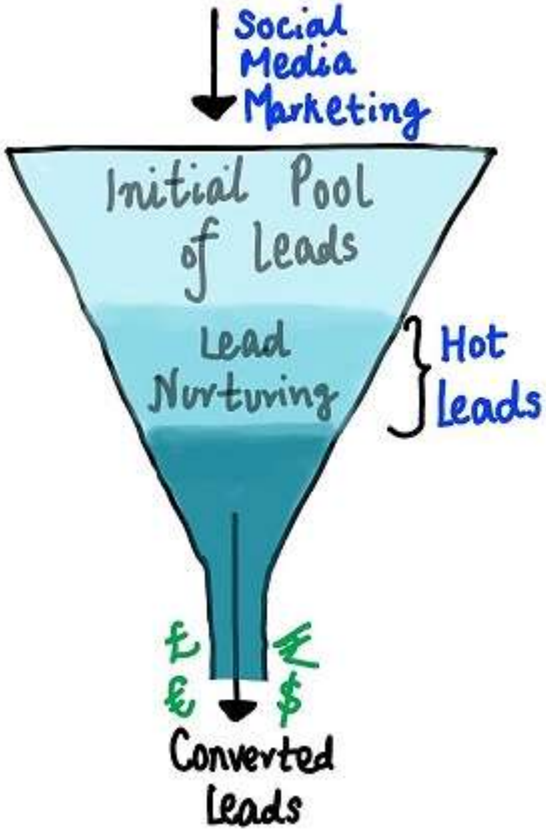
Lead Scoring Case Study

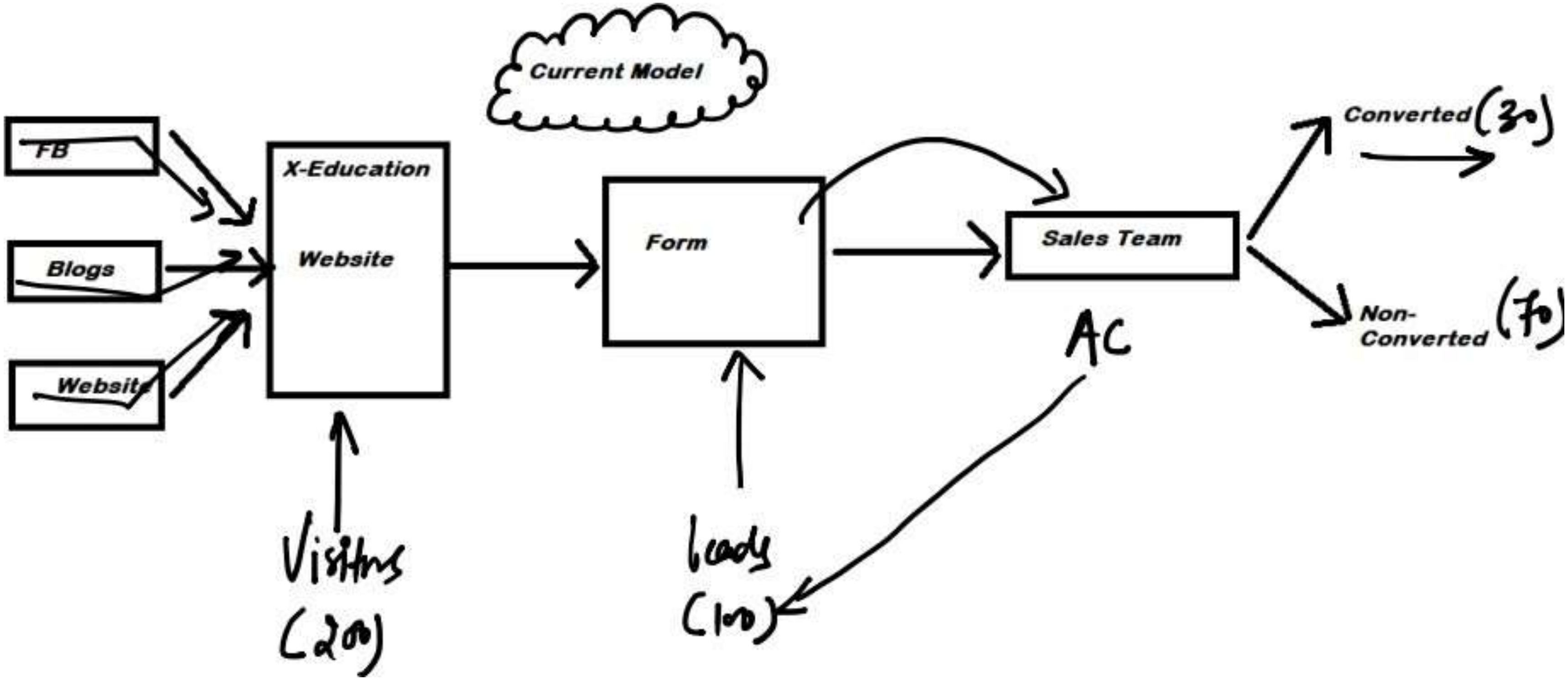
Shrikant Patil

Problem Statement

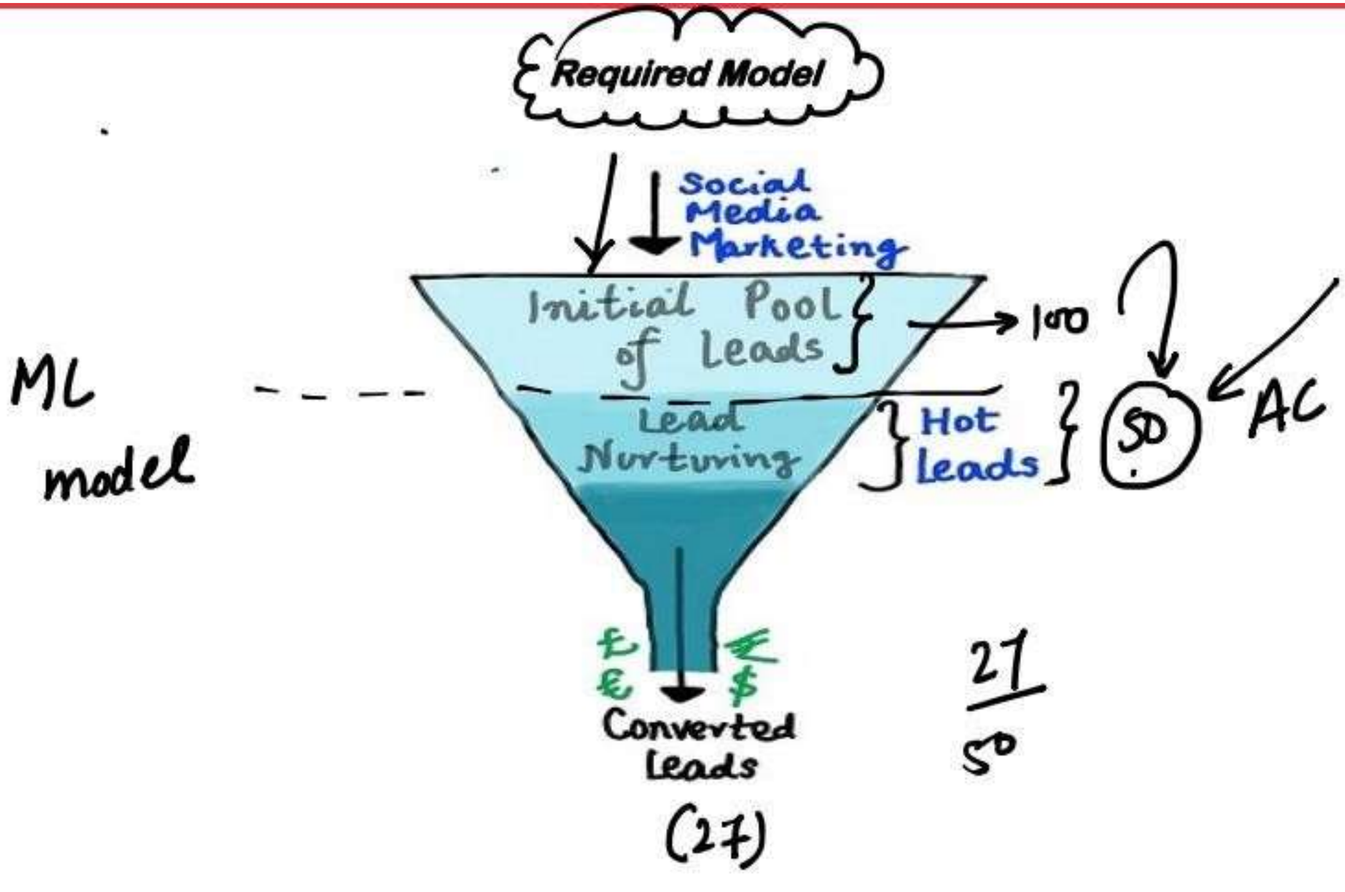
- An education company named X Education sells online courses to industry professionals.
- X Company markets its courses on several websites and search engines like Google. The company call people through their email address or phone number, classified is a **lead**. The typical lead conversion rate at X education is around 30%.
- To increase conversion rate company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- The company requires to build a model wherein need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion Chance.



Current model overview



Company’s CEO expectations



Approach of Data Analysis

- As we aim to predict weather lead is potential lead or not so it's becomes Classification Problem and thus we are using logistic regression model here.
- So we can proceed with analysis in following steps;

✓ Read & Understand Data

✓ Data Cleaning

✓ EDA

✓ Data Preparation

✓ Modeling

✓ Evaluation of Model

Data Understanding

- In the first phase, we began to comprehend the term "Problem Statement."

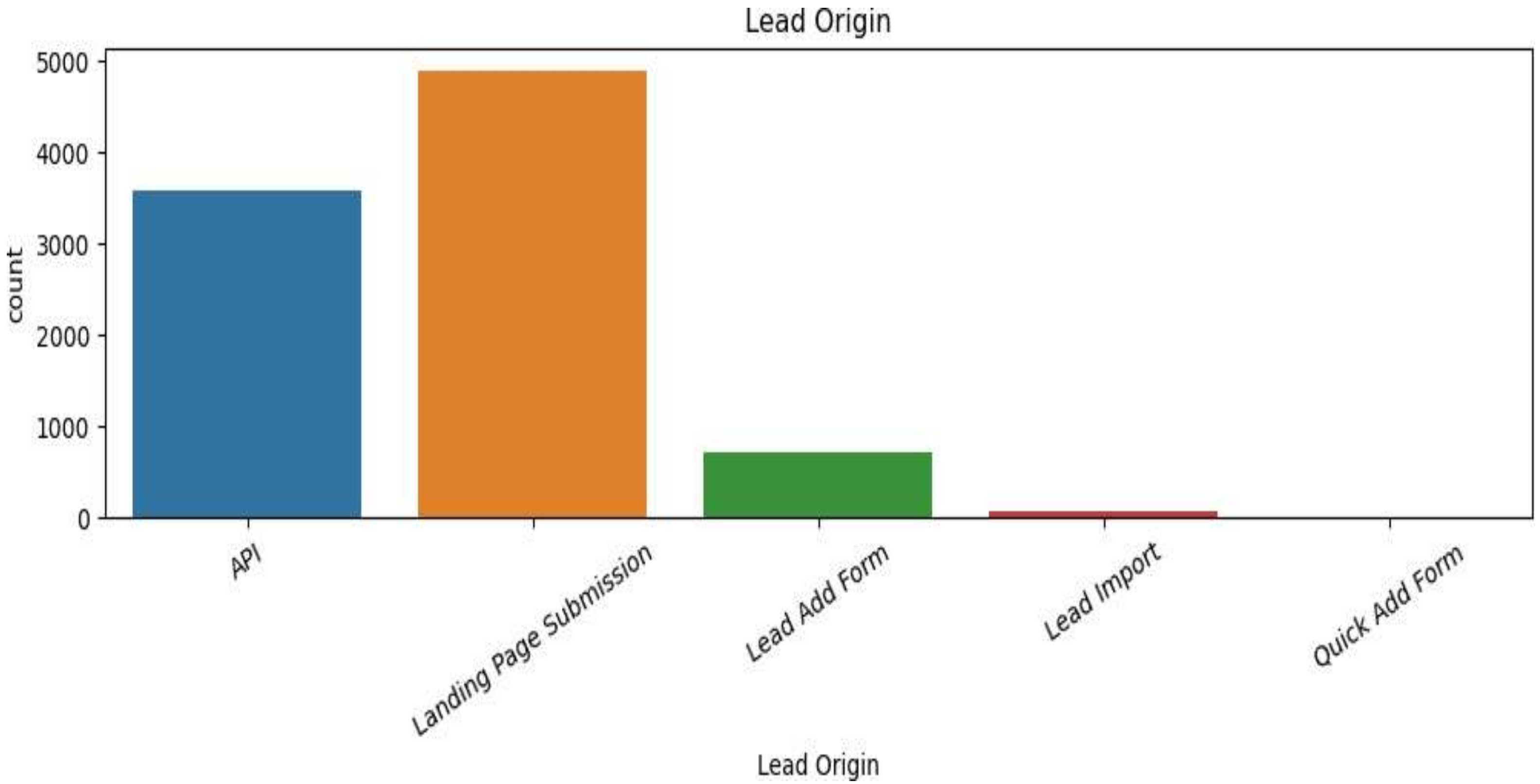
- Then, in order to read and comprehend the data, all necessary libraries were loaded into the Jupyter notebook, and the data dictionary was used to understand the meaning of the data in the file.

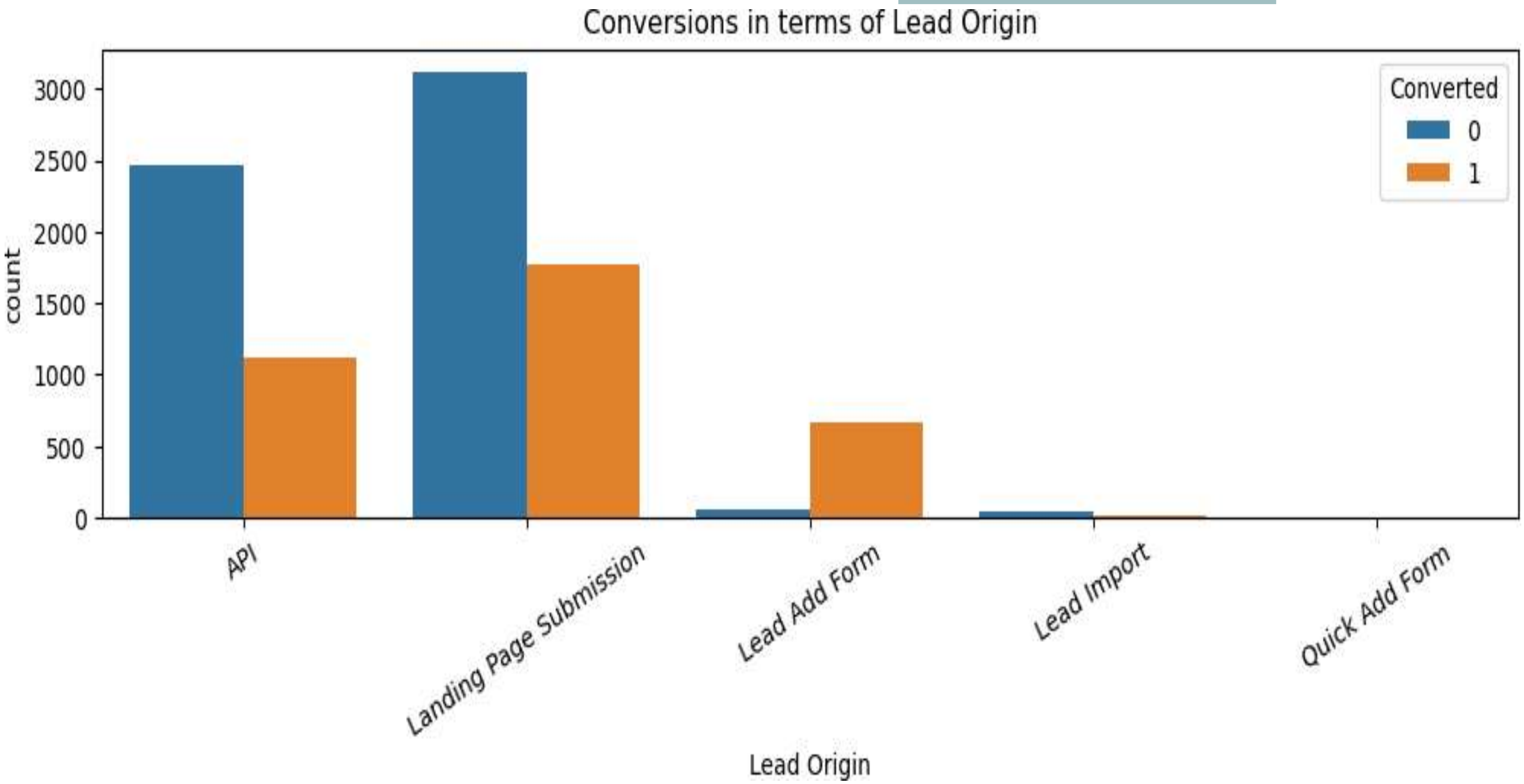
Data Cleaning

- Lot of variables has 'Select' values which are considered to be null values so it replaced with Nan.
- Columns have large number of null values i.e. greater then 40% so we dropped it.
- Since 'Project ID' and 'lead Number' are of no use in regression model and also all have unique values we dropped them.
- Still there was columns with missing value percentage between 25-40% so checked every column one by one & do necessary cleaning as you can see in jupyter file
- So as part of cleaning data we drop some unnecessary column, binning, handle outlier, imputing & so on

EDA (Univariate analysis)

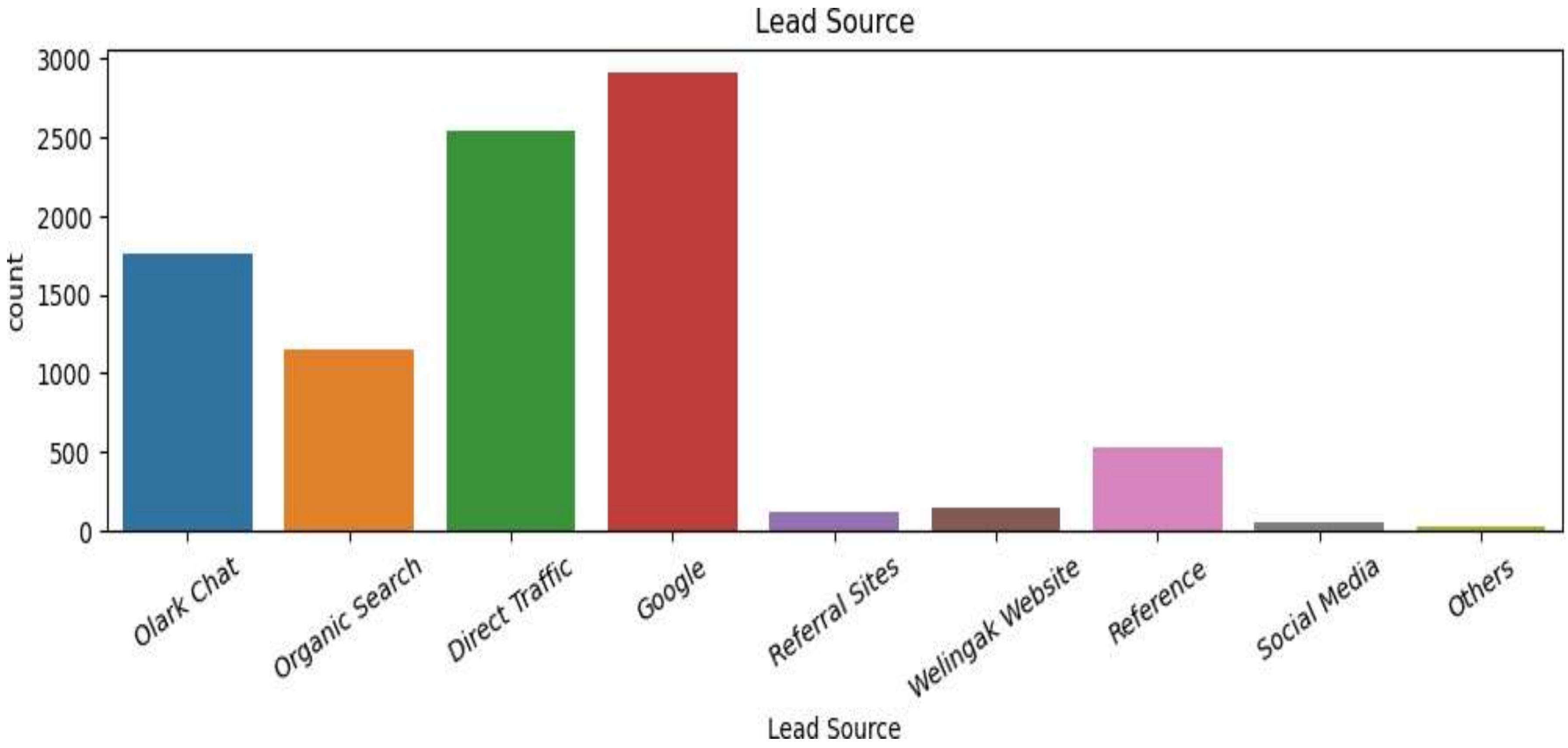
- Lead Origin

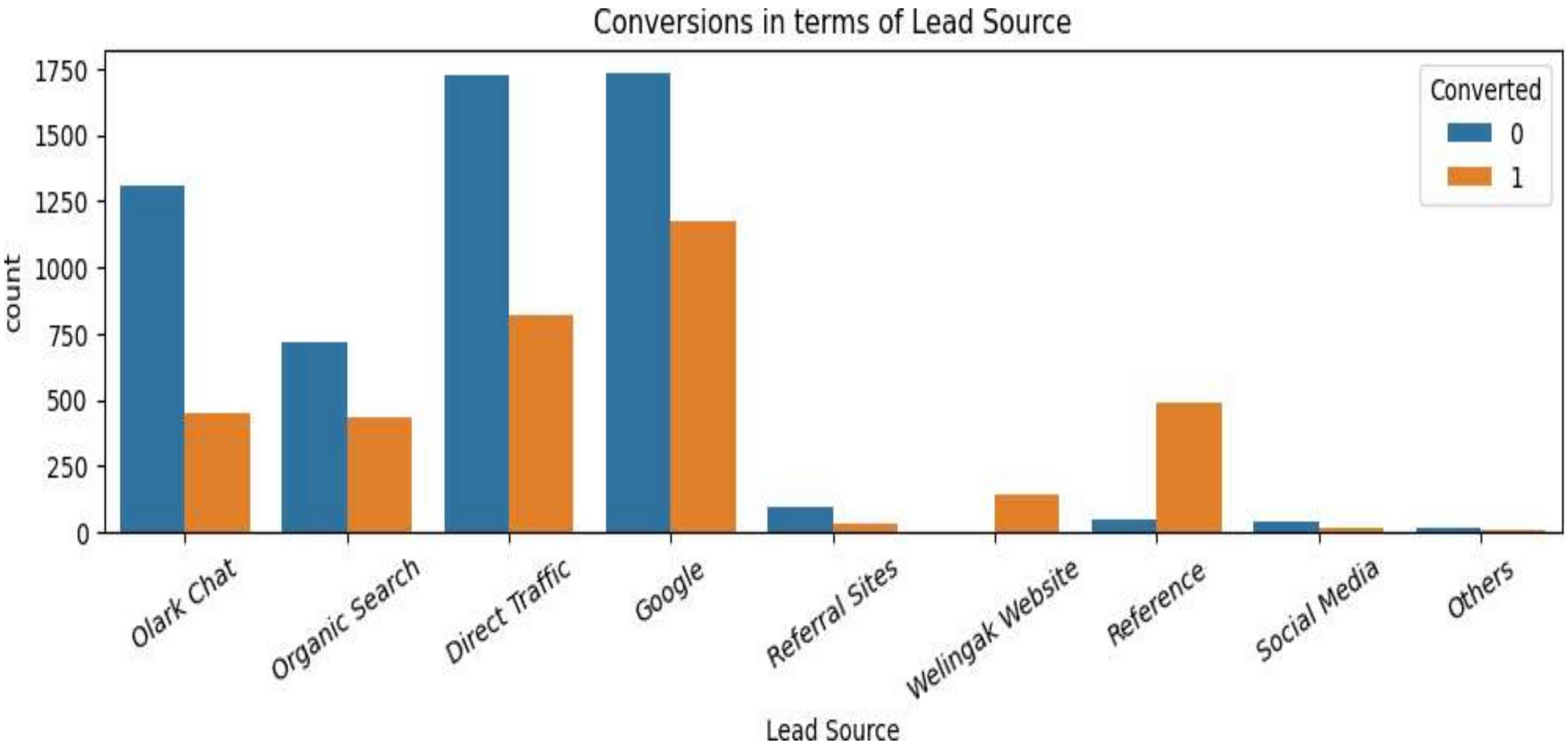




- We can see that API and LP submission has higher number of lead conversions
- Lead add form has higher number of lead conversion rate
- Focus on API and LP submission as they bring higher number of conversion in terms of counts.
- Also try to generate more leads from lead add form as it has higher number of conversion rate

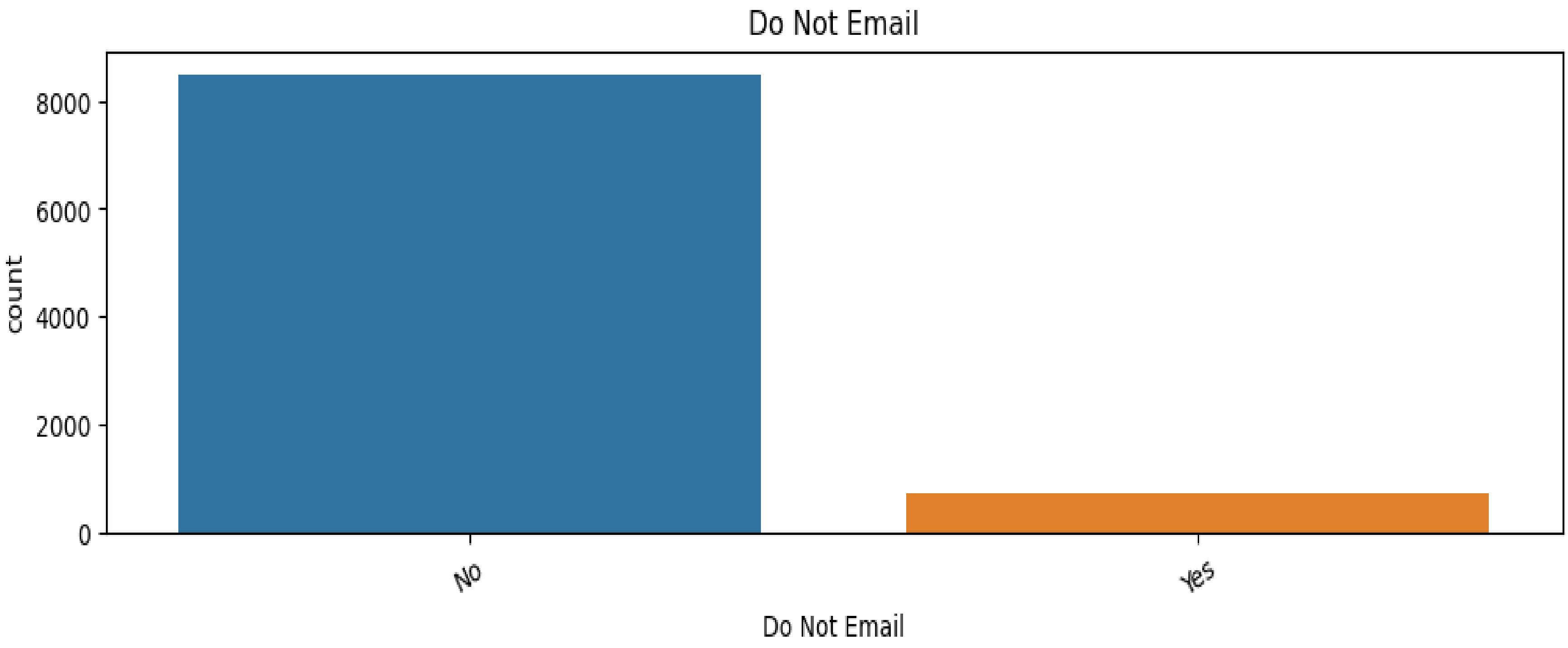
- **Lead Source**

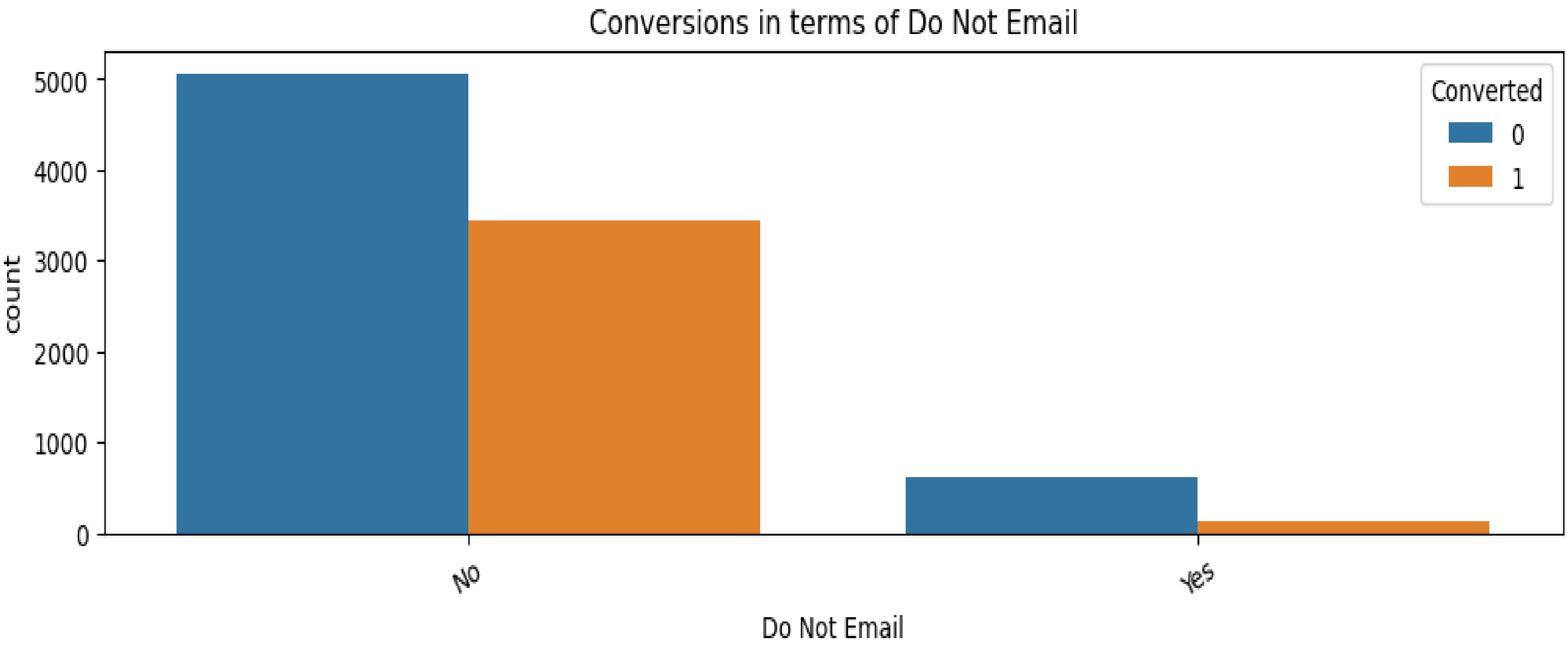

Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Social Media
Others
Lead Source



- We can see that olark chat, organic search, direct traffic, Google has majority of lead conversions in terms of count, so focus on increasing the lead conversions to this categories.
- Welingak website and reference has higher number of lead conversion rate but count is low.
- Main focus should be on 4 major categories.

- **Do Not Email**

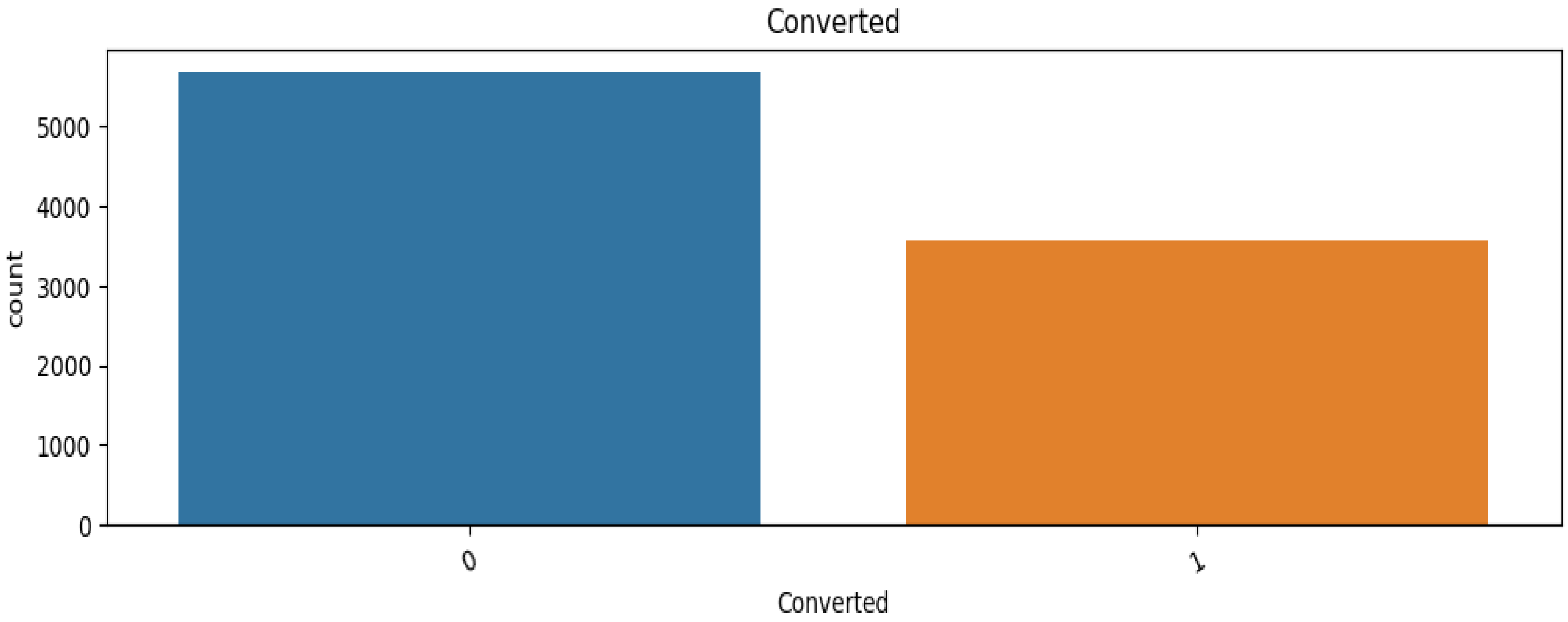




- This shows that most of the customers does not like to be emailed regarding the course.

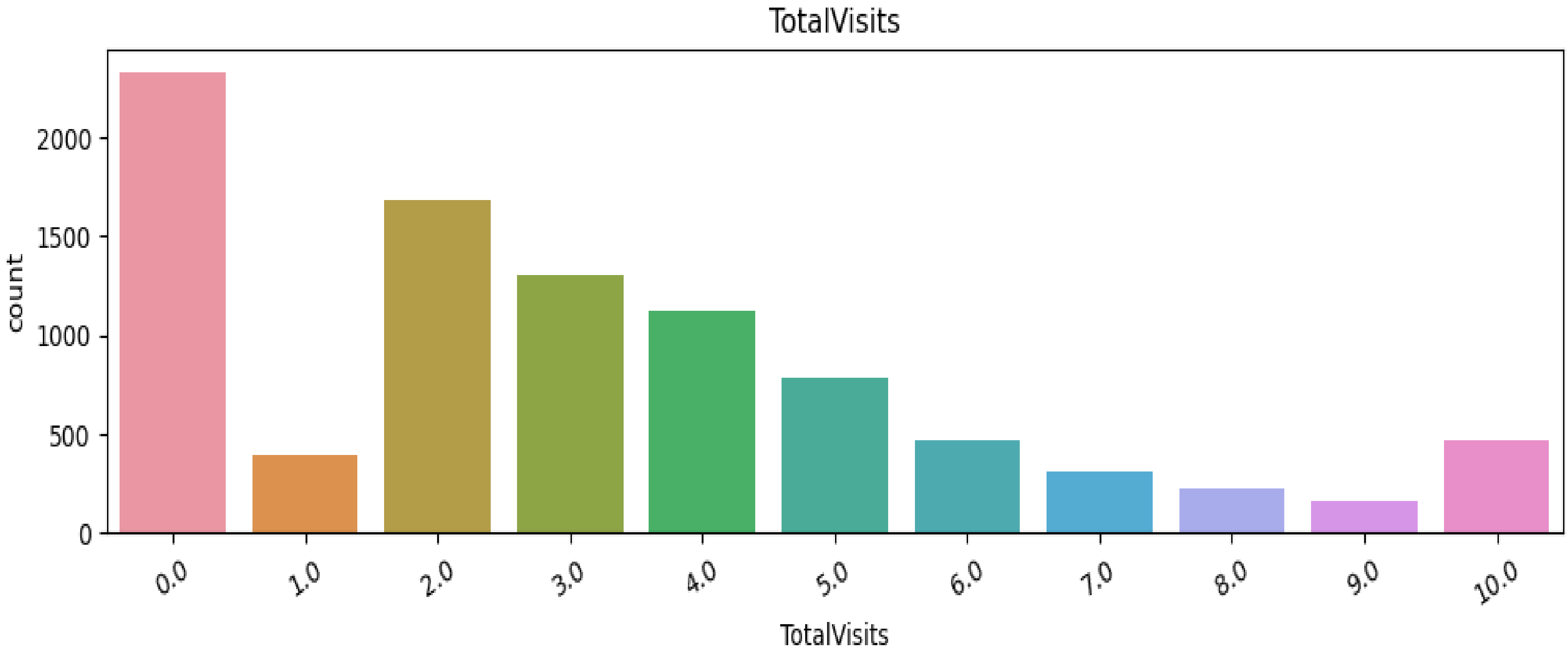
- Also focus on leads showing NO, as the conversion rate is very low for them.

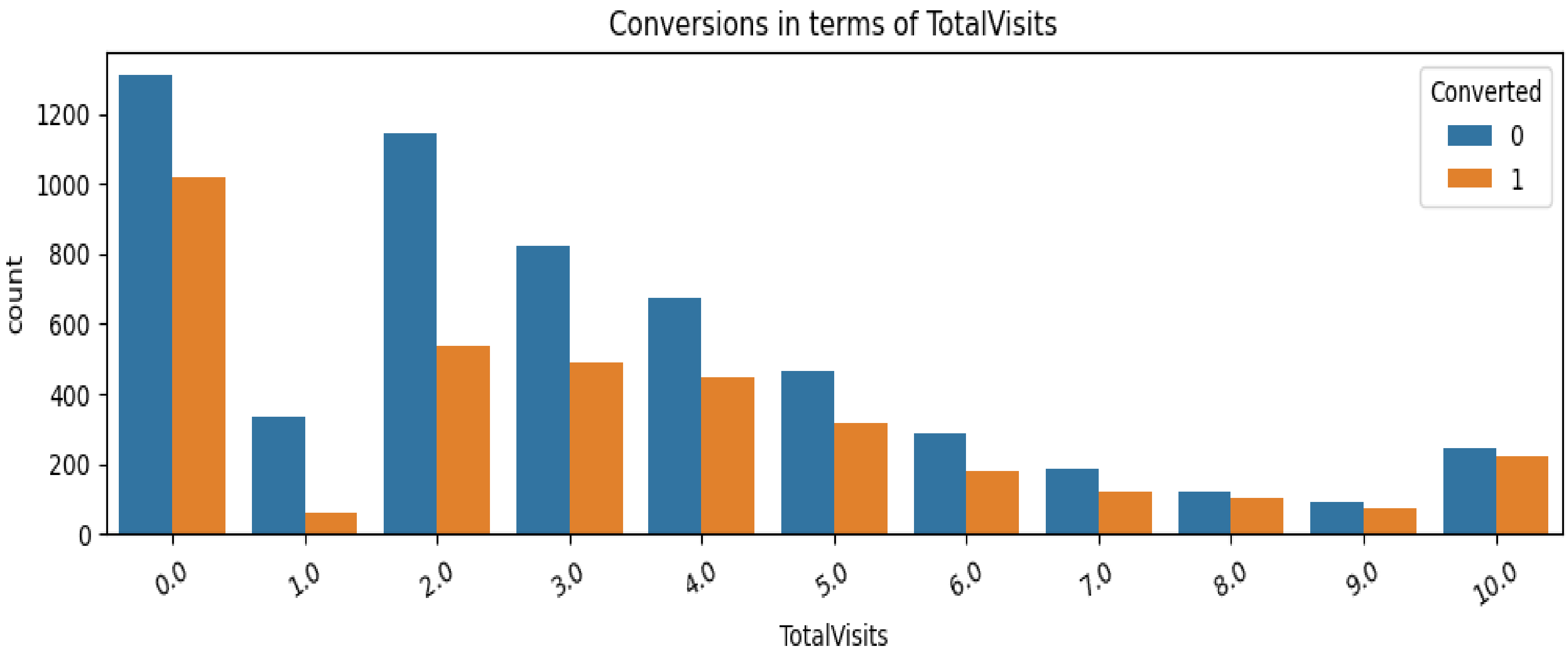
- **Converted**



The overall conversion rate is 38.54 %

- **Total Visits**

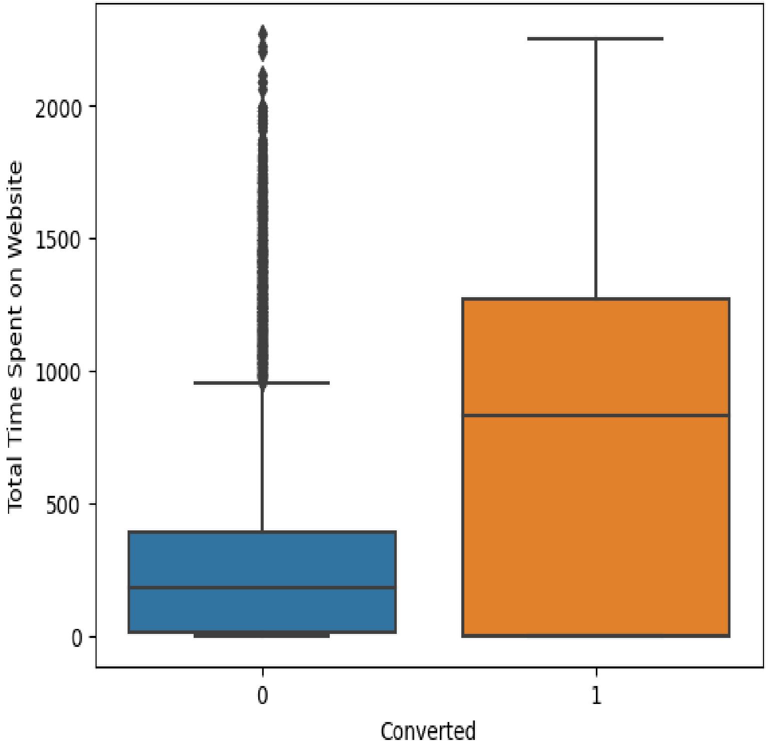


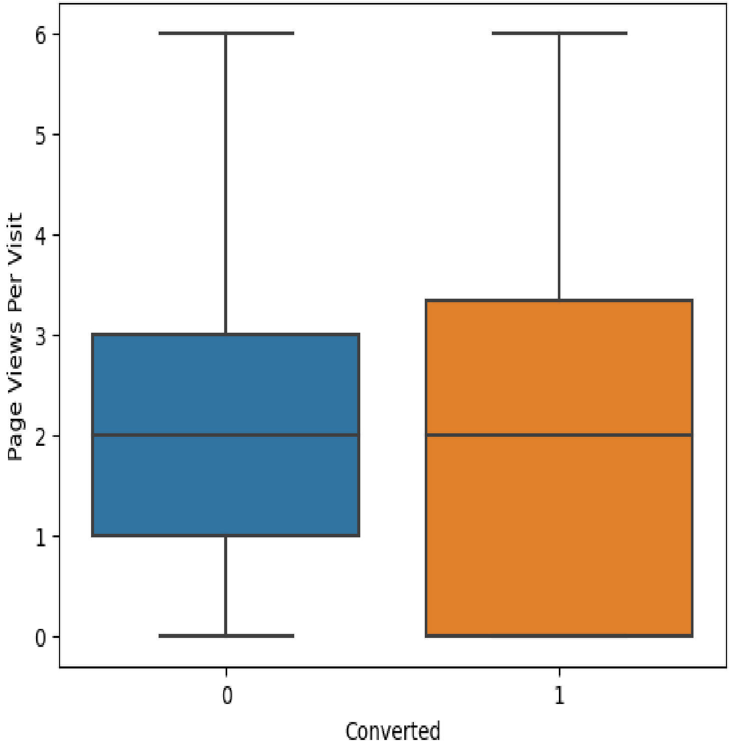


- Most of the leads have not visited, but have highest rate of conversion, as well when lead visited 8-10 times most of time it get converted

- **Total Time Spent on Website**

- **Page Views Per Visit**

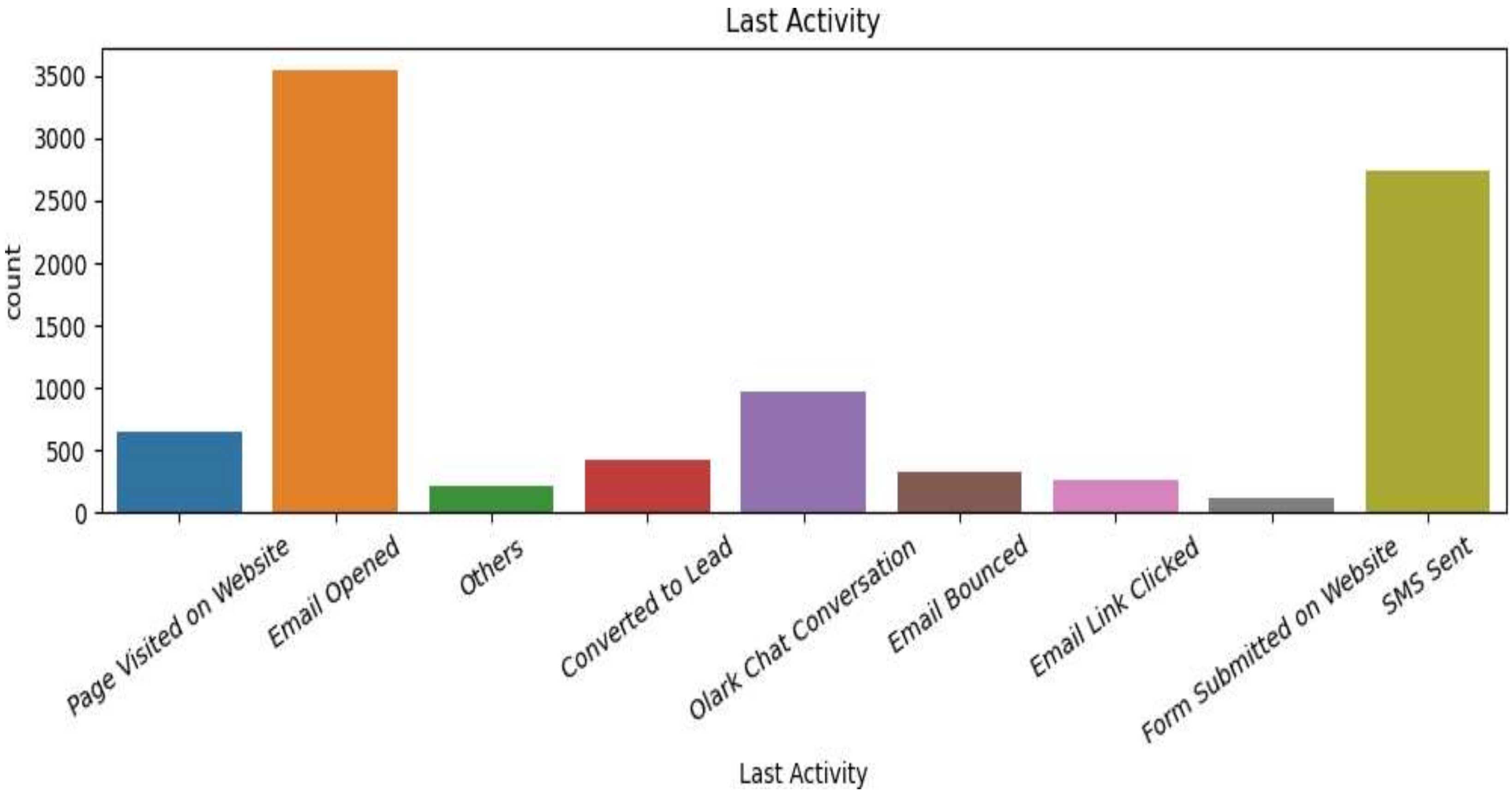


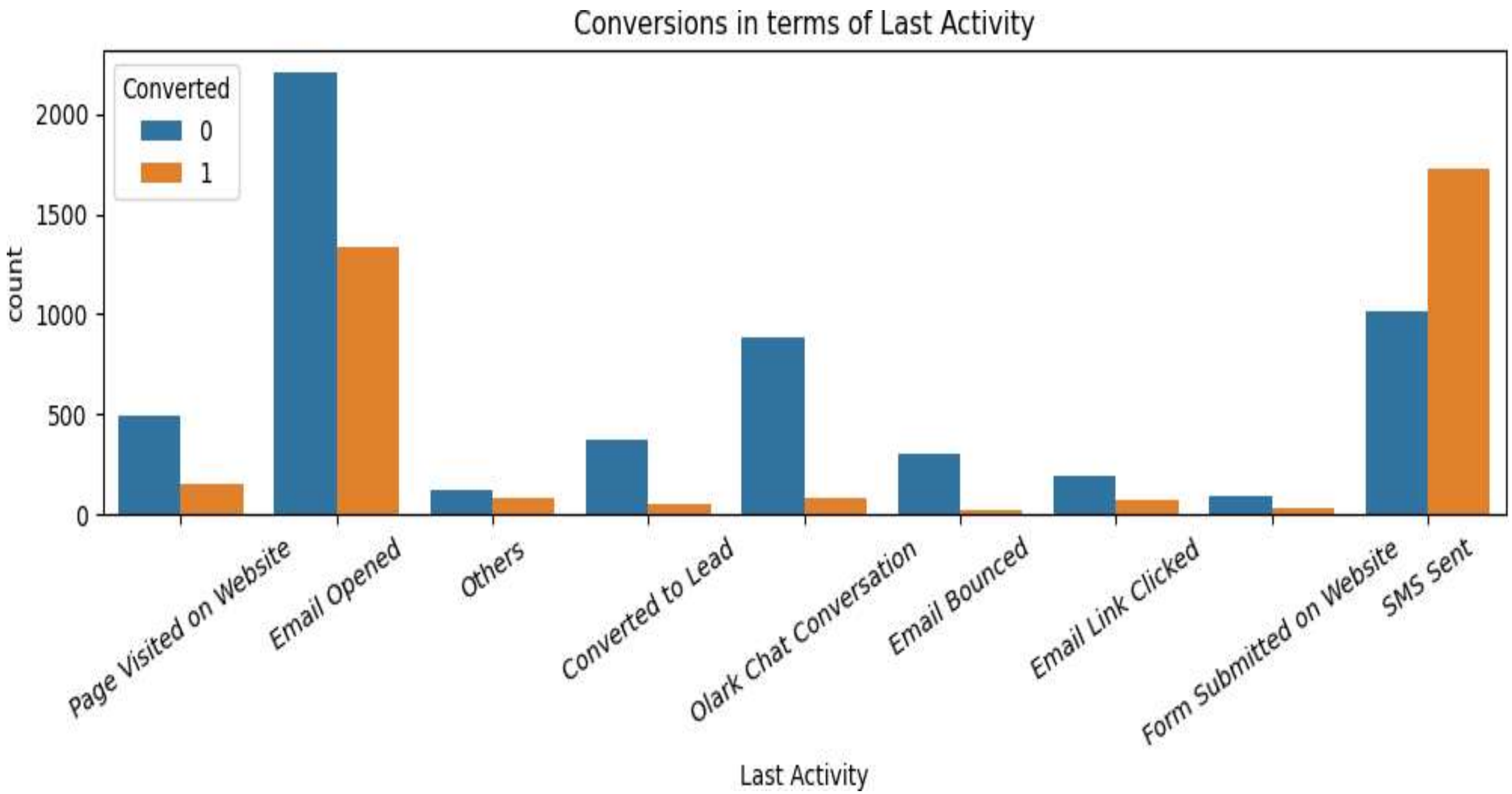


- This shows that more the time spent on website more are the leads that got converted.
- People who have spent less time have not

- Median for both the opted and not opted is same.
- people who have visited 1-3 on an avergare have 50-50 percent chances of getting converted.

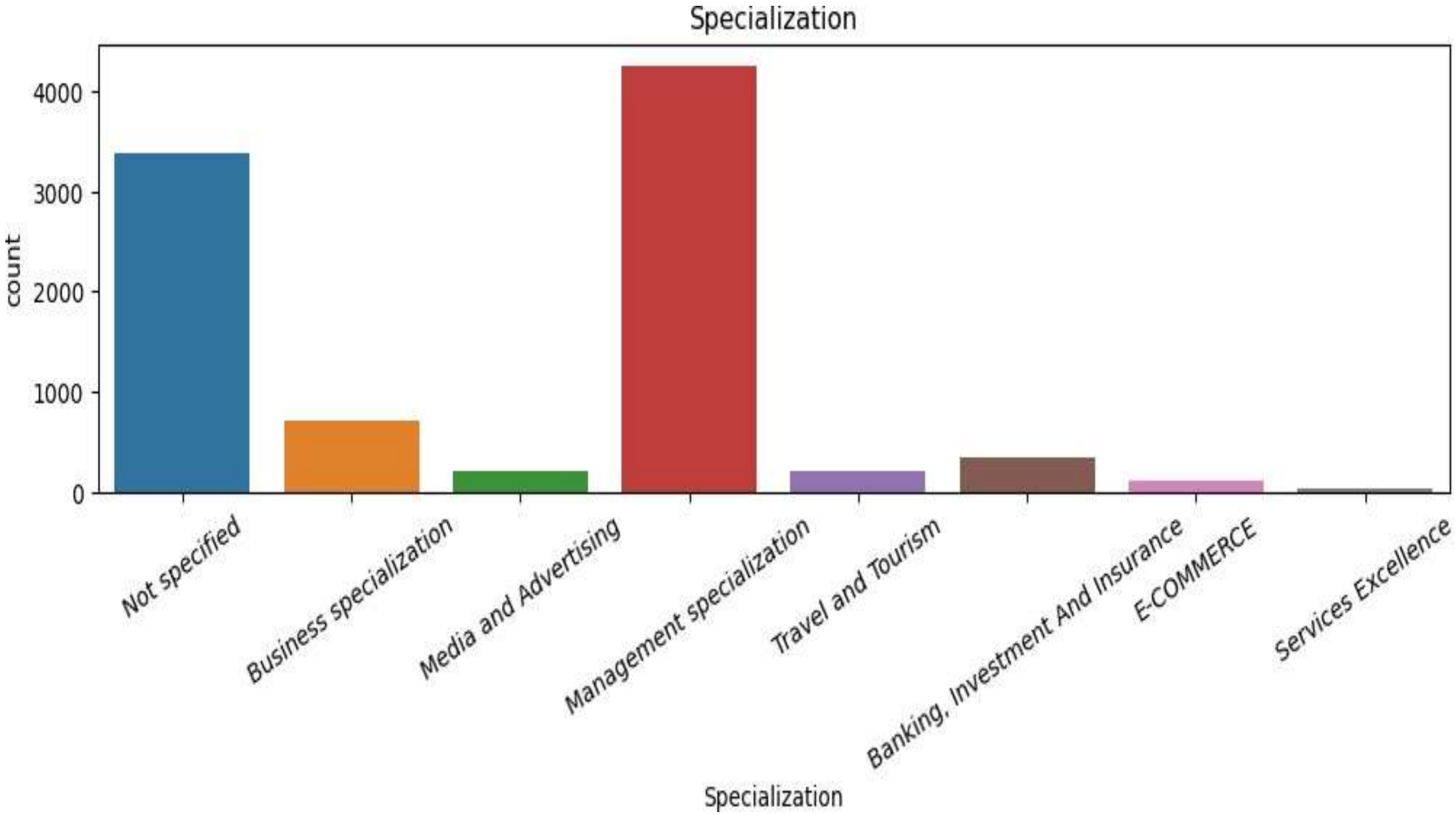
- **Last Activity**

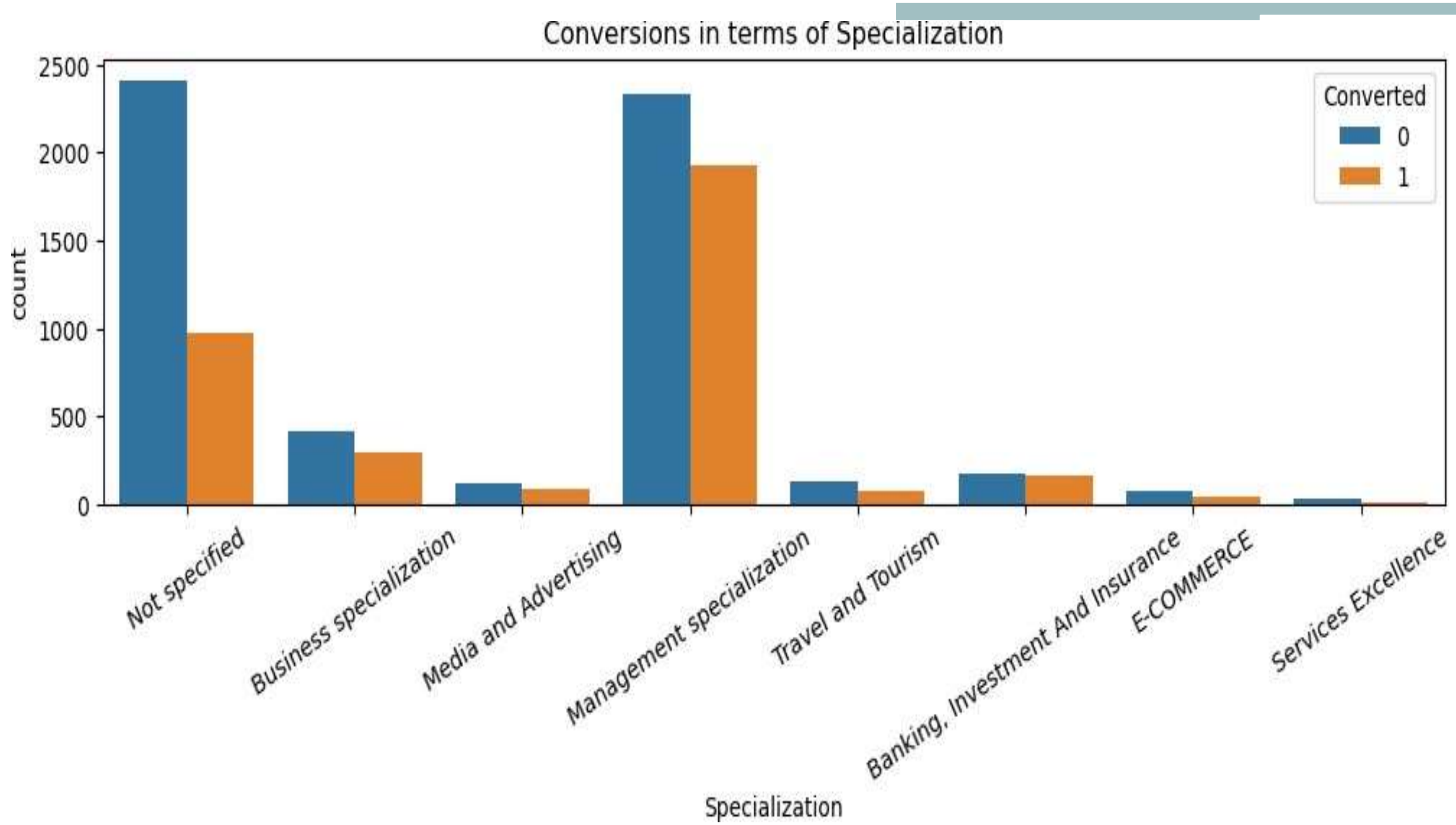




- Most leads are generated by email opened and SMS sent.
- Conversion rate of SMS sent is good so try to improve it more.
- Need to focus more on Olark chat as the conversion rate is poor.
- Focus more on Page visited more on website, Converted to lead, Email bounced.

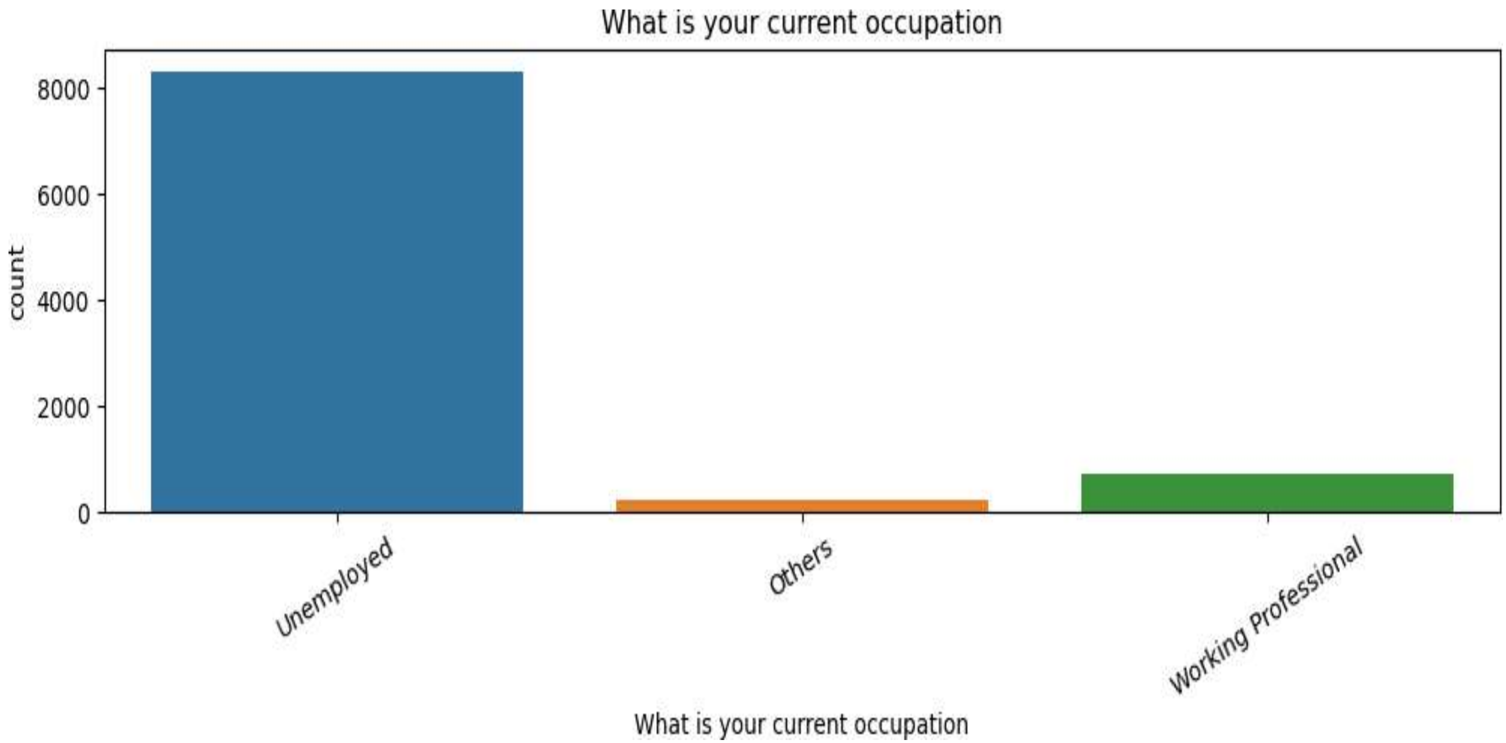
- **Specialization**

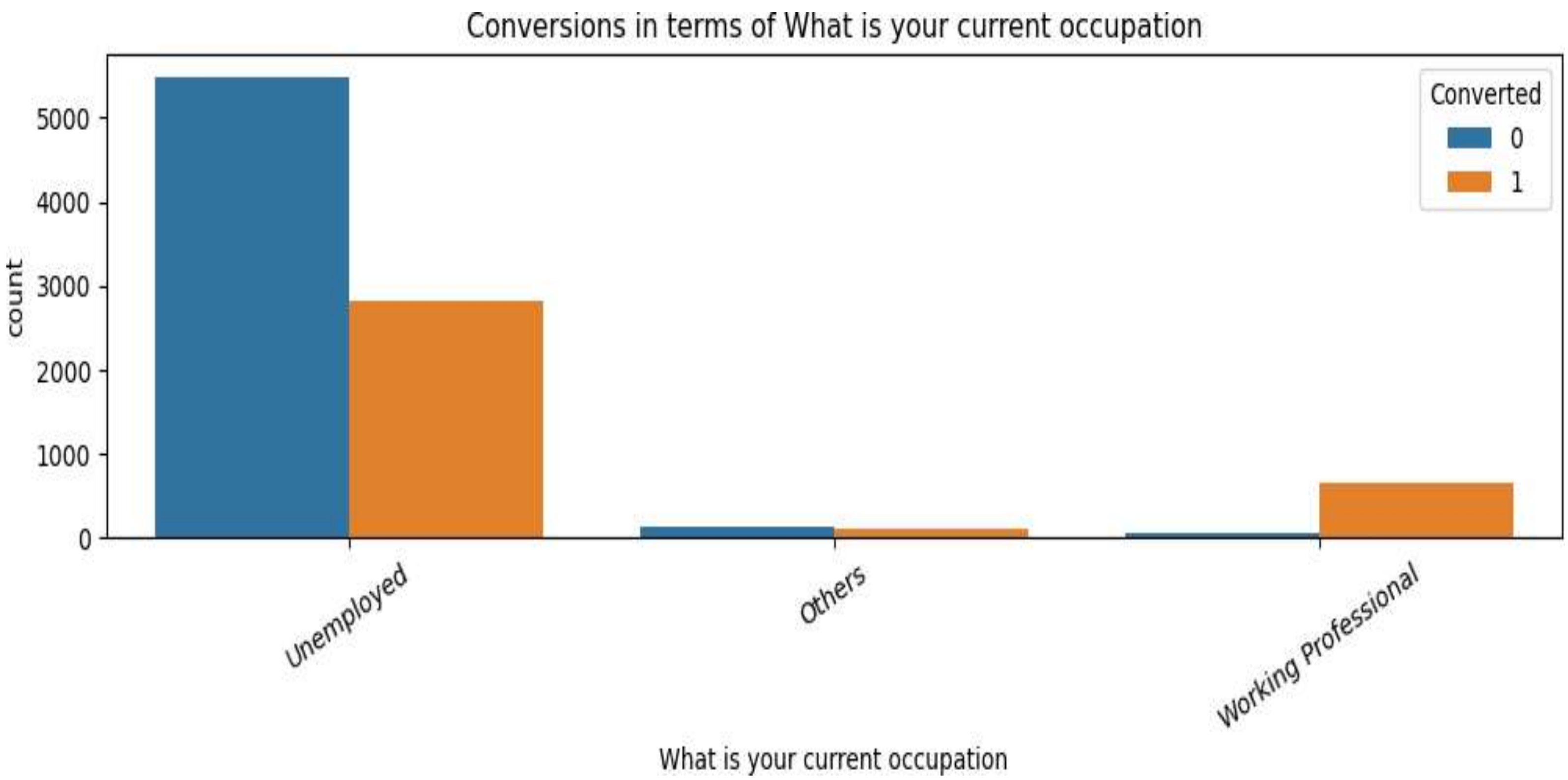




- Not specified and Management specialisation produce the most number of successful
  leads.
- Focus on these 2 categories and try to increase the lead conversion rate.
- Management specialization has highest leads and Service excellence has lowest leads.

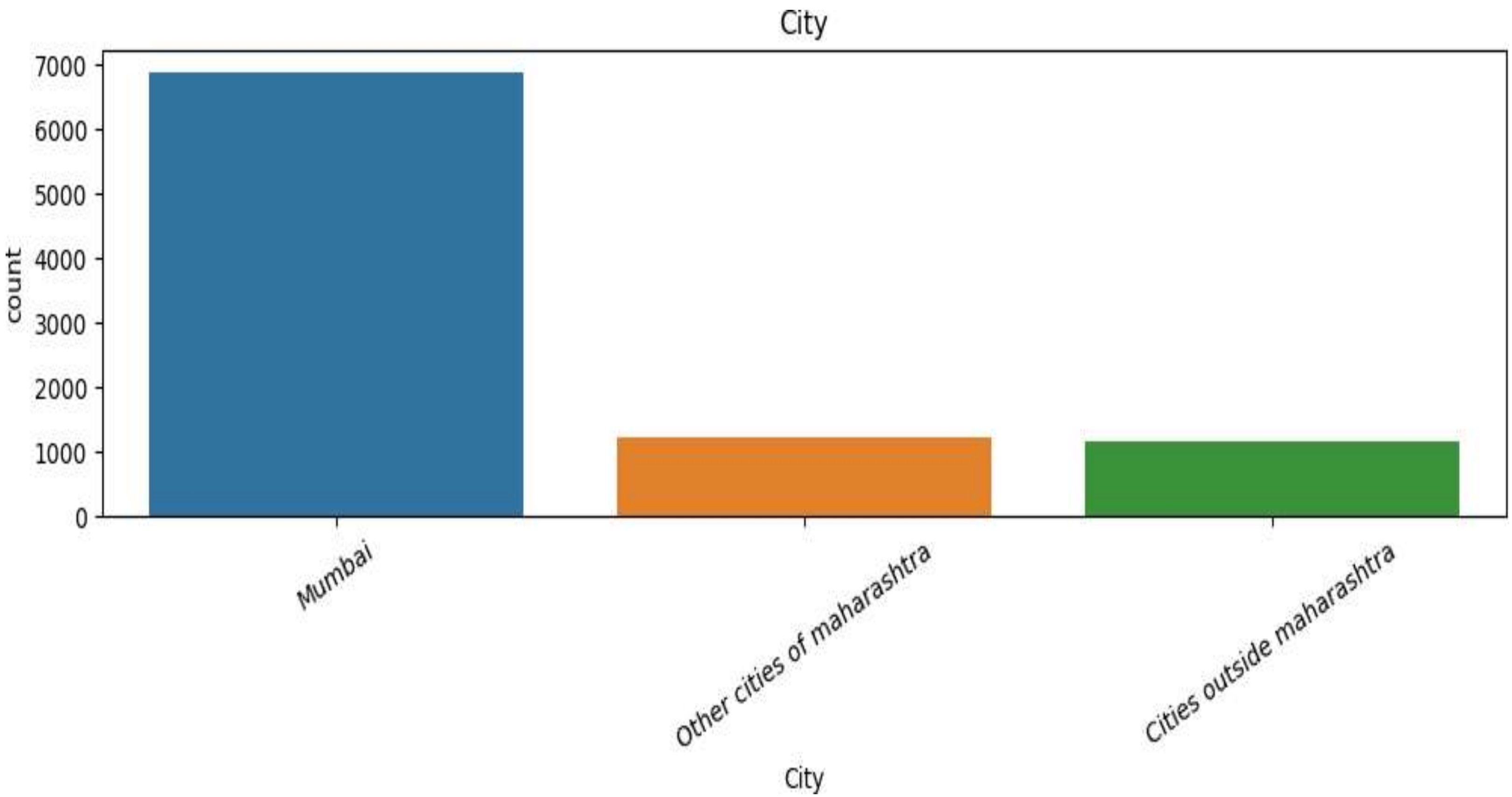
Current occupation

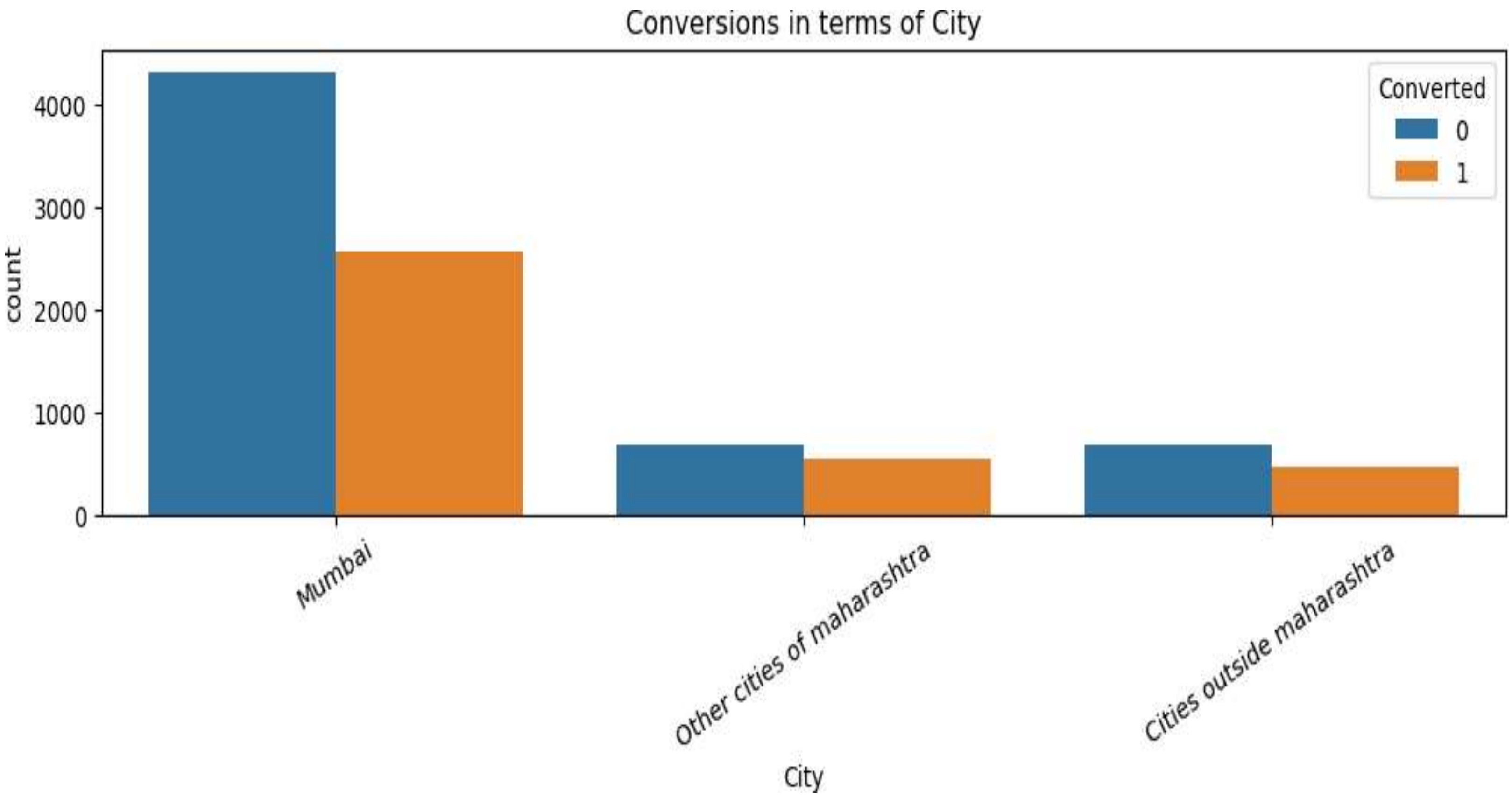




- Most of the costumers are unemployed and the conversion rate is also not good.
- Conversion rate of Working professional is good so try to improve it more.
- Focus on Unemployed and try to increase conversion rate.

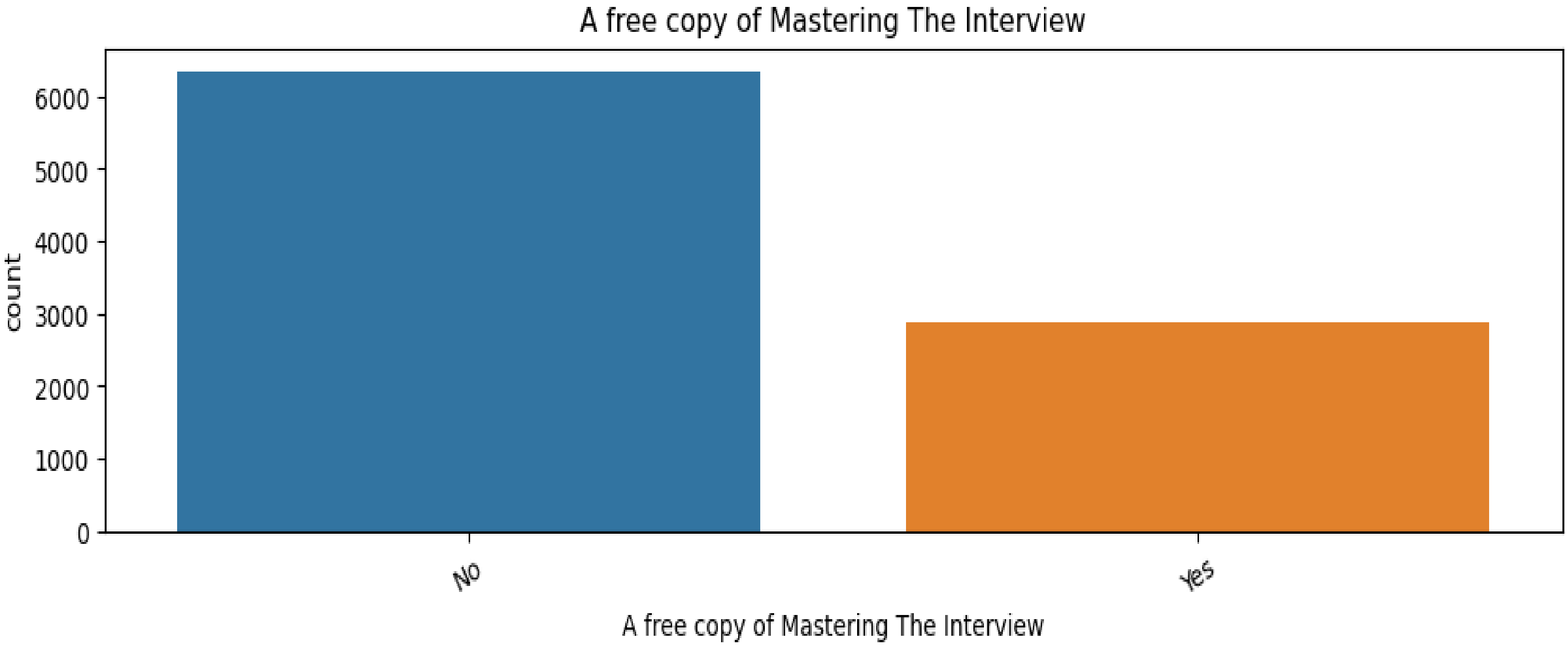
City

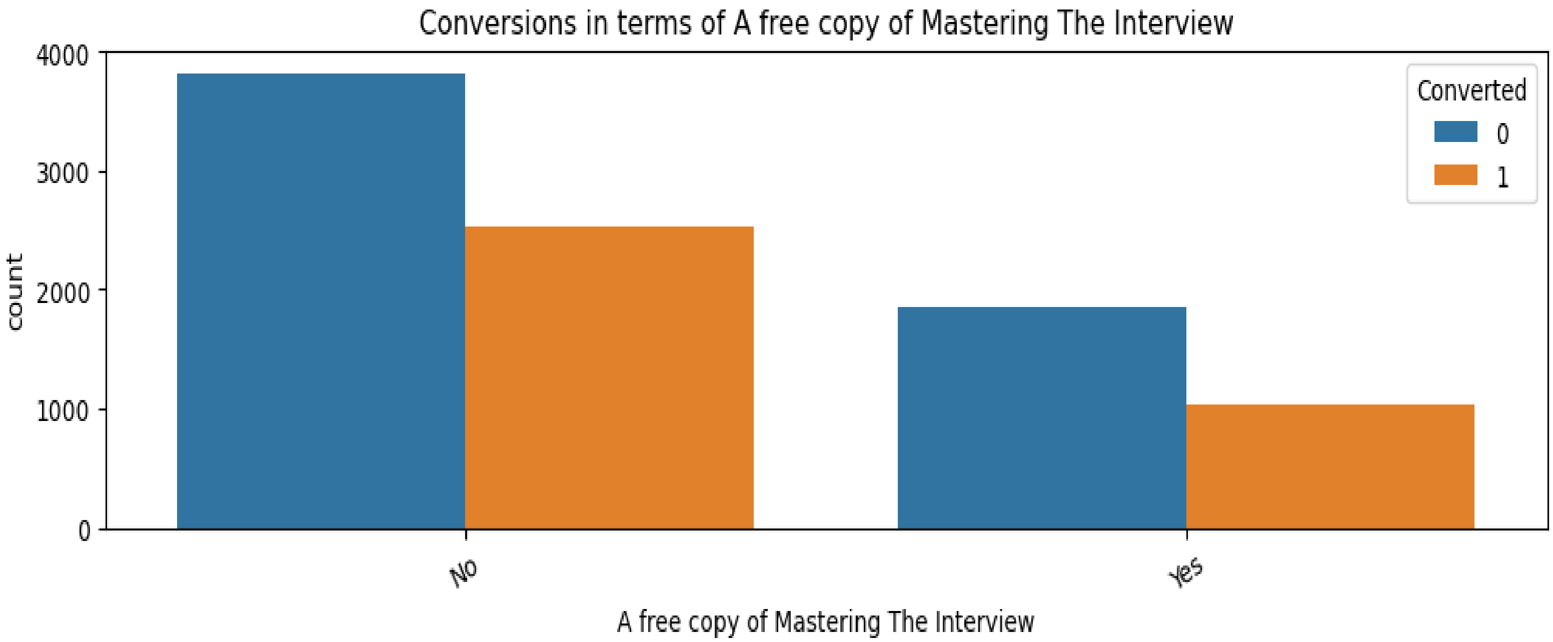

City
7000
6000
5000
4000
3000
2000
1000
0
count
Mumbai
Other cities of maharashtra
Cities outside maharashtra
City



- Most leads are generated in mumbai so focus on increasing lead conversion rate in mumbai.
- Other cities and cities outside maharashtra have equal conversion rates.

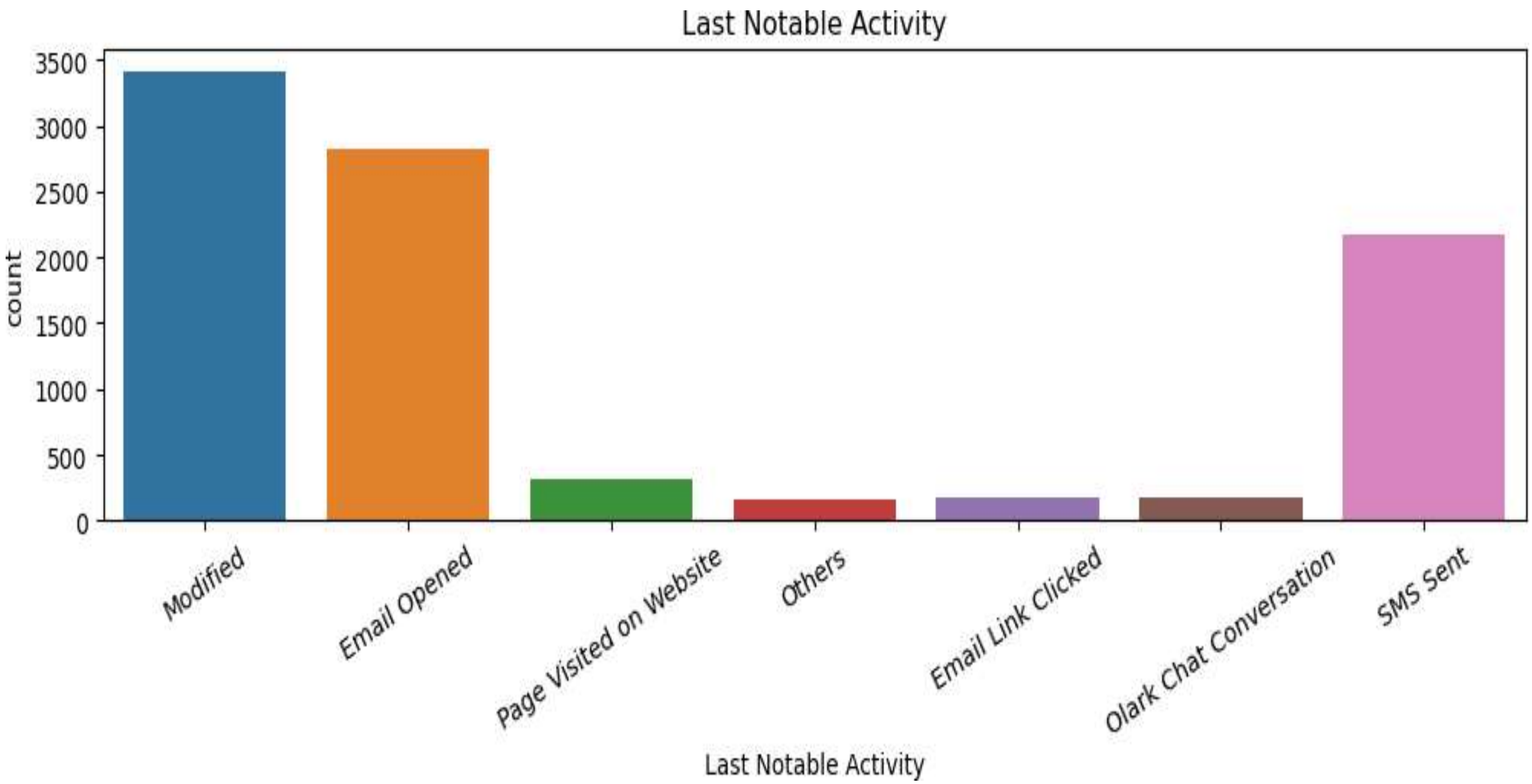
A free copy of Mastering The Interview

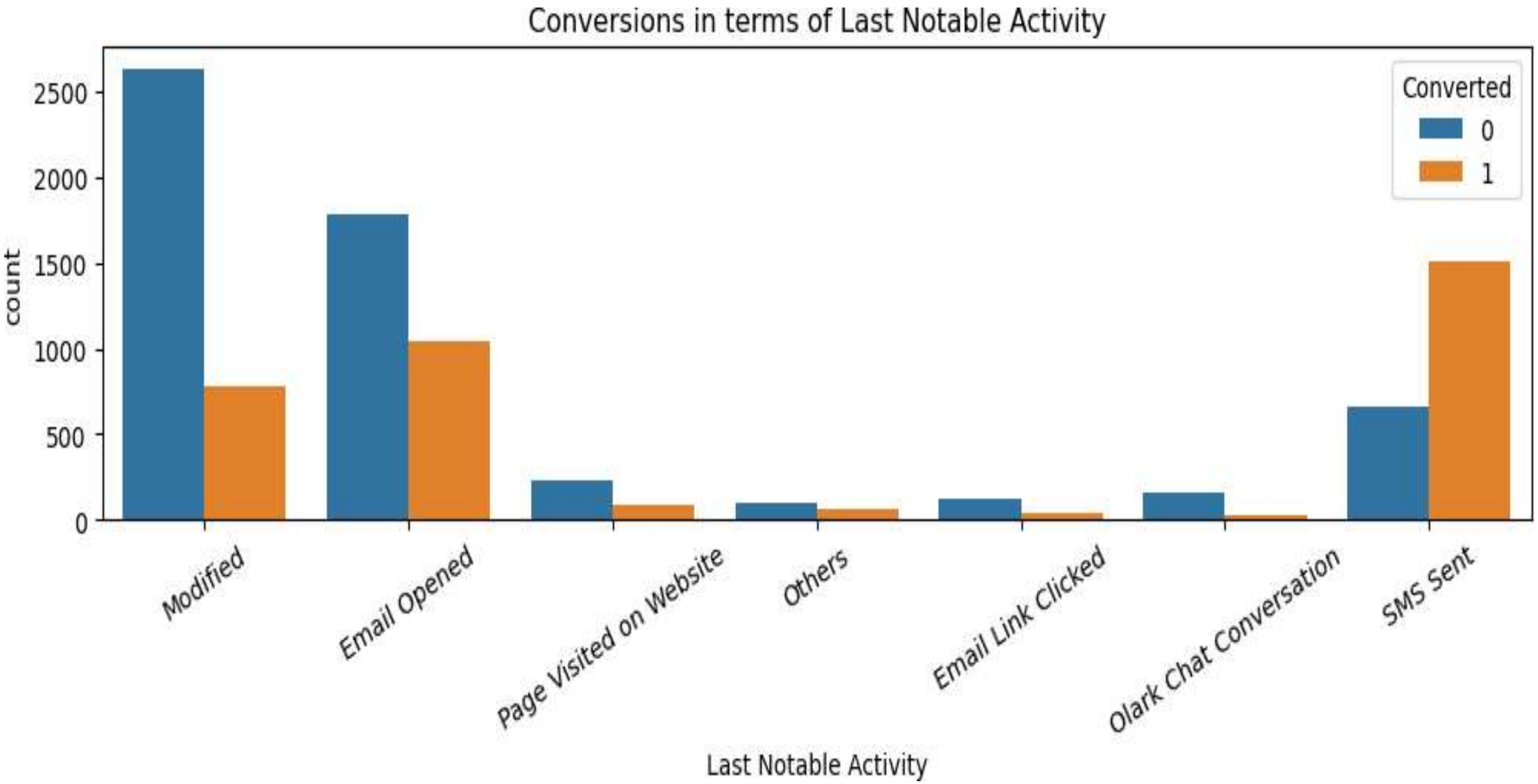




- Most customers have not opted for free copy of mastering the interview.

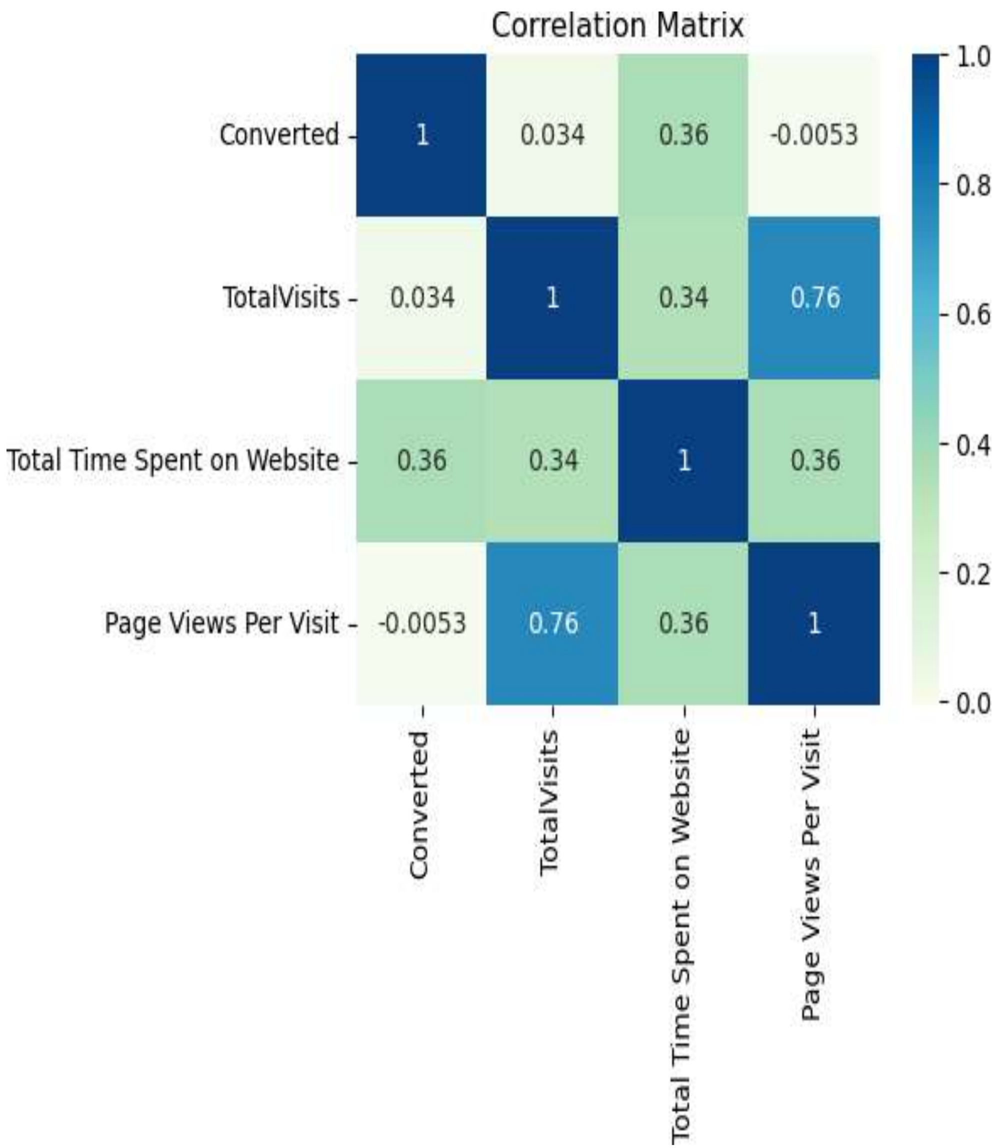
Last Notable Activity





- Modified and Email Opened has most number of leads but lead conversion is not up to the mark.
- Focus on Modified and Email Opened more to increase conversion rate.
- SMS sent has good conversion rate and so there is chance of improvance more.

Checking for numerical variables for collinearity



Correlation Matrix

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1	0.034	0.36	-0.0053
TotalVisits	0.034	1	0.34	0.76
Total Time Spent on Website	0.36	0.34	1	0.36
Page Views Per Visit	-0.0053	0.76	0.36	1

- Since we have high collinearity between Page views per visit and Total visitors we can drop any one of them.

Data Preparation

- Here we convert all data into numeric form (like Yes/No into 1/0)
- Create dummies for all categorical variables
- Performed train-test split &
- Performed scaling using Sklearn Library

Model Building

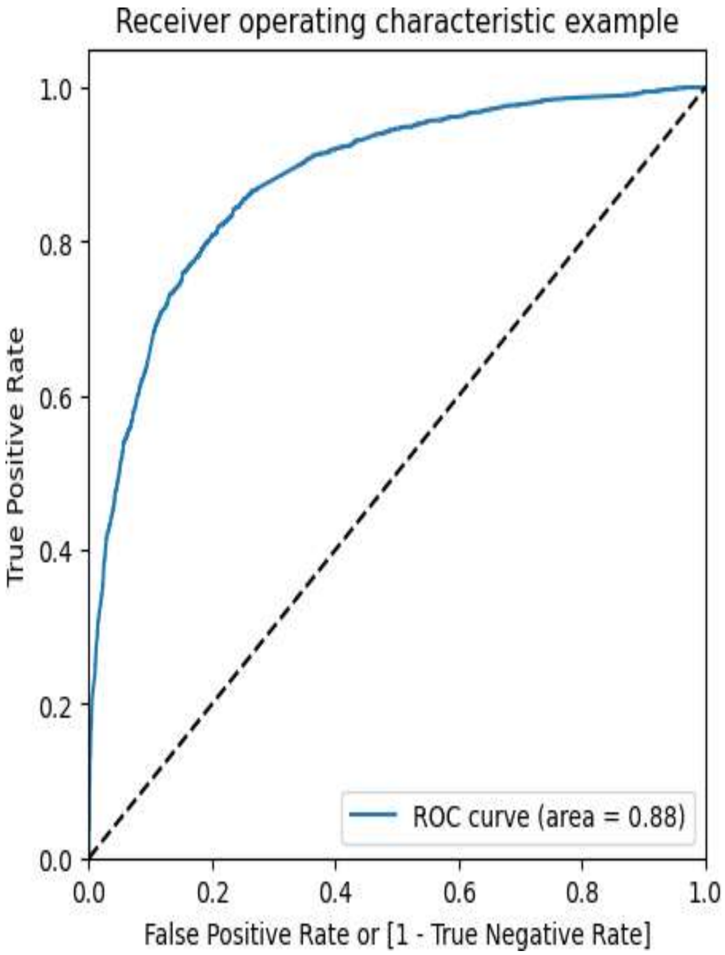
- Here we aim to build logistic regression model
- So we started with feature selection using RFE
- With the help of P value & VIF we reached to most suitable model

Model Evaluation

- Evaluating model with following parameter

Sr.No / Parameter	Parameter	Value
1	Accuracy	0.8157
2	Sensitivity	0.6995
3	Specificity	0.8873
4	False positive rate	0.1126
5	Positive predictive value	0.7927
6	Negative predictive value	0.8273

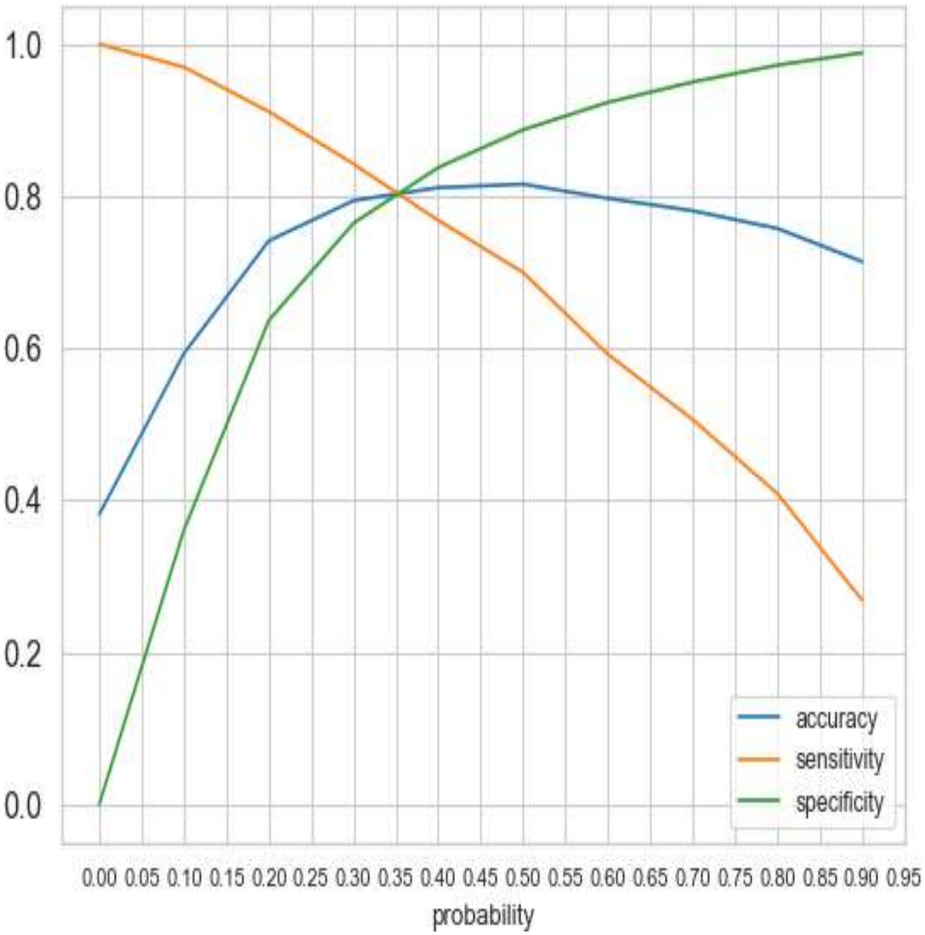
Plotting the ROC curve



- The ROC Curve should be a value close to 1. We are getting a value of 0.88 indicating a good predictive model.

- AUC stands for Area under the ROC curve shows that 88% of the predictions are correct. hence more the AUC better the model.

Finding Optimal Cutoff Point



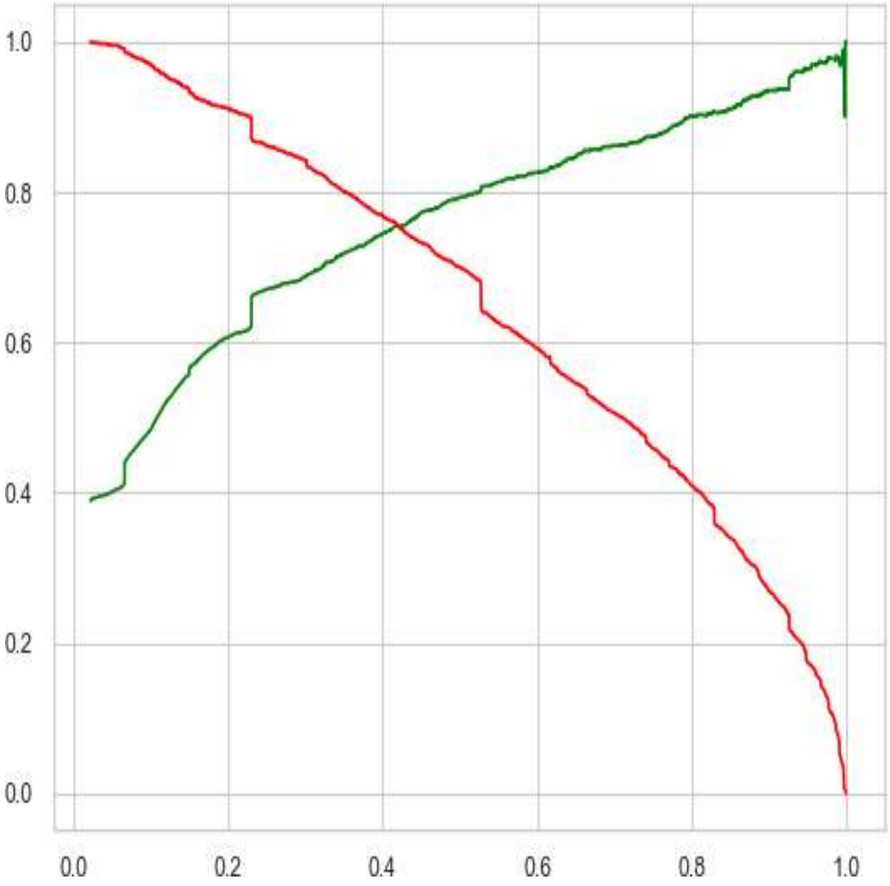
•From the curve, 0.35 is the optimum point to take it as a cutoff probability.

•It means that at this point Accuracy, Sensitivity and Specificity are all same.

Precision and Recall

- **Precision**
  - Precision measures how good our model is when the prediction is positive.
  - We need high precision in places such as recommendation engines, spam mail detection, etc. Where you don’t care about false negatives but focus more on true positives and false positives. It is ok if spam comes into the inbox folder but a really important mail shouldn’t go into the spam folder.
  - TP / TP + FP
  - **We got Precision Value 0.69**
- **Recall**
  - measures how good our model is at correctly predicting positive classes.
  - Models need high recall when you need output-sensitive predictions. For example, predicting cancer or predicting terrorists needs a high recall, in other words, you need to cover false negatives as well. It is ok if a non-cancer tumor is flagged as cancerous but a cancerous tumor should not be labeled non- cancerous.
  - TP / TP + FN
  - **We got Recall Value 0.84**

Precision and Recall tradeoff



- This graph shows that Precision and Recall are inversely related as one increases other decreases.

Final Observation

We make prediction on test data set & got following results compare to train data set

- **Train Data**
- Accuracy - 80.00%
- Sensitivity - 84.22%
- Specificity - 76.43%
- Precision - 69.00%
- Recall - 84.22%

- **Test Data**
- Accuracy - 81.27%
- Sensitivity - 84.22%
- Specificity - 76.43%
- Precision - 77.00%
- Recall - 75.06%

Conclusion

- We got around 1% difference on train and test data's performance metrics.This implies that our final model didn't overfit training data and is performing well.
- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.
- Depending on the business requirement, we can increase or decrease the probability threshold value with in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.

Recommendations:

Total Time Spent on Website

From The Boxplot we can see that number of constumers who have converted spent more time on the website.so try to make them engage in the website and increase their chances of getting converted.

Try to give them offers exclusive on websites so they gets attracted to it.

Lead Source

Olark Chat: We can see that olark chat has majority of lead conversions in terms of count, so focus on increasing the lead conversions to this categories.

Welingak Website: Welingak website has higher number of lead conversion rate but count is low.so you can improve and learn from here that might help you in improving other categories.

Lead Origin

Lead Add Form:

Lead add form has higher number of lead conversion rate so you can improve and learn from here that might help you in improving other categories.

Current Occupation

- *Unemployed:*
    - Unemployed has highest number of conversions by count but conversion rate is low so focus on increasing the conversion rate on unemployed as it is obvious that unemployed will be most interested in getting a new job by learning something new.
- *Others:*
    - Others contains Student, Housewife and Businessmen so focus on them more, try to increase their conversion score to more than 35 which are potential customers.

Last Activity

- *Olark Chat Conversation:*
    - Need to focus more on Olark chat as the conversion rate is poor.
- *Email Bounced:*
    - The conversion rate is not good seems to be very poor so focus here also.
- *Email Opened:*
    - Most leads are generated by email opened so this is an important category that you should focus on. Although having most conversion counts by number the conversion rate is not good.

Last Notable Activity

- SMS Sent:
  - SMS sent has good conversion rate and so there is chance of improvance more.
- Others:
  - Others has many categories such as 'Unreachable','Unsubscribed','Email Bounced','Had a Phone Conversation', 'Approached upfront','View in browser link Clicked','Email Received','Email Marked Spam','Form Submitted on Website', 'Resubscribed to emails'.
  - Since the lead count and converted count is less try on increasing the number as your
  - leads score depends on this.

Specialization

- As most of the customers has opted to not provide their specialization so avoid using this feature. Focus on leads having conversion score more then 35 which are potential leads also known as 'Hot leads'

Thank You