LEAD GENERATION CASE STUDY

Prepared By:

Aswin Chandrasekharan
Vidya Udayabhanu

LEAD GENERATION CASE STUDY

Purpose of Case study:

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.

Dataset provided and its attributes:

1. *'leads.csv'* consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

2. *Target variable is 'Converted'* which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

3. *Initial: Total number of columns= 37 and total number of rows = 9240*

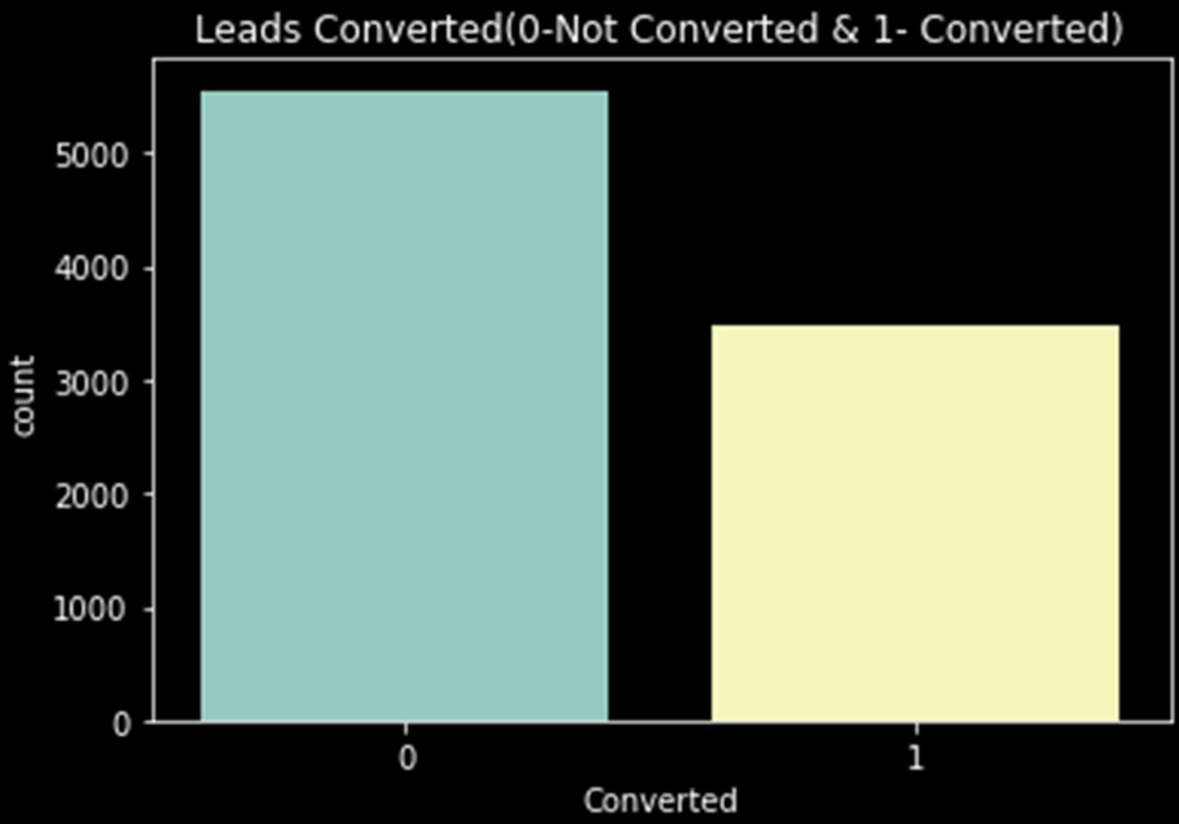
Target Variable

Inference:

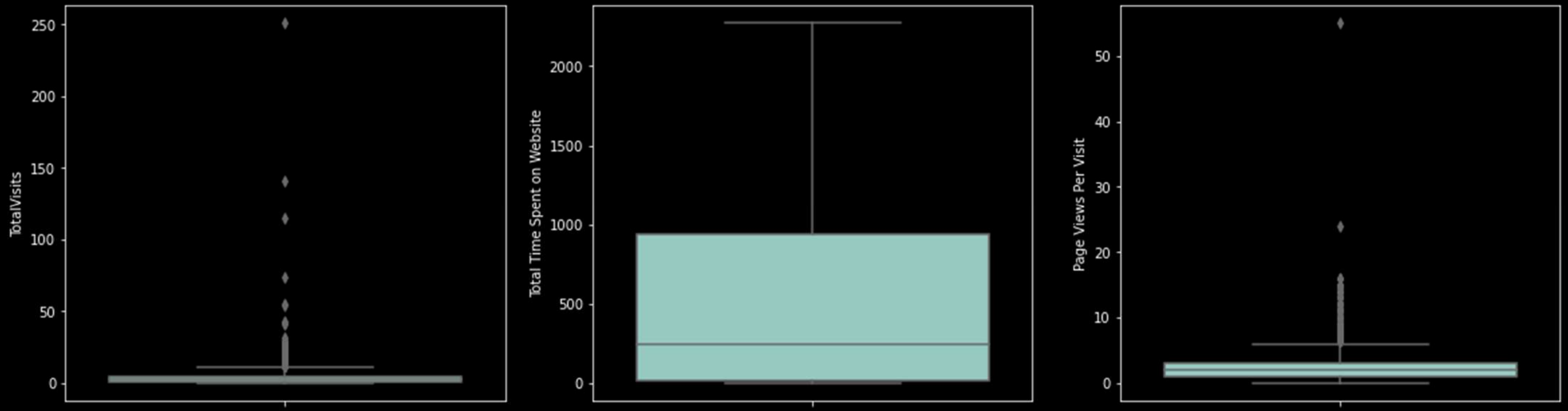
Leads converted : 1

Leads not converted : 0

Conversion rate : 38.51%

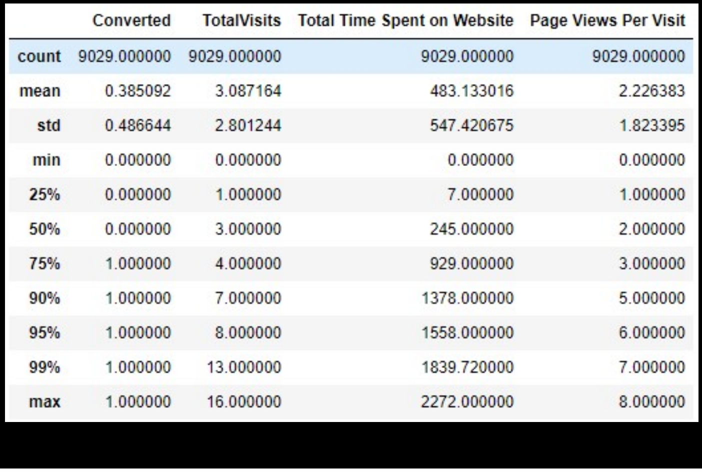


OUTLIER ANALYSIS



Step Taken:

Columns 'Total visits' and 'Page Views Per Visit' clearly have outliers. Therefore, we have removed these outliers that lie beyond the 99th percentile.
The table represents the values after outlier treatment.

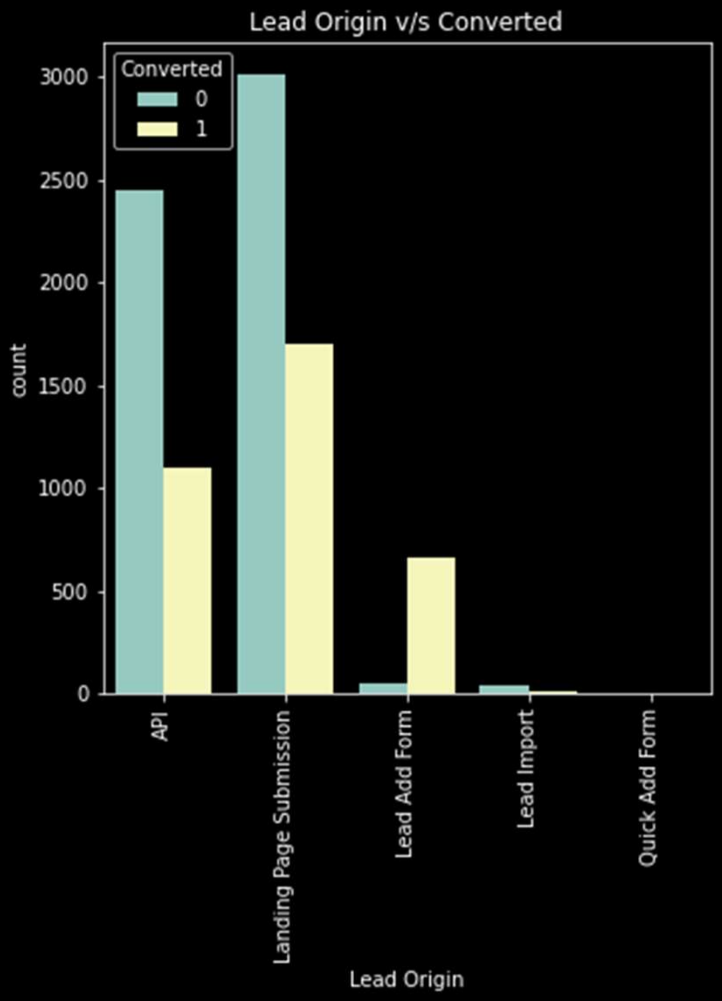


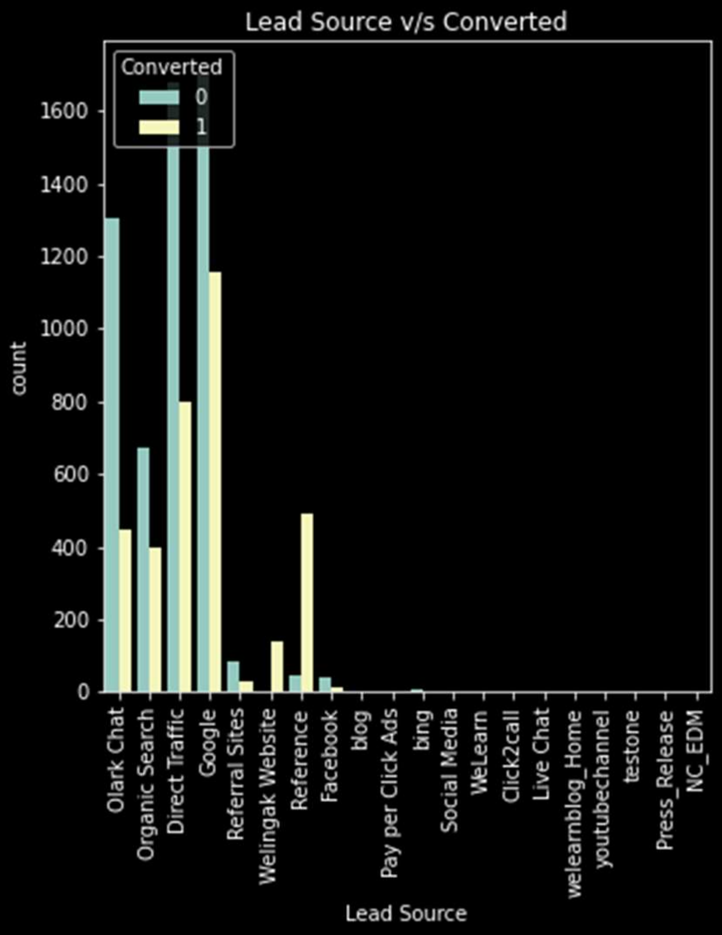
	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9029.000000	9029.000000	9029.000000	9029.000000
mean	0.385092	3.087164	483.133016	2.226383
std	0.486644	2.801244	547.420675	1.823395
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.000000	7.000000	1.000000
50%	0.000000	3.000000	245.000000	2.000000
75%	1.000000	4.000000	929.000000	3.000000
90%	1.000000	7.000000	1378.000000	5.000000
95%	1.000000	8.000000	1558.000000	6.000000
99%	1.000000	13.000000	1839.720000	7.000000
max	1.000000	16.000000	2272.000000	8.000000

BIVARIATE ANALYSIS

VARIABLES ANALYZED w.r.t TARGET VARIABLE (Conversion)

Lead origin
Lead source
Do Not Email
Do not call
Last Activity
Current occupation
Search
What matter most to you in choosing a course
Magazine
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement
Through Recommendation
A few copy of mastering the interview
Receive more updates about our courses
Update me on supply chain content
Get updates on DM content
Last notable Activity.

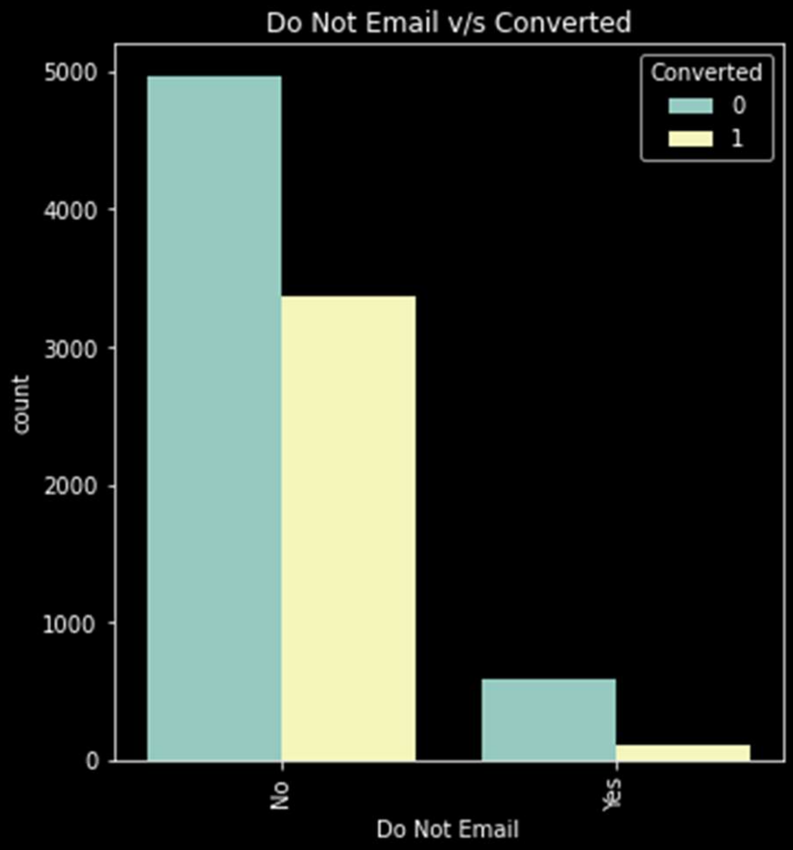


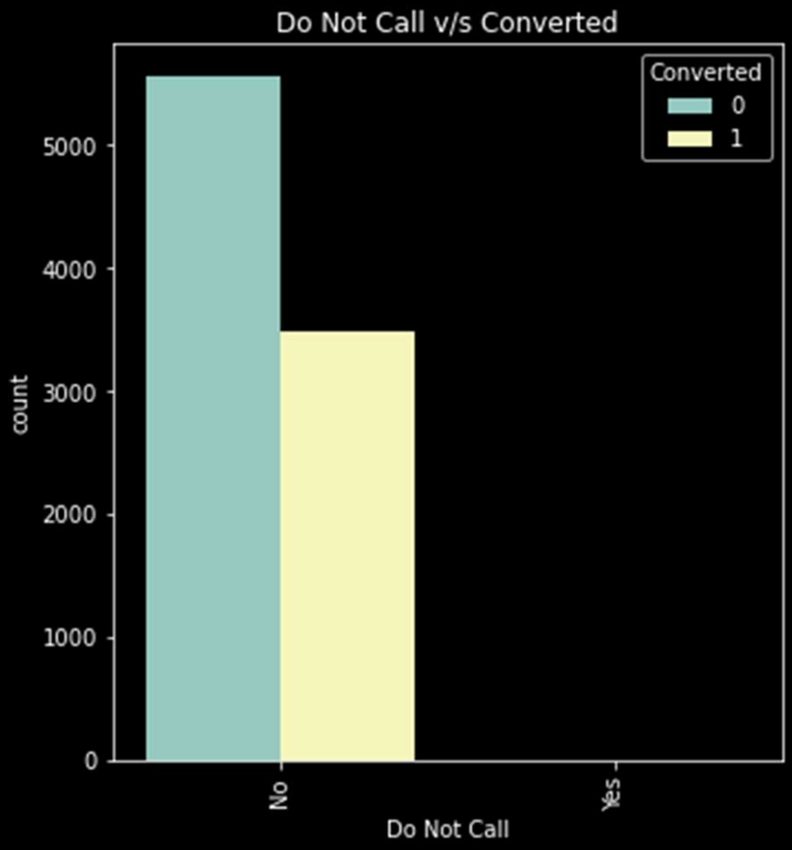


Inference :

From the above analysis, we can infer that in case of Lead Origin, most conversions have been from leads who have submitted information on landing page.

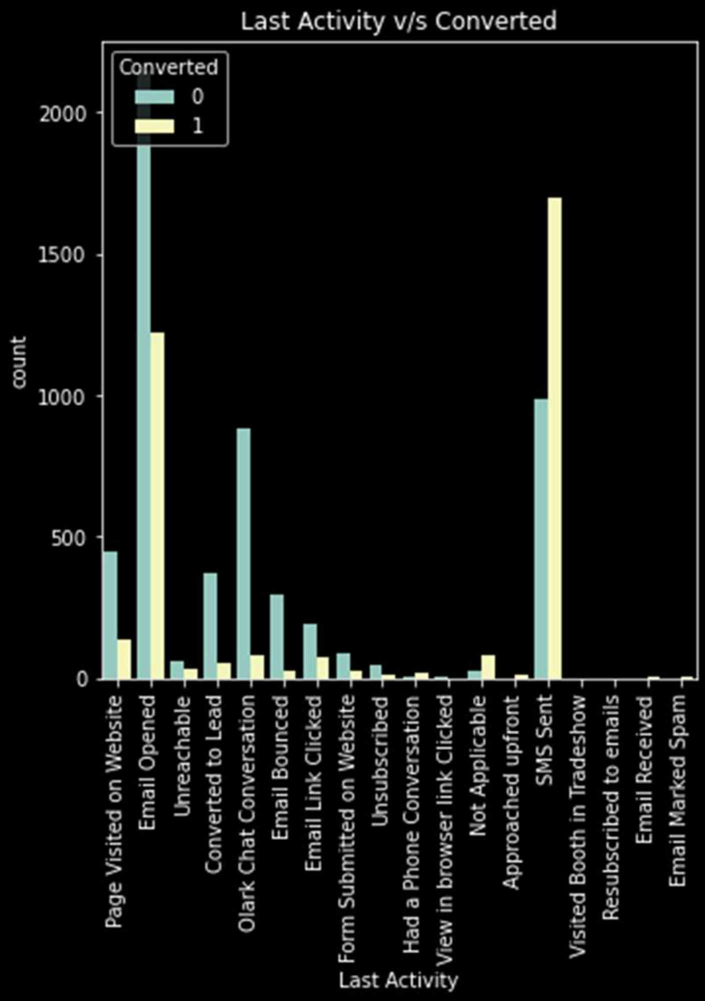
On the other hand, most conversions have been recorded through Google as the lead source.

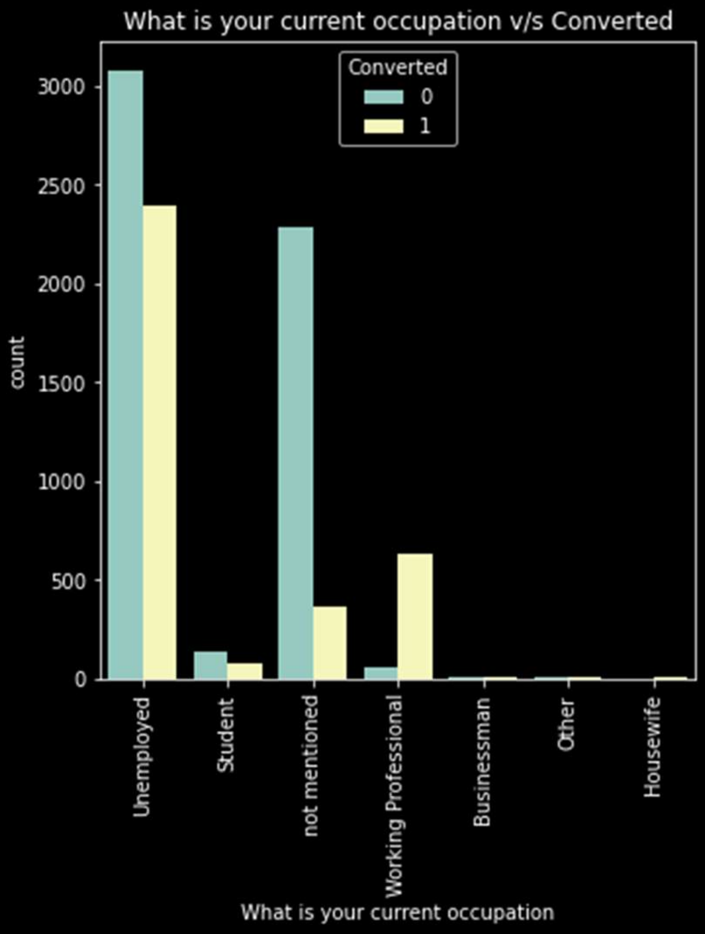




Inference :

From the above analysis, we can infer that most conversions and responses indicate that leads are not in favor of Email notifications and regular calls.

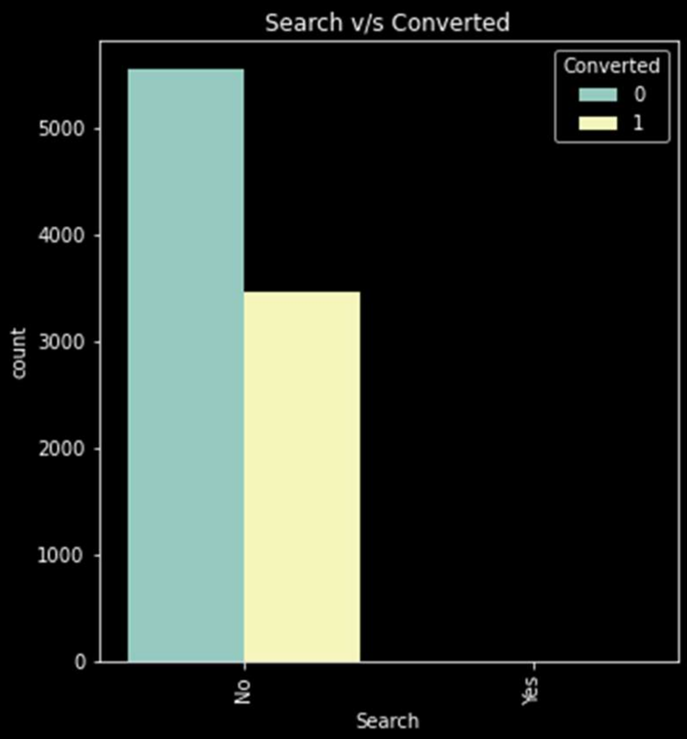


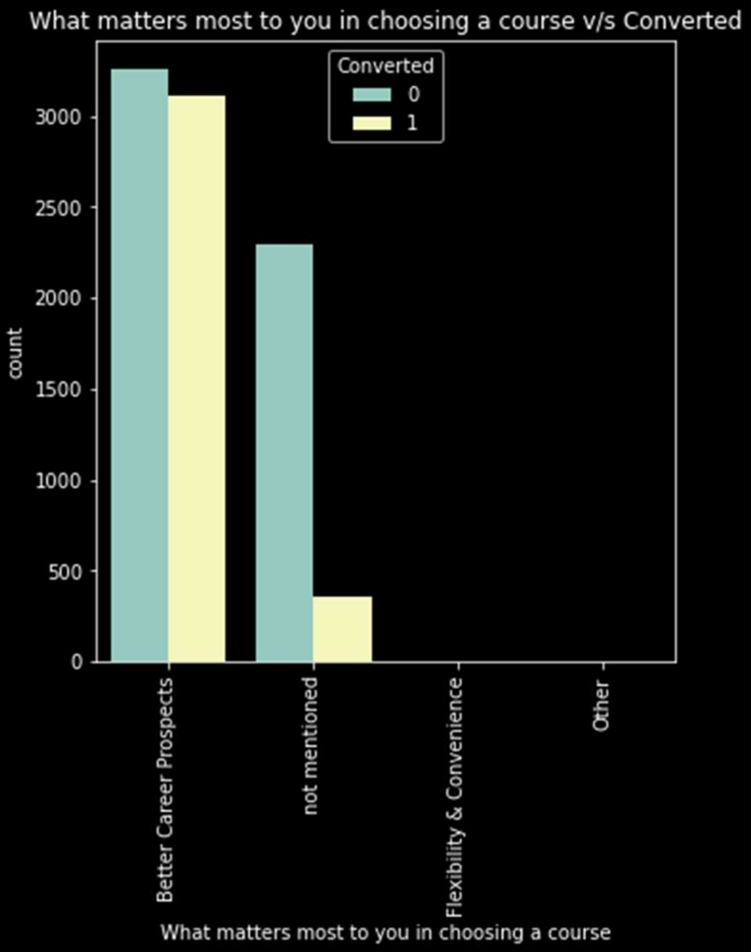


Inference :

From the above analysis, we can infer that among the leads successfully converted, most belong to those who have been sent an SMS.

Also, leads with higher conversion rate have been recorded to be either Unemployed or are working professionals.

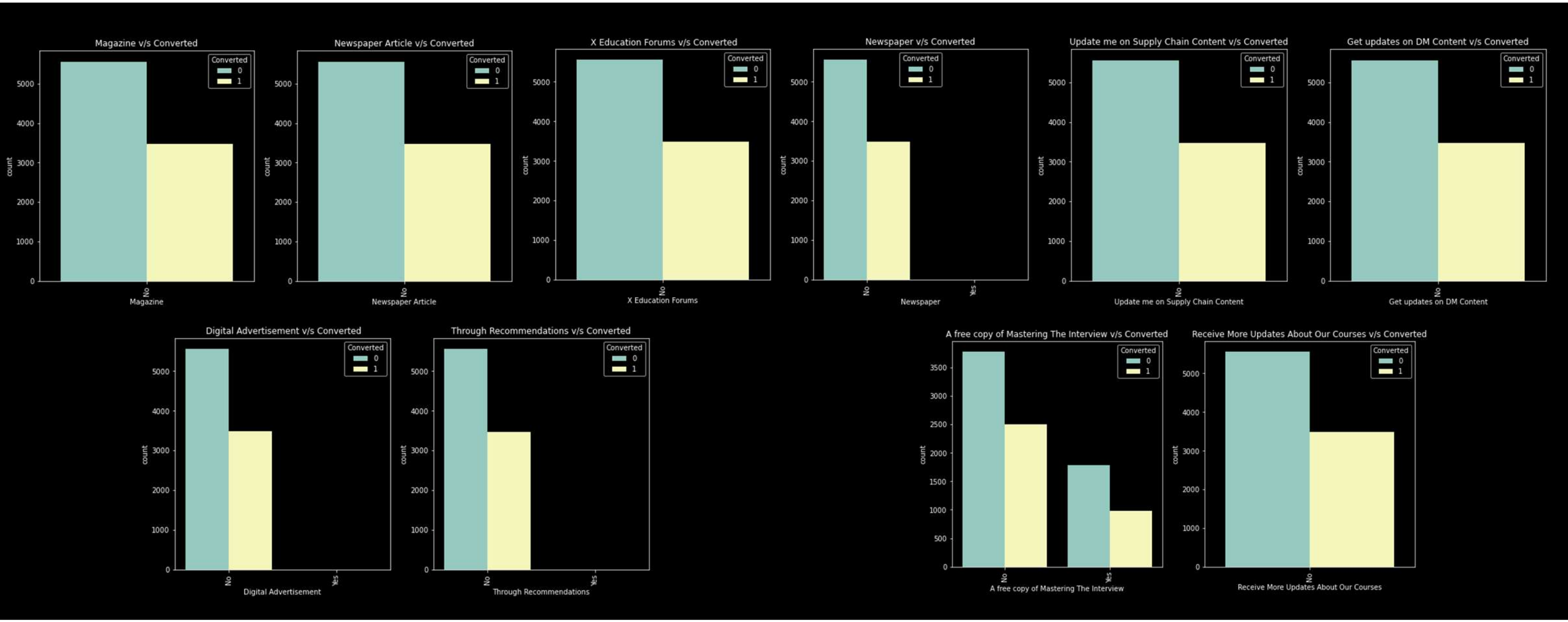




Inference :

This analysis shows that most leads who have been successfully converted look for courses to look for better career prospects.

These leads also tend to not search for the course.



Magazine v/s Converted

Newspaper Article v/s Converted

X Education Forums v/s Converted

Newspaper v/s Converted

Update me on Supply Chain Content v/s Converted

Get updates on DM Content v/s Converted

Digital Advertisement v/s Converted

Through Recommendations v/s Converted

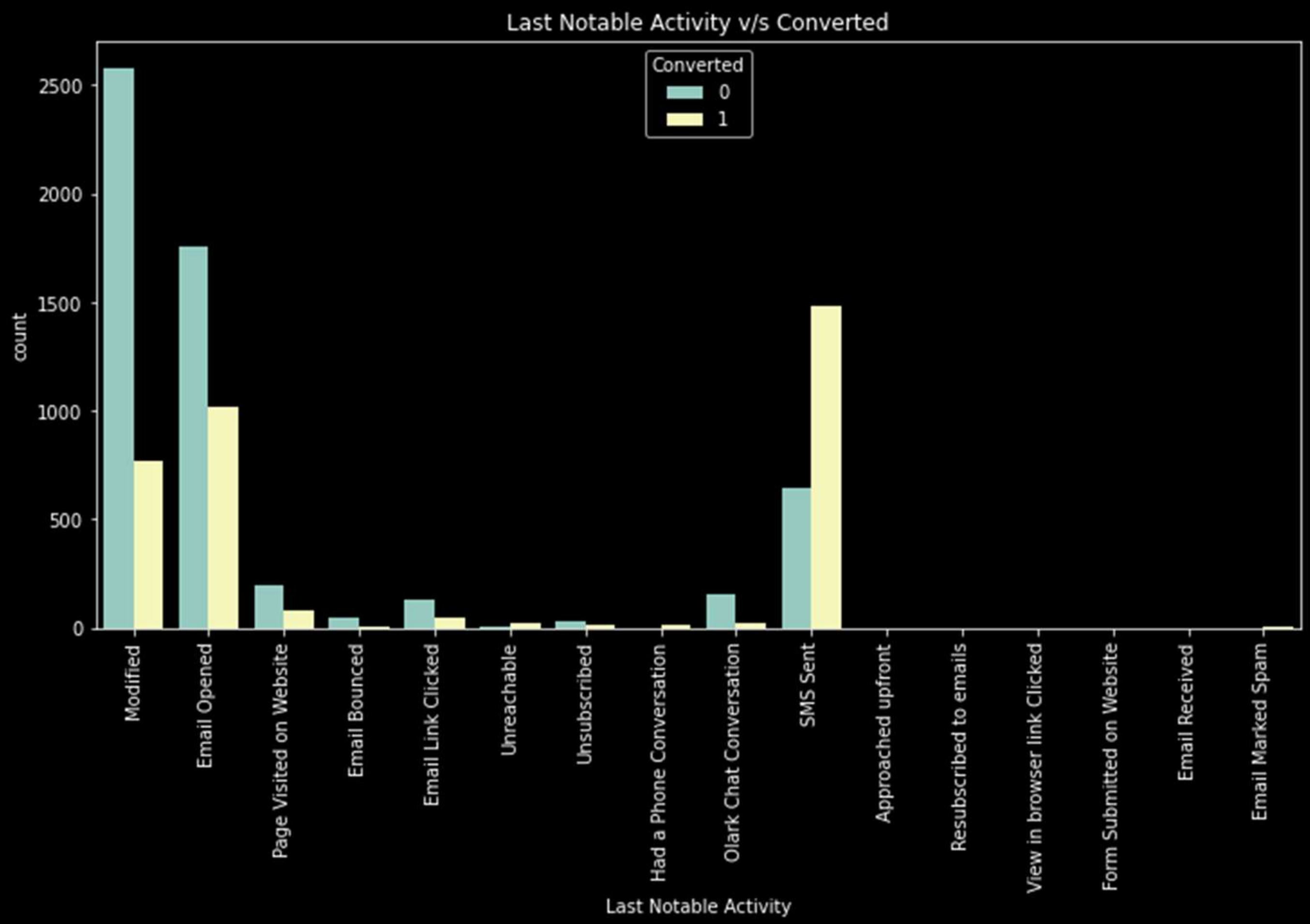
A free copy of Mastering The Interview v/s Converted

Receive More Updates About Our Courses v/s Converted

Inference :

In case of medium of communication with prospective leads, these graphs clearly show that magazines and Newspaper articles are not a great way. Similarly, X education ,Newspapers, any kind of updates, digital advertisements, and recommendations are also not a preferred way of communication or lead generation.

Prospective leads tend to prefer a free copy of mastering the interview and have a higher conversion rate too.

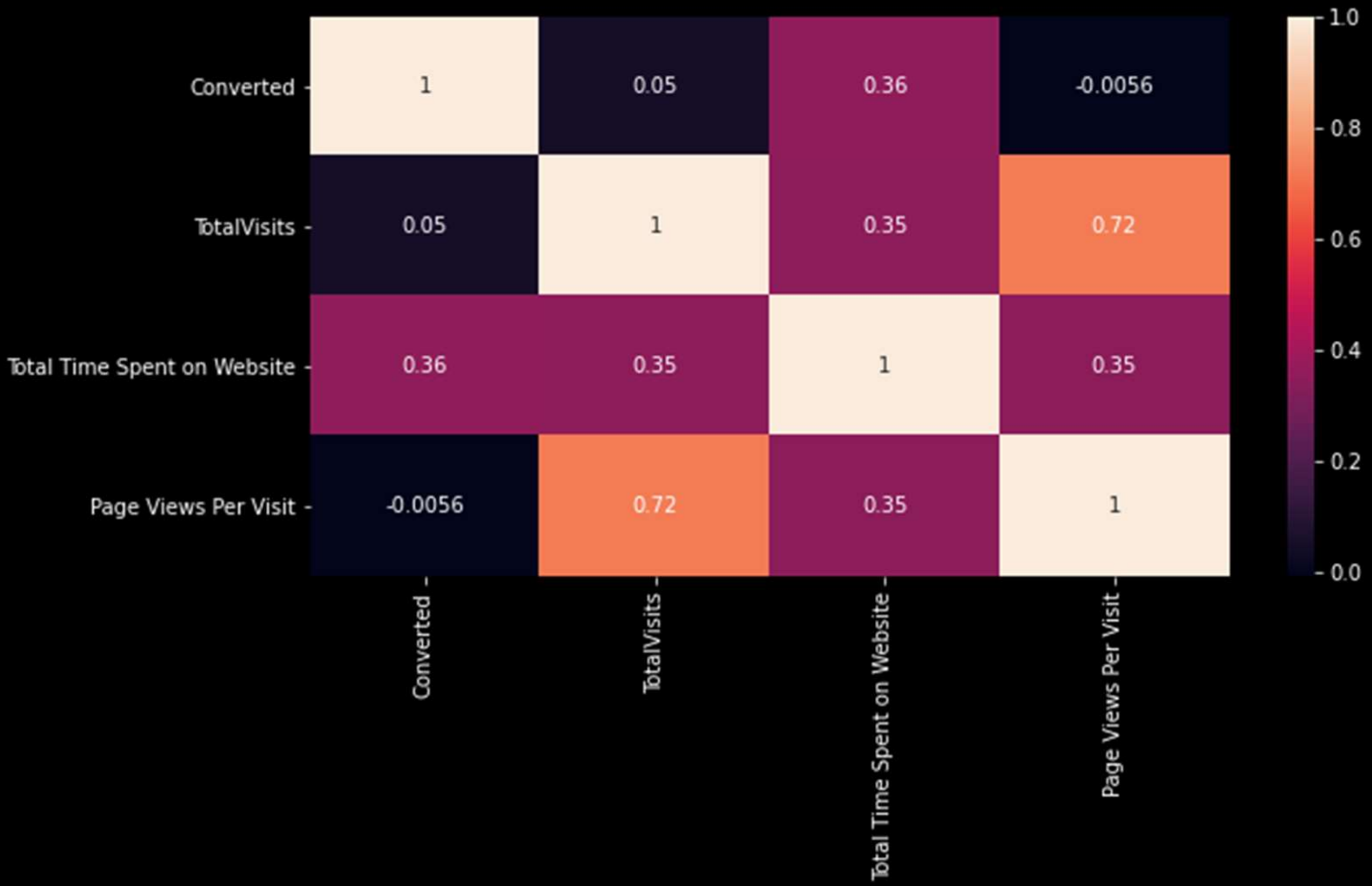


Inference:

From the above analysis, we can infer that among the leads successfully converted, most belong to those who have been sent an SMS. And second most converted leads are those whose last notable activity is Email opened.

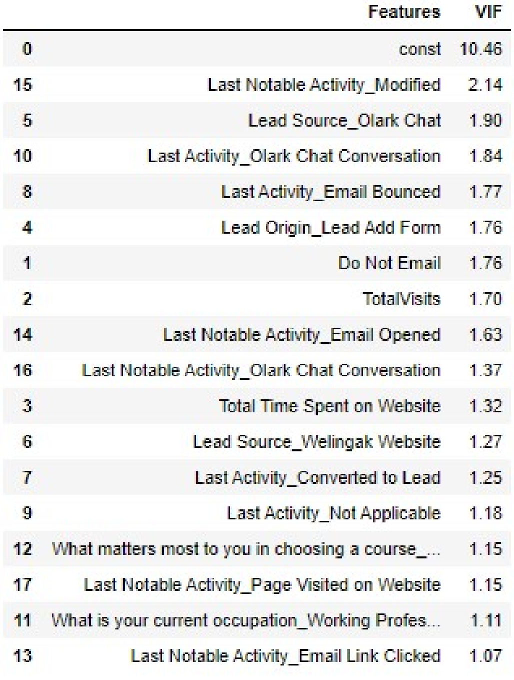
Inference:

In this heatmap, these is a clear correlation between ‘Total Visits’ and ‘Page views per Visit’. Also there is comparatively high correlation between our target variable- Conversions and ‘Total visits’

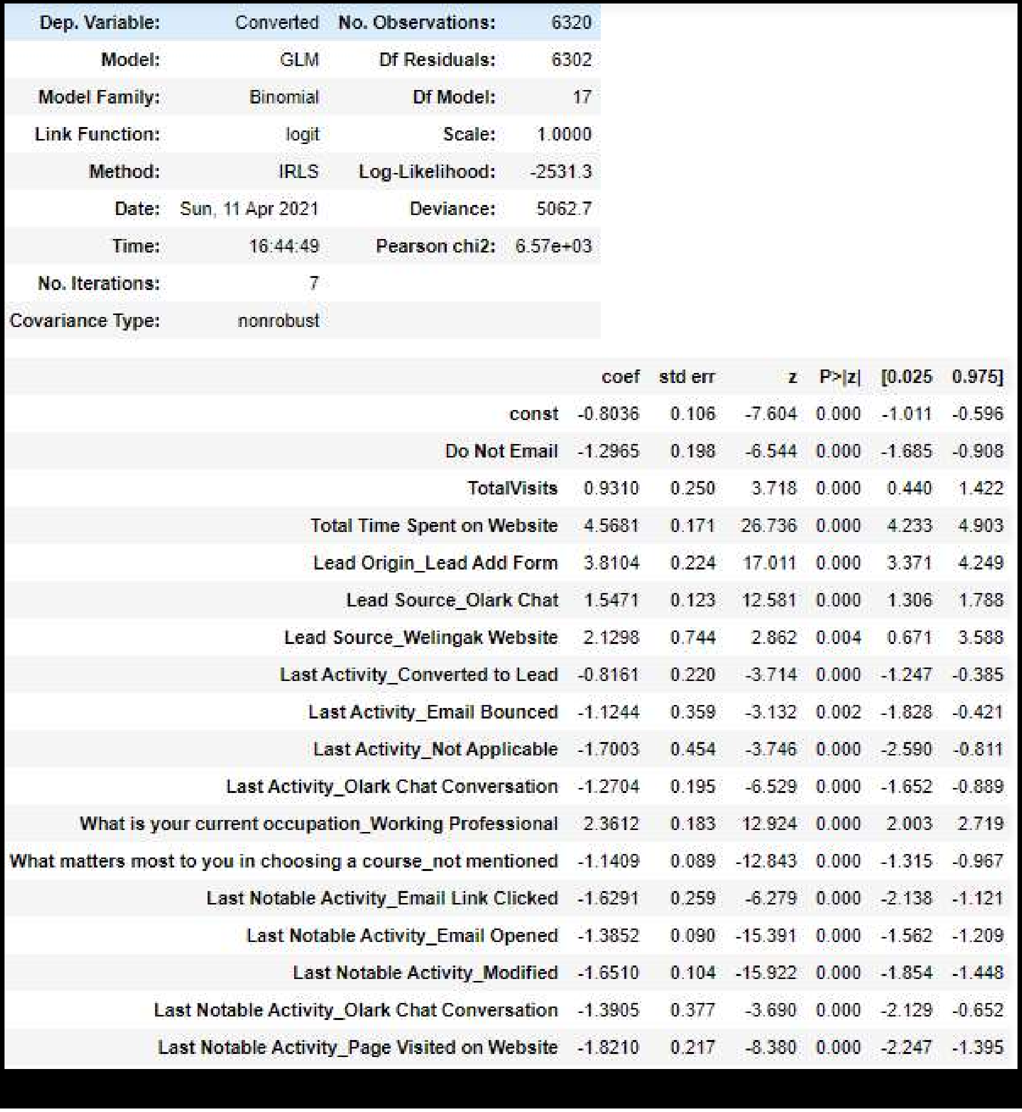


FINAL MODEL

As there is no sign of multicollinearity shown from VIF data frame, we can confirm that [Model 5] is our final model and we are going to use it predict the X train dataset. The P-values are also very sustainable for further analysis.



	Features	VIF
0	const	10.46
15	Last Notable Activity_Modified	2.14
5	Lead Source_Olark Chat	1.90
10	Last Activity_Olark Chat Conversation	1.84
8	Last Activity_Email Bounced	1.77
4	Lead Origin_Lead Add Form	1.76
1	Do Not Email	1.76
2	TotalVisits	1.70
14	Last Notable Activity_Email Opened	1.63
16	Last Notable Activity_Olark Chat Conversation	1.37
3	Total Time Spent on Website	1.32
6	Lead Source_Welingak Website	1.27
7	Last Activity_Converted to Lead	1.25
9	Last Activity_Not Applicable	1.18
12	What matters most to you in choosing a course_...	1.15
17	Last Notable Activity_Page Visited on Website	1.15
11	What is your current occupation_Working Profes...	1.11
13	Last Notable Activity_Email Link Clicked	1.07



Dep. Variable:	Converted	**No. Observations:**	6320
Model:	GLM	**Df Residuals:**	6302
Model Family:	Binomial	**Df Model:**	17
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2531.3
Date:	Sun, 11 Apr 2021	**Deviance:**	5062.7
Time:	16:44:49	**Pearson chi2:**	6.57e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

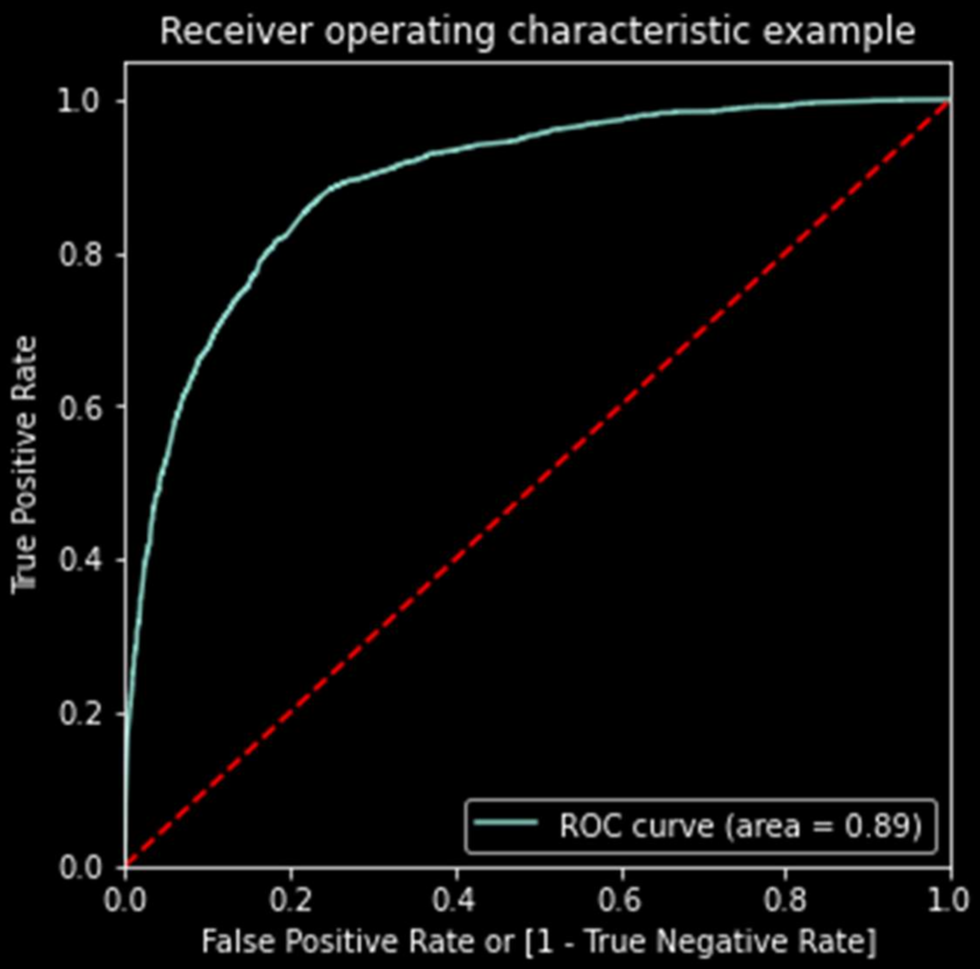
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8036	0.106	-7.604	0.000	-1.011	-0.596
Do Not Email	-1.2965	0.198	-6.544	0.000	-1.685	-0.908
TotalVisits	0.9310	0.250	3.718	0.000	0.440	1.422
Total Time Spent on Website	4.5681	0.171	26.736	0.000	4.233	4.903
Lead Origin_Lead Add Form	3.8104	0.224	17.011	0.000	3.371	4.249
Lead Source_Olark Chat	1.5471	0.123	12.581	0.000	1.306	1.788
Lead Source_Welingak Website	2.1298	0.744	2.862	0.004	0.671	3.588
Last Activity_Converted to Lead	-0.8161	0.220	-3.714	0.000	-1.247	-0.385
Last Activity_Email Bounced	-1.1244	0.359	-3.132	0.002	-1.828	-0.421
Last Activity_Not Applicable	-1.7003	0.454	-3.746	0.000	-2.590	-0.811
Last Activity_Olark Chat Conversation	-1.2704	0.195	-6.529	0.000	-1.652	-0.889
What is your current occupation_Working Professional	2.3612	0.183	12.924	0.000	2.003	2.719
What matters most to you in choosing a course_not mentioned	-1.1409	0.089	-12.843	0.000	-1.315	-0.967
Last Notable Activity_Email Link Clicked	-1.6291	0.259	-6.279	0.000	-2.138	-1.121
Last Notable Activity_Email Opened	-1.3852	0.090	-15.391	0.000	-1.562	-1.209
Last Notable Activity_Modified	-1.6510	0.104	-15.922	0.000	-1.854	-1.448
Last Notable Activity_Olark Chat Conversation	-1.3905	0.377	-3.690	0.000	-2.129	-0.652
Last Notable Activity_Page Visited on Website	-1.8210	0.217	-8.380	0.000	-2.247	-1.395

ROC CURVE

Step taken :

Points to be concluded from above ROC curve –

- The curve is closer to the left side of the border than to the right side hence our model is having a high accuracy, which is great.
- The area under the curve is 89% of the total area which is also high.

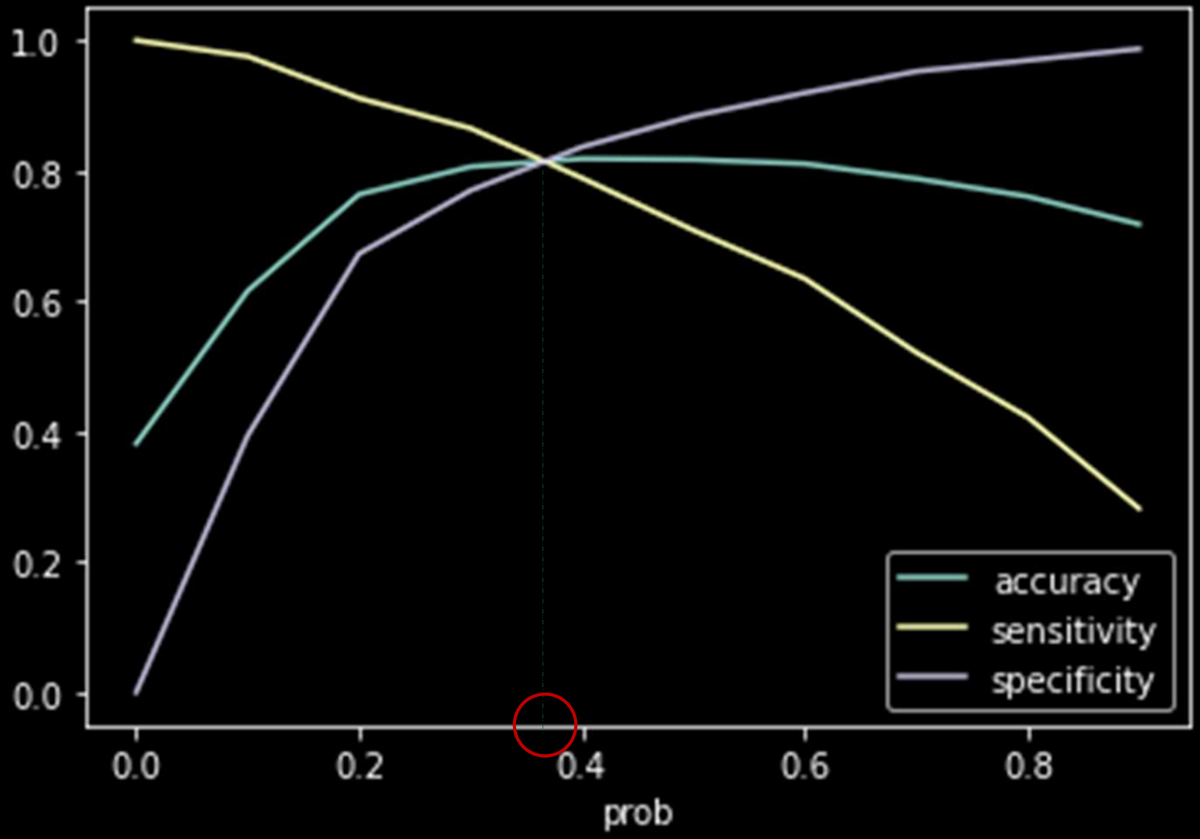


CUT OFF PROBABILITY

Step taken:

From the curve above, 0.38 is the optimum point to take it as a cutoff probability.

This cutoff will be then used for further prediction and analysis.



MODEL EVALUATION (TRAIN SET)

Confusion Matrix

TN 3232	FP 677
FN 477	TP 1934

Accuracy

81.7%

Sensitivity

80.2%

Specificity

82.6%

Precision

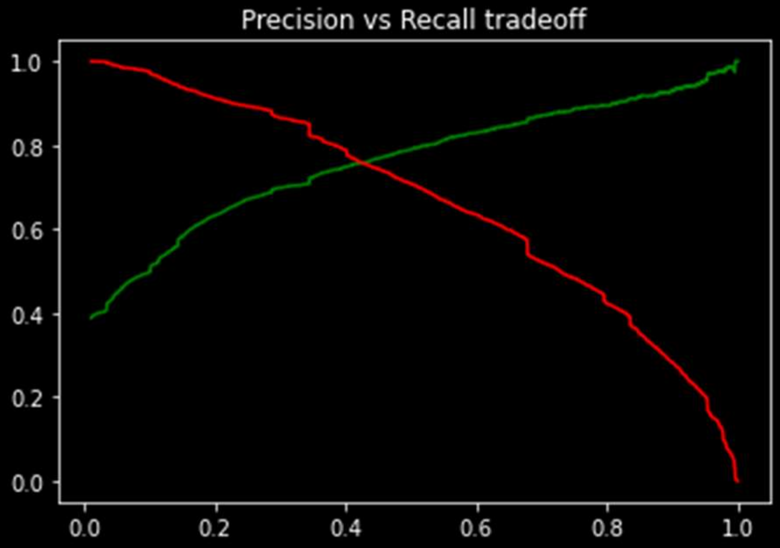
74%

Recall

80.2%

Inference:

Hence, we can see that the final prediction of conversions with a target of 80% (Sensitivity) as per the X Education's requirement . Hence we can confirm that this is a good model as per CEO’s requirement being at least 80%.



MODEL EVALUATION (TEST SET)

Confusion Matrix

TN 1342	FP 301
FN 208	TP 858

Accuracy

82%

Specificity

81.6%

Sensitivity

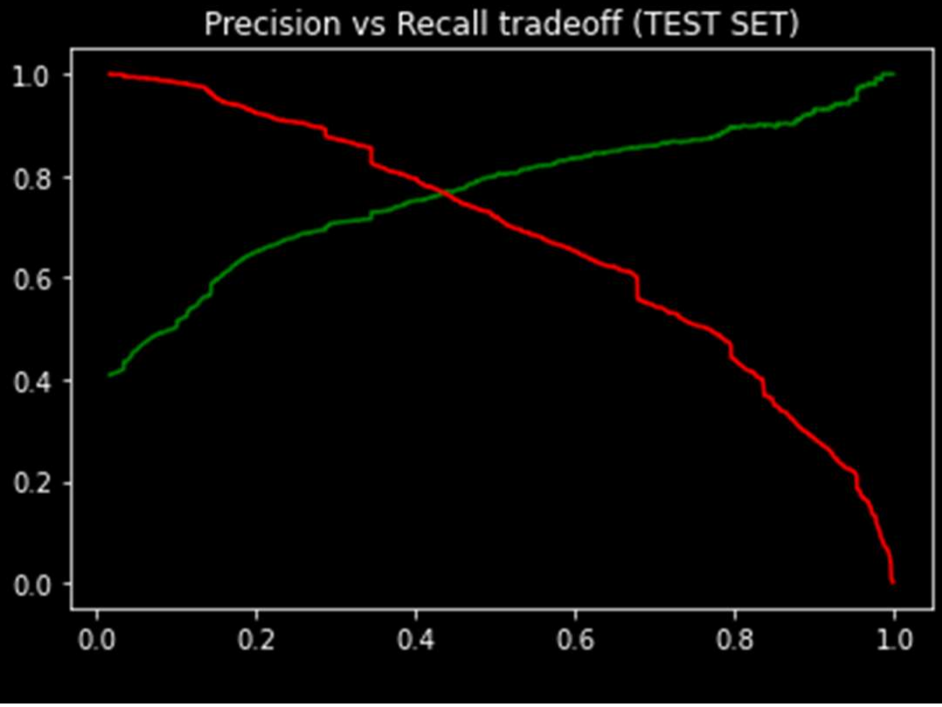
80.4%

Precision

74%

Recall

80.4%



Top 3 Variables

- Total time spent on Website
- Lead origin (Lead Add Form)
- What is your current occupation (Working professional)

Conclusion:

1. While we have checked Sensitivity-Specificity, Precision and Recall Metrics, we have considered the optimal cut off of 0.38 based on the metrics for calculating the final prediction.

2. Accuracy, Sensitivity and Specificity values of test set are around 82%, 80% and 82% which are approximately closer to the respective values calculated using trained set.

3. Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model being around 80%.

4. Hence overall this model seems to be good as the CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.