LEAD SCORING CASE STUDY
(LOGISTIC REGRESSION)

SOLON KUMAR DAS

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

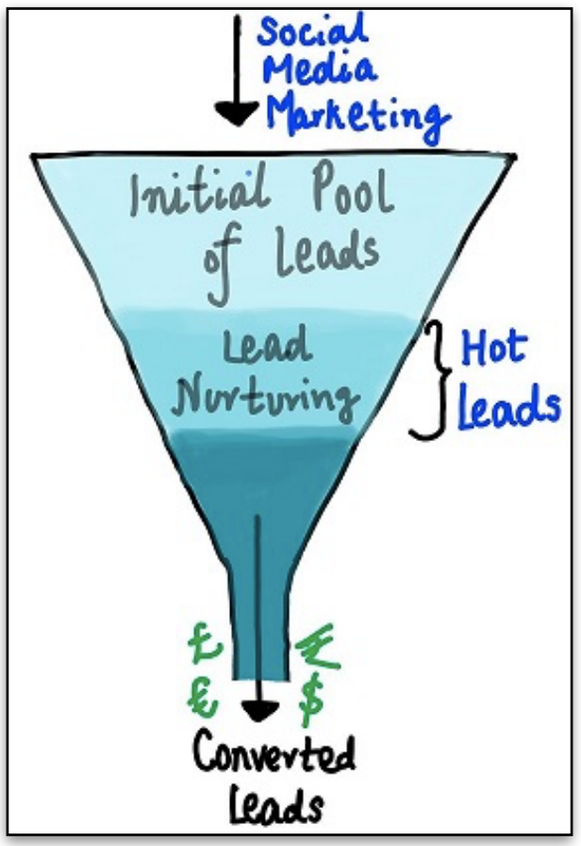
PROBLEM STATEMENT

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

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. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:



As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

GOALS OF THE CASE STUDY

There are quite a few goals for this case study which are as follows:

1. 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.

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

DATA

The data provided (leads dataset) are from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

DATA CLEANING

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

Missing Value Treatment

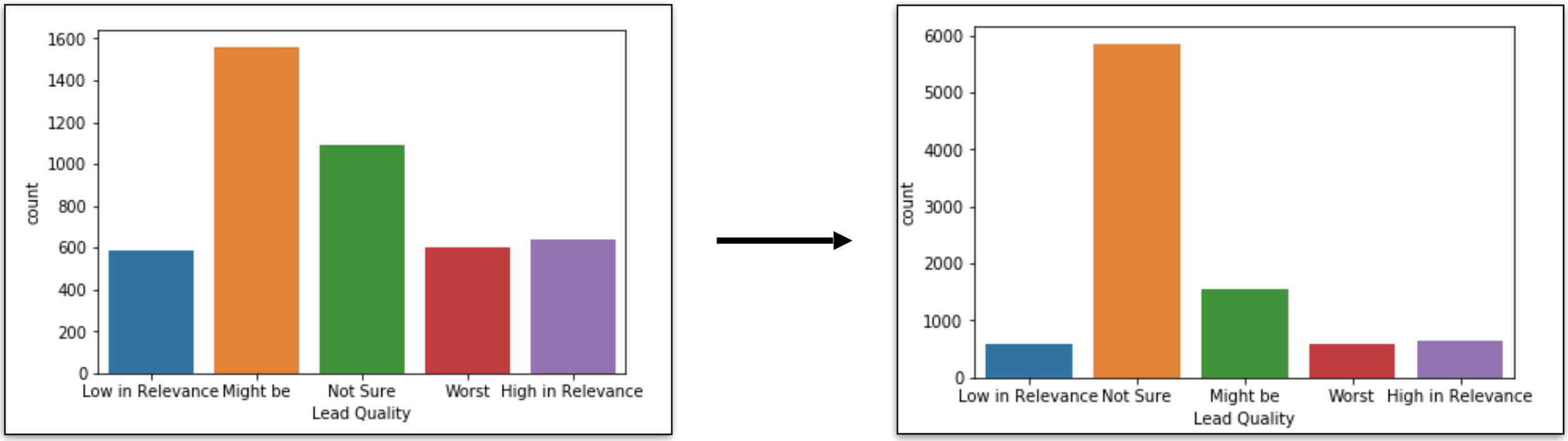
Prospect ID	0.00
Lead Number	0.00
Lead Origin	0.00
Lead Source	0.39
Do Not Email	0.00
Do Not Call	0.00
Converted	0.00
TotalVisits	1.48
Total Time Spent on Website	0.00
Page Views Per Visit	1.48
Last Activity	1.11
Country	26.63
Specialization	36.58
How did you hear about X Education	78.46
What is your current occupation	29.11
What matters most to you in choosing a course	29.32
Search	0.00
Magazine	0.00
Newspaper Article	0.00
X Education Forums	0.00
Newspaper	0.00
Digital Advertisement	0.00
Through Recommendations	0.00
Receive More Updates About Our Courses	0.00
Tags	36.29
Lead Quality	51.59
Update me on Supply Chain Content	0.00
Get updates on DM Content	0.00
Lead Profile	74.19
City	39.71
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65
I agree to pay the amount through cheque	0.00
A free copy of Mastering The Interview	0.00
Last Notable Activity	0.00

Missing value treatment of each column:

1. There were many cells which had ‘Select’ as their values which meant the customer did not select any of the available choices. Hence those values were converted into NULL values.
2. Any column having missing value percentage above 70 % was removed. Eg. Lead Profile.
3. Rest of the columns which have missing value percentage more than 2 % were looked upon individually and the missing values were imputed.

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

Lead Quality(51.59%)



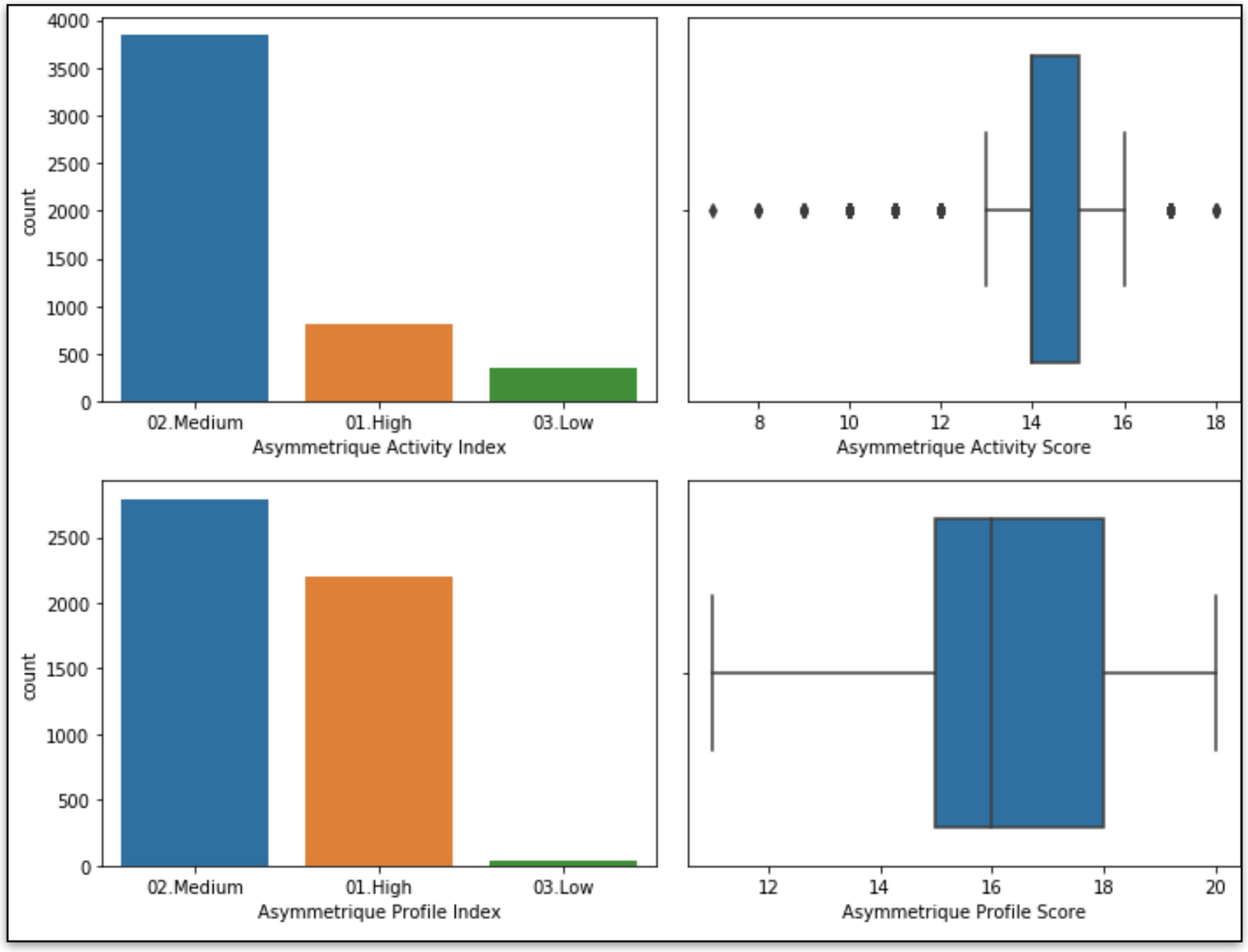
As this column is based on Customer intuition we can impute the missing value to be 'Not Sure'

Asymmetric Activity Index (45.65 %)

Asymmetric Profile Index (45.65 %)

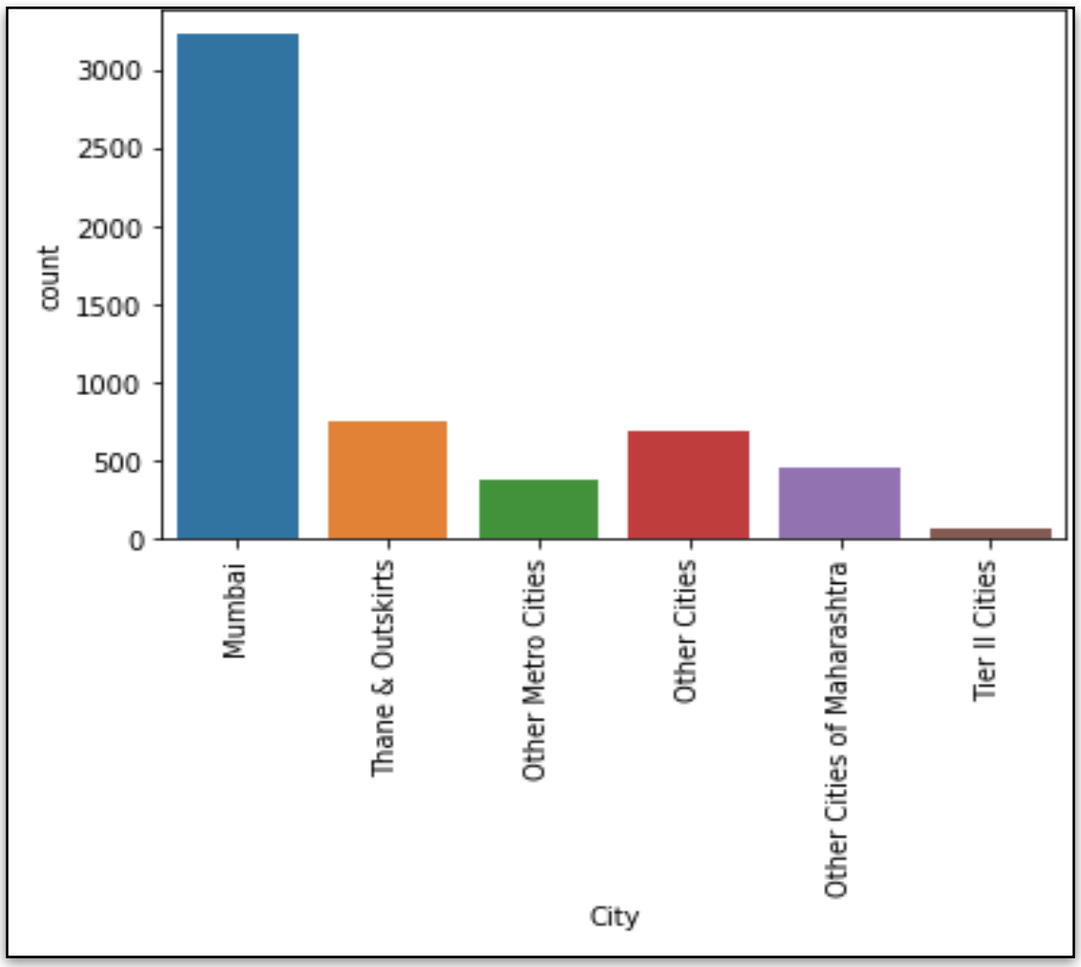
Asymmetric Activity Score (45.65 %)

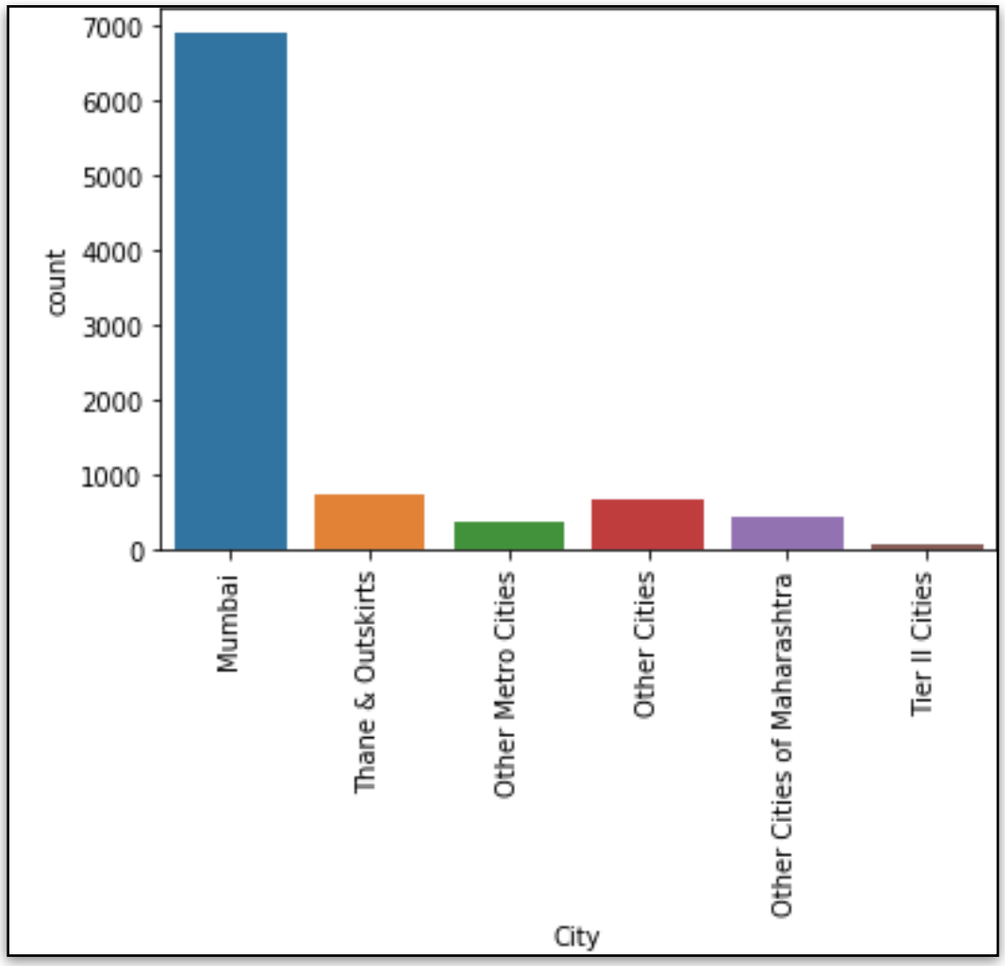
Asymmetric Profile Score (45.65 %)



There is too much variation in thes parameters so its not reliable to impute any value in it.
45% null values means we need to drop these columns.

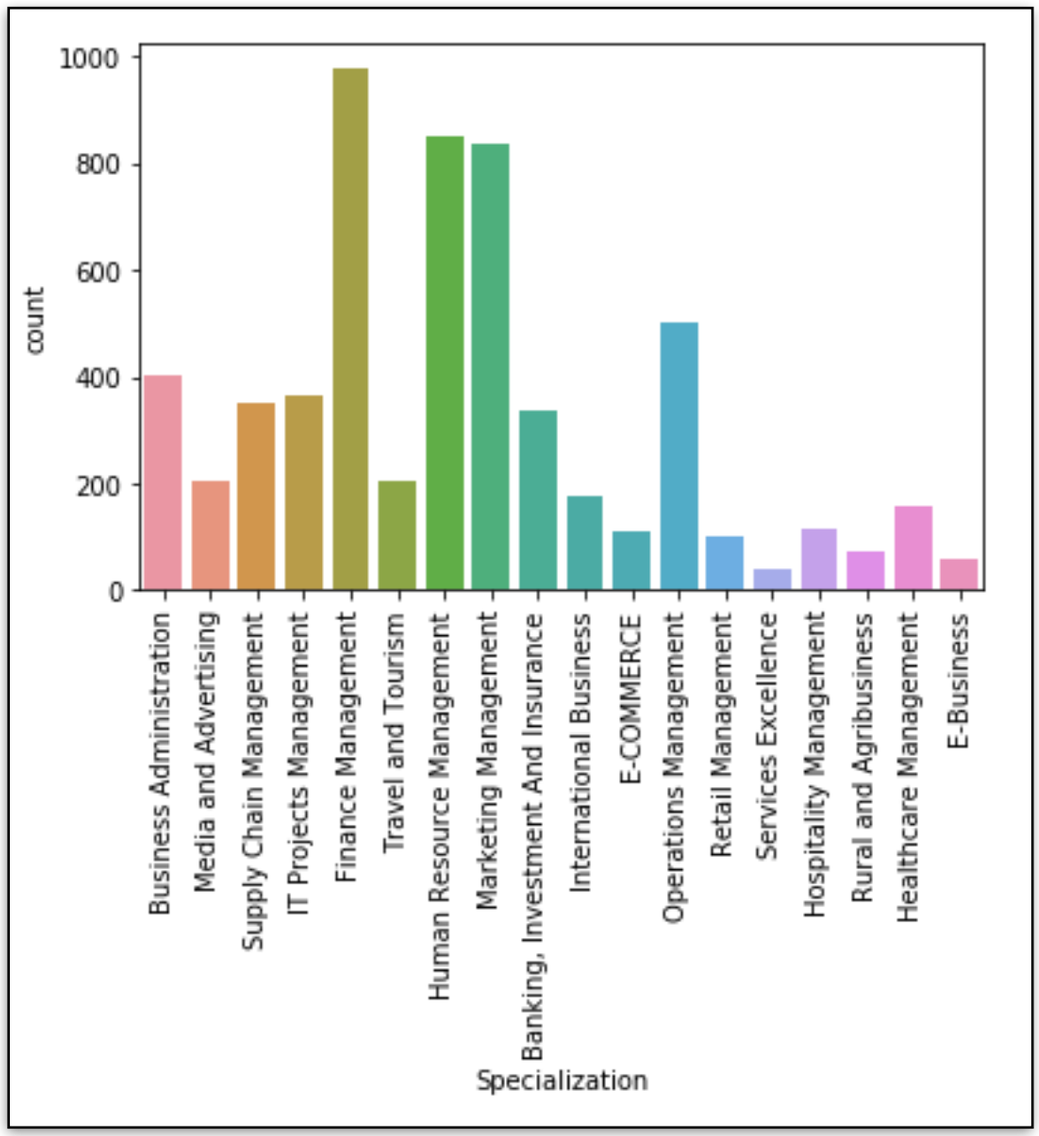
City (39.71%)

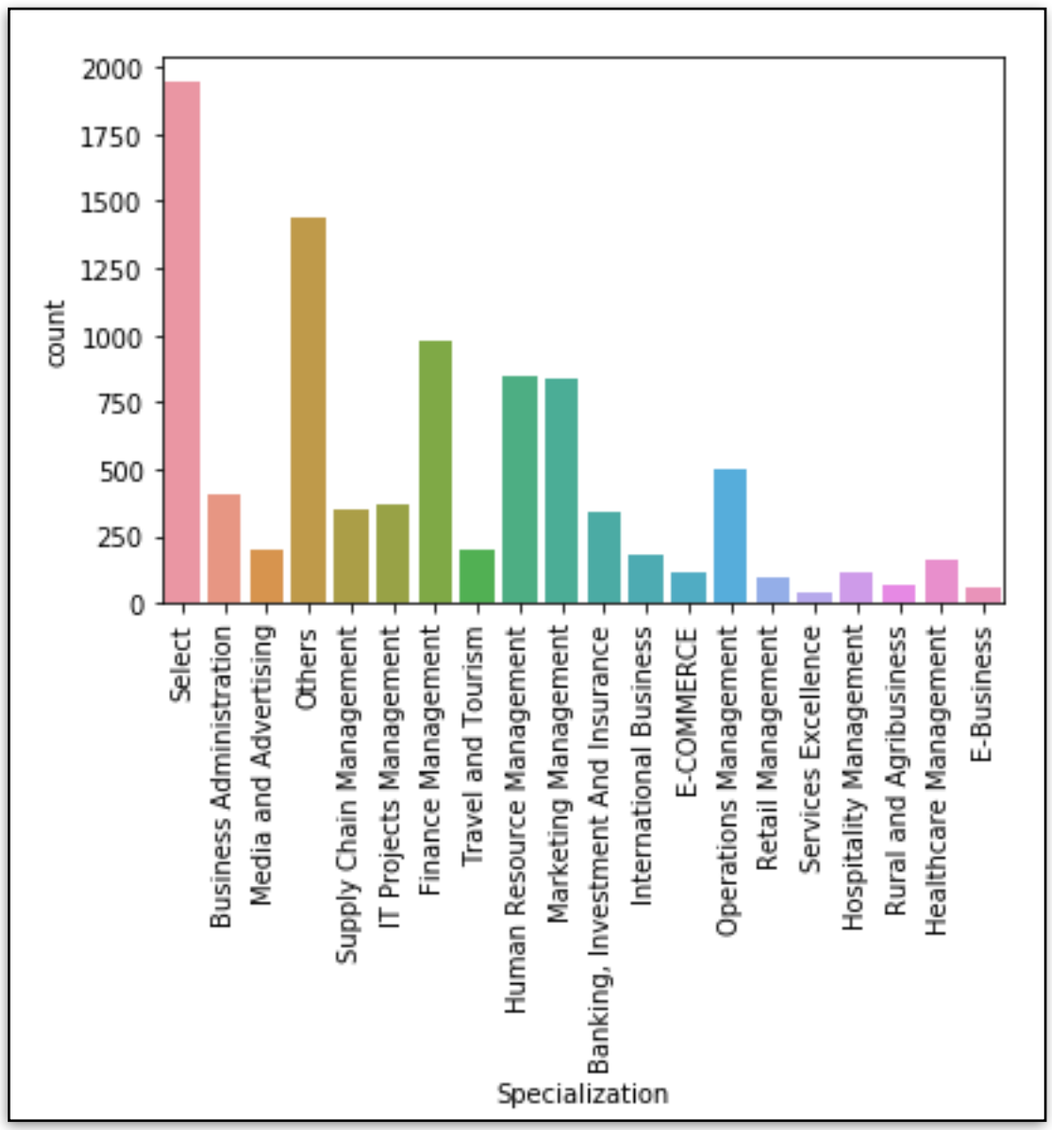




Approximately 60 percent of the City column has Mumbai in it. So we will impute the missing values to Mumbai.

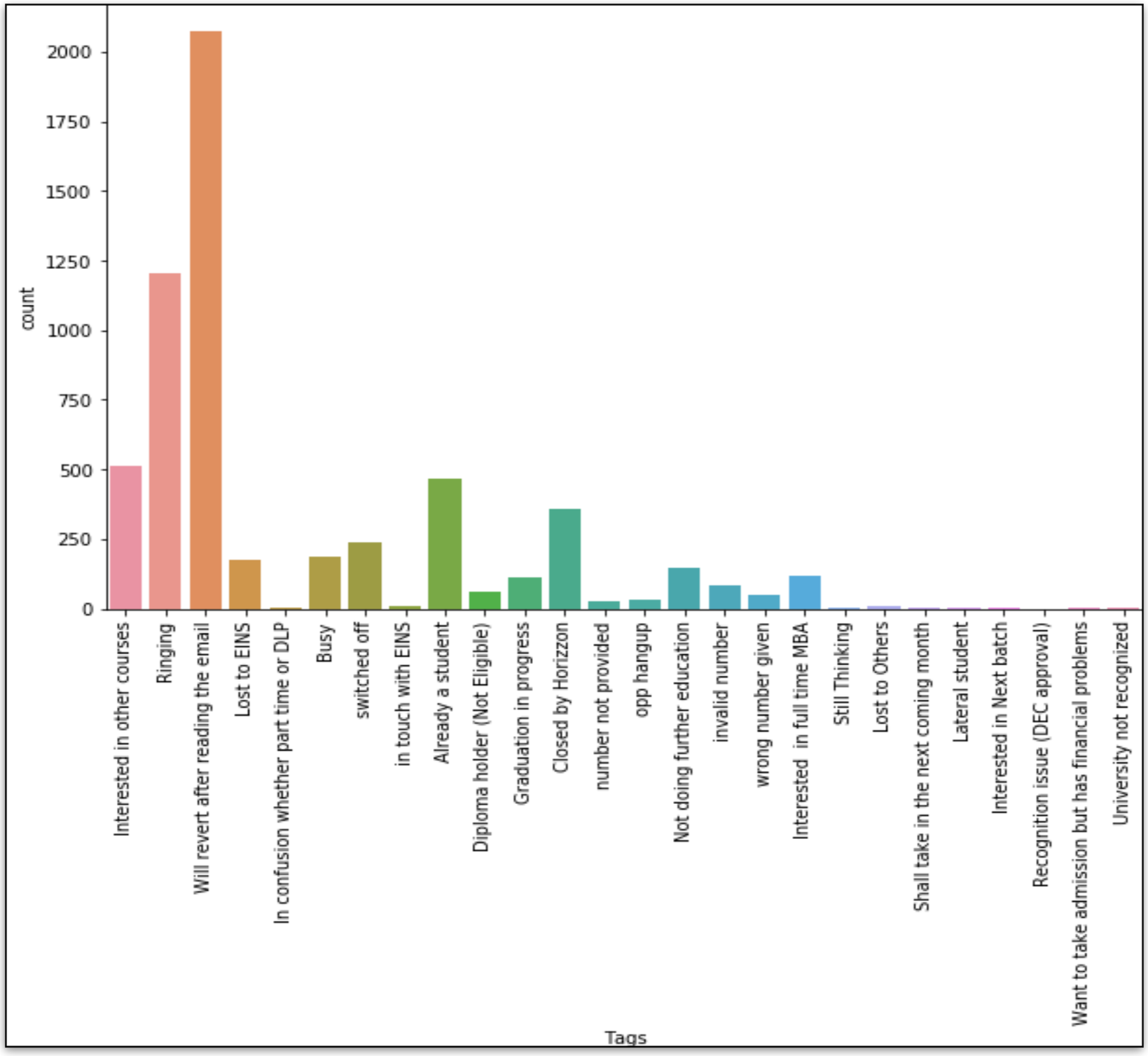
Specialization (36.58%)





It maybe the case that lead has not entered any specialization if their option is not available on the list, may not have any specialization or is a student. Hence we can make a category "Others" for missing values.

Tags (36.29%)



The missing values will be imputed by 'Will revert after reading the email'

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

'What is your current occupation' (29 %)

count	6550
unique	6
top	Unemployed
freq	5600

86% entries are of Unemployed so we can impute "Unemployed" in it. so we will impute the missing value as Unemployed.

Country(26.63%)

count	6779
unique	38
top	India
freq	6492

Country is India for most values so so imputing the same in missing values

What matters most to you in choosing a course(29.32%)

count	6531
unique	3
top	Better Career Prospects
freq	6528

Replacing the missing value by 'Better Career Prospects' as most of the values are of the said value

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

Result of the Treatment

```
Prospect ID                                      0.0
Lead Number                                      0.0
Lead Origin                                      0.0
Lead Source                                      0.0
Do Not Email                                     0.0
Do Not Call                                      0.0
Converted                                        0.0
TotalVisits                                      0.0
Total Time Spent on Website                      0.0
Page Views Per Visit                             0.0
Last Activity                                    0.0
Country                                          0.0
Specialization                                   0.0
What is your current occupation                  0.0
What matters most to you in choosing a course    0.0
Search                                           0.0
Magazine                                         0.0
Newspaper Article                                0.0
X Education Forums                               0.0
Newspaper                                        0.0
Digital Advertisement                            0.0
Through Recommendations                          0.0
Receive More Updates About Our Courses           0.0
Tags                                             0.0
Lead Quality                                     0.0
Update me on Supply Chain Content                0.0
Get updates on DM Content                        0.0
City                                             0.0
I agree to pay the amount through cheque         0.0
A free copy of Mastering The Interview           0.0
Last Notable Activity                            0.0
dtype: float64
```

Now Data is cleaned we can start our Analysis on it

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

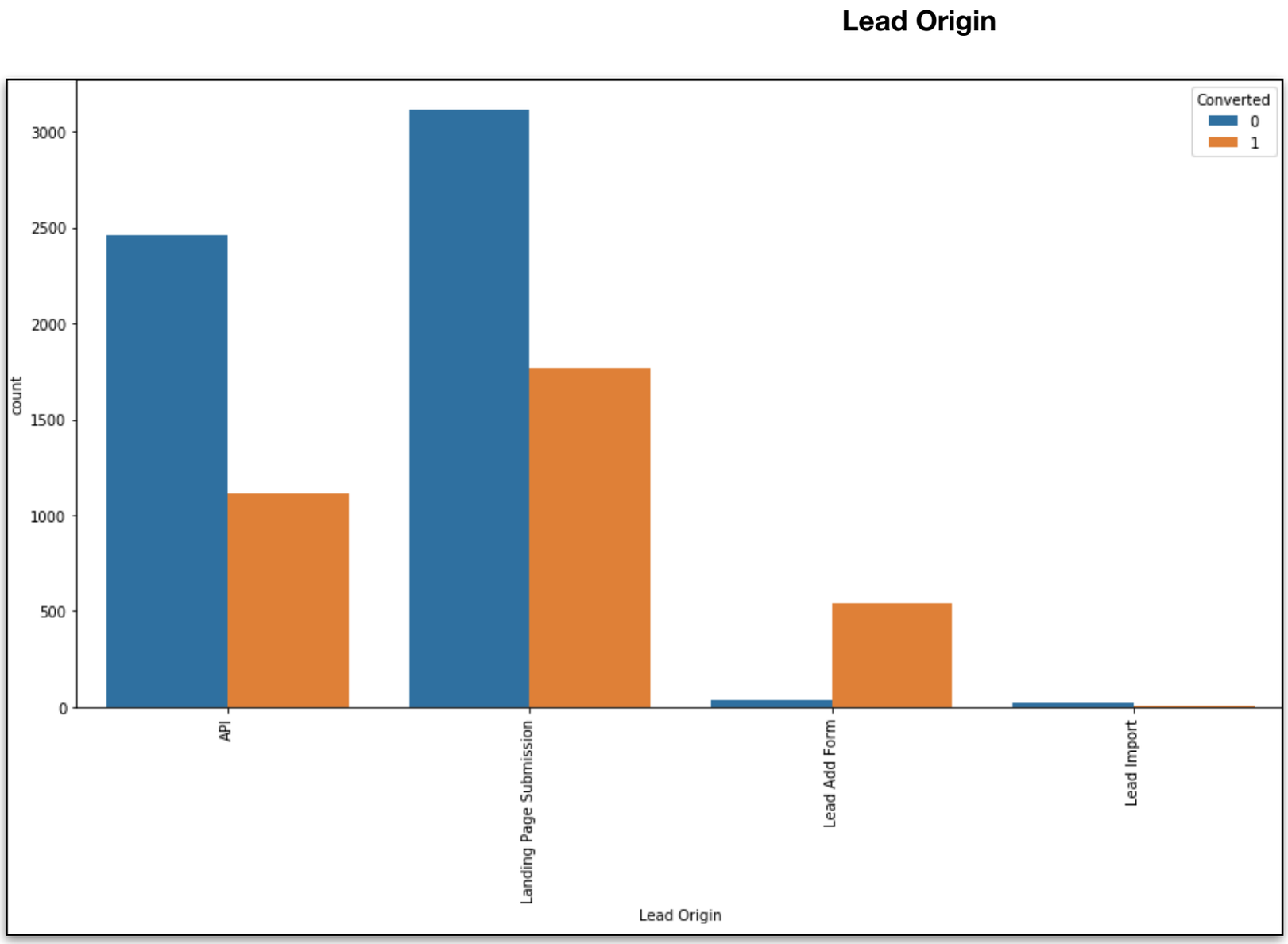
EXPLORATORY DATA ANALYSIS

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

GOALS OF EDA:

EDA is done to check whether each column significantly affects our target variable or not so that we can decide which variables to keep for model building and which variables to discard.

Our Target column is **Converted** which has a 38% Churn rate which means 38% of the 9240 records have churned.

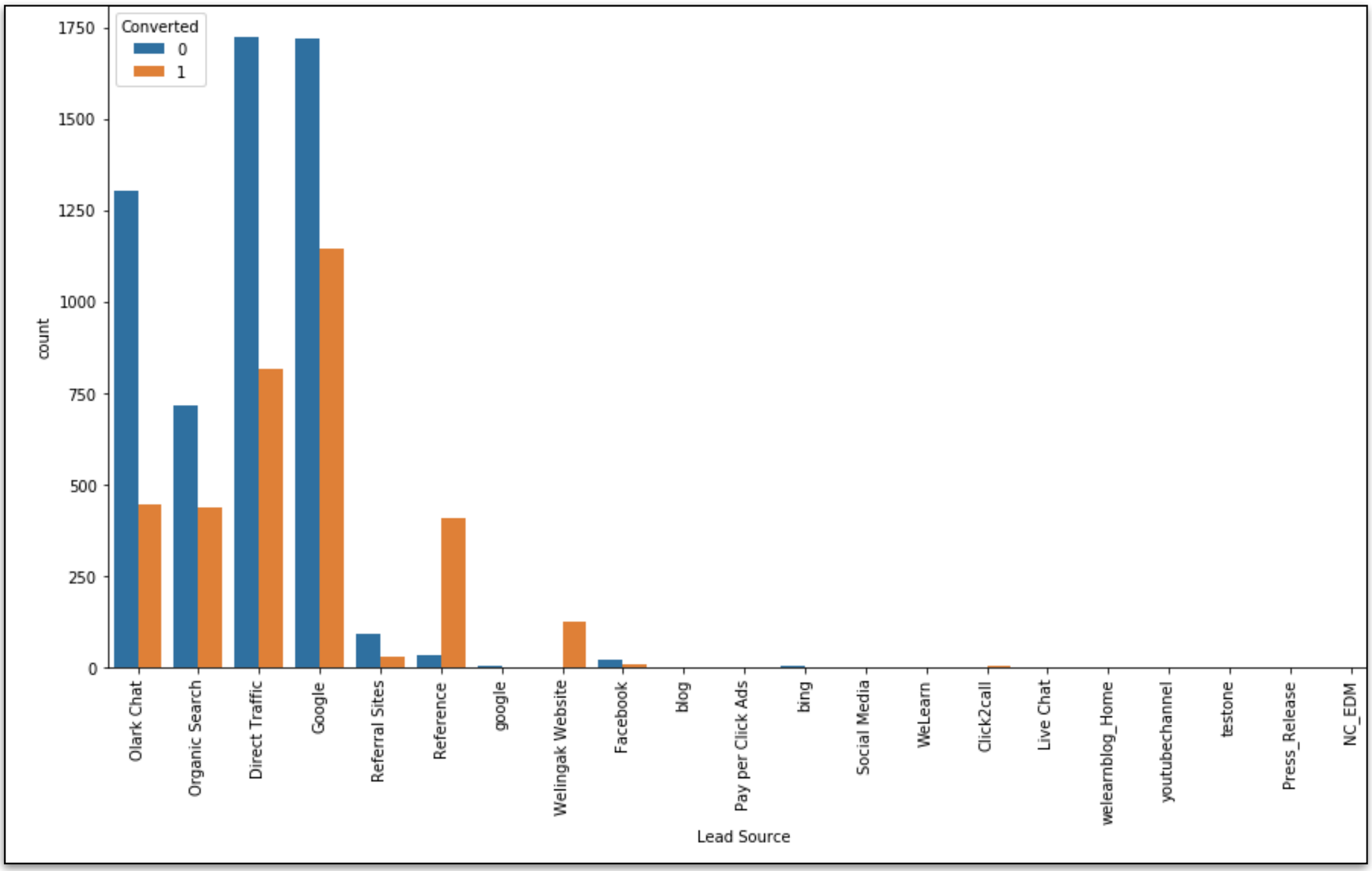


Inference:

1. API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
2. Lead Add Form has more than 90% conversion rate but count of lead are not very high.
3. Lead Import are very less in count.

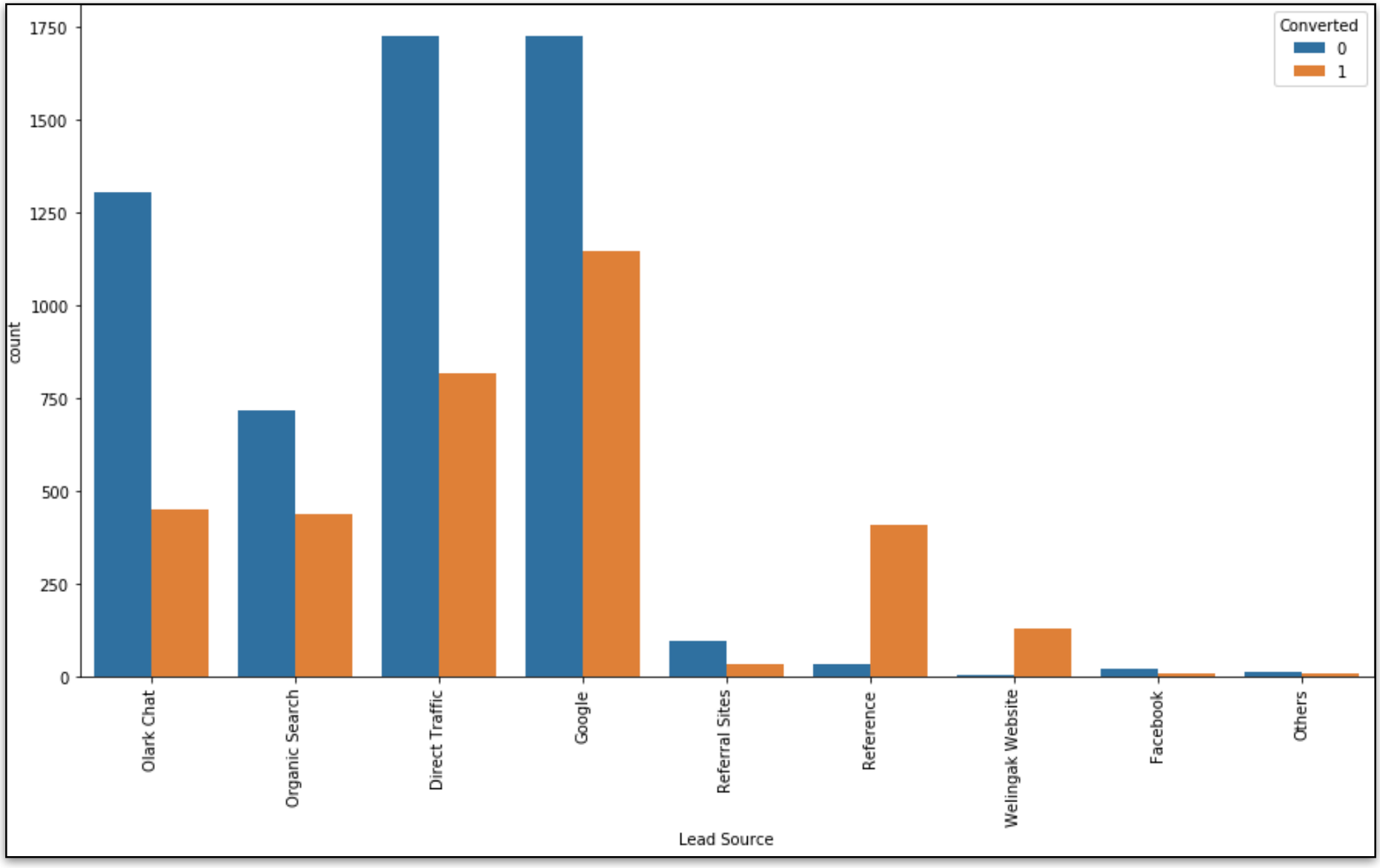
To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)



We see that ‘google’ is redundant and columns like 'Click2call', 'Live Chat', 'NC_EDM', 'Pay per Click Ads', 'Press_Release','Social Media', 'WeLearn', 'bing', 'blog', 'testone', 'welearnblog_Home', 'youtubechannel' have very less lead numbers. So we can classify them together as Others.

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

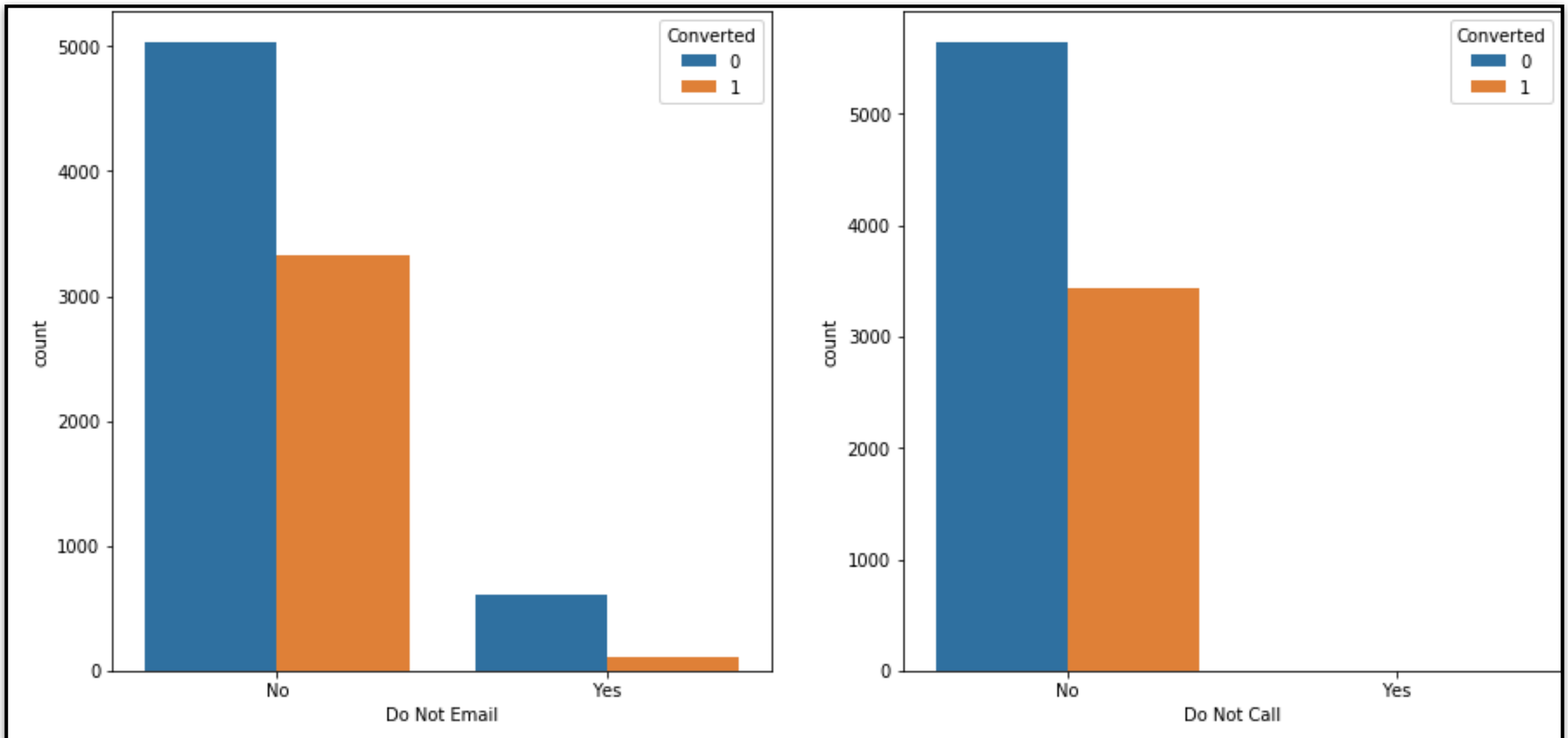


Inference

1. Google and Direct traffic generates maximum number of leads.
2. 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.

Do not Email and Do not Call

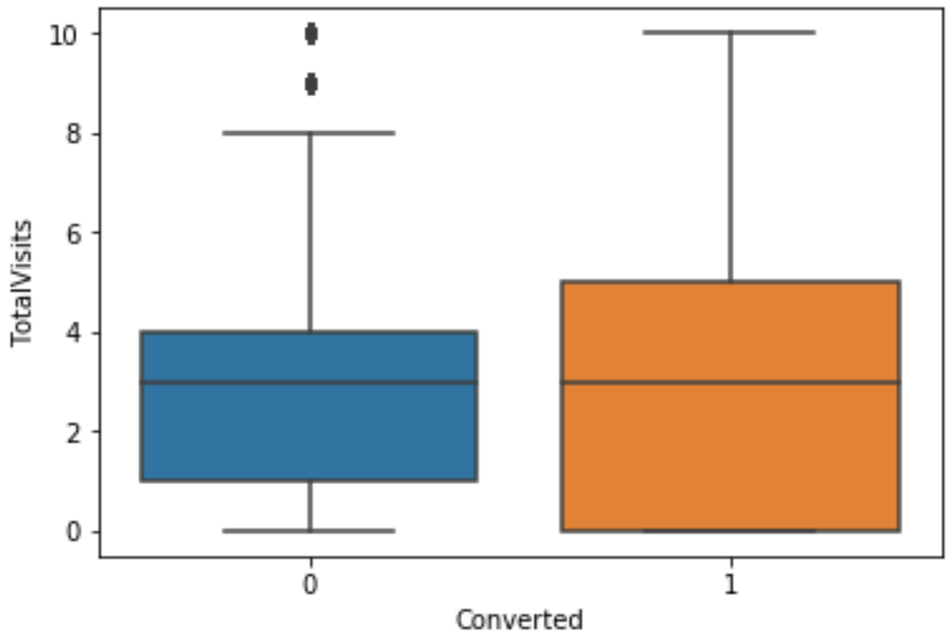


Inference

We see that people who have not opted for these services have a conversion rate

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

Total Visits

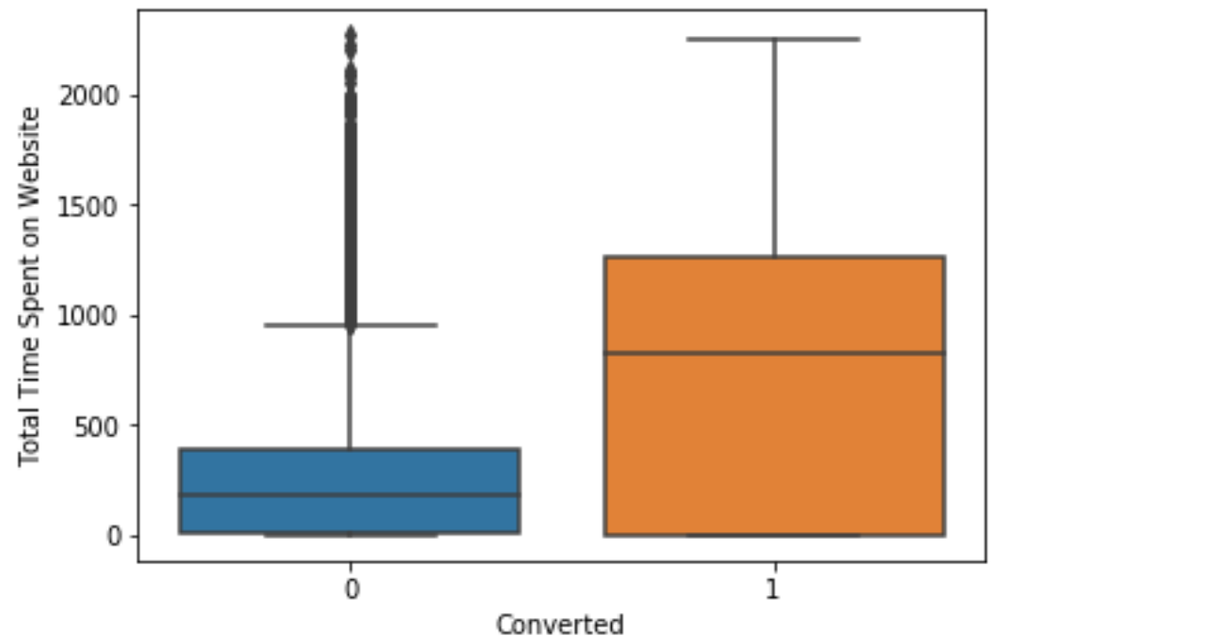


Inference

1. Median for converted and not converted leads are the same.

Nothng conclusive can be said on the basis of Total Visits.

Total time spent on website

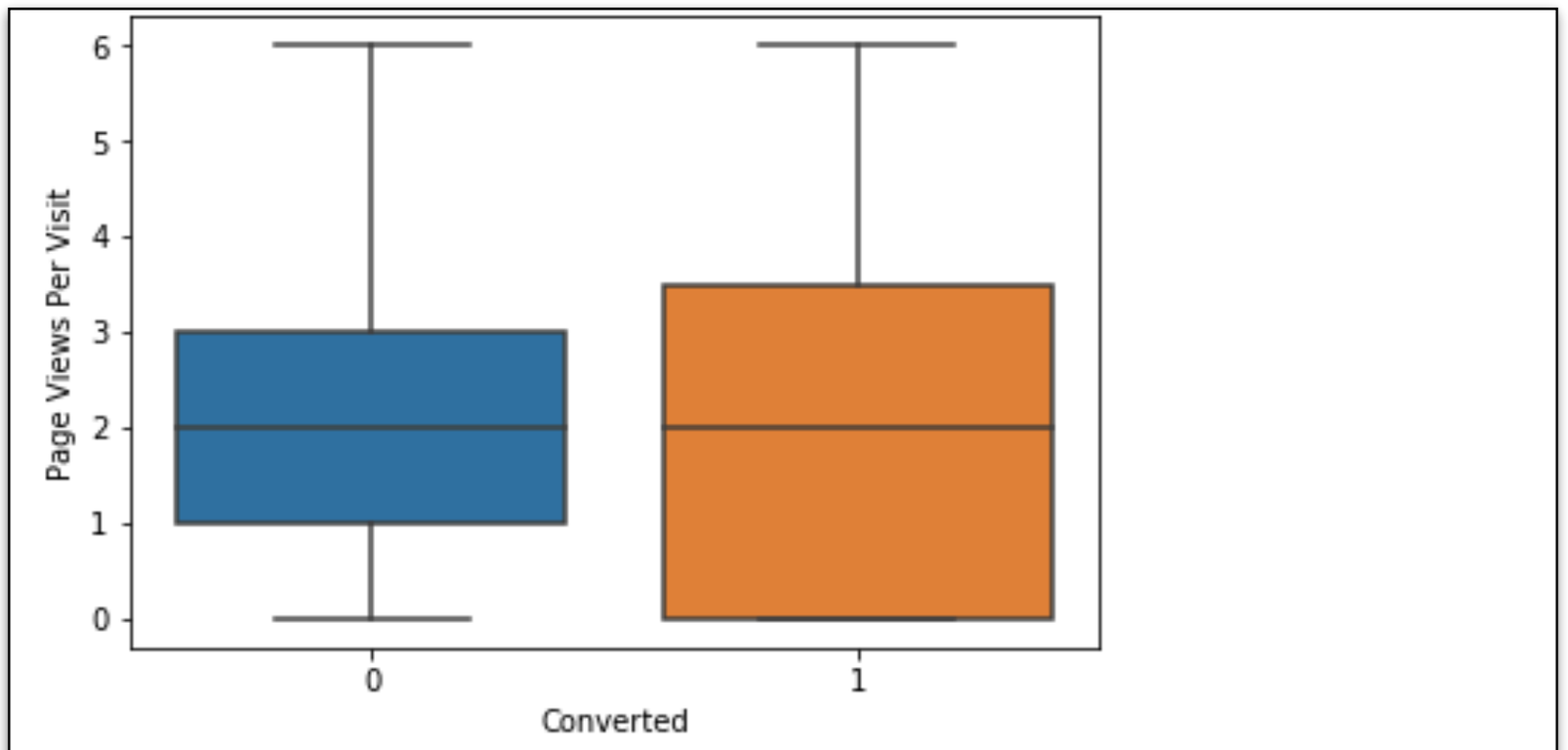


Inference

1. Leads spending more time on the weblise are more likely to be converted.

Website should be made more engaging to make leads spend more time.

Page views per visit

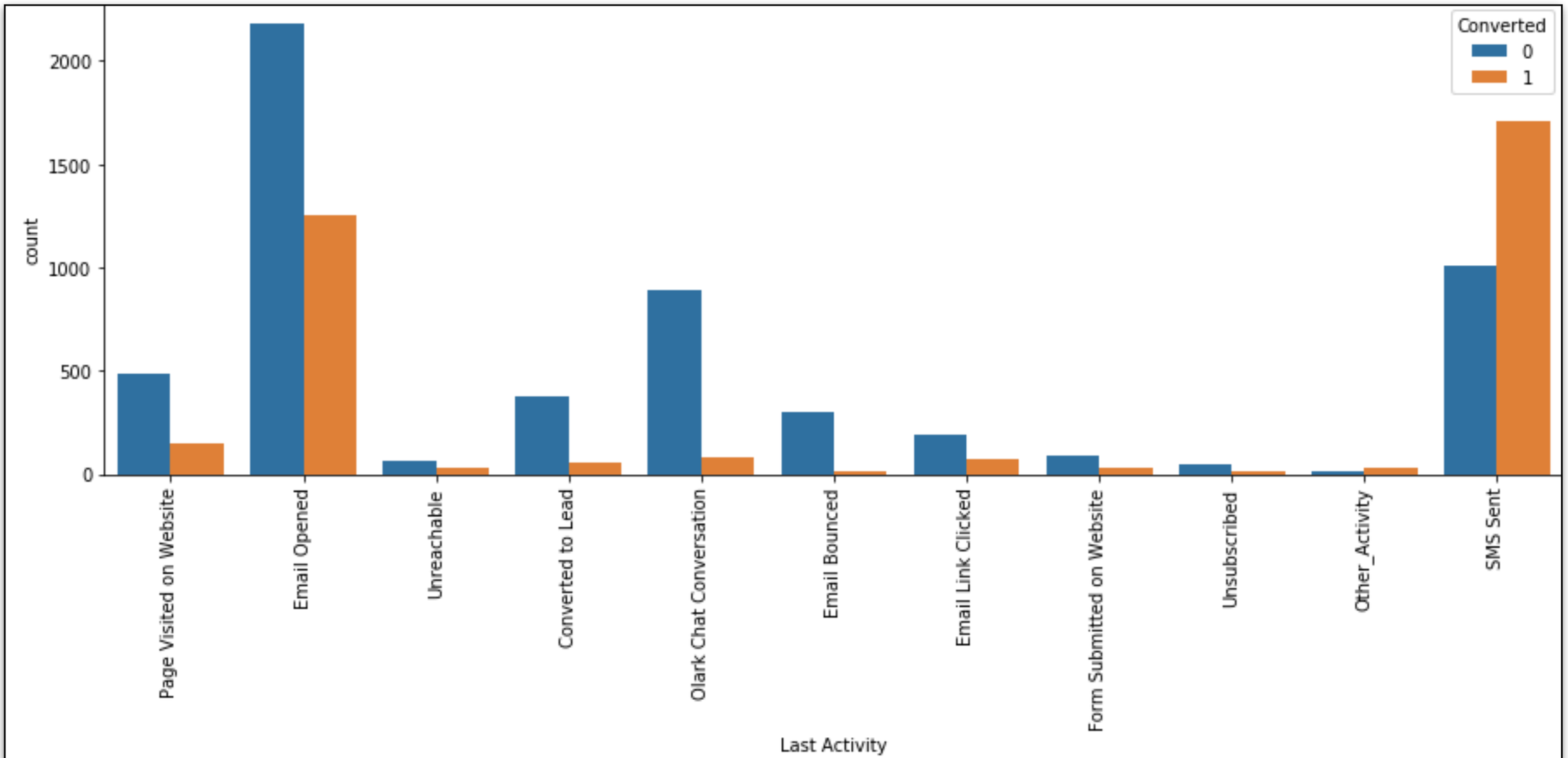


Inference

1. Median for converted and unconverted leads is the same.

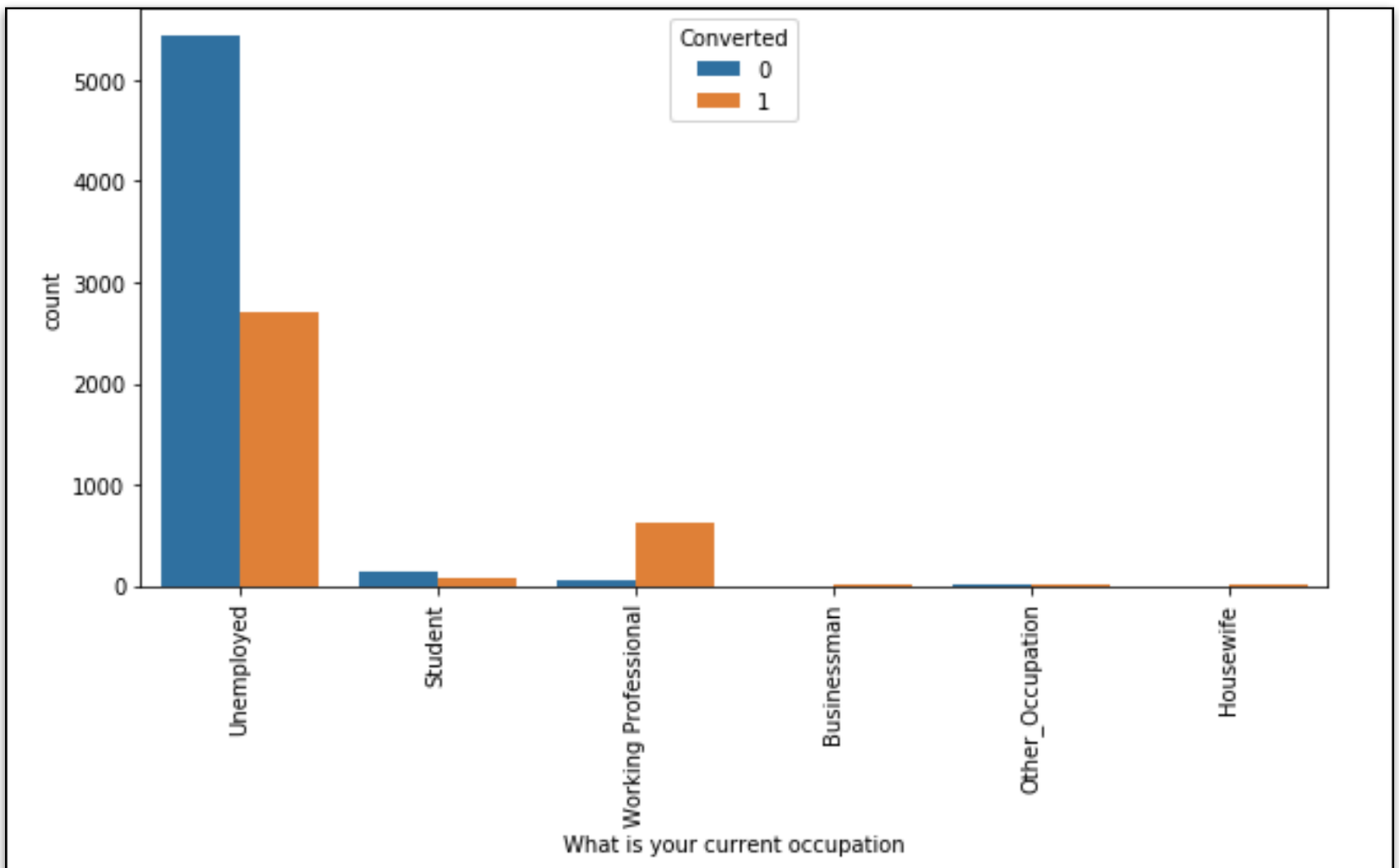
Nothing can be said specifically for lead conversion from Page Views Per Visit

Last Activity



Inference

1. Most of the lead have their Email opened as their last activity.
2. Conversion rate for leads with last activity as SMS Sent is high.



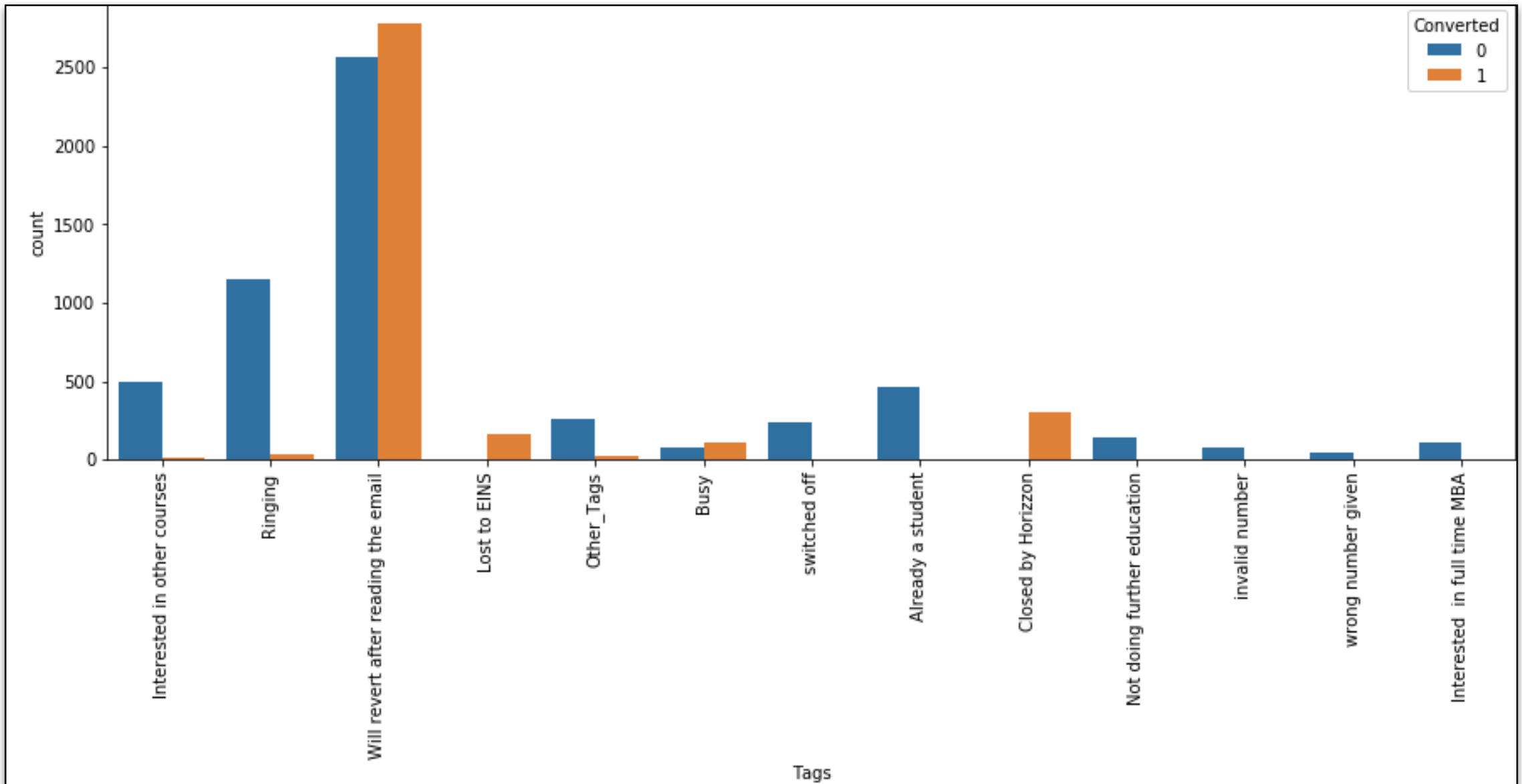
Inference

1. Working Professionals going for the course have high chances of joining it.
2. Unemployed leads are the most in numbers but has a less conversion rate.

Focus should be on increasing leads on Working Professional and converting more unemployed leads

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

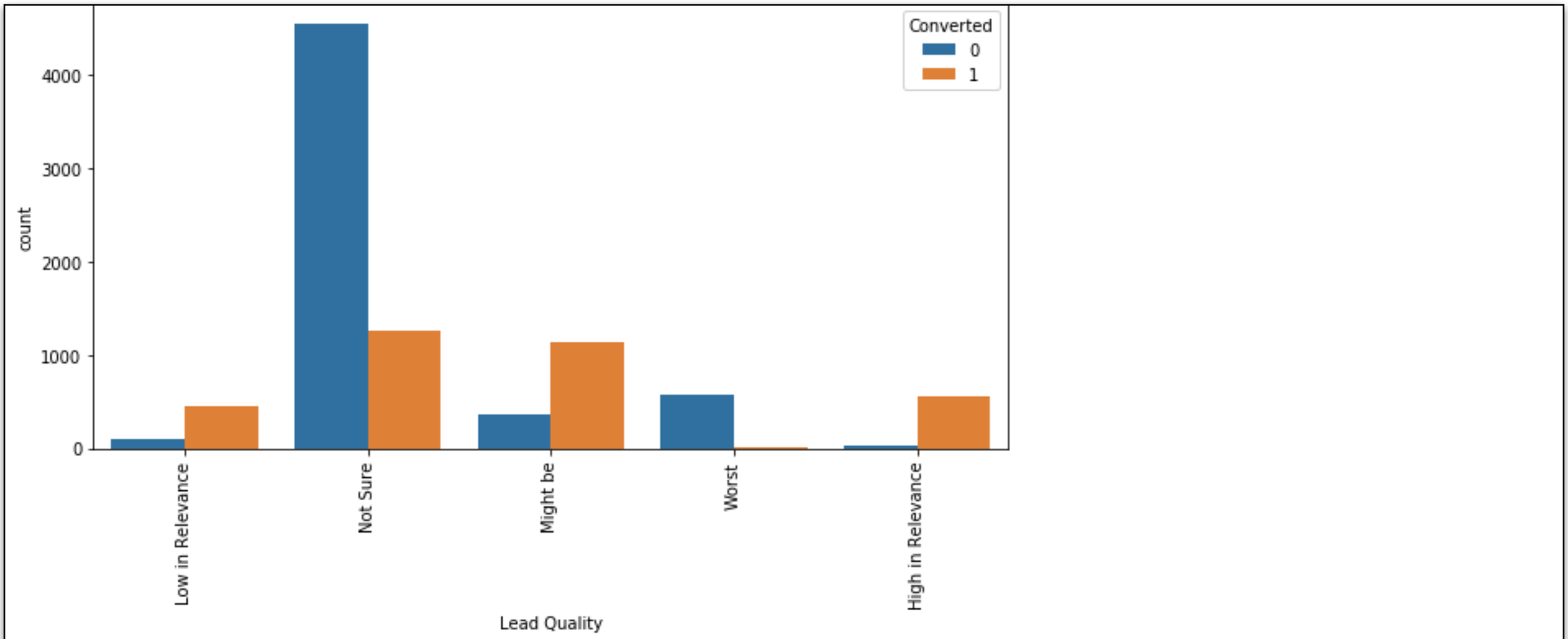
Tags



Inferences:

we see that the tag 'will revert after reading the email' has the highest conversion rate and a large number of leads too.
'Closed by Horizon' has a good amount of conversion rate we should focus on the lead numbers

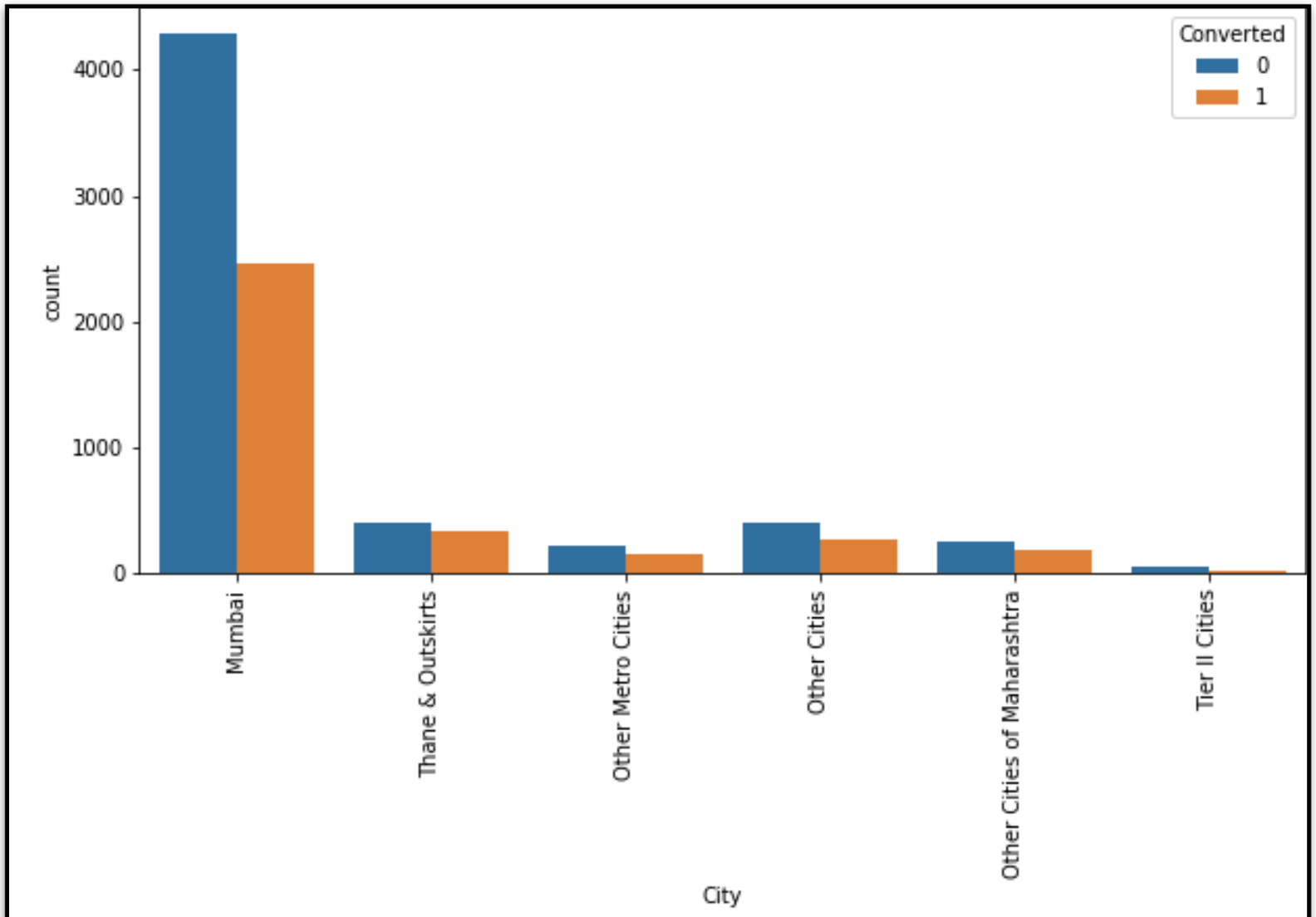
LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)



Inferences:

Most of the people are in the Not Sure bracket while the Might be bracket has a good conversion rate. Should focus on the lead numbers of 'Might be'.

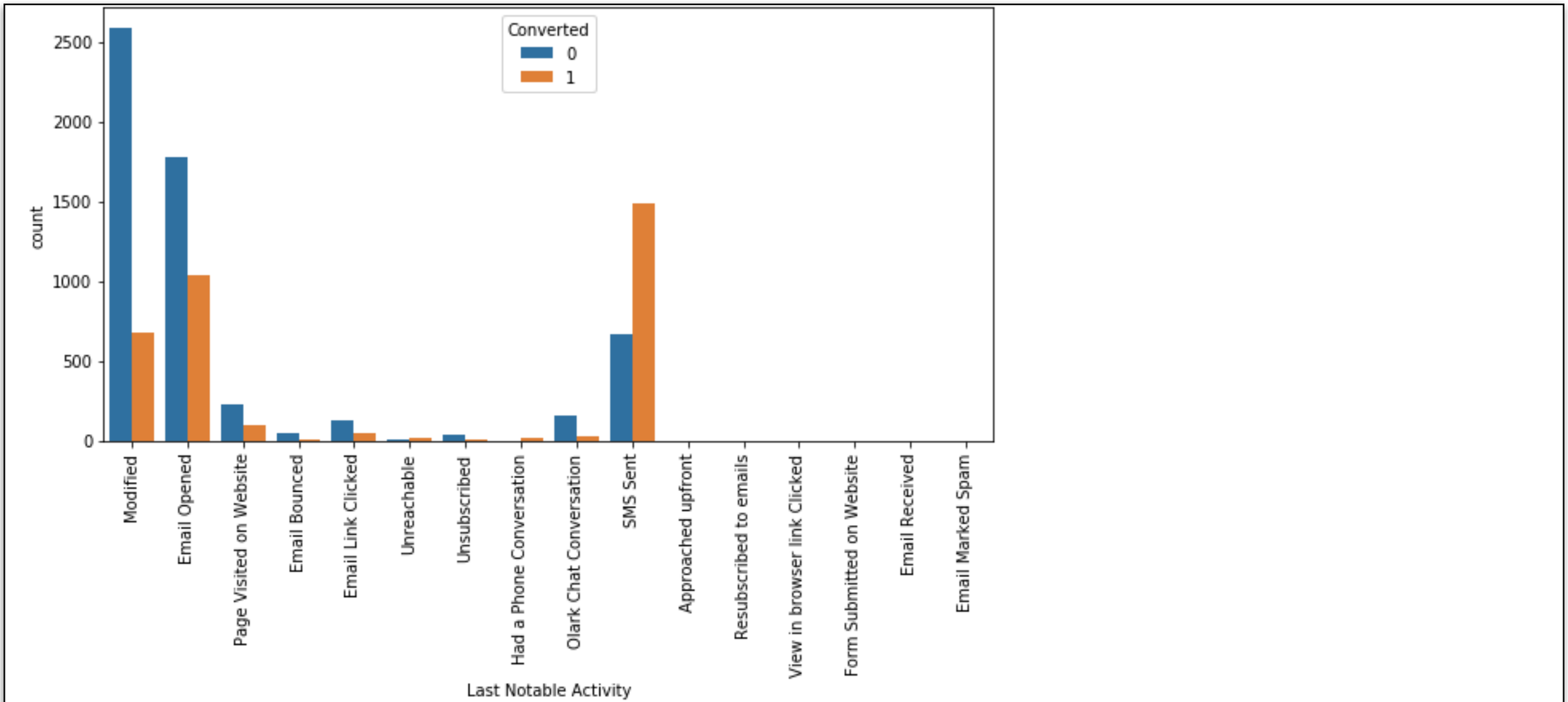
City



Inference

Most leads are from mumbai but conversion rate should be foucused on.

Last Notable Activity



Inference

Most leads are from Modified but very less conversion rate while SMS sent has good conversion rate . Should focus on the number of leads in SMS sent

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

I agree to pay the amount through cheque

Get updates on DM Content

Update me on Supply Chain Content

Receive More Updates About Our Courses

Digital Advertisement

Newspaper

Country

What matters most to you in choosing a course

Search

Magazine

X Education Forums

Newspaper Article

Through Recommendation

A free copy of Mastering The Interview

Rest of these columns were dropped because the values stored in this variables were mostly dominated by one kind of value which would refer nothing in terms of analysis of these columns

**Prospect ID', 'Lead Origin', 'Lead Source', 'Do Not Email',
'Do Not Call', 'Converted', 'TotalVisits',
'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity',
'Specialization', 'What is your current occupation', 'Tags',
'Lead Quality', 'City', 'Last Notable Activity'**

These were the columns that were selected.

MODEL BUILDING

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

- After the Data preparation of these variables were done which included creating binary mapping for yes/no variables and creating dummies for higher level categorical variable.
- The data was left with 87 variables.
- The dataset was spliced into Training Data and Testing Data, then the training data was scaled.
- We used Recursive Feature elimination (RFE) to select only the top 15 variables and build our model

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

The summary report is as follows:

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6335
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1580.6
Date:	Sat, 16 Nov 2019	**Deviance:**	3161.3
Time:	13:47:42	**Pearson chi2:**	3.11e+04
No. Iterations:	24	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.8547	0.215	-8.636	0.000	-2.276	-1.434
Do Not Email	-1.3106	0.213	-6.154	0.000	-1.728	-0.893
Lead Origin_Lead Add Form	1.0452	0.360	2.900	0.004	0.339	1.752
Lead Source_Welingak Website	3.4638	0.817	4.238	0.000	1.862	5.066
What is your current occupation_Working Professional	1.2843	0.287	4.476	0.000	0.722	1.847
Tags_Busy	3.5477	0.332	10.680	0.000	2.897	4.199
Tags_Closed by Horizzon	7.7377	0.762	10.152	0.000	6.244	9.231
Tags_Lost to EINS	8.9540	0.753	11.887	0.000	7.478	10.430
Tags_Ringing	-1.9696	0.340	-5.800	0.000	-2.635	-1.304
Tags_Will revert after reading the email	3.7332	0.228	16.340	0.000	3.285	4.181
Tags_invalid number	-23.4649	2.21e+04	-0.001	0.999	-4.34e+04	4.33e+04
Tags_switched off	-2.5711	0.589	-4.367	0.000	-3.725	-1.417
Tags_wrong number given	-23.0779	3.17e+04	-0.001	0.999	-6.21e+04	6.2e+04
Lead Quality_Not Sure	-3.3496	0.129	-26.033	0.000	-3.602	-3.097
Lead Quality_Worst	-3.7672	0.848	-4.445	0.000	-5.428	-2.106
Last Notable Activity_SMS Sent	2.7931	0.122	22.838	0.000	2.553	3.033

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

Dropping Tags_invalid number and Tags_wrong number given because of high P-value and rebuilding the model

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6337
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1588.8
Date:	Sat, 16 Nov 2019	**Deviance:**	3177.6
Time:	13:49:52	**Pearson chi2:**	3.08e+04
No. Iterations:	8	**Covariance Type:**	nonrobust

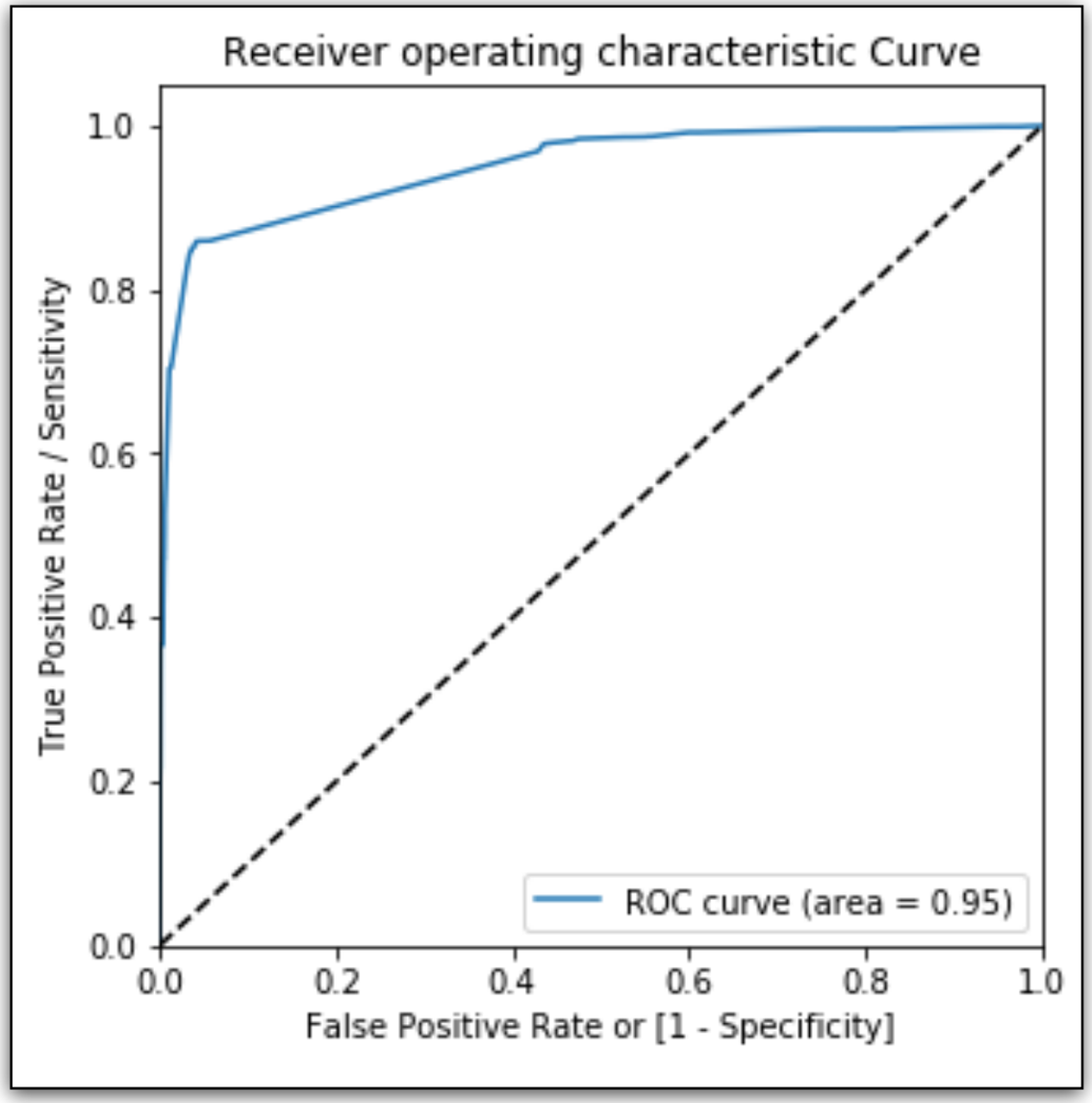
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0888	0.216	-9.654	0.000	-2.513	-1.665
Do Not Email	-1.3012	0.212	-6.134	0.000	-1.717	-0.885
Lead Origin_Lead Add Form	1.0894	0.363	3.001	0.003	0.378	1.801
Lead Source_Welingak Website	3.4138	0.818	4.173	0.000	1.810	5.017
What is your current occupation_Working Professional	1.3403	0.291	4.602	0.000	0.769	1.911
Tags_Busy	3.8040	0.330	11.532	0.000	3.157	4.450
Tags_Closed by Horizzon	7.9562	0.763	10.433	0.000	6.461	9.451
Tags_Lost to EINS	9.1785	0.754	12.177	0.000	7.701	10.656
Tags_Ringing	-1.6947	0.337	-5.036	0.000	-2.354	-1.035
Tags_Will revert after reading the email	3.9665	0.229	17.311	0.000	3.517	4.416
Tags_switched off	-2.2882	0.587	-3.900	0.000	-3.438	-1.138
Lead Quality_Not Sure	-3.3406	0.128	-26.026	0.000	-3.592	-3.089
Lead Quality_Worst	-3.7624	0.850	-4.426	0.000	-5.428	-2.096
Last Notable Activity_SMS Sent	2.7406	0.120	22.847	0.000	2.506	2.976

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

ROC Curve

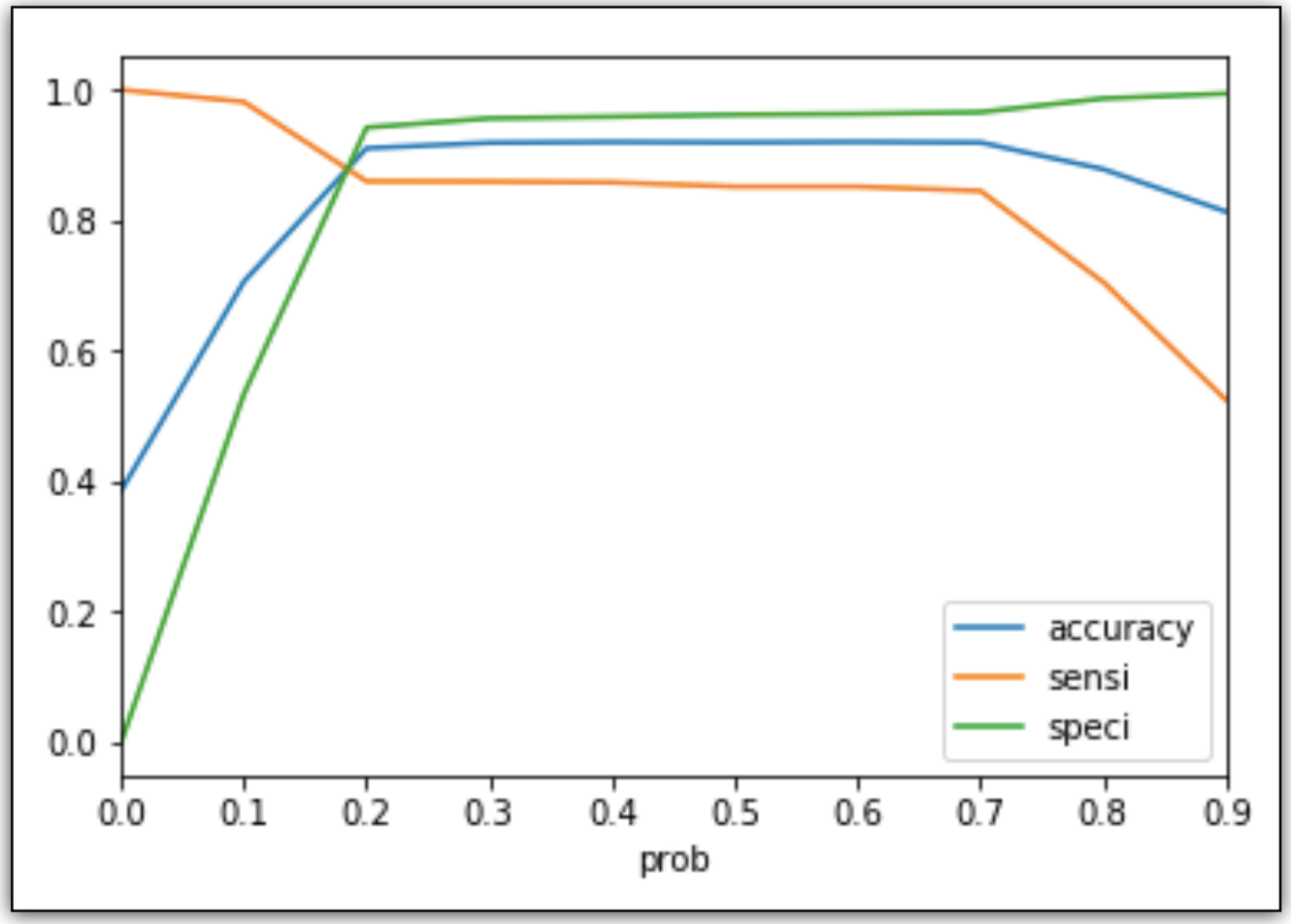
An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- ROC curve shows us the different values of sensitivity and specificity for different threshold values so that we ca



95% of the area is covered by the ROC

Plotting accuracy, sensitivity, and specificity for various probabilities



From the curve, 0.2 is the optimum point to take it as a cutoff probability.

TRAIN SET

Confusion Matrix

Predicted → Actual ↓	Not Churn	Churn
Not Churn	3679 (TN)	226 (FP)
Churn	343 (FN)	2103 (TP)

TP	True Positives
TN	True Negatives
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
PPV	Positive Predicted Value
NPV	Negative Predicted Value

Sensitivity	Specificity	FPR	PPV	NPV
0.8597	0.9421	0.0578	0.9029	0.9147

Precision	Recal
0.9332	0.8515

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

TEST SET

Confusion Matrix

Predicted → Actual ↓	Not Churn	Churn
Not Churn	1635 (TN)	99 (FP)
Churn	155 (FN)	834 (TP)

TP	True Positives
TN	True Negatives
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
PPV	Positive Predicted Value
NPV	Negative Predicted Value

Sensitivity	Specificity	FPR	PPV	NPV
0.8597	0.9421	0.0578	0.9029	0.9147

86% of the values (actual converted) are predicted by the model
94% of the values (actual not converted) are predicted by the model

Precision	Recal
0.8938	0.8432

LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

RESULTS

The top three variables which contribute most towards the probability of a lead getting converted are :

- **Tags,**
- **Lead Source,**
- **Lead Quality.**

The top 3 categorical/dummy variables which should be focused the most in order to increase the probability of lead conversion are:

- **Tags_Lost to EINS,**
- **Tags_Closed by Horizzon**
- **Tags_Will revert after reading the email**

- Focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- Focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
- Websites should be made engaging yet simple to attract customers to spend Tim on it.

THANK YOU