# X EDUCATION - LEAD SCORING CASE STUDY

PRATIKSHA SHIRWALE, PRATIK KALE & KARTHIKEYAN SANKAR

# 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.

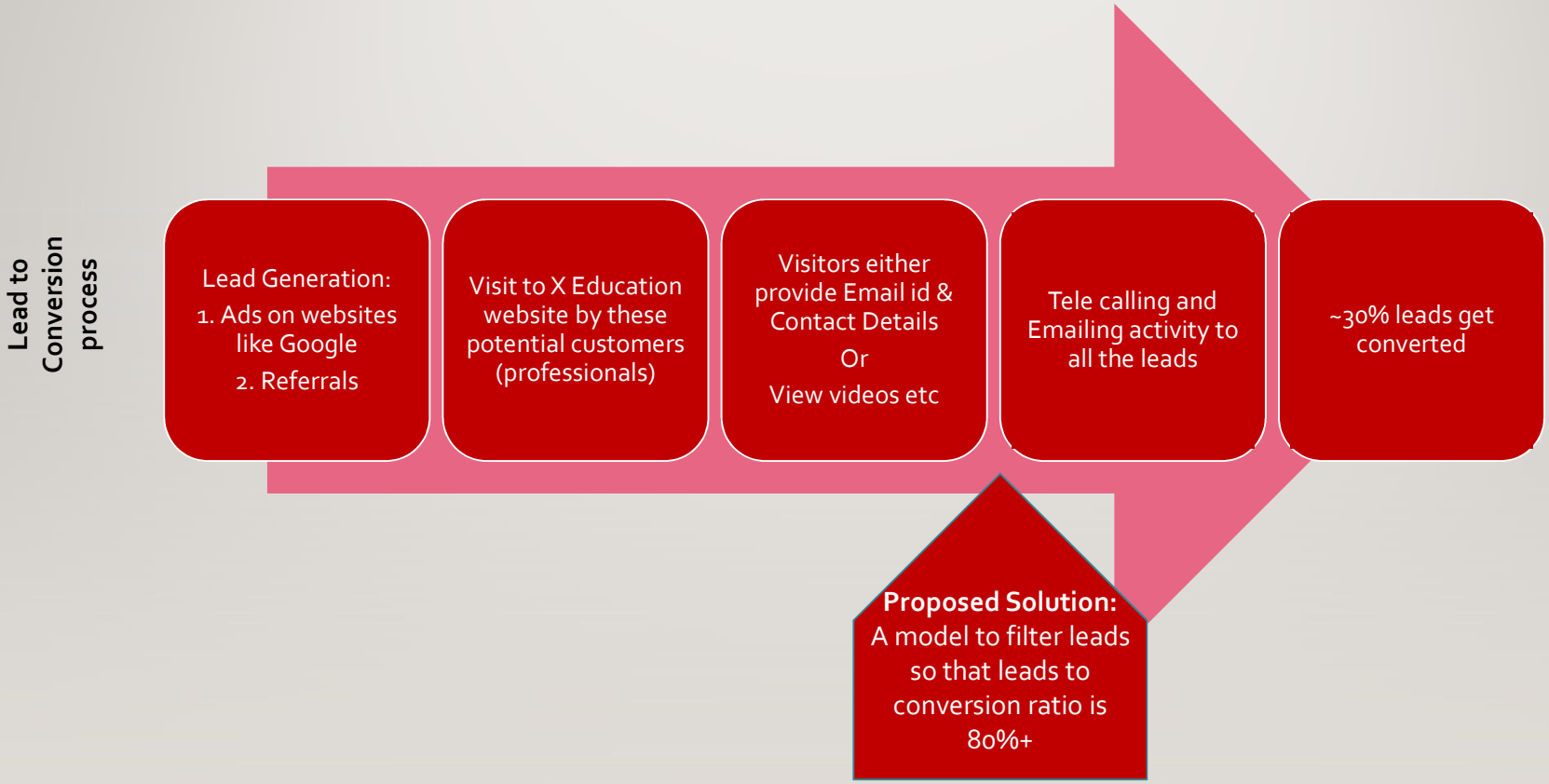
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified 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. 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.

# GOALS OF THE CASE STUDY

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.

# Lead – Conversion Process



# PROPOSED SOLUTION

## Selection of Hot Leads

**Leads Clustering**

We cluster the leads into certain categories based on their tendency or probability to convert, thus, getting a smaller section of hot leads to focus more on.

## Communicating with Hot Leads

**Focus Communication**

Since we would have a smaller set of leads to have communication with, we might make more impact with effective communication.

## Conversion of Hot Leads

**Increase conversion**

Since we focussed on hot leads, which were more probable to convert, we would have a better conversion rate, and hence we can achieve the 80% target.

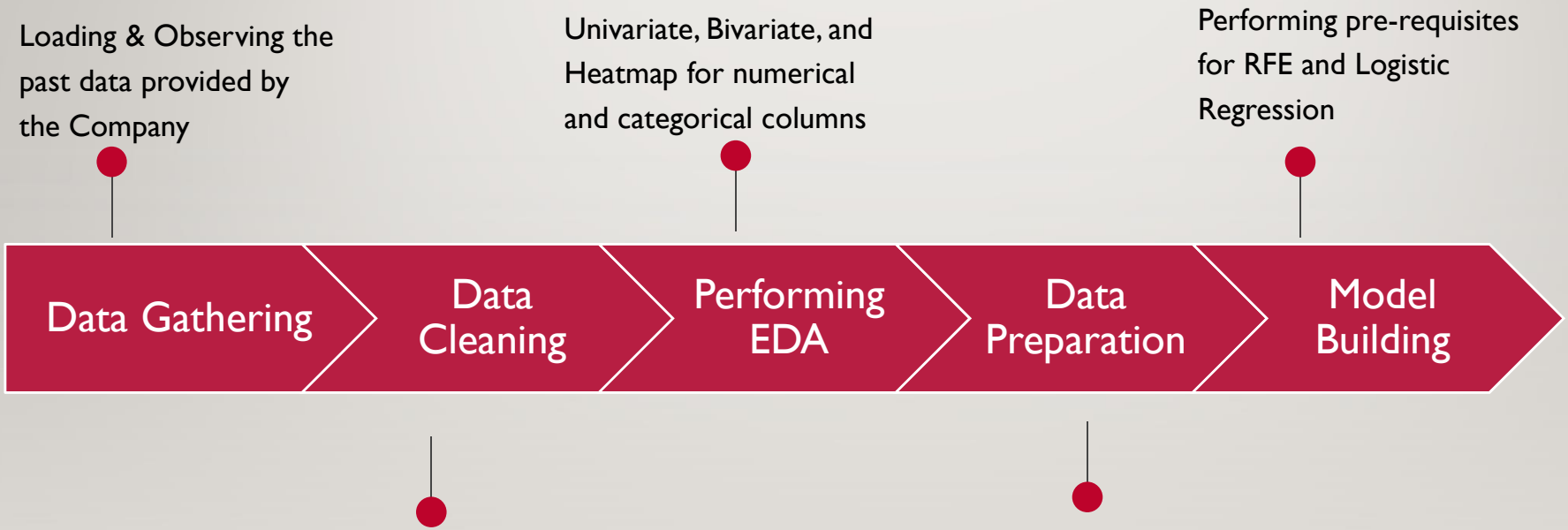
# SOLUTION

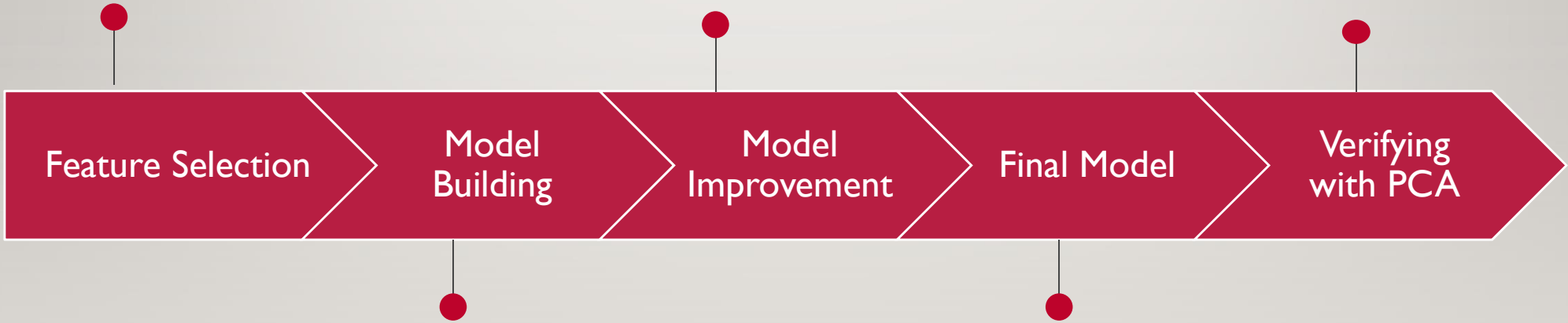
For our Problem Solution, the crucial part is to accurately identify hot leads.

The more accurate we obtain the hot lead, the more chance we get of higher conversion ratio.

Since we have a target of 80% conversion rate, we would want to obtain a high accuracy in obtaining hot leads.

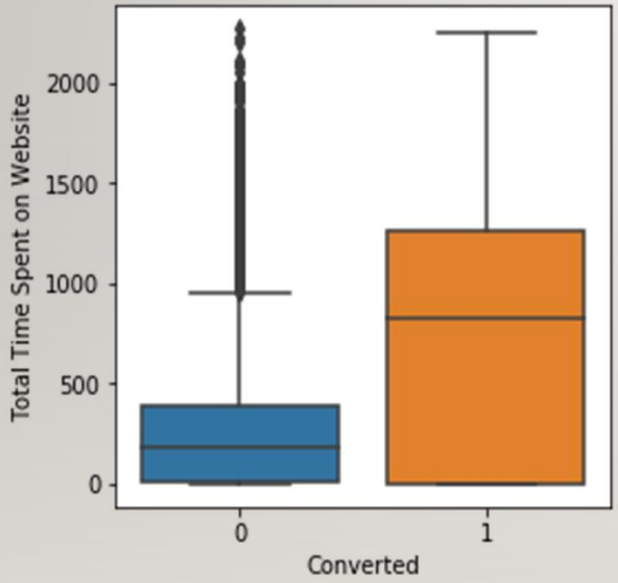
# IMPLEMENTATION

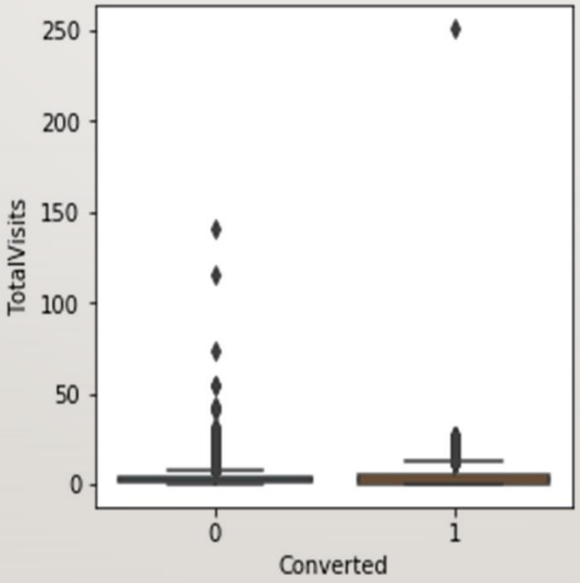


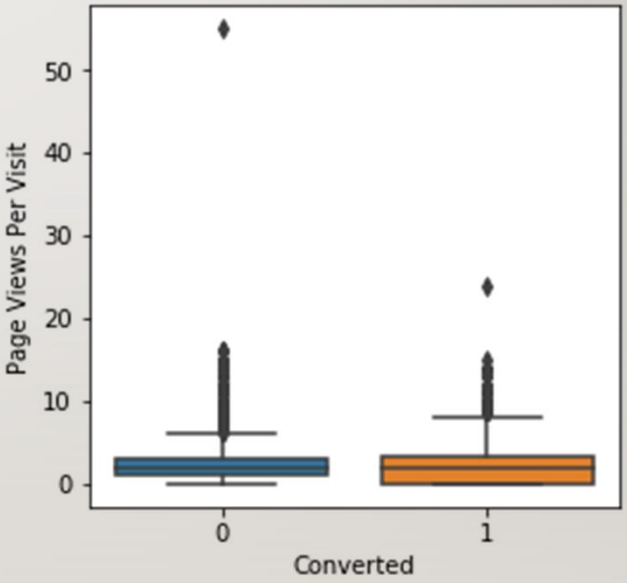

Selection of top 25 features using RFE
Reduction of columns and Model re-building
Verifying our Final Model Accuracy etc. with model built with PCA
Feature Selection
Model Building
Model Improvement
Final Model
Verifying with PCA
Model building using RFE for selected columns
Final Model Analysis and performance on Test Data

# PLOTS

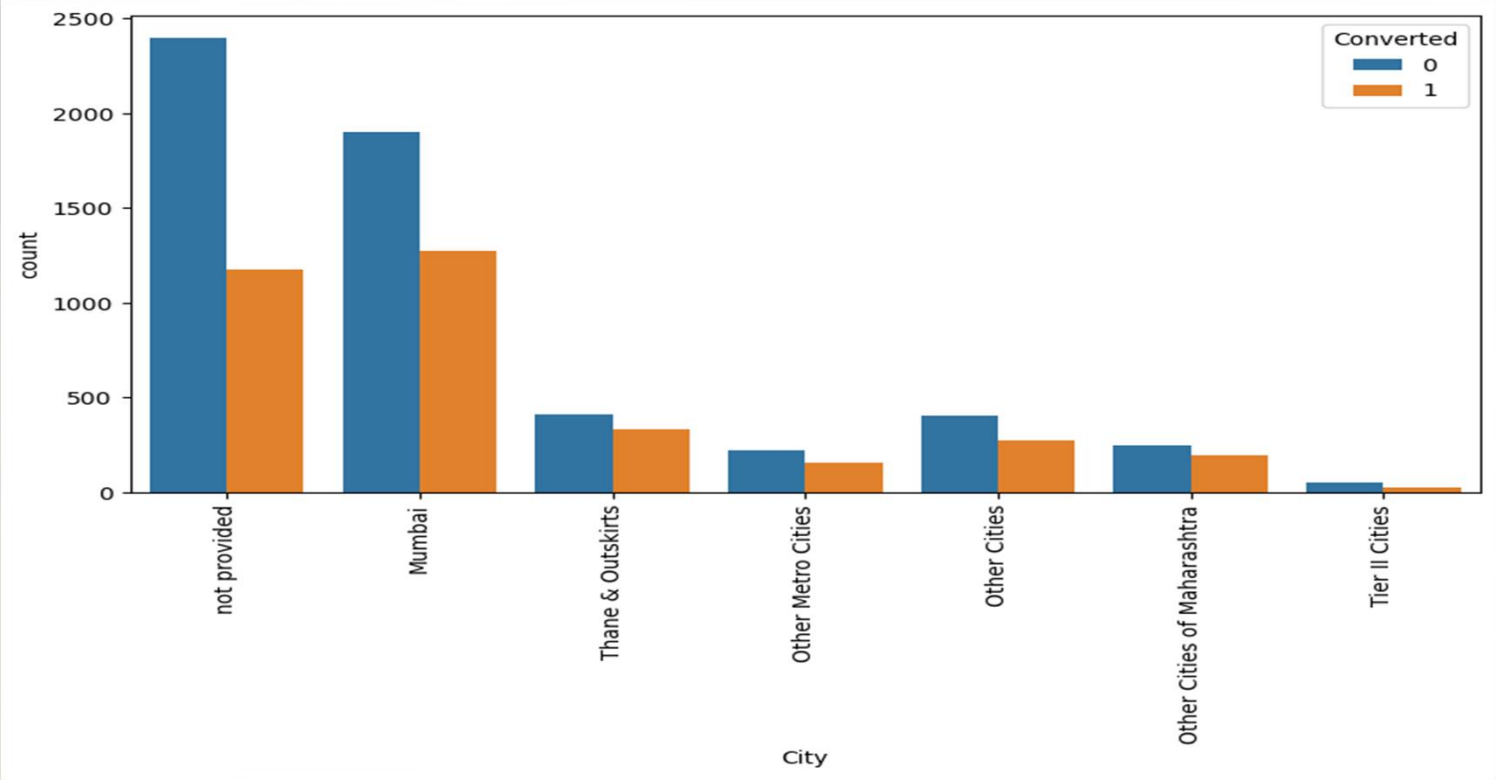
**EDA plots depicting variation in numerical columns for those who Converted and those who didn't.**



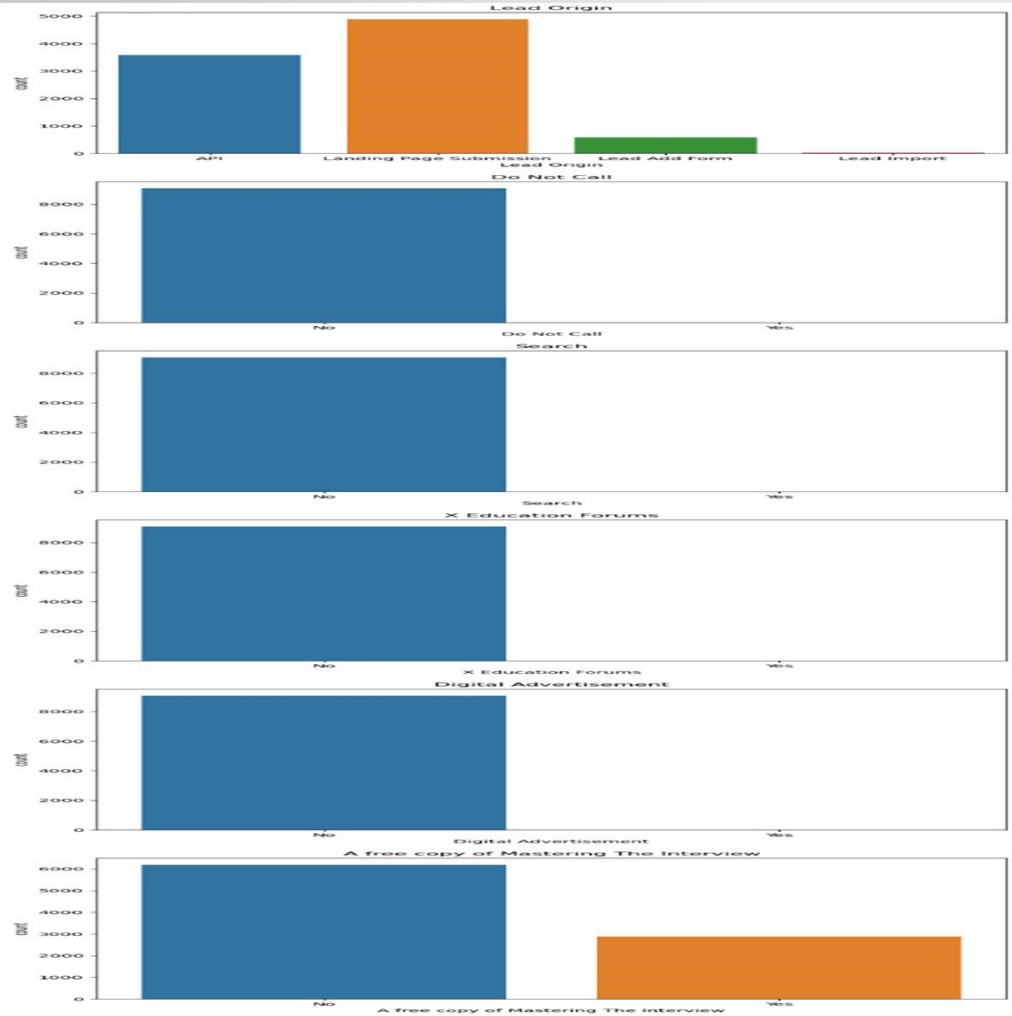



## Spread of Cities



# Observations









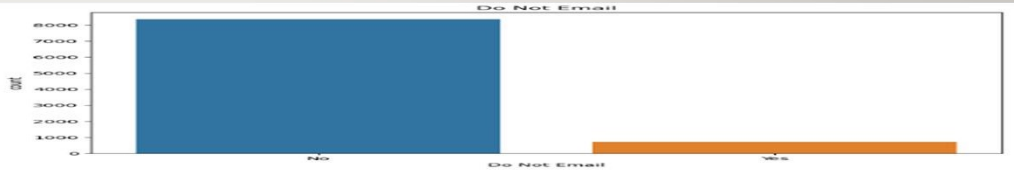


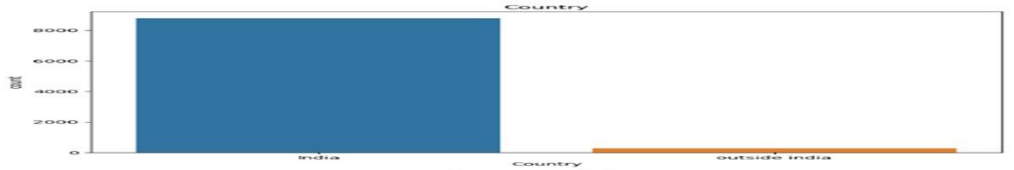


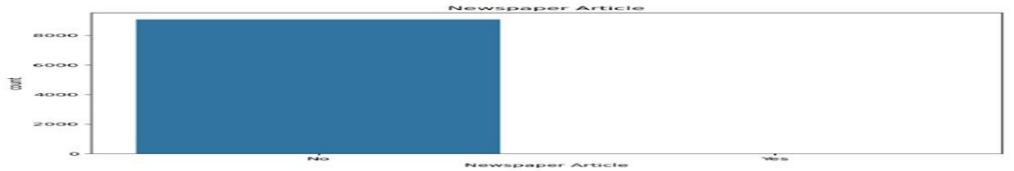


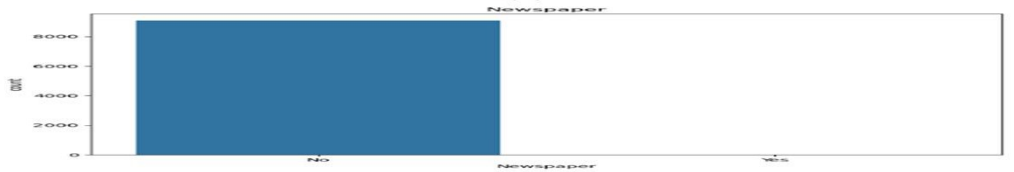


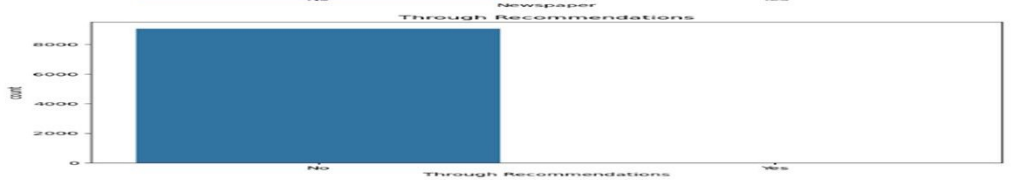


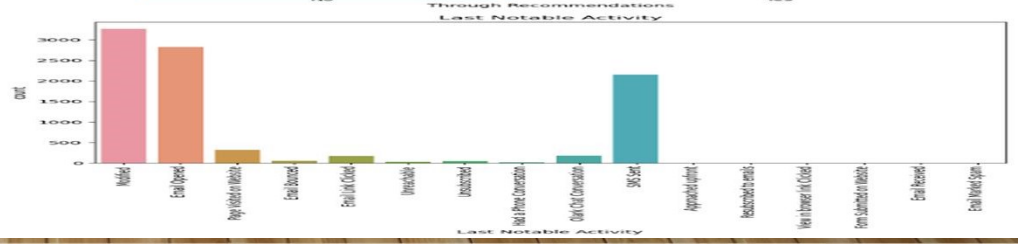


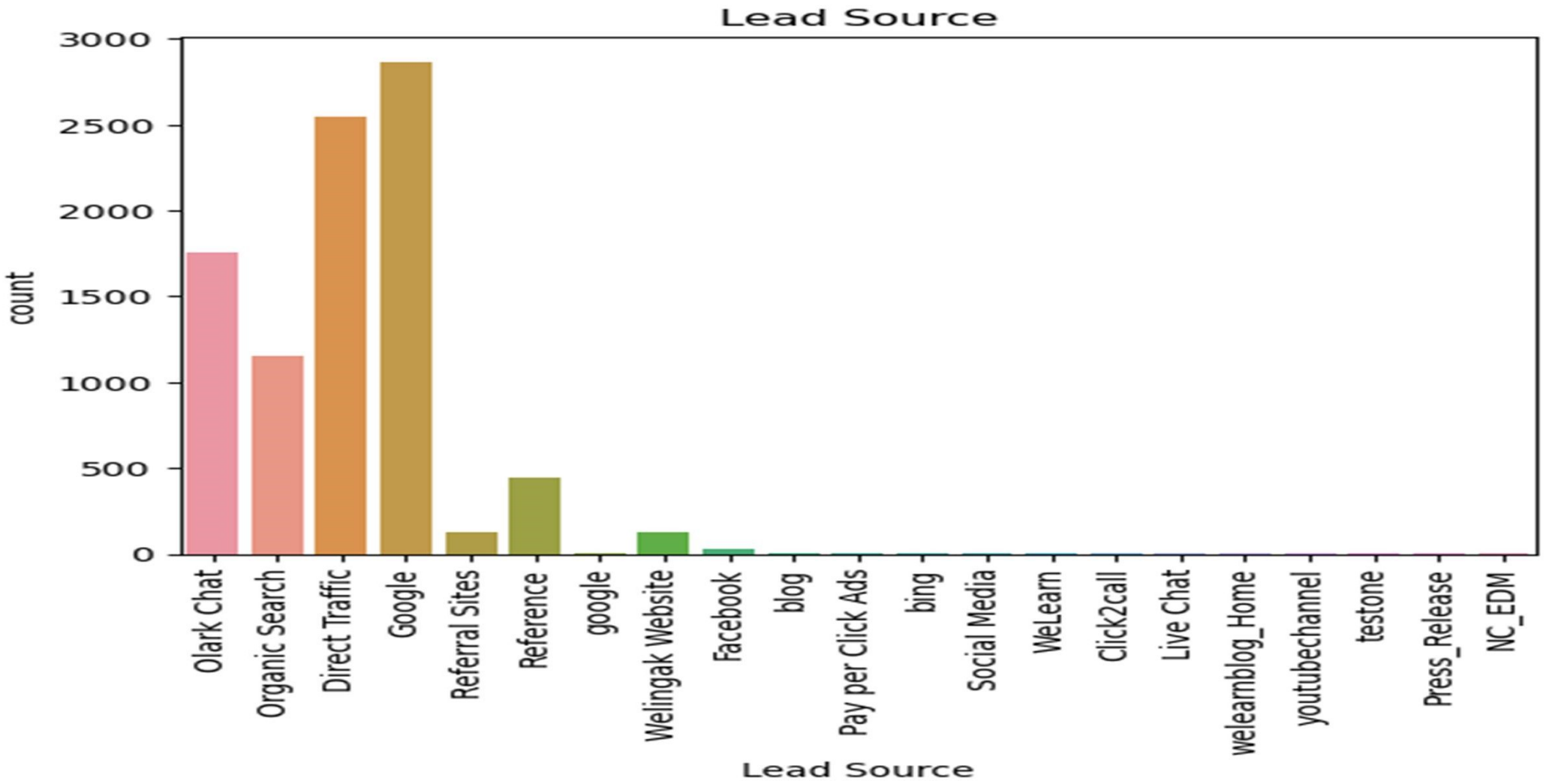



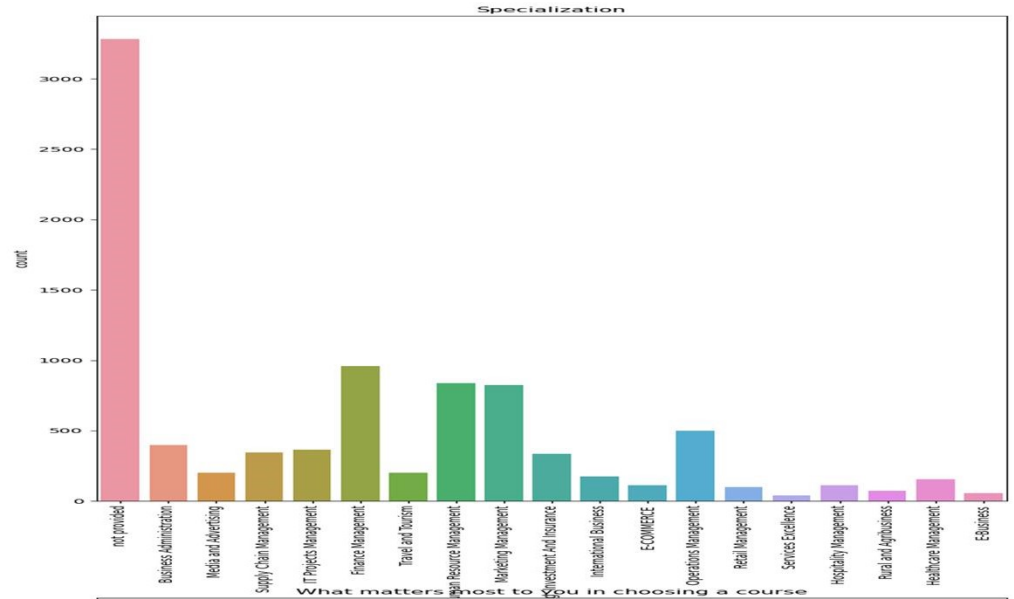



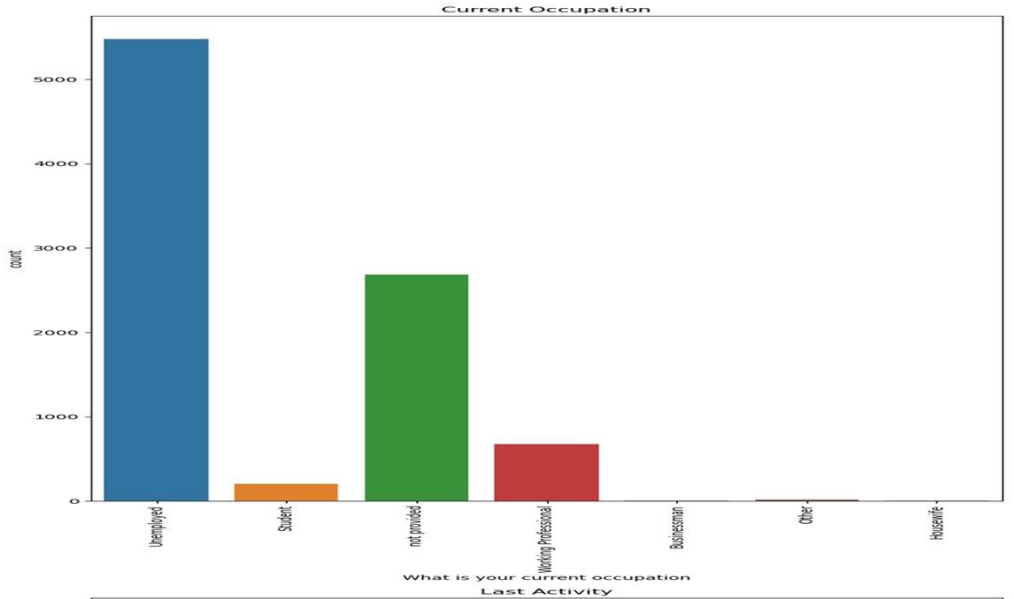



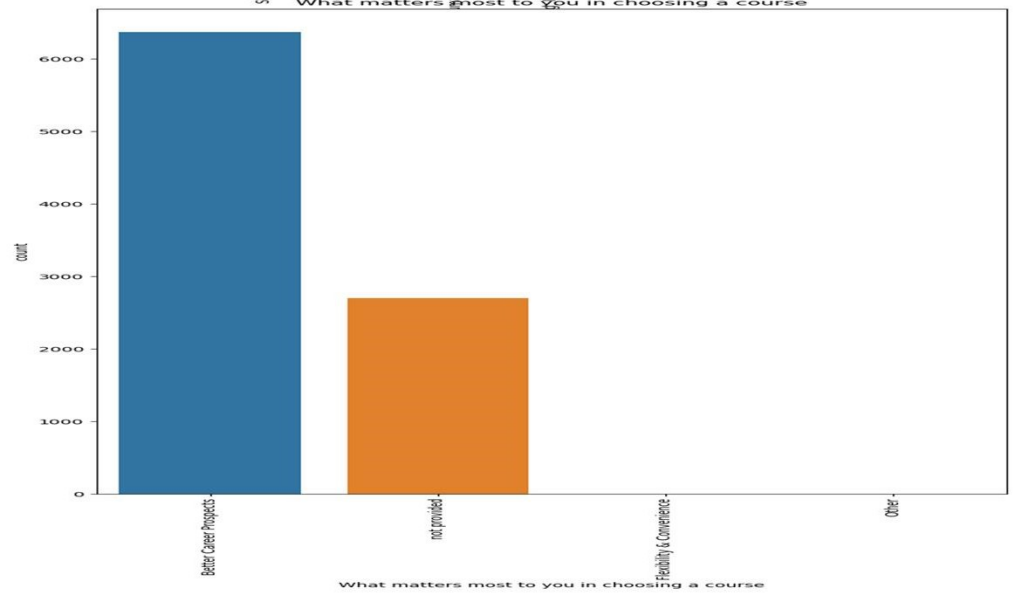



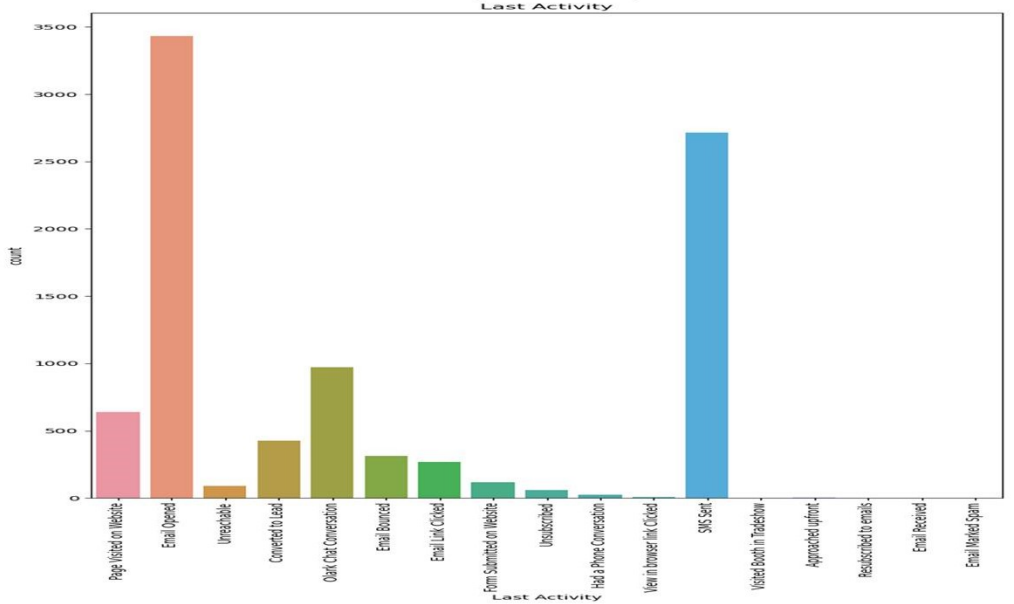
## Lead Source

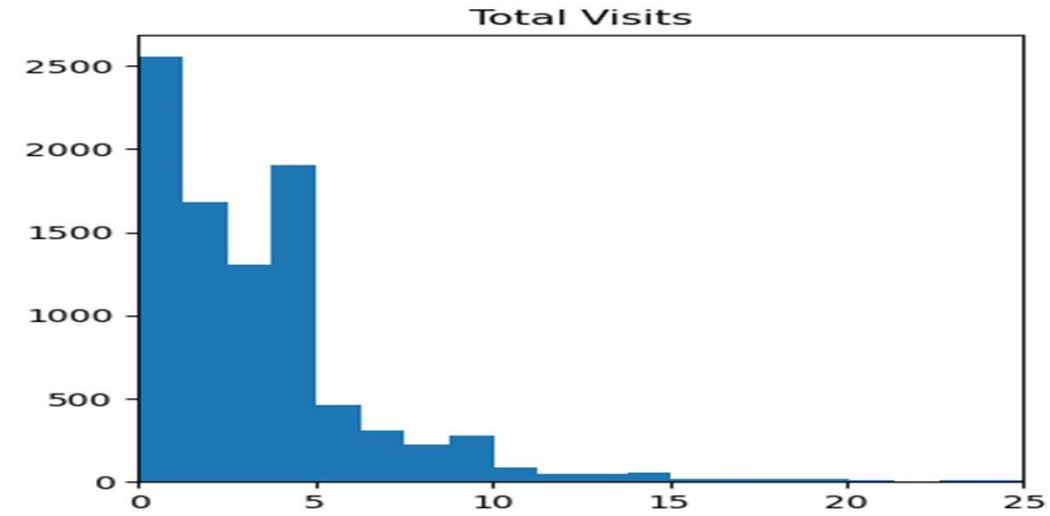

Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

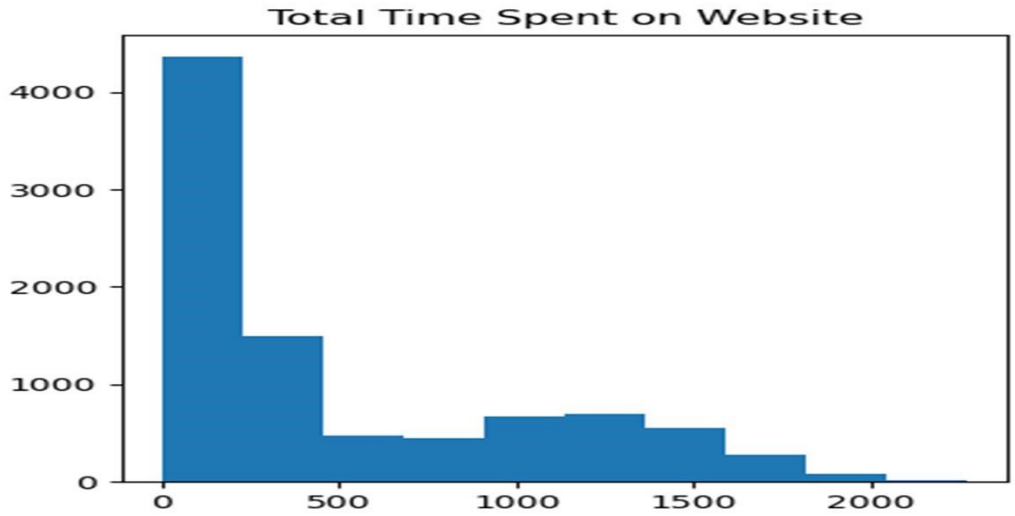

Specialization
count
not provided
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business

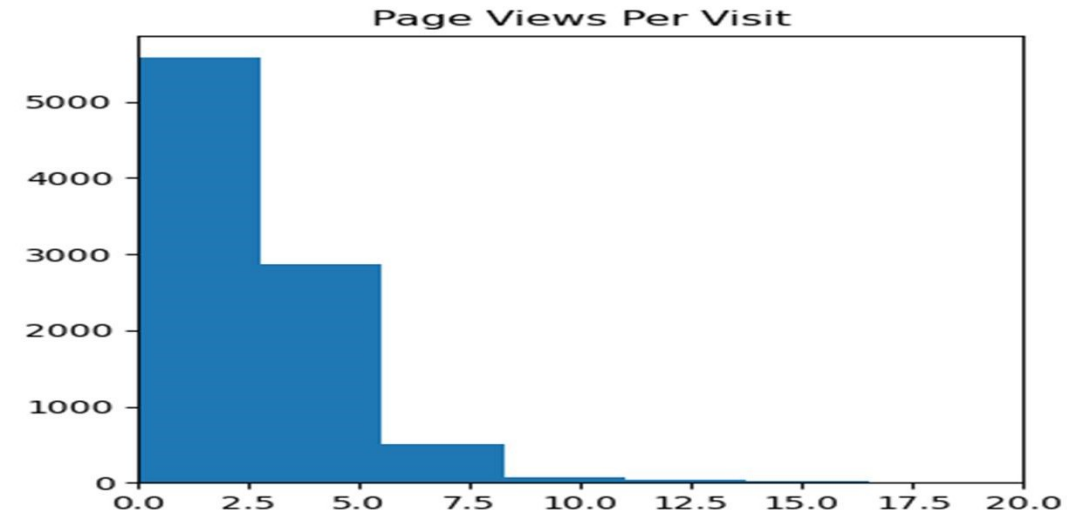

Current Occupation
count
Unemployed
Student
not provided
Working Professional
Businessman
Other
Housewife
What is your current occupation

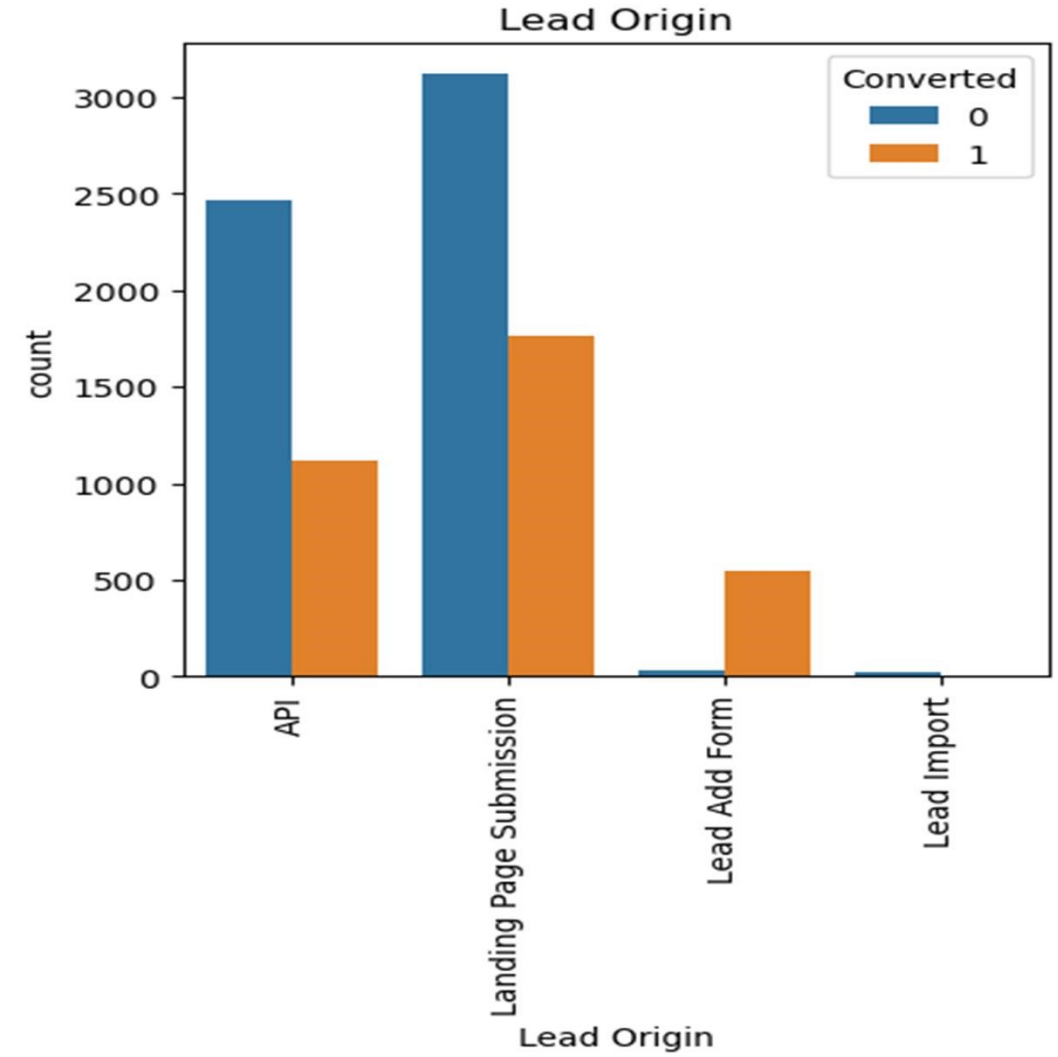

What matters most to you in choosing a course
count
Better Career Prospects
not provided
Flexibility & Convenience
Other
What matters most to you in choosing a course

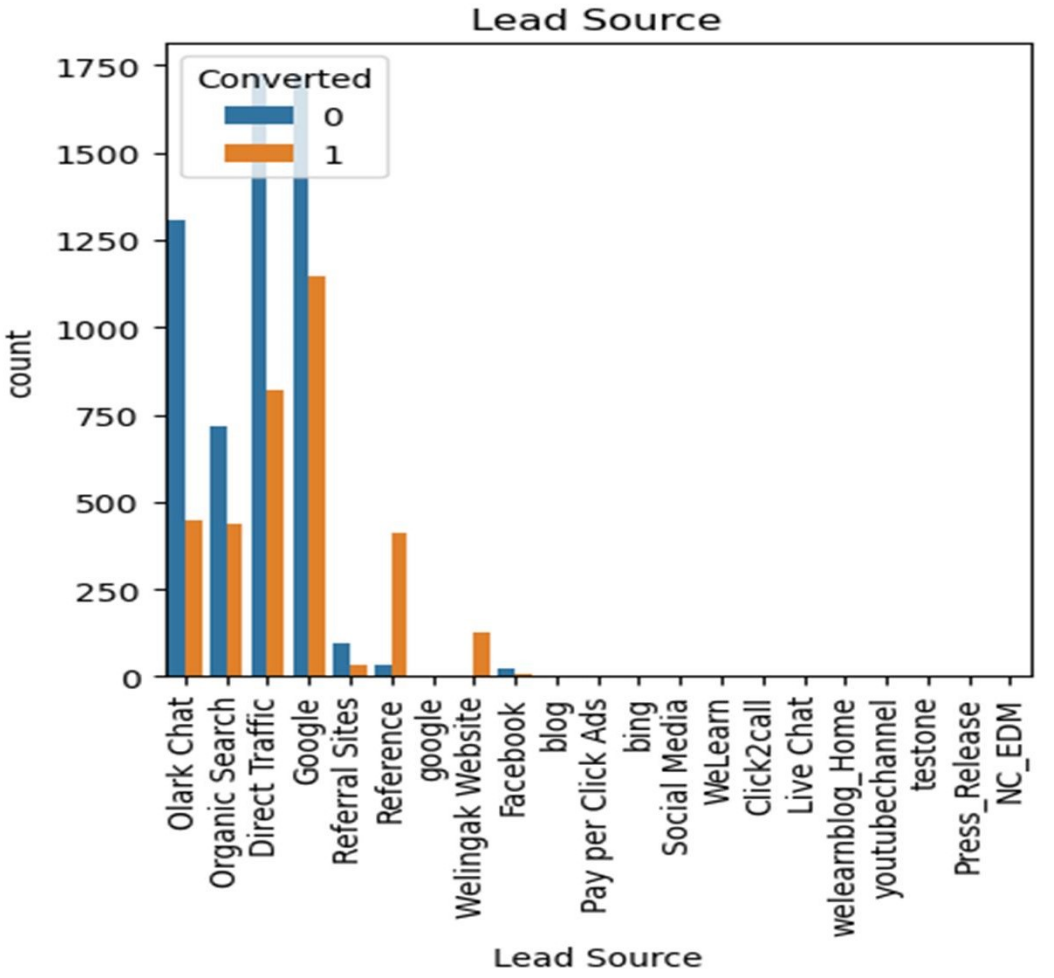

Last Activity
count
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

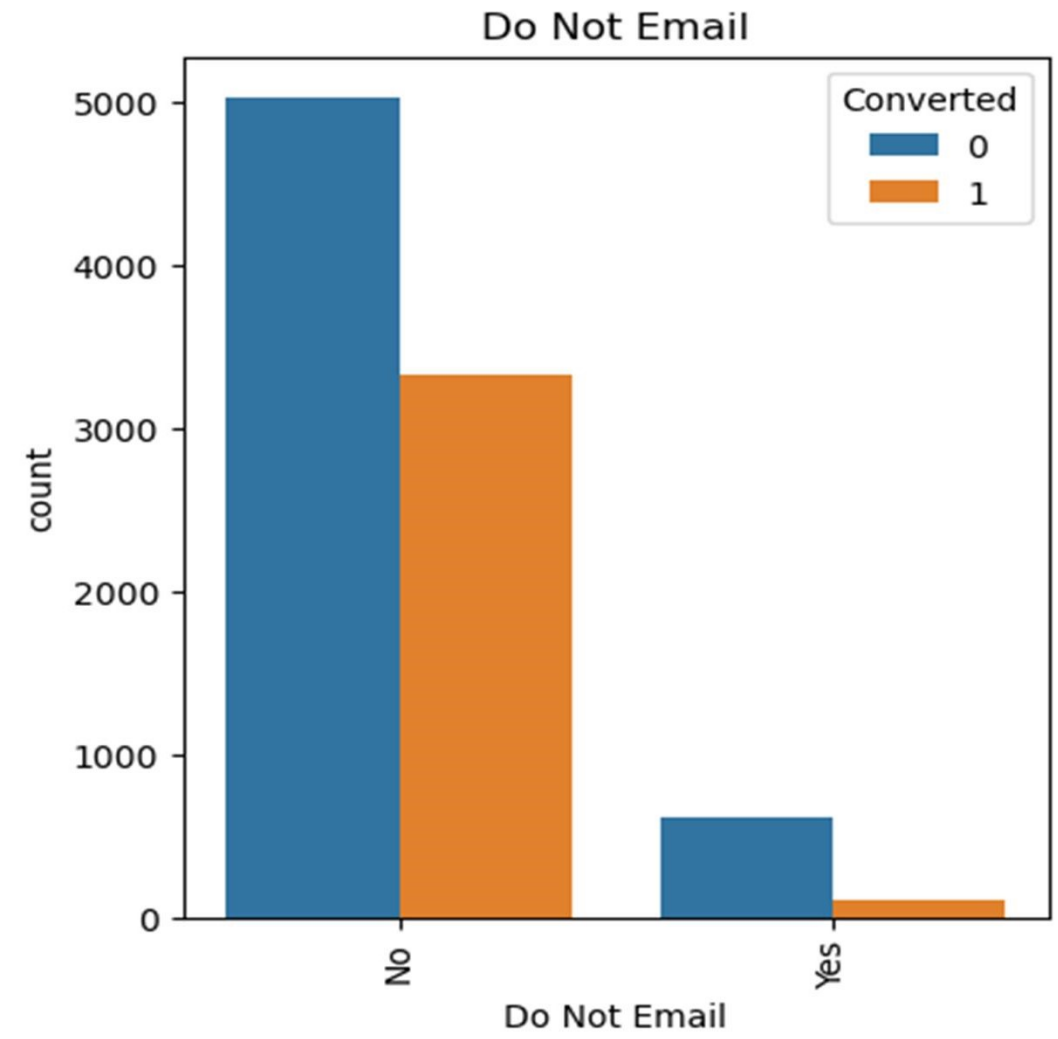

Total Visits
2500
2000
1500
1000
500
0
0
5
10
15
20
25

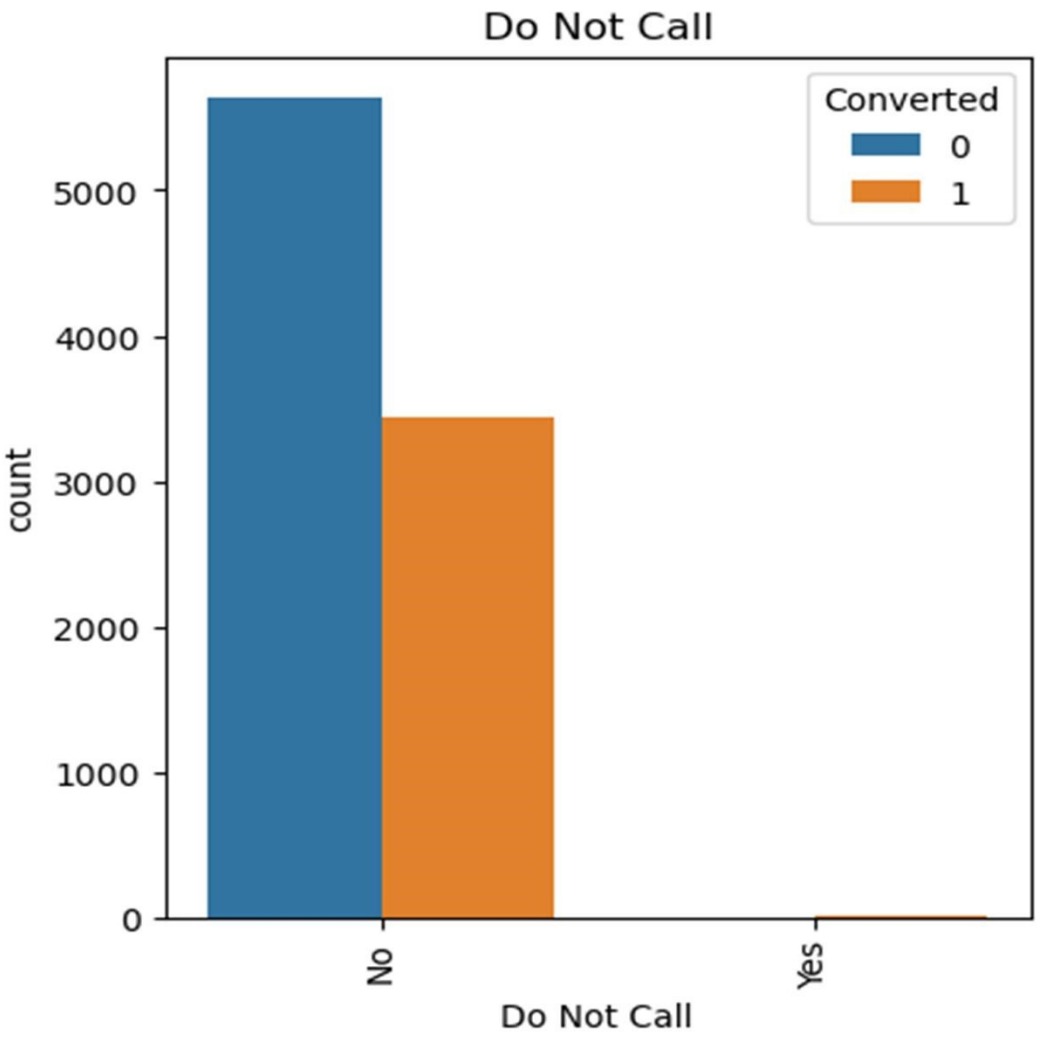

Total Time Spent on Website
4000
3000
2000
1000
0
0
500
1000
1500
2000

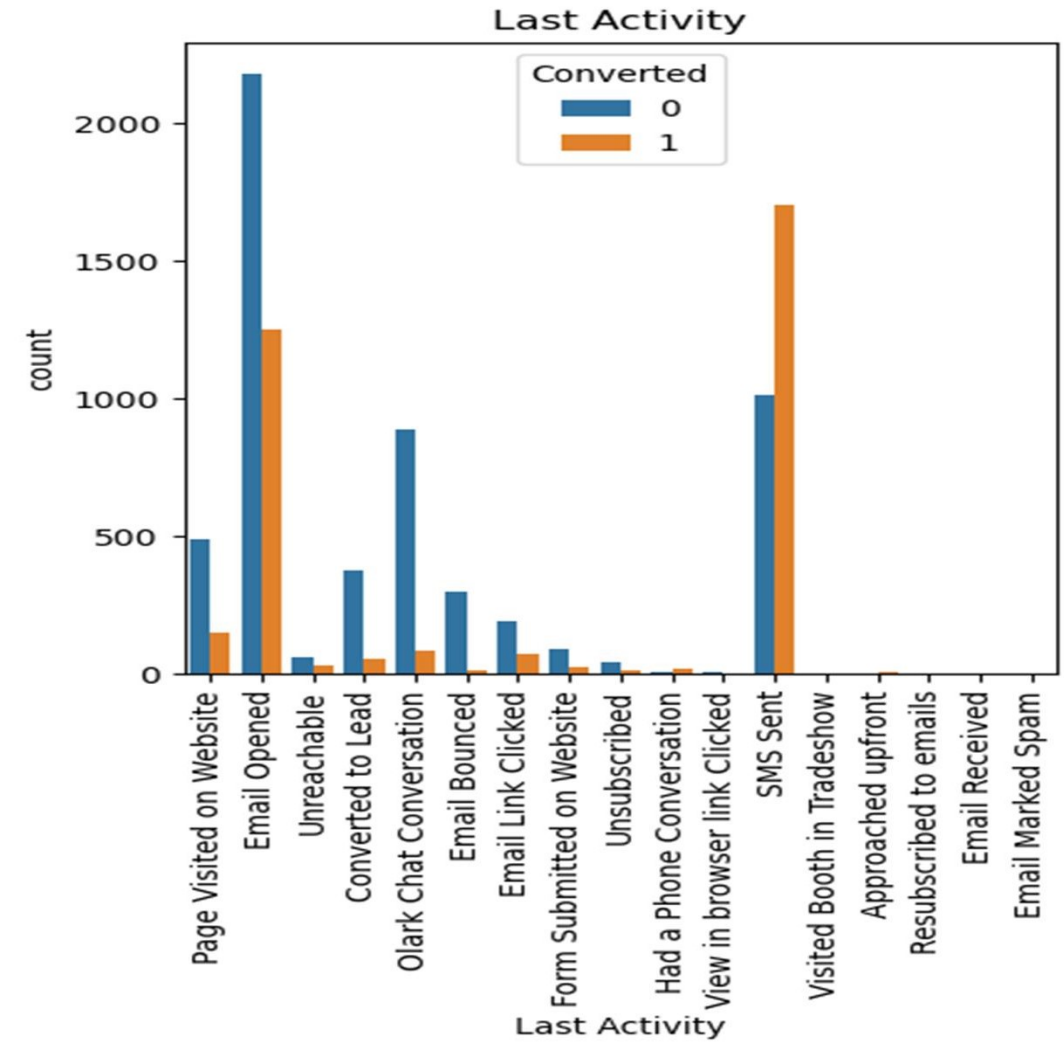

Page Views Per Visit
5000
4000
3000
2000
1000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

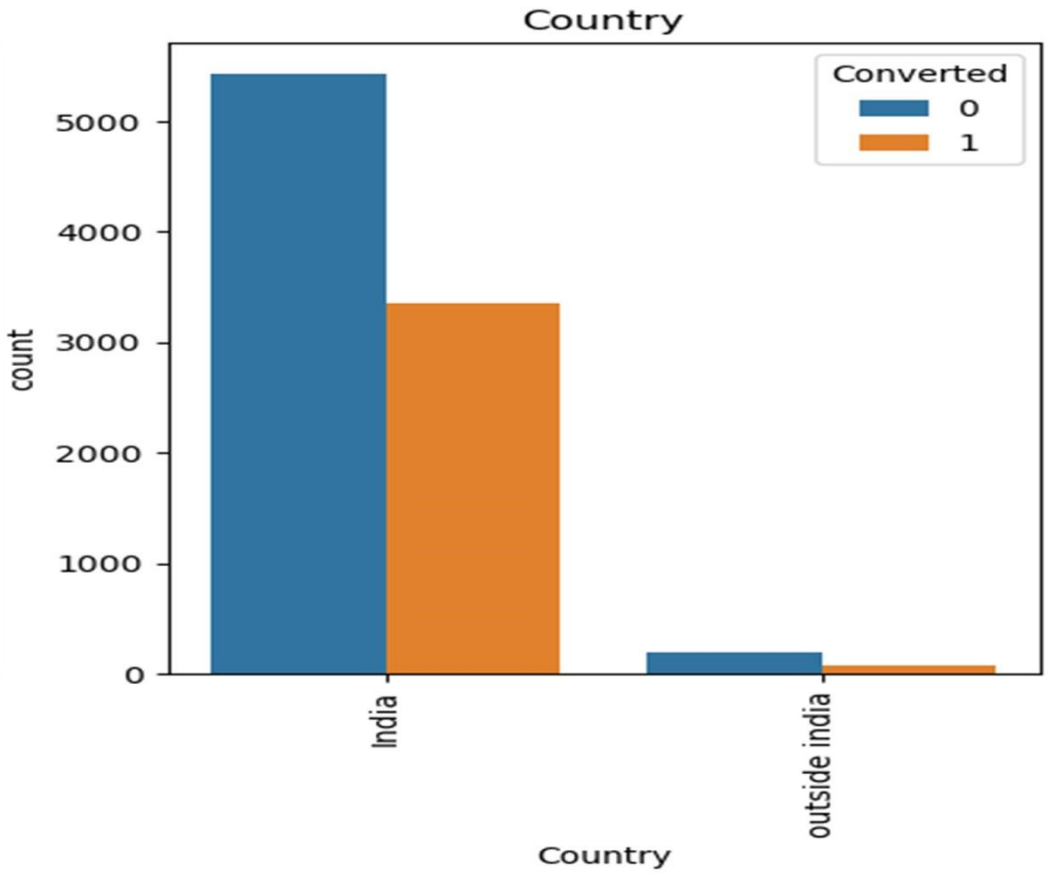

Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

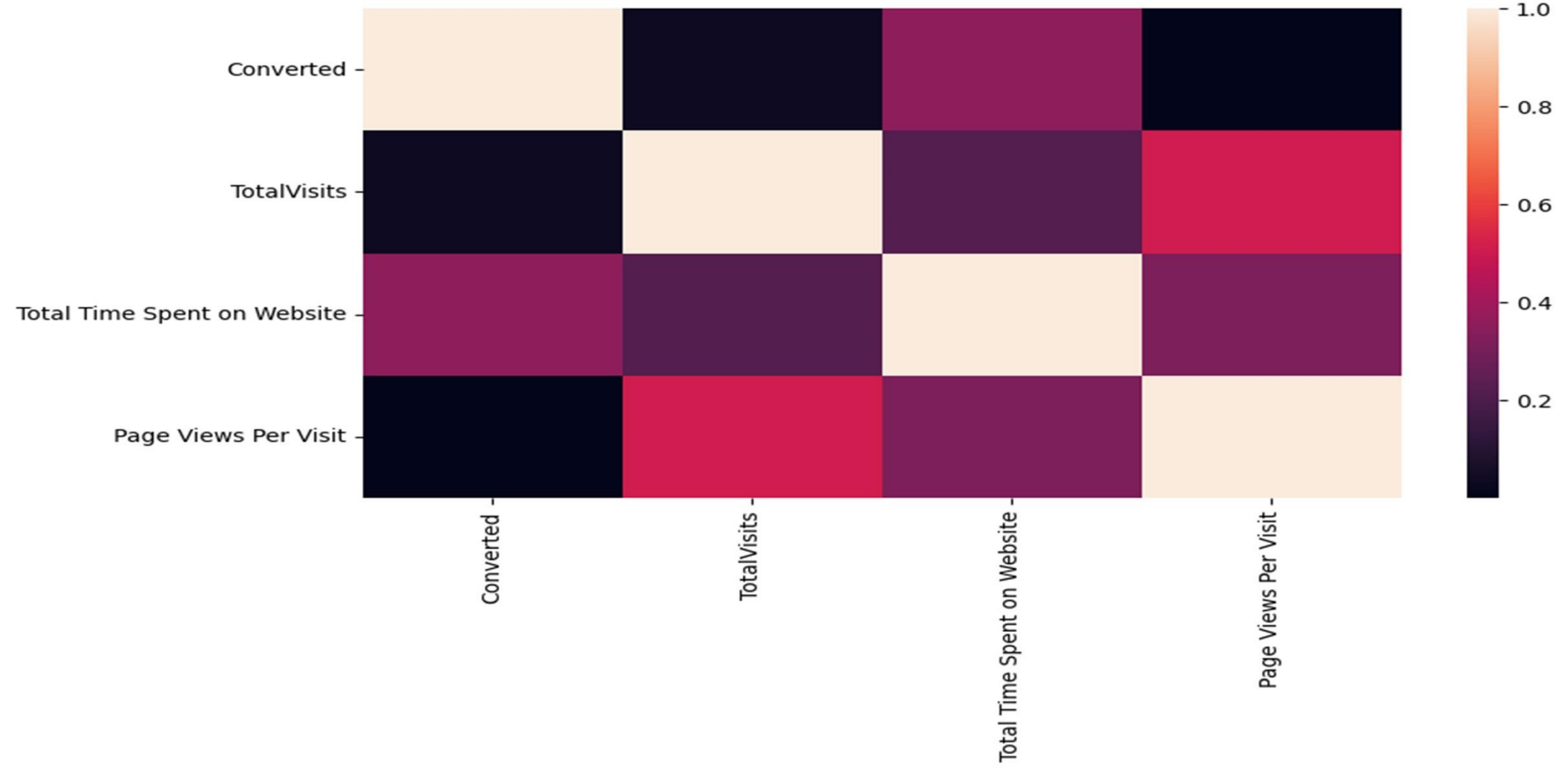

Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

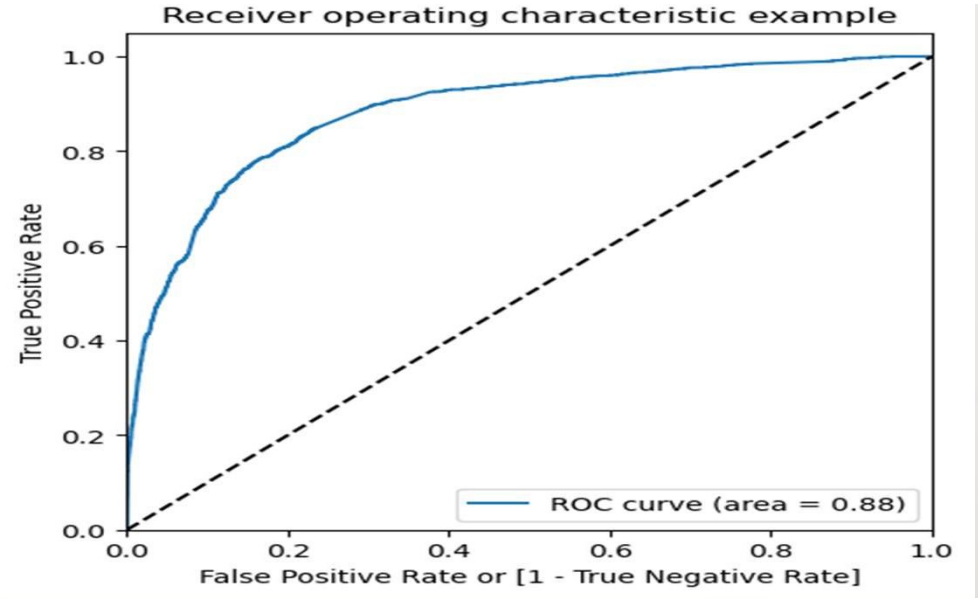

Do Not Email
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Email

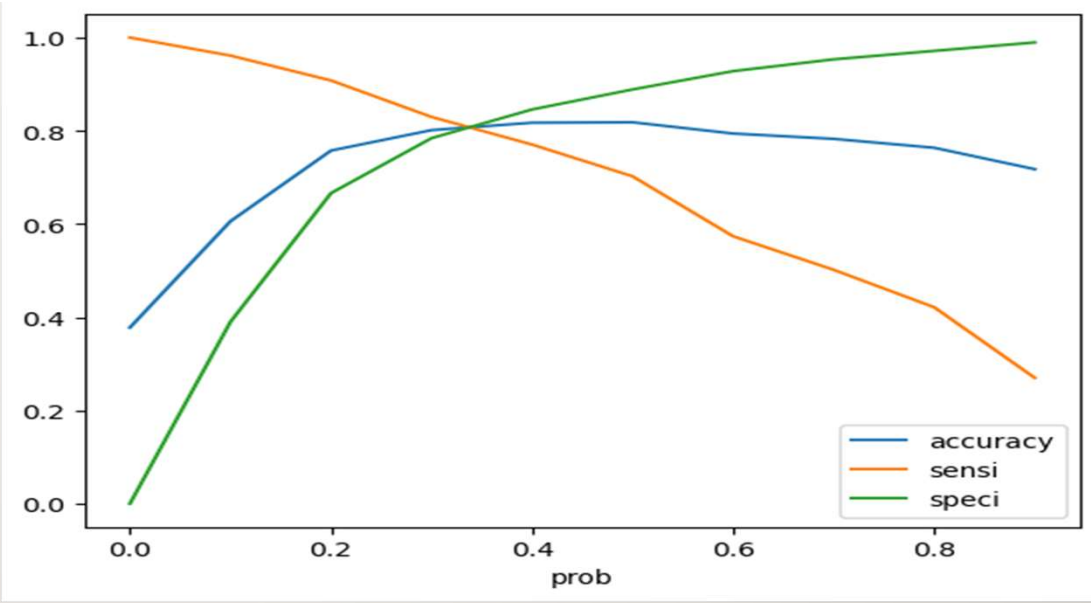

Do Not Call
Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Call


Last Activity
Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
SMS Sent
Visited Booth in Tradeshow
Approached upfront
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

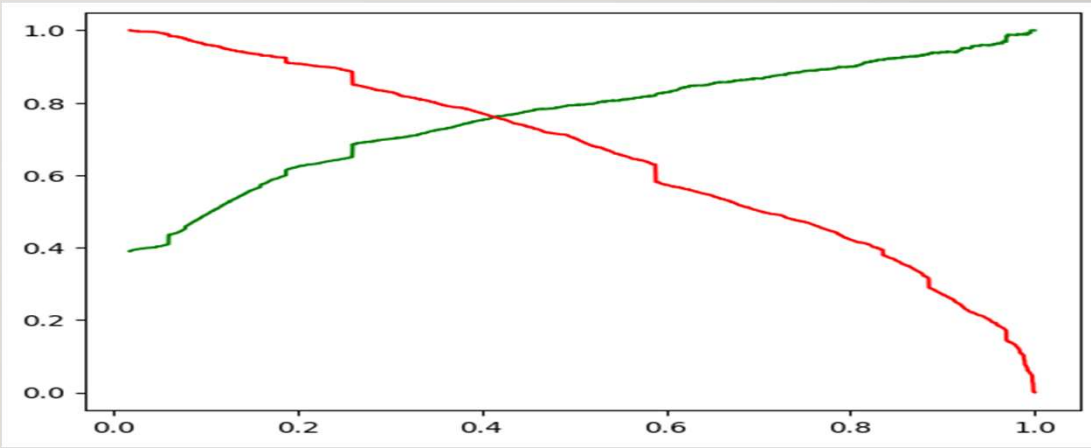

Country
Converted
0
1
count
5000
4000
3000
2000
1000
0
India
outside india
Country


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit




**Linear Regression Final Model Parameters**
**Area under ROC = 0.88**
**Intermediate cut-off = 0.35**
**Final cut-off = 0.44**



# CONCLUSION

TOP VARIABLE CONTRIBUTING TO CONVERSION:

- LEAD SOURCE:
    - Total Visits
    - Total Time Spent on Website
- Lead Origin:
    - Lead Add Form
- Lead source:
    - Direct traffic
    - Google
    - Welingak website
    - Organic search
    - Referral Sites

Last Activity:

- Do Not Email_Yes
- Last Activity_Email Bounced
- Olark chat conversation

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.