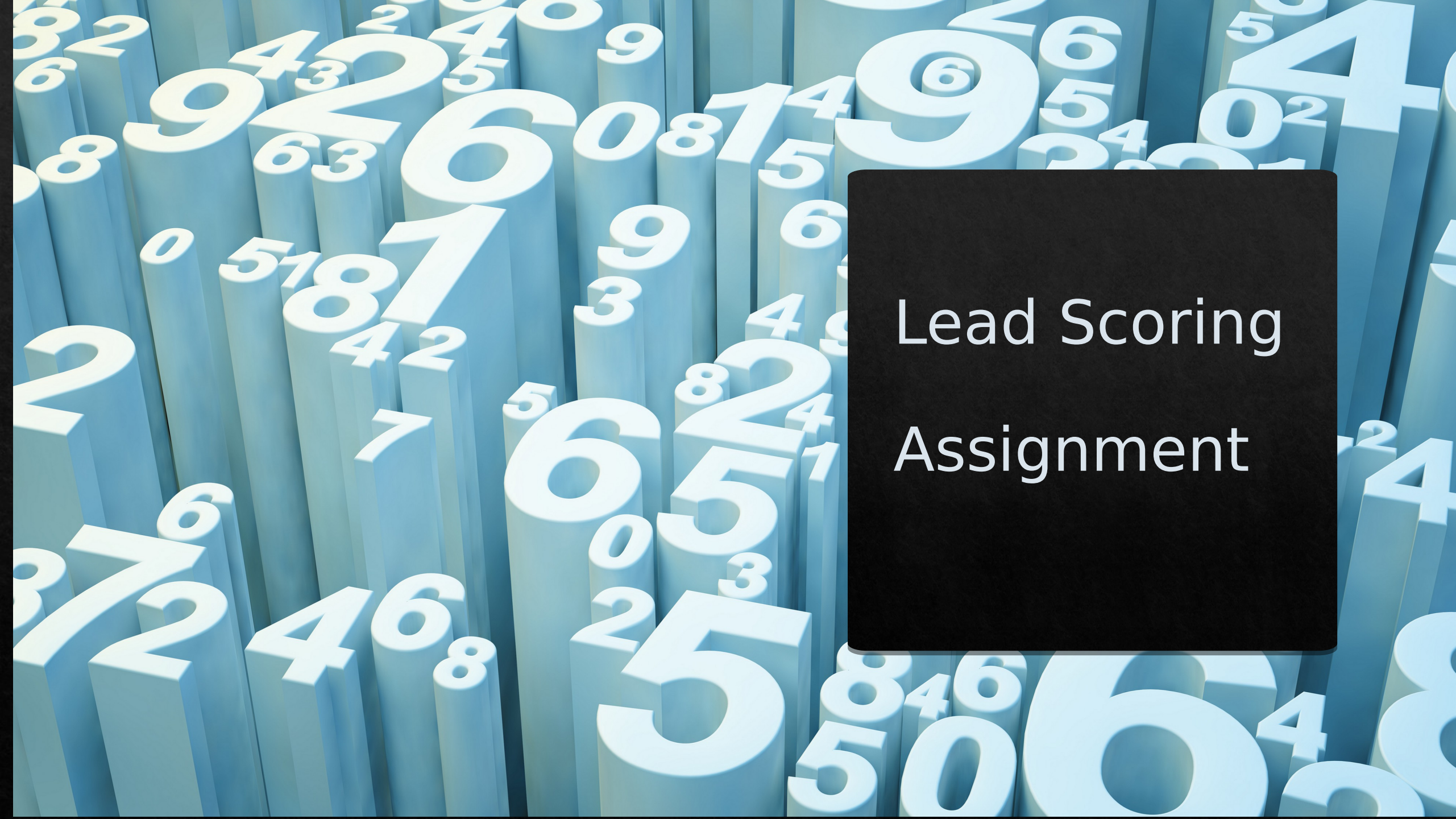

Lead Scoring
Assignment

BUSINESS PROBLEM UNDERSTANDING

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 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%. The company wishes to identify the most potential leads, also known as ‘Hot Leads’, The company requires to build a model and to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%

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.

Provided with a ‘leads’ dataset from the past with around 9000 data points which 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. We have inspected the data thoroughly and done the following process step by step

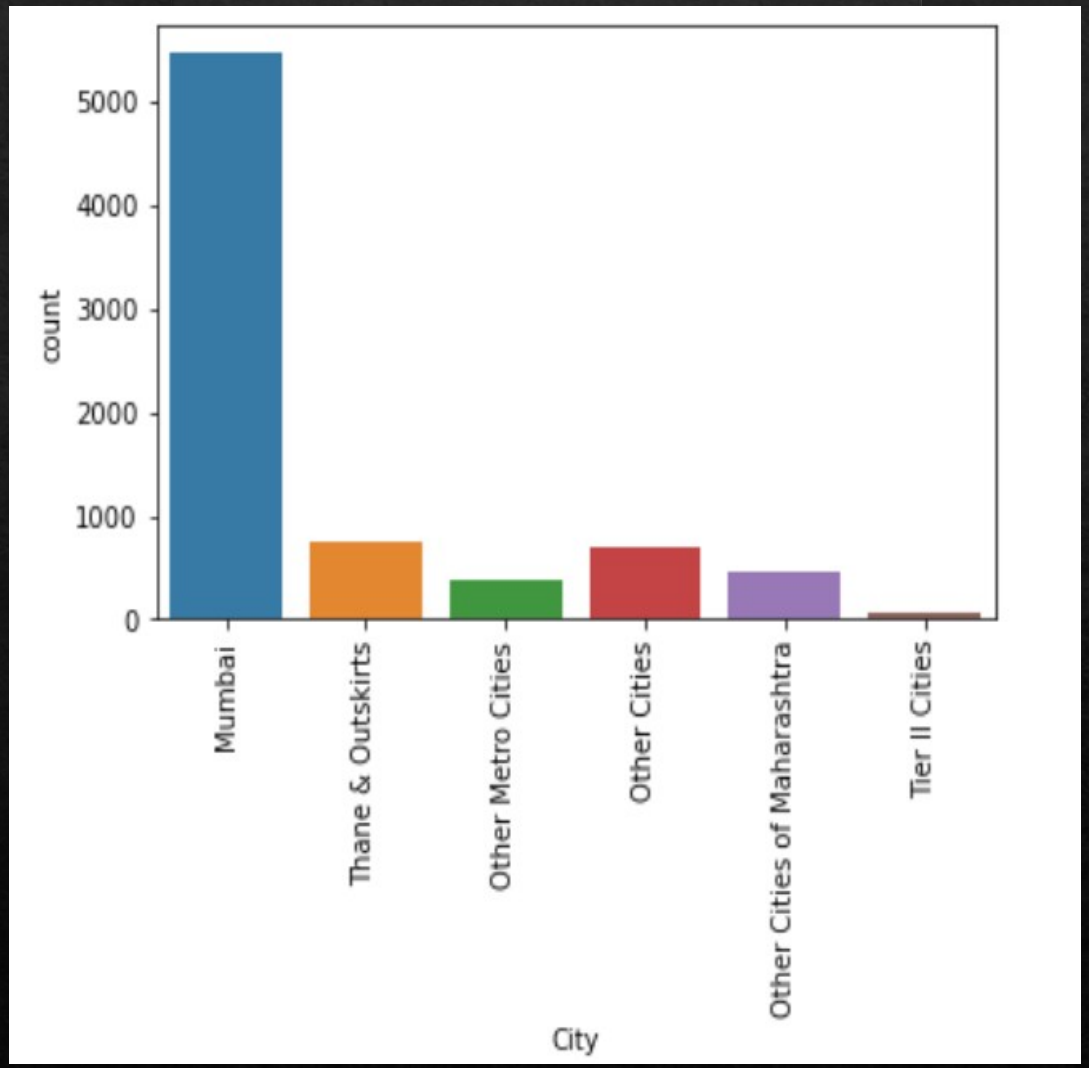
- Data Reading
- Cleaning data – for null values
- Outlier Handling
- EDA for better understanding of variables
- Model Building
- Evaluation of the model

City vs Leads

Insights on leads as per cities :

- The city Mumbai is having the highest leads
- Thane and outskirts along with other cities in Maharashtra are having leads

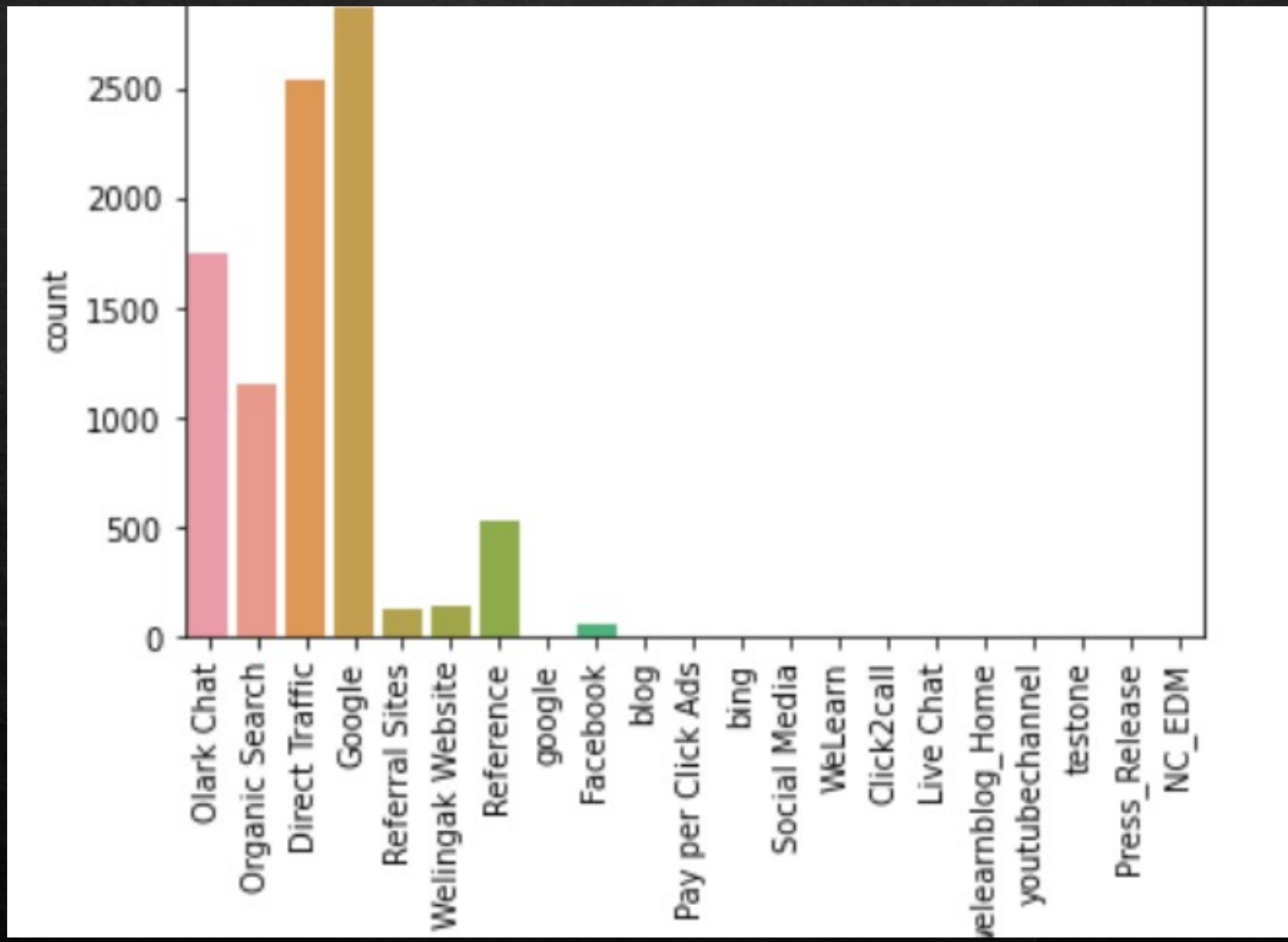
Since Mumbai has the highest number of leads among other cities, Mumbai should be targeted



Lead Source

Where do the leads come from ?

- As per the graph most of the leads are sourced from Google search
- It is followed by Olark Chat and Direct Traffic
- While other sources have seen no impact



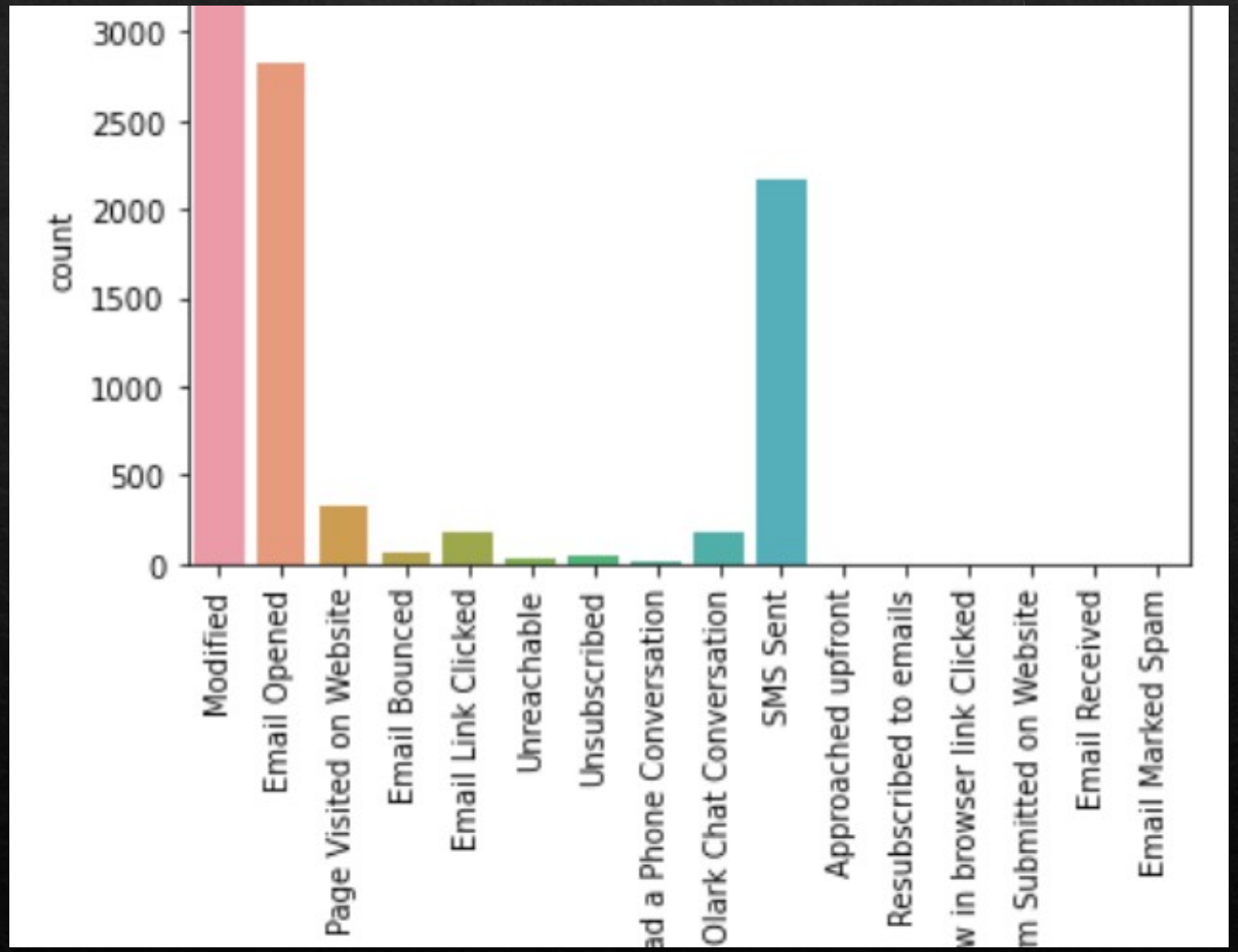
Lead Activity

What was the activity before being a lead?

There are 3 effective activities to get leads :

- Those opening emails and reading it
- SMS
- Modified

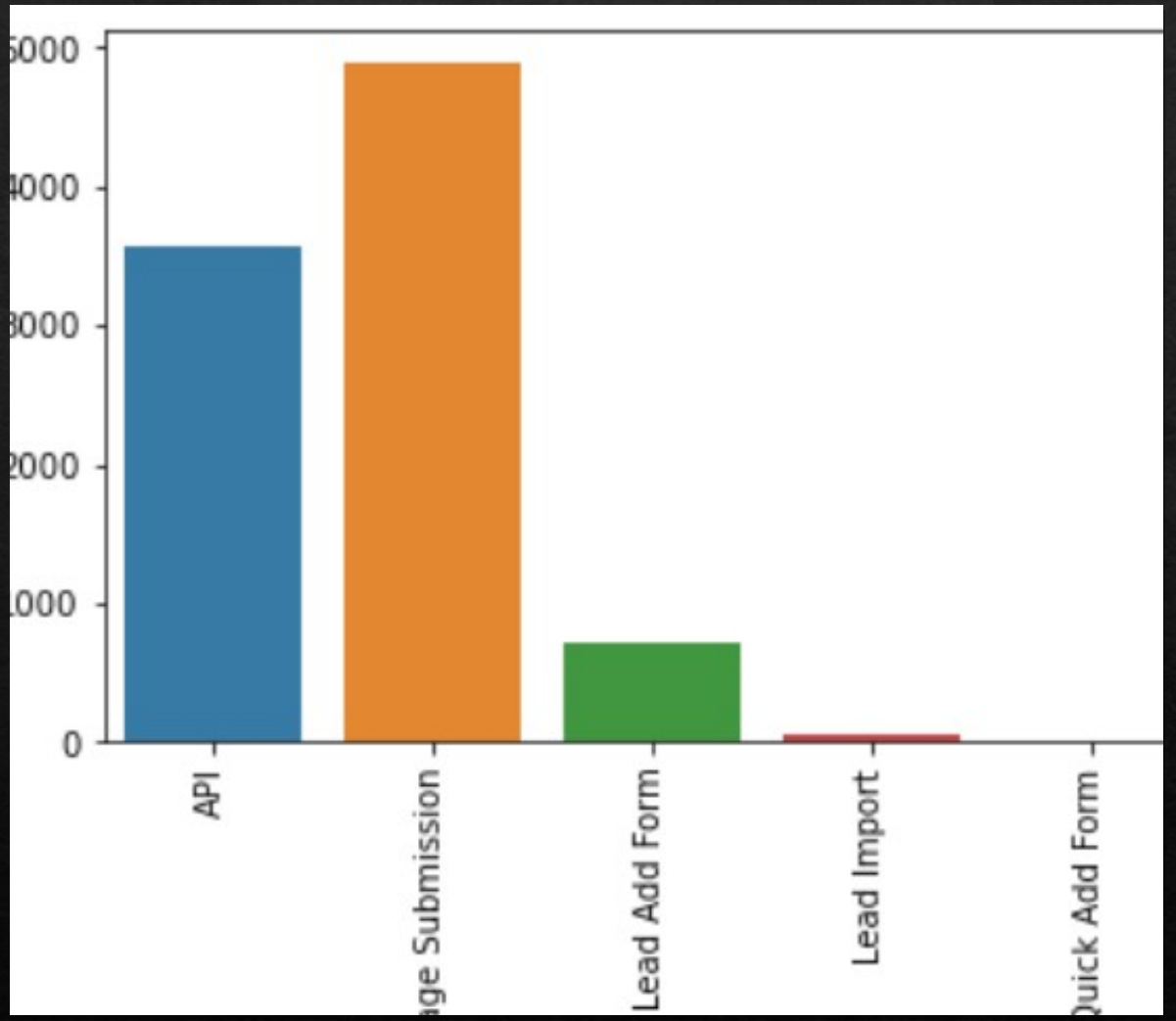
To get leads Email and SMS are the best way to get leads



Lead’s Origin

The origin means where the customer was identified to be a lead :

- Land page submission has the lead ability
- Application Interface (API) follows the conversion to lead

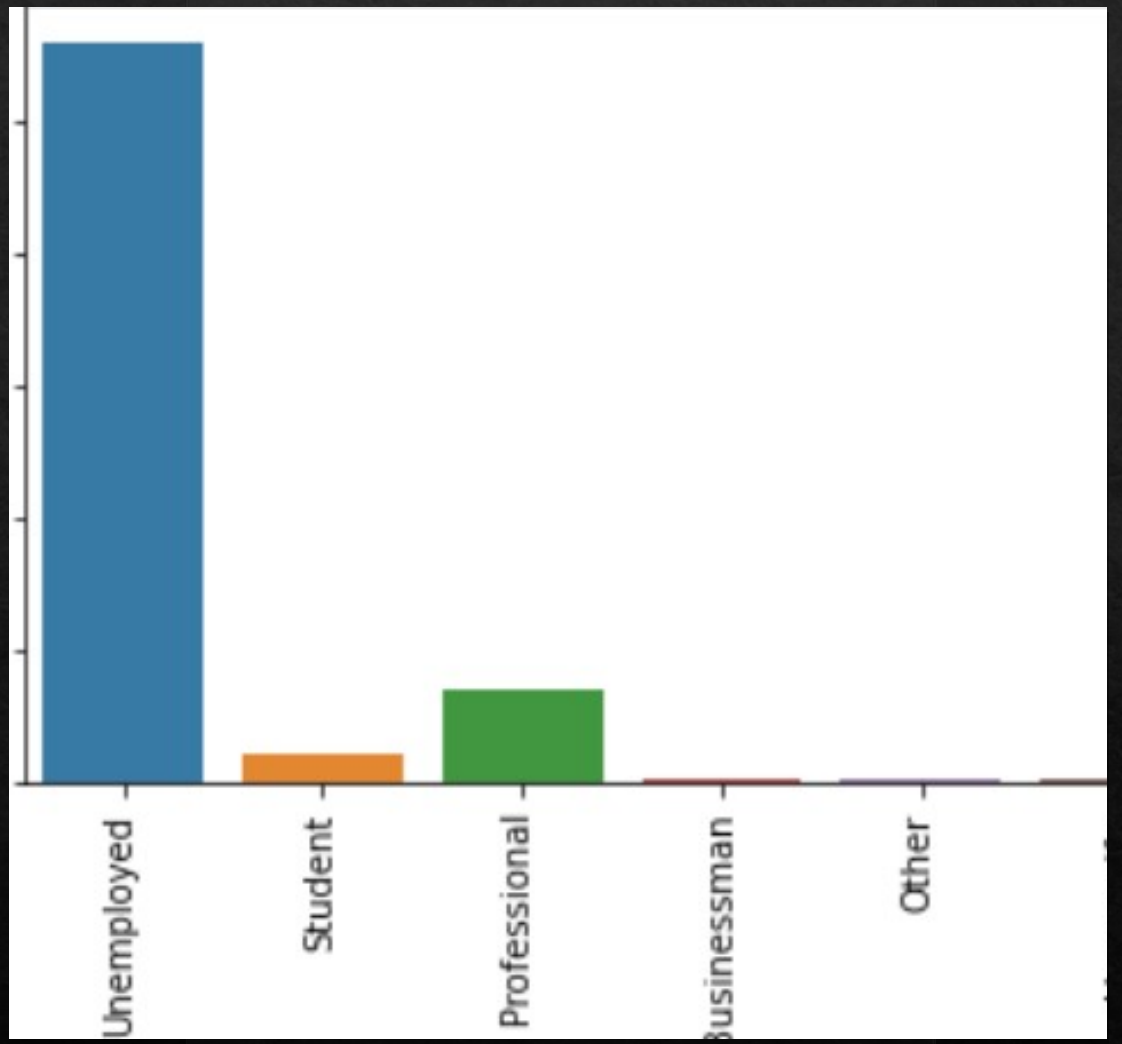


Occupation

What was the occupation :

- Most unemployed are the people want to do the course
- Along with them the working professionals

The company should work on upskilling and be flexible with timing



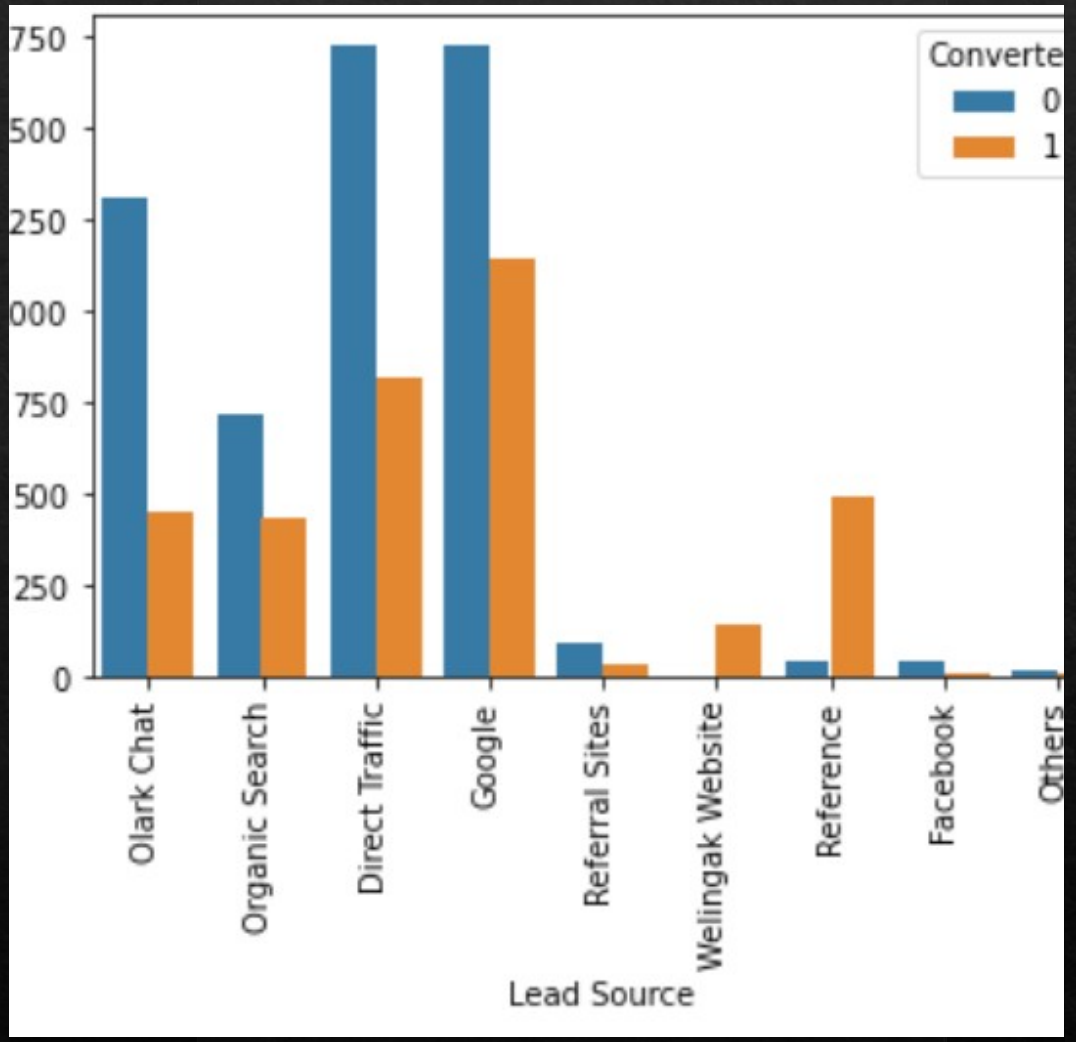
Features Vs Conversion

Lead Source Vs Conversion

Conversion rate as per lead sources :

- Google , Direct traffic , Olark chat are good source for getting leads
- The conversion rate is high for google search

The presence of the company should be maintained on google

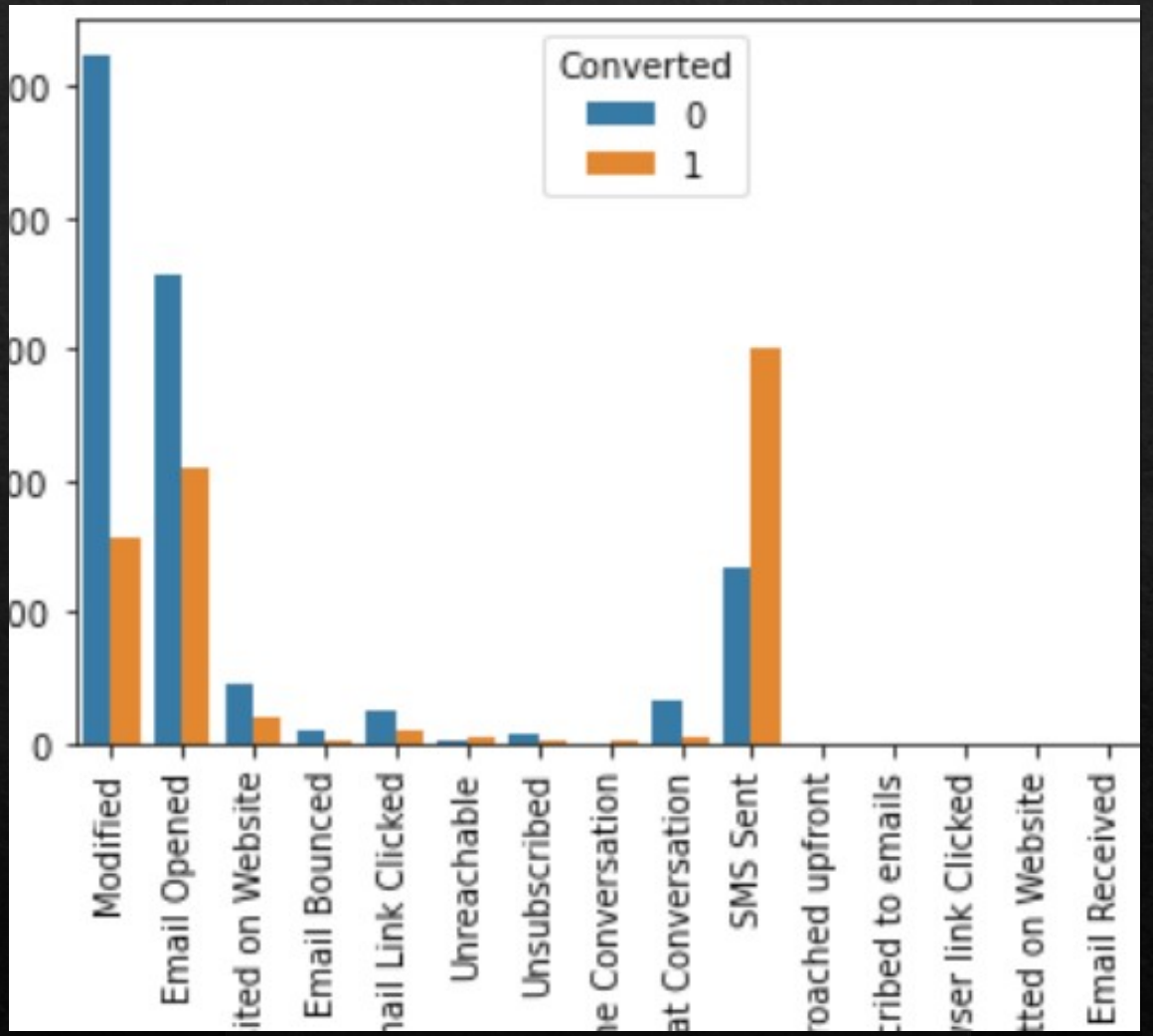


Activity Vs Conversion

Which activity has the highest lead conversion :

- SMS , Email are having the highest response
- But SMS has a high conversion rate

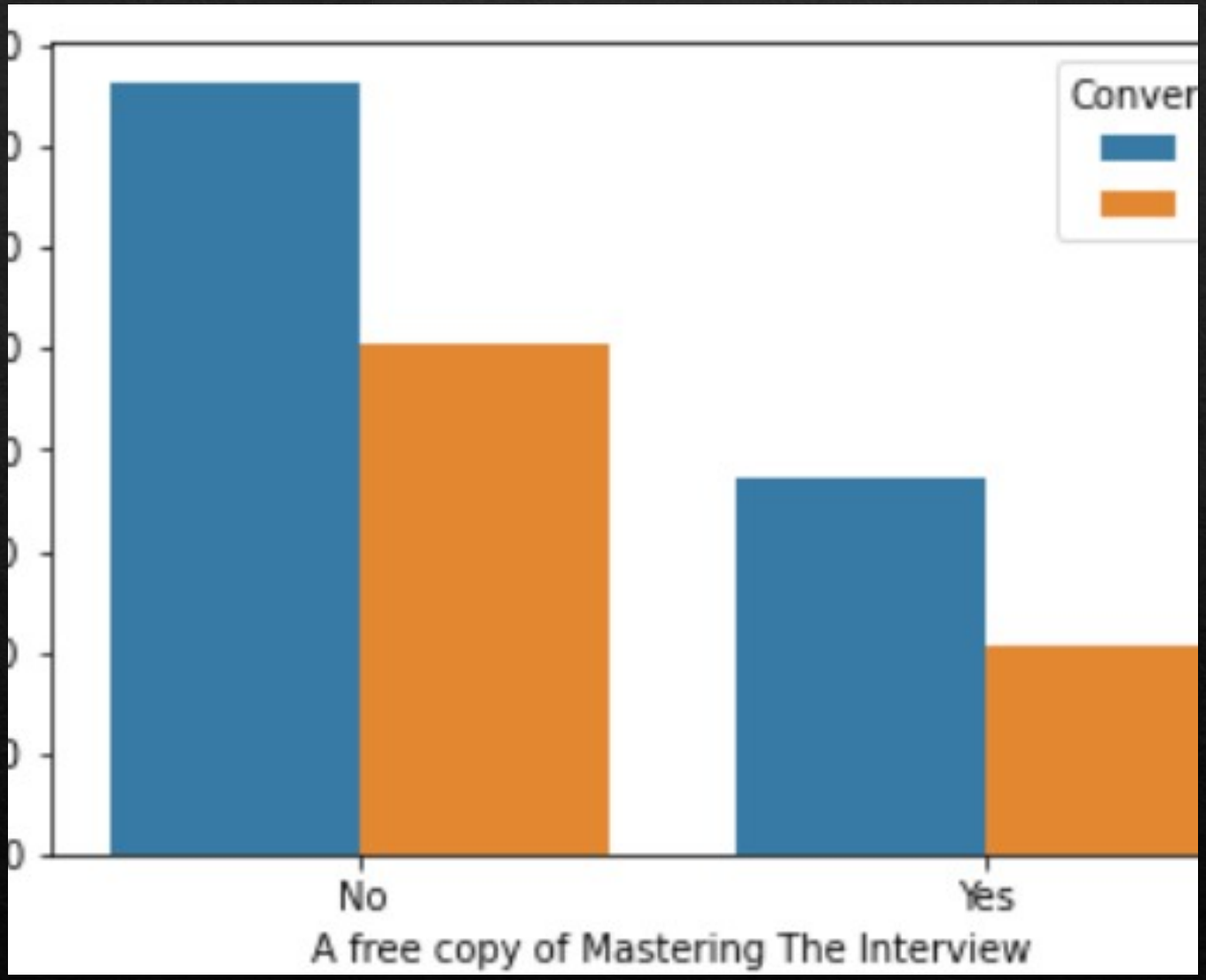
The company should focus on email marketing and SMS



Interview Copy

Whether they want the free copy of interview or not ?

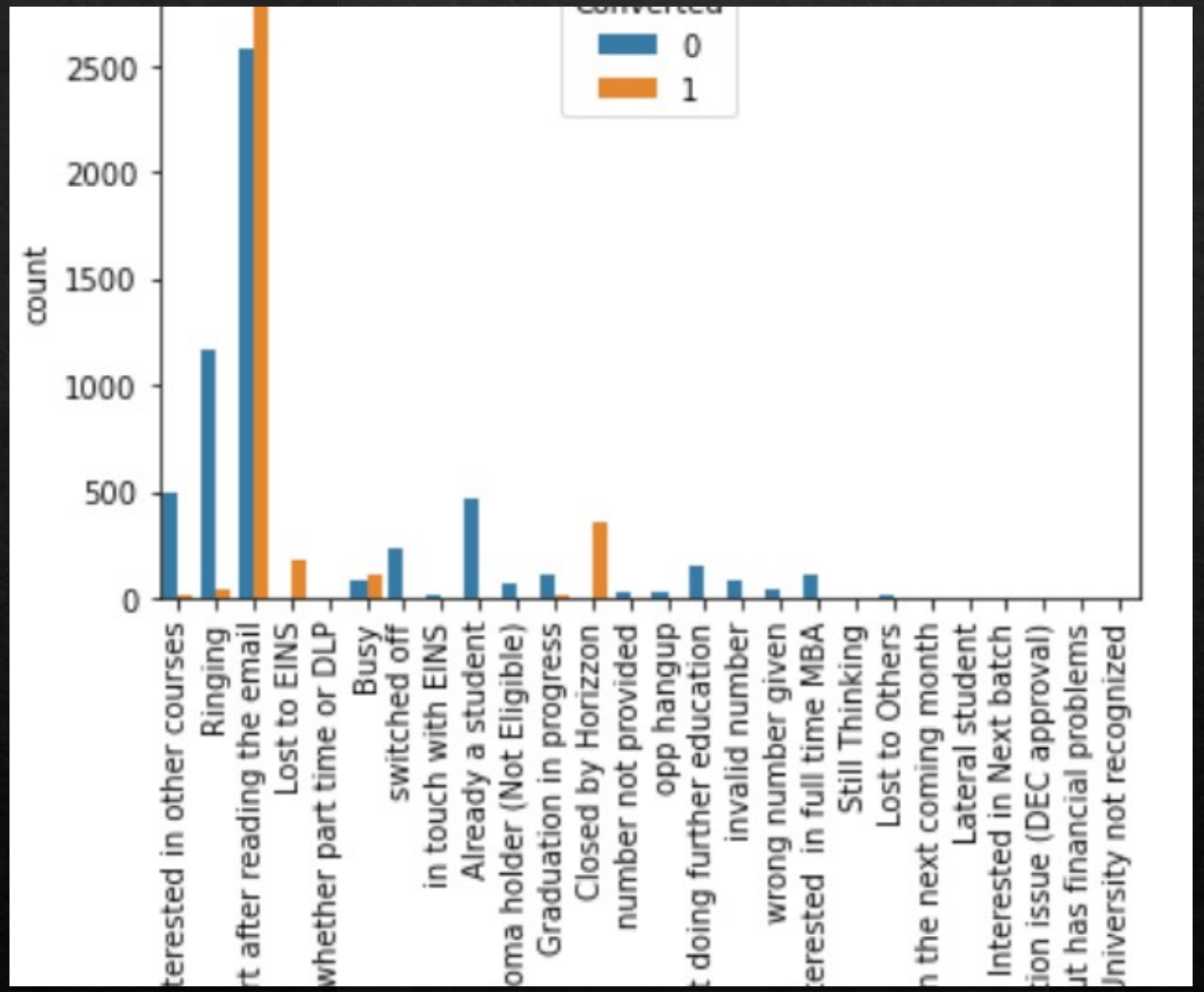
- Since there are mostly unemployed they are interested in a copy of interview
- While those who did not want a copy have seen a huge conversion



Tags vs Conversion

Tags is the current status of the lead

- The lead with the status :
- Will revert after reading email are high responsive to get converted



Model Building

Model Prediction

Process :

We built 9 models for the prediction

While reaching model 9 we deleted 20 insignificant variables

As per model 9 the significant variables are :

- Tags Closed by Horizzon
- Tags lost to EINS
- Revert after reading mail
- Lead Source Reference

	coef	std err	z	P>|z|	[0.02
const	-2.8661	0.230	-12.435	0.000	-3.31
otal Time Spent on Website	0.9108	0.106	8.590	0.000	0.70
Lead Source_Olark Chat	1.0741	0.335	3.208	0.001	0.41
Lead Source_Reference	2.2196	0.588	3.772	0.000	1.06
Last Activity_SMS Sent	1.2539	0.208	6.024	0.000	0.84
Country_Qatar	-3.3910	1.474	-2.300	0.021	-6.28
Tags_Busy	2.3373	0.302	7.742	0.000	1.74
Tags_Closed by Horizzon	7.4335	1.033	7.194	0.000	5.40
Tags_Lost to EINS	7.1175	1.109	6.417	0.000	4.94
Tags_Ringing	-1.2669	0.335	-3.786	0.000	-1.92
evert after reading the email	5.3416	0.257	20.759	0.000	4.83
Tags_switched off	-1.6842	0.635	-2.653	0.008	-2.92
Lead Quality_Worst	-2.7232	0.665	-4.094	0.000	-4.02

Confusion matrix

Actual/ Predicted	Not Converted	Converted
Not Converted	1304	64
Converted	81	1598

As per the Confusion Matrix :

- 2902 are rightly predicted out of which1304 are not converted and 1598 are truly converted
- Whereas 64 are actually converted but the model predicted them as not converted
- Also coming to actually converted the model predicted them as 81

0.9612864800476474
In [181]:

Sensitivity and Specificity

Sensitivity of our model is :

TP/ (TP+FN)

- 0.9612

Specificity of our model is :

TN/(TN+FP)

- 0.9451

Metrics accuracy :

- 0.9524

Precision and Recall

Precision :

Precision = TP/TP+FP

- 0.9555

Recall :

Recall = TP/TP+FN

- 0.9612

Precision and Recall Score

We also used sci-kit-learn to find the same

Precision Score :

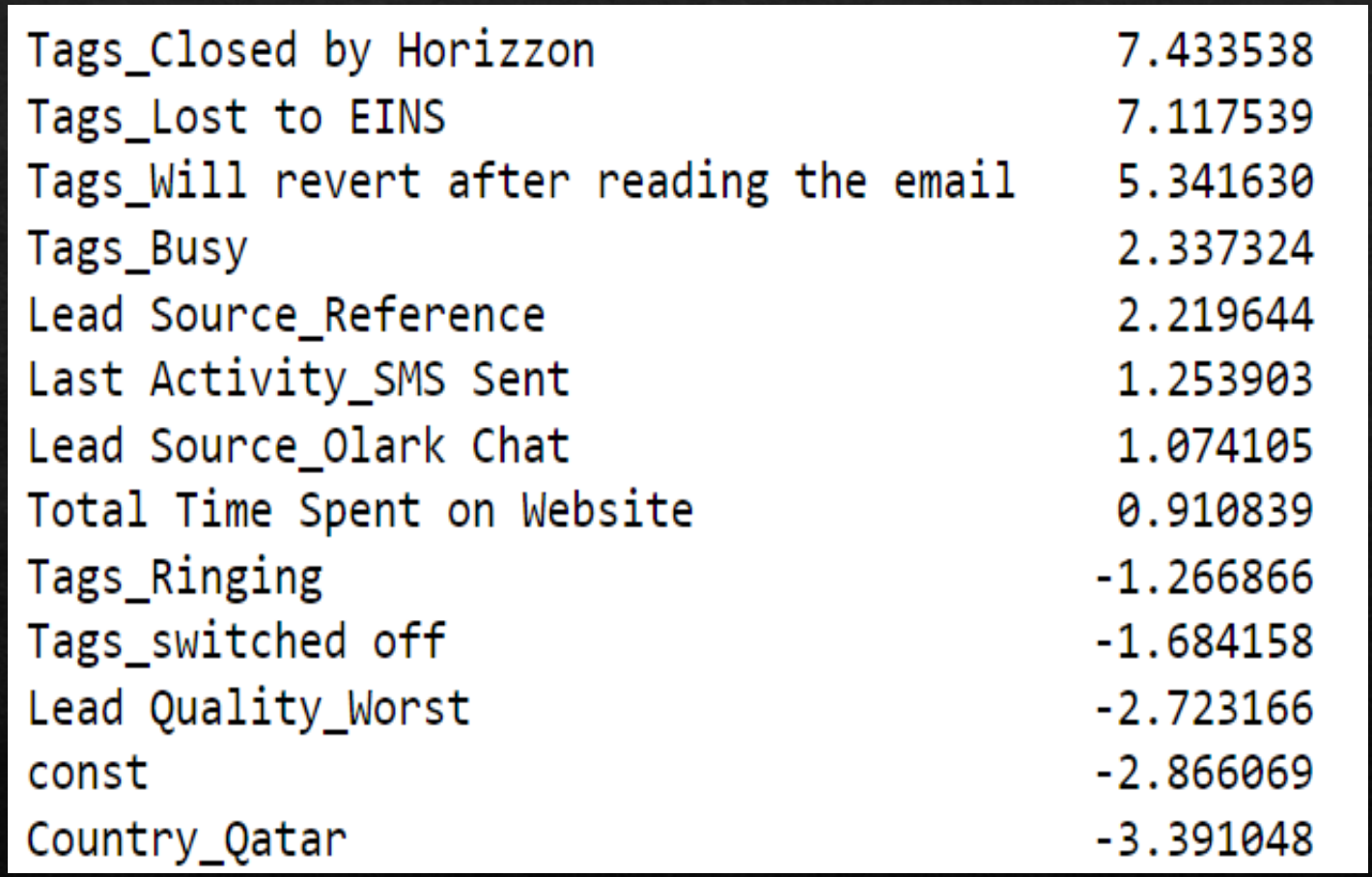
- 0.9614

Recall Score :

- 0.9517

Most Important Features

1. Tags closed by horizon :
   - 7.433
2. Tag Lost to EINS :
   - 7.1175
3. Will revert after reading mail :
   - 5.431
4. Tag Busy :
   - 2.337
5. Lead source as reference :
   - 2.219
6. Activity as SMS sent :
   - 1.2539



Tags_Closed by Horizzon	7.433538
Tags_Lost to EINS	7.117539
Tags_Will revert after reading the email	5.341630
Tags_Busy	2.337324
Lead Source_Reference	2.219644
Last Activity_SMS Sent	1.253903
Lead Source_Olark Chat	1.074105
Total Time Spent on Website	0.910839
Tags_Ringing	-1.266866
Tags_switched off	-1.684158
Lead Quality_Worst	-2.723166
const	-2.866069
Country_Qatar	-3.391048

The company can make calls to them and concentrate on Lead Sources and activity as SMS and email