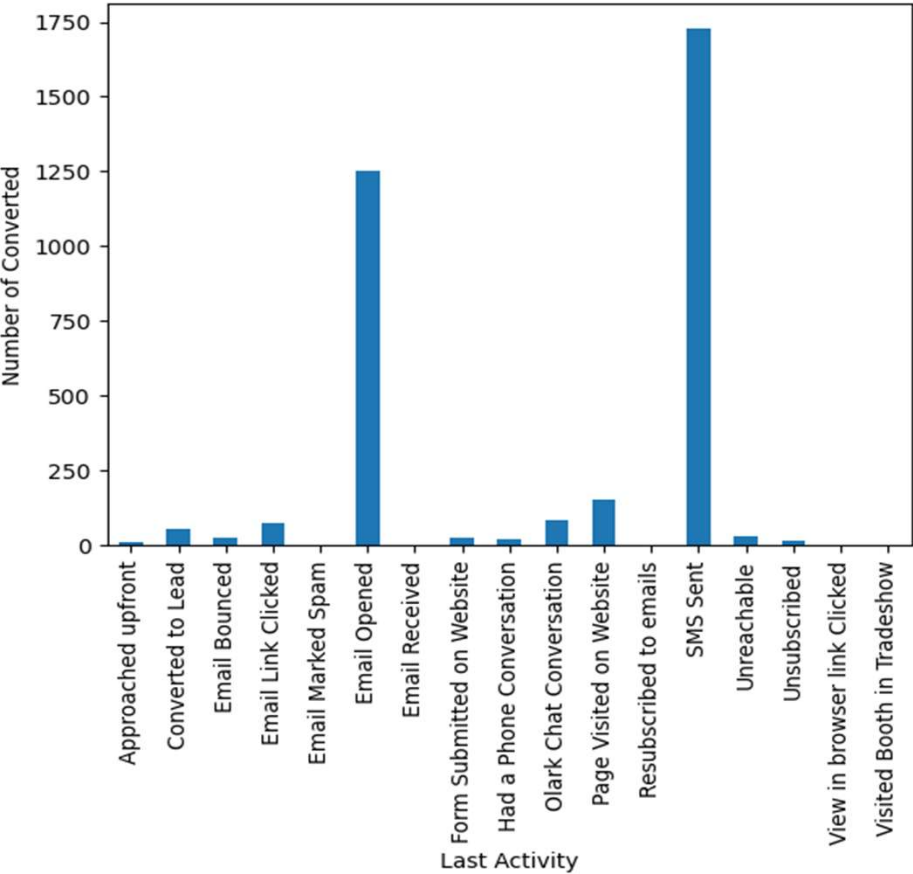
Distribution of Lead Source Vs Converted

Here is the distribution of Number of converted and Percentage of Converted of Lead Source. Here we can't conclude any specific conclusion.



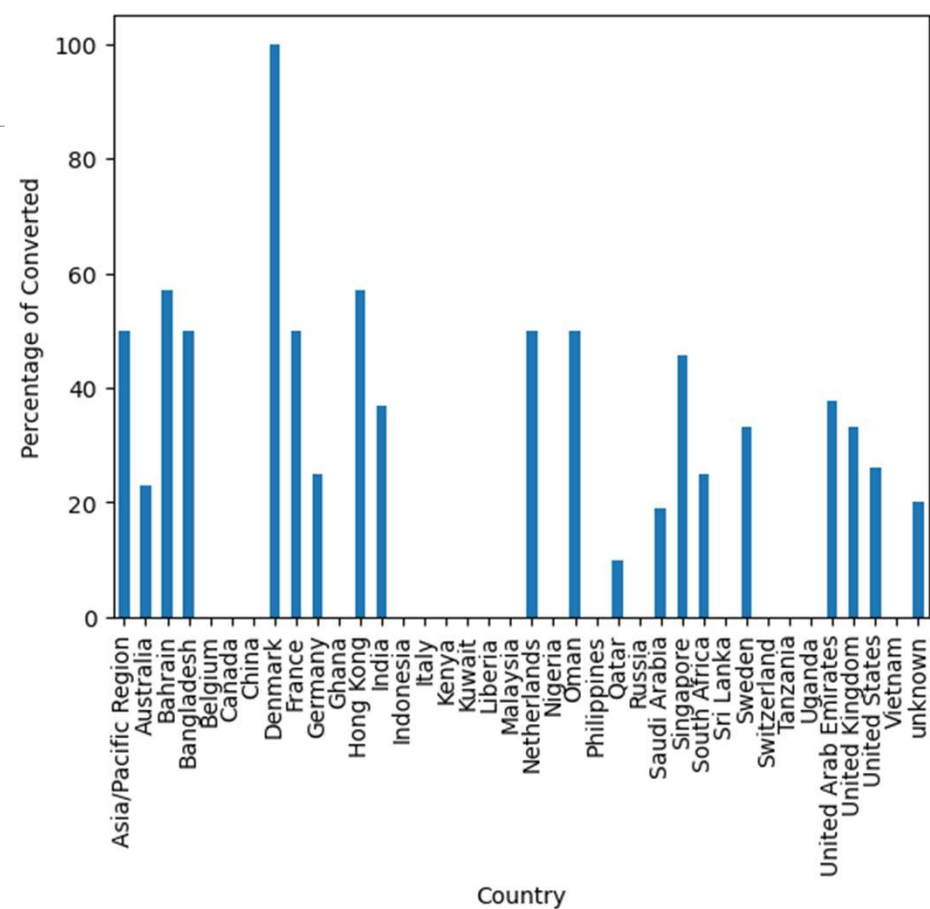
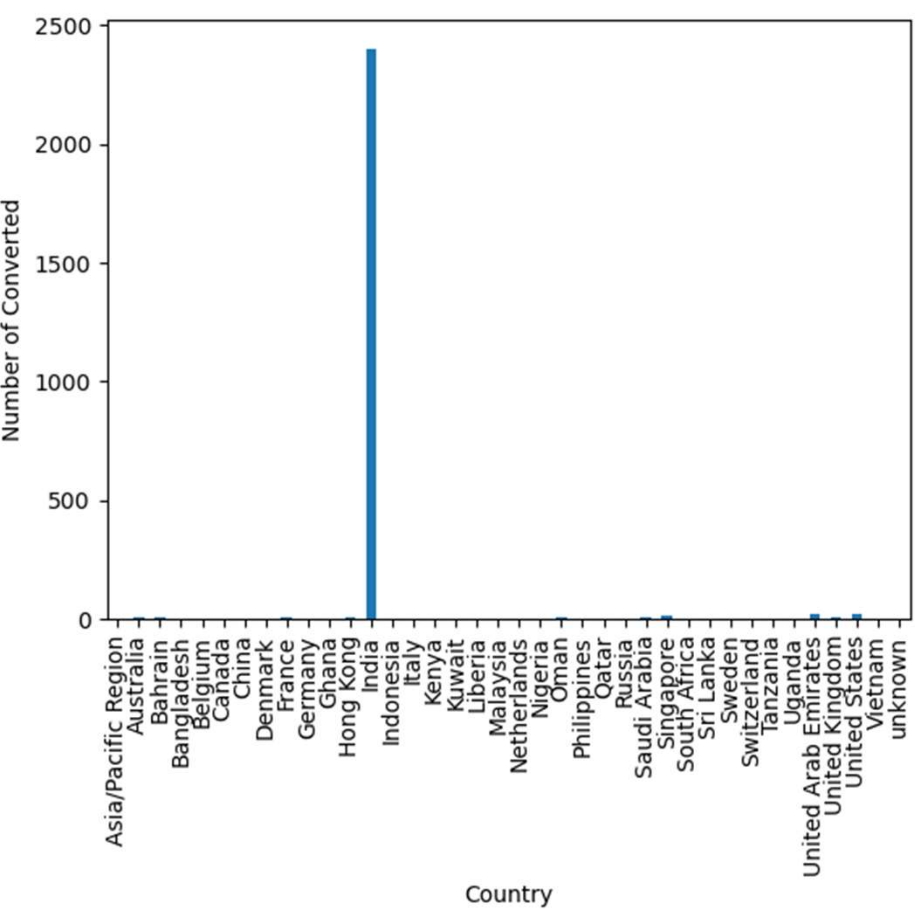
Distribution of Last Activity Vs Converted

Here is the distribution of Number of converted and Percentage of Converted of Last Activity of the Customer. Here we can't conclude any specific conclusion.



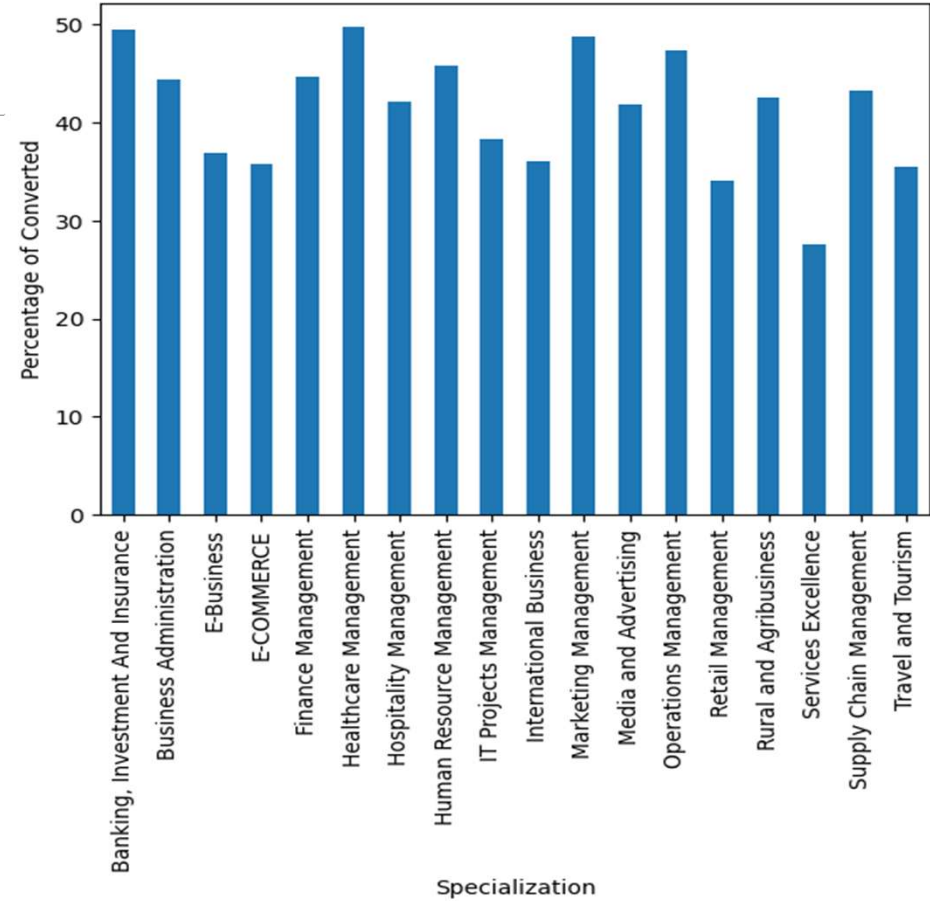
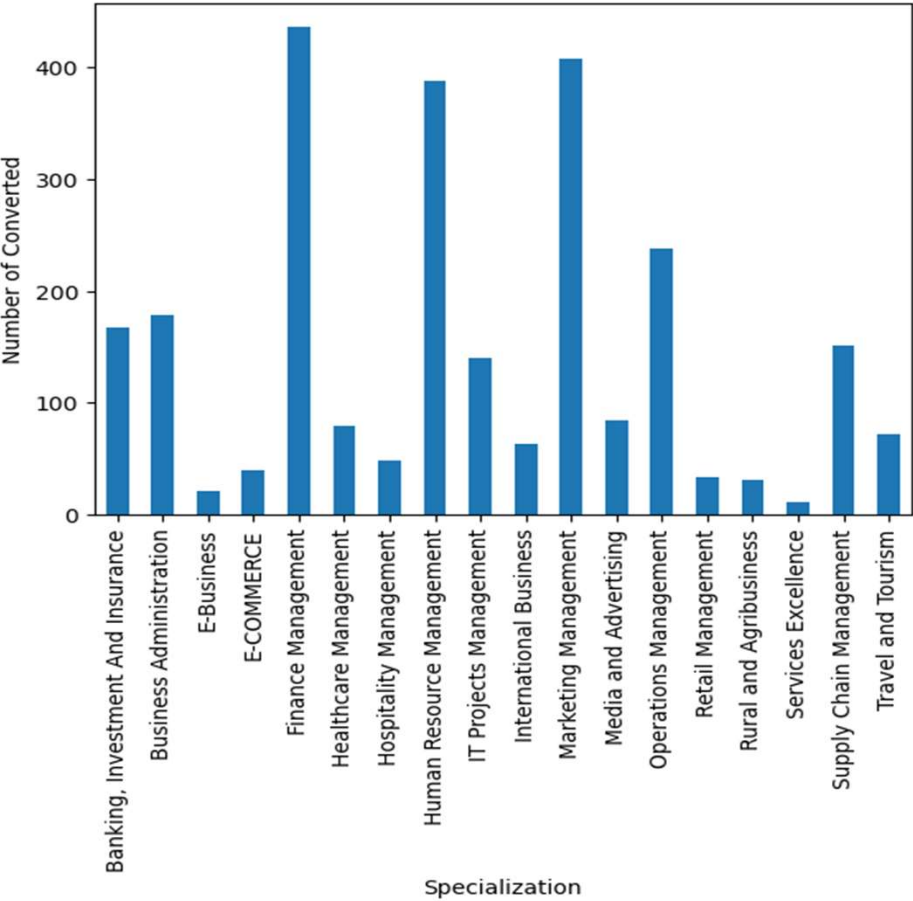
Distribution of Country Vs Converted

Here is the distribution of Number of converted and Percentage of Converted of each country we can clearly see here that the majority of the Customer are coming from India, Hence we can drop this column.



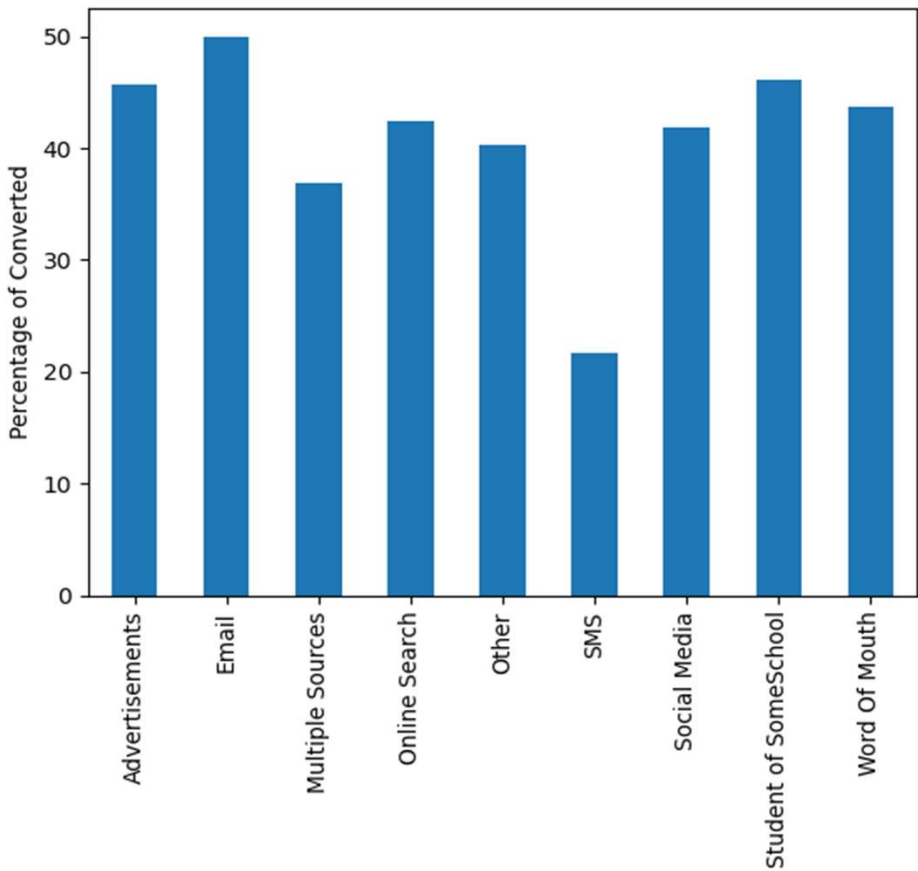
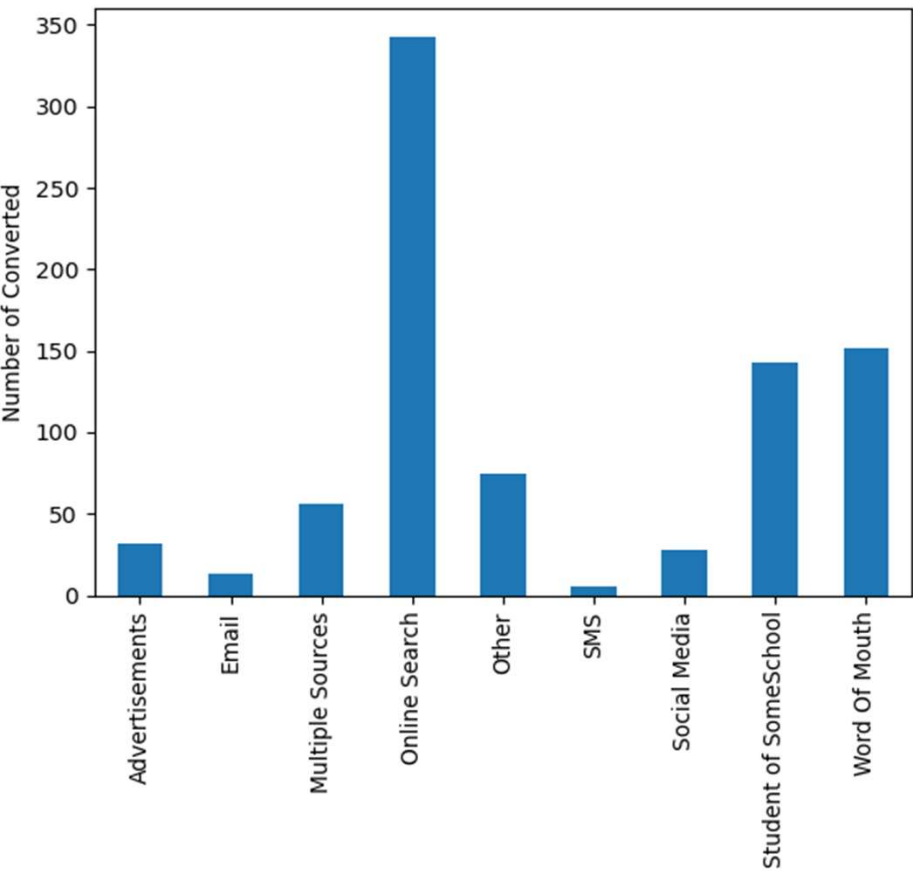
Distribution of Specialization Vs Converted

Here we can see that conversion rate of every Specialization and none of that crossing the line of 50% and only one is below 30% that is Service Excellence.



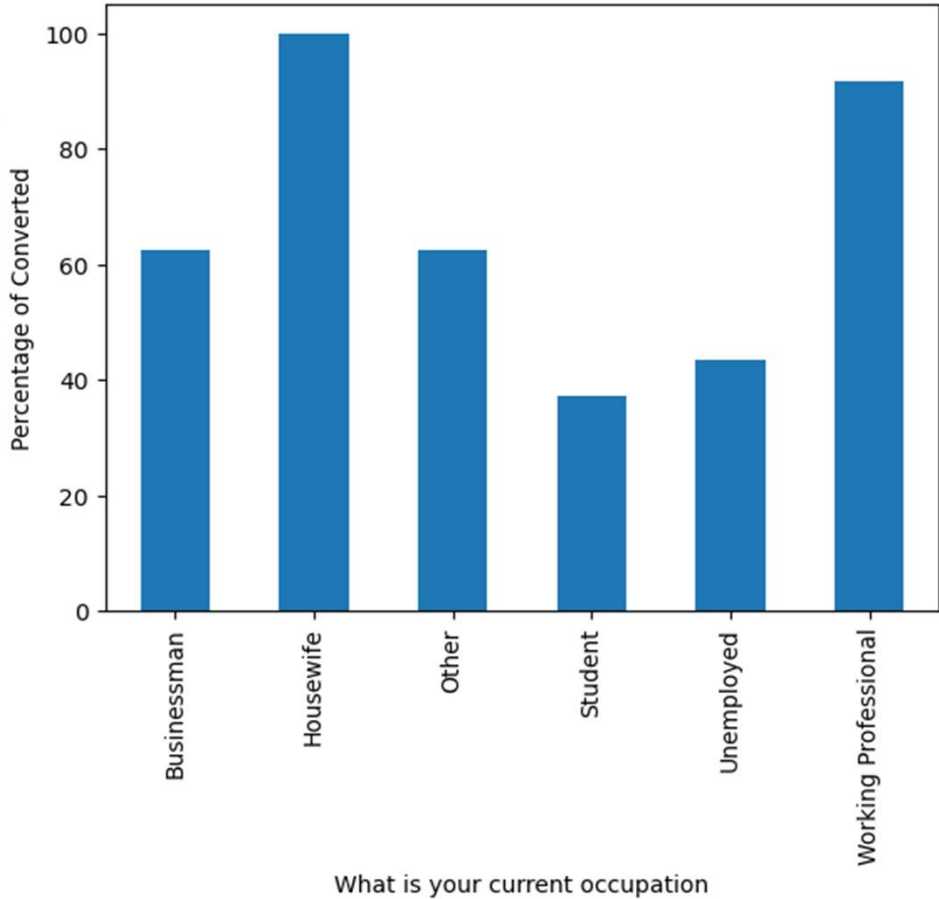
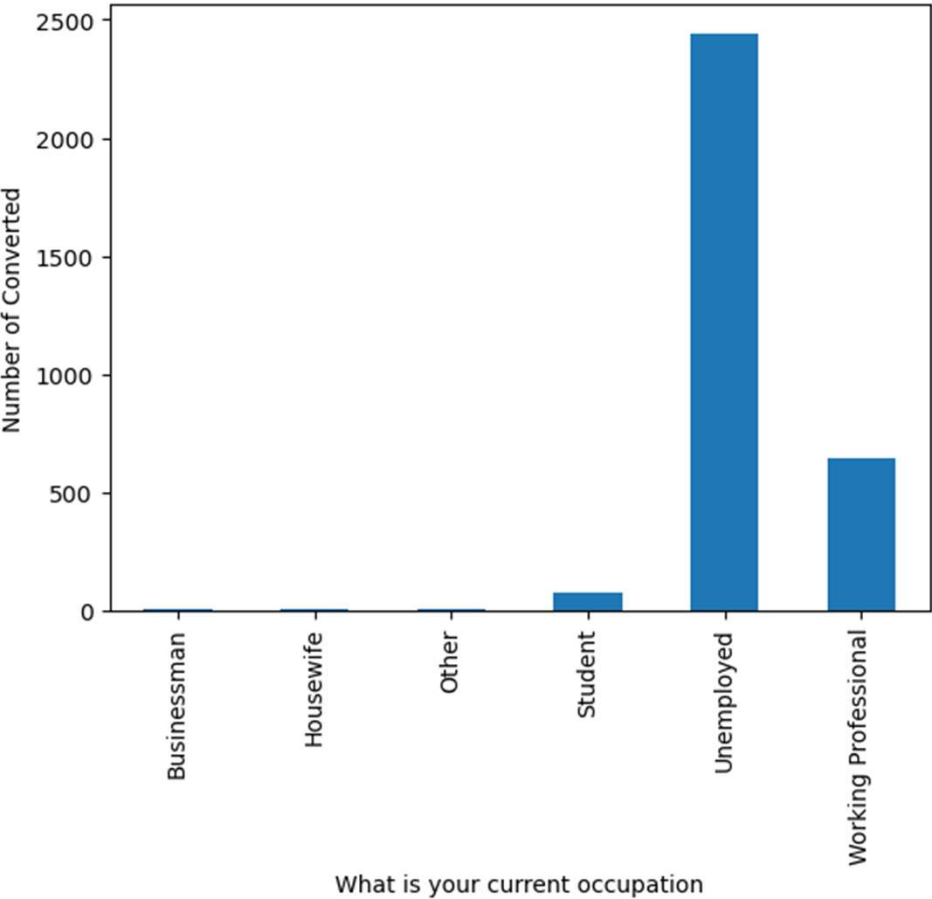
Distribution of How did you hear about X Education Vs Converted

Customer who heard about the X Education are most are from Online Search and have a around 43% of conversion rate.



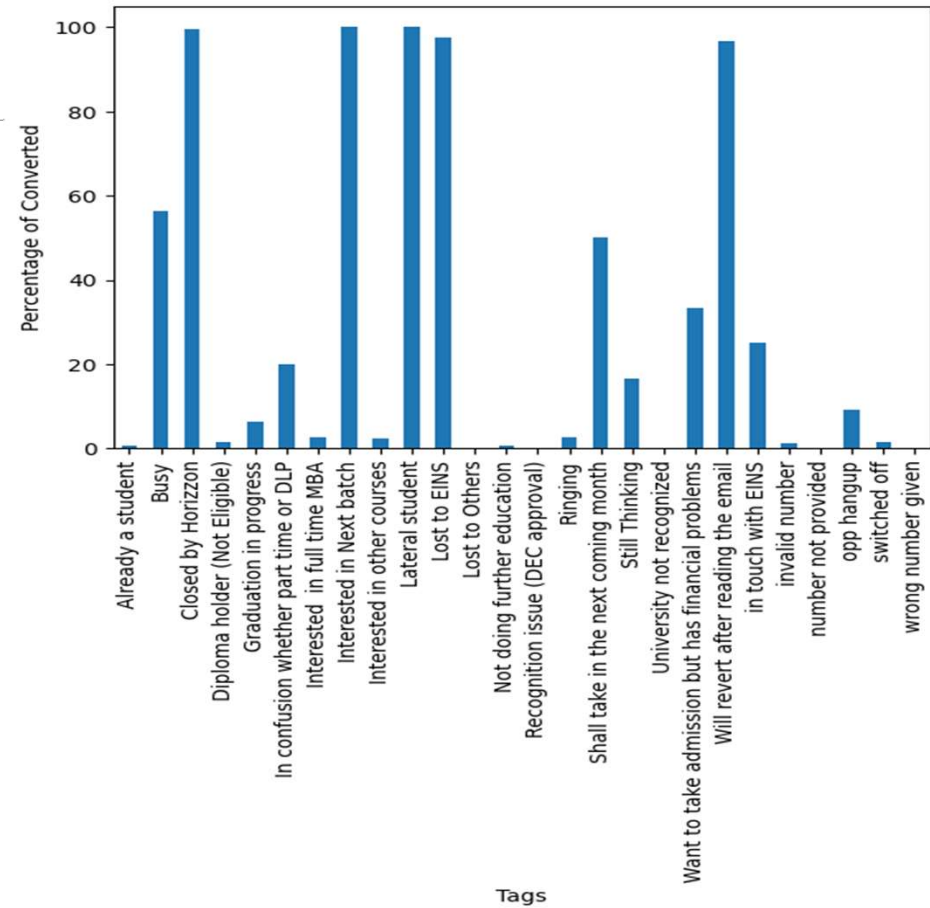
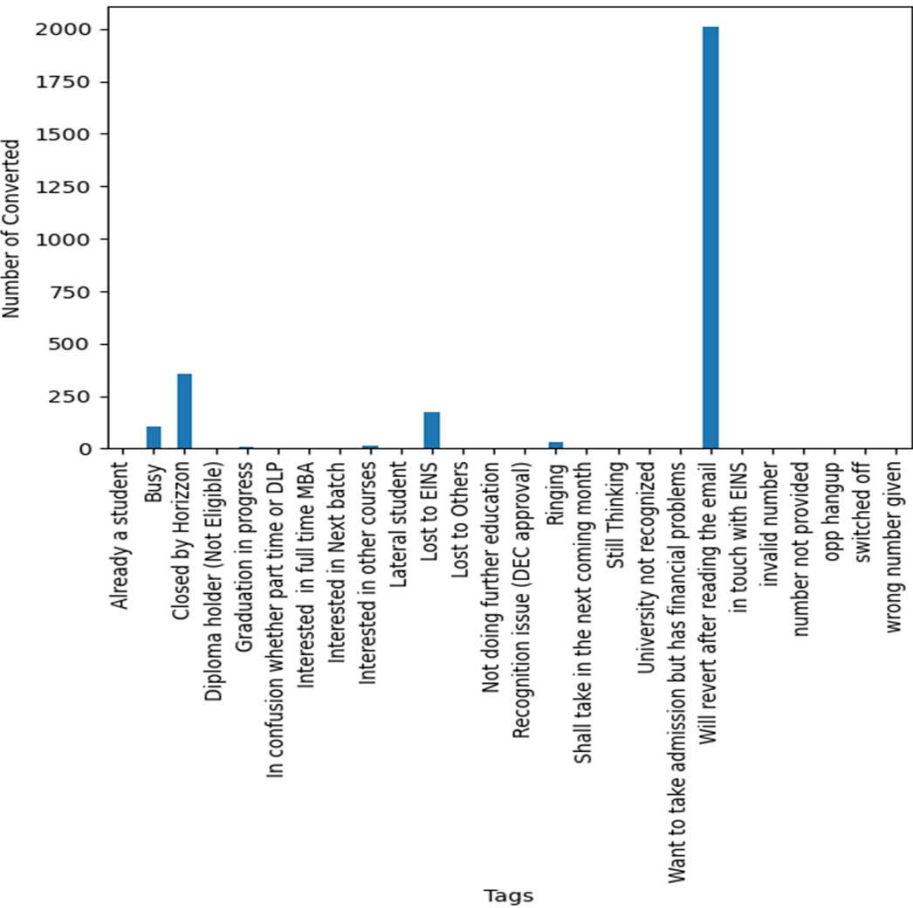
Distribution of What is your current occupation Vs Converted

Here we have most number of customer are unemployed but their conversion rate is low on other hand working professional are low in number but have high in conversion rate.



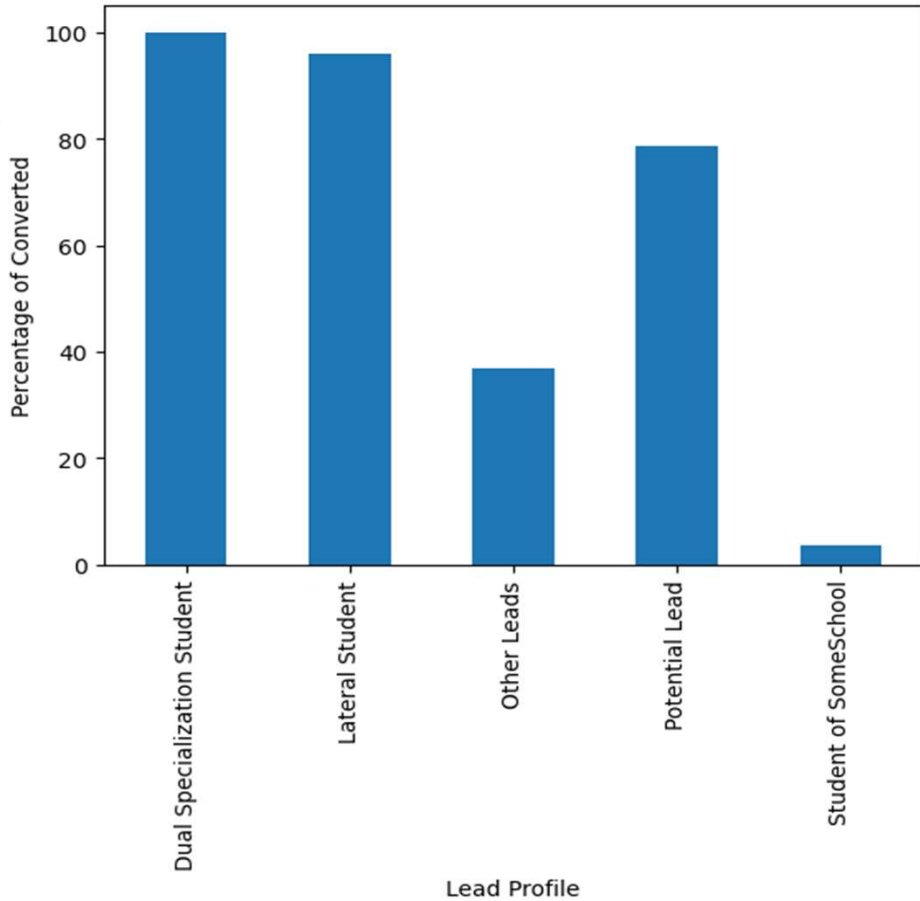
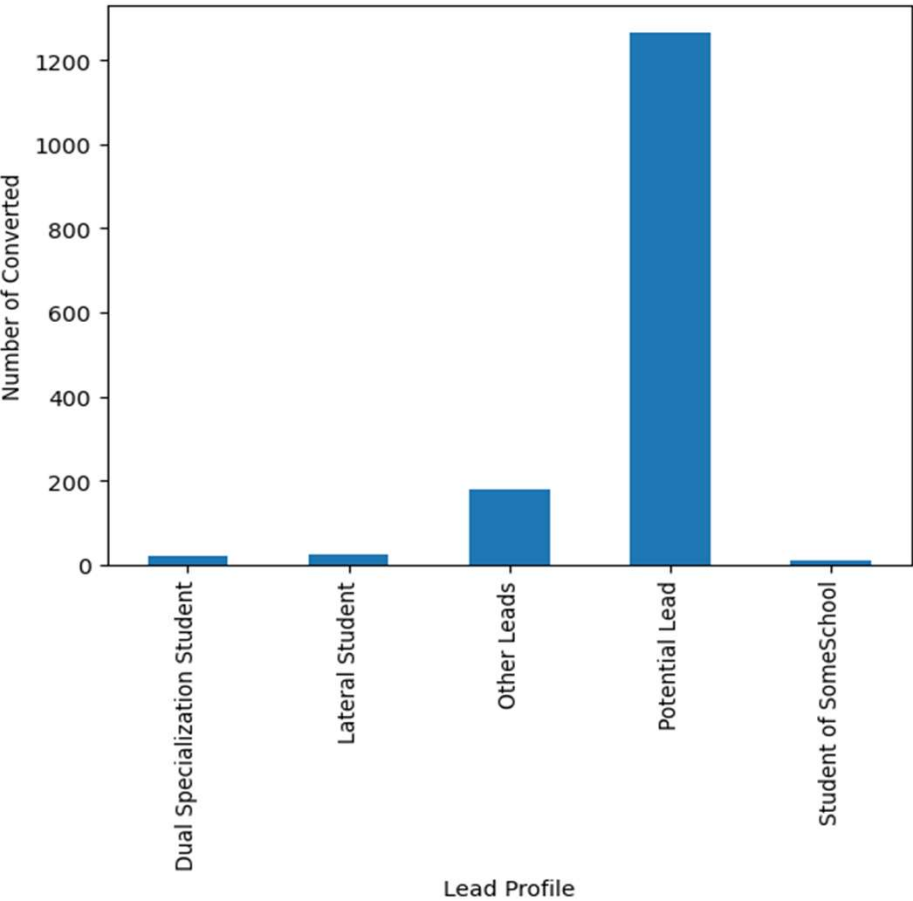
Distribution of Tags Vs Converted

This plots shows the distribution of Tags on the number of converted customer and percentage of converted customer.



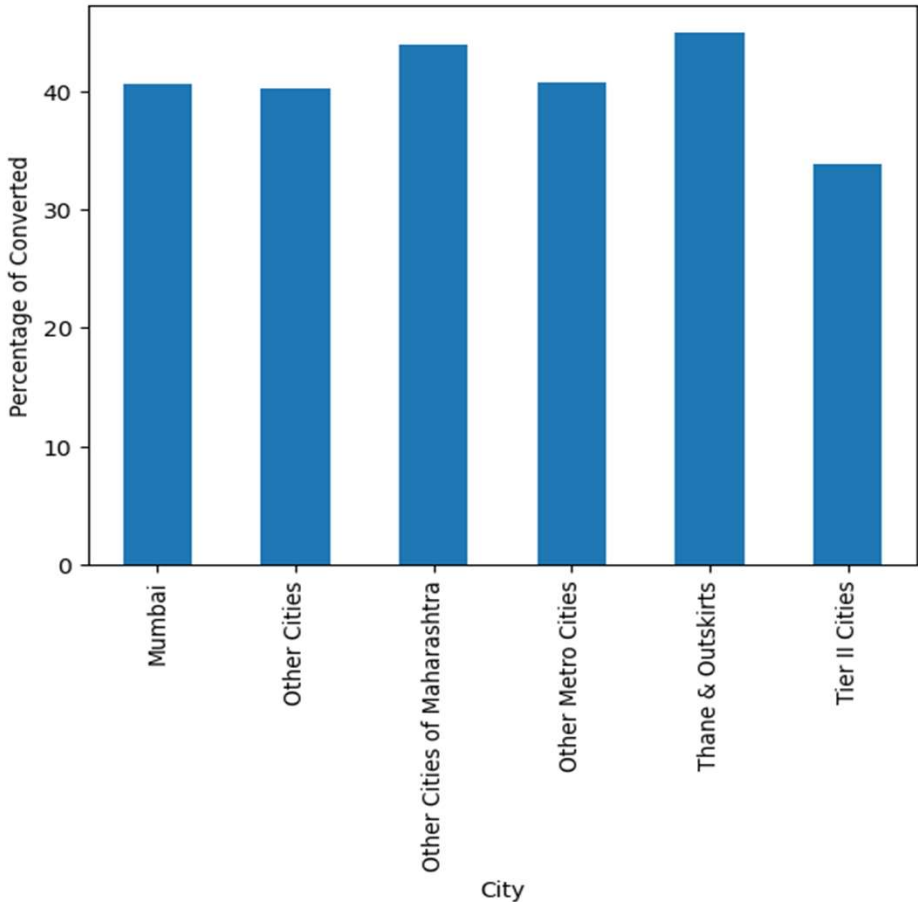
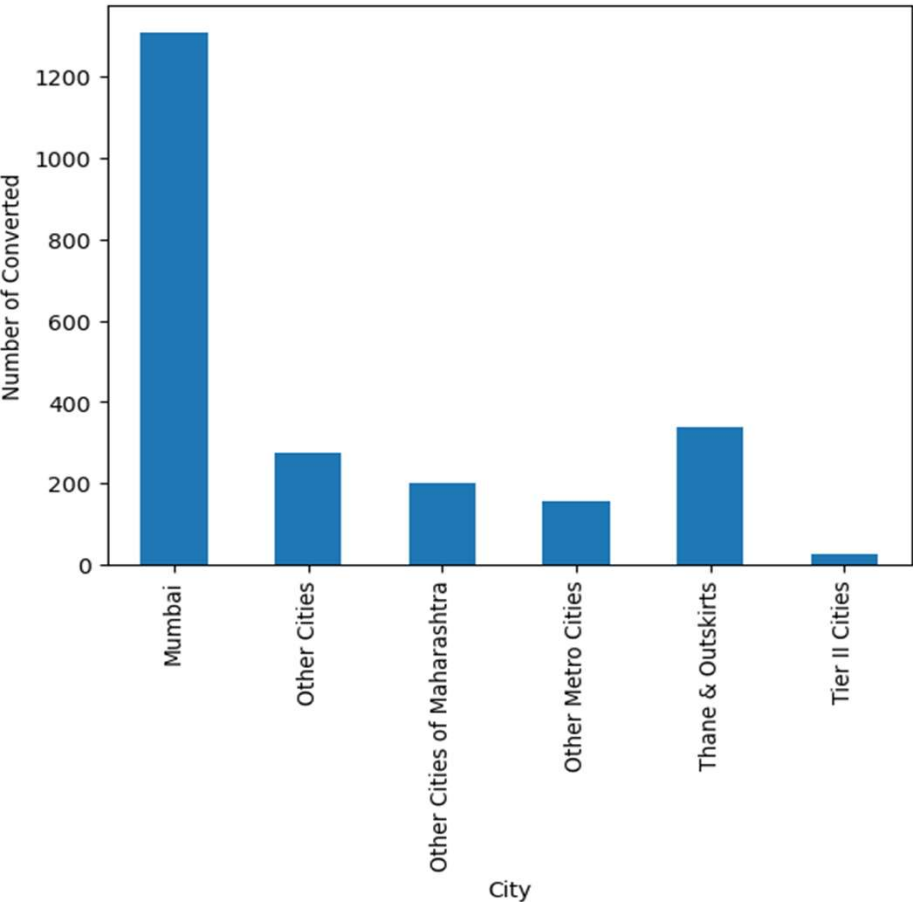
Distribution of Lead Profile Vs Converted

This plots shows the distribution of Lead Profile on the number of converted customer and percentage of converted customer.



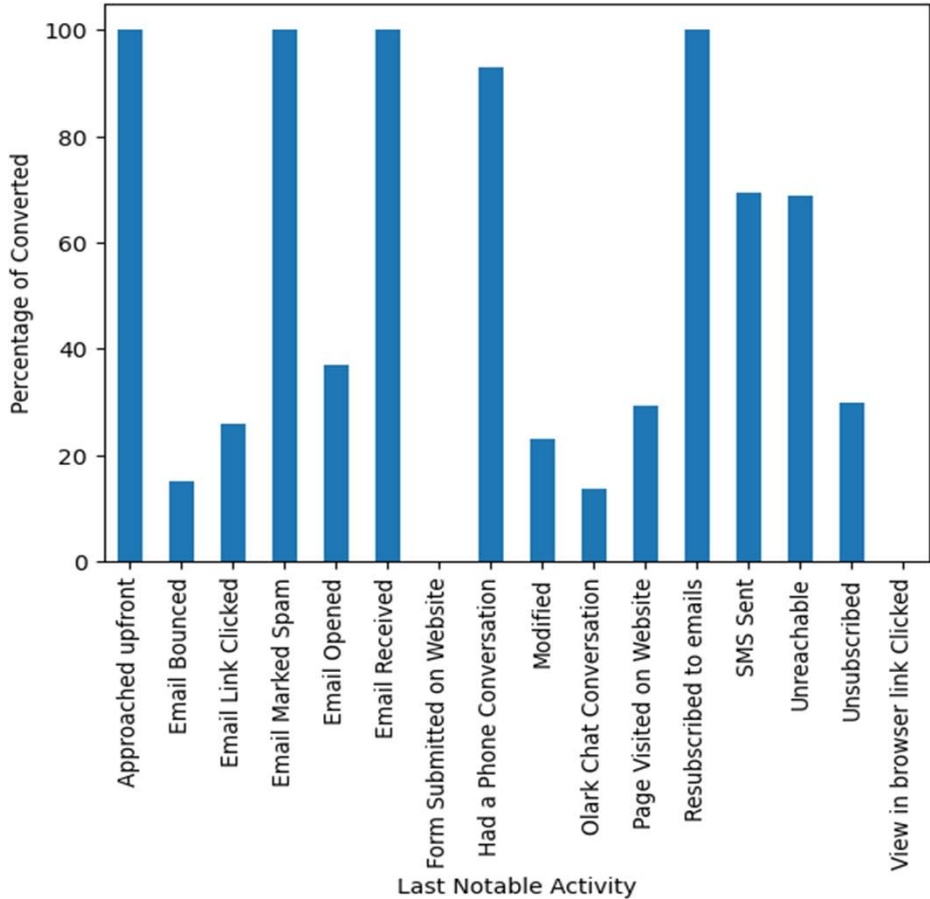
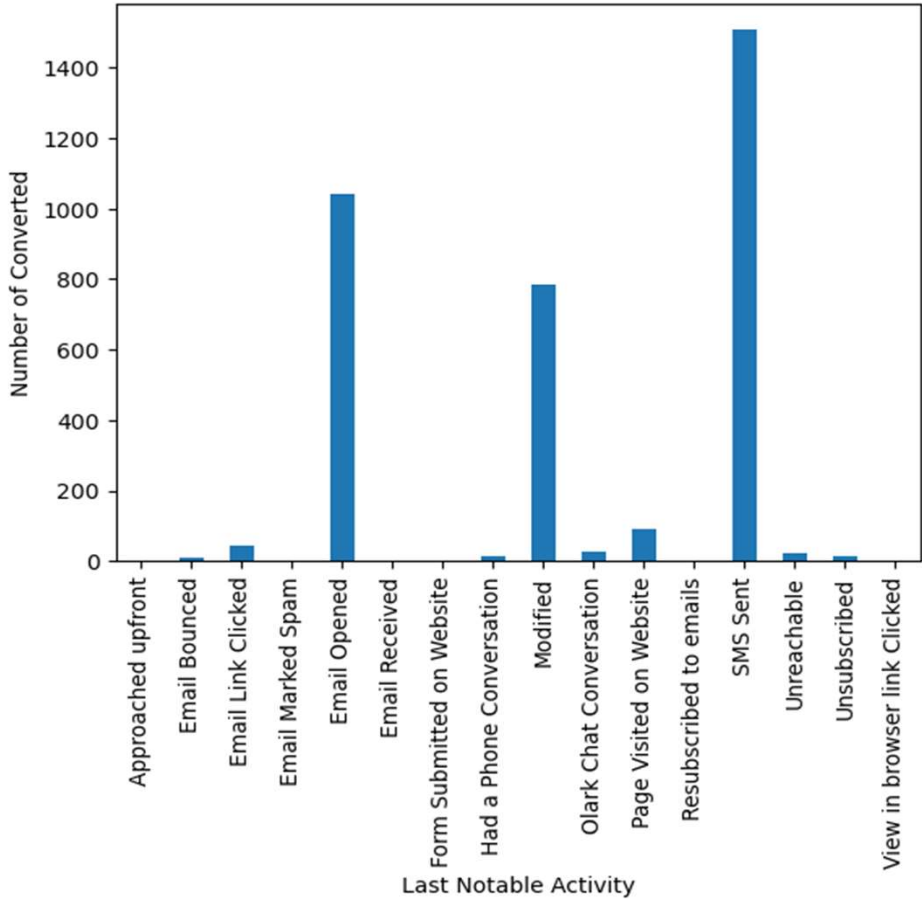
Distribution of City Vs Converted

This plots shows the distribution of City on the number of converted customer and percentage of converted customer.

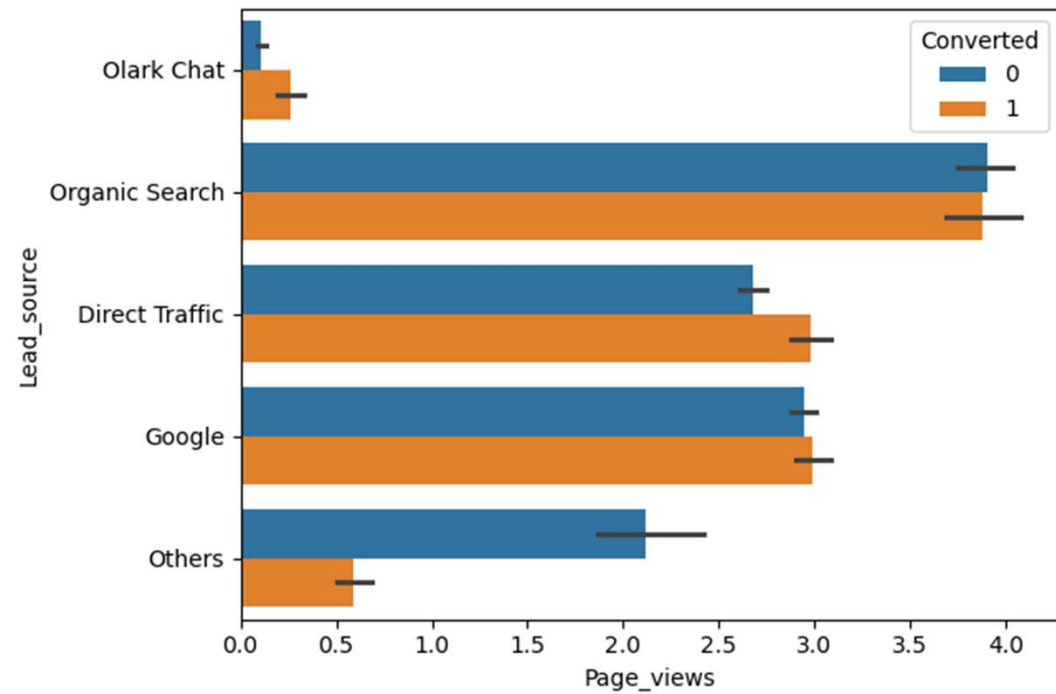


Distribution of Last Notable Activity Vs Converted

This plots shows the distribution of Last Notable Activity on the number of converted customer and percentage of converted customer.

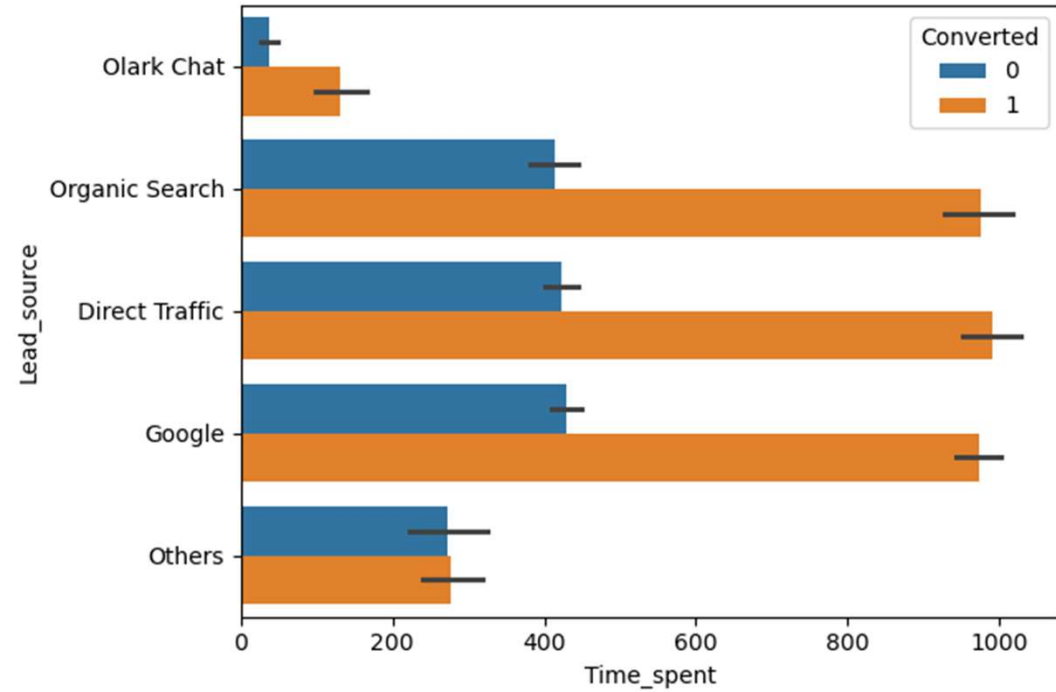


Lead Source Vs Page Views



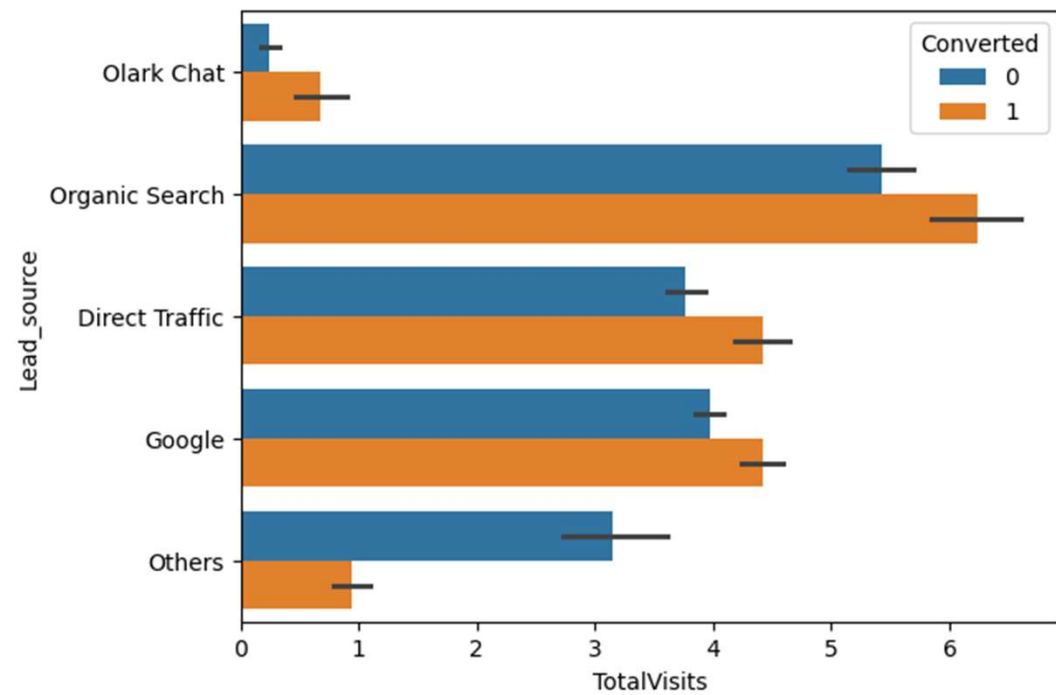
Organic Search are having most number of customer and around 50% of conversion rate also.

Lead Source Vs Time Spend



Organic Search, Direct Traffic and Google have high conversion rate.

Lead Source Vs Total Visits



Organic Search, Direct Traffic and Google have high conversion rate.

Columns Manipulation

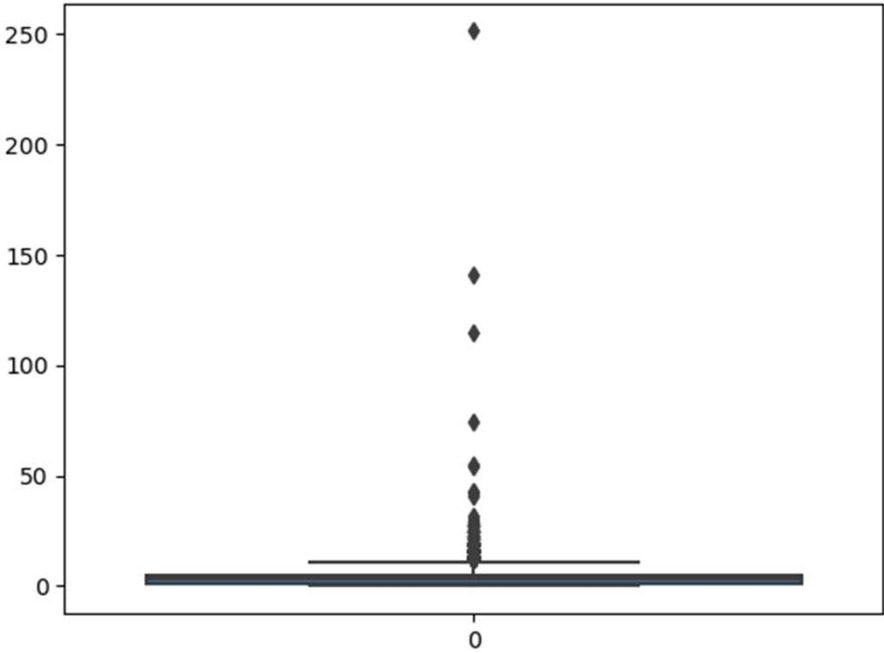
- Dropped those columns which have more then 45% of null values.
- Dropped those columns which has majorly only one category.
- Converted all categorical columns to dummy variable .
- Filling null values with mean if it was a numerical column
- Filling null values with mode if it was a categorical column.

#	Column	Non-Null Count	Dtype
0	Lead Origin	9240 non-null	object
1	Lead Source	9204 non-null	object
2	Converted	9240 non-null	int64
3	TotalVisits	9103 non-null	float64
4	Total Time Spent on Website	9240 non-null	int64
5	Page Views Per Visit	9103 non-null	float64
6	Last Activity	9137 non-null	object
7	Specialization	5860 non-null	object
8	What is your current occupation	6550 non-null	object
9	Tags	5887 non-null	object
10	City	5571 non-null	object
11	A free copy of Mastering The Interview	9240 non-null	object
12	Last Notable Activity	9240 non-null	object

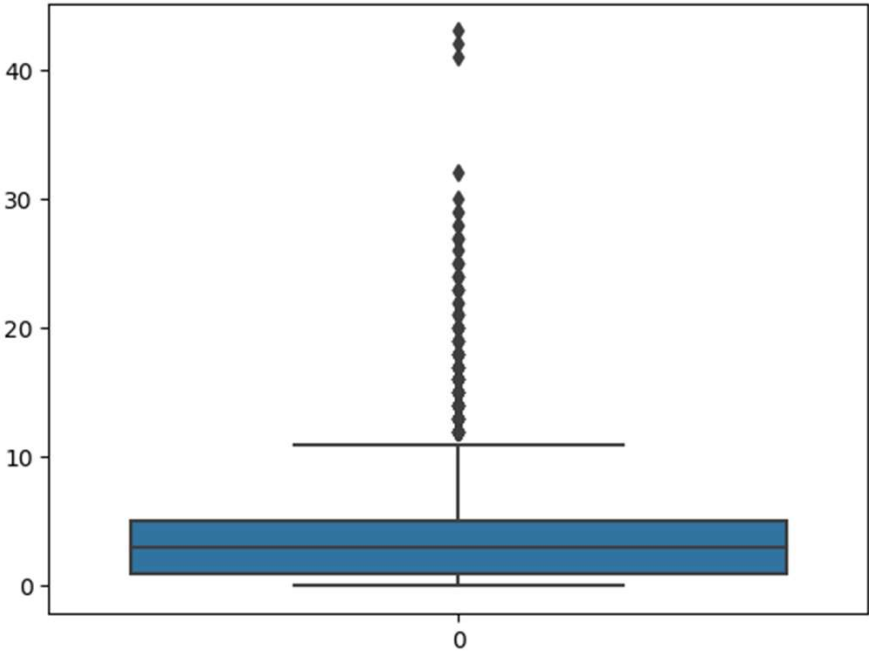
Outliers Treatment

Total Visits

Before:-

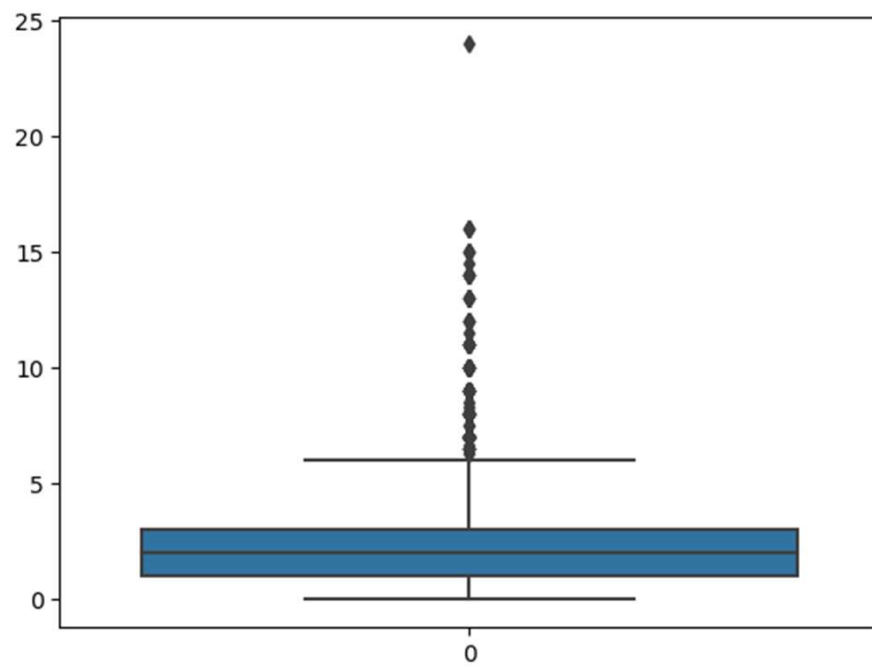


After:-

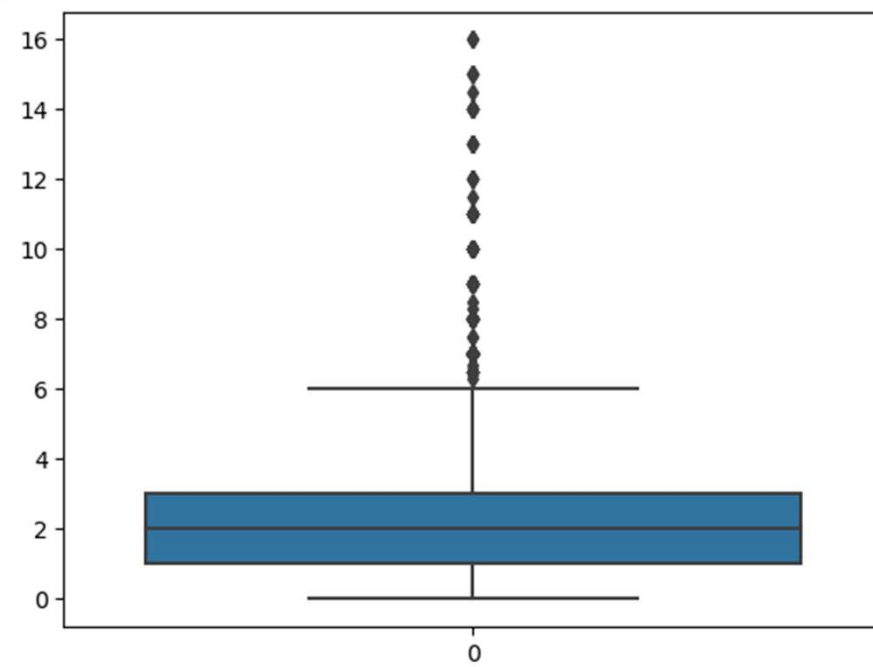


Page Views Per Visit

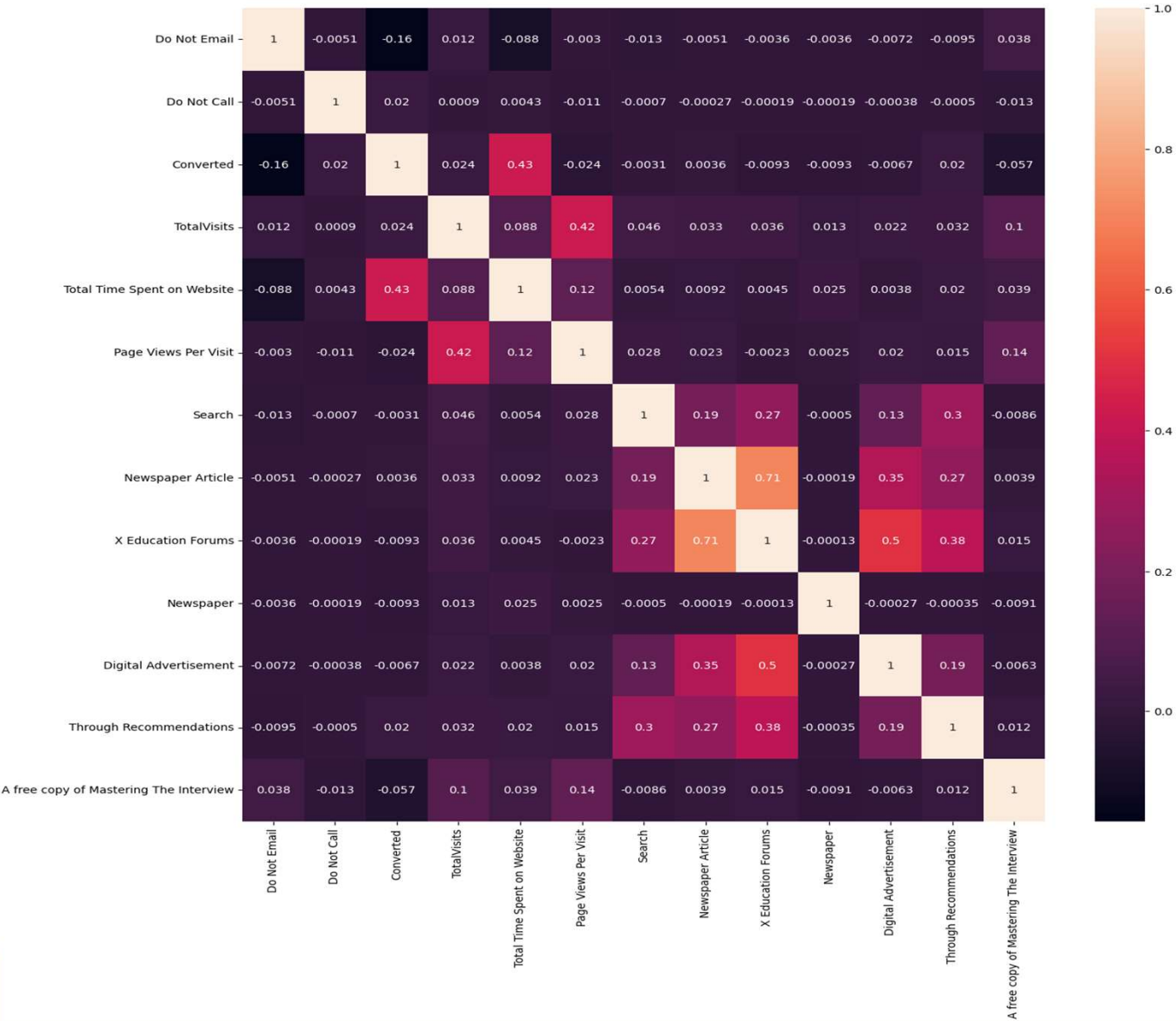
Before:-



After:-



Heatmap with some more variables



VIF Analysis

	feature	VIF
98	Last Notable Activity_Email Marked Spam	inf
106	Last Notable Activity_Resubscribed to emails	inf
37	Last Activity_Resubscribed to emails	inf
30	Last Activity_Email Marked Spam	inf
103	Last Notable Activity_Modified	1130.205825
...
71	Tags_Interested in Next batch	1.023595
69	Tags_In confusion whether part time or DLP	1.020261
81	Tags_University not recognized	1.016872
73	Tags_Lateral student	1.015602
79	Tags_Shall take in the next coming month	1.010650

	feature	VIF
62	What is your current occupation_Unemployed	636.187350
9	Lead Source_Google	365.790019
7	Lead Source_Direct Traffic	324.423963
31	Last Activity_Email Opened	286.511781
37	Last Activity_SMS Sent	219.709183
...
68	Tags_In confusion whether part time or DLP	1.020164
80	Tags_University not recognized	1.016829
72	Tags_Lateral student	1.015587
99	Last Notable Activity_Form Submitted on Website	1.013755
78	Tags_Shall take in the next coming month	1.010610

	feature	VIF
37	Last Activity_Resubscribed to emails	inf
105	Last Notable Activity_Resubscribed to emails	inf
102	Last Notable Activity_Modified	1130.205825
98	Last Notable Activity_Email Opened	1004.486994
63	What is your current occupation_Unemployed	771.843206
...
71	Tags_Interested in Next batch	1.023595
69	Tags_In confusion whether part time or DLP	1.020261
81	Tags_University not recognized	1.016872
73	Tags_Lateral student	1.015602
79	Tags_Shall take in the next coming month	1.010650

	feature	VIF
31	Last Activity_Email Opened	255.411777
9	Lead Source_Google	252.466281
7	Lead Source_Direct Traffic	224.477755
37	Last Activity_SMS Sent	195.822022
13	Lead Source_Organic Search	102.543502
...
79	Tags_University not recognized	1.016774
71	Tags_Lateral student	1.015569
59	What is your current occupation_Housewife	1.015267
98	Last Notable Activity_Form Submitted on Website	1.013732
77	Tags_Shall take in the next coming month	1.010574

	feature	VIF
101	Last Notable Activity_Modified	1130.205825
97	Last Notable Activity_Email Opened	1004.486994
62	What is your current occupation_Unemployed	771.843206
105	Last Notable Activity_SMS Sent	739.487650
9	Lead Source_Google	577.079659
...
70	Tags_Interested in Next batch	1.023595
68	Tags_In confusion whether part time or DLP	1.020261
80	Tags_University not recognized	1.016872
72	Tags_Lateral student	1.015602
78	Tags_Shall take in the next coming month	1.010650

	feature	VIF
9	Lead Source_Google	26.913720
8	Lead Source_Facebook	25.565901
5	Lead Origin_Lead Import	25.496329
7	Lead Source_Direct Traffic	24.770865
4	Lead Origin_Lead Add Form	16.593449
...
22	Lead Source_blog	1.013013
11	Lead Source_NC_EDM	1.012070
76	Tags_Shall take in the next coming month	1.010515
101	Last Notable Activity_Resubscribed to emails	1.010277
30	Last Activity_Email Marked Spam	1.008679

	feature	VIF
5	Lead Origin_Lead Import	25.461292
8	Lead Source_Facebook	25.415604
4	Lead Origin_Lead Add Form	14.531554
15	Lead Source_Reference	12.525302
79	Tags_Will revert after reading the email	9.747848
...
13	Lead Source_Pay per Click Ads	1.004590
24	Lead Source_welearnblog_Home	1.004412
21	Lead Source_blog	1.003510
18	Lead Source_WeLearn	1.003432
10	Lead Source_NC_EDM	1.002566

	feature	VIF
4	Lead Origin_Lead Add Form	14.531310
14	Lead Source_Reference	12.525302
78	Tags_Will revert after reading the email	9.746840
34	Last Activity_SMS Sent	7.847761
3	Lead Origin_Landing Page Submission	6.566749
...
12	Lead Source_Pay per Click Ads	1.004590
23	Lead Source_welearnblog_Home	1.004412
20	Lead Source_blog	1.003510
17	Lead Source_WeLearn	1.003432
9	Lead Source_NC_EDM	1.002566

Now it is looking good from the VIF perspective so this is the final for the modelling.

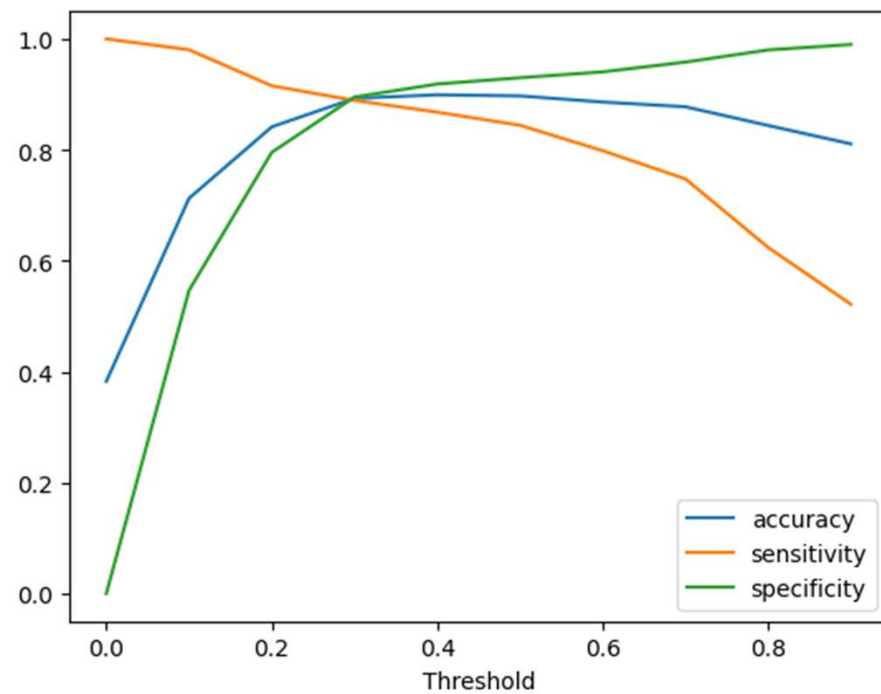
Modelling

Dep. Variable:	Converted	No. Observations:	5623							
Model:	GLM	Df Residuals:	5522		coef	std err	z	P> z	[0.025	0.975]
Model Family:	Binomial	Df Model:	100	const	-1.009e+15	7.47e+06	-1.35e+08	0.000	-1.01e+15	-1.01e+15
Link Function:	Logit	Scale:	1.0000	TotalVisits	2.069e+13	1.17e+06	1.77e+07	0.000	2.07e+13	2.07e+13
Method:	IRLS	Log-Likelihood:	nan	Total Time Spent on Website	5.996e+14	9.7e+05	6.18e+08	0.000	6e+14	6e+14
Date:	Fri, 15 Nov 2024	Deviance:	98643.	Page Views Per Visit	-6.527e+12	1.21e+06	-5.41e+06	0.000	-6.53e+12	-6.53e+12
Time:	22:48:06	Pearson chi2:	4.82e+18	Lead Origin_Landing Page Submission	-1.884e+14	2.93e+06	-6.42e+07	0.000	-1.88e+14	-1.88e+14
No. Iterations:	100	Pseudo R-squ. (CS):	nan	Lead Origin_Lead Add Form	1.048e+15	1.61e+07	6.51e+07	0.000	1.05e+15	1.05e+15
Covariance Type:	nonrobust			Lead Origin_Quick Add Form	1.012e+14	6.8e+07	1.49e+06	0.000	1.01e+14	1.01e+14
				Lead Source_Direct Traffic	1.592e+13	2.73e+06	5.84e+06	0.000	1.59e+13	1.59e+13
				Lead Source_Facebook	4.17e+14	1.16e+07	3.59e+07	0.000	4.17e+14	4.17e+14
				Lead Source_Live Chat	2.779e+15	6.9e+07	4.03e+07	0.000	2.78e+15	2.78e+15
				Lead Source_NC_EDM	-7.9148	6.31e-07	-1.25e+07	0.000	-7.915	-7.915
				Lead Source_Olark Chat	4.143e+14	4.83e+06	8.57e+07	0.000	4.14e+14	4.14e+14
				Lead Source_Organic Search	1.17e+14	3.02e+06	3.88e+07	0.000	1.17e+14	1.17e+14
				Lead Source_Pay per Click Ads	31.6273	4.62e-07	6.85e+07	0.000	31.627	31.627
				Lead Source_Press_Release	-1.265e+15	4.79e+07	-2.64e+07	0.000	-1.27e+15	-1.27e+15
				Lead Source_Reference	-1.114e+13	1.67e+07	-6.65e+05	0.000	-1.11e+13	-1.11e+13
				Lead Source_Referral Sites	-8.462e+13	7.24e+06	-1.17e+07	0.000	-8.46e+13	-8.46e+13
				Lead Source_Social Media	-2.456e+15	6.77e+07	-3.63e+07	0.000	-2.46e+15	-2.46e+15

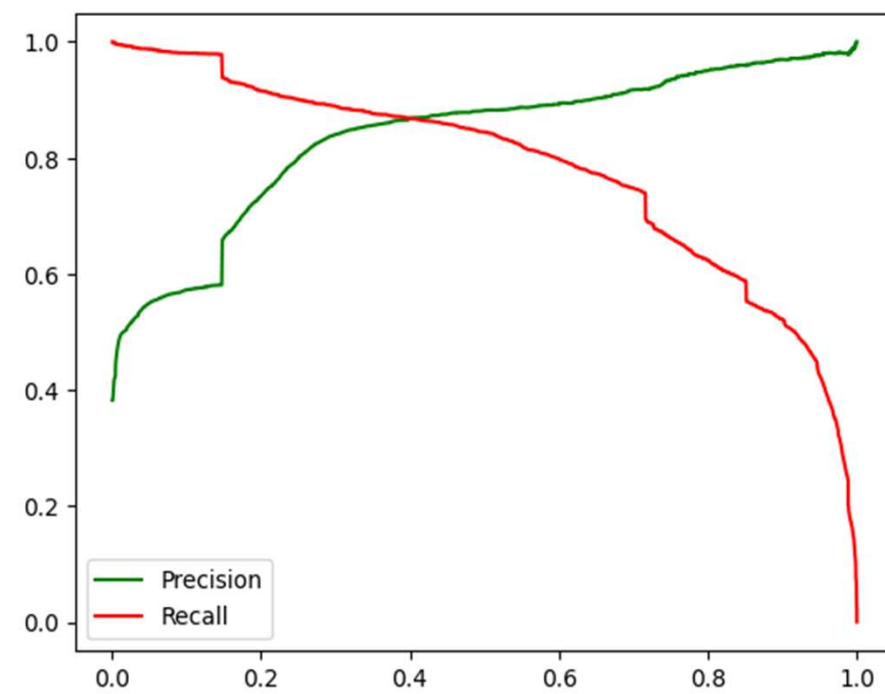
Here no column has error more then 0 so everything is looking good for modelling.

- We have split our dataset into train and test by 75% to train and 25% to test set by train_test_split method
- We have scaled our data using StandardScaler() Method.

Relation between the metrics



Precision and Recall



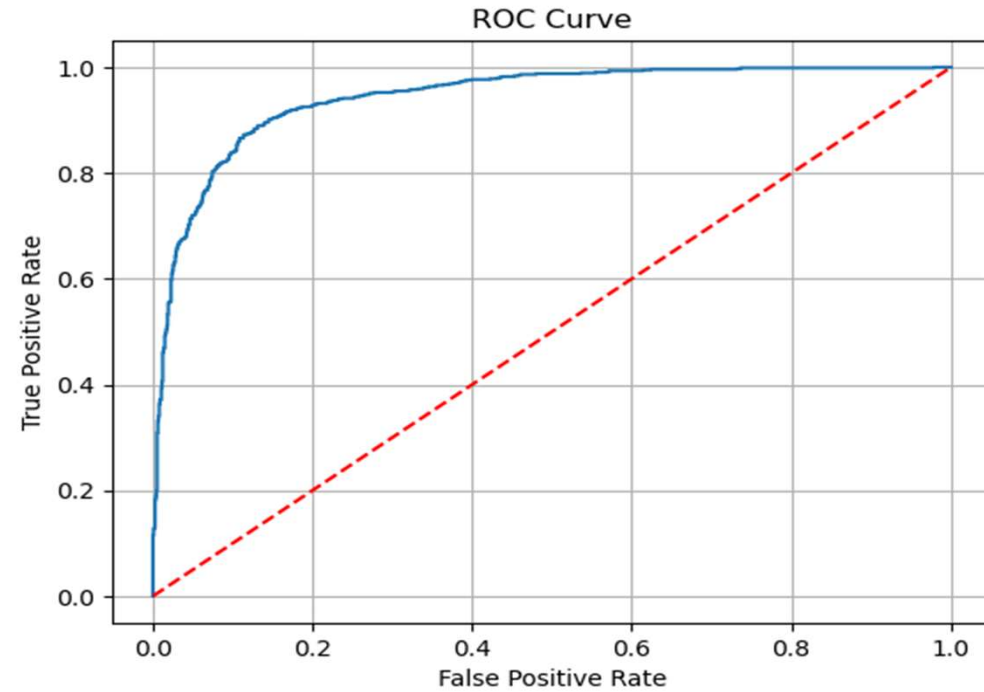
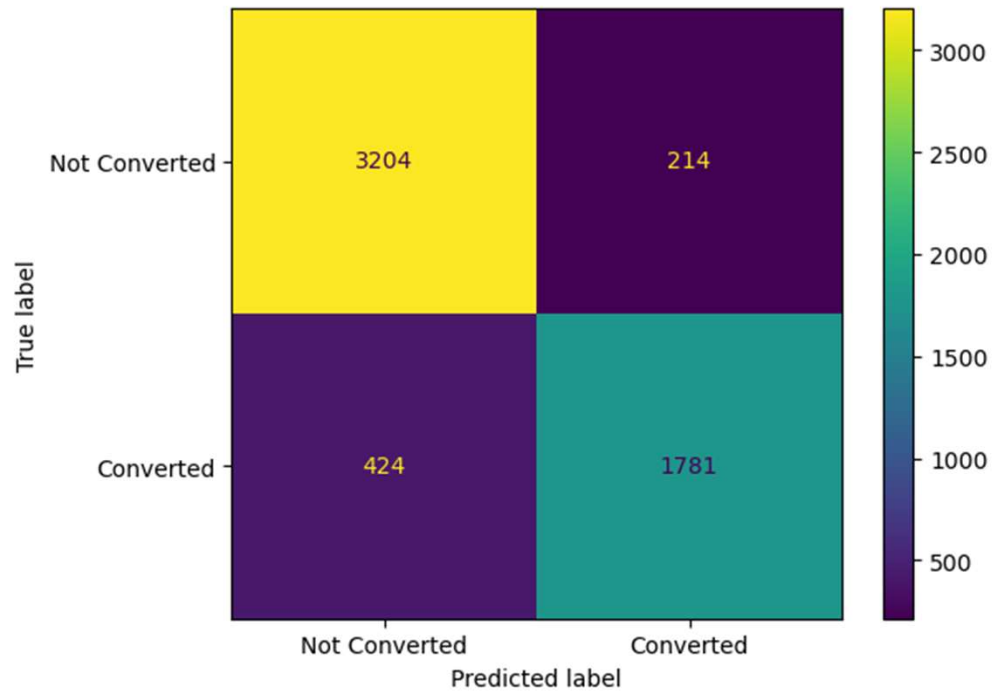
Train Set:-

Threshold	Accuracy	Precision	Recall
0.3	88.32	80.45	92.76
0.4	89.74	84.88	89.84
0.5	89.51	87.26	85.76
0.6	88.65	89.27	80.77

Test Set:-

Threshold	Accuracy	Precision	Recall
0.3	86.88	79.00	90.44
0.4	88.00	83.58	86.20
0.5	87.68	85.94	81.83
0.6	86.83	88.43	76.23

From the plots and the evaluation metrics tables we find 0.4 Threshold should be good for our model.
So, the final model is like.



Train set:-

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.88	0.94	0.91	3418
1	0.89	0.81	0.85	2205

accuracy			0.89	5623
----------	--	--	------	------

macro avg	0.89	0.87	0.88	5623
-----------	------	------	------	------

weighted avg	0.89	0.89	0.89	5623
--------------	------	------	------	------

Test set:-

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.91	0.89	0.90	1143
1	0.84	0.86	0.85	732

accuracy			0.88	1875
----------	--	--	------	------

macro avg	0.87	0.88	0.87	1875
-----------	------	------	------	------

weighted avg	0.88	0.88	0.88	1875
--------------	------	------	------	------

Thank You

