Lead Scoring

CASE STUDY

By
Sanjukta Sengupta
Priyanka Arunachalam

Problem Statement

1) To identify the most potential leads, also known as ‘Hot Leads’ .

These leads can be from referrals, manual entry by sales team or through online portal filling by customers.

2) The lead conversion rate should be around 80% .

3) If the company's requirement changes in the future our model should be capable of doing this analysis dynamically .

Goal For The X Education

-> 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.

-> There are some more problems presented by the company which our 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.

Steps for Data Analysis

1. Importing the essential Libraries
2. Understanding the Dataset
3. Handling missing values in the training as well as Test set
4. Defining the Dependent and Independent Variables and exploring the relationship.
5. Defining the model
6. Model Evaluation
7. Prediction using Test Set

By Priyanka Arunachalam & Sanjukta Sengupta

LET’S NOW SEE HOW FACTORS WILL BE AFFECTING THE CONVERSION OF LEADS

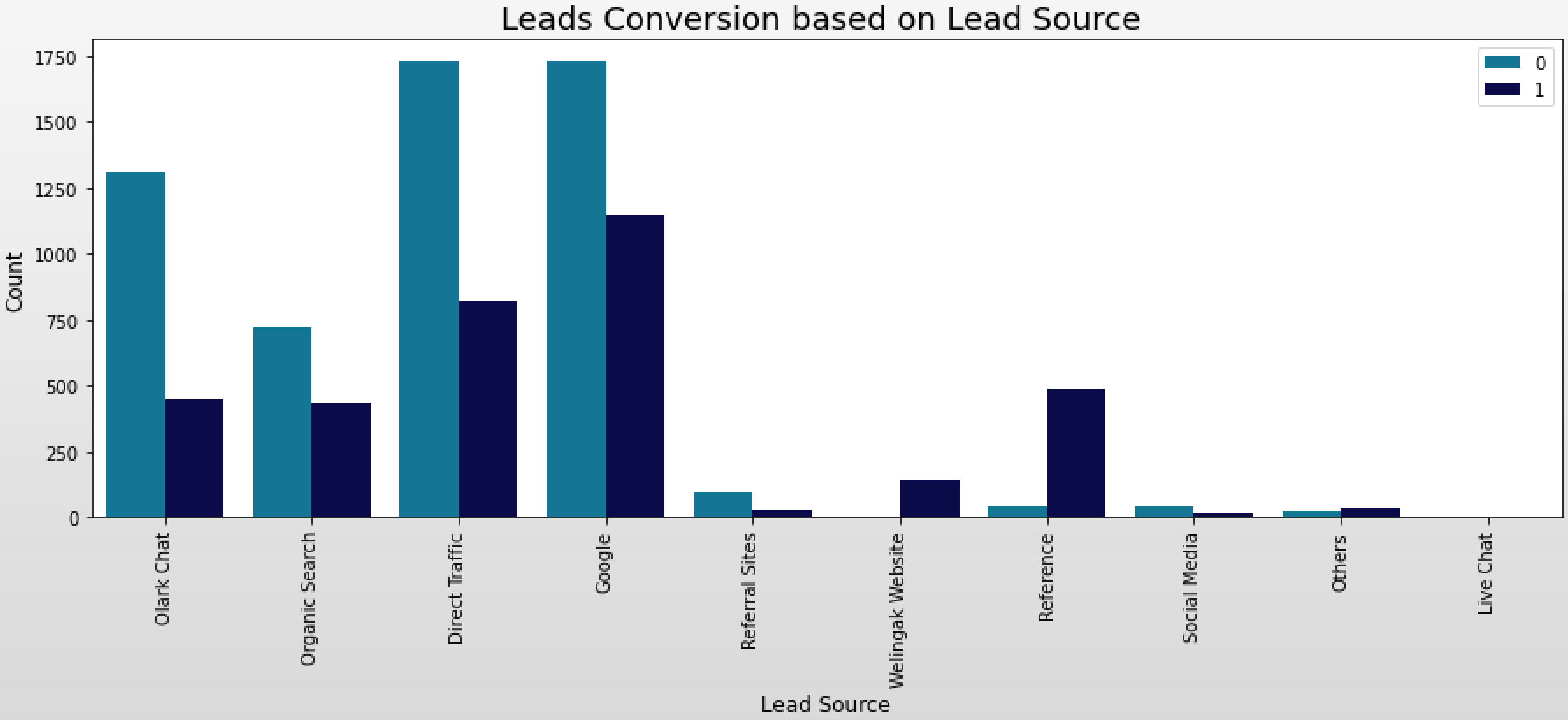
How much of Lead conversion is taking place through a Lead Source?

INFERENCES:

1.Highest number of the leads are generated from Google and Direct Traffic.

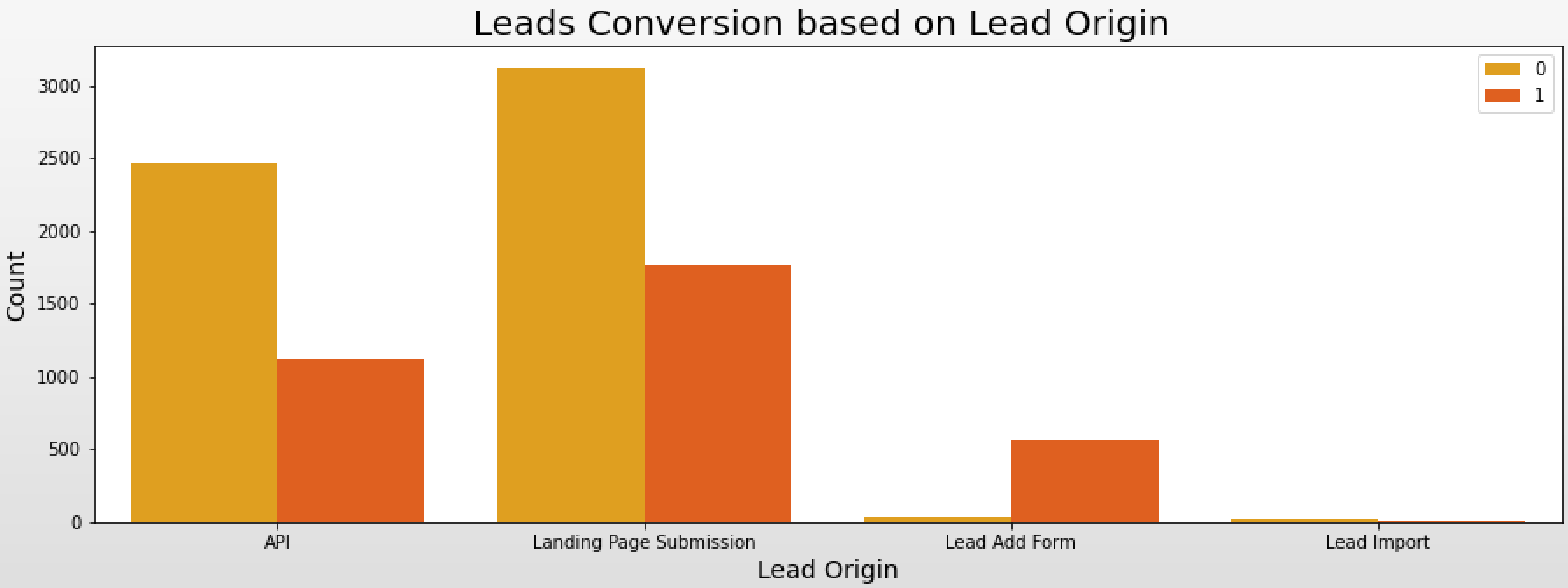
2. While the least being the Live chat attribute through which leads can be generated.

3.Leads can be maximized from References and other Welingak Website.



By Priyanka Arunachalam & Sanjukta Sengupta

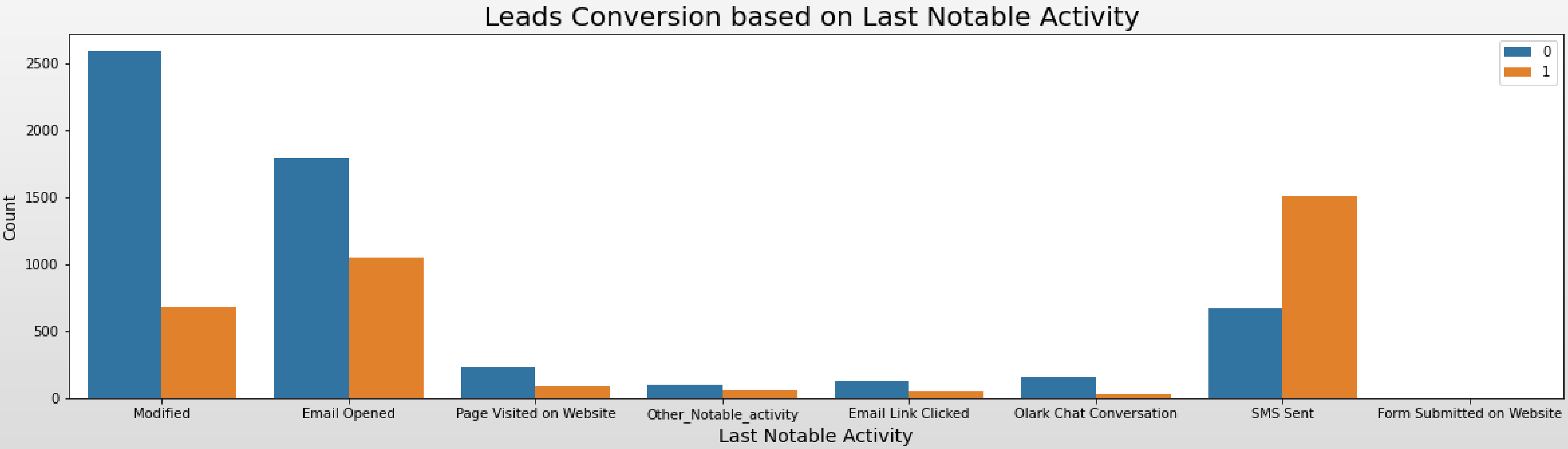
How much of Lead conversion is taking place on the basis of Lead Origin?



INFERENCES:

1. API and Landing page is giving away the most leads for conversion.
2. Lead Add form has high number leads for conversion but the count of the leads is less.
3. Higher count of leads in Lead Add form can lead to higher conversion

How much of Lead conversion is affected on the basis of Last Notable Activity?



INFERENCE:

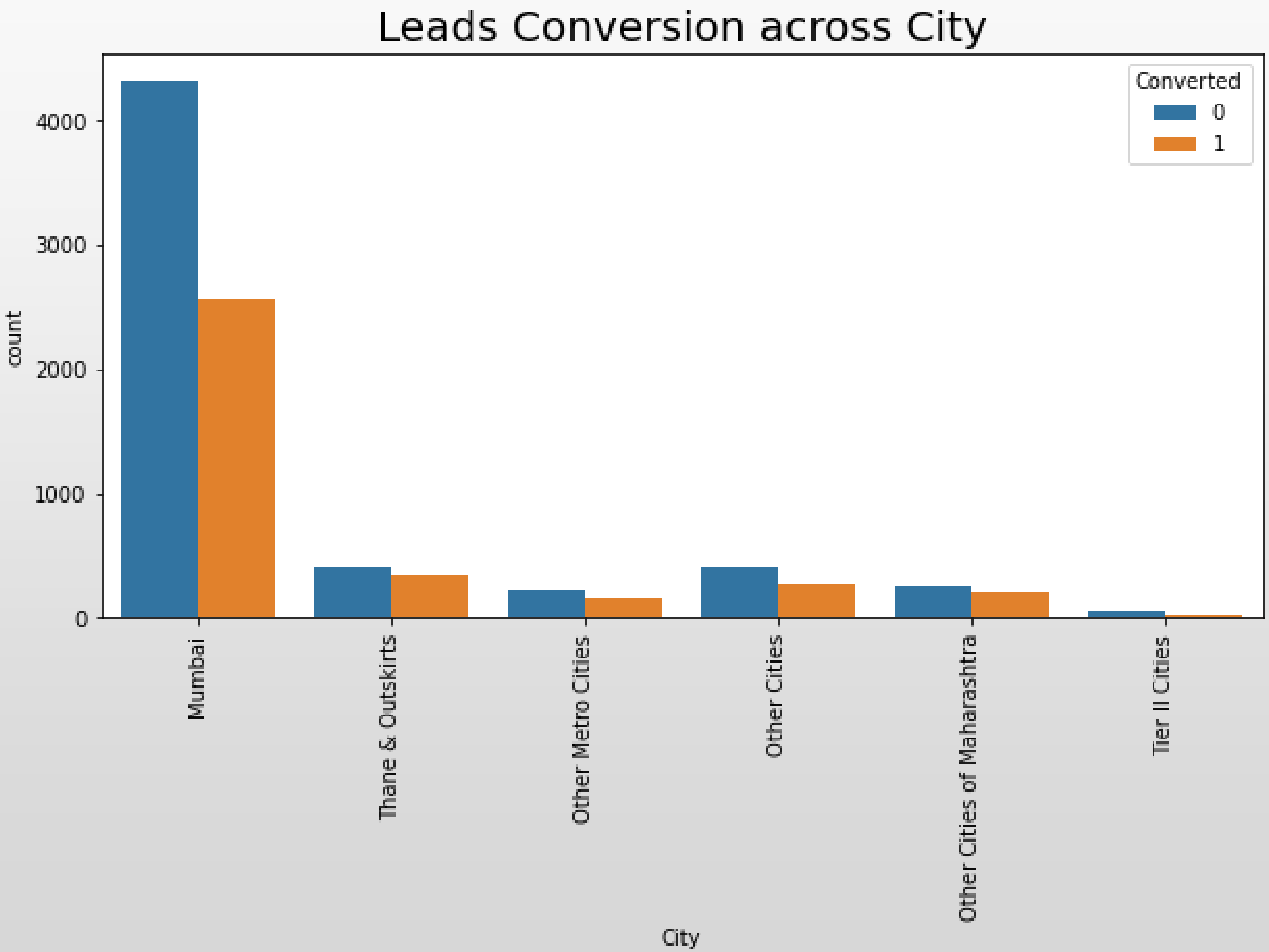
SMS Sent and Email opened is having higher number of leads as compared to others.
Modified has the least number of leads conversion ration

By Priyanka Arunachalam & Sanjukta Sengupta

Bivariate Analysis of Lead Conversion Across Cities

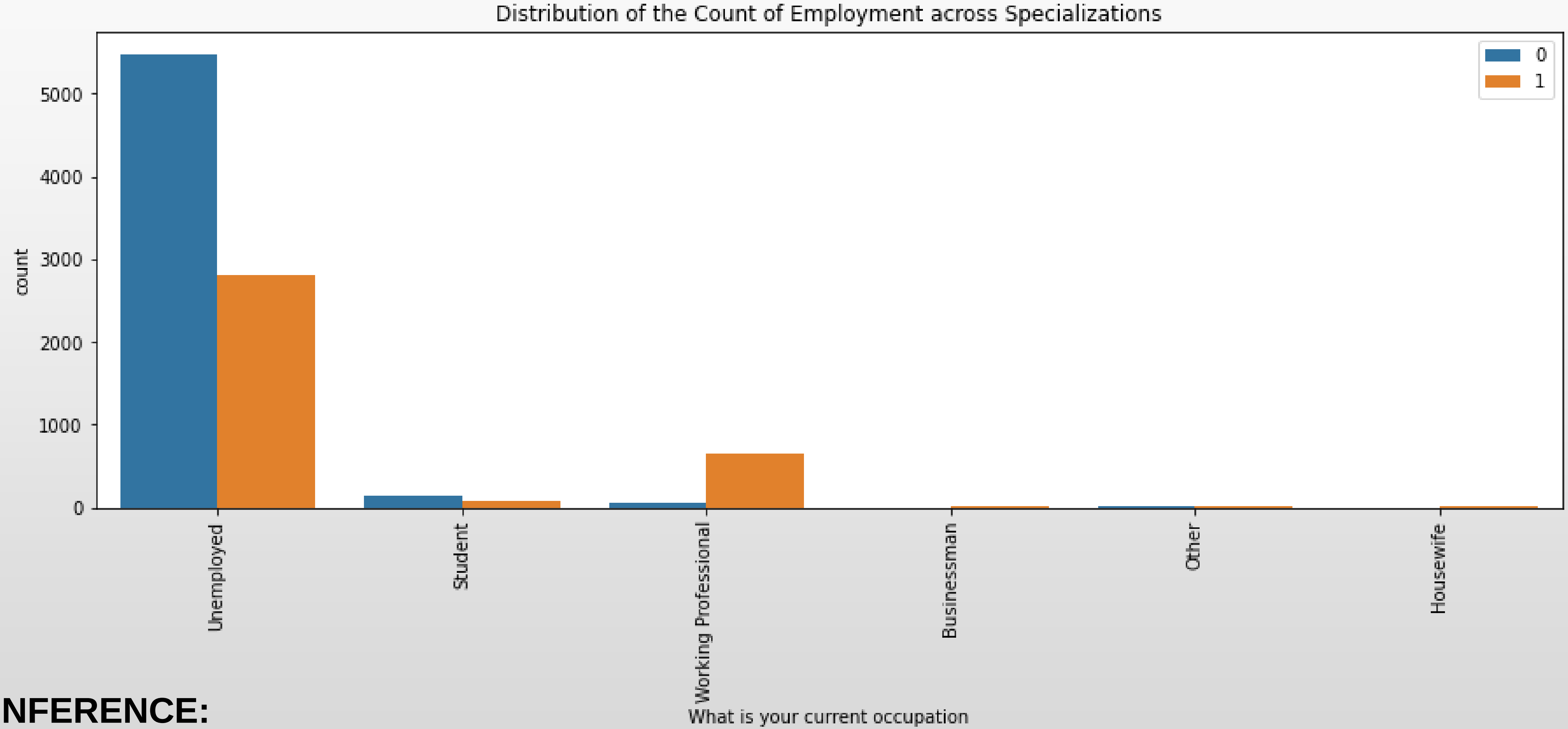
INFERENCE:

1. Here we can see that Mumbai is the most occurring city in the Dataset.



By Priyanka Arunachalam & Sanjukta Sengupta

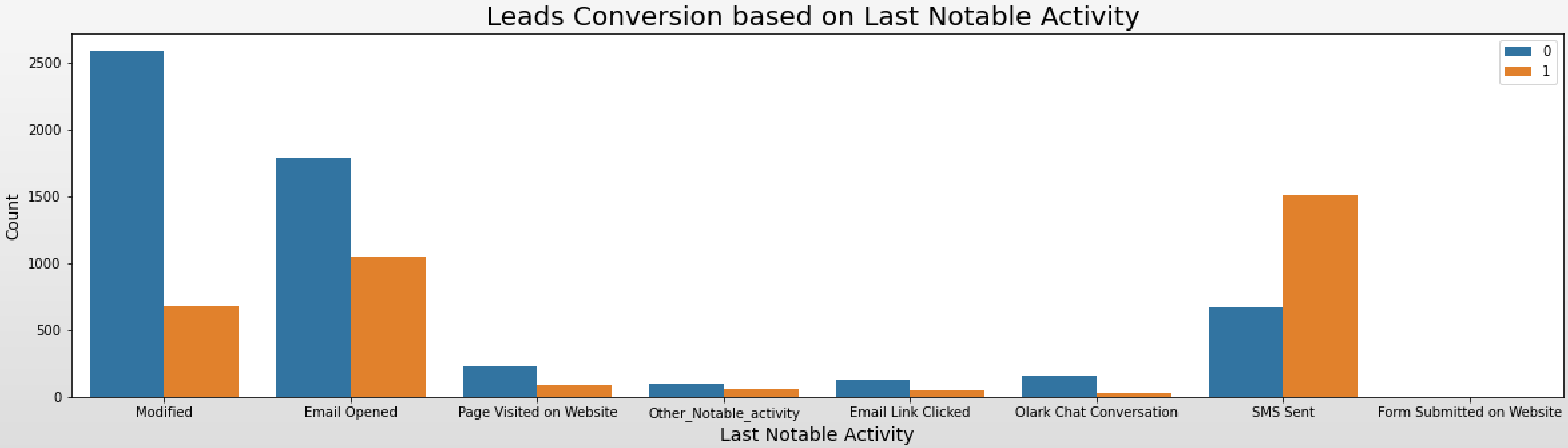
Bivariate Analysis of Current Occupation vs Number of Leads



INFERENCE:

Here Working Professional can go for the course as the chances are high
Unemployed leads are the highest over here.

Leads Conversion on the Basis on Last Notable Activity



INFERENCE:

SMS Sent and Email opened is having higher number of leads as compared to others.
Modified has the least number of leads conversion ration

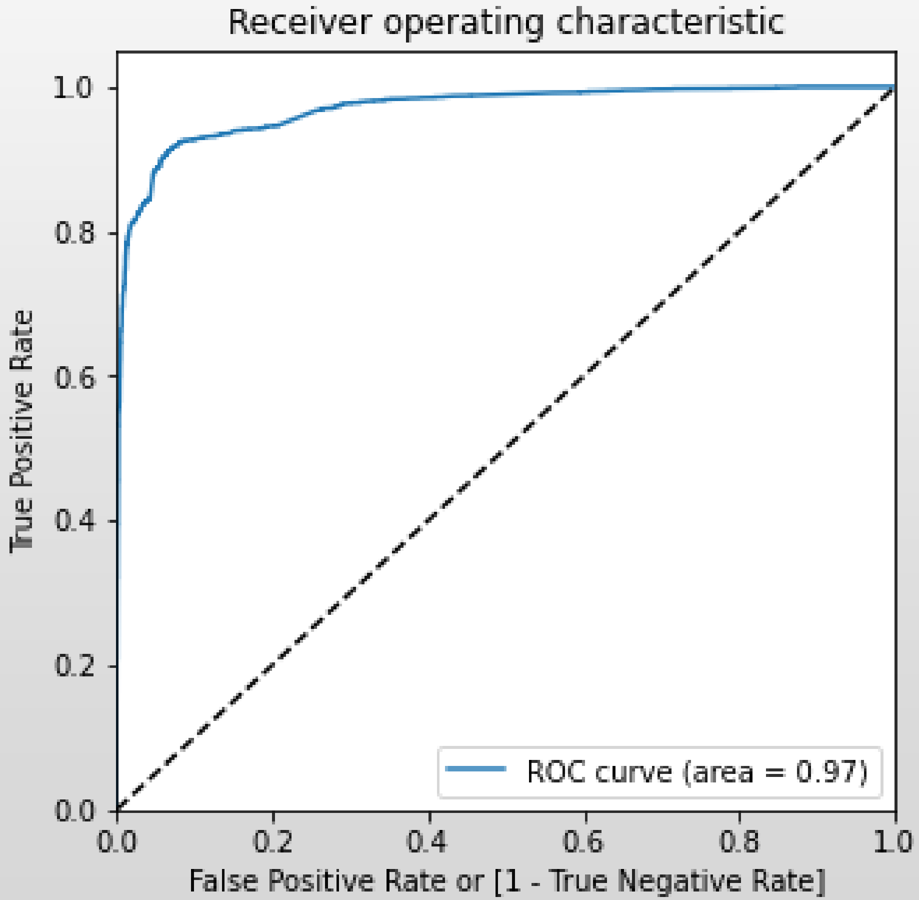
By Priyanka Arunachalam & Sanjukta Sengupta

Plotting an ROC Curve to see how good a model can perform

INFERENCE:
As we have seen that ROC value should have the value close to 1 but here we can see that it is 0.97 which is close to 1.
It clearly indicates that it will be a good model

So as we can clearly notice that the above model is performing well. The ROC curve also shows a value of of 0.97, which is very good value. Now we have found the following values for the Train Data:
Accuracy : 92.30%
Sensitivity : 91.7%
Specificity : 92.66%

ROC CURVE



By Priyanka Arunachalam & Sanjukta Sengupta

LOOKING AT VARIANCE INFLATION FACTOR

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

INFERENCE:

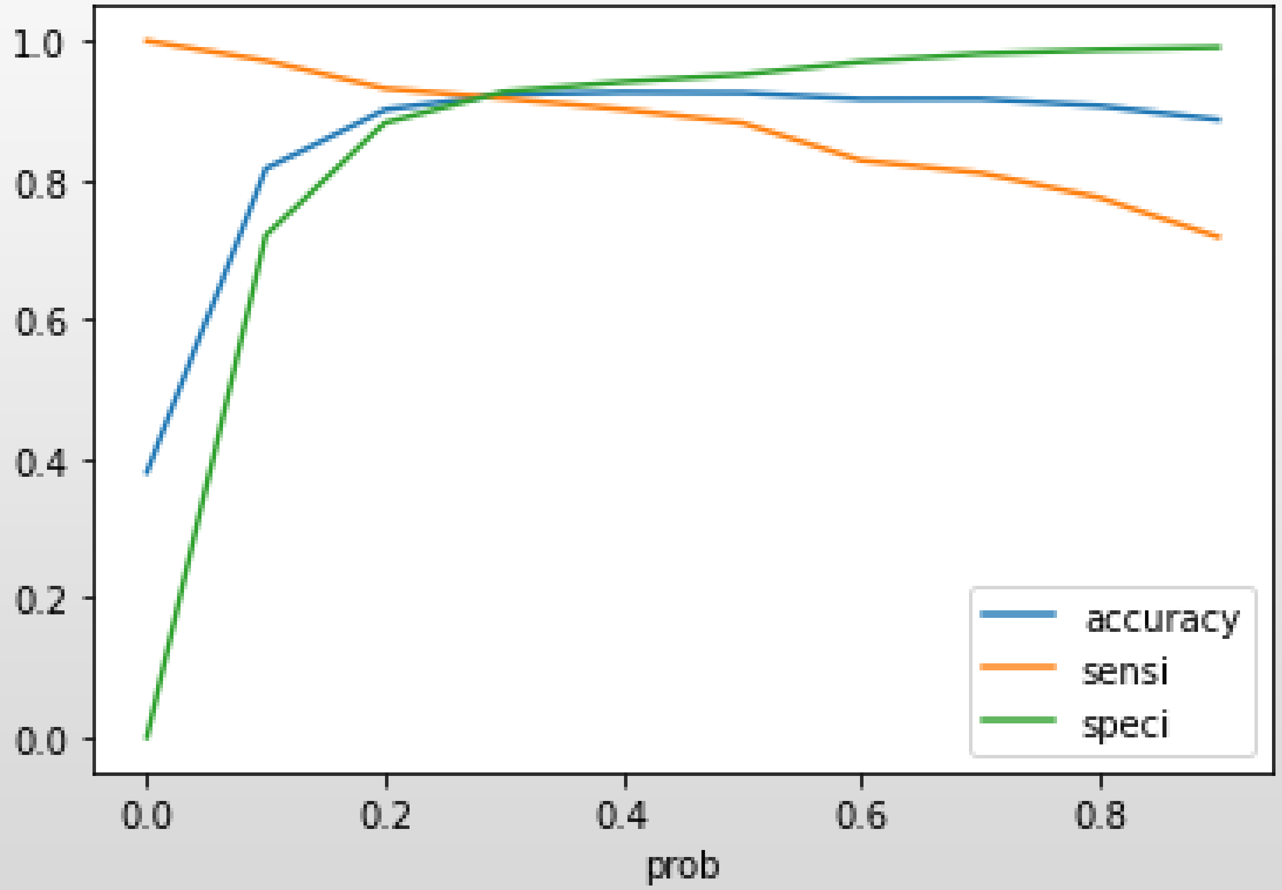
Here we can see that the VIF value of each variable is below 5. Hence it is suitable for the prediction of leads.

By Priyanka Arunachalam & Sanjukta Sengupta

Accuracy, Sensitivity, and Specificity Trade-Off

INFERENCE:

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



Whether Agenda is achieved or Not?

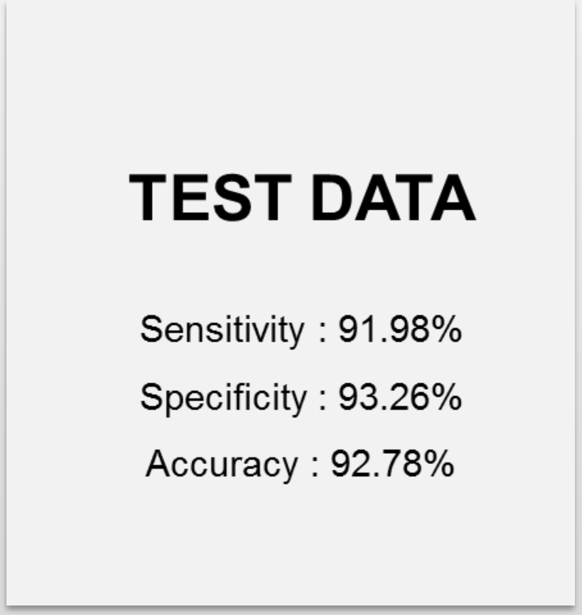
	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	7681	0	0.02	2	0
1	984	0	0.03	3	0
2	8135	0	0.69	69	1
3	6915	0	0.01	1	0
4	2712	1	0.95	95	1

INFERENCE:

Here we can see that the final Predicted value above 30% is considered as 1 and less than 30% is 0. Hence we have achieved a lead conversion of over 80%.

FINAL OBSERVATION

Now we will be comparing the values obtained from Train & Test Set:



TEST DATA

Sensitivity : 91.98%

Specificity : 93.26%

Accuracy : 92.78%

TRAIN DATA

Accuracy : 92.29%

Sensitivity : 91.70%

Specificity : 92.66%

Here we can see that our Model has achieved over 90% Accuracy, Sensitivity and Specificity, which clearly depicts that our Lead conversion will increase by 80%.

THANK YOU

THE END

By Priyanka Arunachalam & Sanjukta Sengupta