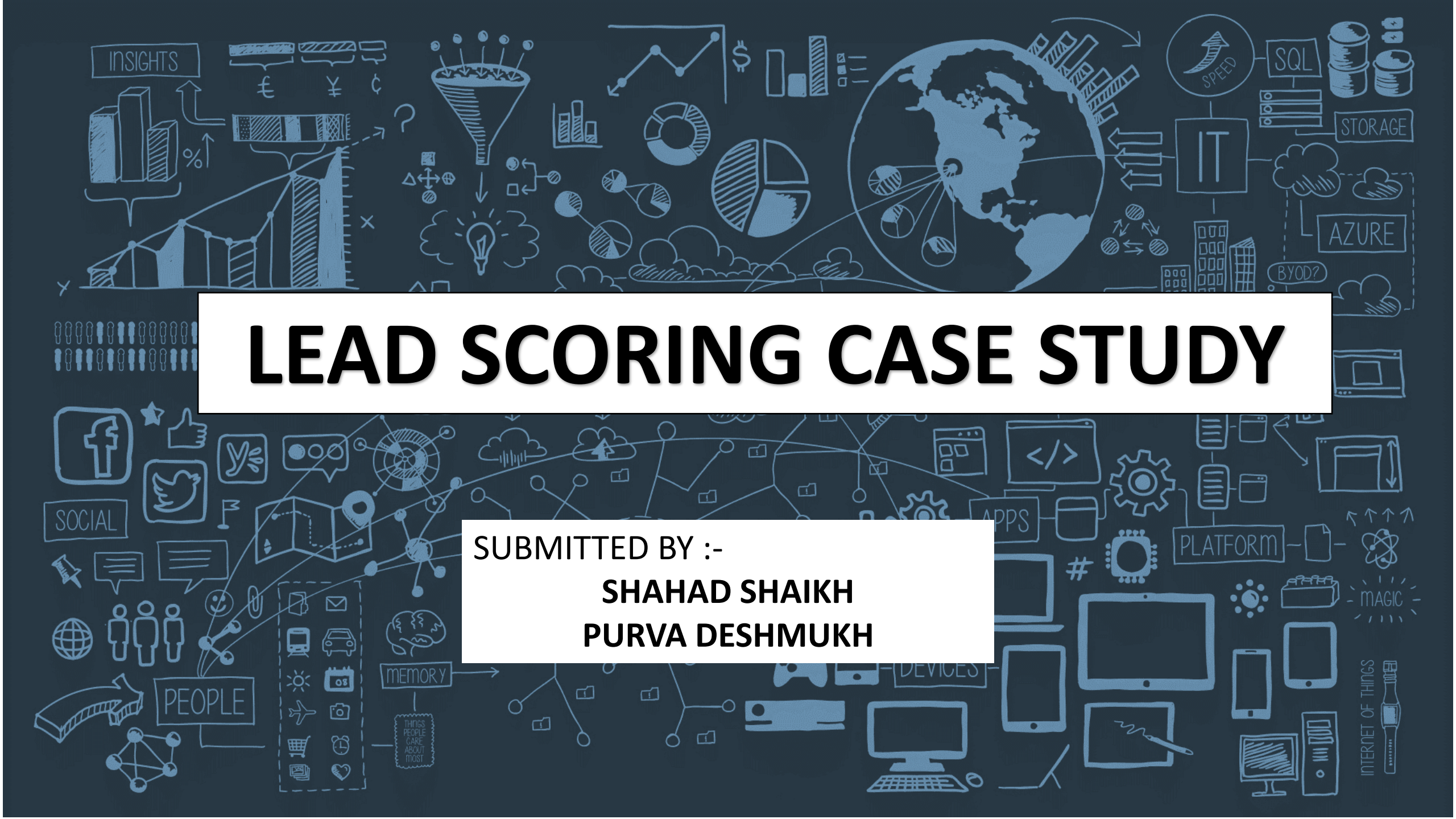


LEAD SCORING CASE STUDY

SUBMITTED BY :-

SHAHAD SHAIKH

PURVA DESHMUKH

PROBLEM STATEMENT

An education company named X Education sells online courses 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. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if say, they acquire 100 leads in a day, only about 30 of them are converted.

- X Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- To get the solution we have to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

ANALYSIS APPROACH

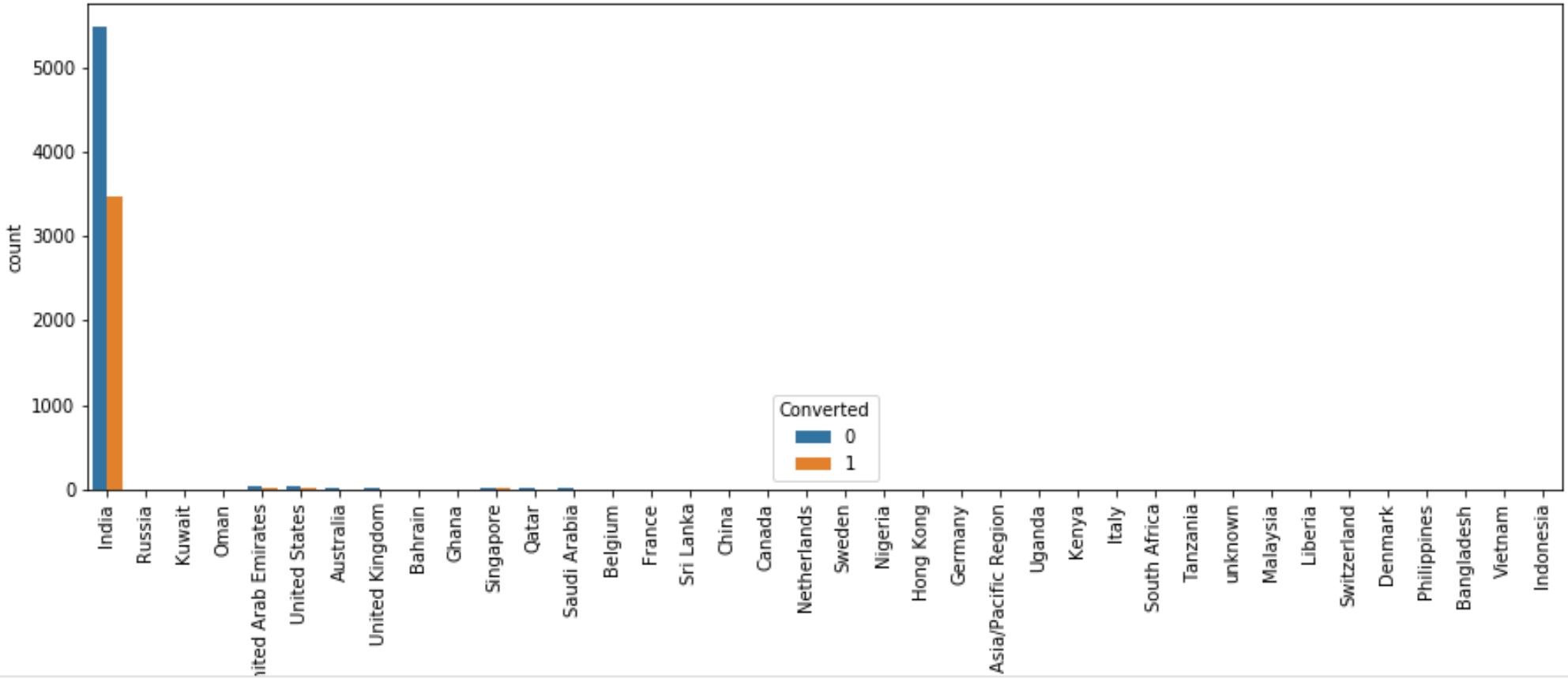
1. DATA CLEANING AND TREATMENT

- The select level present in many categorical variables where handled.
- Columns with missing values more than 45% missing values were dropped.
- Prospect ID & Lead Number are two variables that are just indicative of the ID number of the Contacted People & hence were dropped.

2. CATEGORICAL ATTRIBUTE ANALYSIS

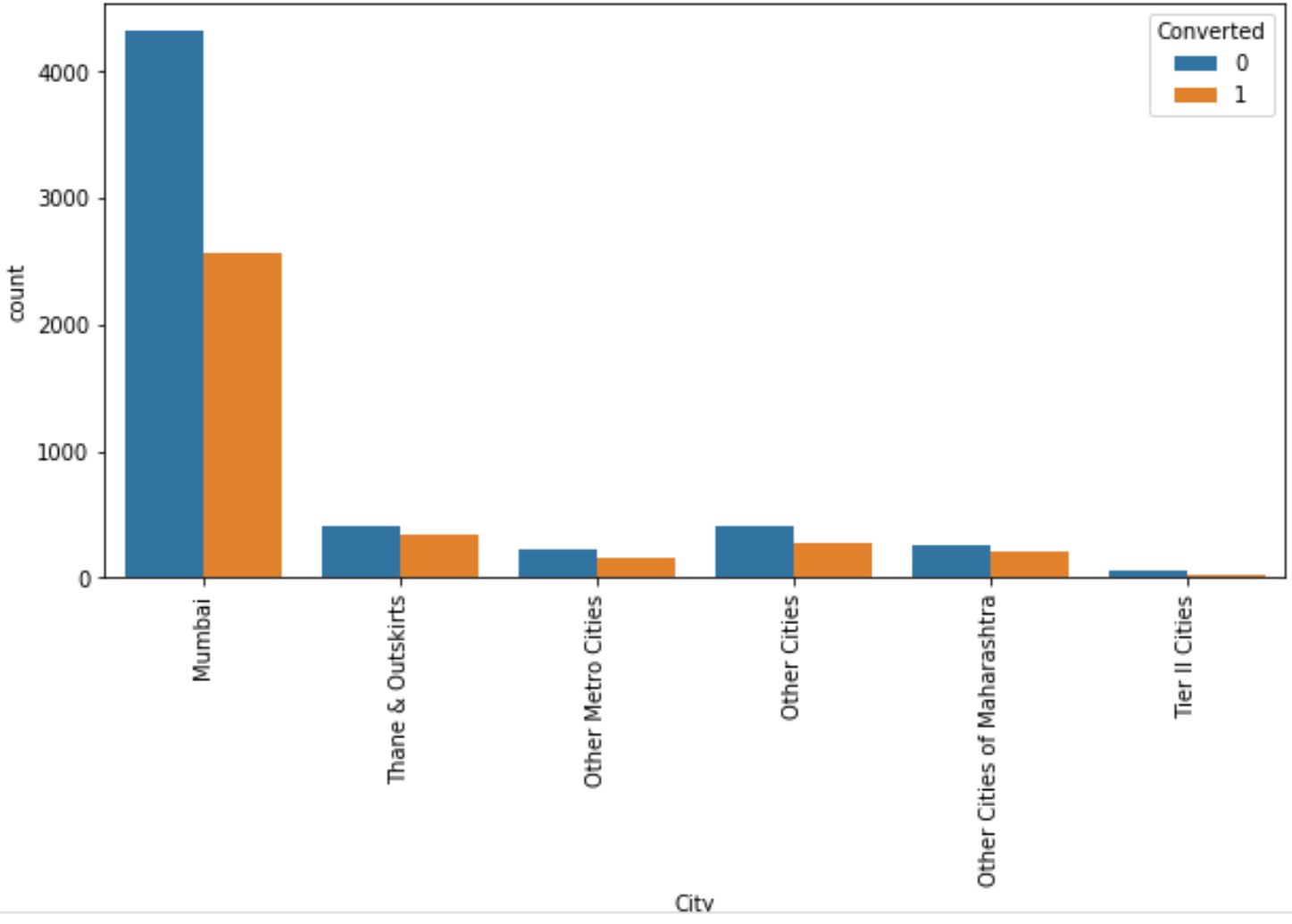
2.1. Plot of Country wise distribution of converted and non-converted leads.

Here, most common occurence among the non-missing values was “India” hence we imputed the missing values with “India”. After Imputation we can see the Number of Values for India are quite high (nearly 97% of the Data), this column can be dropped as it has highly skewed data



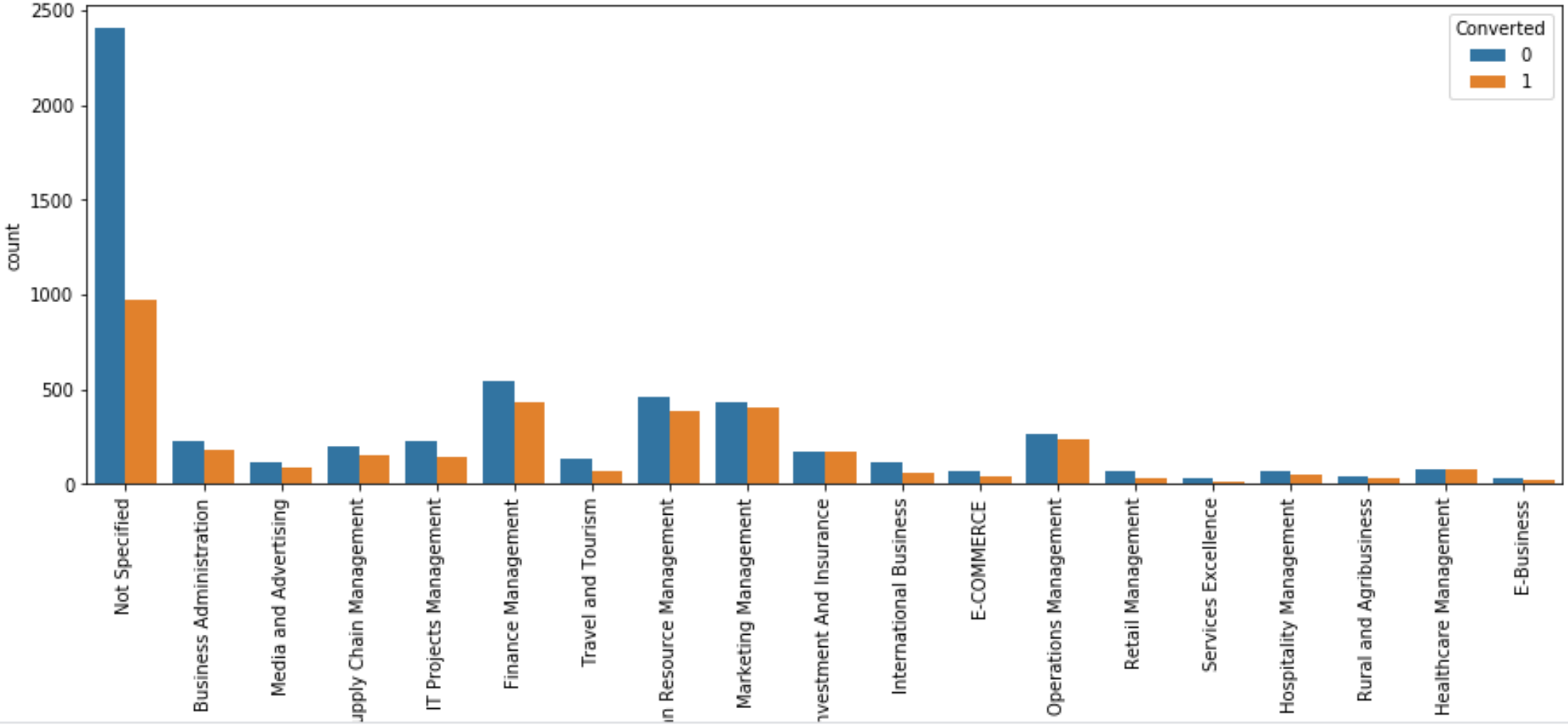
2.2. Plotting City wise distribution of Converted and Non-converted Leads

Here we can see Mumbai City has Maximum number of Converted and non-converted Leads.

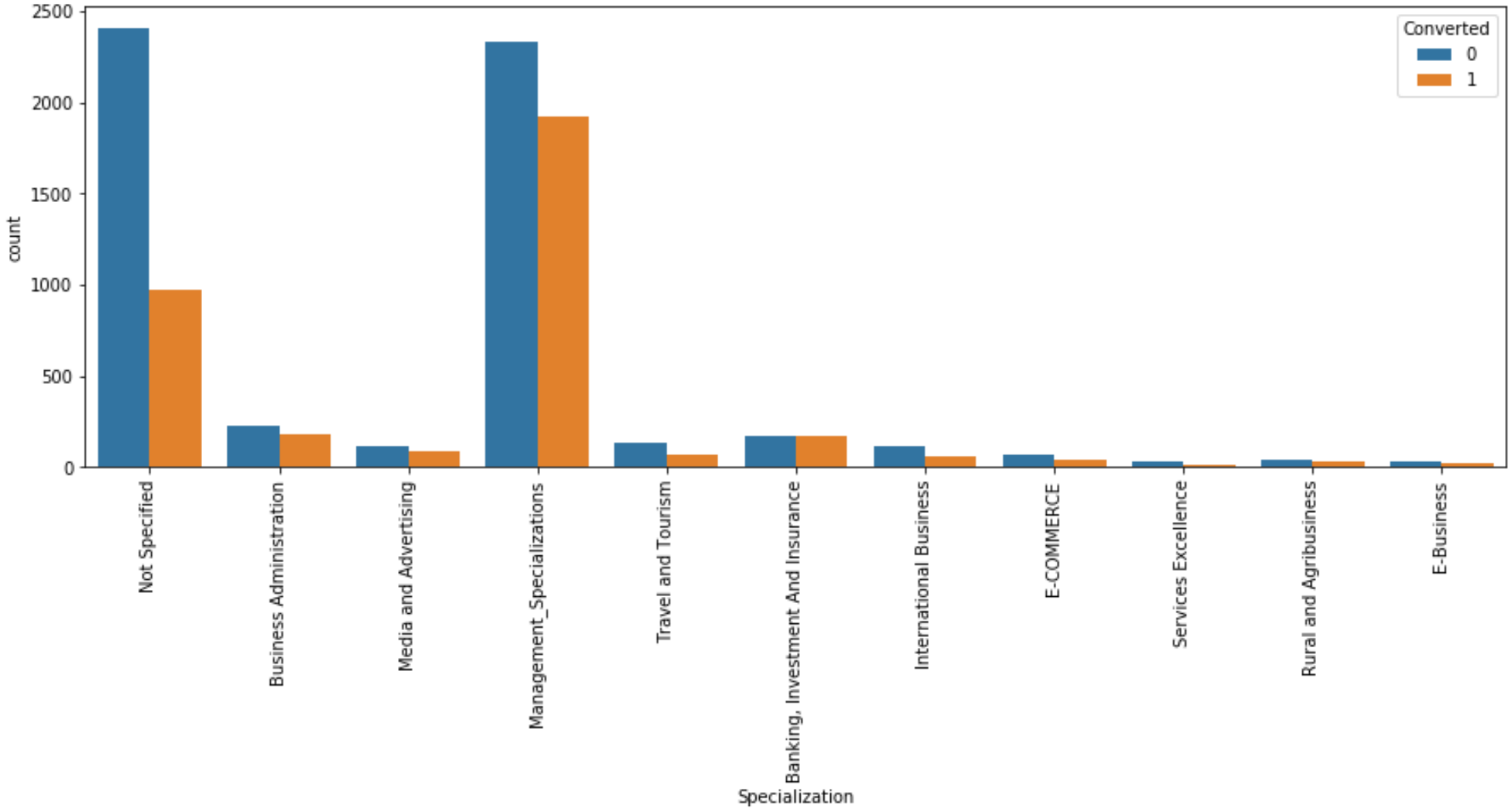


2.3. Plotting Specialization wise distribution of Converted and Non-converted Leads

As we can see that maximum number of leads have not specified their Specialization, maybe because they might not have filled it they are students.

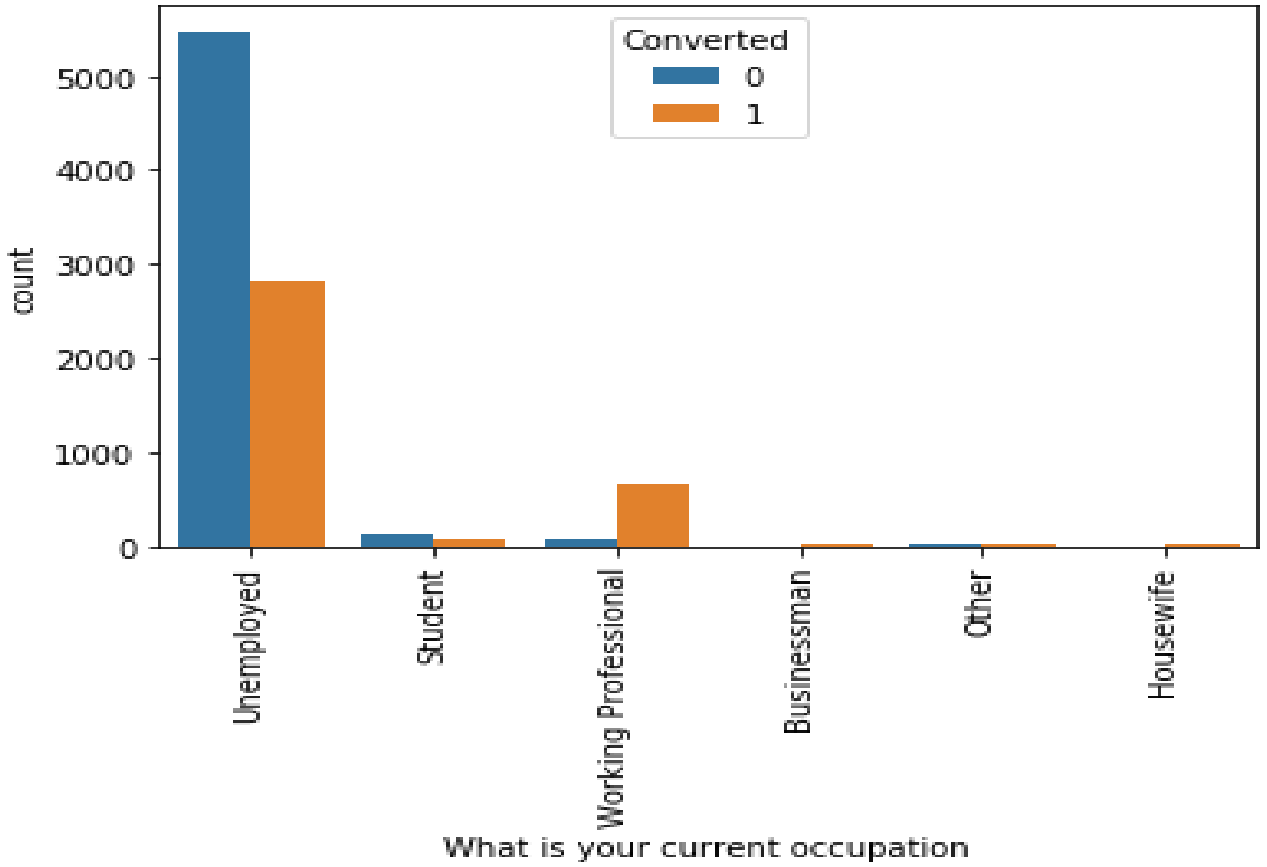


Leads having specialization in management have higher number of converted and non-converted Leads, Hence Management related courses should be given high importance.

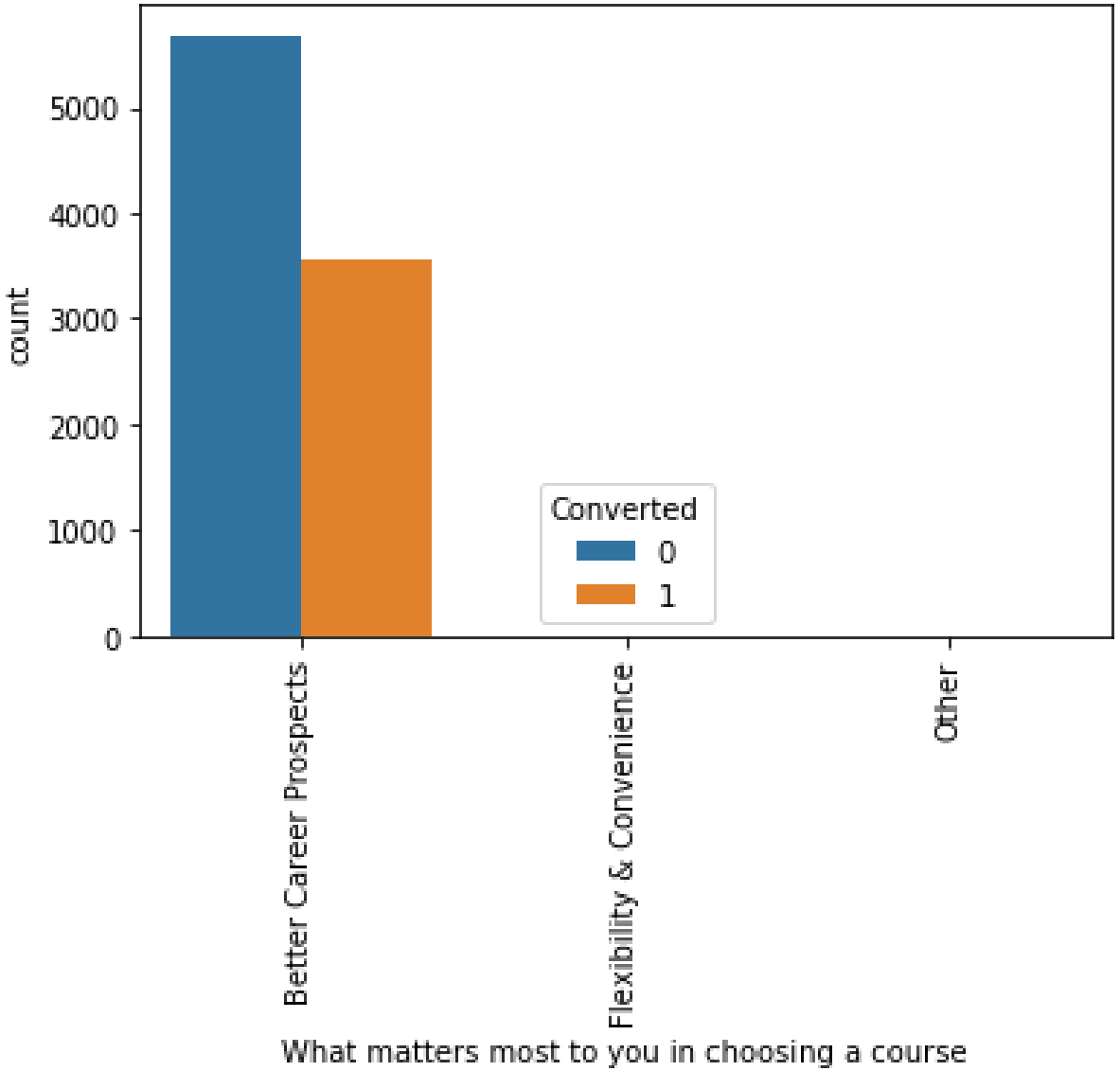


2.4. Occupation wise Distribution of Converted and Non-Converted Leads.

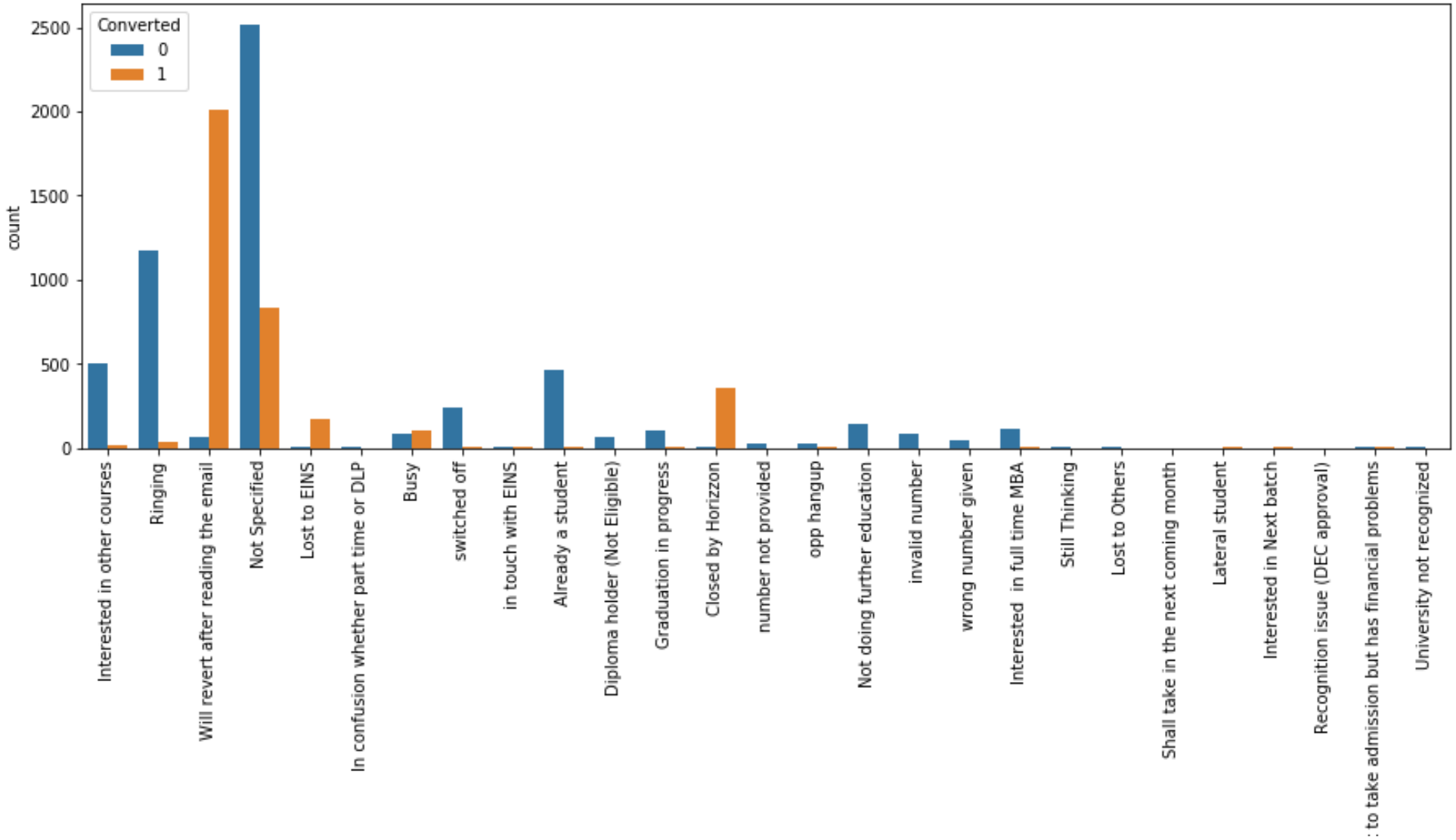
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.



2.5. Most of the Converted and non-converted leads choose the course based on better Career prospects

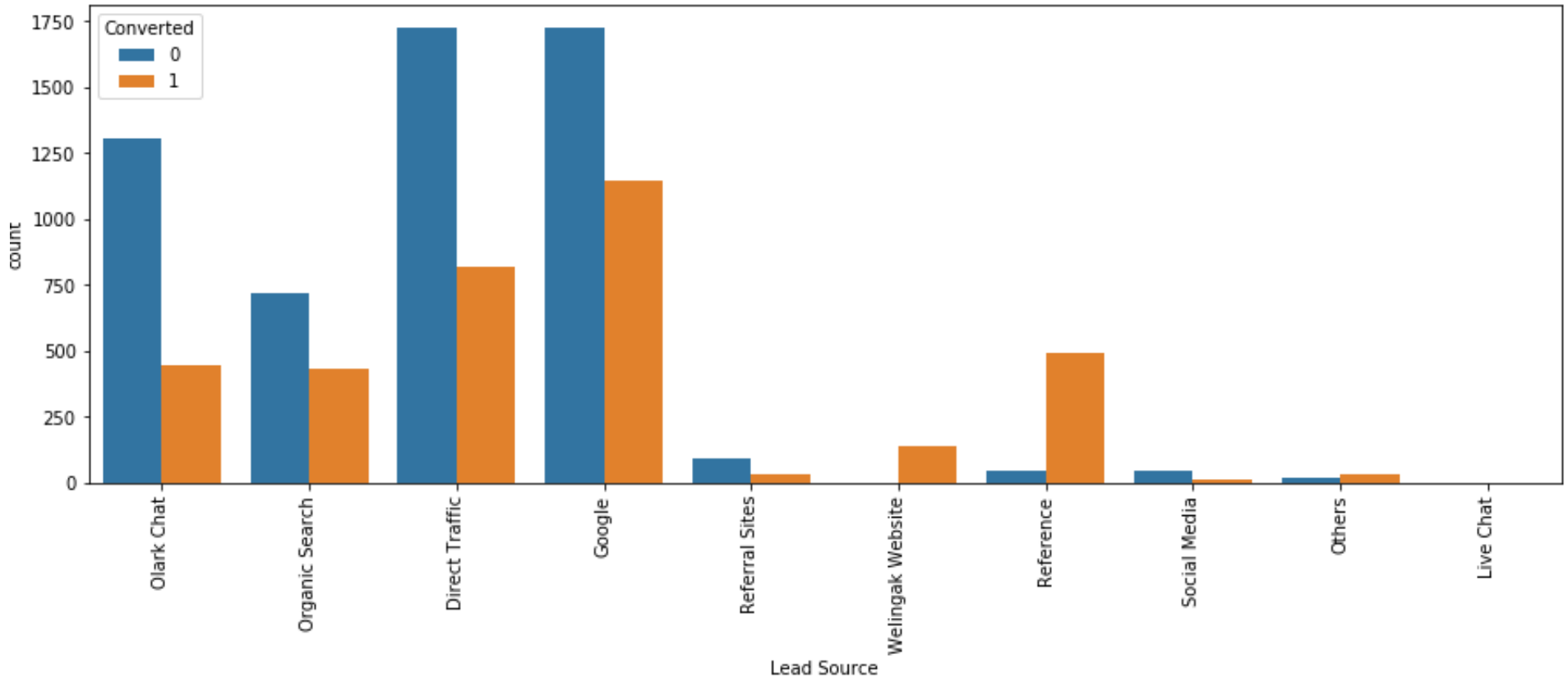


2.6. Most of the leads that got converted are most likely to revert after reading the email advertisement.



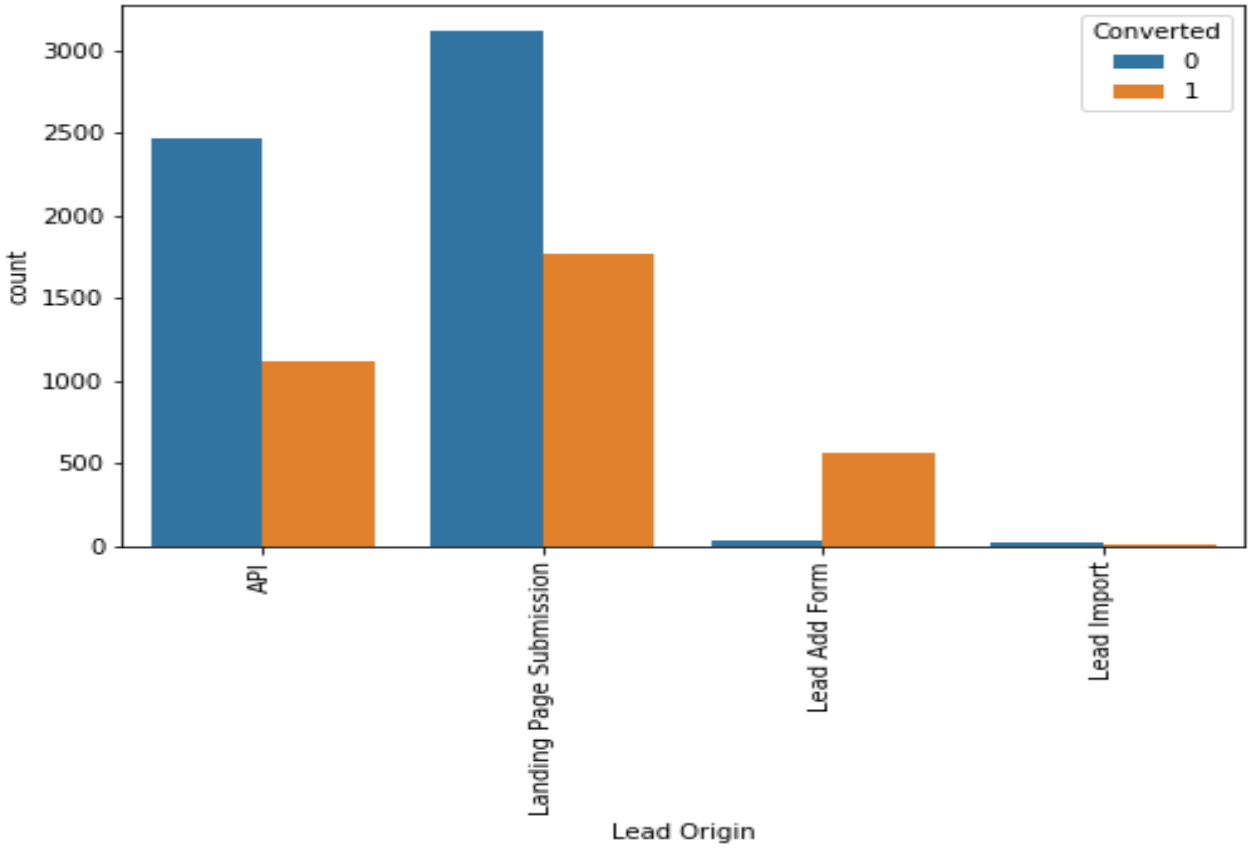
2.7. Source wise distribution of Converted and Non-Converted Leads

- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

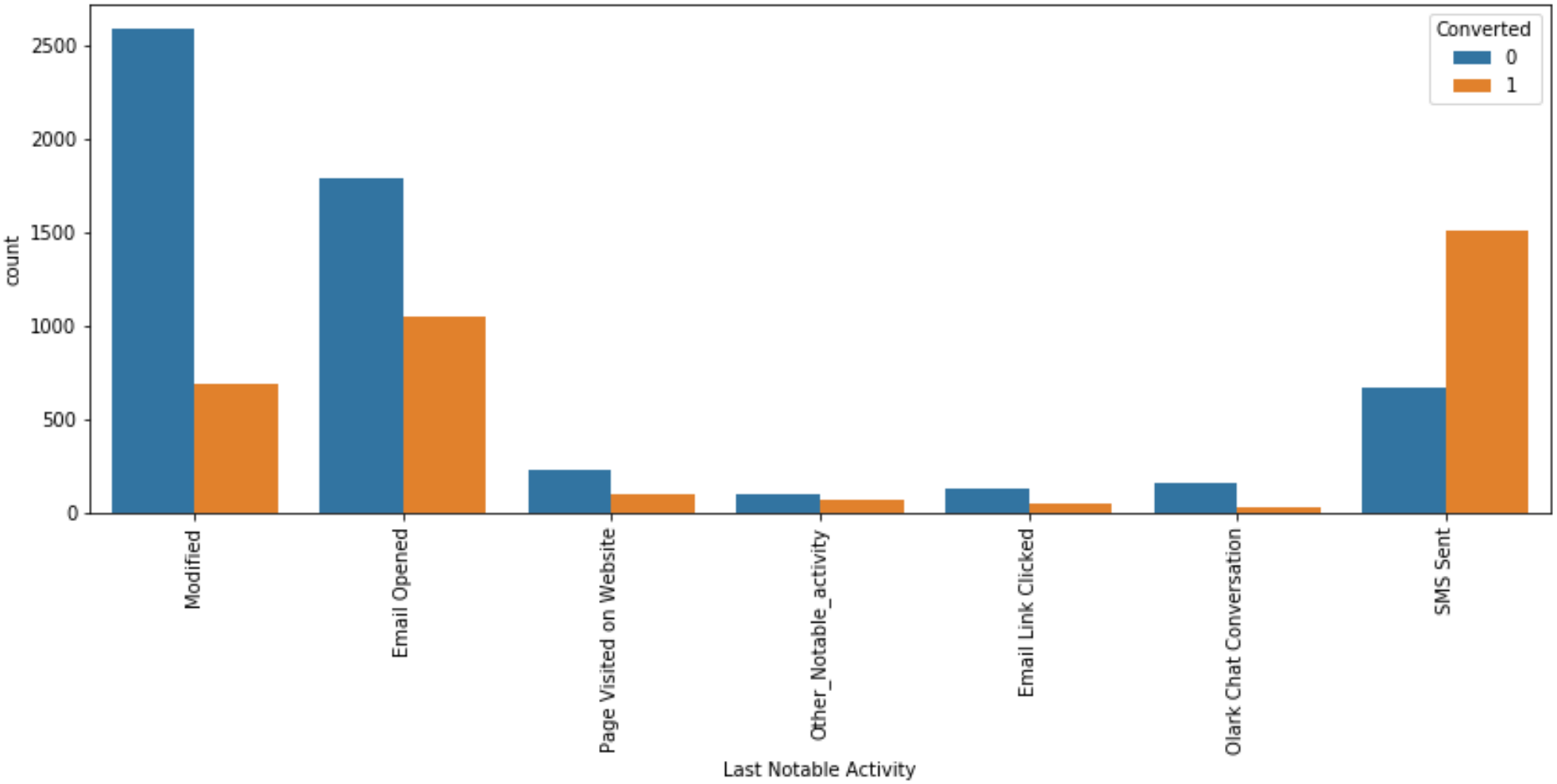


2.8 Lead Origin distribution of Converted and Non-Converted Leads

- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

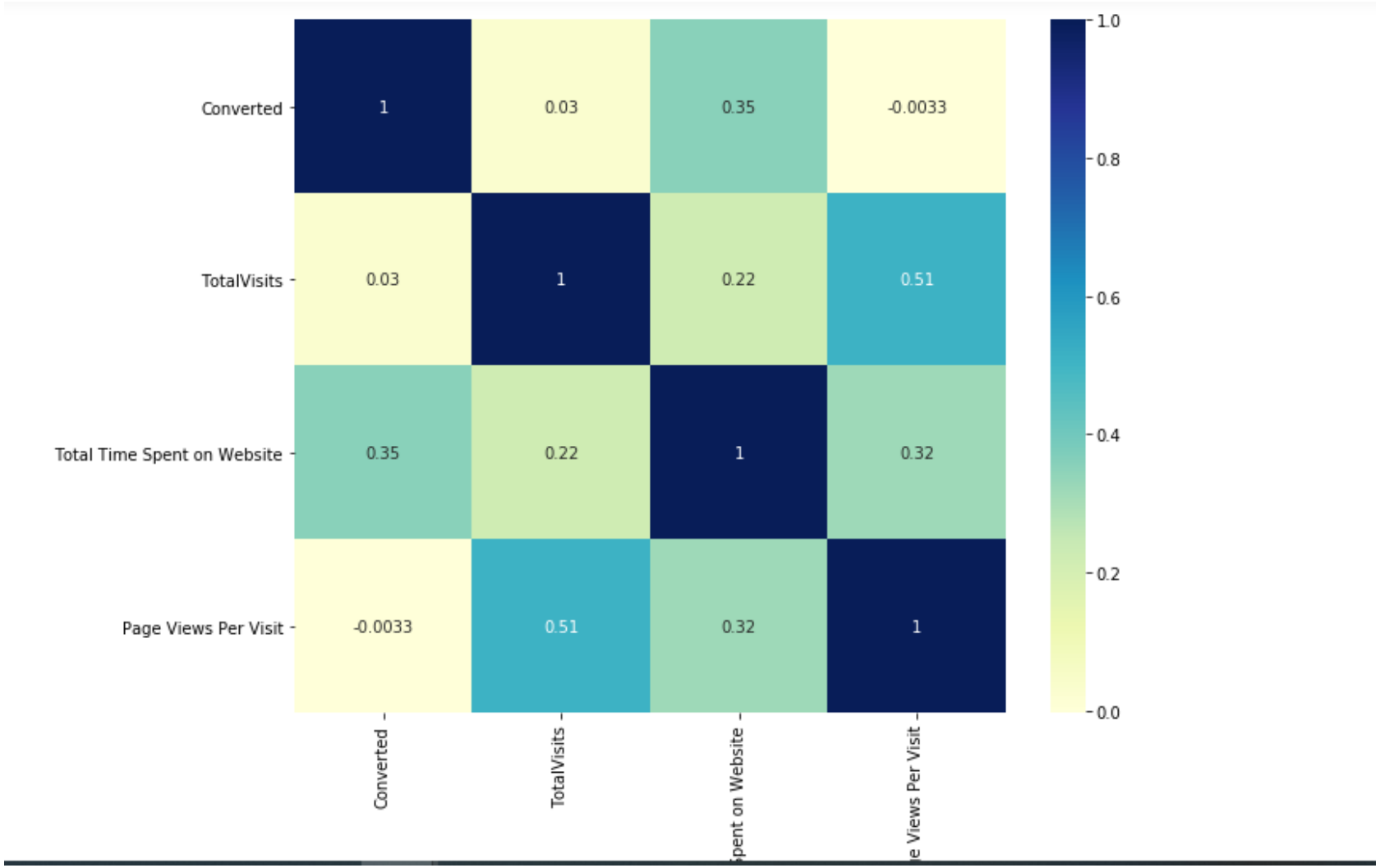


2.9. Based on the last notable acitivity performed by the student, maximum number of converted and non-converted leads came from SMS and E-mail



3. NUMERICAL ATTRIBUTES ANALYSIS

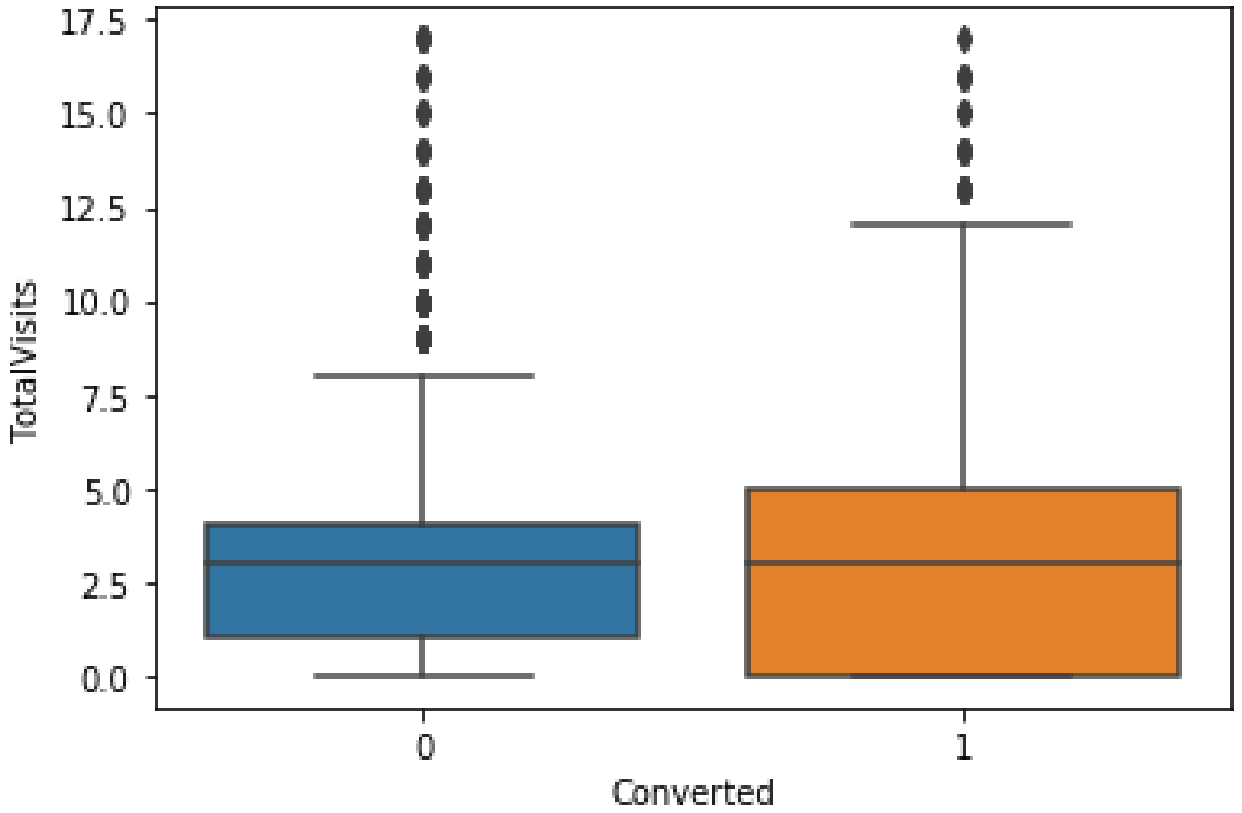
3.1. Numerical Columns correlation is plotted using heatmap.



3.2. Plot of Total visits Vs the Converted Leads

Inference

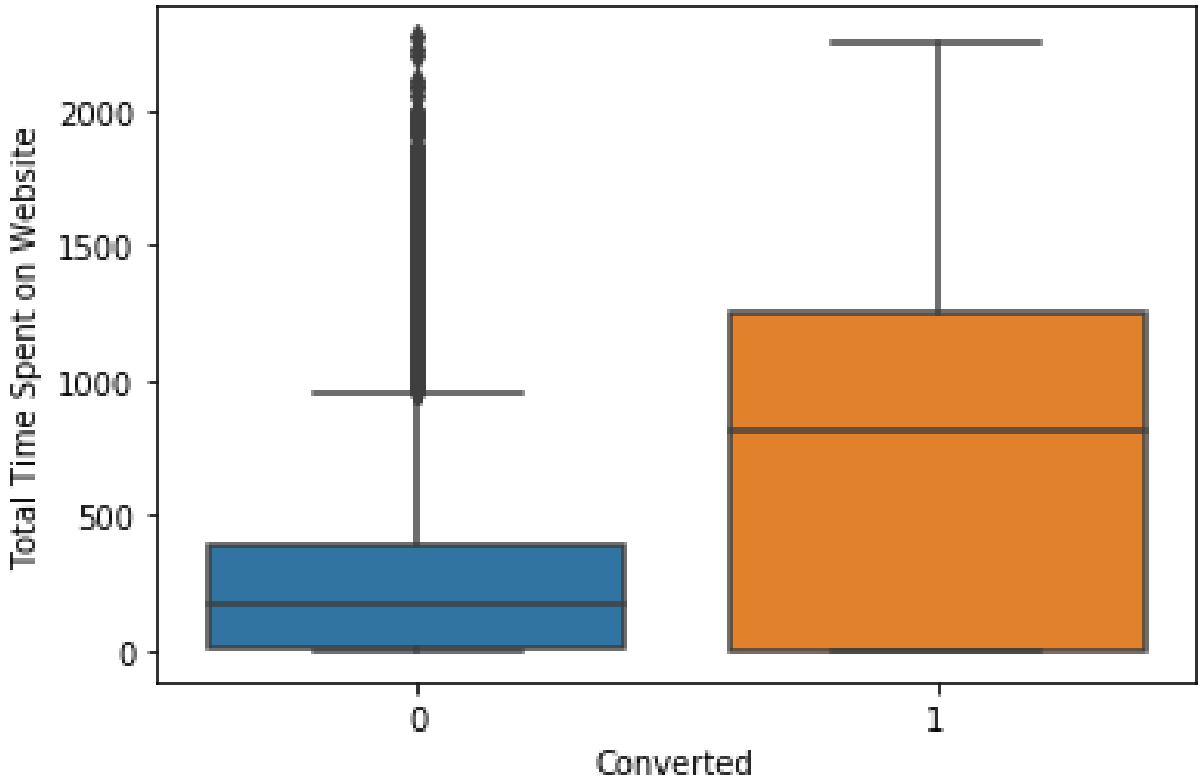
- Median for converted and not converted leads are the close.
- Nothng conclusive can be said on the basis of Total Visits



3.3. Plot of Total Time spent on Website Vs the Converted Leads

Inference

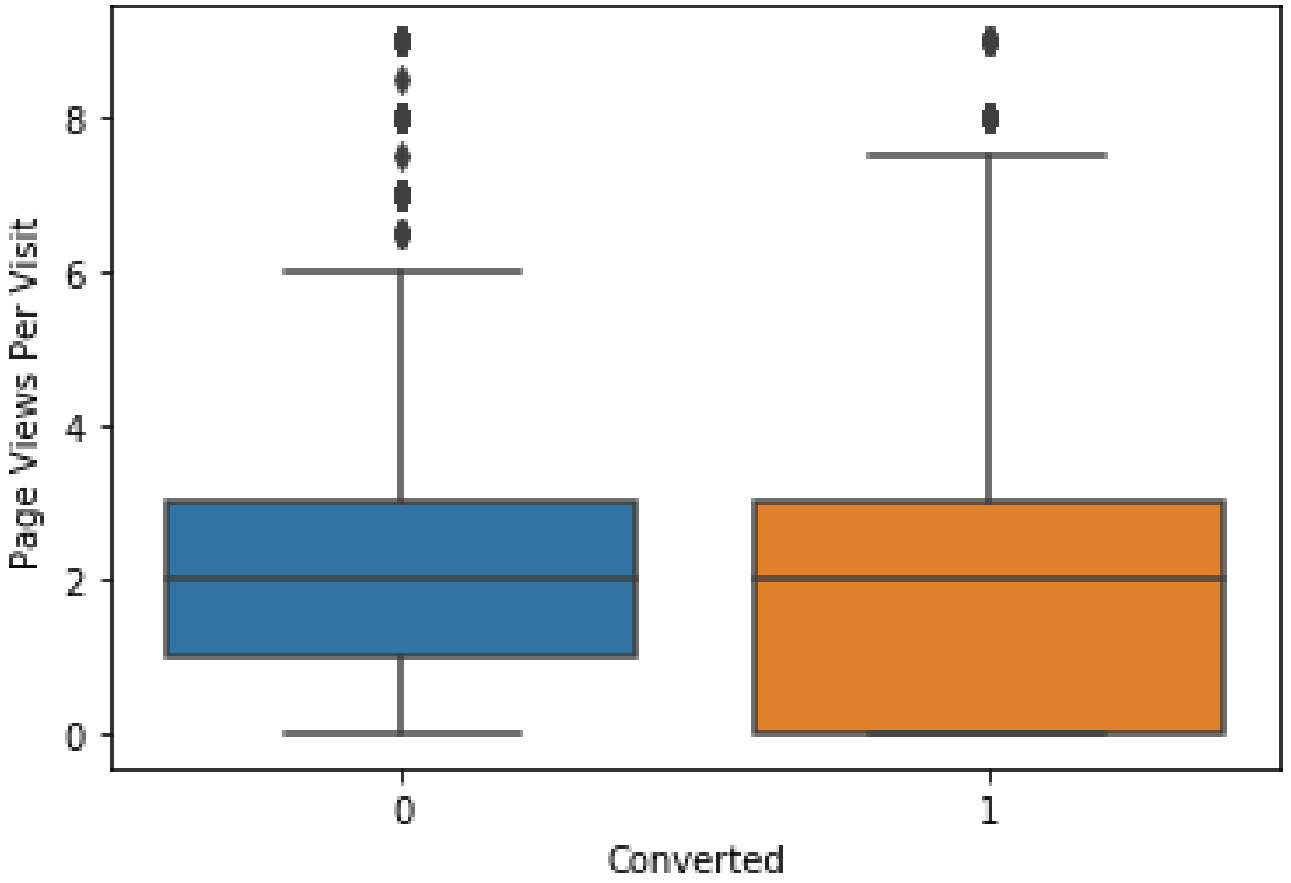
- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.



3.4. Plot of Page Views Per Visit on Website Vs the Converted Leads

Inference

- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit



4. Logistic Regression Model Building

To solve this problem we have to 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.

4.1 We Trained the logistic regression model on 70% of complete data and tested it on 30% of the remaining data.

4.2 Running RFE with 15 variables as output.

Final Output for RFE model building

Generalized Linear Model Regression Results

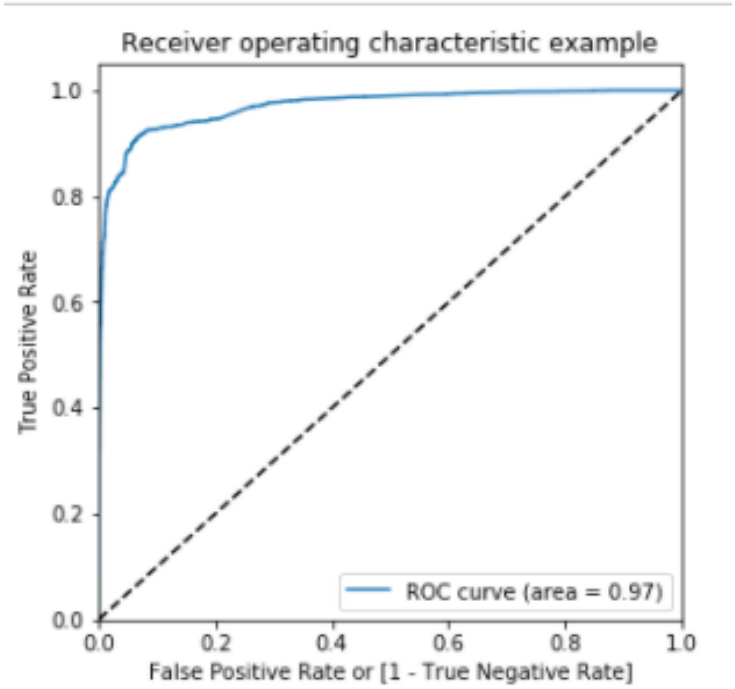
Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6253
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1263.3
Date:	Mon, 09 Aug 2021	**Deviance:**	2526.6
Time:	00:33:24	**Pearson chi2:**	8.51e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1179	0.084	-13.382	0.000	-1.282	-0.954
Total Time Spent on Website	0.8896	0.053	16.907	0.000	0.786	0.993
Lead Origin_Lead Add Form	1.6630	0.455	3.657	0.000	0.772	2.554
Lead Source_Direct Traffic	-0.8212	0.127	-6.471	0.000	-1.070	-0.572
Lead Source_Welingak Website	3.8845	1.114	3.488	0.000	1.701	6.068
Last Activity_SMS Sent	1.9981	0.113	17.718	0.000	1.777	2.219
Last Notable Activity_Modified	-1.6525	0.124	-13.279	0.000	-1.896	-1.409
Last Notable Activity_Olark Chat Conversation	-1.8023	0.491	-3.669	0.000	-2.765	-0.839
Tags_Closed by Horizzon	7.1955	1.020	7.053	0.000	5.196	9.195
Tags_Interested in other courses	-2.1318	0.406	-5.253	0.000	-2.927	-1.336
Tags_Lost to EINS	5.9177	0.611	9.689	0.000	4.721	7.115
Tags_Other_Tags	-2.3737	0.206	-11.507	0.000	-2.778	-1.969
Tags_Ringing	-3.4531	0.238	-14.532	0.000	-3.919	-2.987
Tags_Will revert after reading the email	4.5070	0.188	24.002	0.000	4.139	4.875

	Features	VIF
1	Lead Origin_Lead Add Form	1.82
12	Tags_Will revert after reading the email	1.56
4	Last Activity_SMS Sent	1.46
5	Last Notable Activity_Modified	1.40
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
10	Tags_Other_Tags	1.25
0	Total Time Spent on Website	1.22
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01

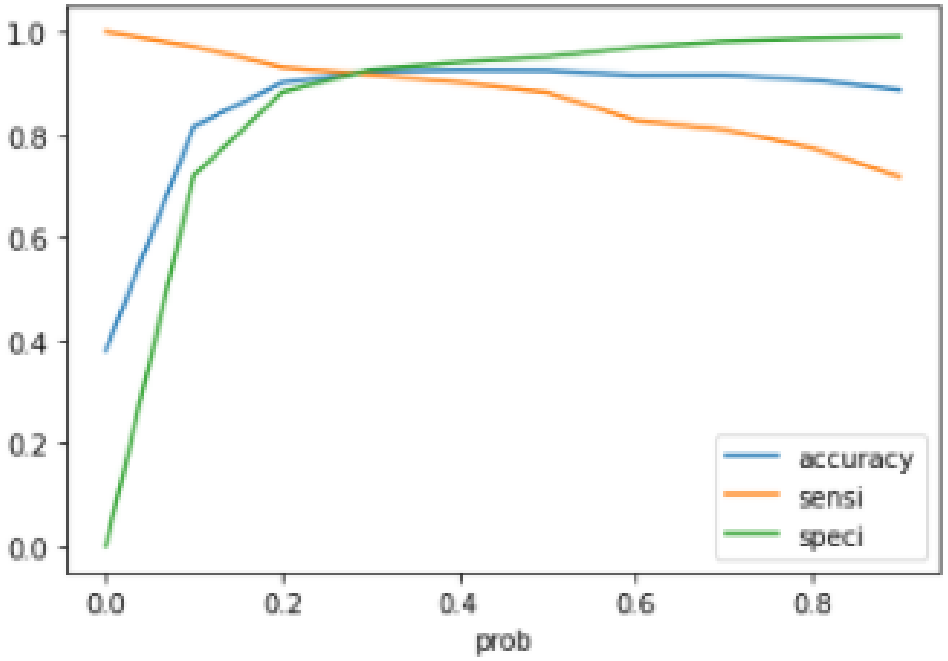
5. ROC Curve

The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.



Optimal Cut-Off point

From the curve below, 0.3 is the optimum point to take it as a cutoff probability.



Observation

- So as we can see above the model seems to be performing well. The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:
- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%
- Some of the other Stats are derived below, indicating the False Positive Rate, Positive Predictive Value,Negative Predictive Values, Precision & Recall.

Final Observation

- Let us compare the values obtained for Train & Test:
- **Train Data:**
- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%
- **Test Data:**
- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%
- The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

Conclusion

- It was found that the variables that mattered the most are the following -:
- Lead Origin_Lead Add Form
- Last Activity_SMS Sent
- Tags_Will revert after reading the email

- The important 3 categorical variables in the model which should be focused are -:
- Lead Origin
- Lead Source
- Last Activity

- The X education company should use the following strategies to increase the probability of lead conversion -:

- Target leads which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
- Target leads where the last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
- Target leads that have read the emails or replied on it. This shows their interest.
- Students can be approached, but they will have a lower probability of converting due to the course being industry based. However, this can also be a motivating factor to ensure industry readiness by the time they complete their education