LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

-POOJA AGRAWAL

NEHA SHUKLA

PROBLEM STATEMENT

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 Us To Help Them Select The Most Promising Leads, I.E. The Leads That Are Most Likely To Convert Into Paying Customers. The Company Requires Us 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. The CEO, In Particular, Has Given A Ballpark Of The Target Lead Conversion Rate To Be Around 80%.

GOALS OF THE 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.

The data provided (leads dataset) are from the past with around 9000 data points. 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.

DATA INSPECTION & CLEANING

The dataset has 9240 records and 37 Variables

No duplicate values are available

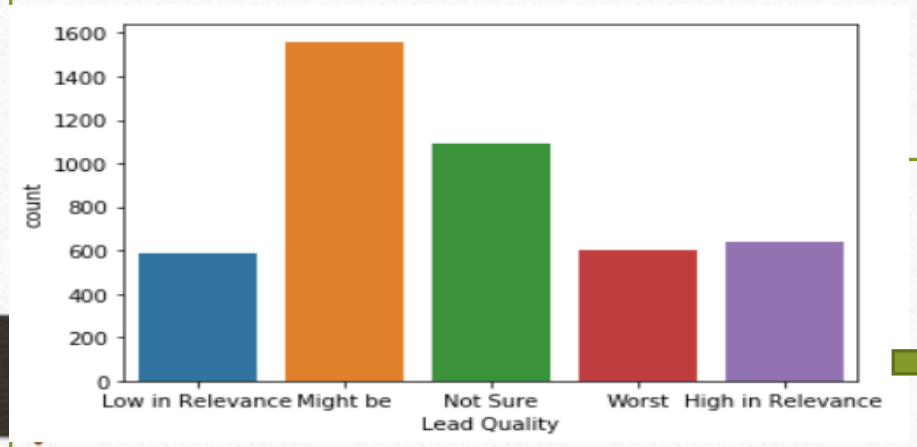
There were many cells which had 'Select' as their values which meant the customer did not select any of the available choices. Hence those values were converted into NULL values.

Any column having missing value percentage above 70 % was removed. Eg. Lead Profile.

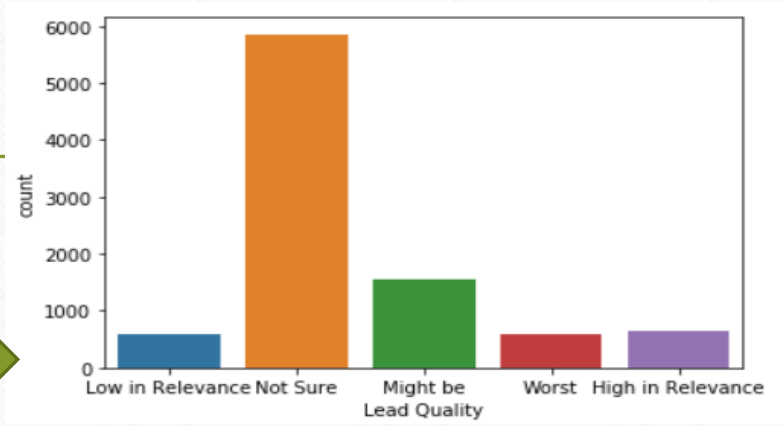
Rest of the columns which have missing value percentage more than 2 % were looked upon individually and the missing values were imputed.

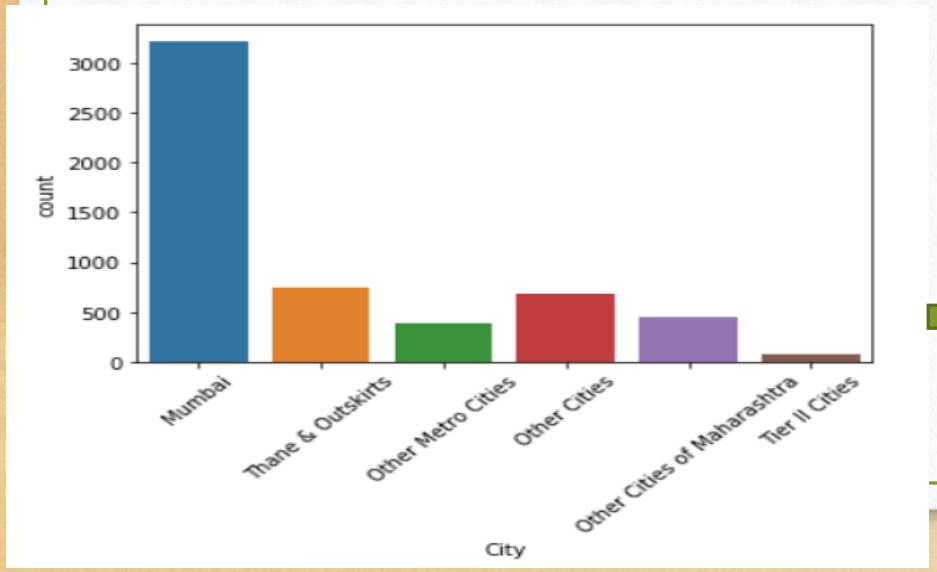
Prospect ID	0.00
Lead Number	0.00
Lead Origin	0.00
Lead Source	0.39
Do Not Email	0.00
Do Not Call	0.00
Converted	0.00
TotalVisits	1.48
Total Time Spent on Website	0.00
Page Views Per Visit	1.48
Last Activity	1.11
Country	26.63
Specialization	15.56
How did you hear about X Education	23.89
What is your 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	29.32
City	15.37
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

Missing Values Imputation

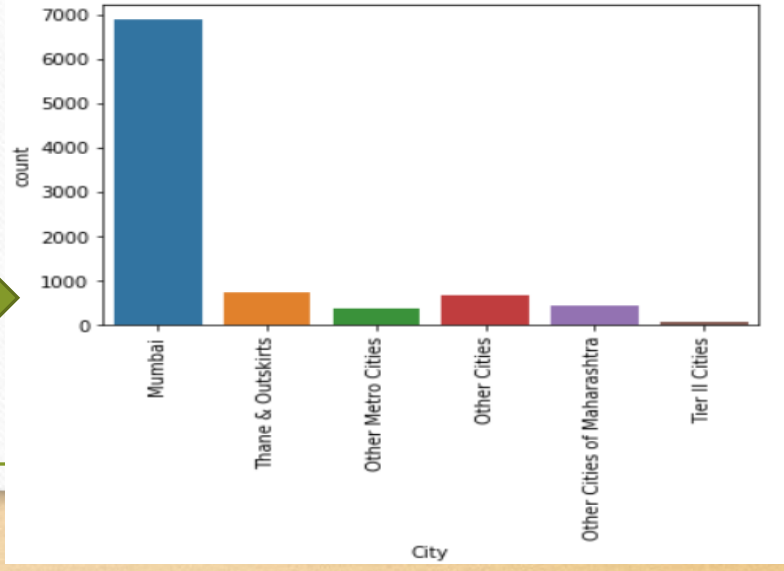


Lead quality is based on customer intuition. Imputed missing values with 'Not Sure'

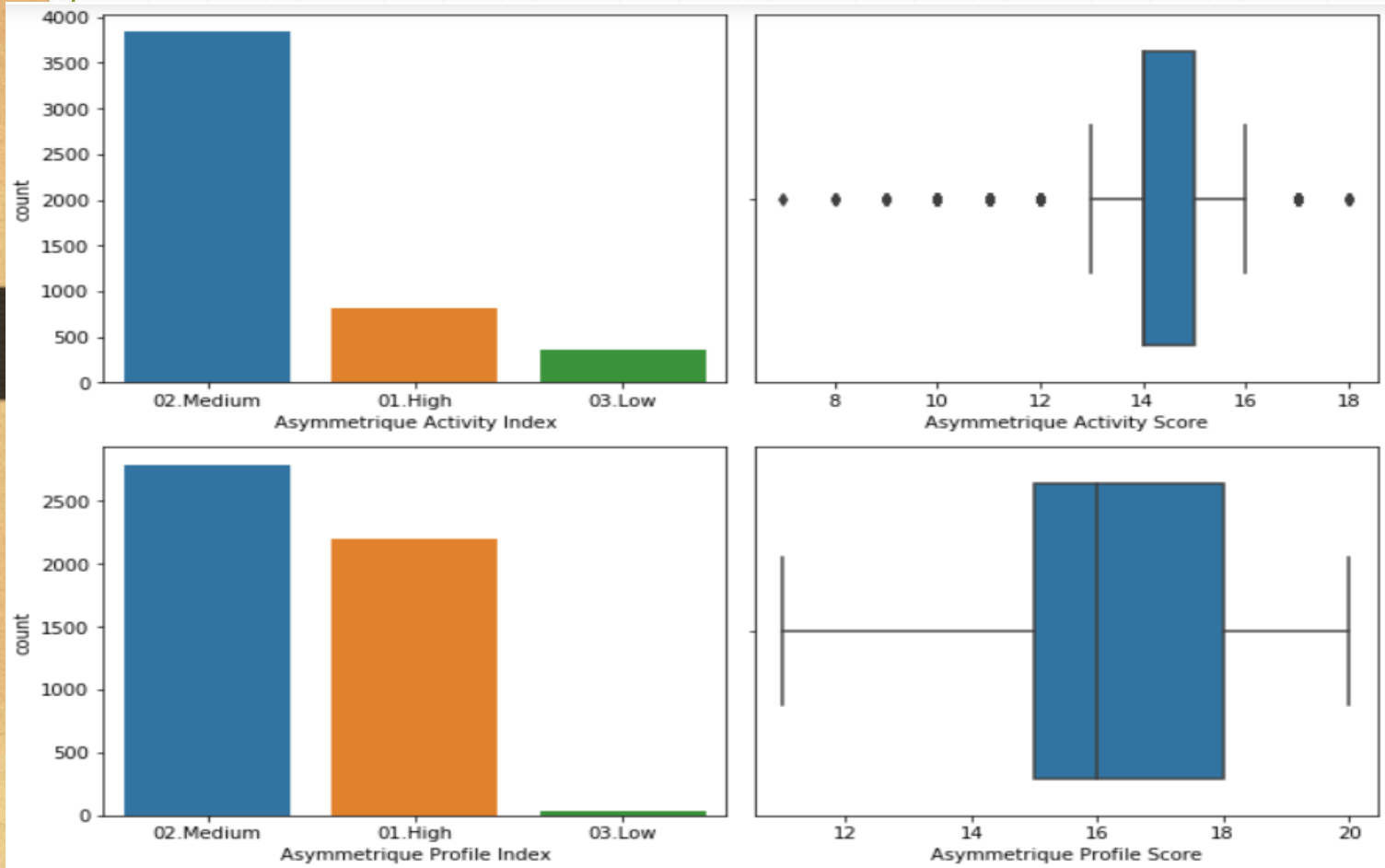




60% of the **City** column has Mumbai in it. Imputed the missing values to Mumbai

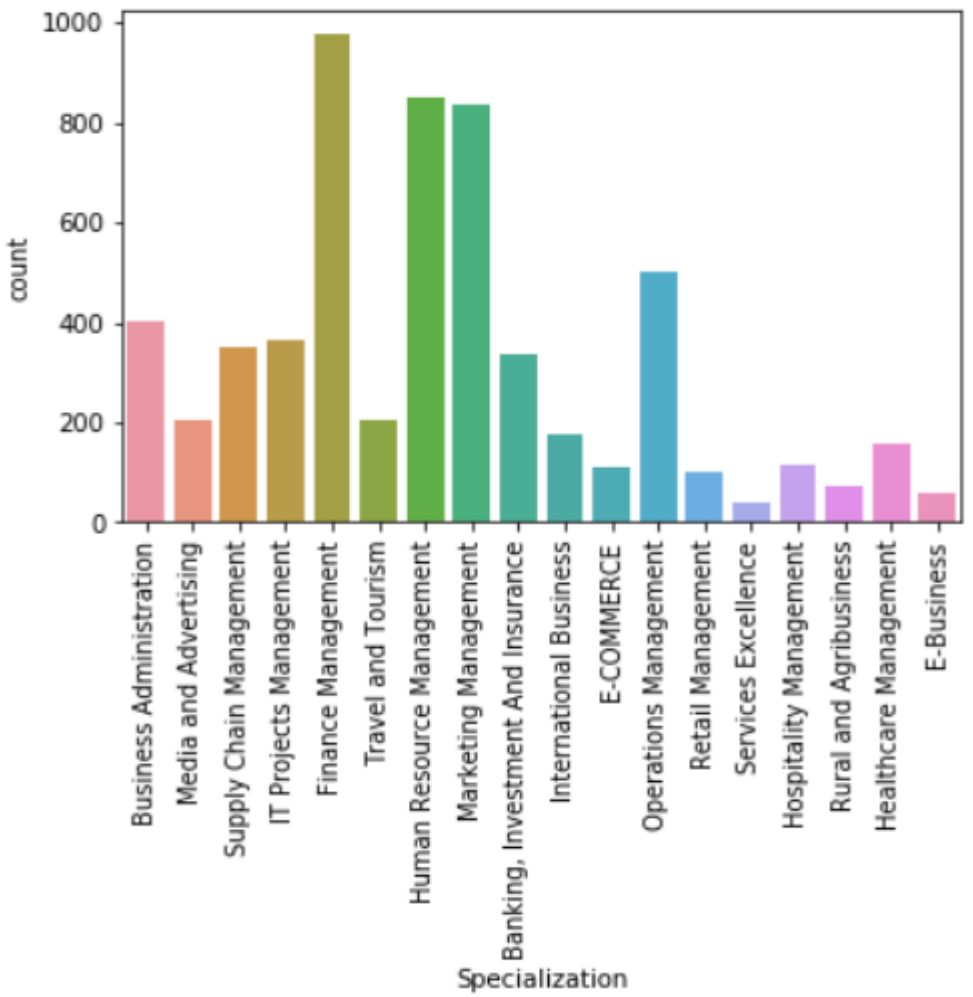


Asymmetrique Activity Index(45.65%) Asymmetrique Activity Score(45.65%) Asymmetrique Profile Index(45.65%) Asymmetrique Profile Score(45.65%)

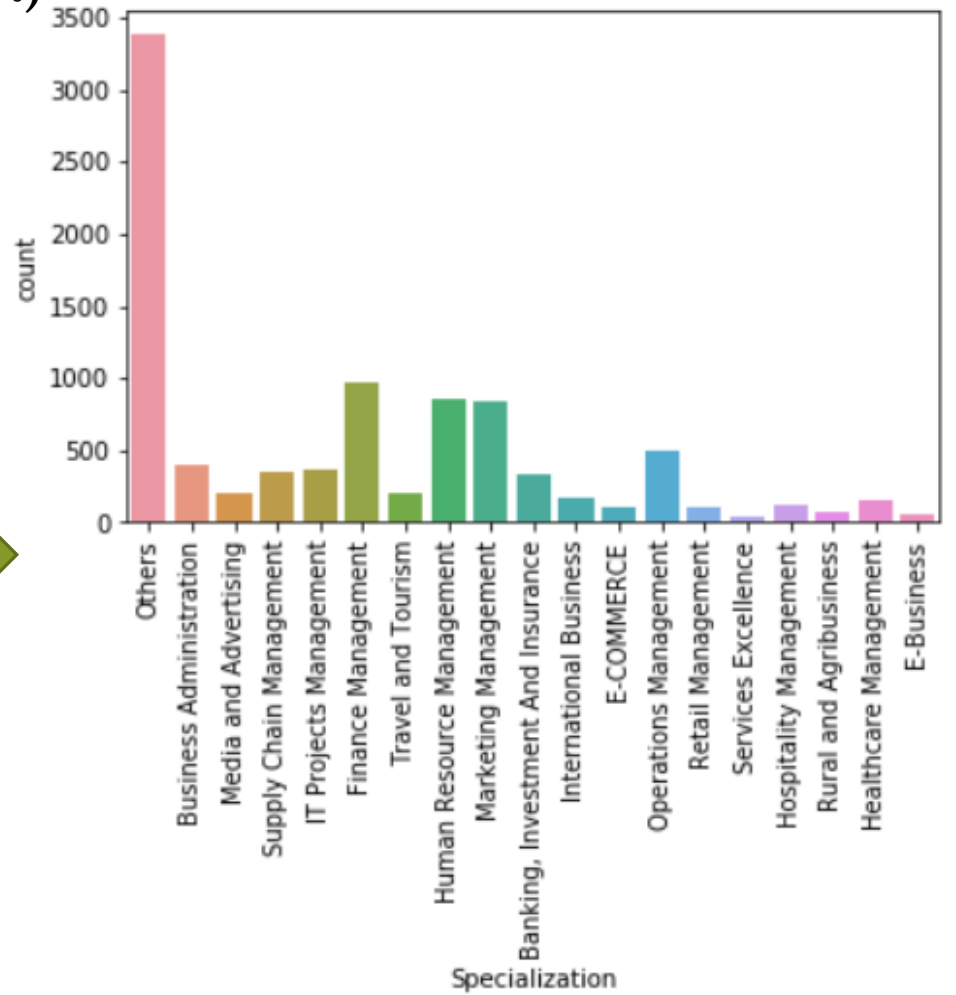


There is too much variation in these parameters so its not reliable to impute any value in it. 45% null values means we need to drop these columns.

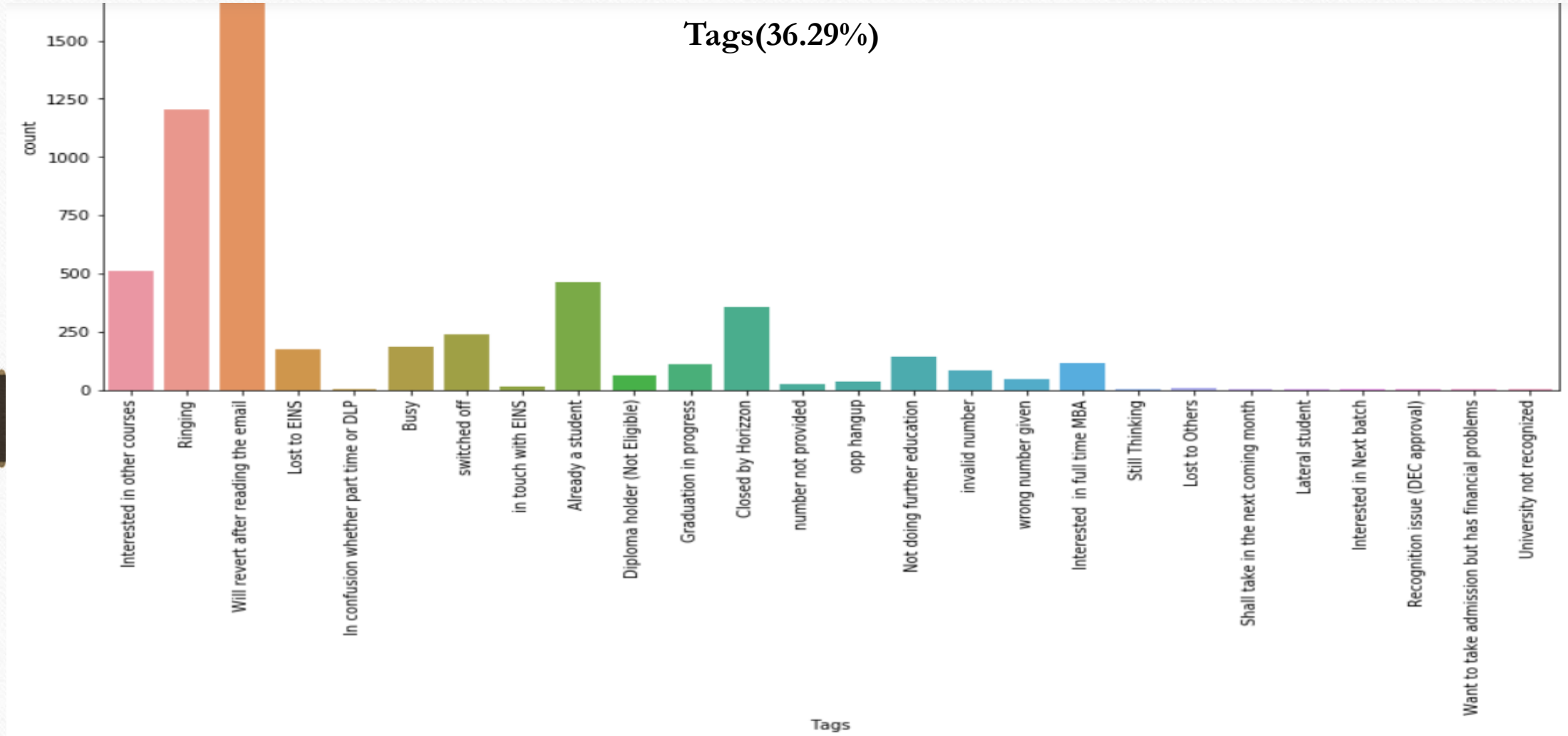
Specialization(36.58%)



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'

'What is your current occupation' (29 %)

count	6550
unique	6
top	Unemployed
freq	5600

86% entries are unemployed. So, imputed missing value with "Unemployed"

Country(26.63%)

count	6779
unique	38
top	India
freq	6492

For most values, country is India. So imputing the missing values with "India"

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

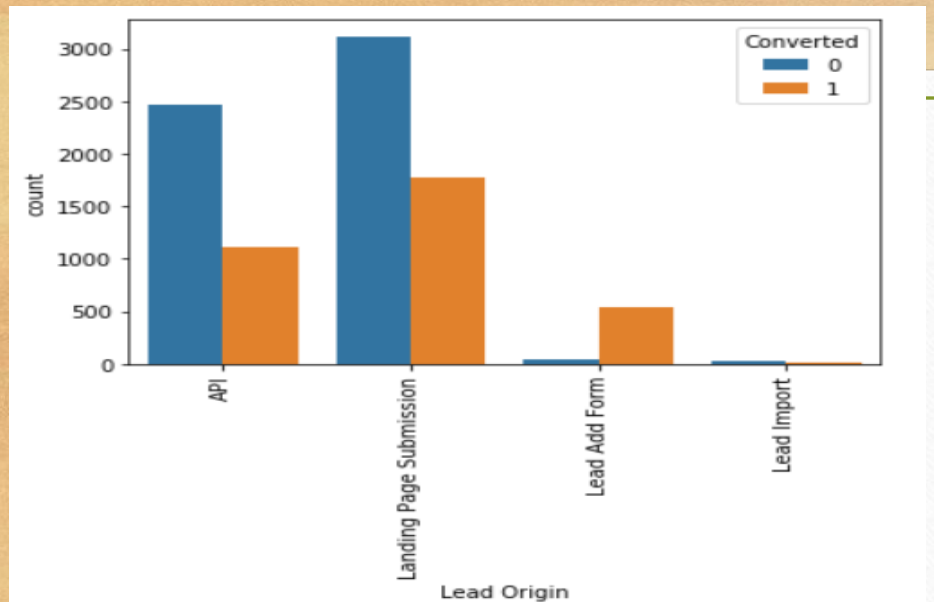
count	6531
unique	3
top	Better Career Prospects
freq	6528

Blanks in this column can be imputed by "Better Career Prospects"

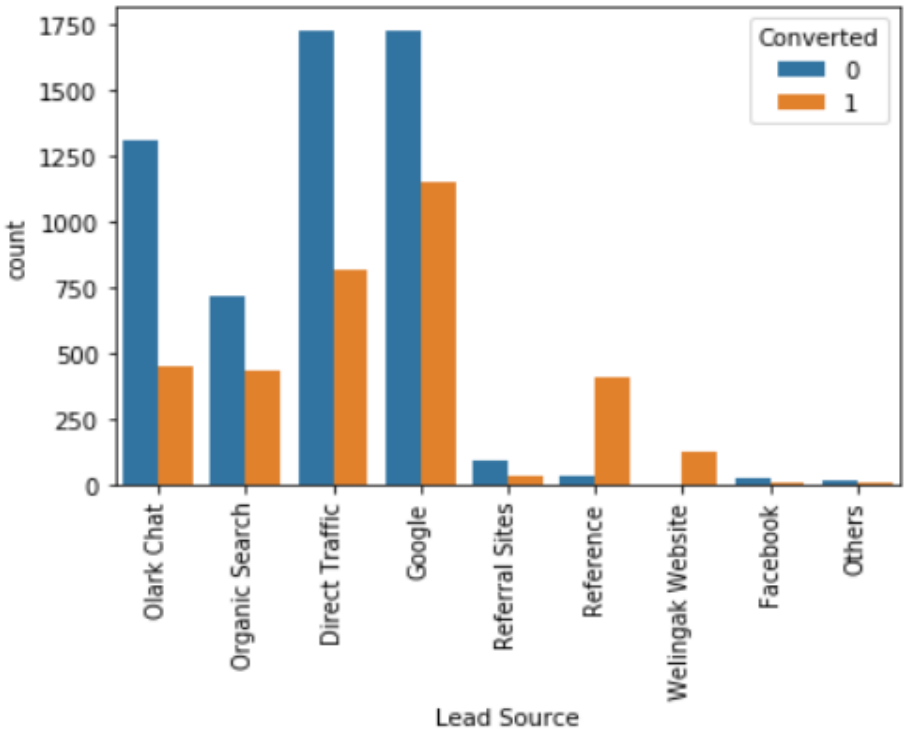
Prospect ID	0.00
Lead Number	0.00
Lead Origin	0.00
Lead Source	0.39
Do Not Email	0.00
Do Not Call	0.00
Converted	0.00
TotalVisits	1.48
Total Time Spent on Website	0.00
Page Views Per Visit	1.48
Last Activity	1.11
Country	0.00
Specialization	0.00
What is your current occupation	0.00
What matters most to you in choosing a course	0.00
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	0.00
Lead Quality	0.00
Update me on Supply Chain Content	0.00
Get updates on DM Content	0.00
City	0.00
I agree to pay the amount through cheque	0.00
A free copy of Mastering The Interview	0.00
Last Notable Activity	0.00

Rest missing values are under 2% so we have dropped these rows.

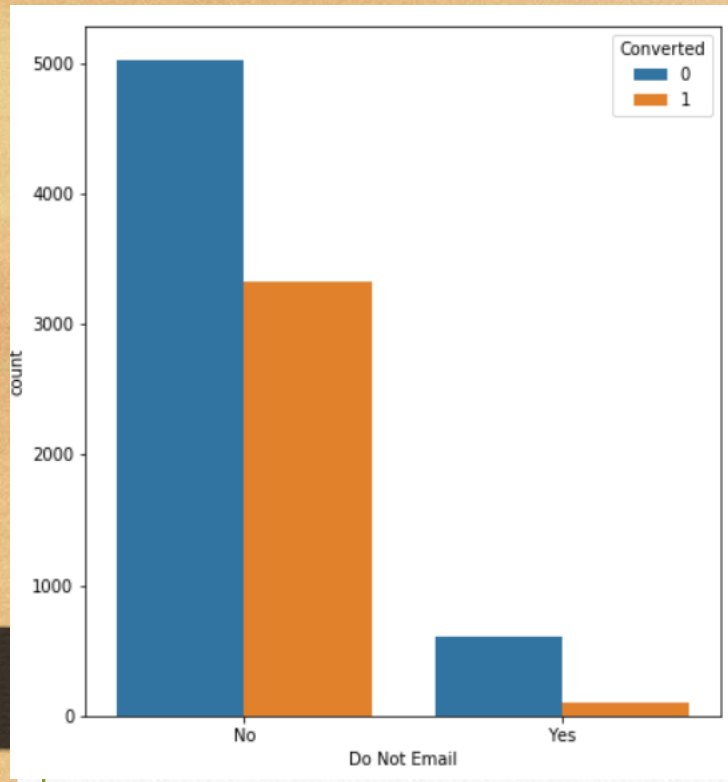
Prospect ID	0.0
Lead Number	0.0
Lead Origin	0.0
Lead Source	0.0
Do Not Email	0.0
Do Not Call	0.0
Converted	0.0
TotalVisits	0.0
Total Time Spent on Website	0.0
Page Views Per Visit	0.0
Last Activity	0.0
Country	0.0
Specialization	0.0
What is your current occupation	0.0
What matters most to you in choosing 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

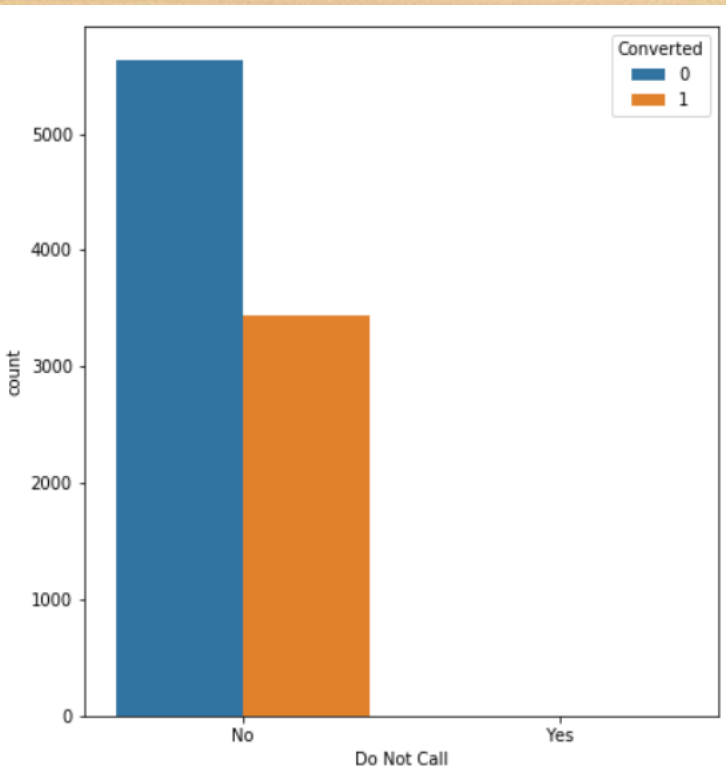


- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

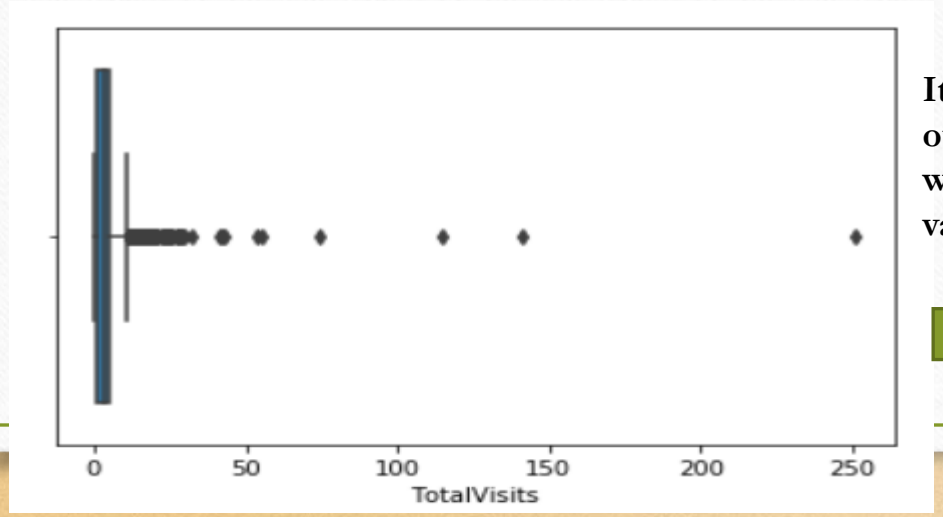


- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, we should focus on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

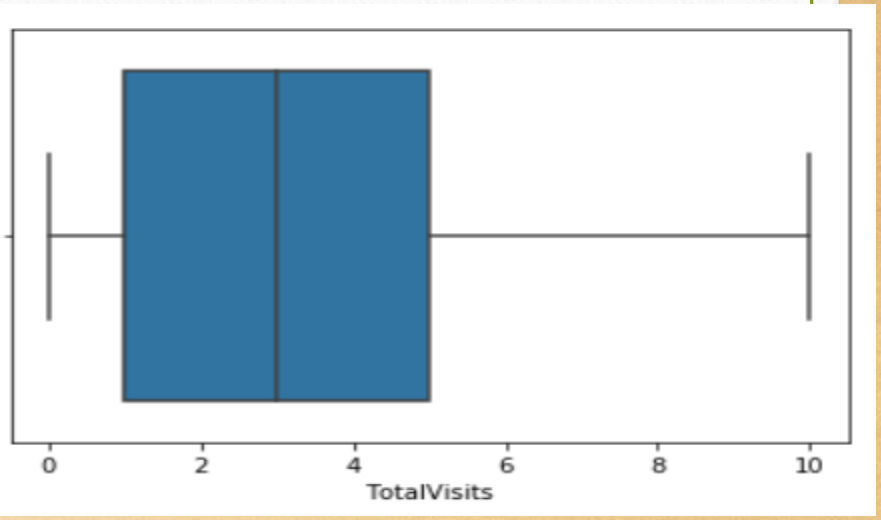


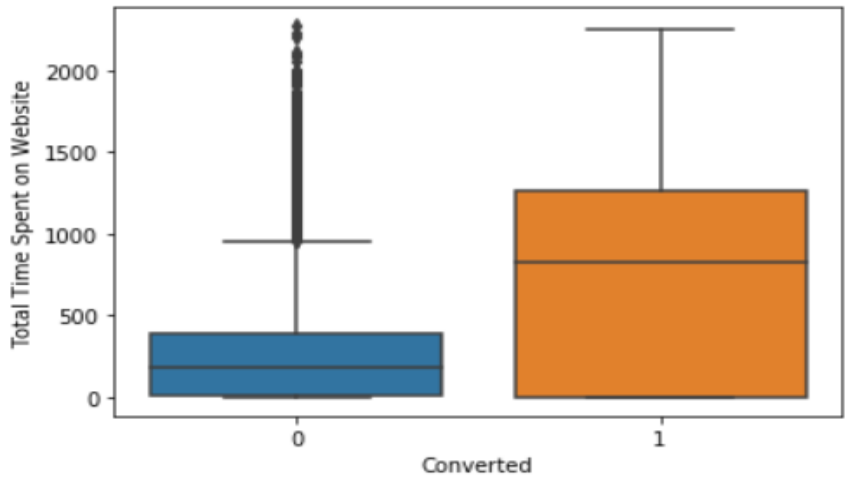


People who haven't opted for these services have a high conversion rate

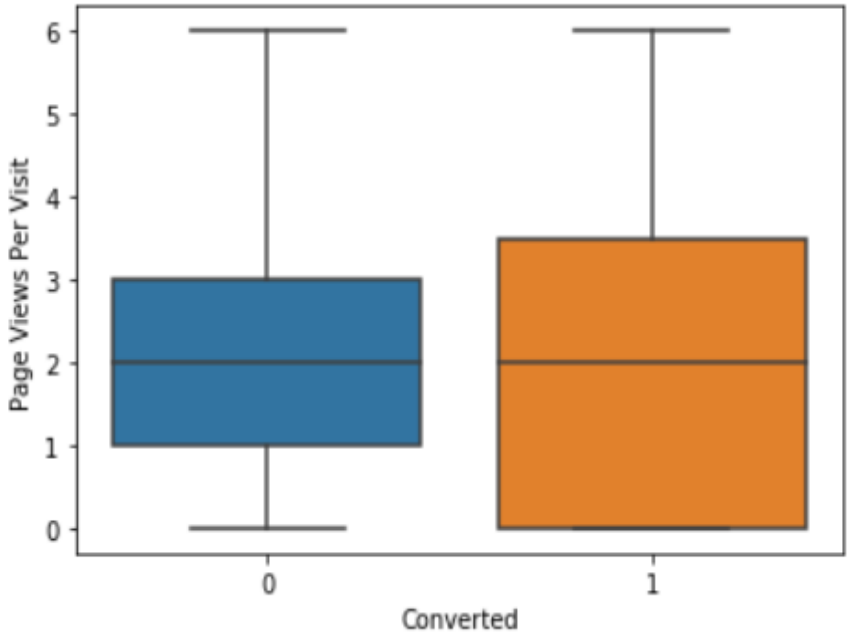


It's found that there are many outliers in the data. Therefore, we will cap the outliers to 95% value for analysis.

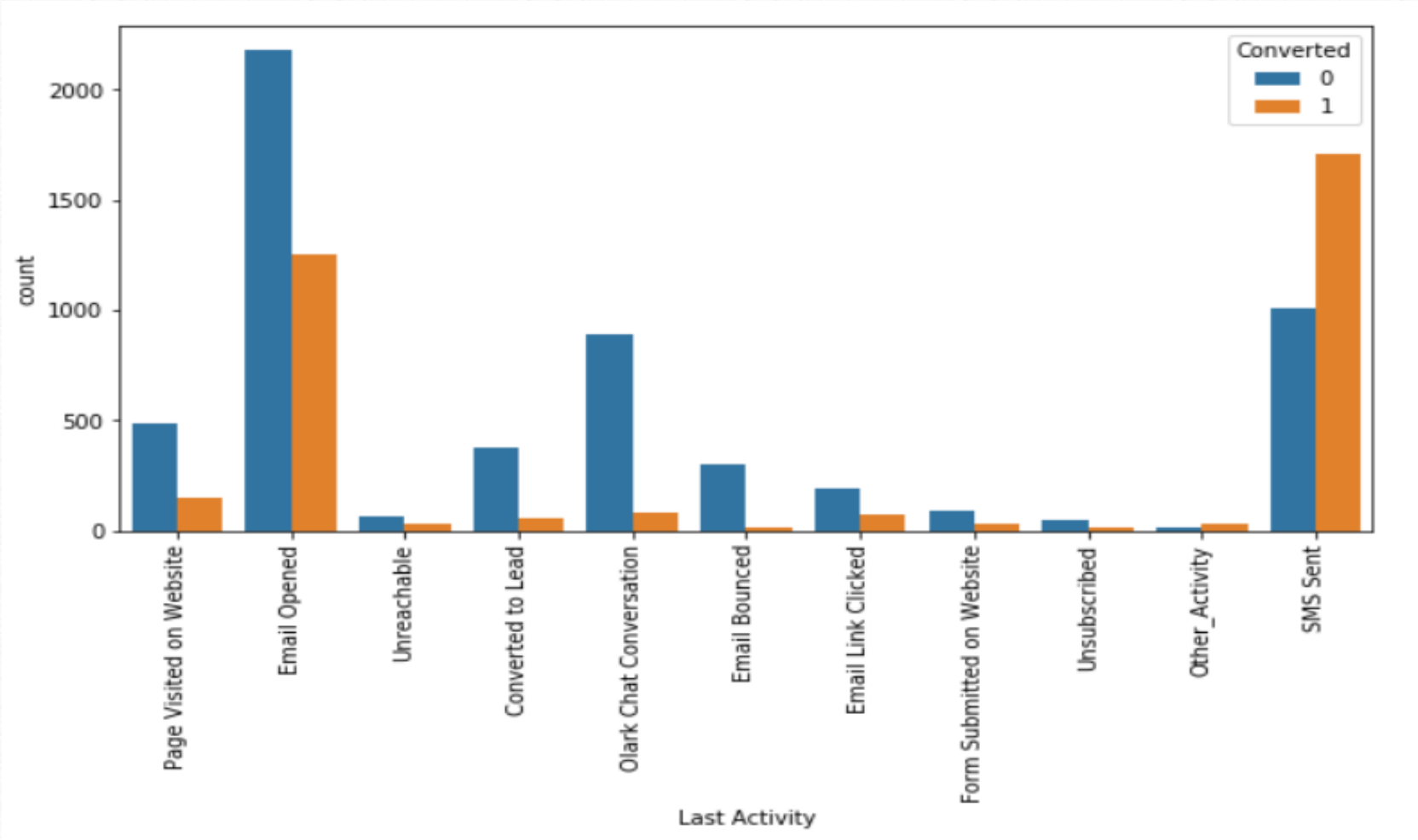




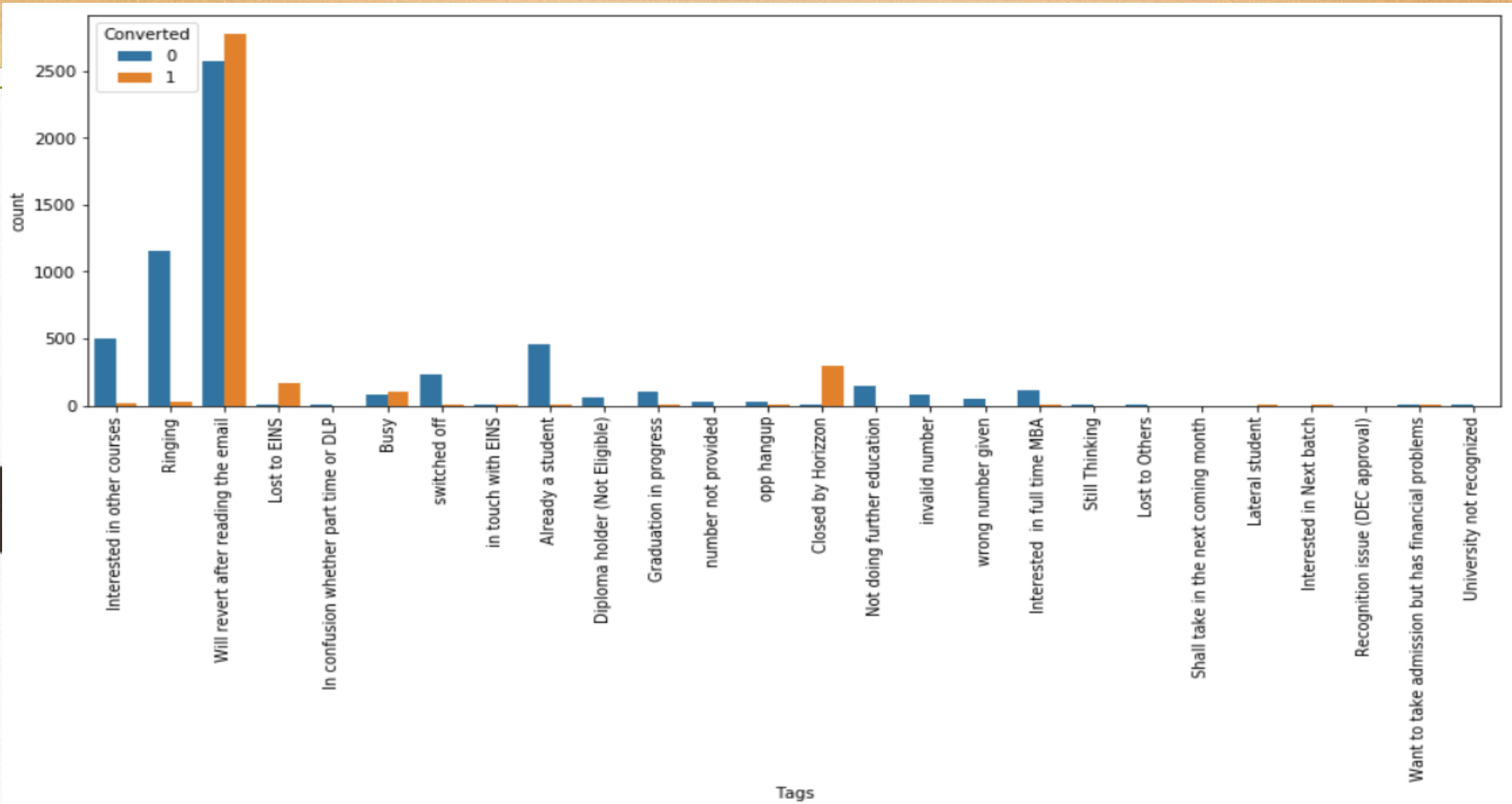
Leads spending more time on the website are more likely to be converted.



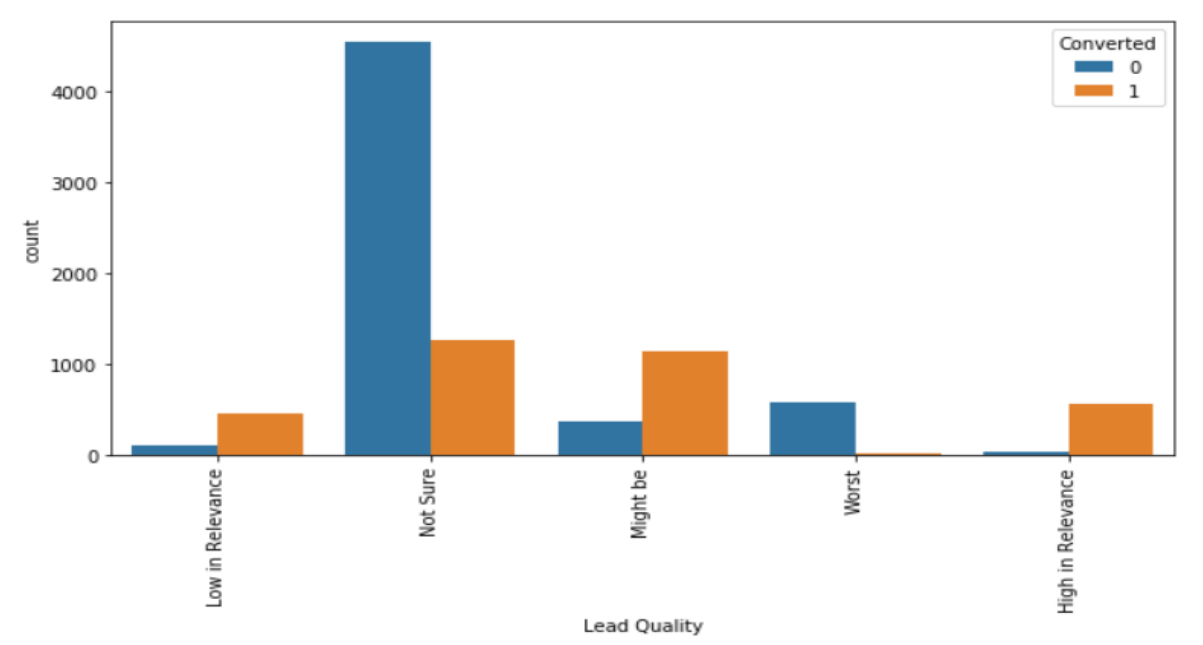
Median for converted and unconverted leads is the same. **Nothing can be said specifically for lead conversion from Page Views Per Visit**



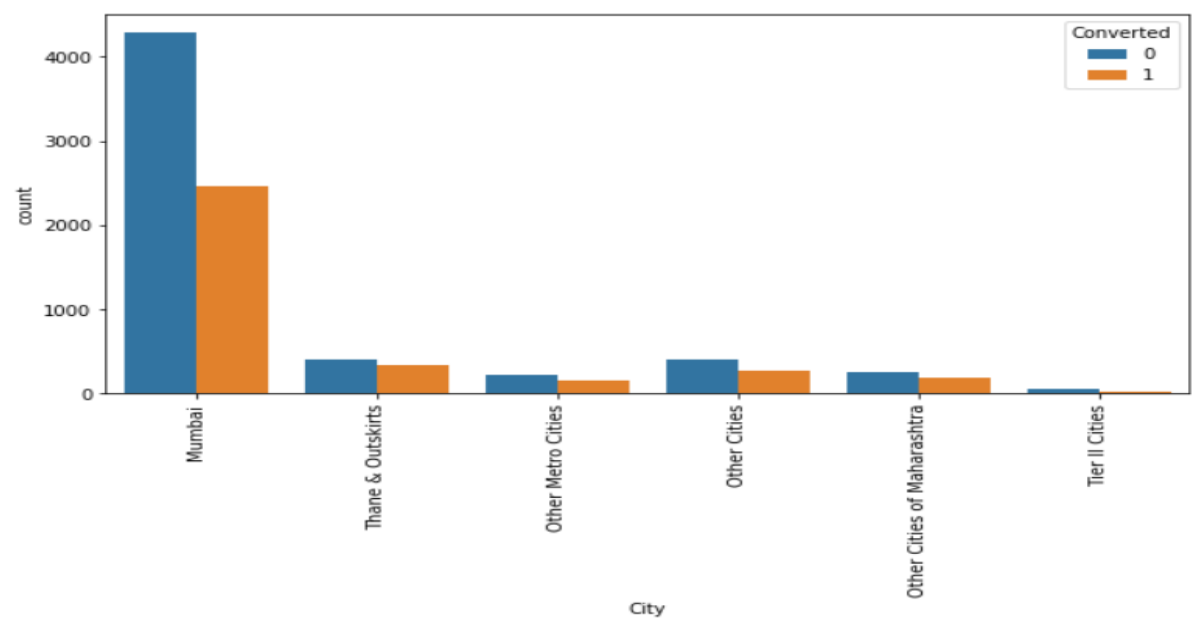
Most of the lead have their Email opened as their last activity.
Conversion rate for leads with last activity as SMS Sent is almost 60%.



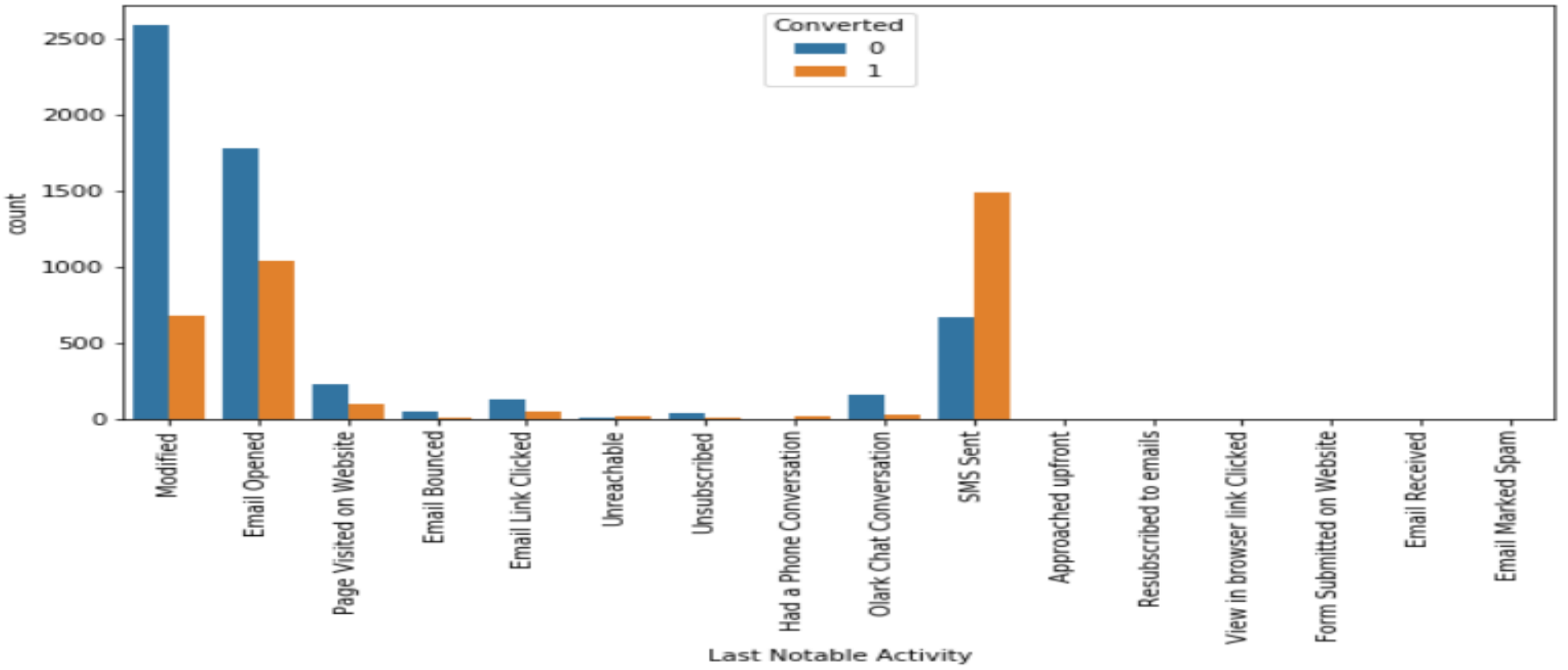
The Leads who revert after reading the email have the highest conversion rate



The conversion rate of “Might be” is highest



Most leads are from Mumbai with around 30% conversion rate



SMS Sent has the highest rate of conversion

it is observed that many columns(listed below) are not adding any information to the model. Therefore, we can drop them for further analysis

'Lead Number', 'What matters most to you in choosing a course','Search','Magazine','Newspaper Article', 'X Education Forums','Newspaper','Digital Advertisement',' Through Recommendations', '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', 'A free copy of Mastering The Interview', 'Country'

Data Preparation

- Converted some binary variables(Yes/No) to 1/0. For example, 'Do Not Email', 'Do Not Call'
- For categorical variables with multiple levels, created dummy features and dropped the respective original column
- The data set is left with 87 variables now

Test-Train Split

- Test-Train split was performed on data set by putting feature variables to X and response variable('Converted') to y

Feature Scaling

- We performed feature scaling on train data
- The conversion rate is 38%

MODEL BUILDING

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

Dropping tags_invalid number and tags_wrong number given because of high p-value and rebuilding the MODEL

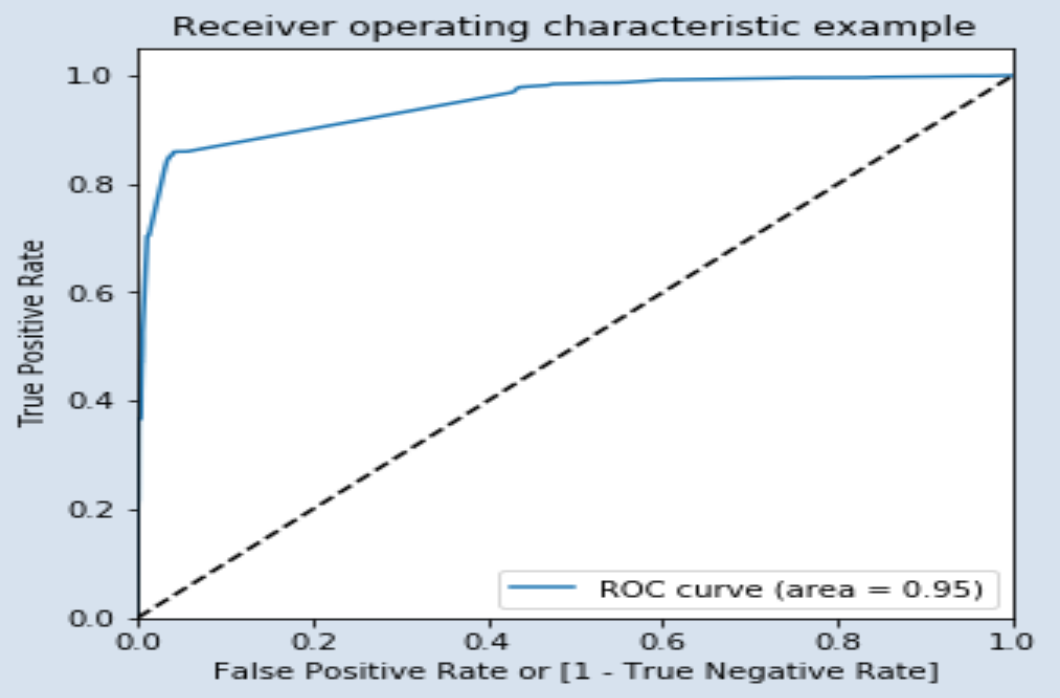
Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6337
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1588.8
Date:	Sat, 16 Nov 2019	**Deviance:**	3177.6
Time:	13:49:52	**Pearson chi2:**	3.08e+04
No. Iterations: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

ROC CURVE

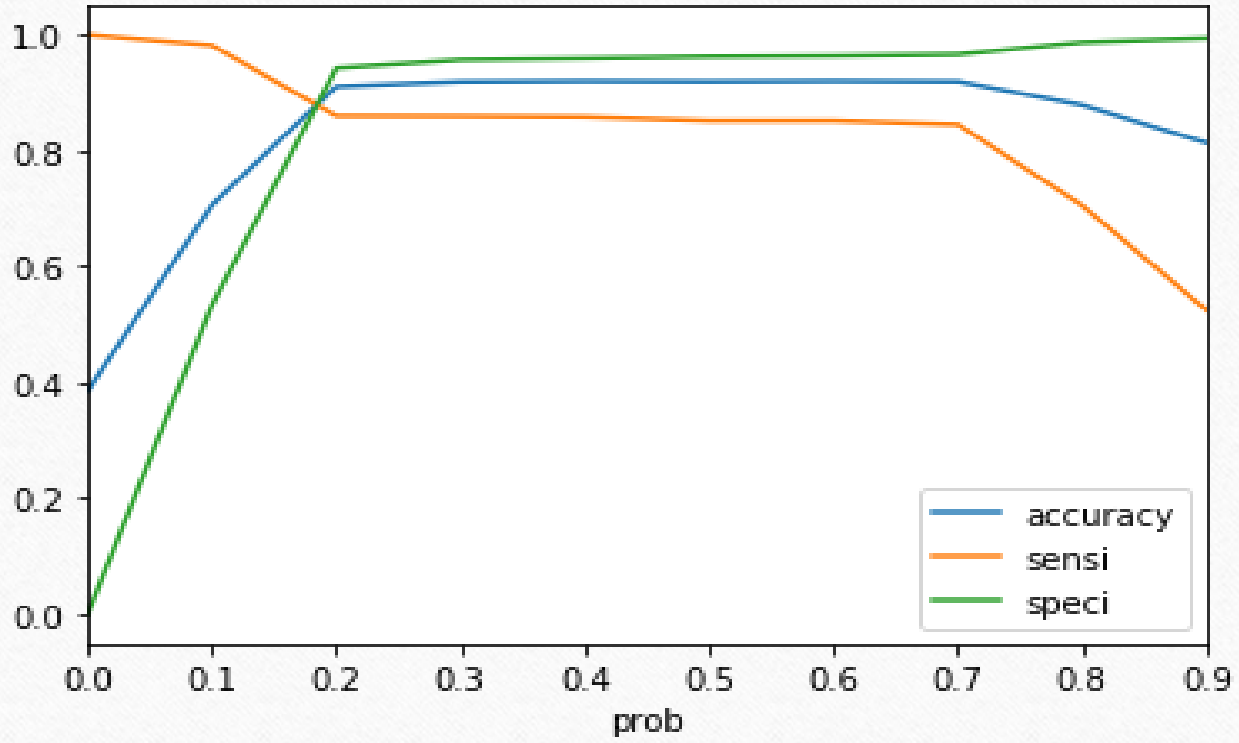
An ROC curve demonstrates several things:

- 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



95% of the area is covered by the ROC

Plotting accuracy, sensitivity, and specificity for various probabilities



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

Metrics on 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

Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
0.9104	0.8597	0.9421	0.9029	0.8597	0.8808

Precision Recall

Metrics on 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

Accuracy	Sensitivity	Specificity	Precision	Recall	F1 Score
0.9067	0.8432	0.9429	0.8938	0.8432	0.8678

84% of the values (actual converted) are predicted by the model
94% of the values (actual not converted) are predicted by the model

CONCLUSION

- 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 should be 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 Time on it.

Thank You…