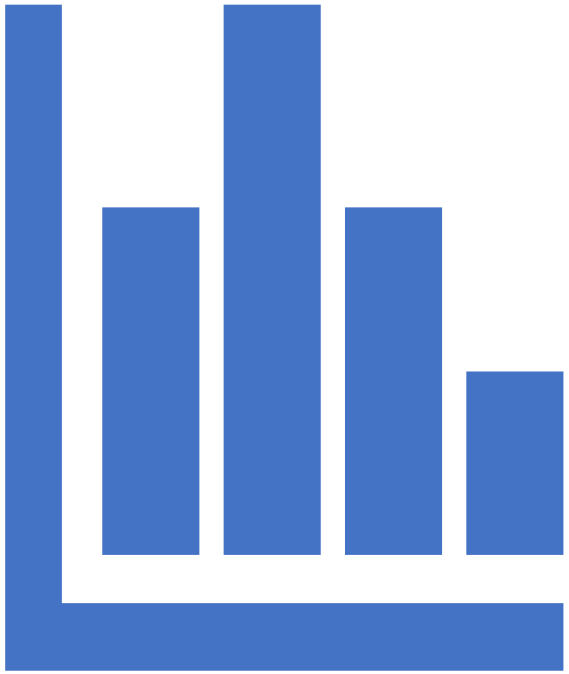


LEAD SCORING CASE STUDY

By:

Archit Srivastava

Rohith Shankar

OBJECTIVE:

This case study aims to help X Education solve their problems by –

- Building a logistic regression model to assign a lead score between 0 and 100 for each leads which can be used by the company to target hot leads to improve the conversion rates. Have a prediction accuracy of 80%.
- Suggest solutions to problems that are expected to arise in the future

DATA SET

Leads Data :

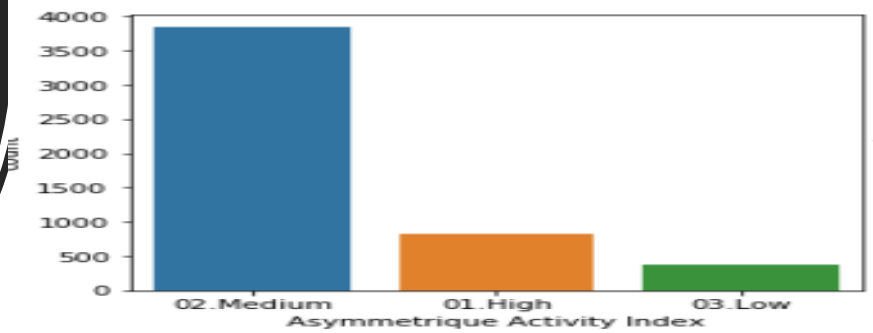
It contains information of the various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The ‘Converted’ column explains if lead was converted or not.

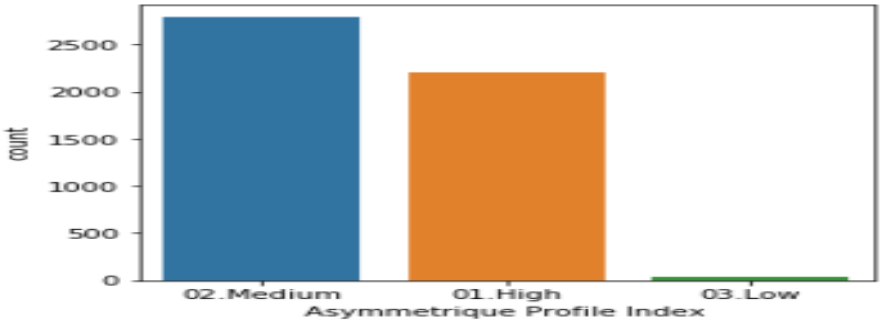
DATA MANIPULATION

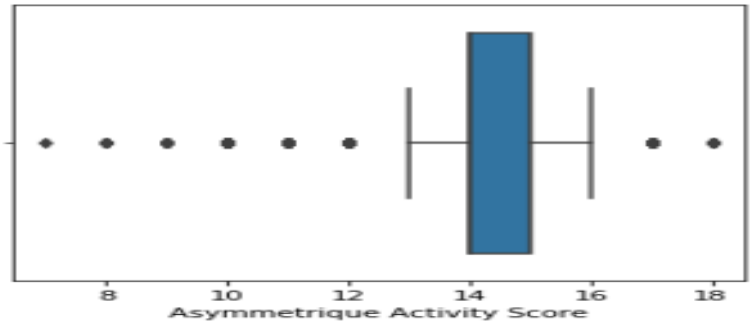
1. The Fields were the ”Select” has been populated, it is replaced with NaN, i.e. it is considered as a null value.
2. The variables don’t require data type transformation.
3. Columns with high Null values are dropped –
   - How did you hear about X Education
   - Lead Profile
4. Columns which have only one value hence will not make impact to analysis are dropped –
   - Magazine
   - Receive More Updates About Our Courses
   - Update me on Supply Chain Content
   - Get updates on DM Content
   - I agree to pay the amount through cheque

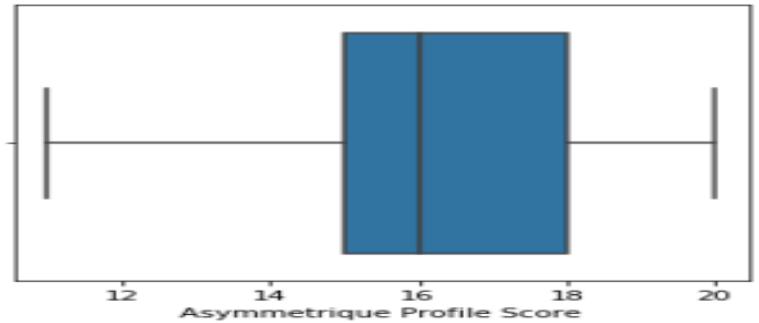
DATA MANIPULATION

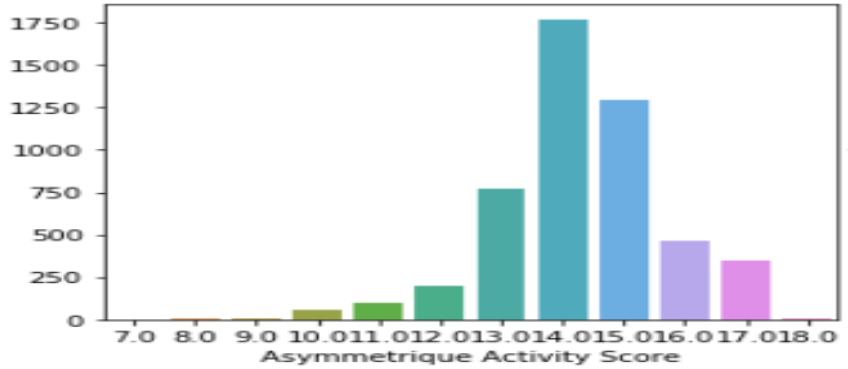
5. The 4 Asymmetrique columns are observed to have 45 % Null values and lot of variance and no pattern emerges. Hence these columns can be dropped.

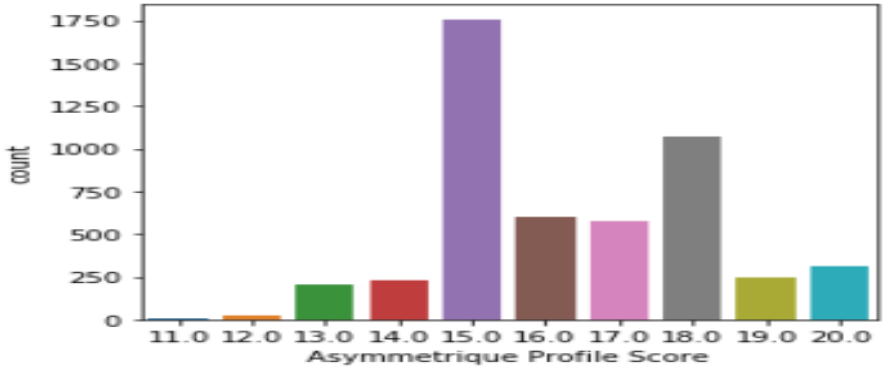










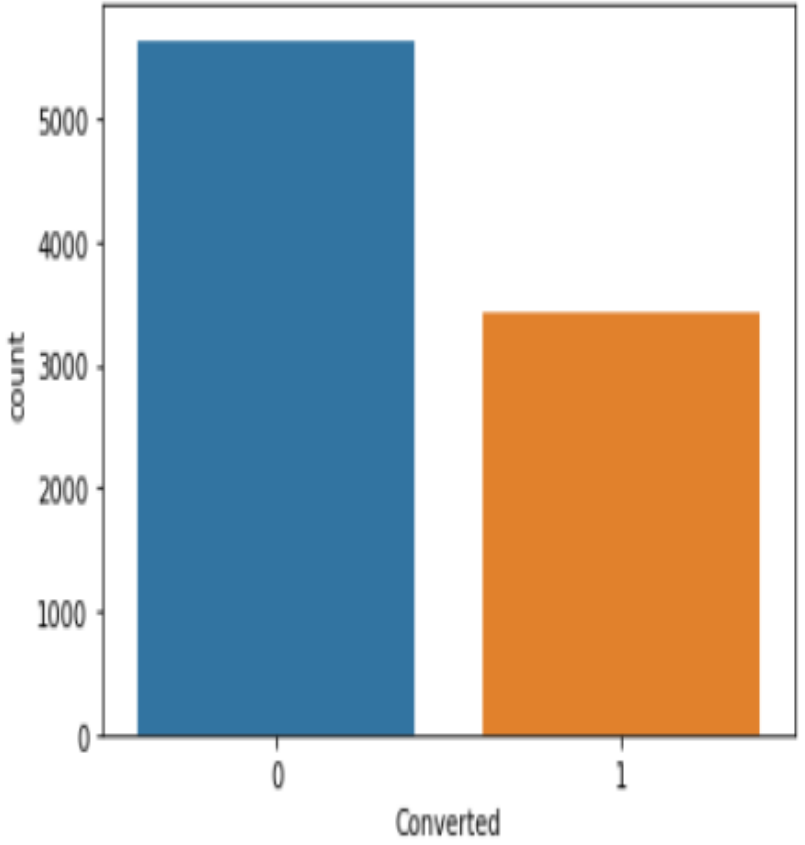


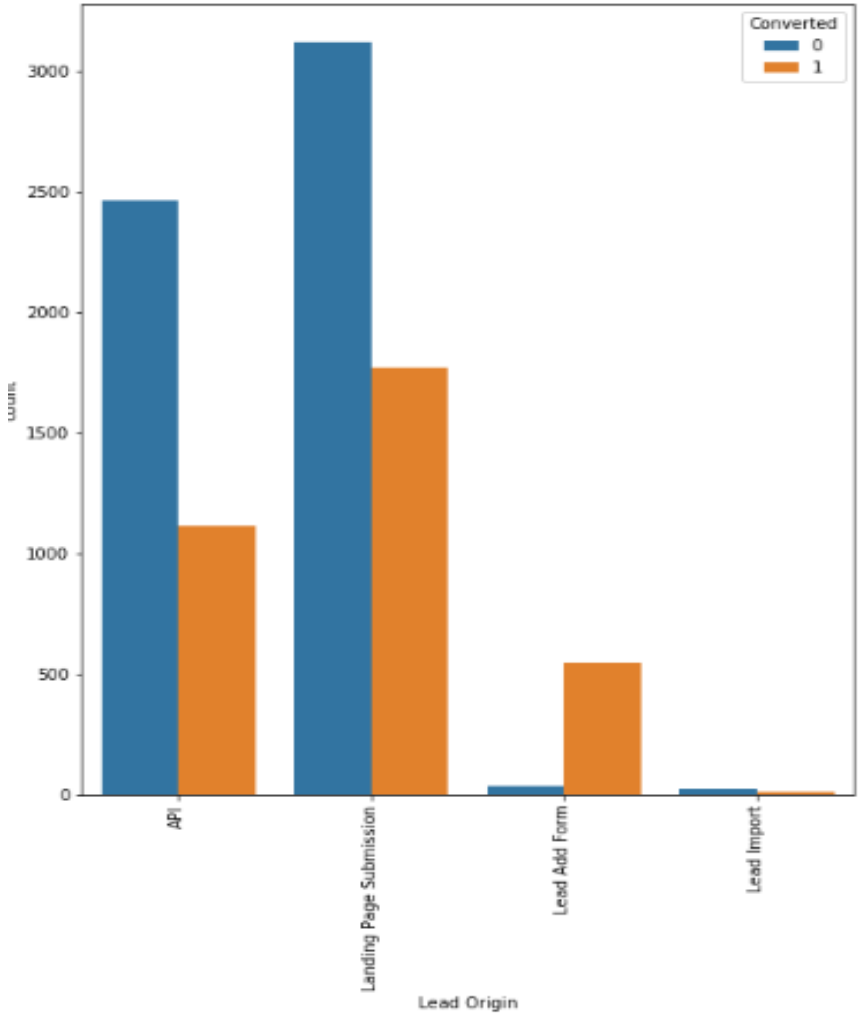
DATA MANIPULATION

6. Lead Quality Column replace the NaN with ‘Not Sure’ as both can be assumed to be the same.
7. City column the NaN value is replaced with Mumbai which is the highest value and imputing it won’t affect the analysis.
8. Tags column the fields where value hasn’t been selected we can assume it to be ‘Will revert after reading the email’.
9. What matters most to you in choosing a course column on plotting a graph shows to be nearly a single value hence dropping the variable.
10. What is your current occupation column the Null values can be replaced with Unemployed value as it has the highest percentage.
11. Specialization column Null values can be imputed with ‘Specialization_other’
12. Country column Null value imputed with value ‘India’.
13. Dropping rows where null value are present as the percentage of these are very low.

UNIVARIATE ANALYSIS

1. The ratio of conversion from plot indicates conversion to be lower.
2. Lead Conversion Univariate analysis shows conversion from various mediums. Lead Ad Form has greater conversion rates.

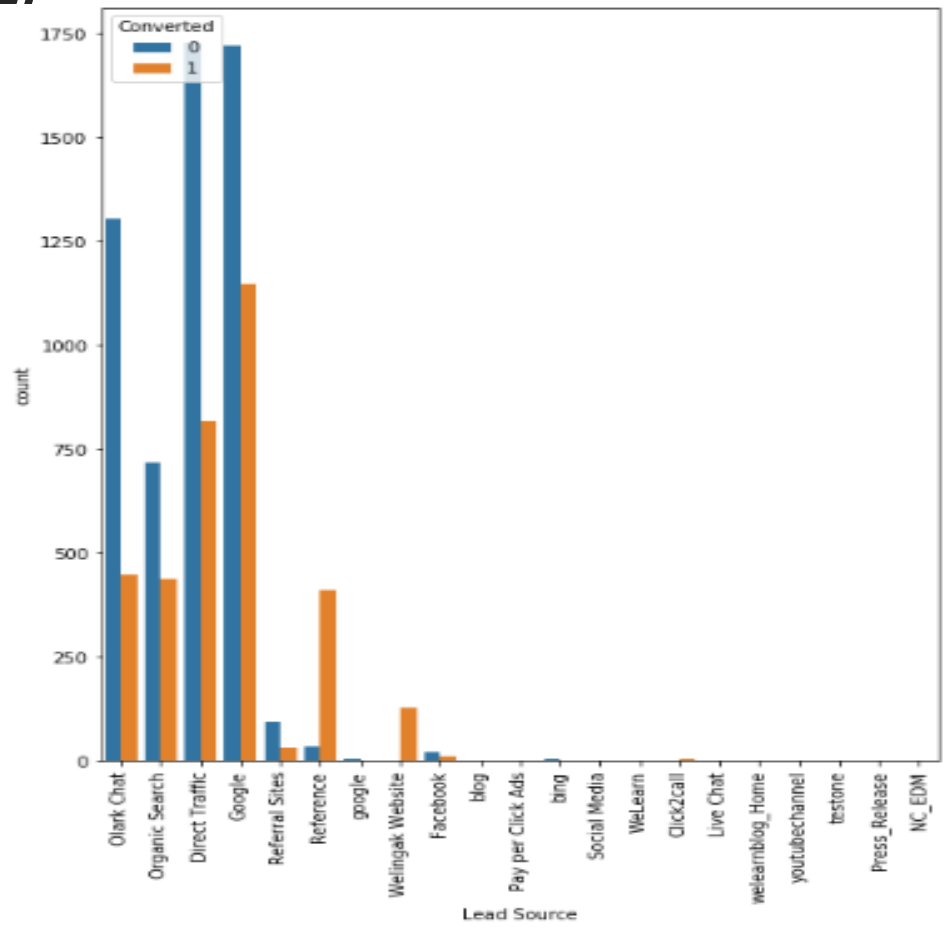


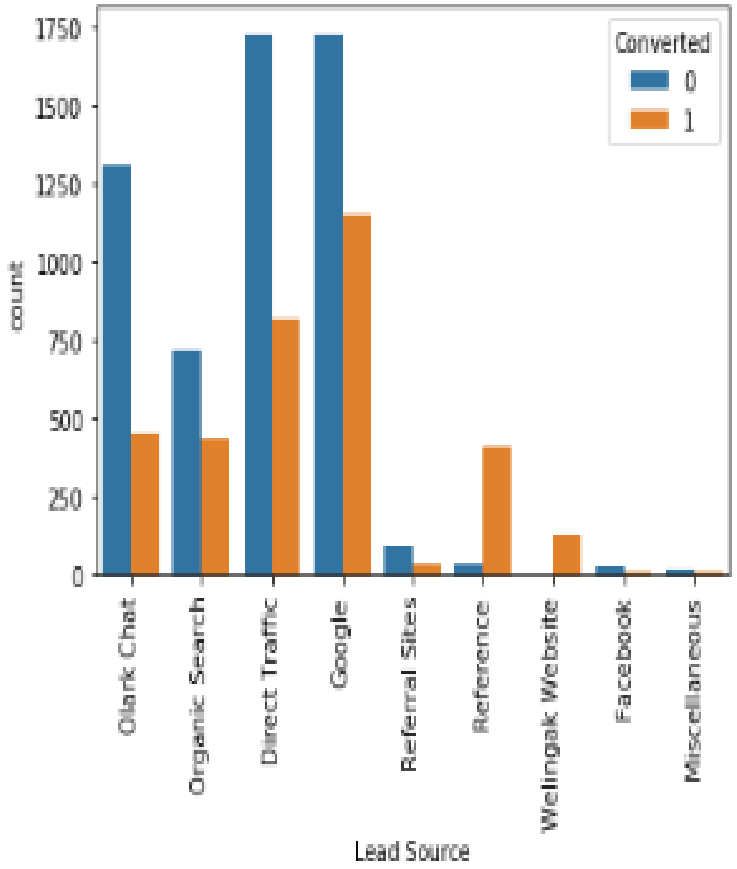


UNIVARIATE ANALYSIS

Lead Source Univariate analysis –
The first graph helps to identify best Lead Source
The second graph is a univariate analysis of best Lead Source.

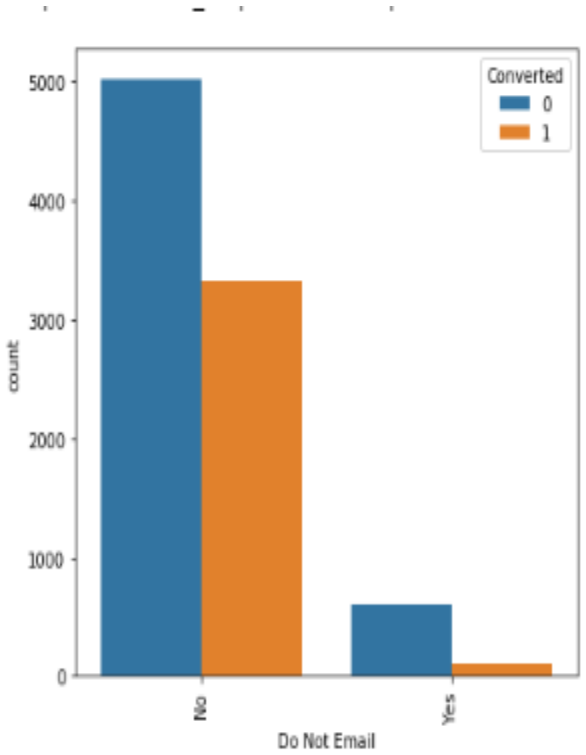
From this we can say that Google gives the highest conversion.

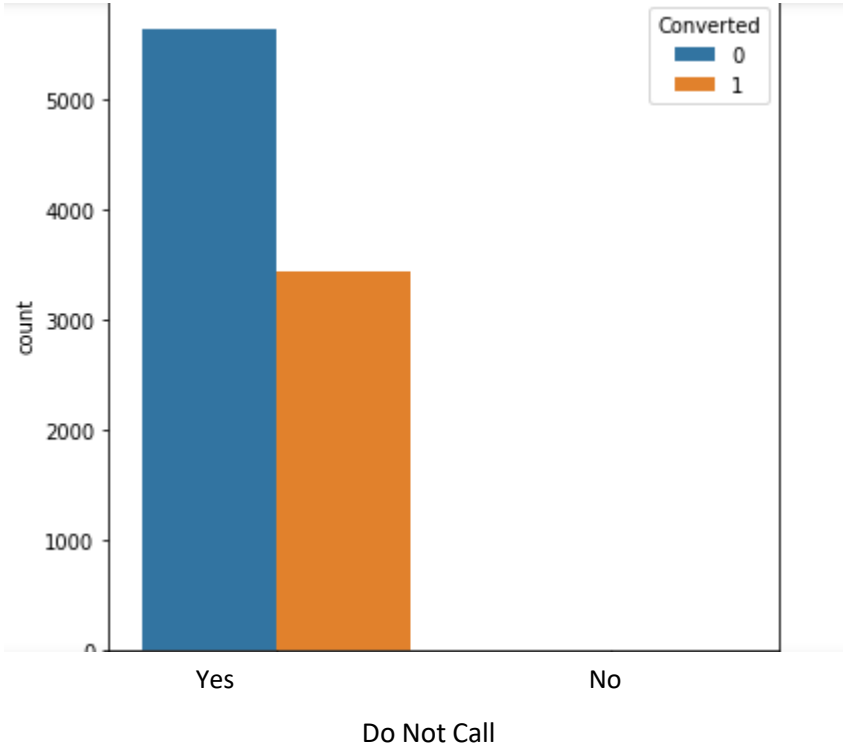




UNIVARIATE ANALYSIS

The univariate analysis of people who have requested not to be Called or Emailed shows –
People who request to be emailed convert higher
People who request no to be called have higher conversion.

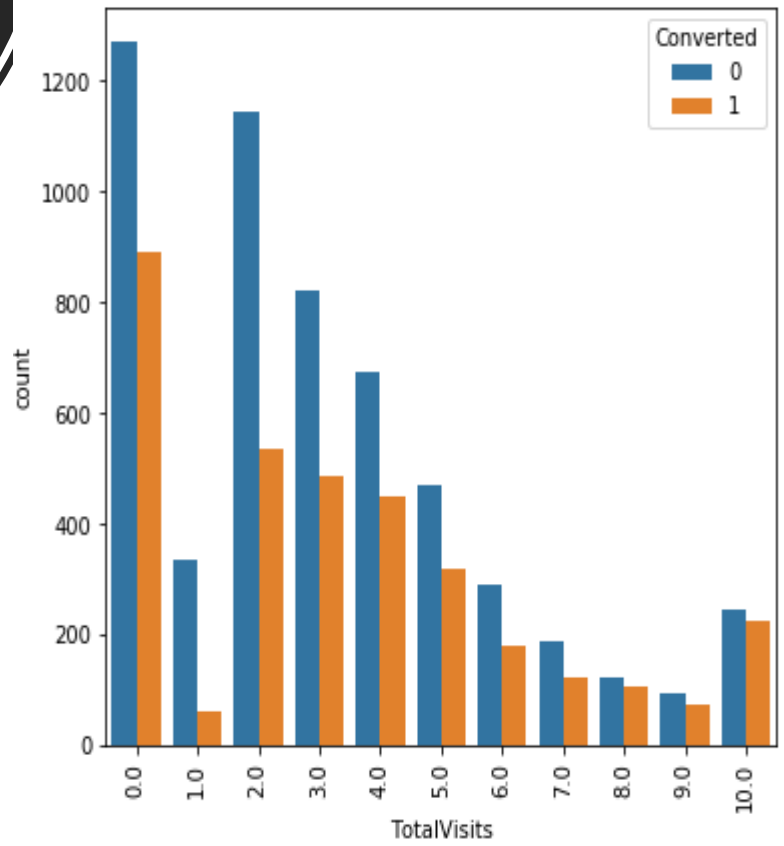


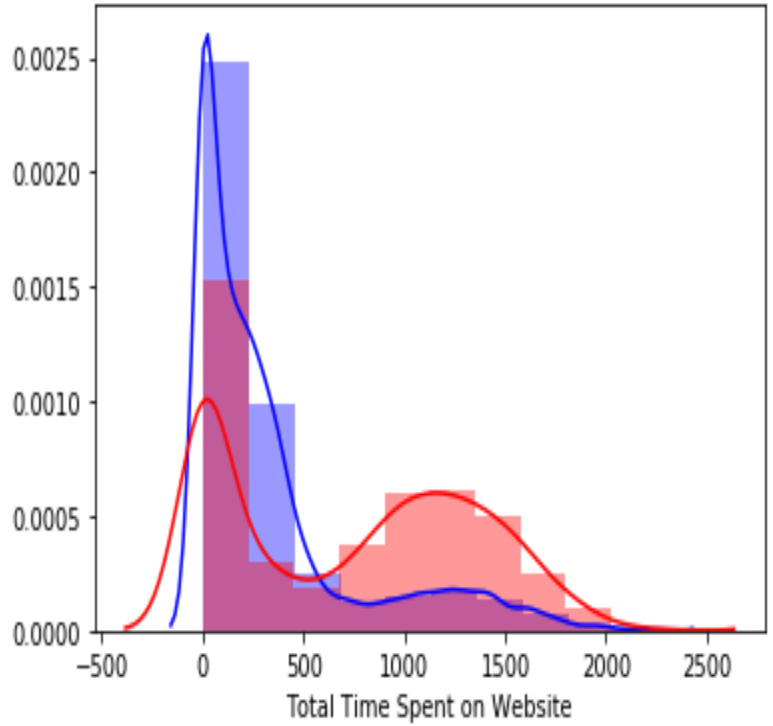


UNIVARIATE ANALYSIS

Total number of visits doesn't give any clear insights.

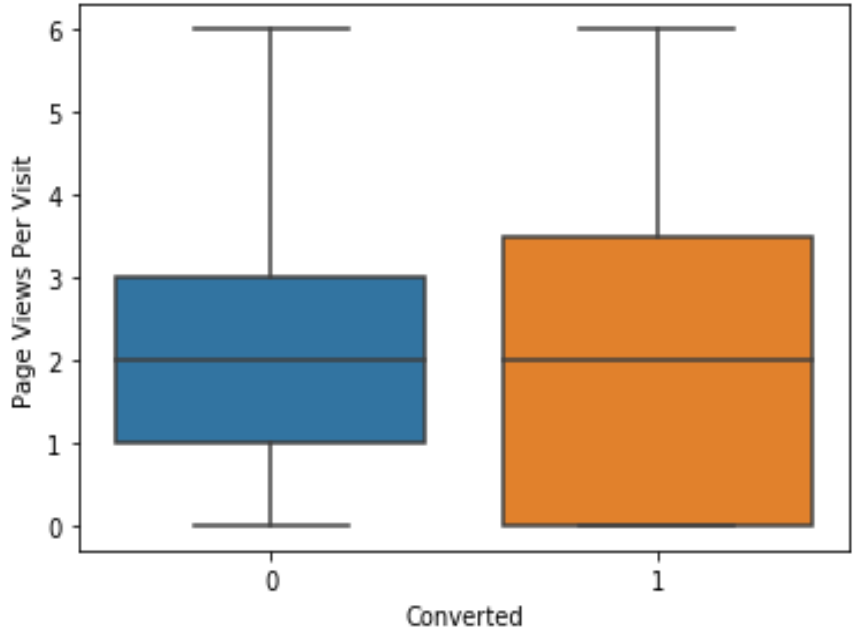
Total Time spent shows that people who spend more time tend to convert.

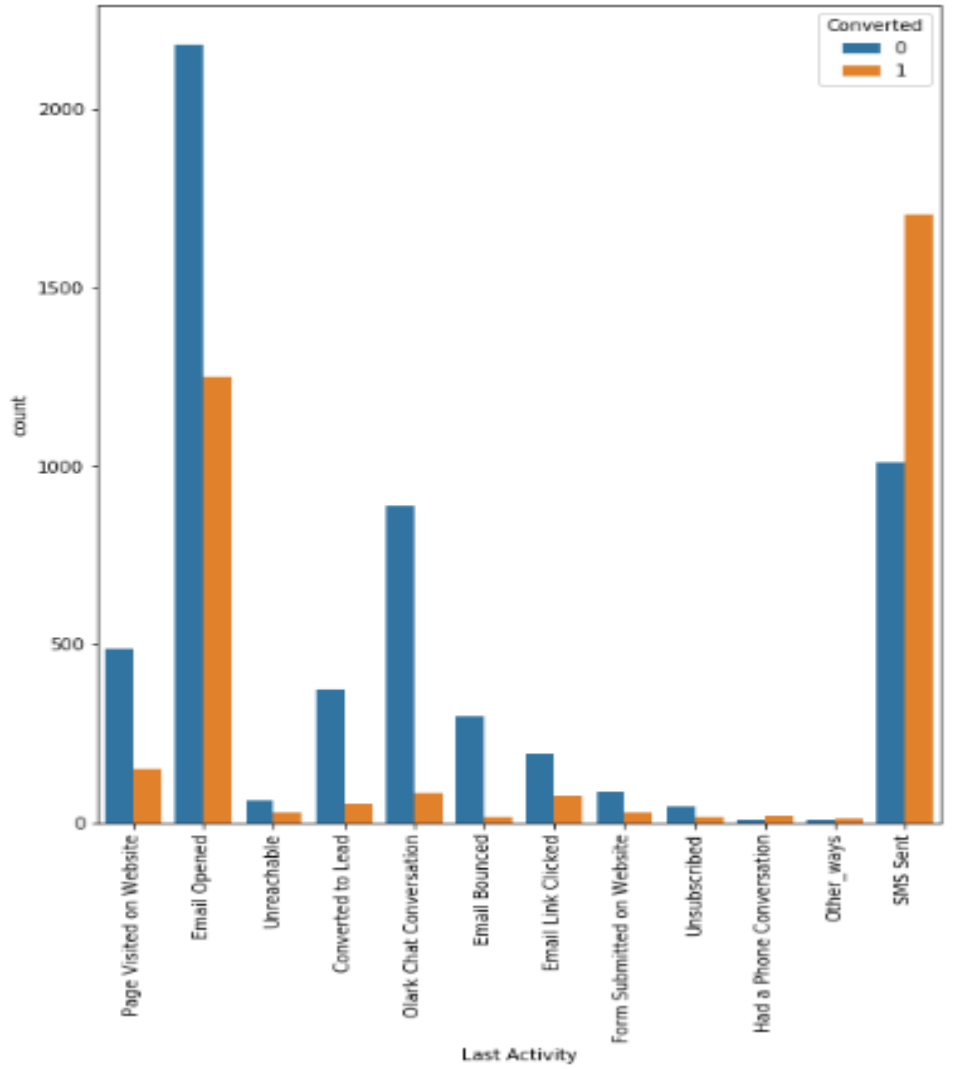




UNIVARIATE ANALYSIS

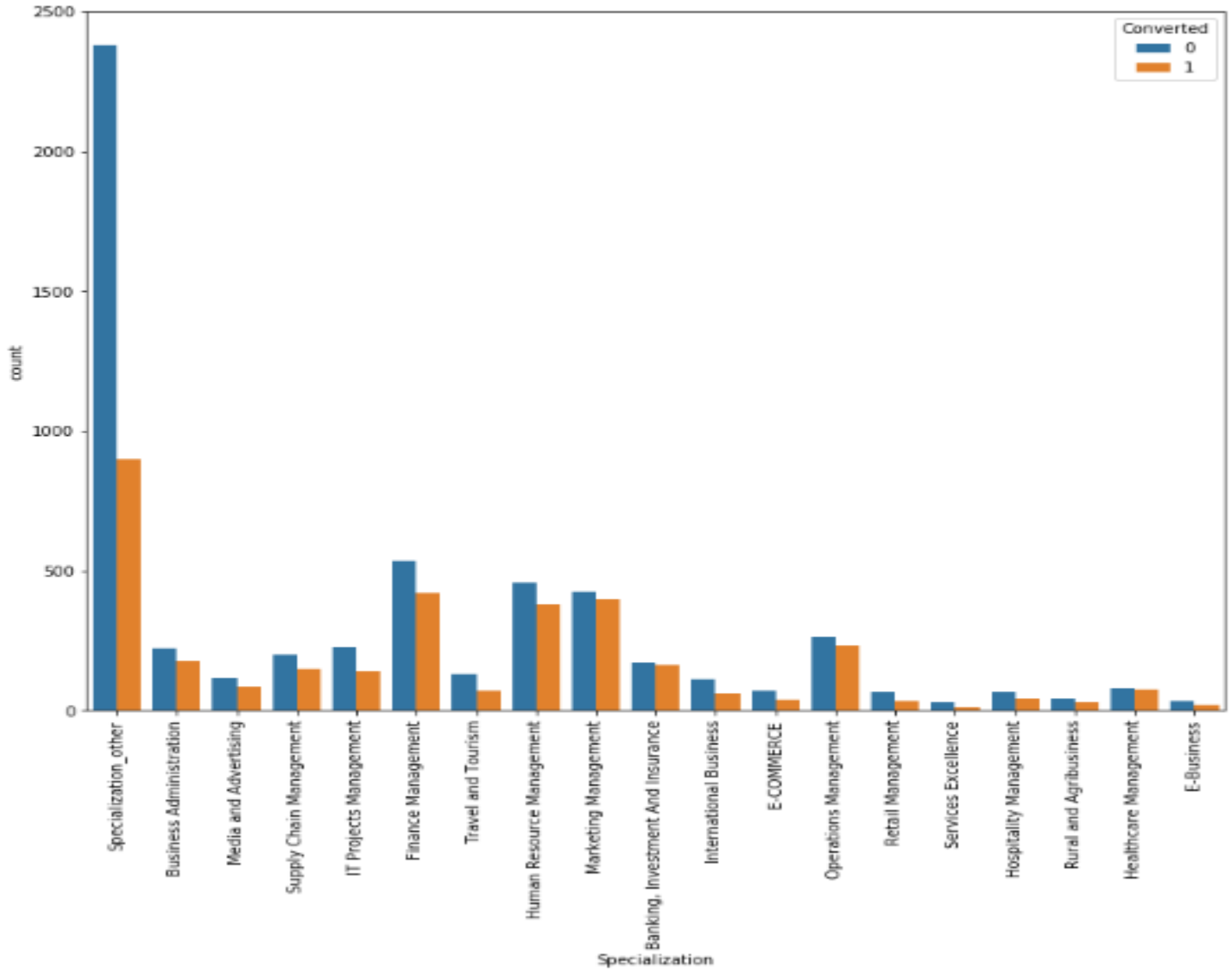
Pages viewed per visit doesn't affect conversion.
Last Activity which has highest conversion is SMS Sent





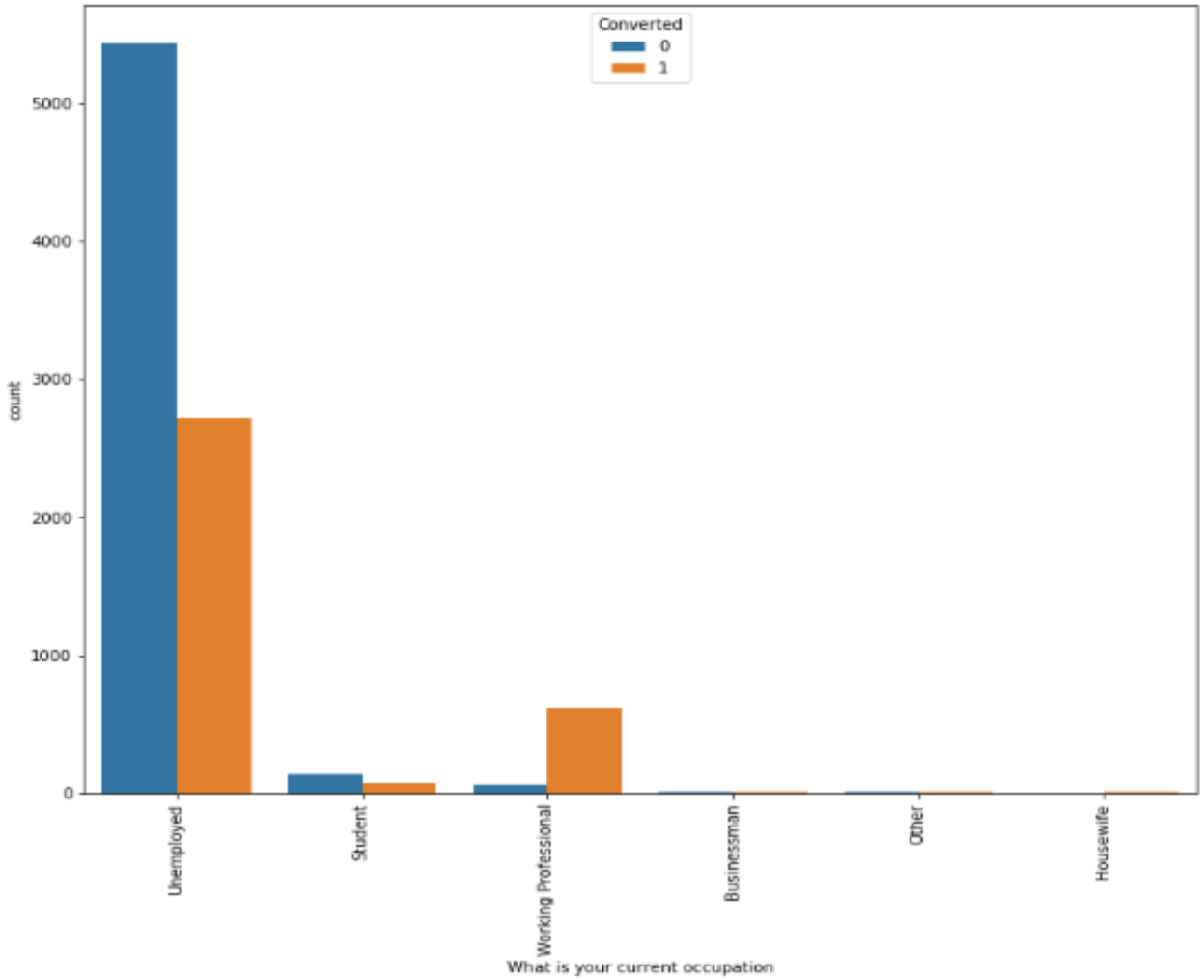
UNIVARIATE ANALYSIS

Specializations Finance ,Marketing and HRM are most converted



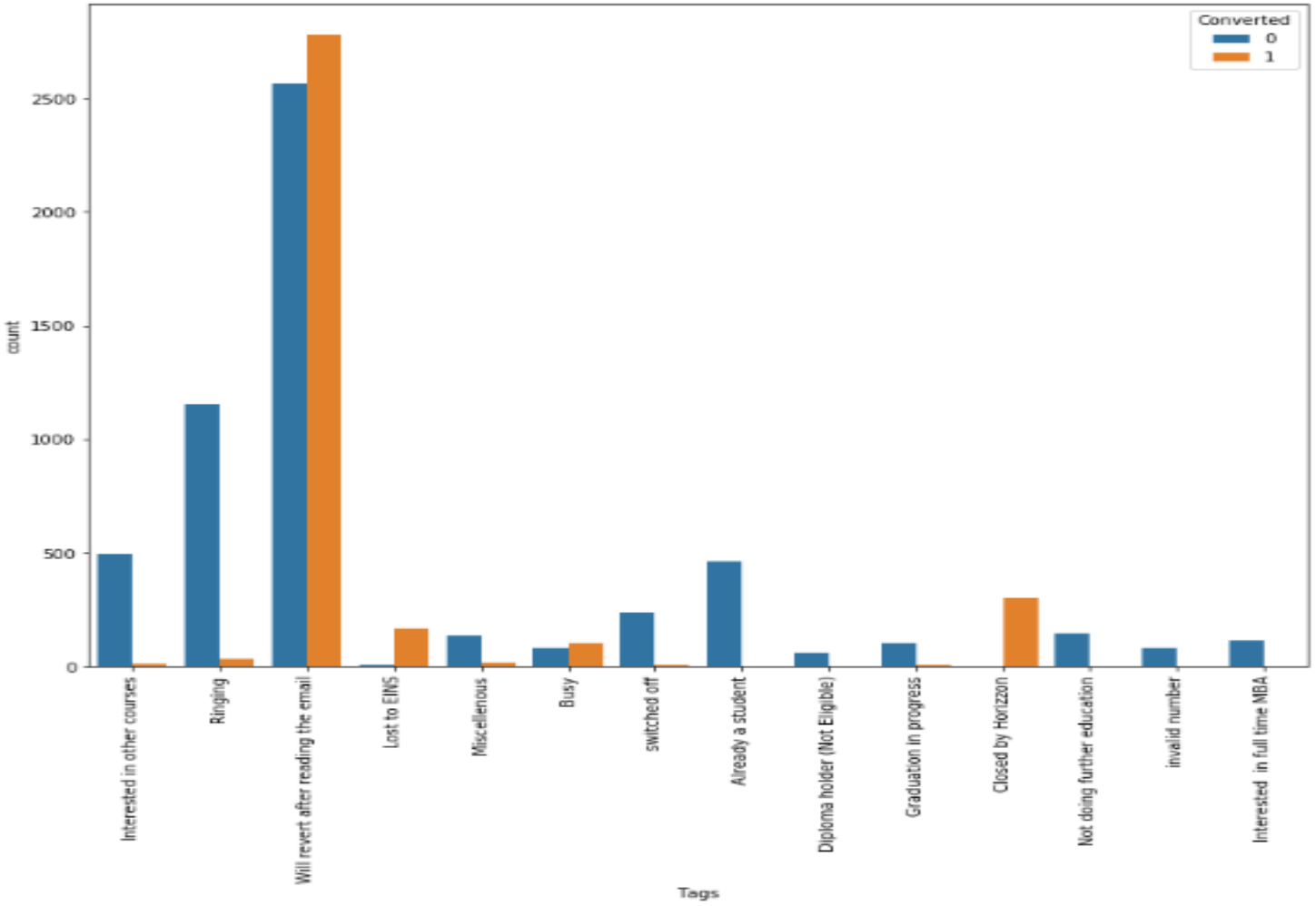
UNIVARIATE ANALYSIS

Employed professionals are most converted but Unemployed have the largest count hence chance of improvement in that sector.



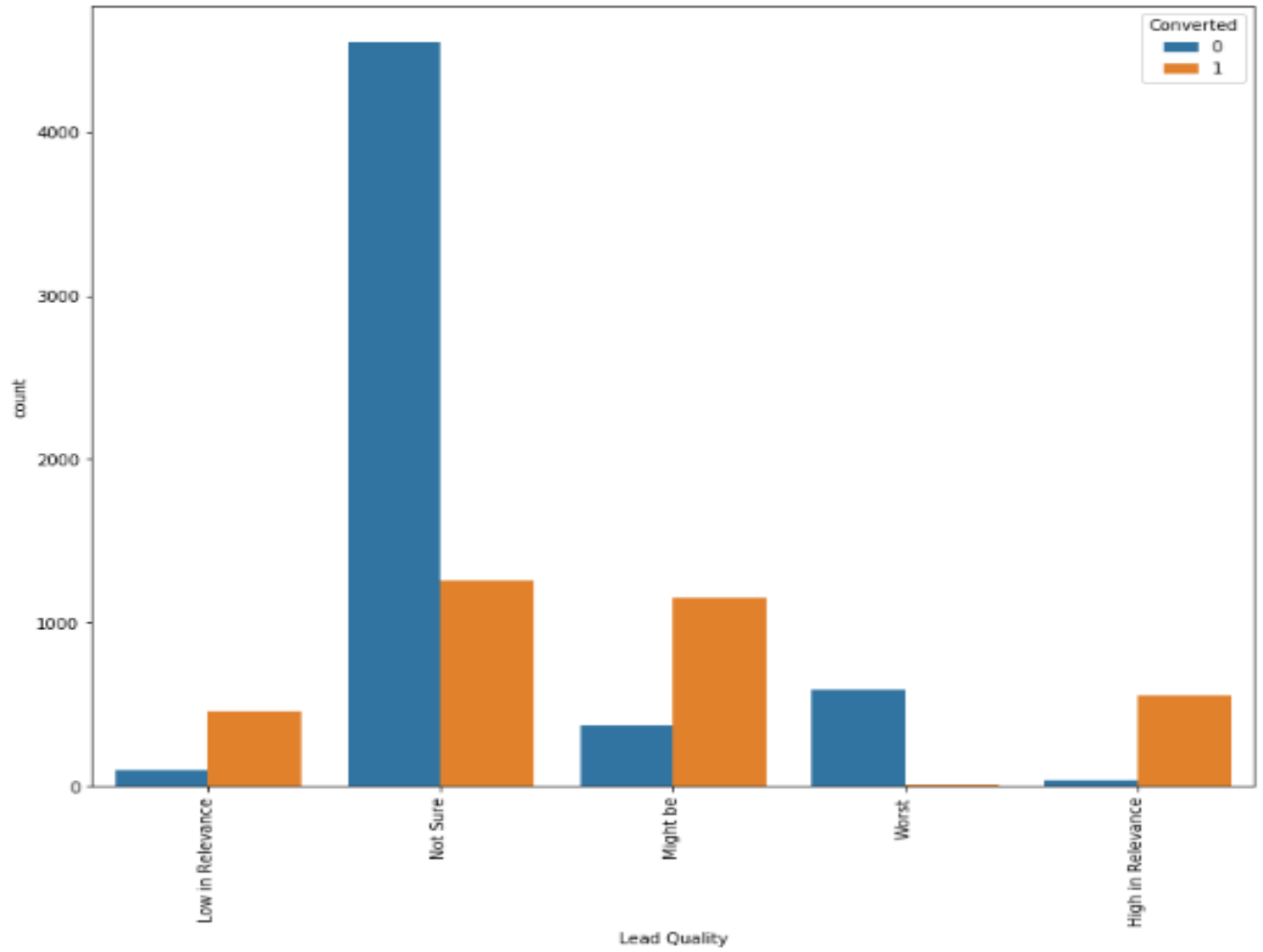
UNIVARIATE ANALYSIS

People who mention to revert after reading mail have highest conversion.



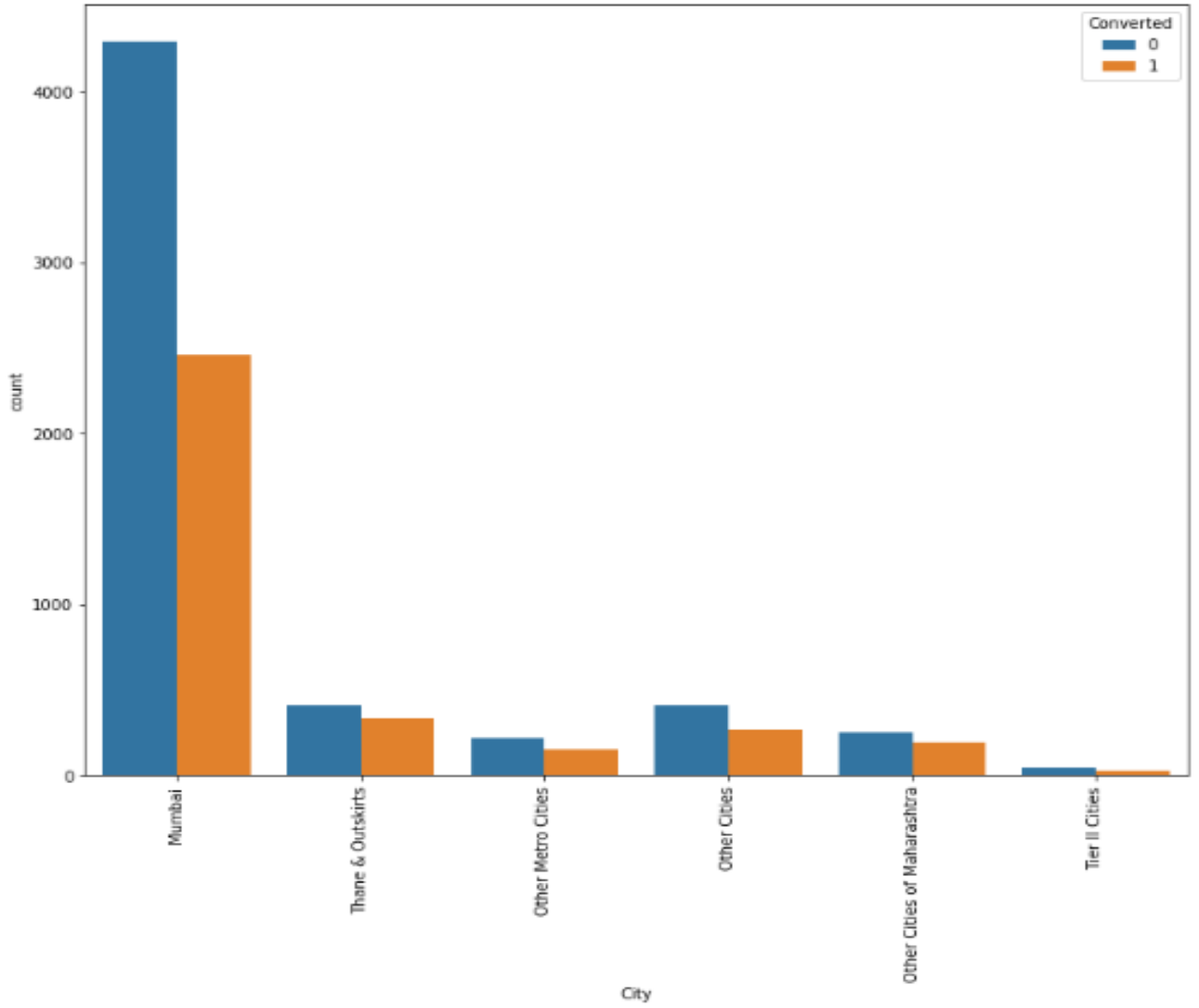
UNIVARIATE ANALYSIS

Lead Quality conversion is highest among Might Be and Not Sure



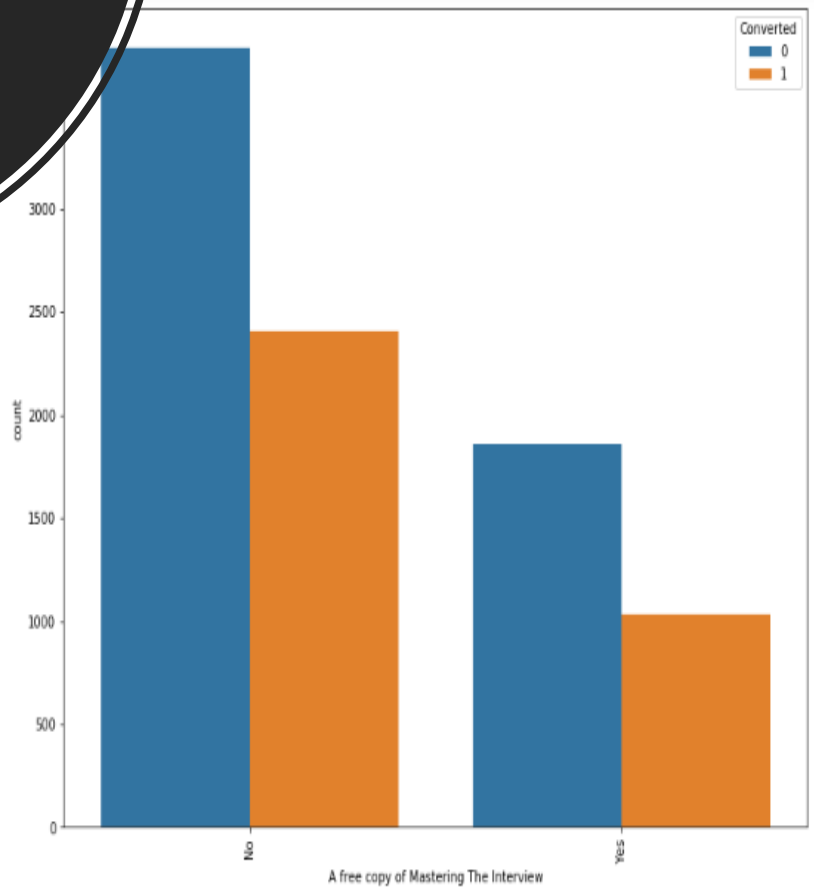
UNIVARIATE ANALYSIS

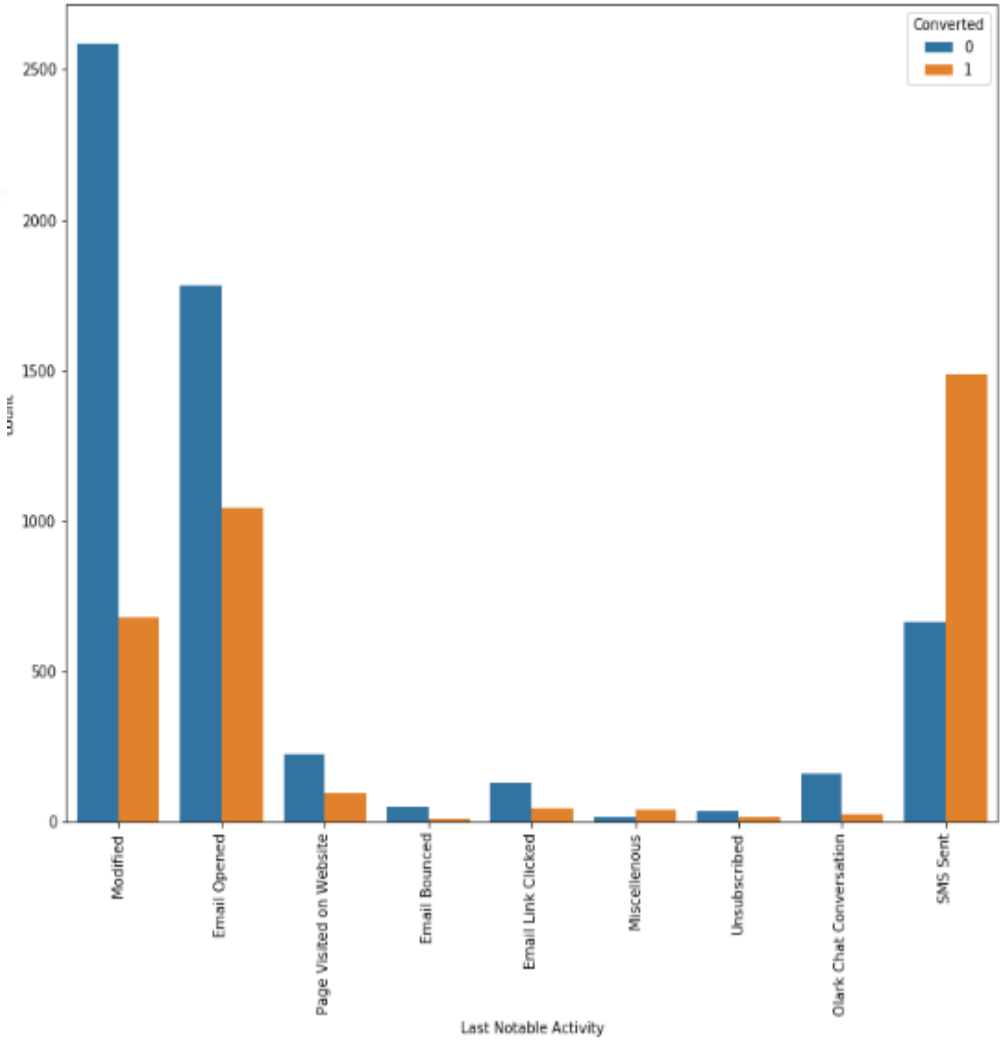
Mumbai has highest conversion.



UNIVARIATE ANALYSIS

Copy of Mastering Interview didn't seem to affect the conversion considerably.
Last notable Activity where conversion rate is high is for SMS sent.





MODEL SELECTION

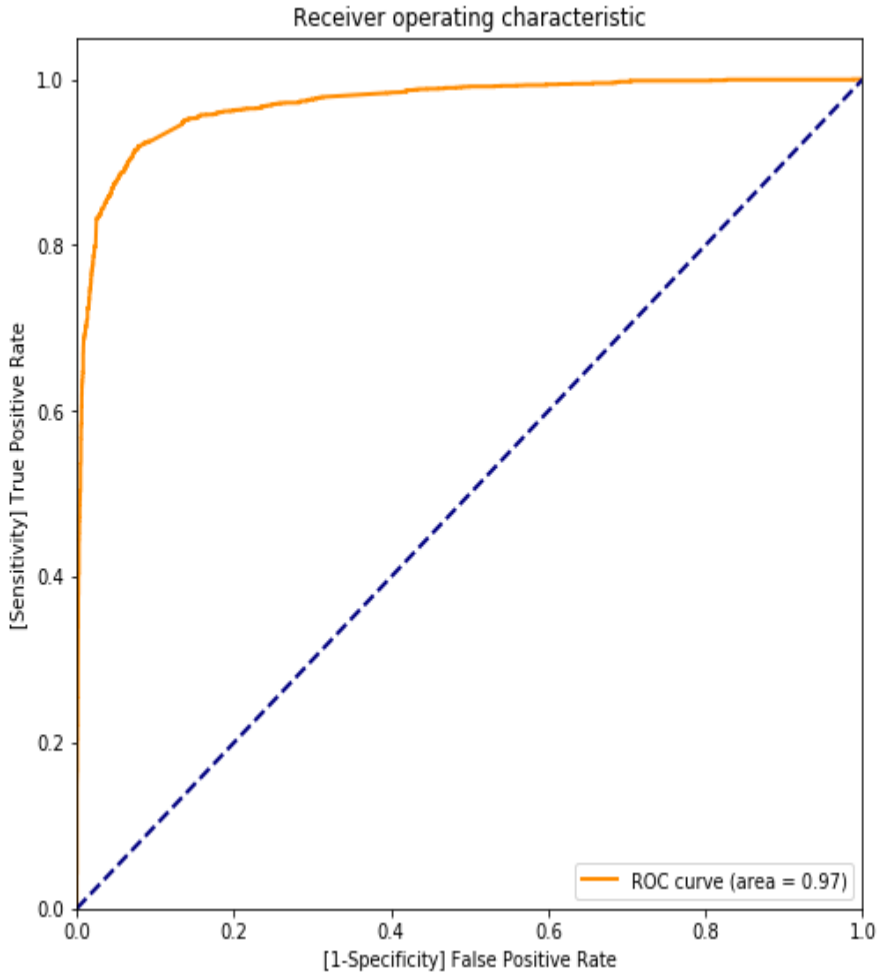
Select the best fit model in which all the variables have low P value and VIF value less than 5.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8018	0.237	-7.606	0.000	-2.266	-1.337
Do Not Email	-1.2106	0.243	-4.972	0.000	-1.688	-0.733
Total Time Spent on Website	1.1500	0.061	18.939	0.000	1.031	1.269
Lead Origin_Lead Add Form	2.1776	0.370	5.888	0.000	1.453	2.902
Lead Origin_Lead Import	1.4771	0.746	1.980	0.048	0.015	2.939
Lead Source_Olark Chat	1.2391	0.140	8.879	0.000	0.966	1.513
Lead Source_Welingak Website	3.2383	0.817	3.966	0.000	1.638	4.839
Last Activity_SMS Sent	1.8756	0.112	16.719	0.000	1.656	2.095
What is your current occupation_Working Professional	1.3746	0.328	4.195	0.000	0.732	2.017
Tags_Busy	3.4104	0.337	10.130	0.000	2.751	4.070
Tags_Closed by Horizzon	8.5041	0.807	10.537	0.000	6.922	10.086
Tags_Lost to EINS	9.2013	0.777	11.848	0.000	7.679	10.723
Tags_Ringing	-1.6163	0.345	-4.679	0.000	-2.293	-0.939
Tags_Will revert after reading the email	3.6903	0.241	15.336	0.000	3.219	4.162
Tags_switched off	-2.3103	0.628	-3.676	0.000	-3.542	-1.079
Lead Quality_Not Sure	-3.1460	0.142	-22.132	0.000	-3.425	-2.867
Lead Quality_Worst	-3.9338	0.852	-4.619	0.000	-5.603	-2.265
Last Notable Activity_Modified	-1.6528	0.116	-14.254	0.000	-1.880	-1.426
Last Notable Activity_Olark Chat Conversation	-1.7924	0.407	-4.408	0.000	-2.589	-0.995

MODEL SELECTION

The ROC curve of the model selected and the VIF values of the variables

	Features	VIF
14	Lead Quality_Not Sure	3.28
12	Tags_Will revert after reading the email	3.19
2	Lead Origin_Lead Add Form	1.82
16	Last Notable Activity_Modified	1.70
6	Last Activity_SMS Sent	1.65
4	Lead Source_Olark Chat	1.64
11	Tags_Ringing	1.53
1	Total Time Spent on Website	1.42
5	Lead Source_Welingak Website	1.37
7	What is your current occupation_Working Profes...	1.26
9	Tags_Closed by Horizzon	1.22
15	Lead Quality_Worst	1.13
0	Do Not Email	1.13
8	Tags_Busy	1.12
13	Tags_switched off	1.10
10	Tags_Lost to EINS	1.09
17	Last Notable Activity_Olark Chat Conversation	1.08
3	Lead Origin_Lead Import	1.01



CONCLUSION

The Variables of Optimum Model

1	Do Not Email
2	Total Time Spent on Website
3	Lead Origin_Lead Add Form
4	Lead Origin_Lead Import
5	Lead Source_Olark Chat
6	Lead Source_Welingak Website
7	Last Activity_SMS Sent
8	What is your current occupation_Working Professional
9	Tags_Busy
10	Tags_Closed by Horizzon
11	Tags_Lost to EINS
12	Tags_Ringing
13	Tags_Will revert after reading the email
14	Tags_switched off
15	Lead Quality_Not Sure
16	Lead Quality_Worst
17	Last Notable Activity_Modified
18	Last Notable Activity_Olark Chat Conversation

CONCLUSION

The Model is able to predict with an accuracy of 84% which was the requirement.
The Model suggests to look for leads whose last activity was SMS, leads who are working professional, who mention to revert after reading the mail. Leads who request not to be e-mailed, Switched off has a negative conversion.