

Lead Scoring Case Study


by - Yash Kumar Roy



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

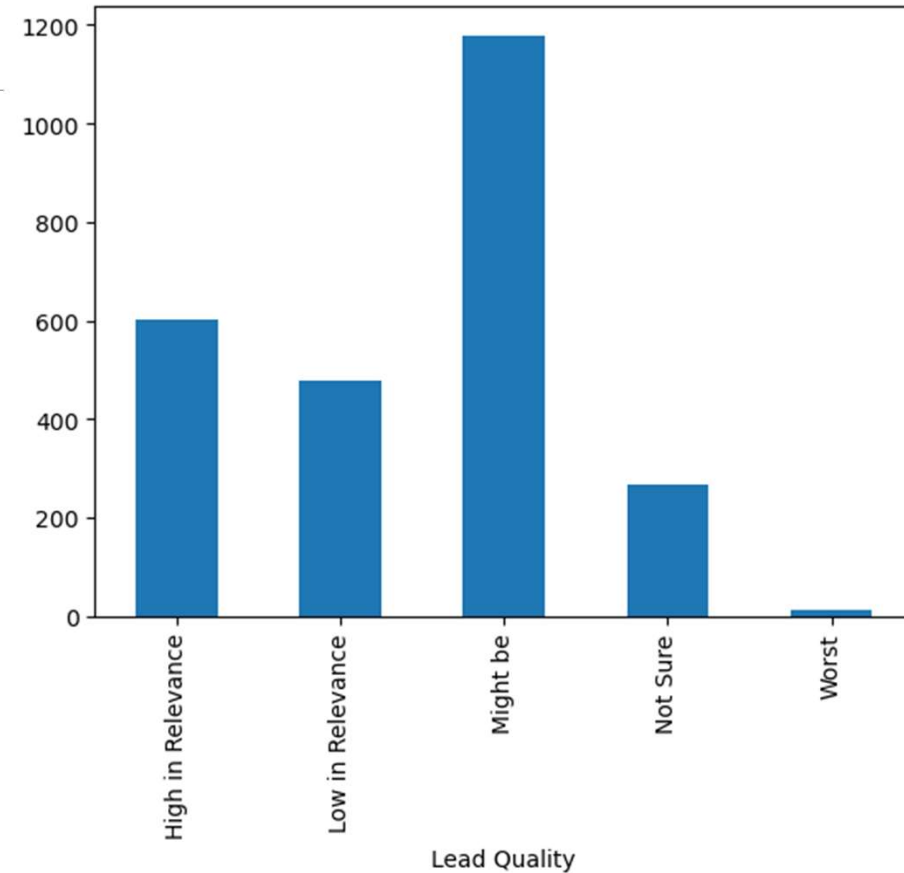
← Is showing the number of null Values in the columns

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

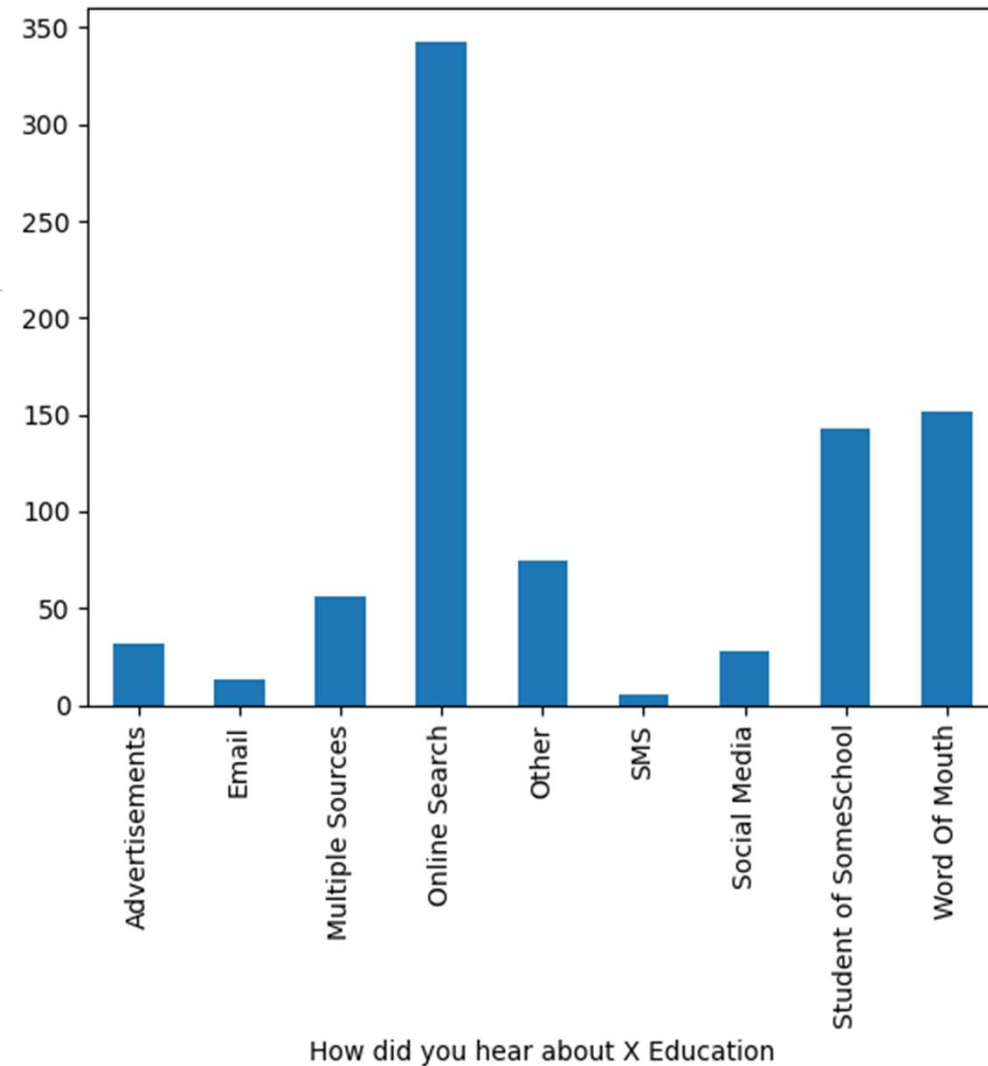
Distribution of Lead Quality and number of Converted

We can clearly see here that number of customer who filled the “Might be” in the Lead Quality section are most number of converted customer.

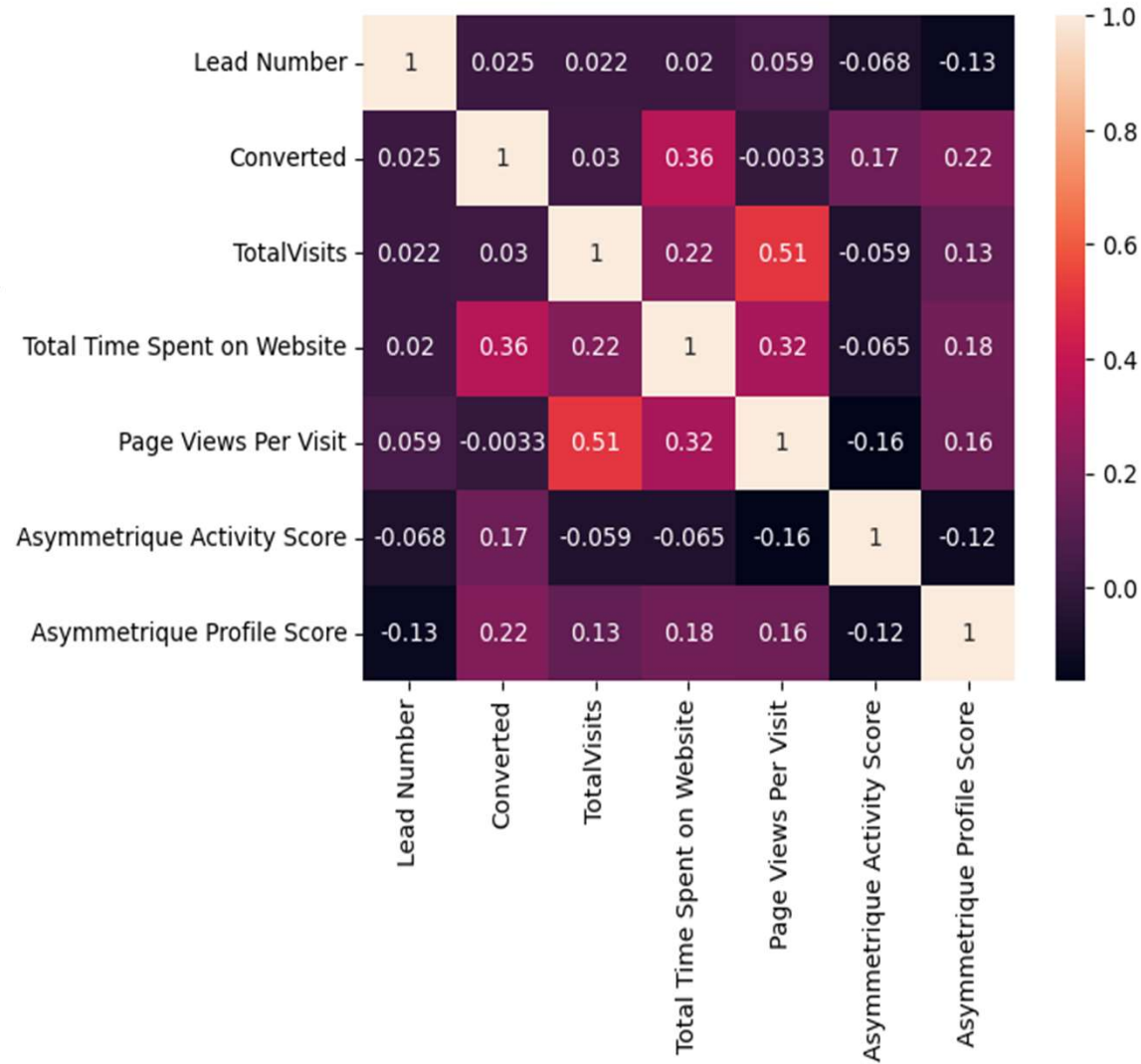


How did you hear about X Education vs Converted

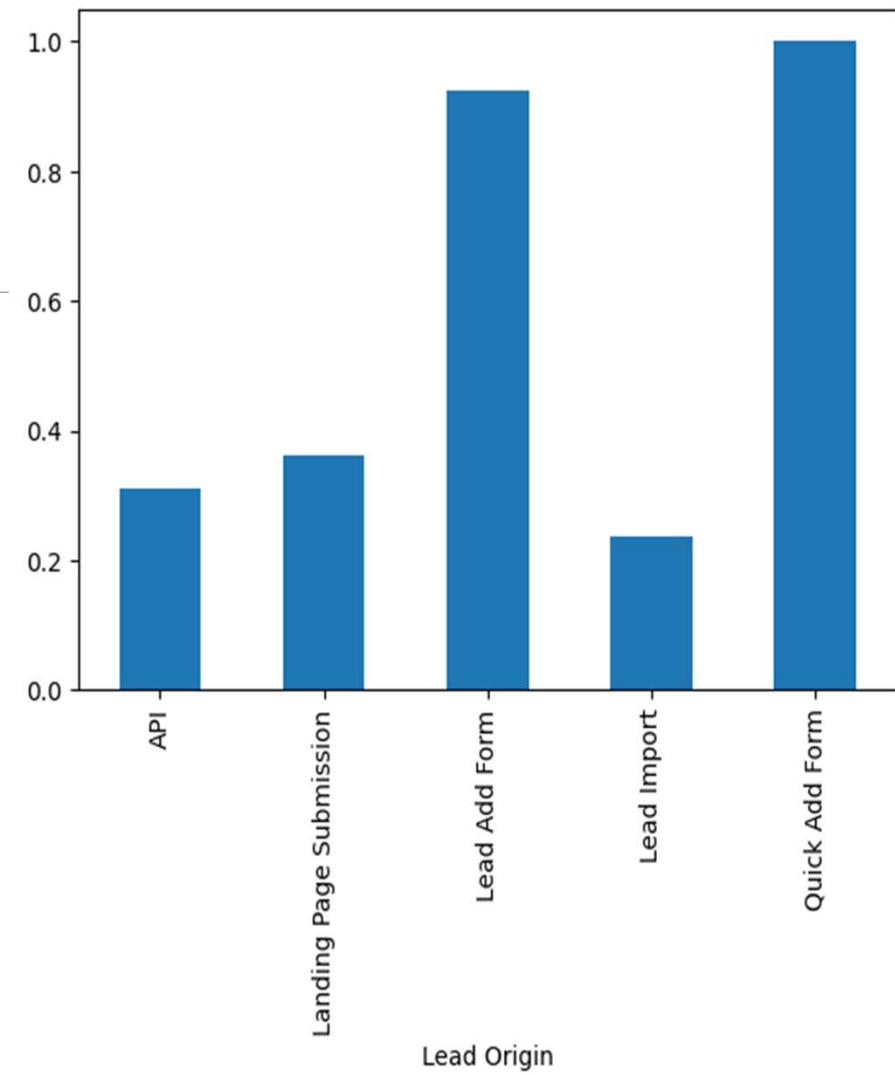
This graph is showing that customer who heard about the X Education company by the Online Search are very high of converted customer.



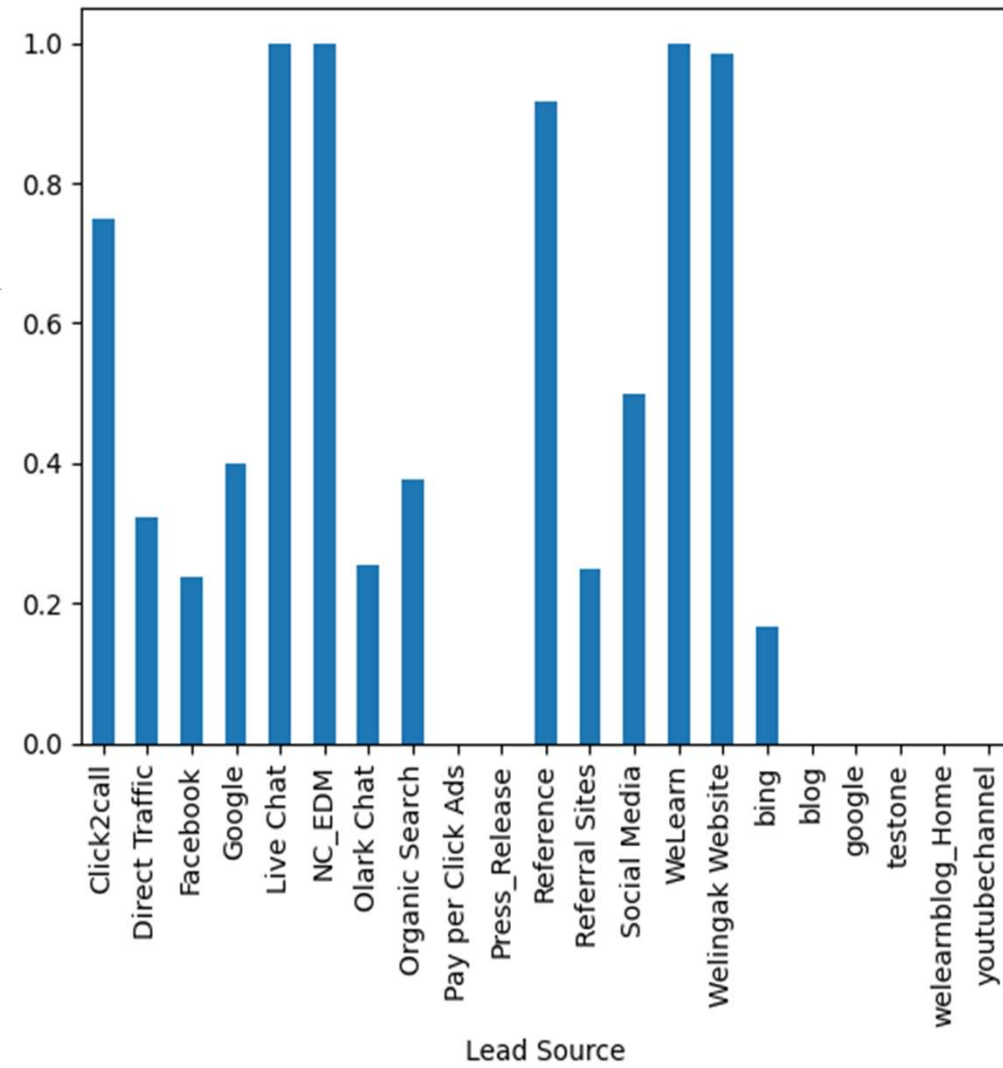
Heatmap showing the correlations between all numeric columns



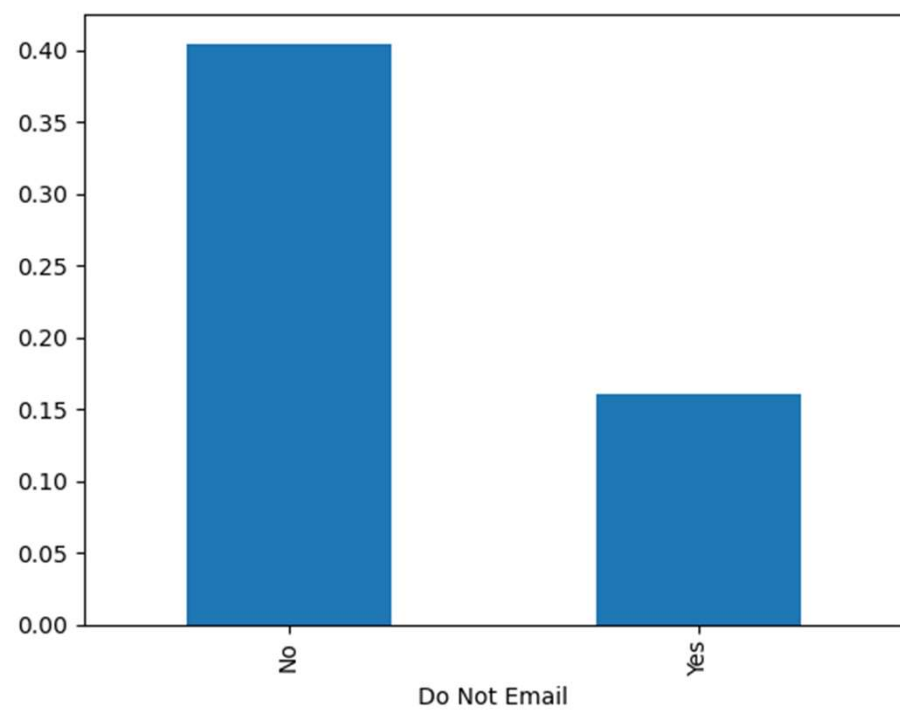
This graph shows the distribution of Lead Origin vs Converted



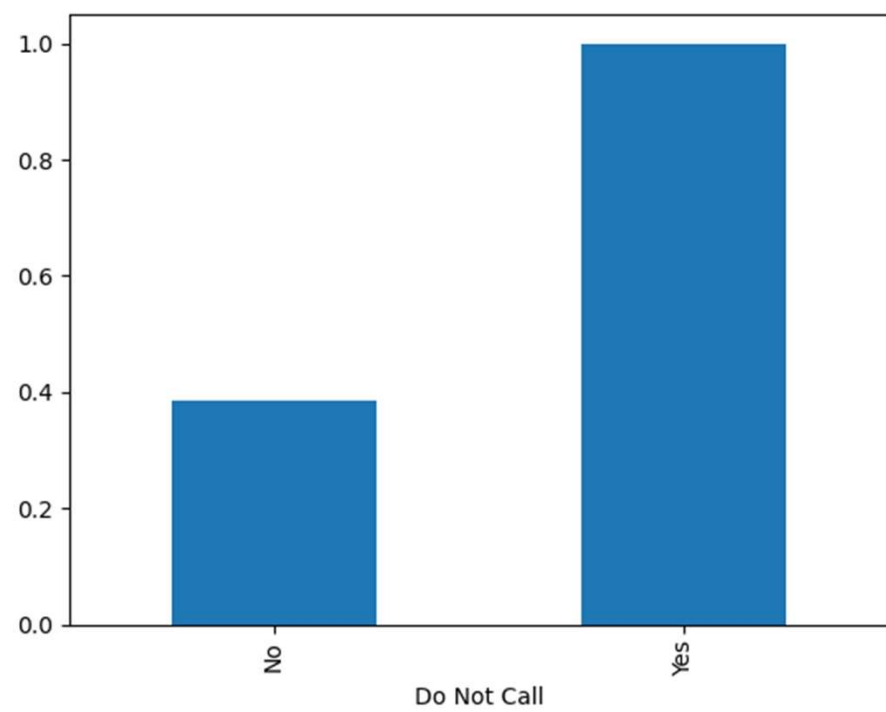
Lead Source vs Converted




Do Not Email vs Converted

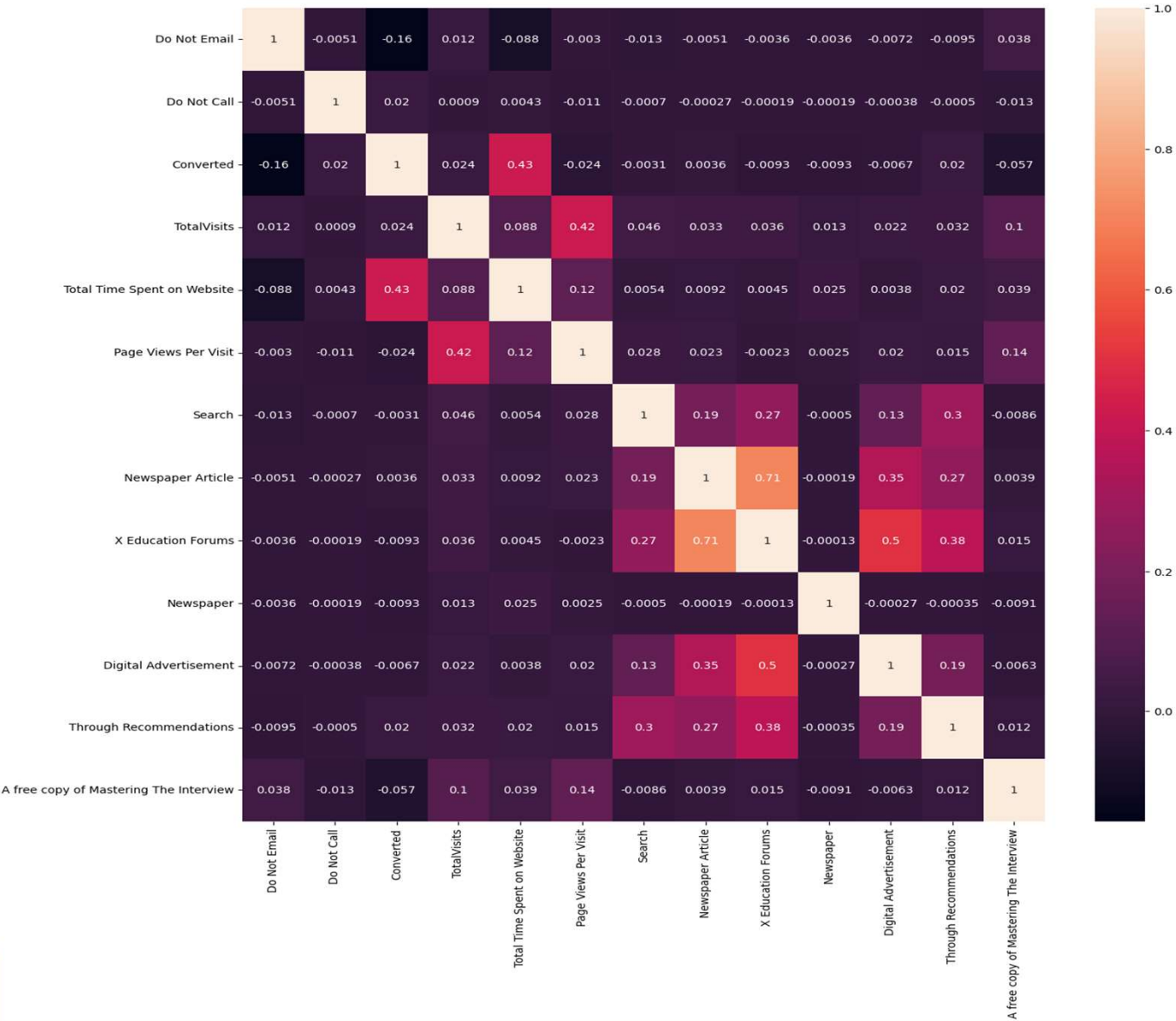


Do Not Call vs Converted

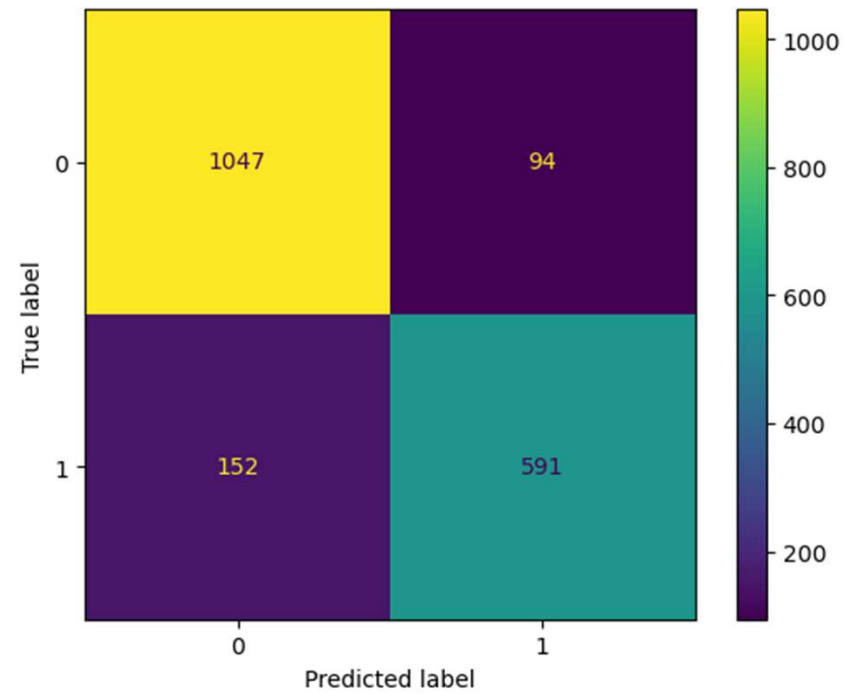


-
- Dropped those columns which have more than 45% of null values.
 - Converted all categorical columns to dummy variable .
 - Filling null values with mean if it was a numerical column
 - Filling null values with mode if it was a categorical column.
- 

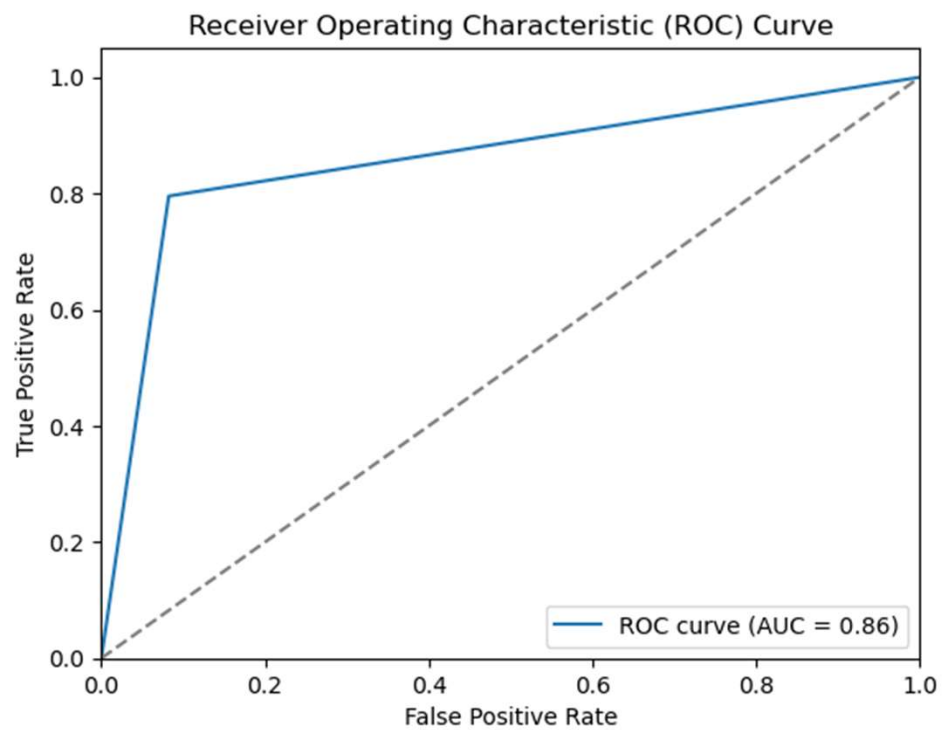
Heatmap with some more variables



This is the confusion matrix of the model of Logistic Regression Model



Roc Curve



- The accuracy of the model is 86.94%

Thank You

A solid orange horizontal bar spanning the width of the slide, located at the bottom.