Lead Scoring Case study

Submitted by:
Nidhi Misra
Vibhuti Dalal

Problem Statement:

- X Education sells online courses to industry professionals. They market it's courses in multiple websites and based on number of people filling the form, leads gets created . Their current lead conversion rate is 30%,need to build a model which will help in increasing the lead conversion rate to 80% . X education wants a model which will assign a lead score to leads and higher the lead score, more chances of lead conversion.

- Customers who have higher probability of buying the course are known as 'hot leads'

Assumptions

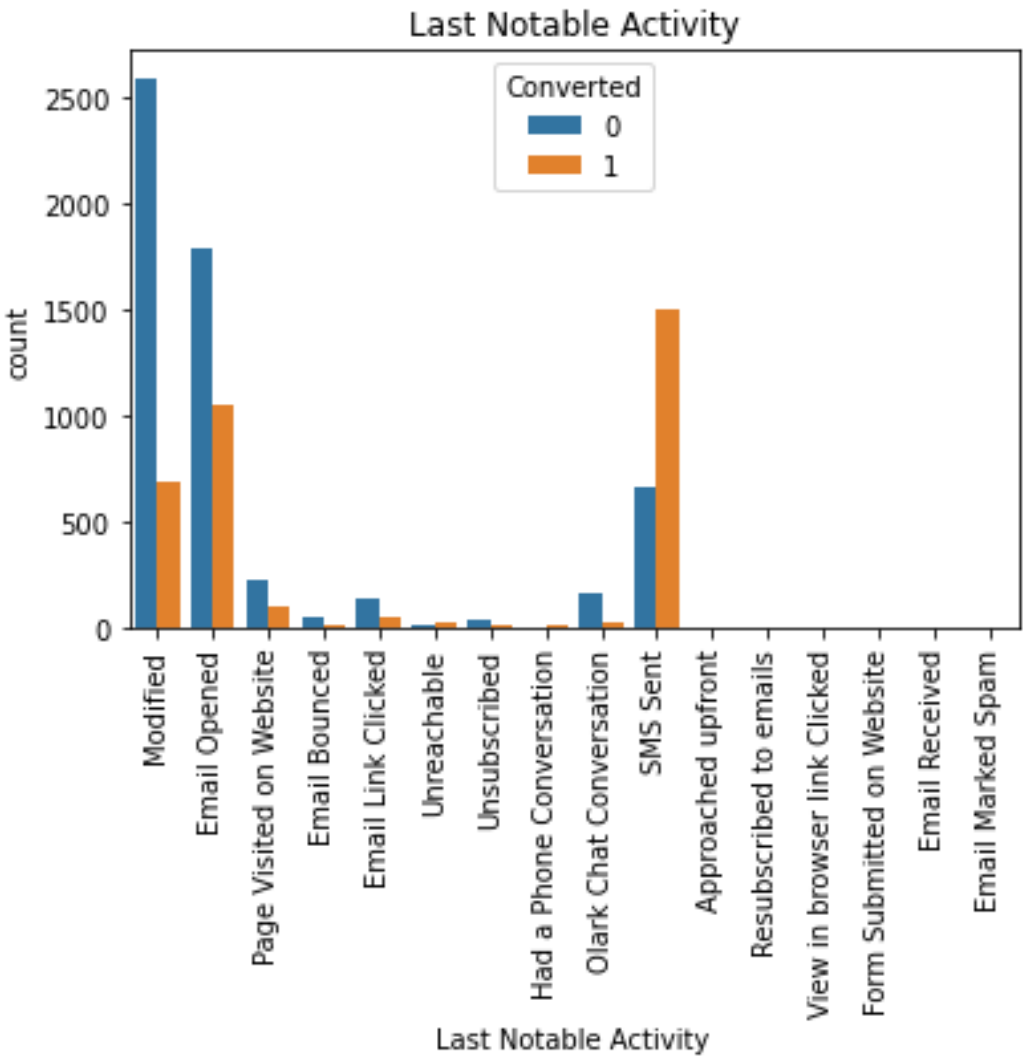
- X Education has a period of 2 months every year during which they hire some interns. The sales team has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible.
- the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls.

Approach

- Data Cleaning and preparation
- Select hypothesis
- Validating hypothesis by building regression model
- Train\Test technique
- Share solution based on recall, sensitivity and other matrix

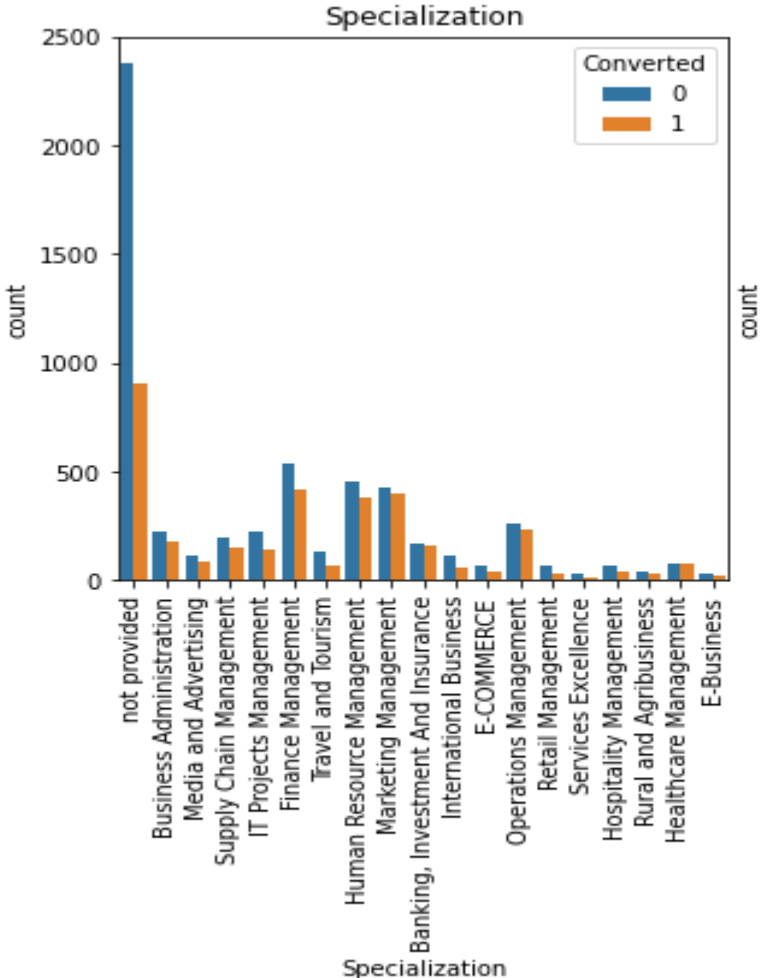
Data preparation

- Found the columns with unique values and dropped them.
- Calculated null values and columns having more than 40% of null values were dropped.
- Categorical data analysis



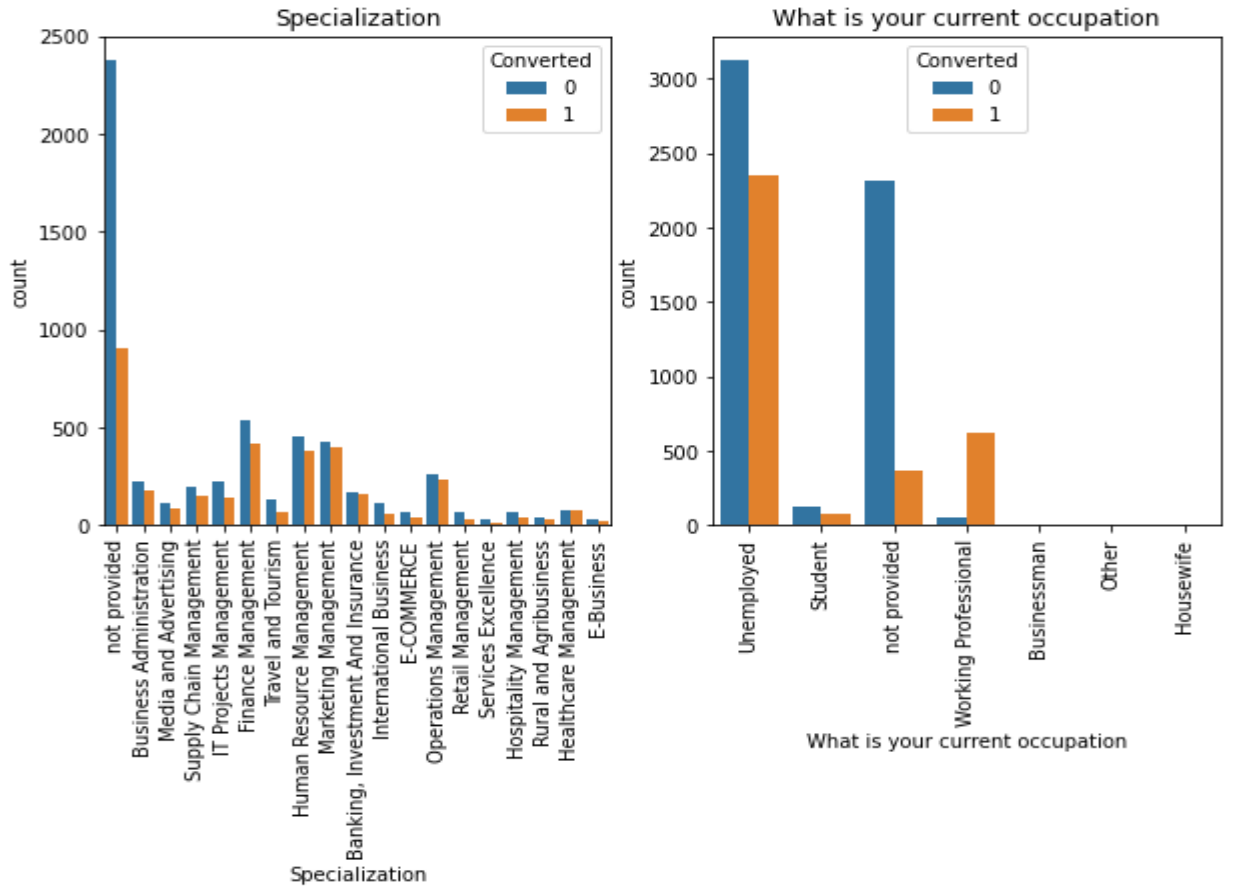
Categorical Data Analysis

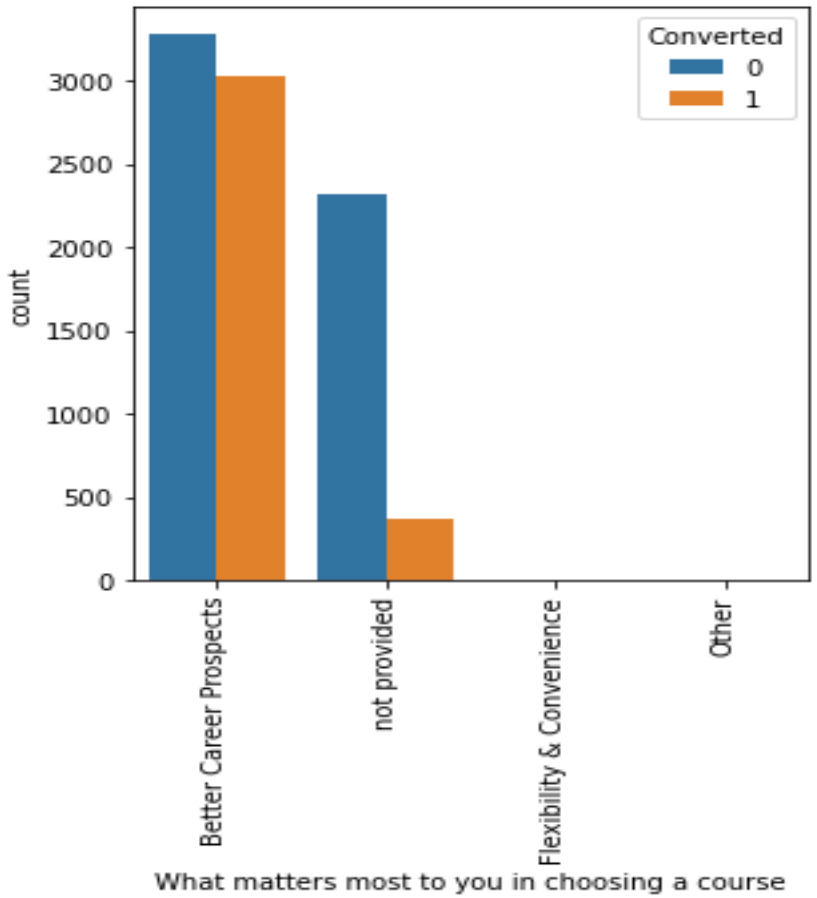
- Many leads have not selected the specialization

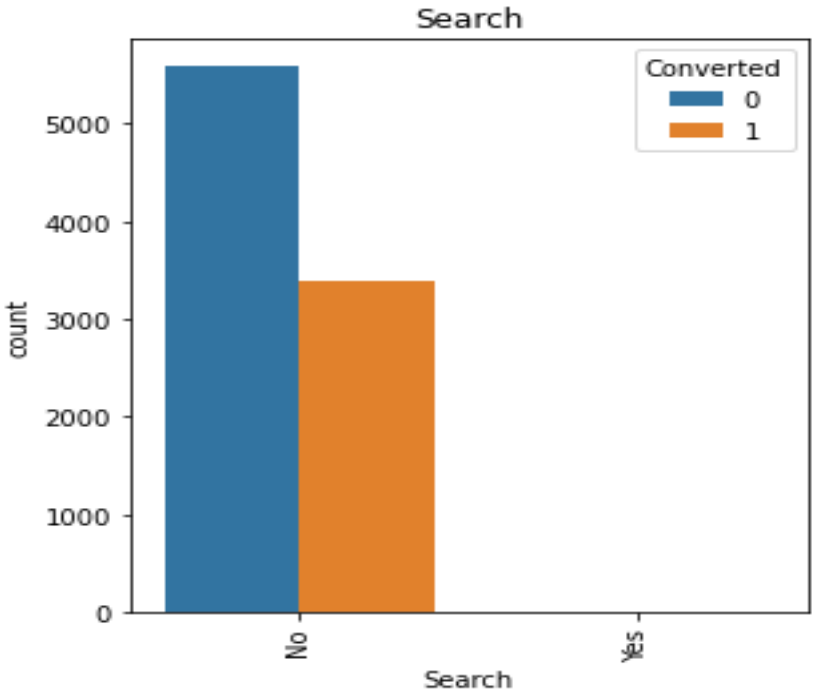


Categorical Data Analysis

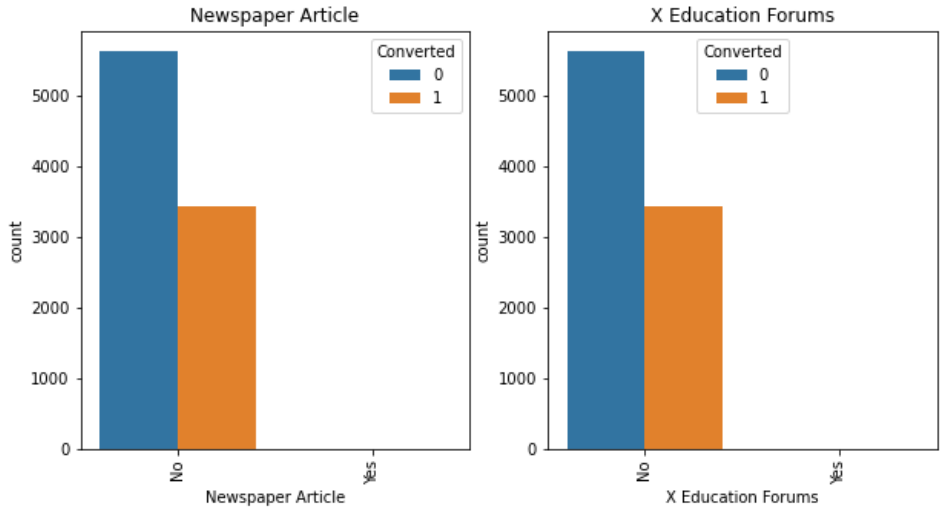
- Occupation column contains most of the percentage with Unemployed.

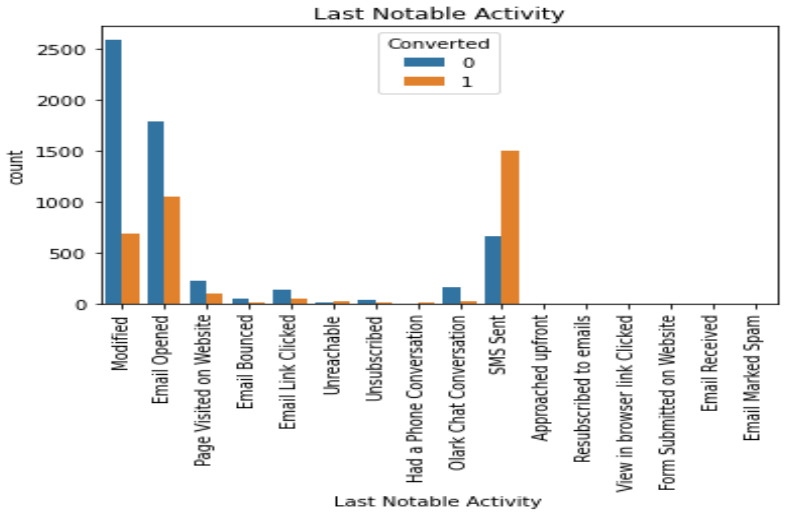



Converted
0
1
3000
2500
2000
1500
1000
500
0
count
Better Career Prospects
not provided
Flexibility & Convenience
Other
What matters most to you in choosing a course

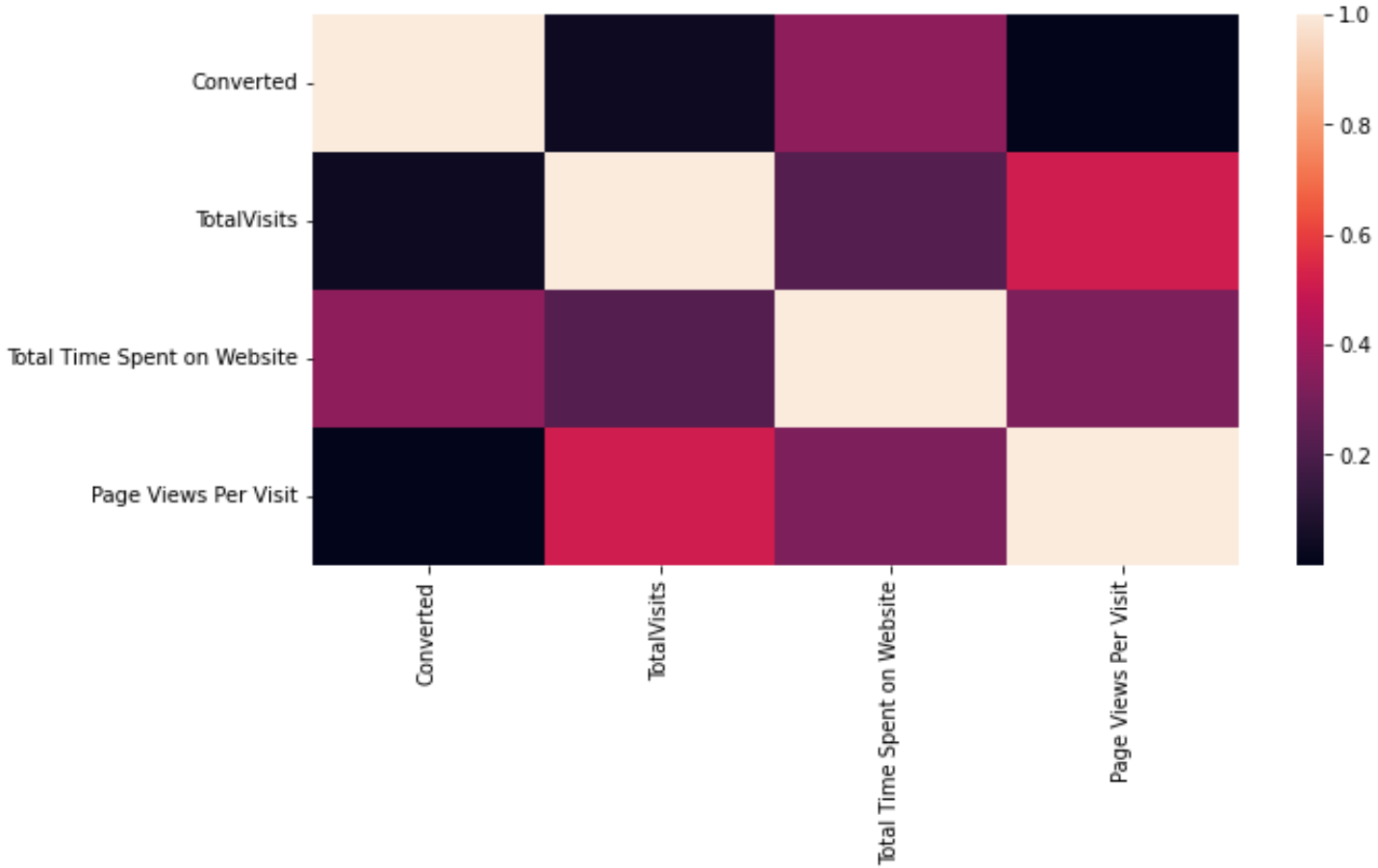

Search
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Search

As we have greater number of No and very few yes, we dropped the column

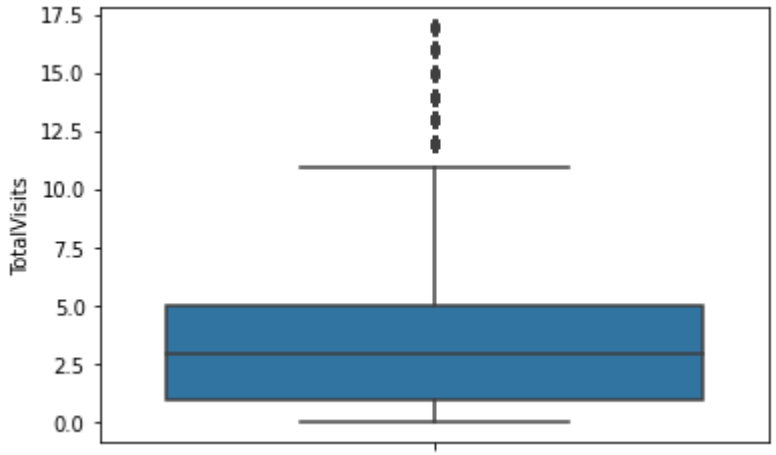


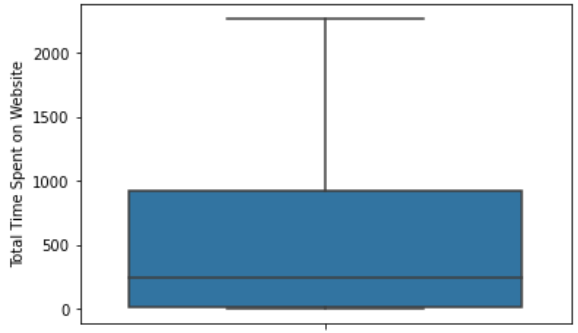


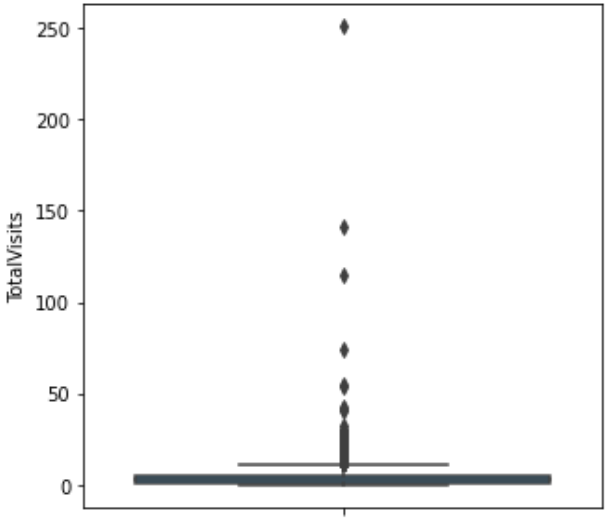
Numerical data

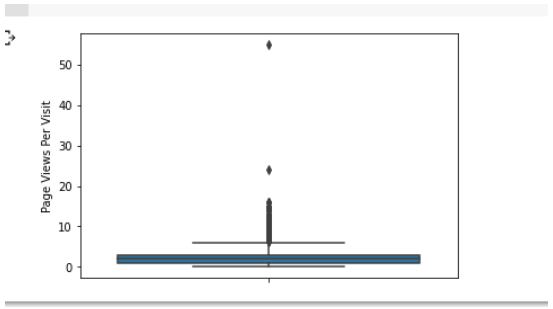


Outliers


TotalVisits
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


Total Time Spent on Website
2000
1500
1000
500
0


TotalVisits
250
200
150
100
50
0


Page Views Per Visit
50
40
30
20
10
0

Model Building

Dep. Variable:	Converted	**No. Observations:**	6293
Model:	GLM	**Df Residuals:**	6272
Model Family:	Binomial	**Df Model:**	20
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2573.2
Date:	Tue, 18 Oct 2022	**Deviance:**	5146.4
Time:	23:02:50	**Pearson chi2:**	6.52e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.3920	0.101	3.876	0.000	0.194	0.590
TotalVisits	1.9299	0.301	6.405	0.000	1.339	2.520
Total Time Spent on Website	4.7035	0.170	27.617	0.000	4.370	5.037
Page Views Per Visit	-2.0243	0.444	-4.558	0.000	-2.895	-1.154
Lead Origin_Lead Add Form	3.0451	0.256	11.896	0.000	2.543	3.547
Lead Source_Direct Traffic	-1.5377	0.132	-11.617	0.000	-1.797	-1.278
Lead Source_Google	-1.1120	0.129	-8.598	0.000	-1.366	-0.859
Lead Source_Organic Search	-1.4285	0.165	-8.657	0.000	-1.752	-1.105
Lead Source_Referral Sites	-1.3511	0.334	-4.049	0.000	-2.005	-0.697
Lead Source_Welingak Website	2.4662	1.039	2.373	0.018	0.429	4.503
Do Not Email_Yes	-1.4273	0.206	-6.916	0.000	-1.832	-1.023
Last Activity_Email Bounced	-1.1159	0.396	-2.820	0.005	-1.891	-0.340
Last Activity_Olark Chat Conversation	-1.2987	0.193	-6.723	0.000	-1.677	-0.920
What is your current occupation_Housewife	23.3558	2.89e+04	0.001	0.999	-5.66e+04	5.66e+04
What is your current occupation_Working Professional	2.7793	0.191	14.523	0.000	2.404	3.154
Last Notable Activity_Email Link Clicked	-2.0672	0.266	-7.760	0.000	-2.589	-1.545
Last Notable Activity_Email Opened	-1.4274	0.090	-15.864	0.000	-1.604	-1.251
Last Notable Activity_Had a Phone Conversation	22.3270	2.19e+04	0.001	0.999	-4.28e+04	4.29e+04
Last Notable Activity_Modified	-1.8466	0.099	-18.632	0.000	-2.041	-1.652
Last Notable Activity_Olark Chat Conversation	-1.6187	0.372	-4.347	0.000	-2.348	-0.889
Last Notable Activity_Page Visited on Website	-2.1305	0.216	-9.849	0.000	-2.554	-1.707

Conclusion

- On **Training Data**
- With the cutoff of 0.35 we get the Precision & Recall of 79.29% & 70.22% respectively.
- So to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of **0.44** which gave
- Accuracy 81.80%
- Precision 75.71%
- Recall 76.32%

- Prediction on **Test Data**
- Accuracy 80.57%
- Precision 74.87%
- Recall 73.26%

Recommendations

- if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be **0.35**

- If we go with Precision – Recall Evaluation the optimal cut off value would be **0.44**