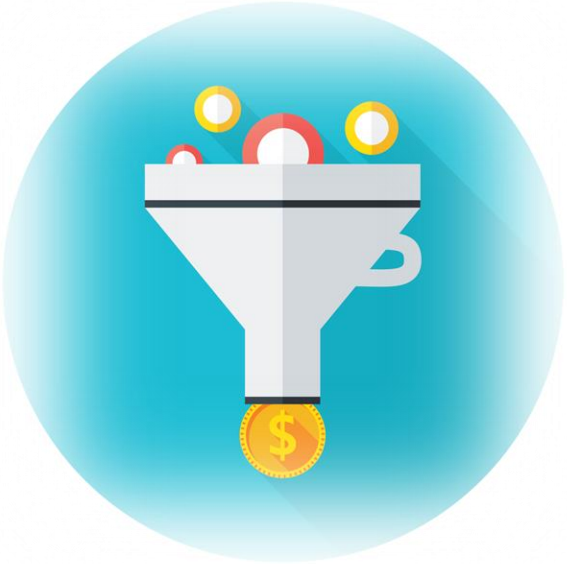
Lead Scoring

BY SABYASACHI PARIDA

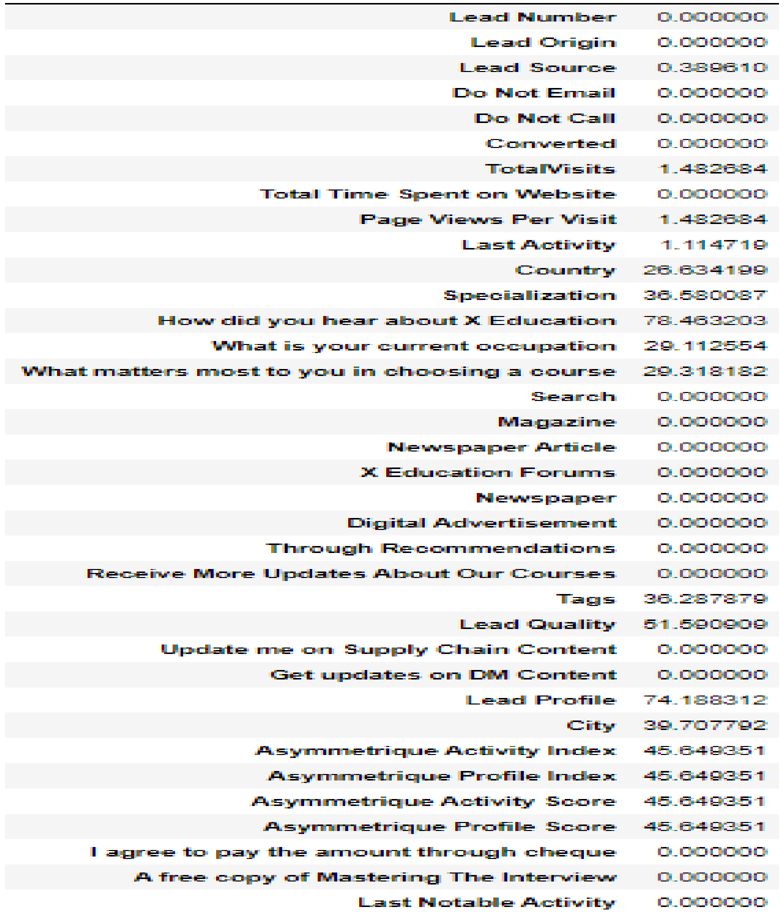
Objective

I. 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.

II. 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.



Data Distribution



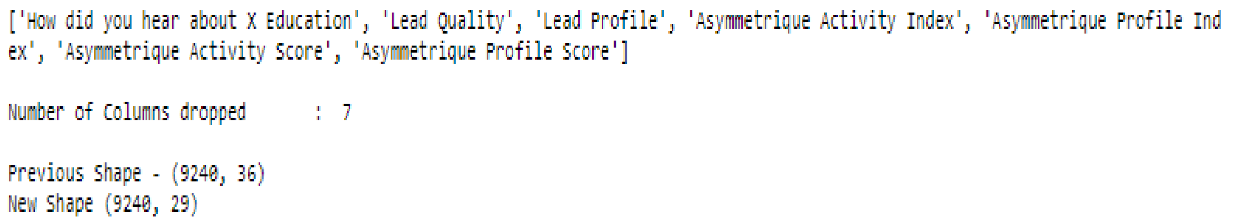
Lead Number	0.000000
Lead Origin	0.000000
Lead Source	0.389610
Do Not Email	0.000000
Do Not Call	0.000000
Converted	0.000000
TotalVisits	1.482684
Total Time Spent on Website	0.000000
Page Views Per Visit	1.482684
Last Activity	1.114719
Country	26.634199
Specialization	36.580087
How did you hear about X Education	78.463203
What is your current occupation	29.112554
What matters most to you in choosing a course	29.318182
Search	0.000000
Magazine	0.000000
Newspaper Article	0.000000
X Education Forums	0.000000
Newspaper	0.000000
Digital Advertisement	0.000000
Through Recommendations	0.000000
Receive More Updates About Our Courses	0.000000
Tags	36.287879
Lead Quality	51.590909
Update me on Supply Chain Content	0.000000
Get updates on DM Content	0.000000
Lead Profile	74.188312
City	39.707792
Asymmetrique Activity Index	45.649351
Asymmetrique Profile Index	45.649351
Asymmetrique Activity Score	45.649351
Asymmetrique Profile Score	45.649351
I agree to pay the amount through cheque	0.000000
A free copy of Mastering The Interview	0.000000
Last Notable Activity	0.000000

Observation

I. There are few columns which have more than 45% Null Values.

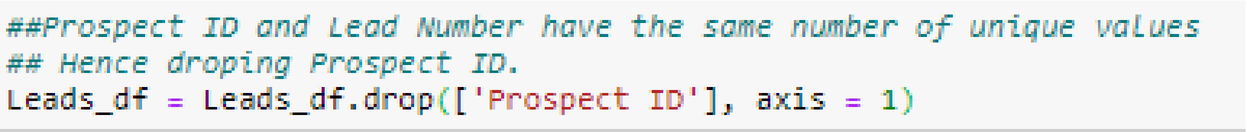
II. They wont add any value to our analysis and modeling because of their large null values.

III. Dropping those Columns .



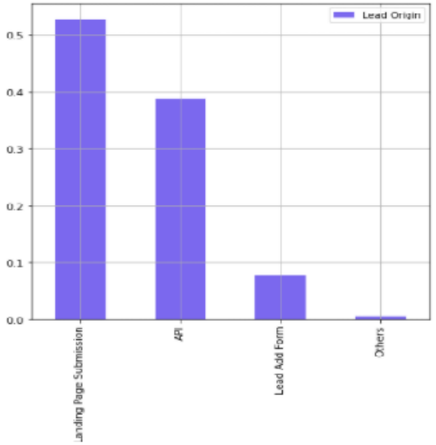
```
['How did you hear about X Education', 'Lead Quality', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique Profile Ind
ex', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score']

Number of Columns dropped      :  7

Previous Shape - (9240, 36)
New Shape (9240, 29)
```

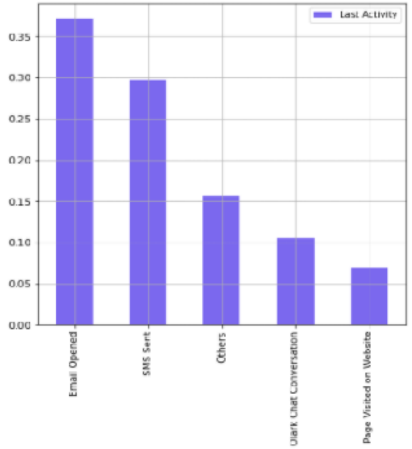
I. It was observed that Prospect ID and Lead Number have unique values and any of them can be used to uniquely identify any lead. So **ProspectID** was dropped

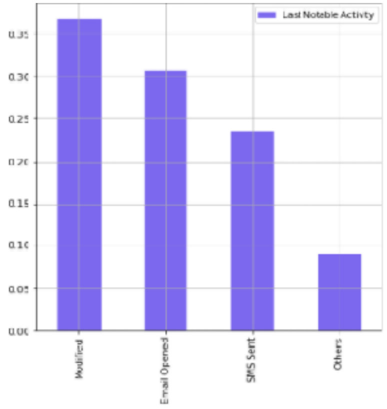


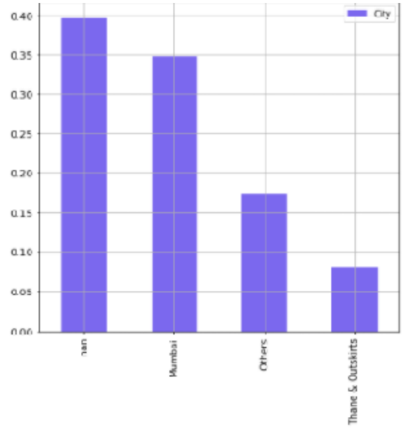
```
##Prospect ID and Lead Number have the same number of unique values
## Hence droping Prospect ID.
Leads_df = Leads_df.drop(['Prospect ID'], axis = 1)
```

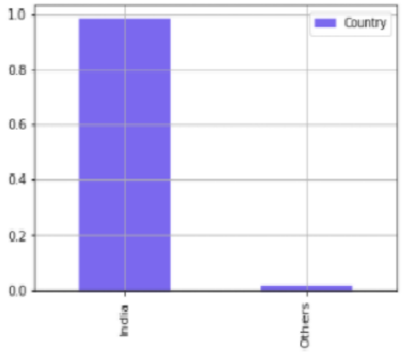
Data Distribution – A Brief Highlight

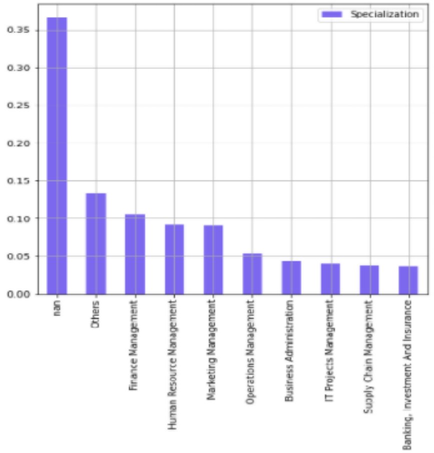


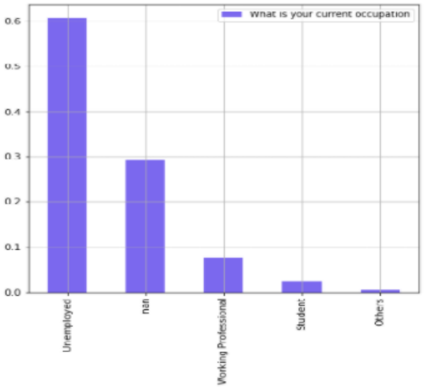


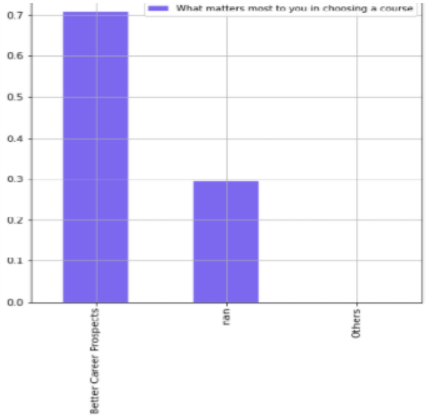










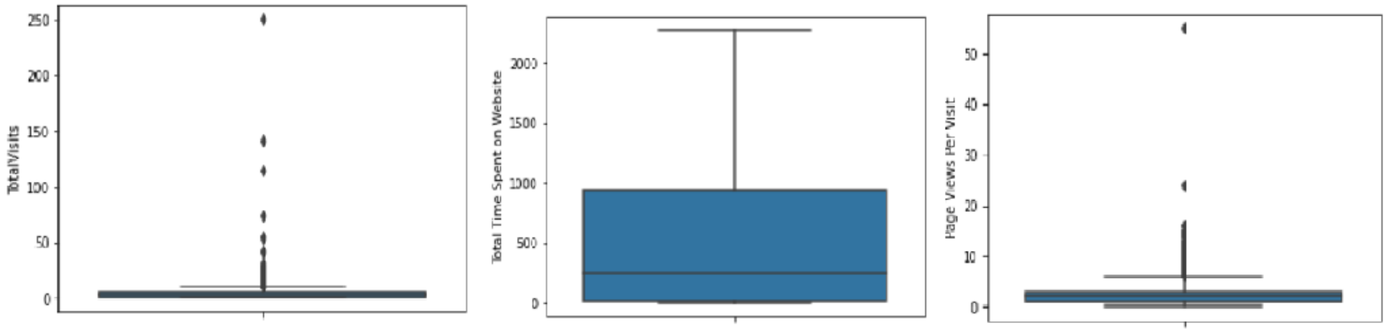


Observation

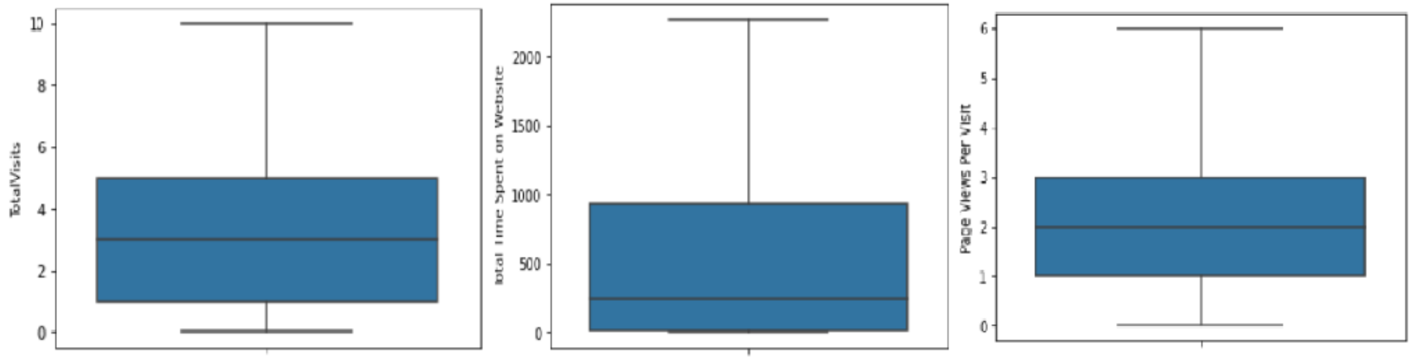
I. Most of the lead's origin is Landing page Submission (52.9%).
II. Most of the Lead source(31.1%) is from Google.
III. In the case of specialization, Not Specified has the highest count followed by Others followed by Finance Management.
IV. Most of the Leads are from Mumbai and had country India.
V. Most of the leads were un-employed.
VI. Most of the Leads were there for better Career Opportunity.
VII. At all those places where the number of variables was of too less percentage. It was grouped into group- Others.

Outlier Analysis – Numerical Variables

As Is



After Capping



Observation

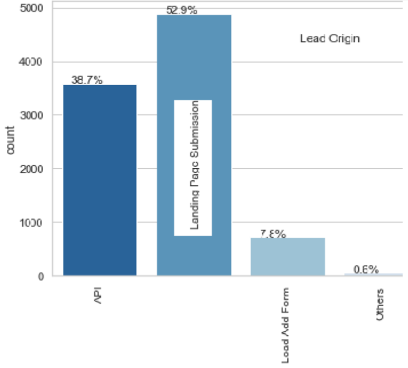
I. There are lot of outliers present in Total visits .Mean wont be the best metrics to impute them rather Median would be the best metric to impute.

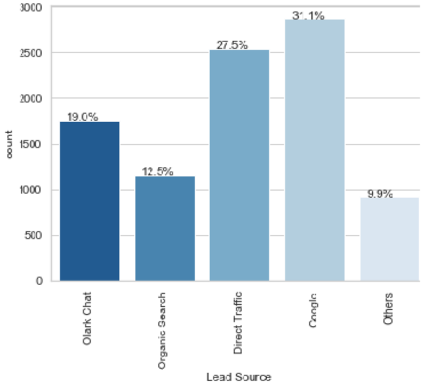
II. Same is the case for Page views per visit.

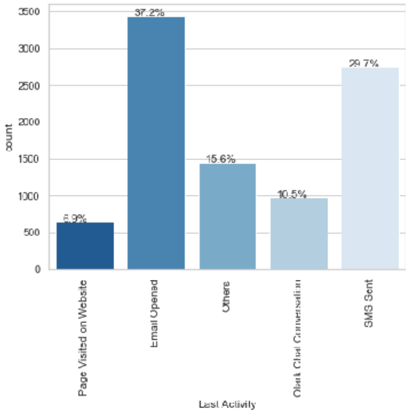
III. There are a lot of values outside 95% in the higher bound.

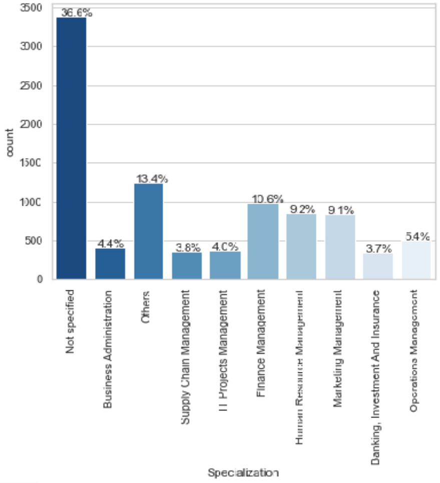
IV. Capping was done in order to treat them such that to high values don’t influence the model.

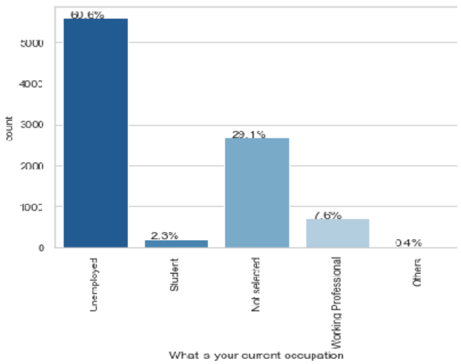
Univariate Analysis – Variables with multiple categories

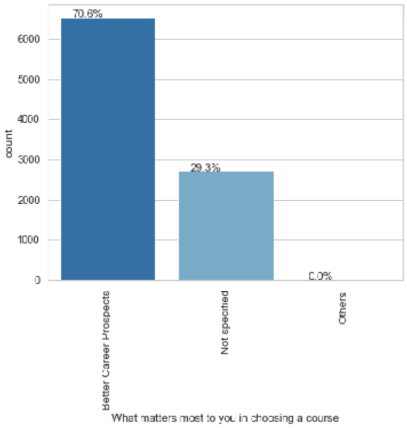


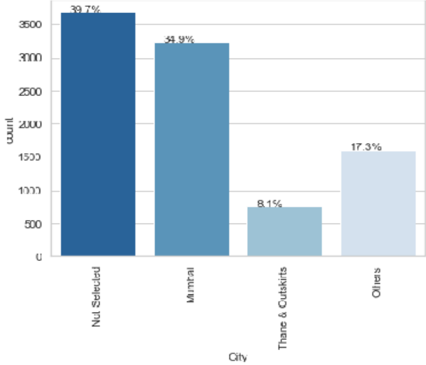


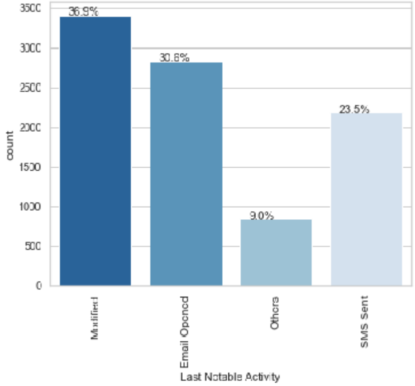












Observation

I. Most of the lead's origin is Landing page Submission and google was the source.

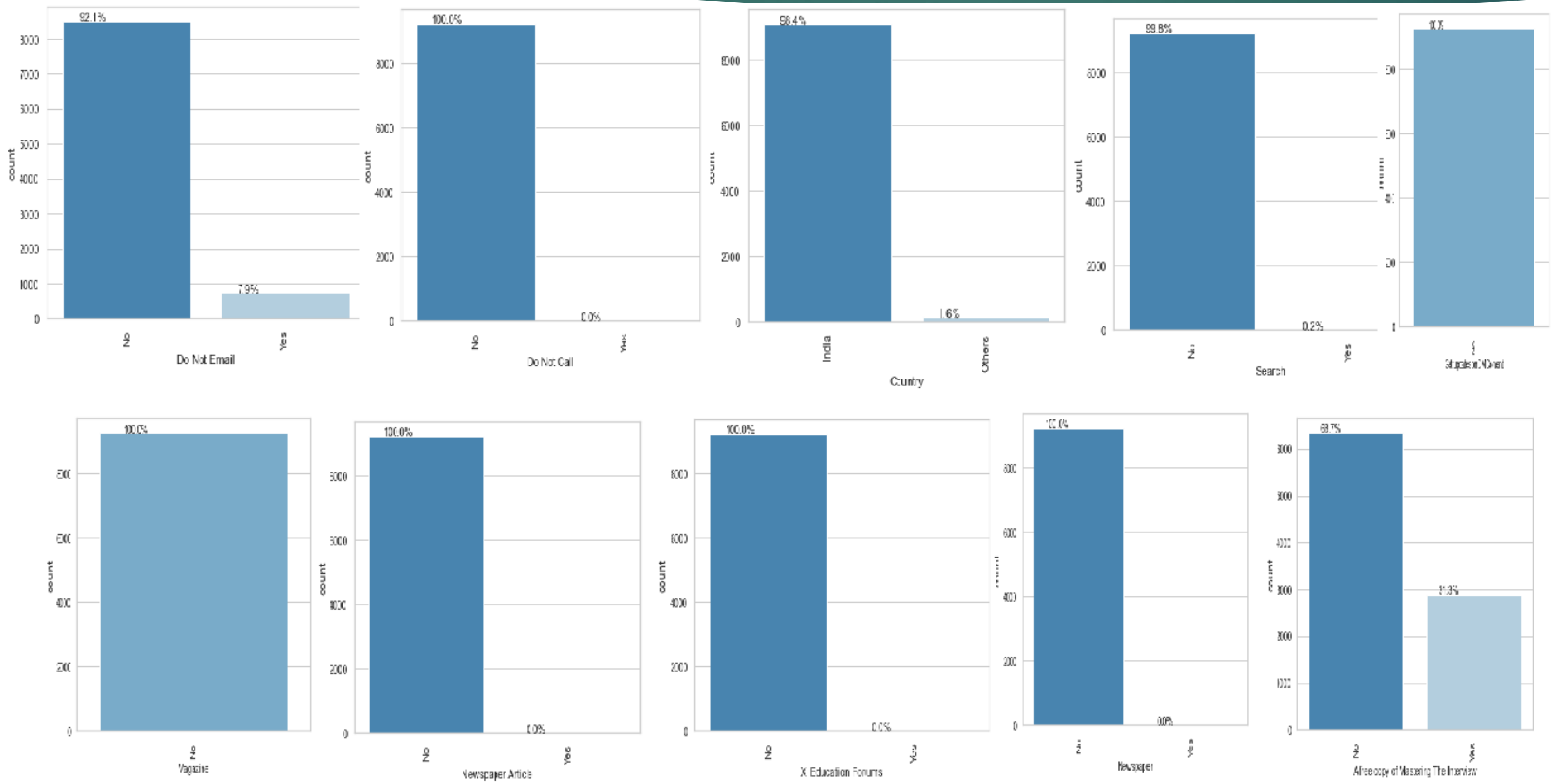
II. Email opened is the highest Last activity.

III. In the case of specialization, Not Specified has the highest count followed by Others followed by Finance Management.

IV. Most of the leads are unemployed, and about 29.1% did not select any option and most of the leads are here in search of better career option.

V. In the case of Last Notable Activity modified had 36.9% followed by 30.6% .

Univariate Analysis – Variables with two categories

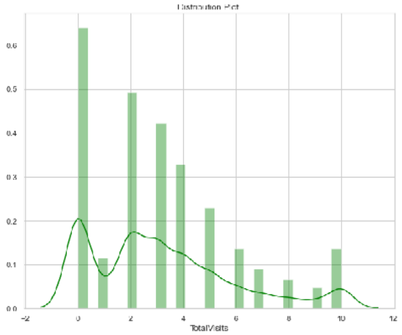


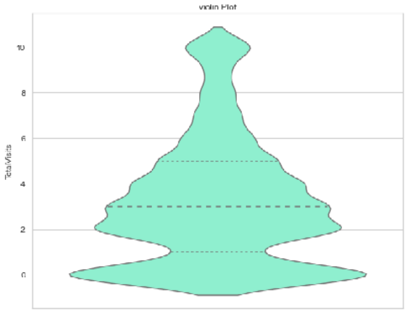
Observation

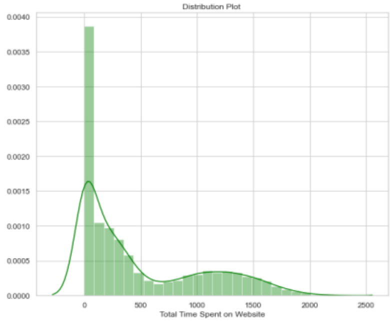
I. 92% selected not to email them ,100% not to call them ,99.8% selected Search as No,100% selected No for Magazine, newspaper article education forum, Digital Adds, Updates on courses, and 100 % selected not to pay through cheque.

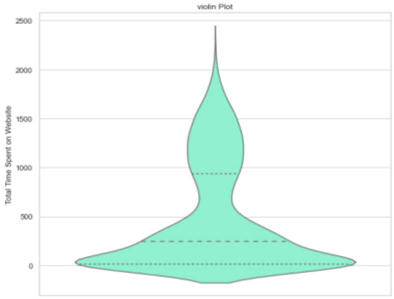
II. 31.3% selected yes to get a copy of Mastering Interview .

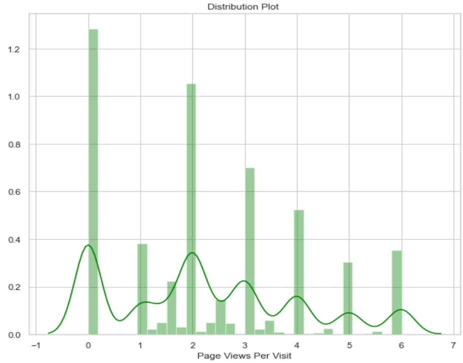
Univariate Analysis- Numerical variables

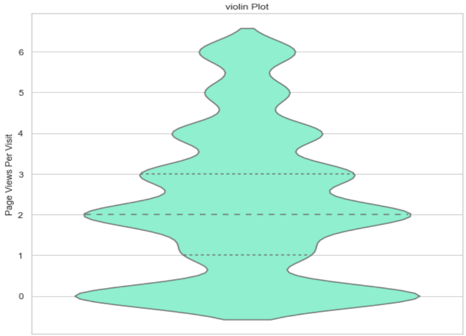








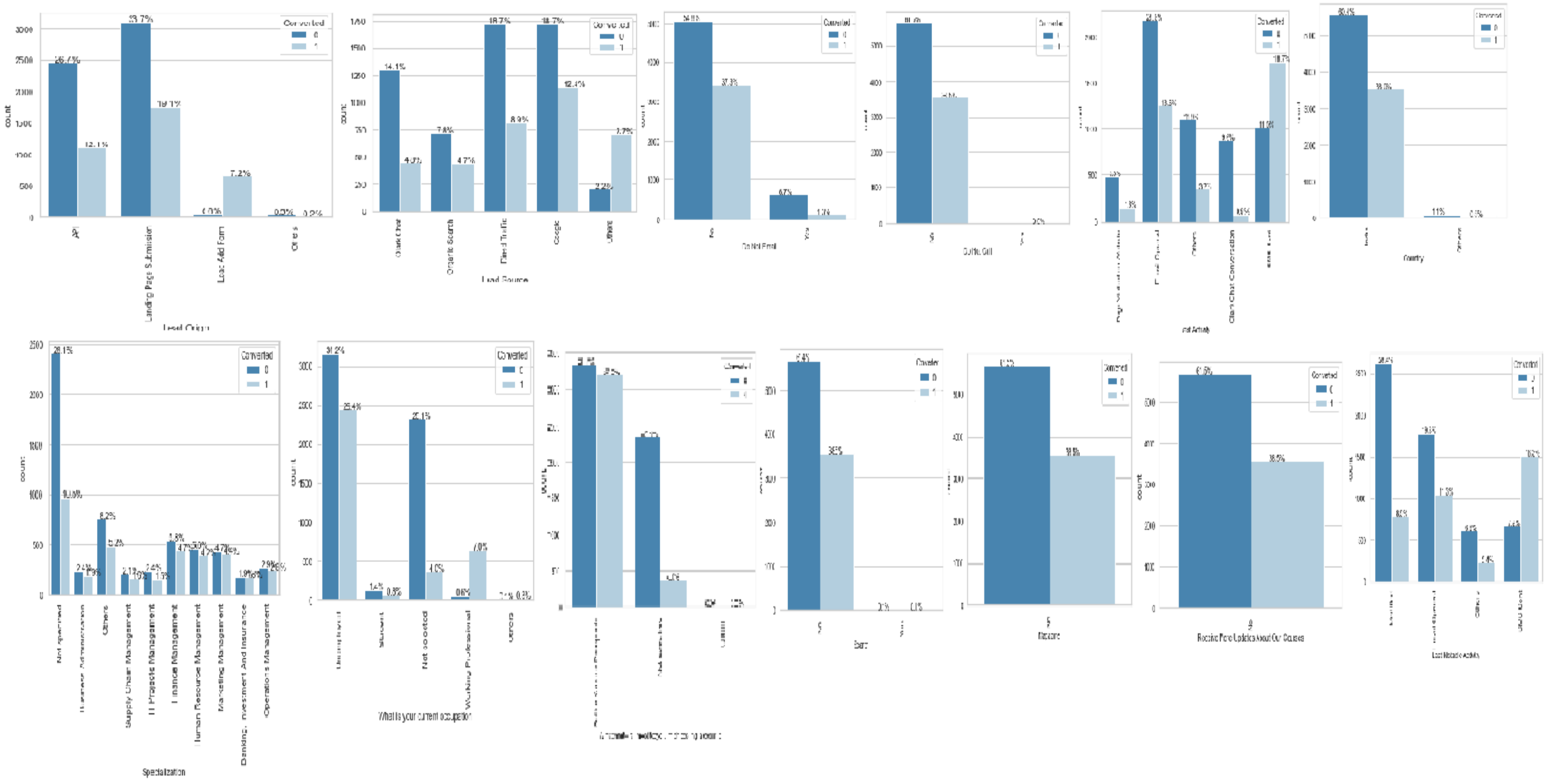




Observation

I. The variables are not normally distributed

Bivariate Analysis



Observation

I. Landing page submission has the highest conversion rate as well as non conversion rate.

II. Most of the converted leads has google as source.

III. About 38% of leads who selected do not call or email got converted.

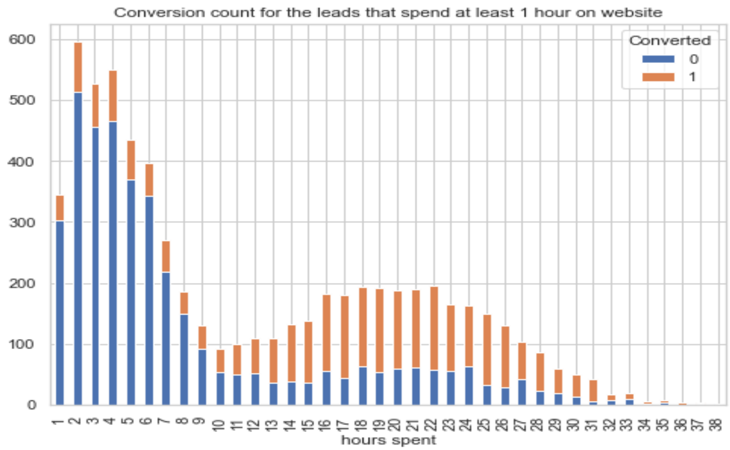
IV. SMS sent activity had the highest conversion rate.

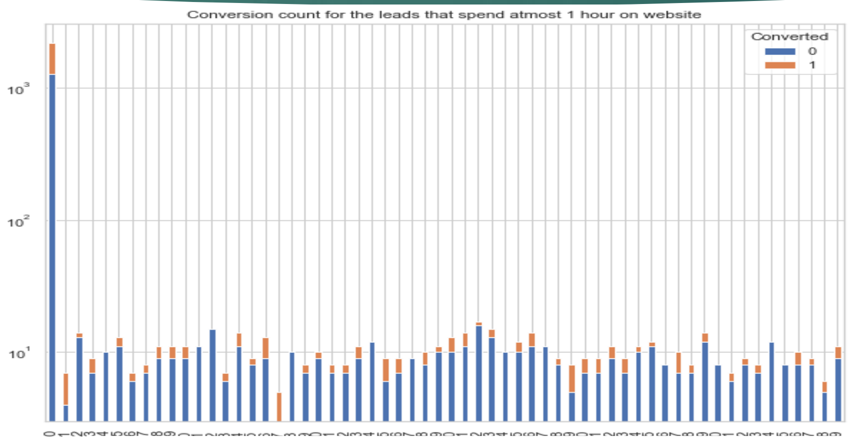
V. 26% of the leads which were converted were unemployed and 34 % of the leads were here for better career prospects.

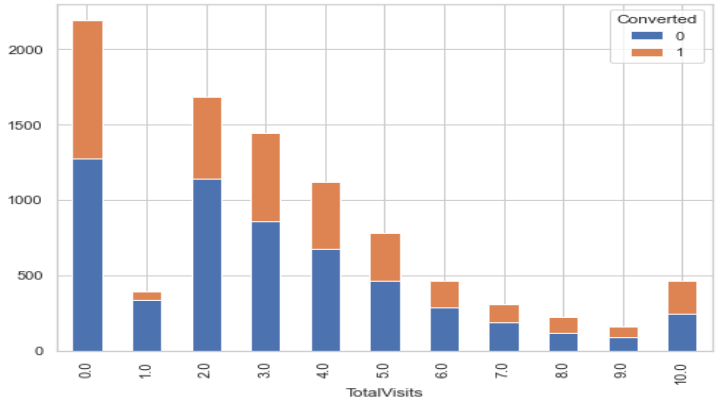
VI. About 39% of leads who selected NO to magazine, Newspaper etc got converted.

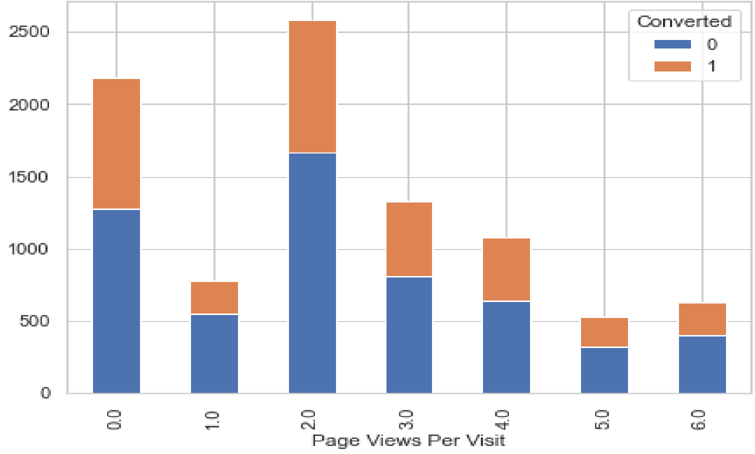
VII. SMS sent was the last notable activity which got converted which is about 17%.

Bivariate analysis – Numerical variables









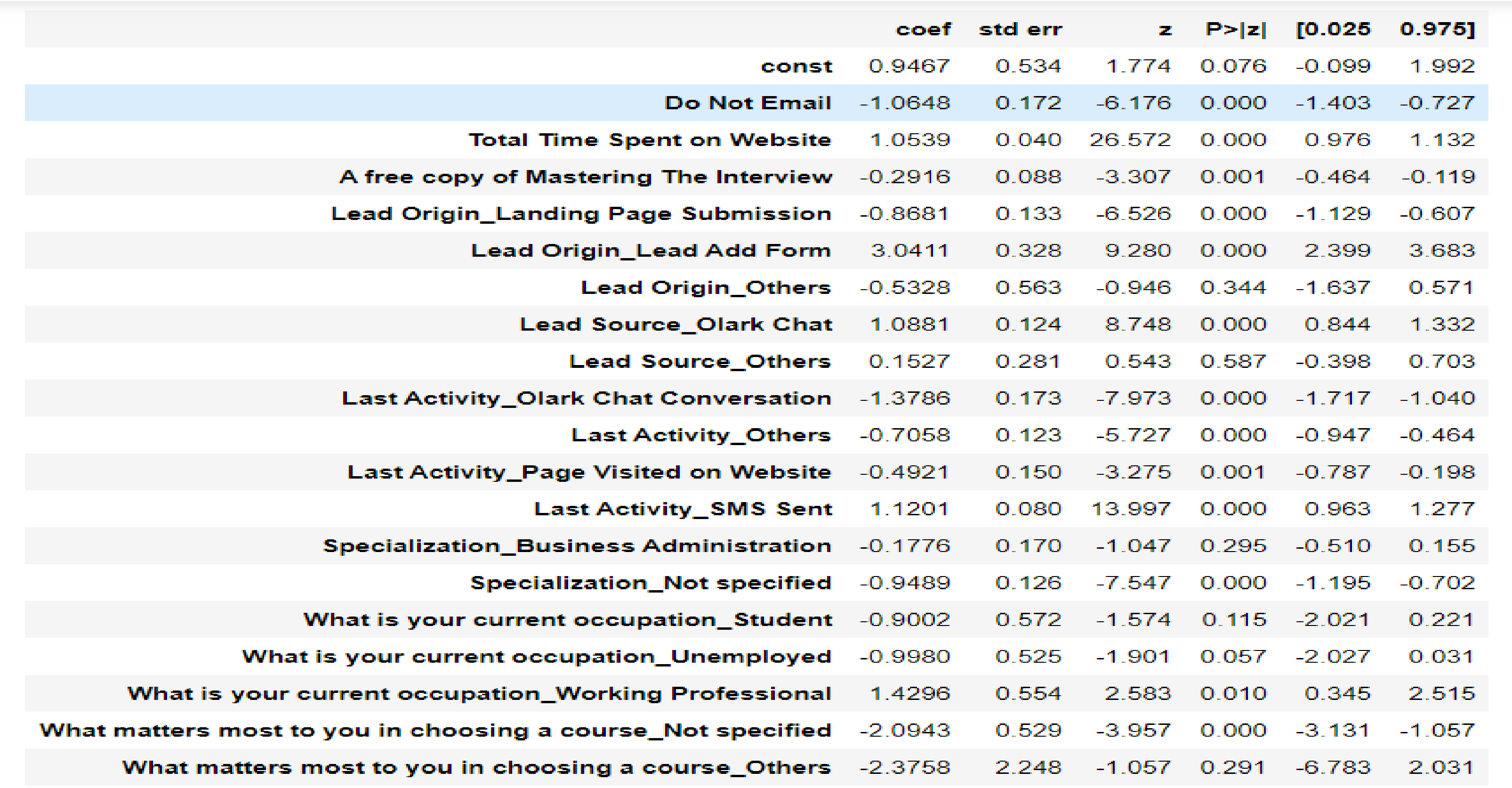
Observation

I. 17 mins has the highest conversion rate (i.e. all) and in case of leads above 1 hrs. those who spent 22 hrs. have a better conversion rate.

II. Those who have visited the page 2 times have a better conversion rate.

III. 0 - 1 times total visits have a better conversion rate.

Model results after RFE and selecting 20 Variables



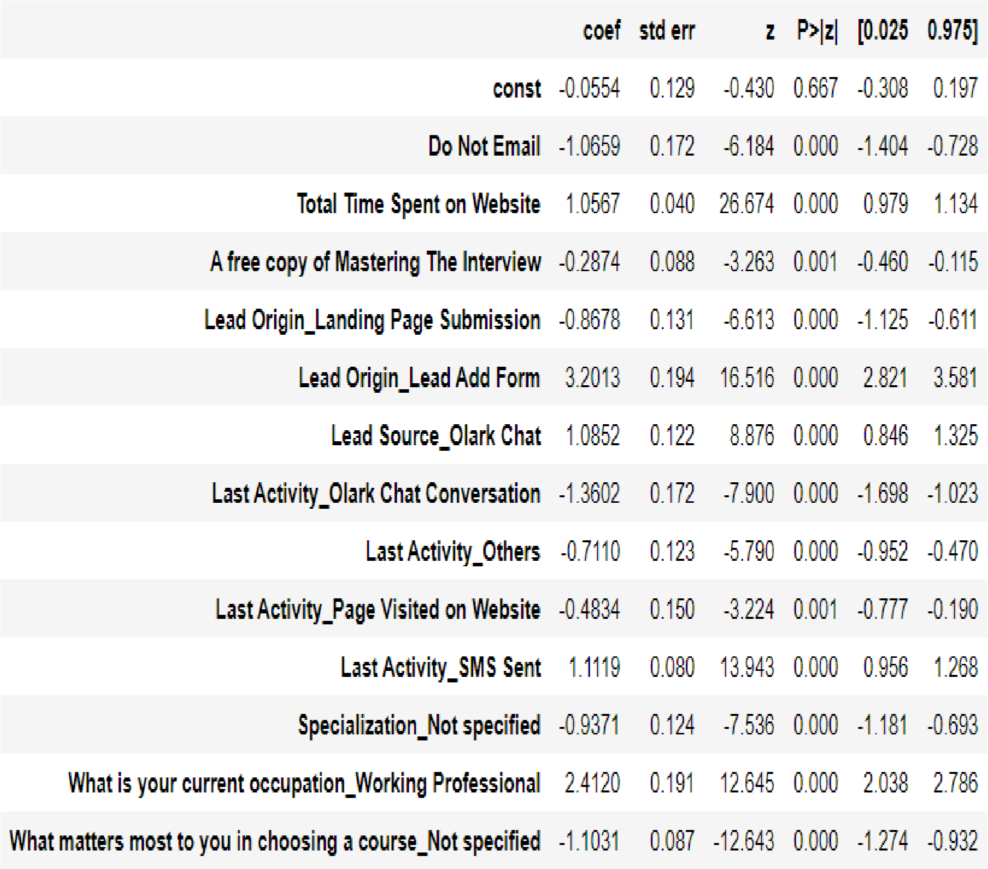
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9467	0.534	1.774	0.076	-0.099	1.992
Do Not Email	-1.0648	0.172	-6.176	0.000	-1.403	-0.727
Total Time Spent on Website	1.0539	0.040	26.572	0.000	0.976	1.132
A free copy of Mastering The Interview	-0.2916	0.088	-3.307	0.001	-0.464	-0.119
Lead Origin_Landing Page Submission	-0.8681	0.133	-6.526	0.000	-1.129	-0.607
Lead Origin_Lead Add Form	3.0411	0.328	9.280	0.000	2.399	3.683
Lead Origin_Others	-0.5328	0.563	-0.946	0.344	-1.637	0.571
Lead Source_Olark Chat	1.0881	0.124	8.748	0.000	0.844	1.332
Lead Source_Others	0.1527	0.281	0.543	0.587	-0.398	0.703
Last Activity_Olark Chat Conversation	-1.3786	0.173	-7.973	0.000	-1.717	-1.040
Last Activity_Others	-0.7058	0.123	-5.727	0.000	-0.947	-0.464
Last Activity_Page Visited on Website	-0.4921	0.150	-3.275	0.001	-0.787	-0.198
Last Activity_SMS Sent	1.1201	0.080	13.997	0.000	0.963	1.277
Specialization_Business Administration	-0.1776	0.170	-1.047	0.295	-0.510	0.155
Specialization_Not specified	-0.9489	0.126	-7.547	0.000	-1.195	-0.702
What is your current occupation_Student	-0.9002	0.572	-1.574	0.115	-2.021	0.221
What is your current occupation_Unemployed	-0.9980	0.525	-1.901	0.057	-2.027	0.031
What is your current occupation_Working Professional	1.4296	0.554	2.583	0.010	0.345	2.515
What matters most to you in choosing a course_Not specified	-2.0943	0.529	-3.957	0.000	-3.131	-1.057
What matters most to you in choosing a course_Others	-2.3758	2.248	-1.057	0.291	-6.783	2.031

Observation

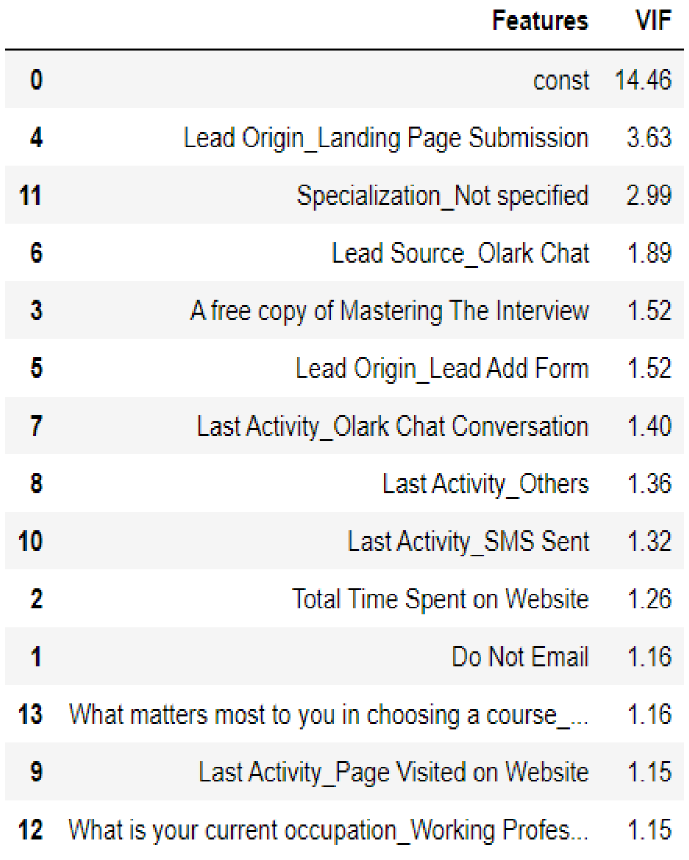
I. The following have high p value .

- What is your current occupation_Others
- What is your current occupation_Student
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional
- What matters most to you in choosing a course_Not specified

Final Model- After Manual feature Elimination



	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0554	0.129	-0.430	0.667	-0.308	0.197
Do Not Email	-1.0659	0.172	-6.184	0.000	-1.404	-0.728
Total Time Spent on Website	1.0567	0.040	26.674	0.000	0.979	1.134
A free copy of Mastering The Interview	-0.2874	0.088	-3.263	0.001	-0.460	-0.115
Lead Origin_Landing Page Submission	-0.8678	0.131	-6.613	0.000	-1.125	-0.611
Lead Origin_Lead Add Form	3.2013	0.194	16.516	0.000	2.821	3.581
Lead Source_Olark Chat	1.0852	0.122	8.876	0.000	0.846	1.325
Last Activity_Olark Chat Conversation	-1.3602	0.172	-7.900	0.000	-1.698	-1.023
Last Activity_Others	-0.7110	0.123	-5.790	0.000	-0.952	-0.470
Last Activity_Page Visited on Website	-0.4834	0.150	-3.224	0.001	-0.777	-0.190
Last Activity_SMS Sent	1.1119	0.080	13.943	0.000	0.956	1.268
Specialization_Not specified	-0.9371	0.124	-7.536	0.000	-1.181	-0.693
What is your current occupation_Working Professional	2.4120	0.191	12.645	0.000	2.038	2.786
What matters most to you in choosing a course_Not specified	-1.1031	0.087	-12.643	0.000	-1.274	-0.932

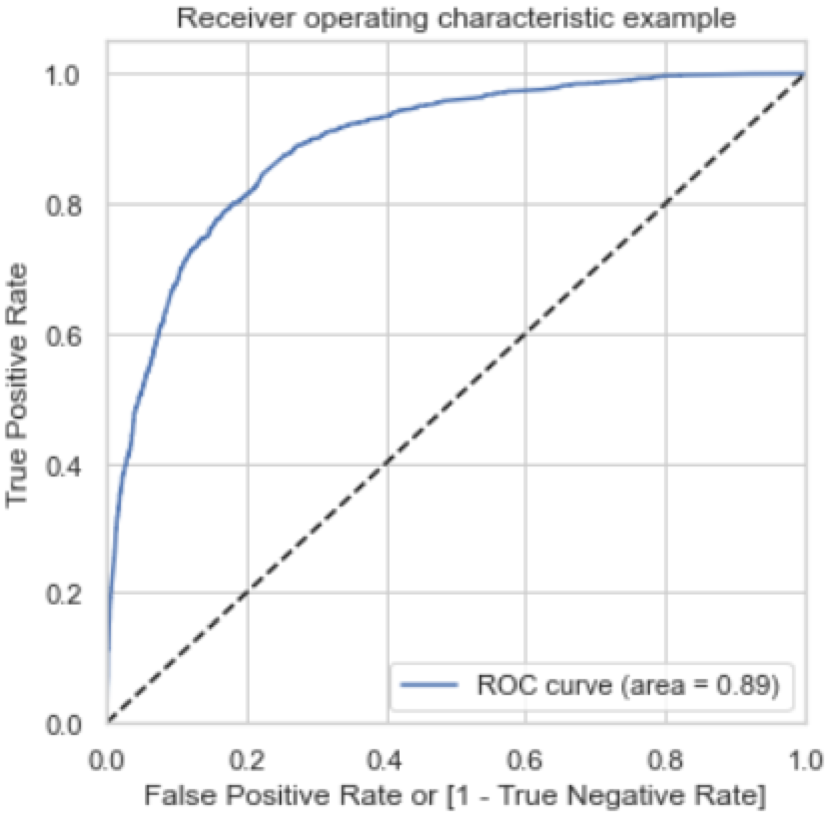


	Features	VIF
0	const	14.46
4	Lead Origin_Landing Page Submission	3.63
11	Specialization_Not specified	2.99
6	Lead Source_Olark Chat	1.89
3	A free copy of Mastering The Interview	1.52
5	Lead Origin_Lead Add Form	1.52
7	Last Activity_Olark Chat Conversation	1.40
8	Last Activity_Others	1.36
10	Last Activity_SMS Sent	1.32
2	Total Time Spent on Website	1.26
1	Do Not Email	1.16
13	What matters most to you in choosing a course_...	1.16
9	Last Activity_Page Visited on Website	1.15
12	What is your current occupation_Working Profes...	1.15

Observation

I. All the VIFs and P values are within the range .

ROC Curve



Observation

I. An ROC curve demonstrates several things:

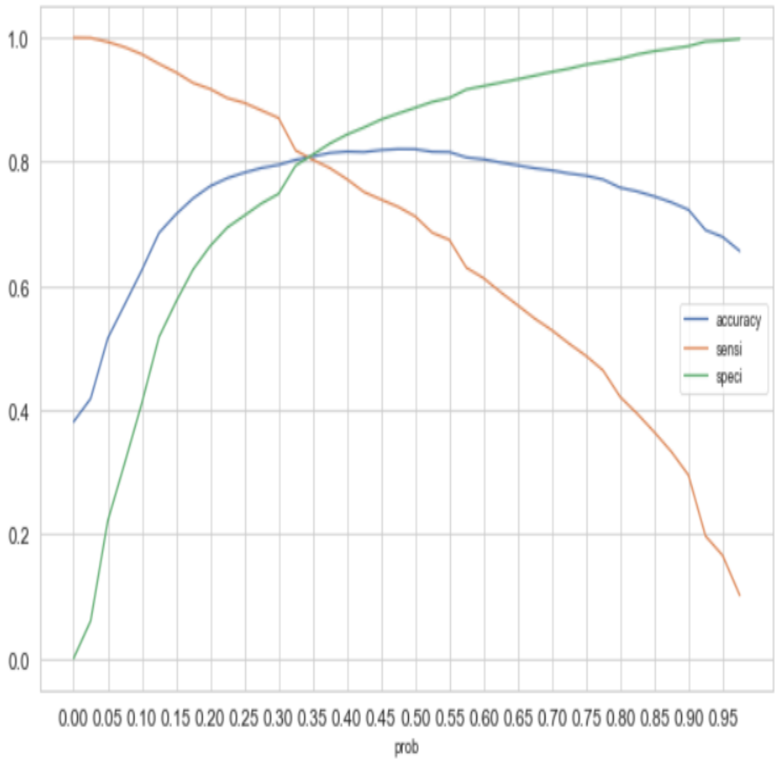
II. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

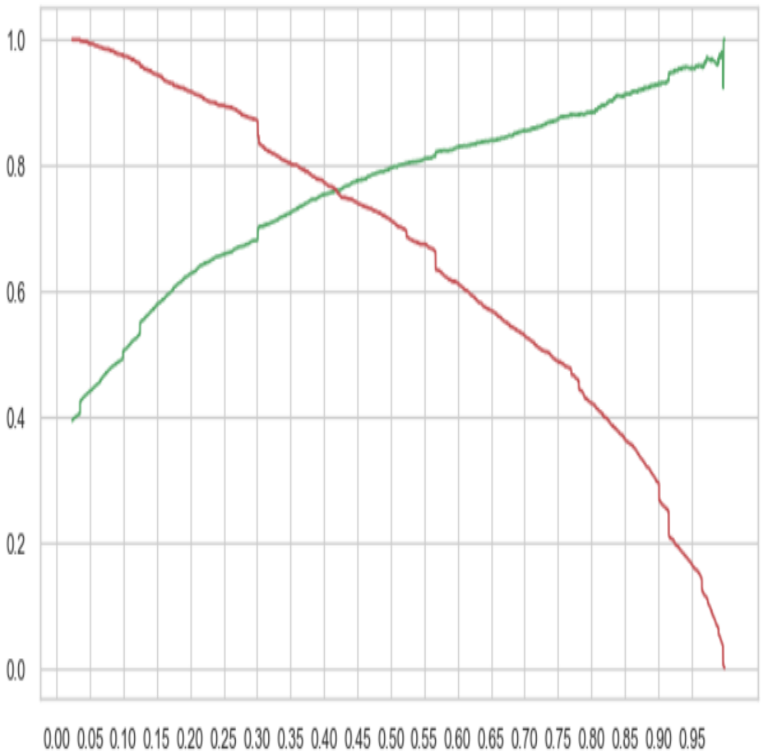
III. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

IV. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

V. Area under the Curve is 0.89

Optimal Cut-off Point & Precision - recall trade-off





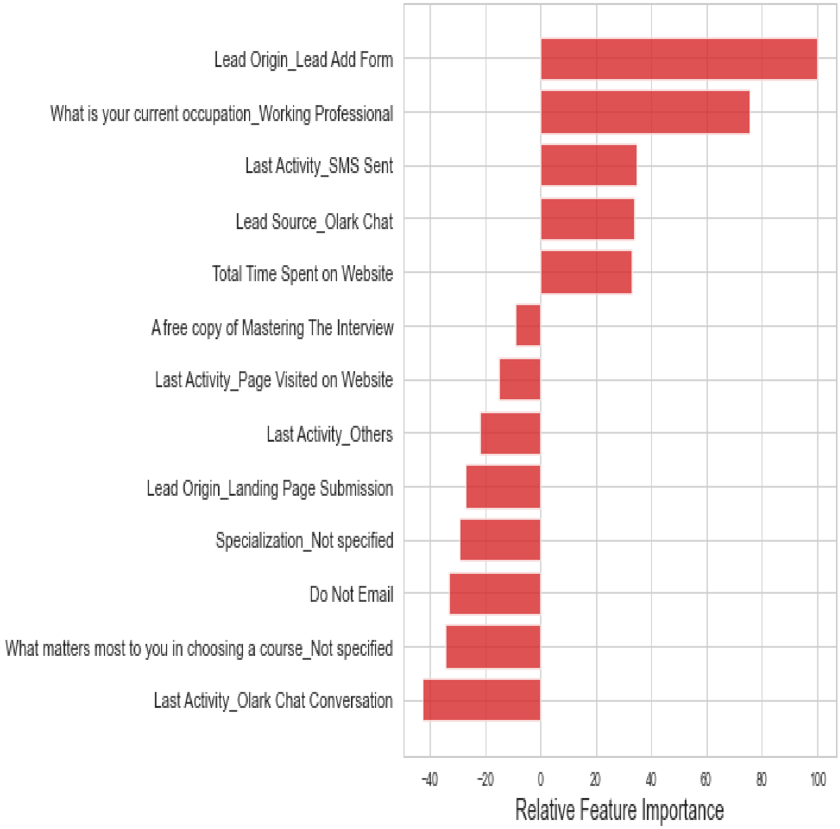
Observation

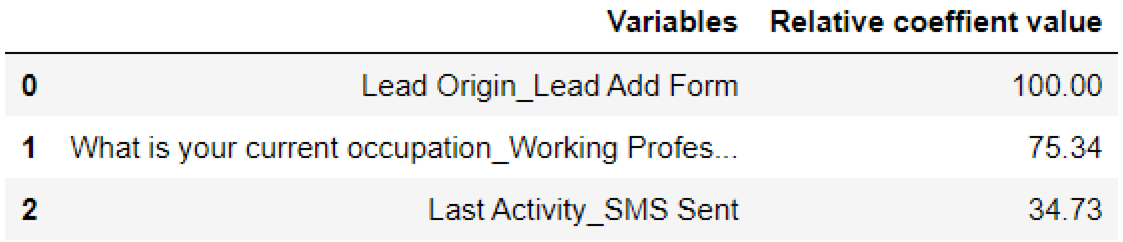
I. The Optimal cutoff seems to be at 0.34 from the graph where we will get high Accuracy, Sensitivity and Specificity.

II. The Higher the precision lower the Recall and vice versa.

III. We will take our previous cutoff i.e. 0.34 as we have a good model with high accuracy and sensitivity

Relative Feature Importance





	Variables	Relative coeffient value
0	Lead Origin_Lead Add Form	100.00
1	What is your current occupation_Working Profes...	75.34
2	Last Activity_SMS Sent	34.73

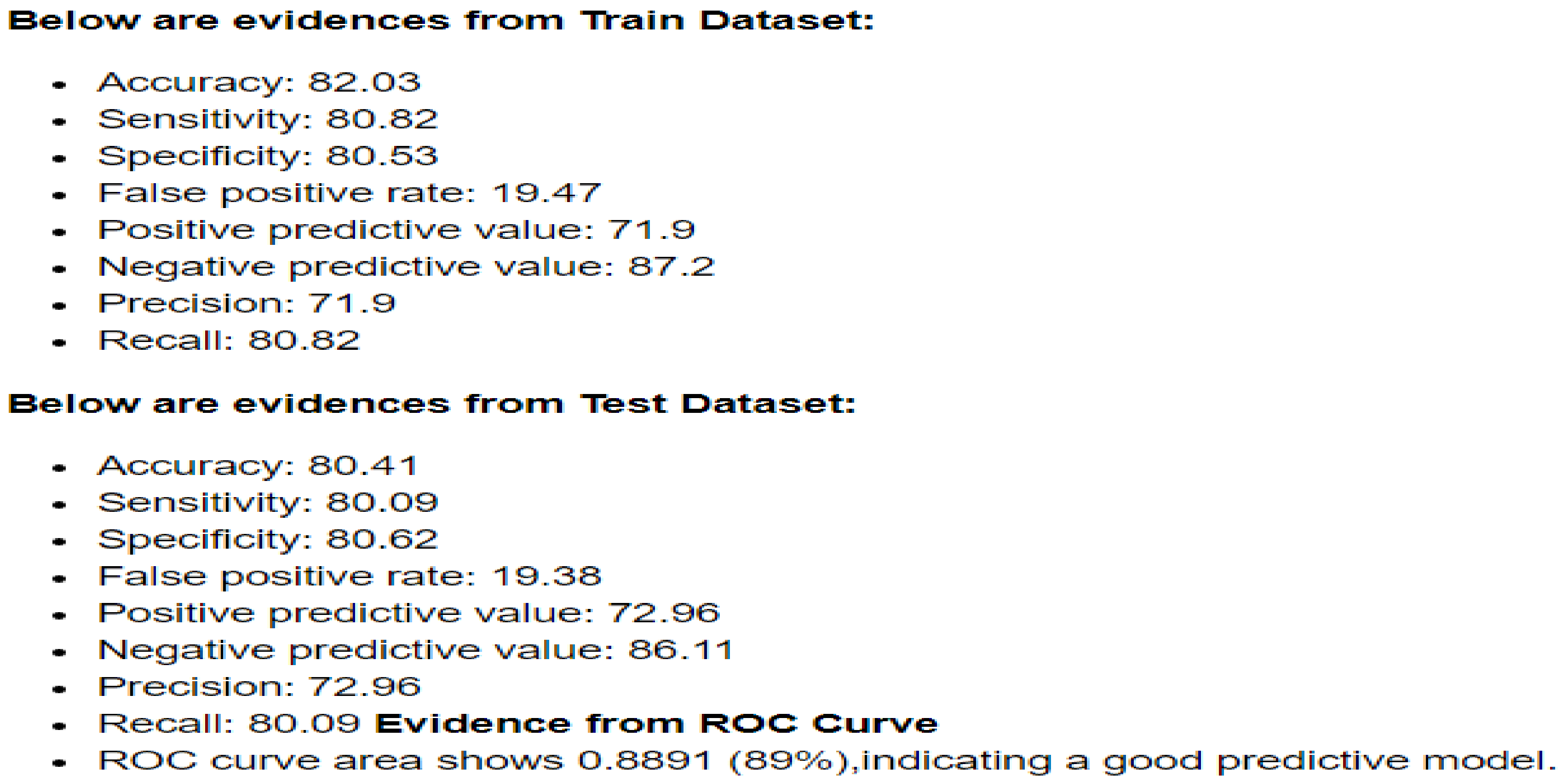
Observation

I. Lead Origin_Lead Add Form - This has 100 coefficient value

II. What is your current occupation_Working Professional - It has 75.34 coefficient value

III. Last Activity_SMS Sent - It has 34.73 coefficient value

Conclusion



Below are evidences from Train Dataset:

- Accuracy: 82.03
- Sensitivity: 80.82
- Specificity: 80.53
- False positive rate: 19.47
- Positive predictive value: 71.9
- Negative predictive value: 87.2
- Precision: 71.9
- Recall: 80.82

Below are evidences from Test Dataset:

- Accuracy: 80.41
- Sensitivity: 80.09
- Specificity: 80.62
- False positive rate: 19.38
- Positive predictive value: 72.96
- Negative predictive value: 86.11
- Precision: 72.96
- Recall: 80.09 **Evidence from ROC Curve**
- ROC curve area shows 0.8891 (89%),indicating a good predictive model.