# Lead Scoring Case Study

Submitted by  - Godwin Neal
Email : gnneal96@gmail.com

# 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.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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 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%.

# Strategy

- ➢Source the data for analysis
- ➢Clean and prepare the data
- ➢Exploratory Data Analysis.
- ➢Feature Scaling
- ➢Splitting the data into Test and Train dataset.
- ➢Building a logistic Regression model and calculate Lead Score.
- ➢Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- ➢Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

# Problem Solving Methodology

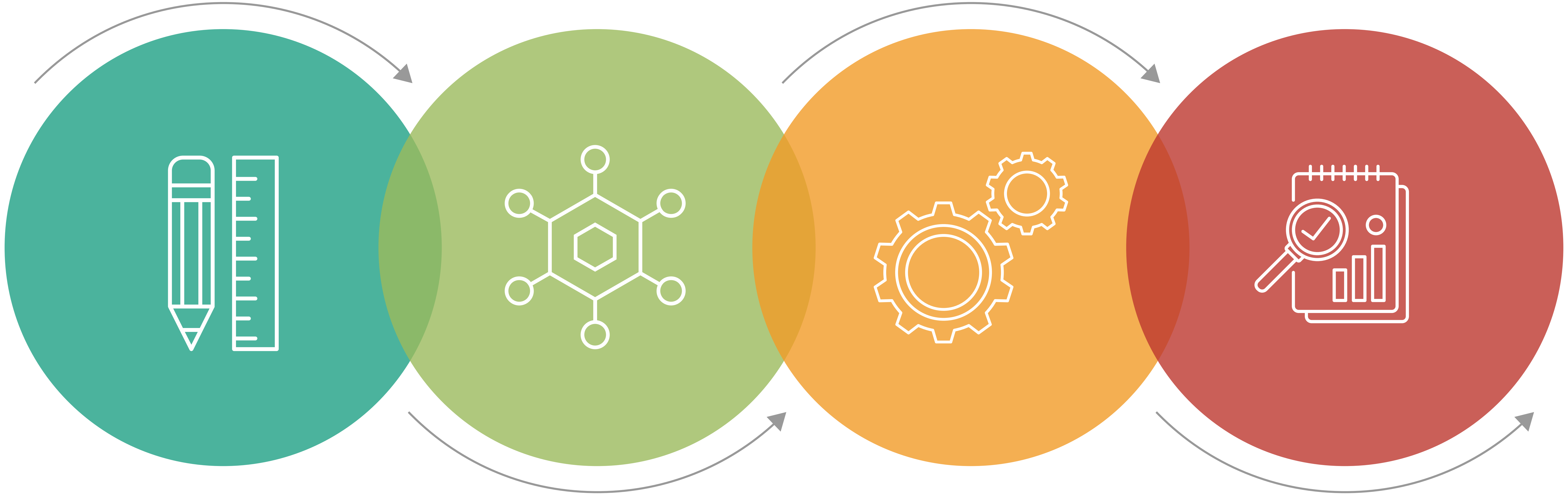
**Data Sourcing , Cleaning and Preparation**

- Read the Data from Source
- Convert data into clean format suitable

  for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

**Model Building**

- Feature Selection using RFE
- Determine the optimal model
- using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.



Feature Scaling and Splitting Train and Test Sets

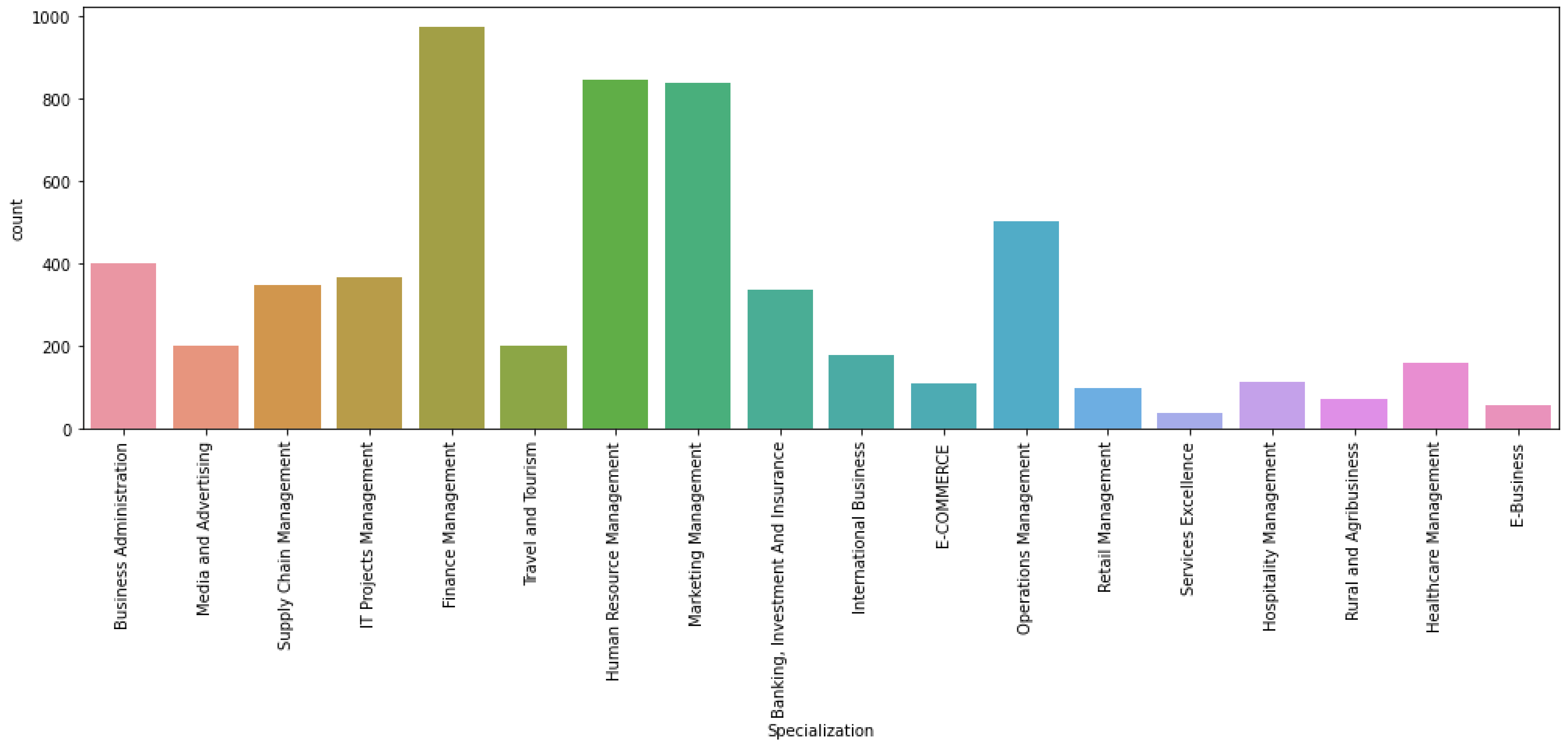
- Feature Scaling of Numeric
- data
- Splitting data into train and test set.

Result

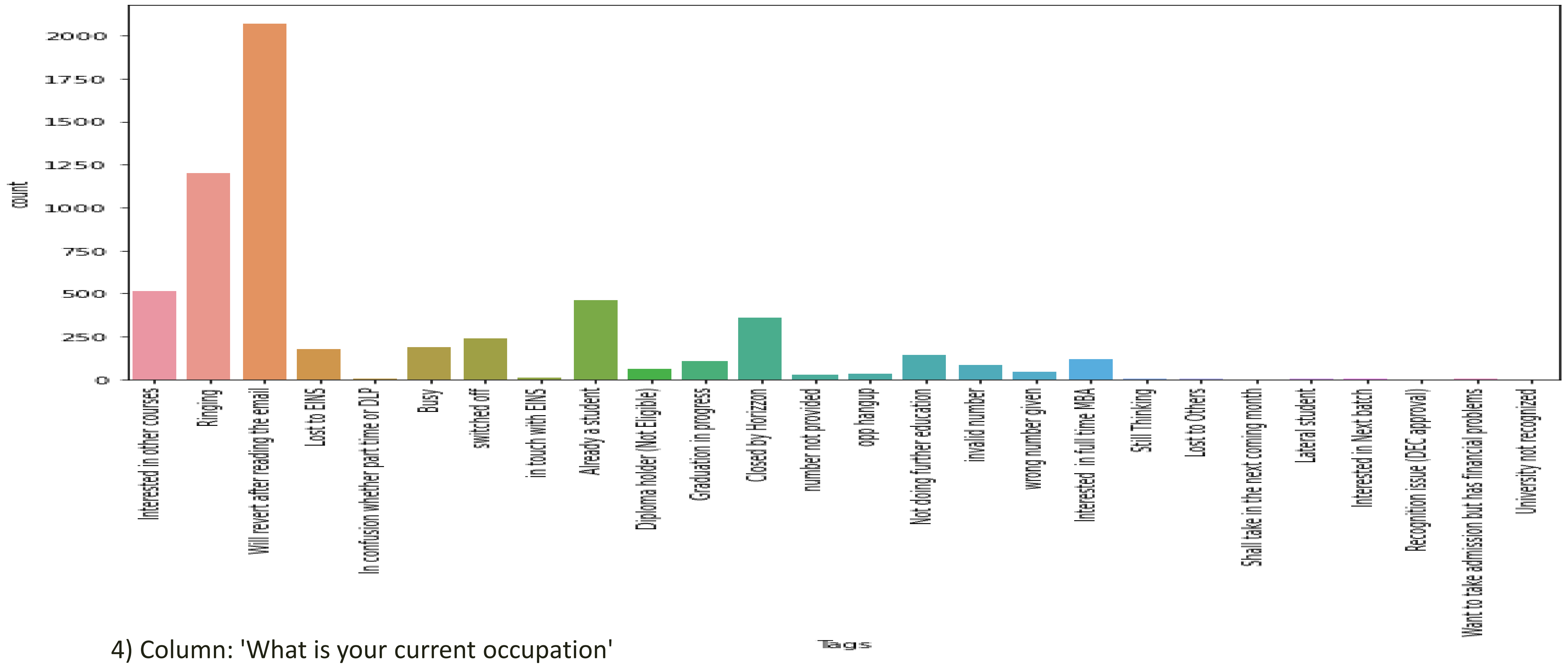
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

# Data Inspection

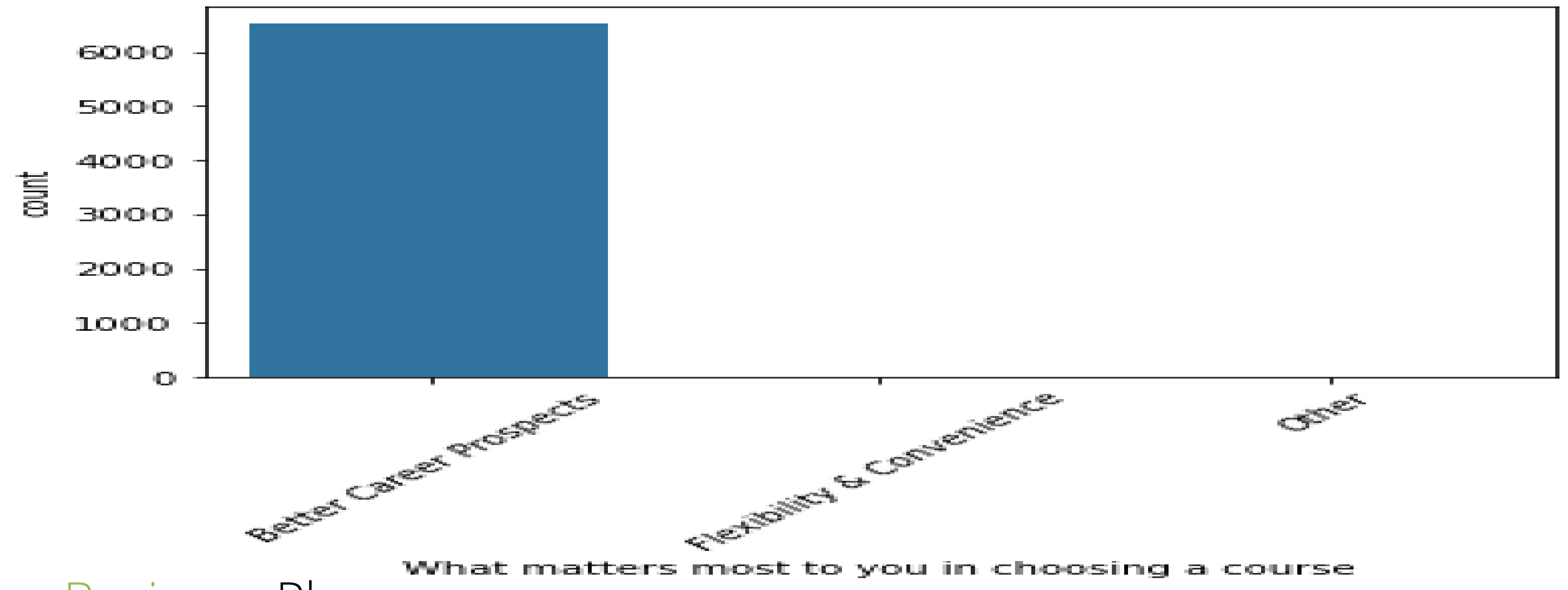
1) Column: 'Specialization'
This column has 37% missing values



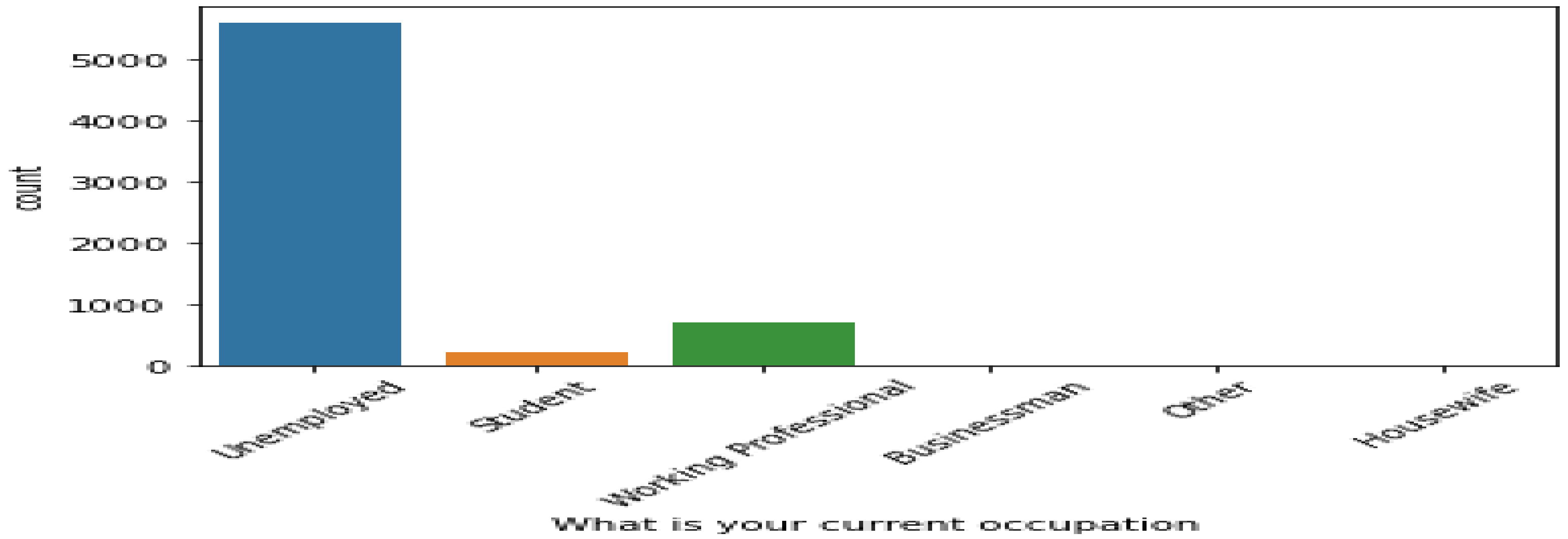
2) Tags column
'Tags' column has 36% missing values



3) Column: 'What matters most to you in choosing a course'
this column has 29% missing values

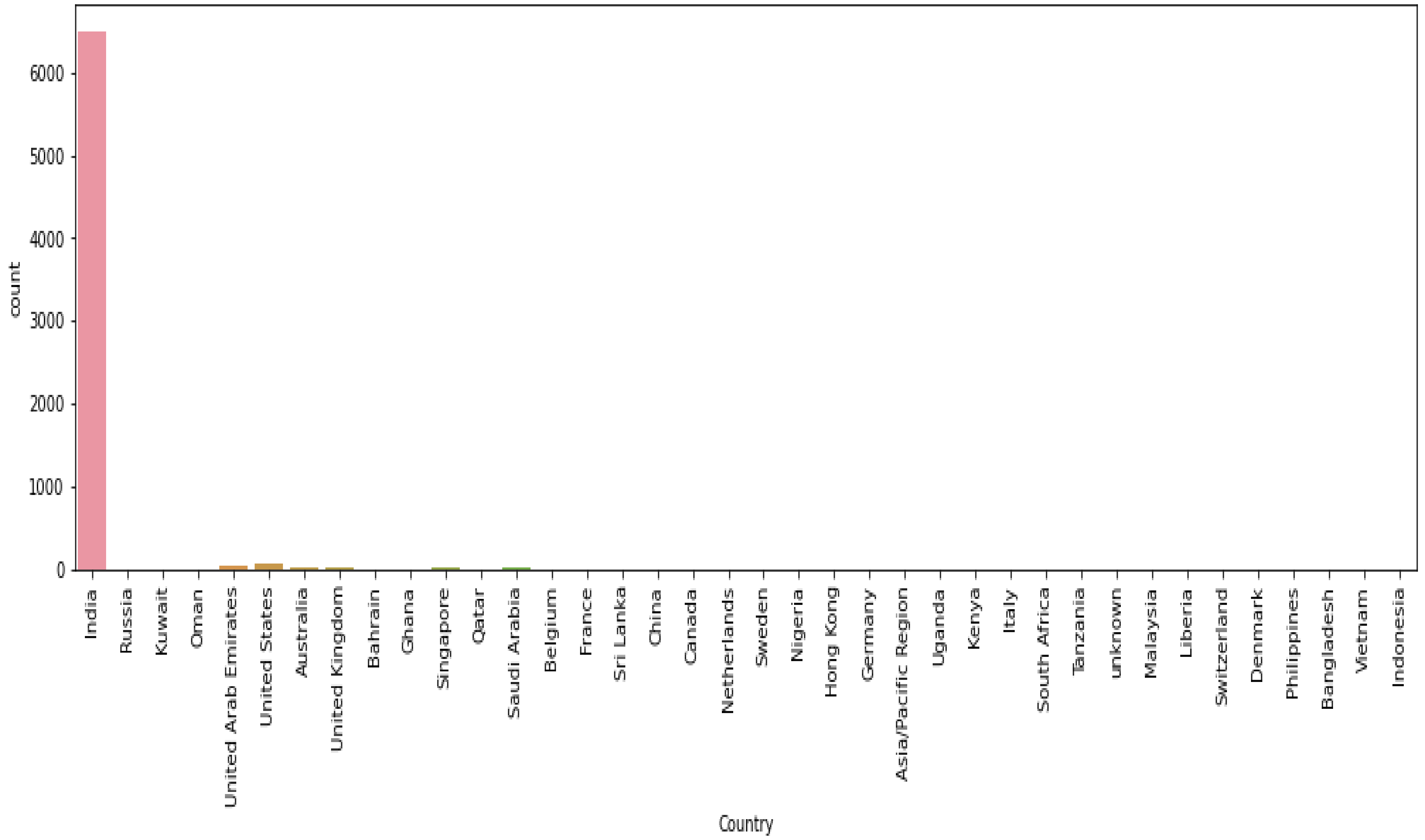


4) Column: 'What is your current occupation'
this column has 29% missing values

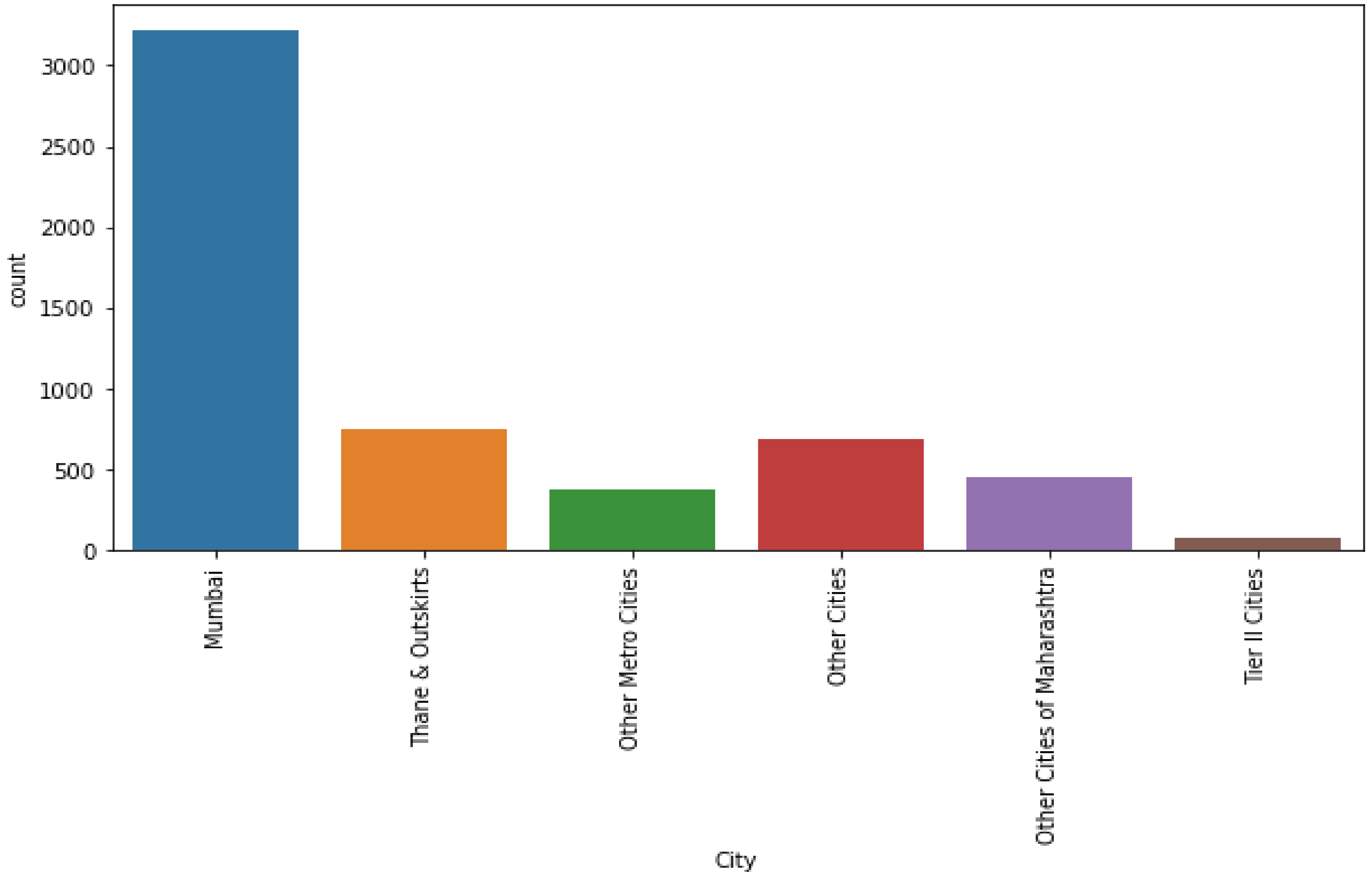


# Data Inspection

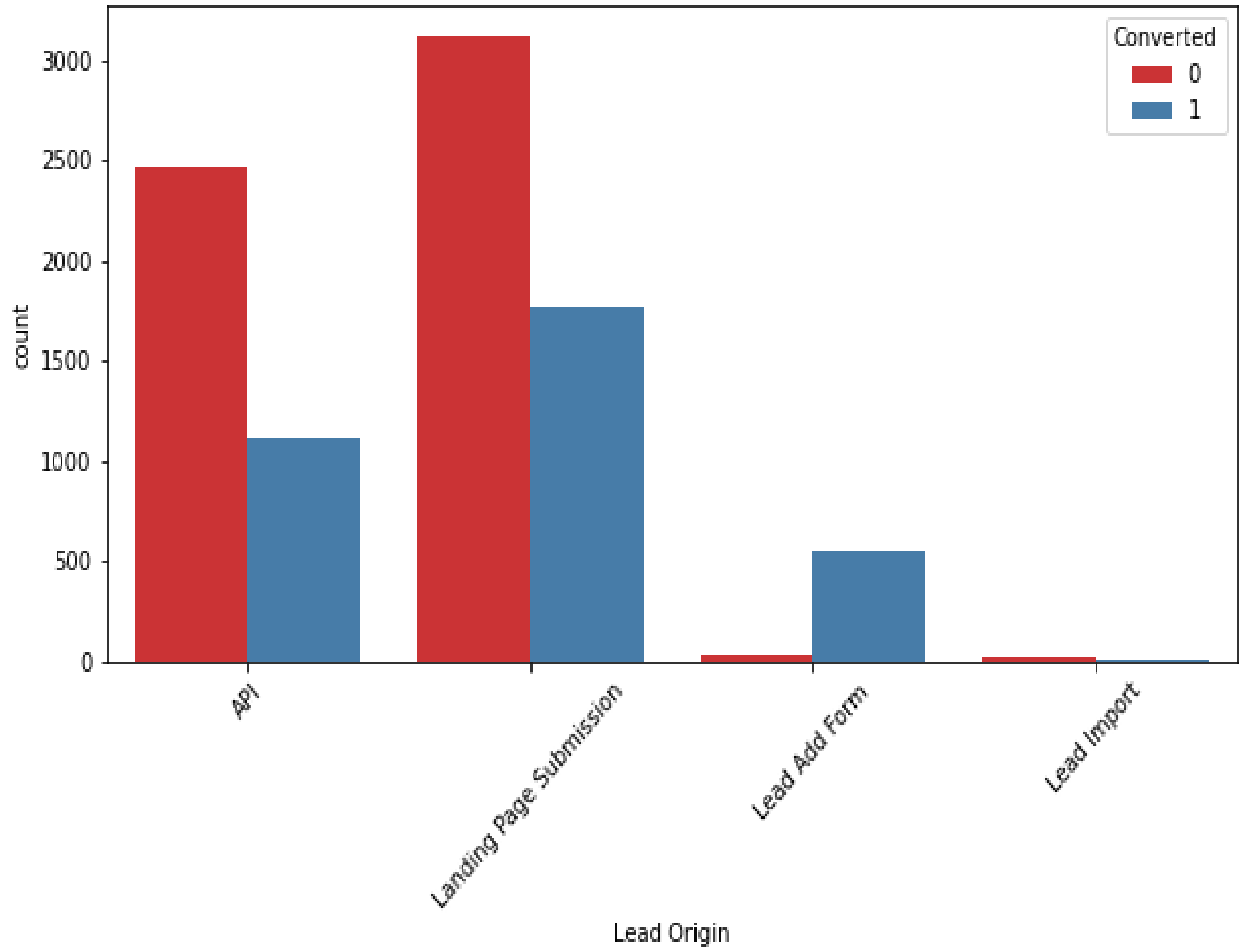
5) Column: 'Country'
This column has 27% missing values



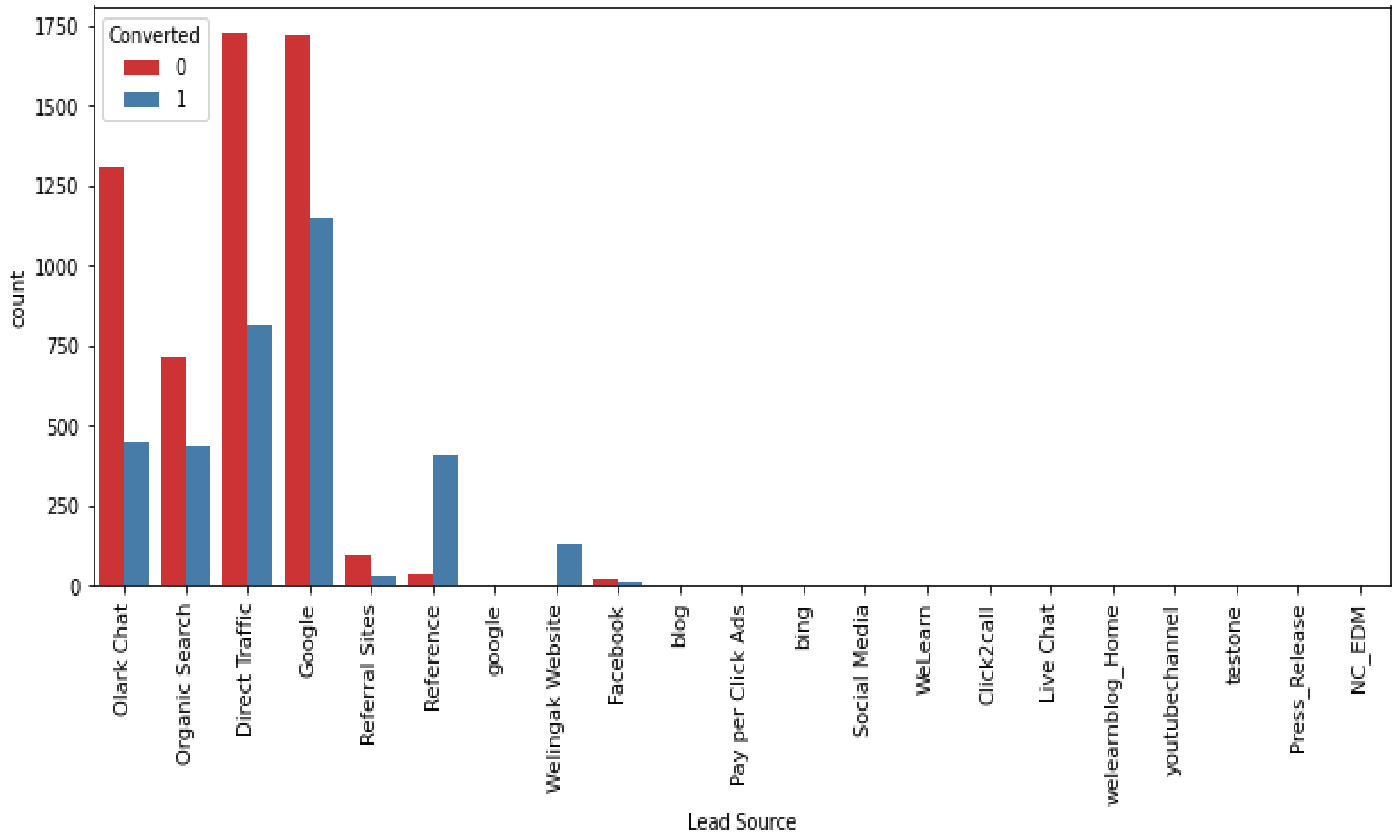
6) Column: 'City'
This column has 40% missing values



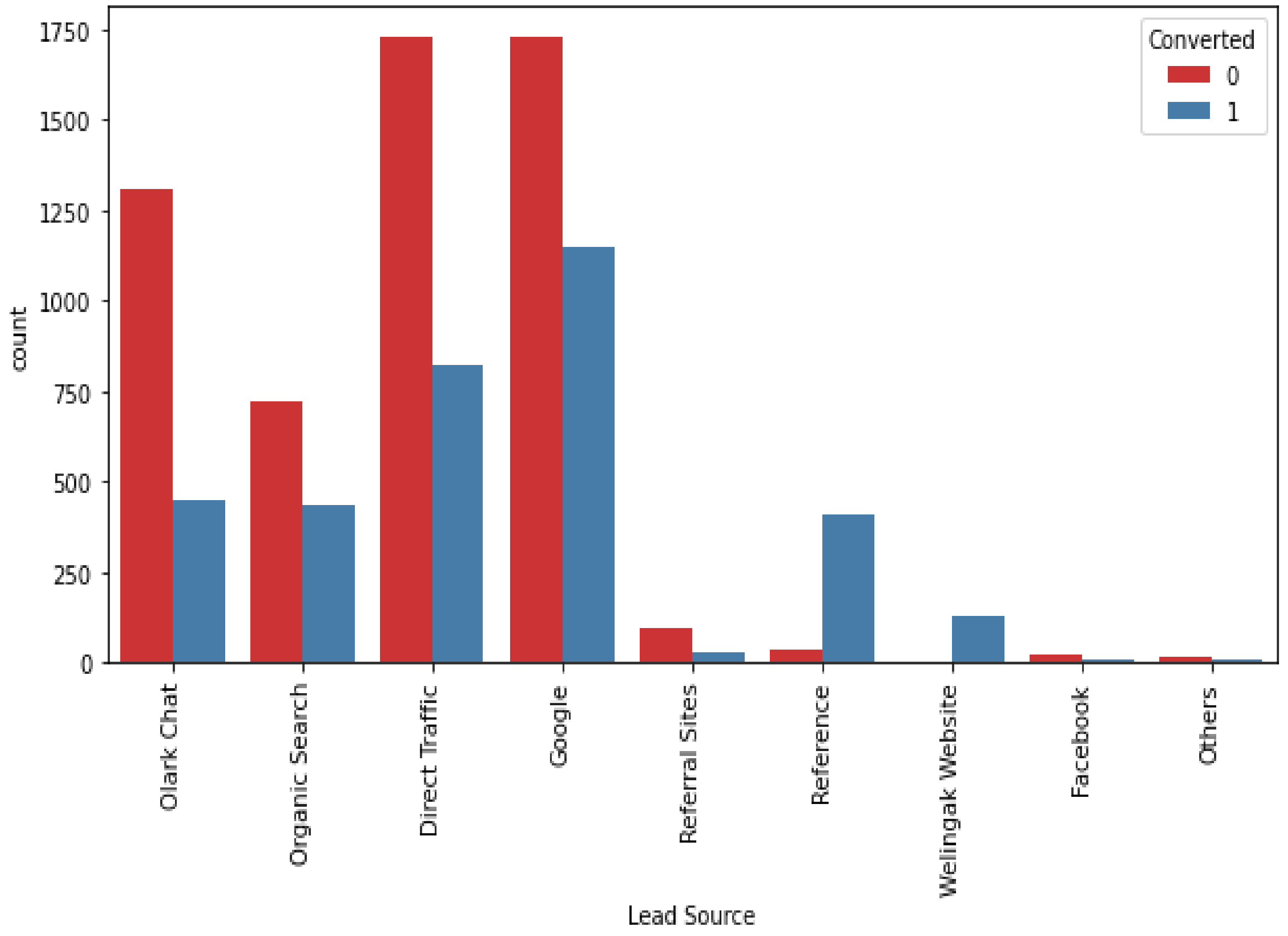
# EDA



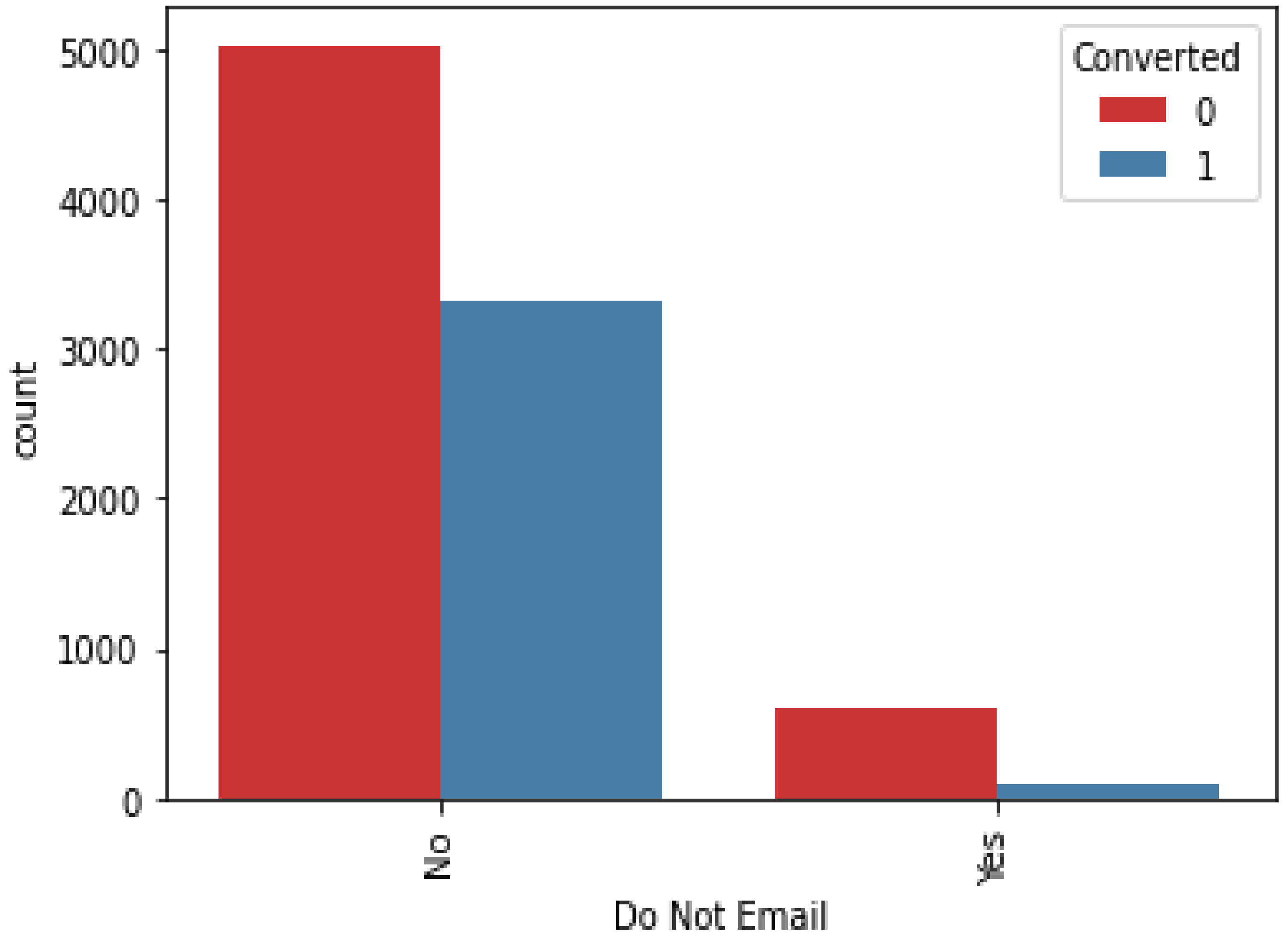
Converted
🟥 0
🟦 1
3000
2500
2000
1500
1000
500
0
count
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin



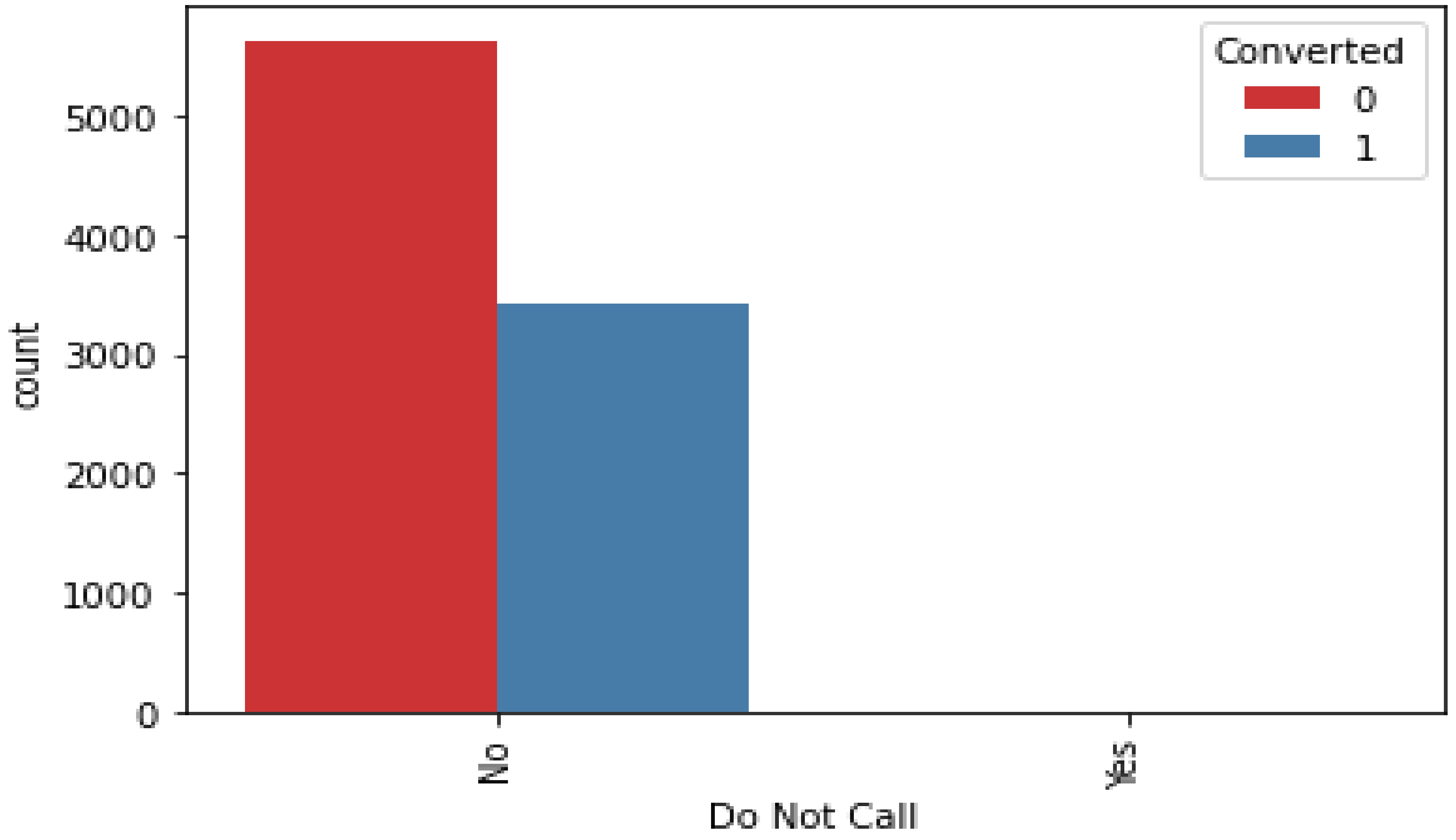
Converted
🟥 0
🟦 1
1750
1500
1250
1000
750
500
250
0
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source



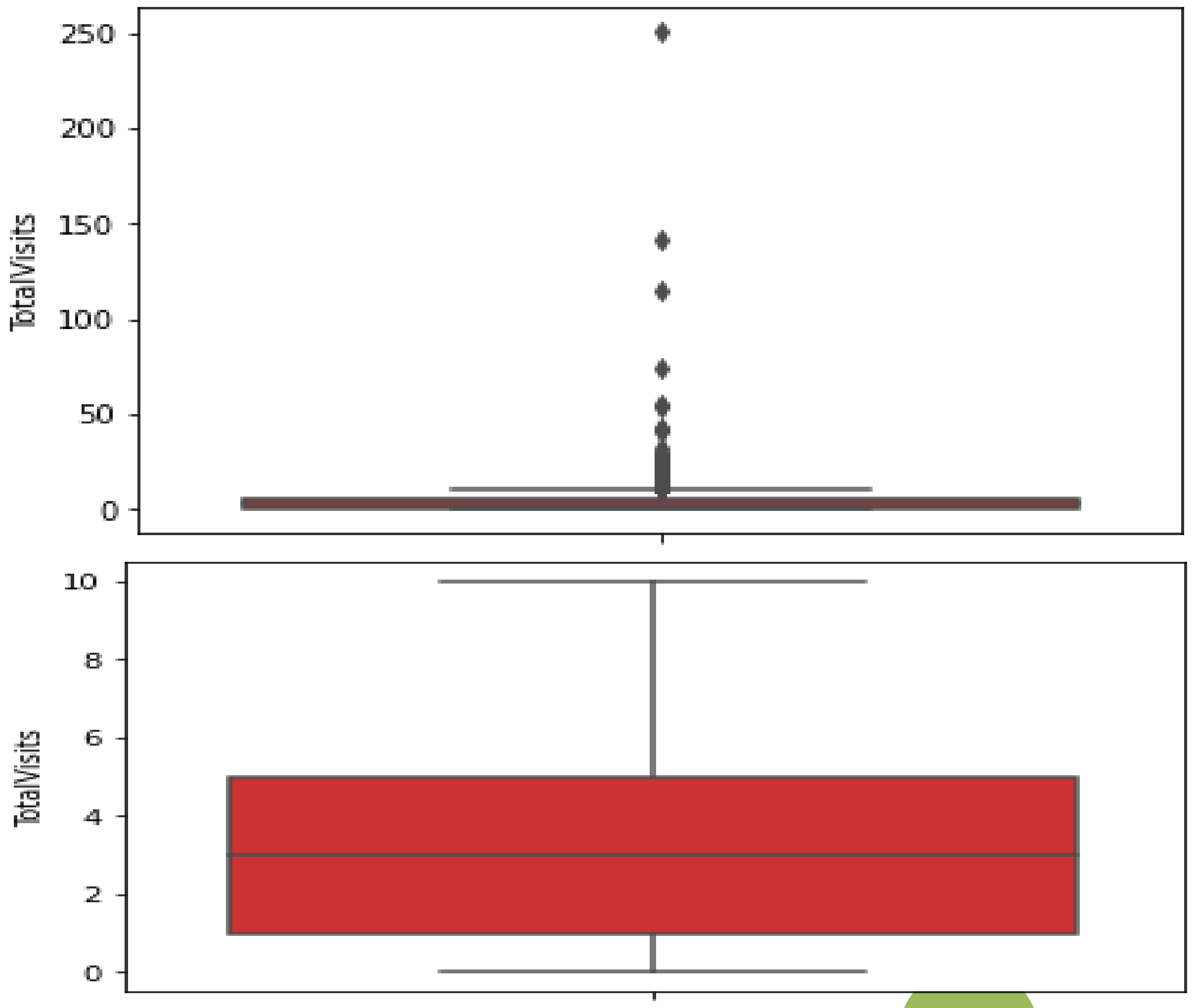
Converted
🟥 0
🟦 1
1750
1500
1250
1000
750
500
250
0
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source



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

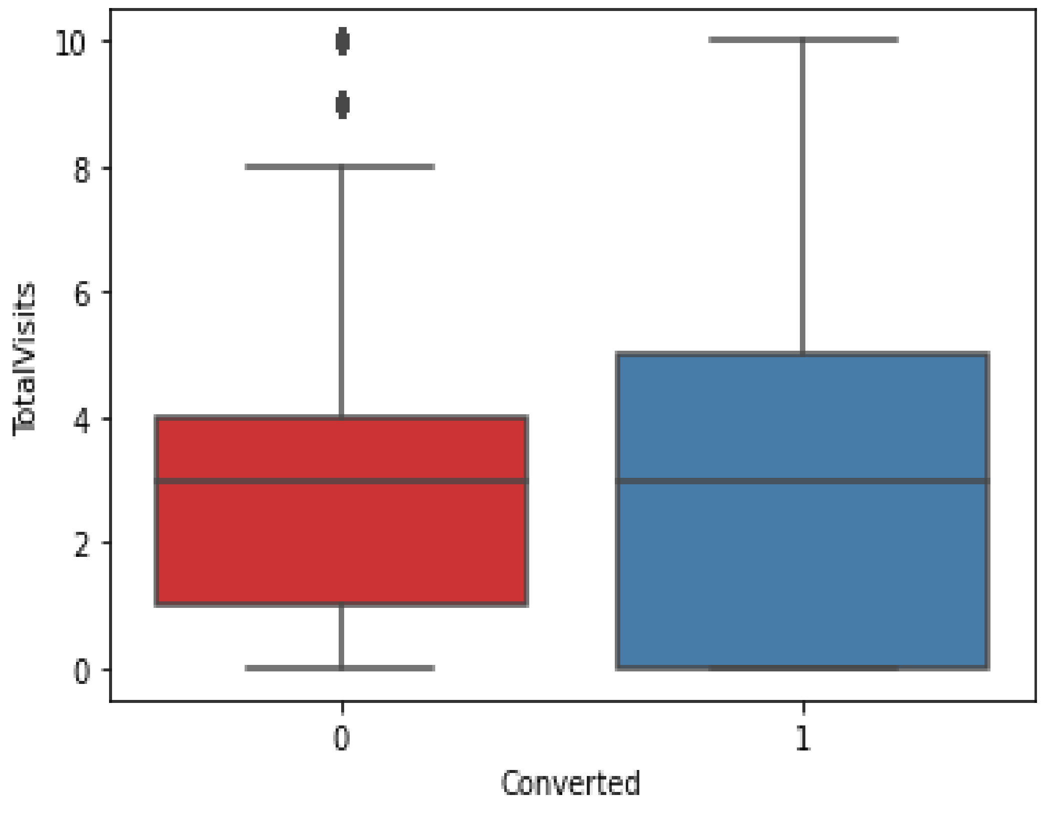


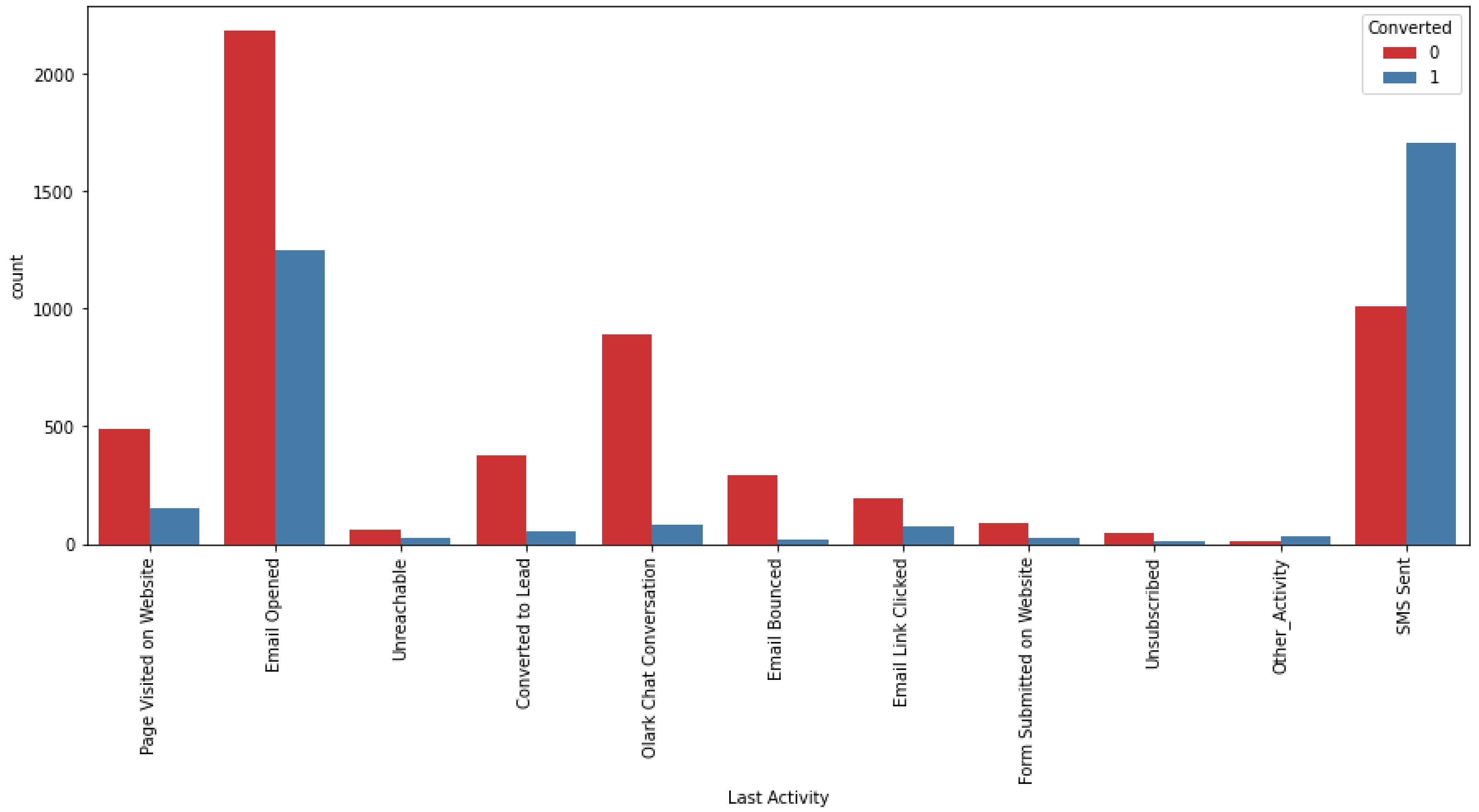
Converted
🟥 0
🟦 1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Call

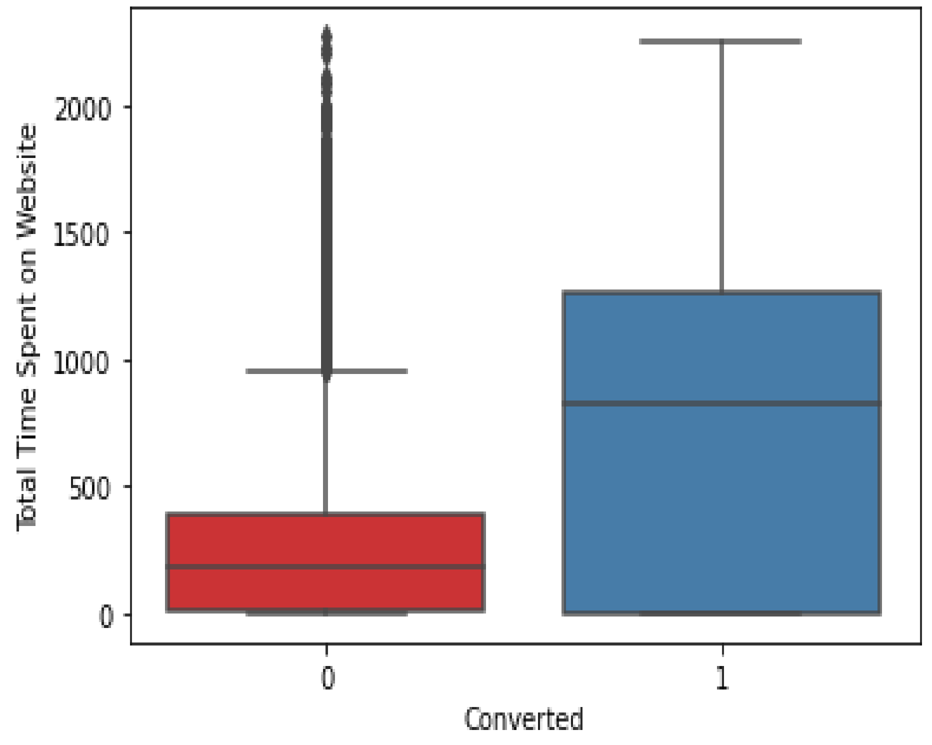


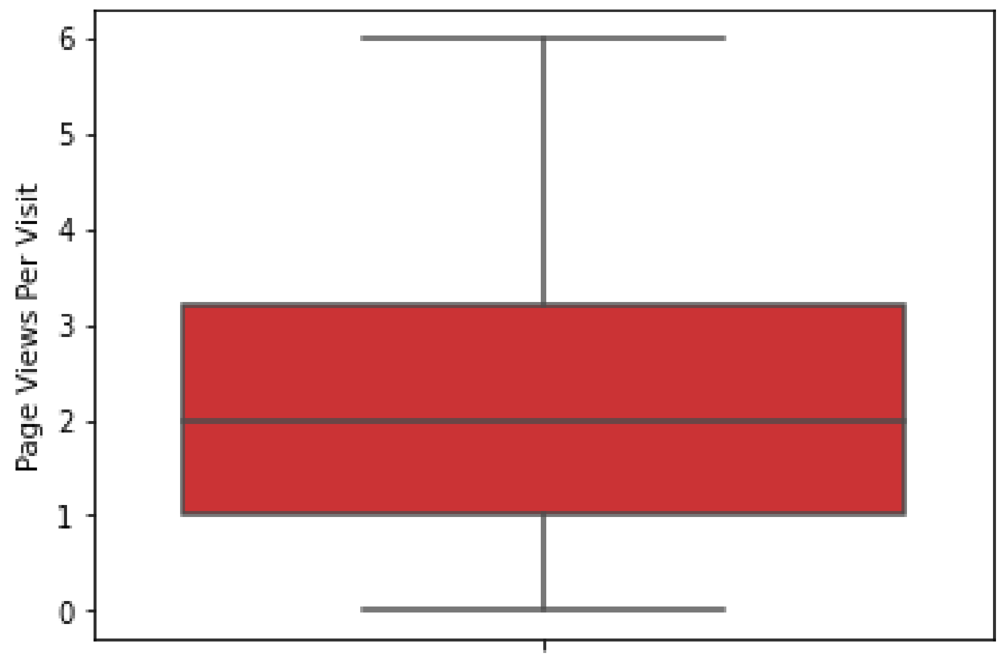
250
200
150
100
50
0
TotalVisits
10
8
6
4
2
0
TotalVisits

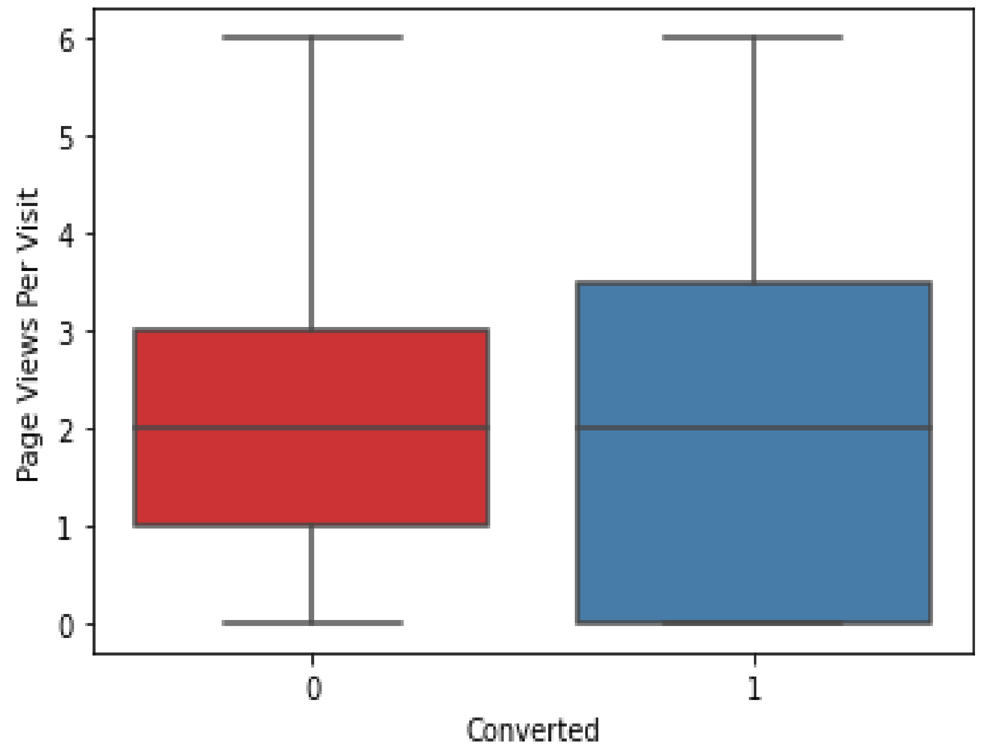
# EDA

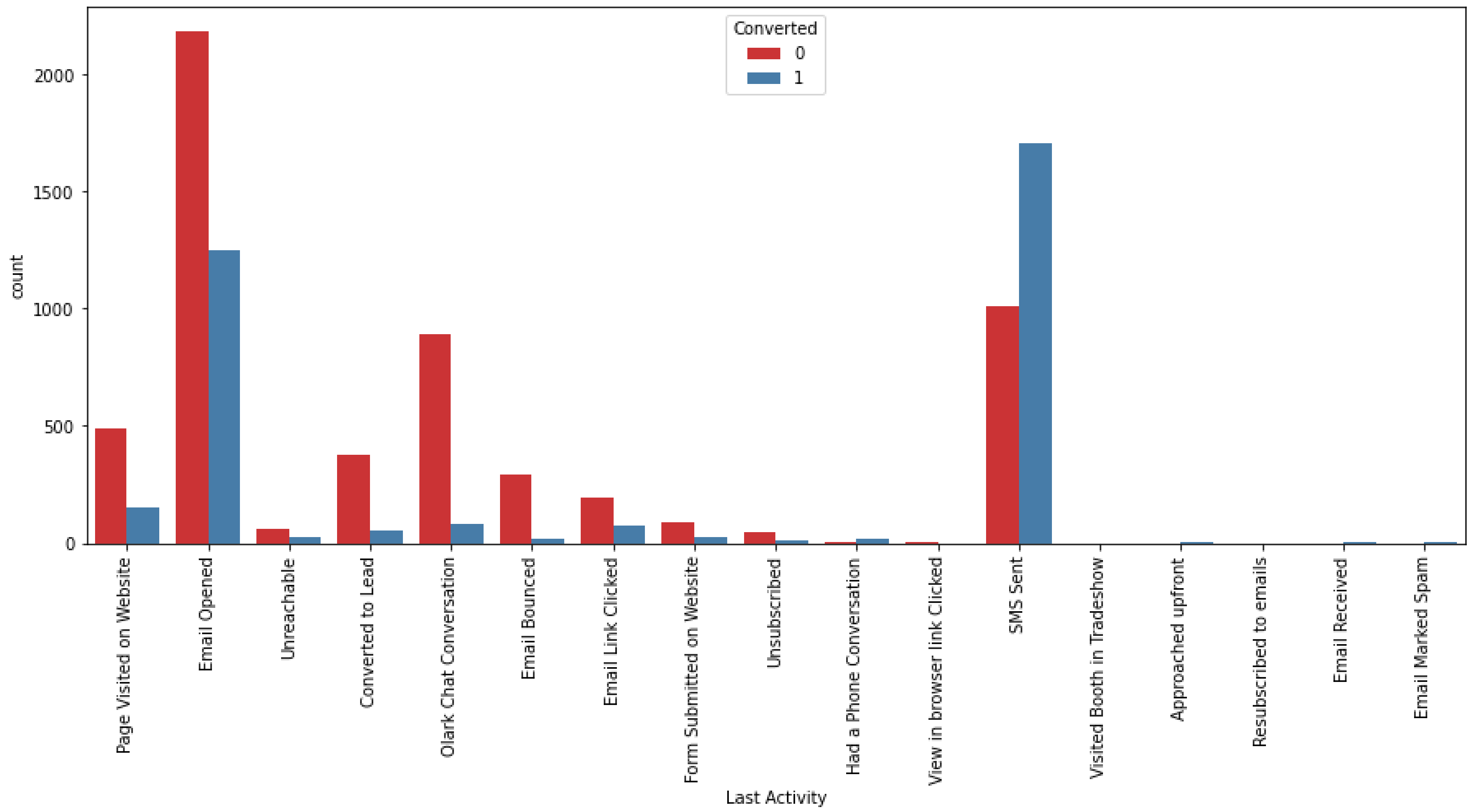


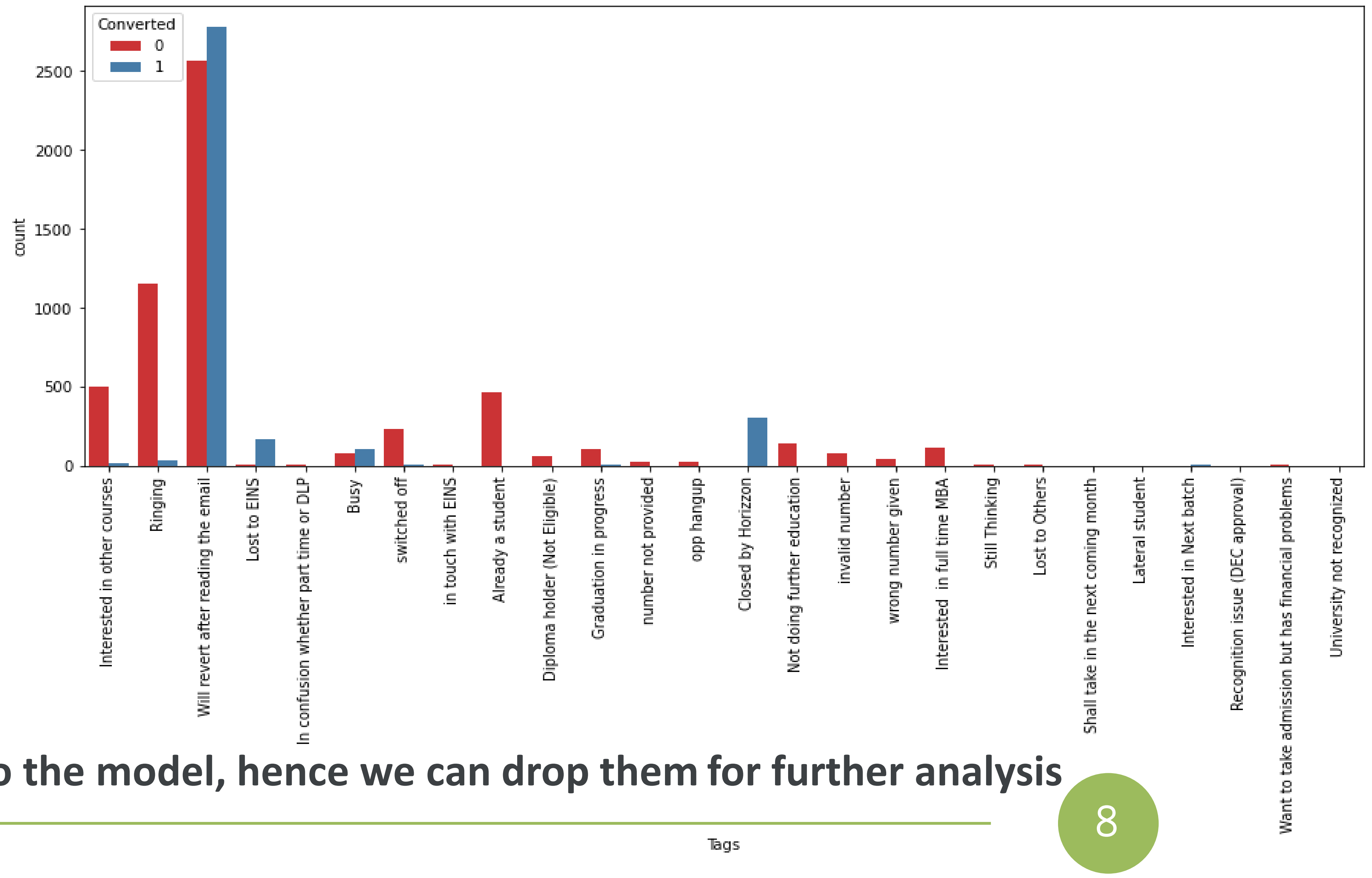












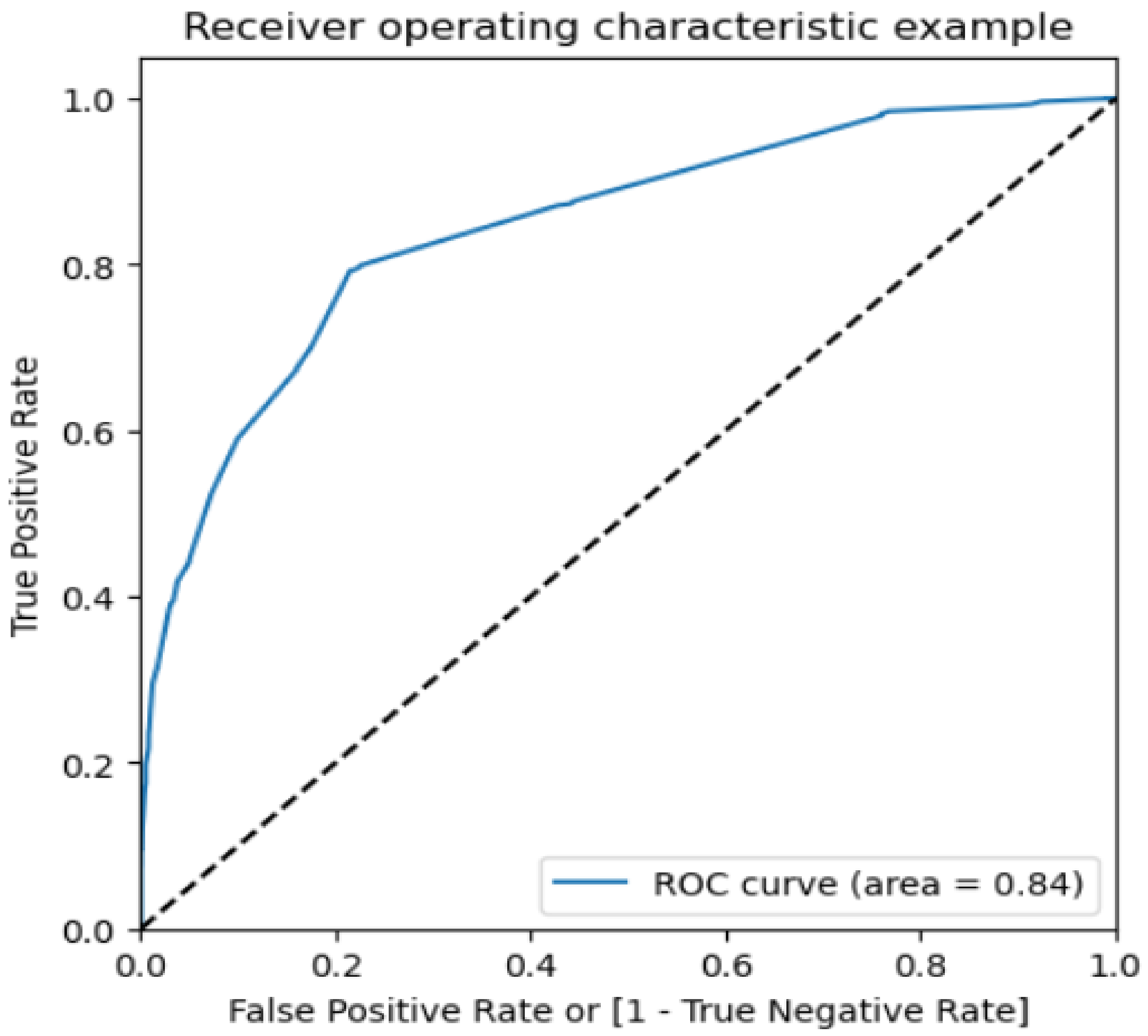
**Based on the univariate analysis we have seen that many columns are not adding any information to the model, hence we can drop them for further analysis**

# Model

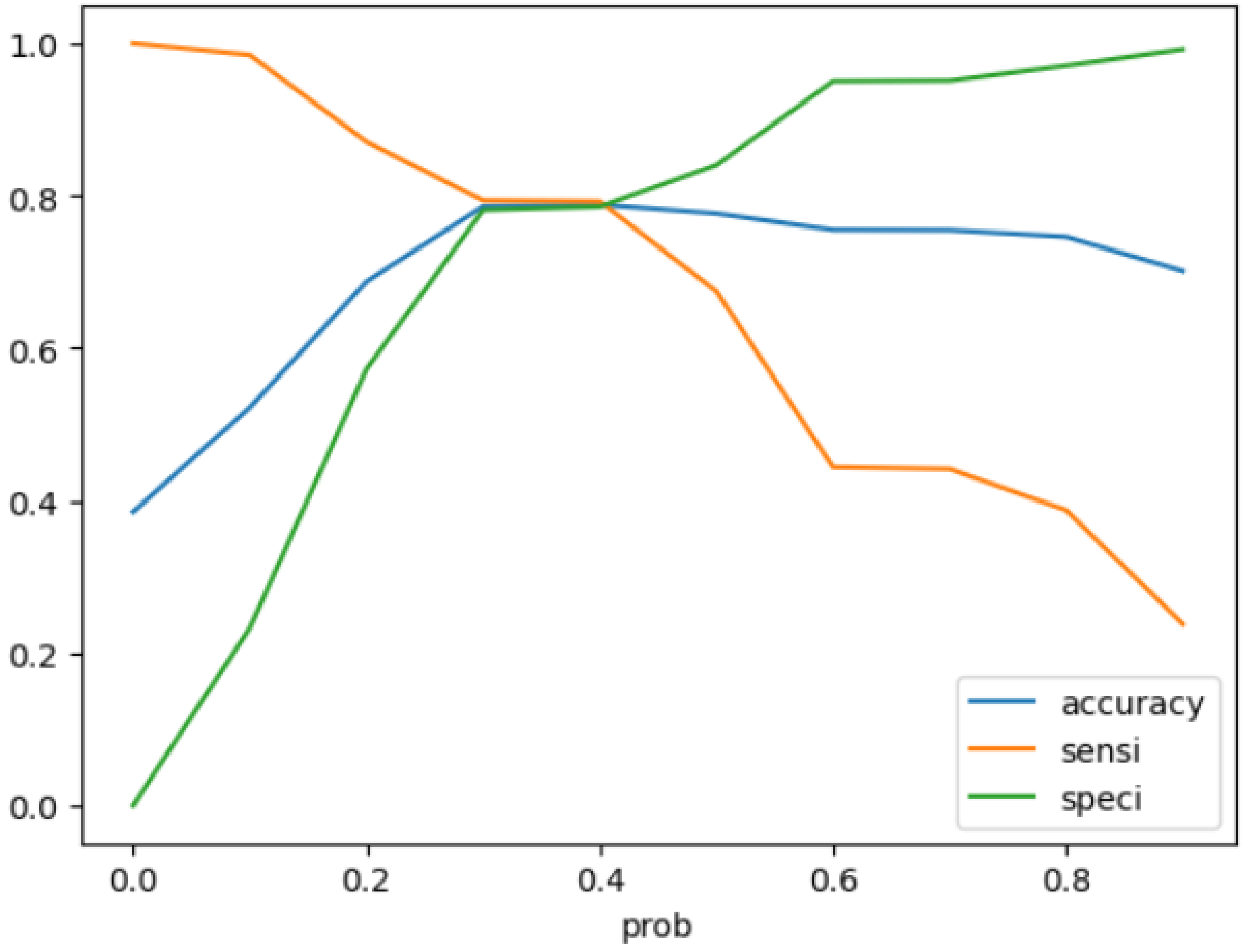
## Plotting the ROC Curve

An ROC curve demonstrates several things:

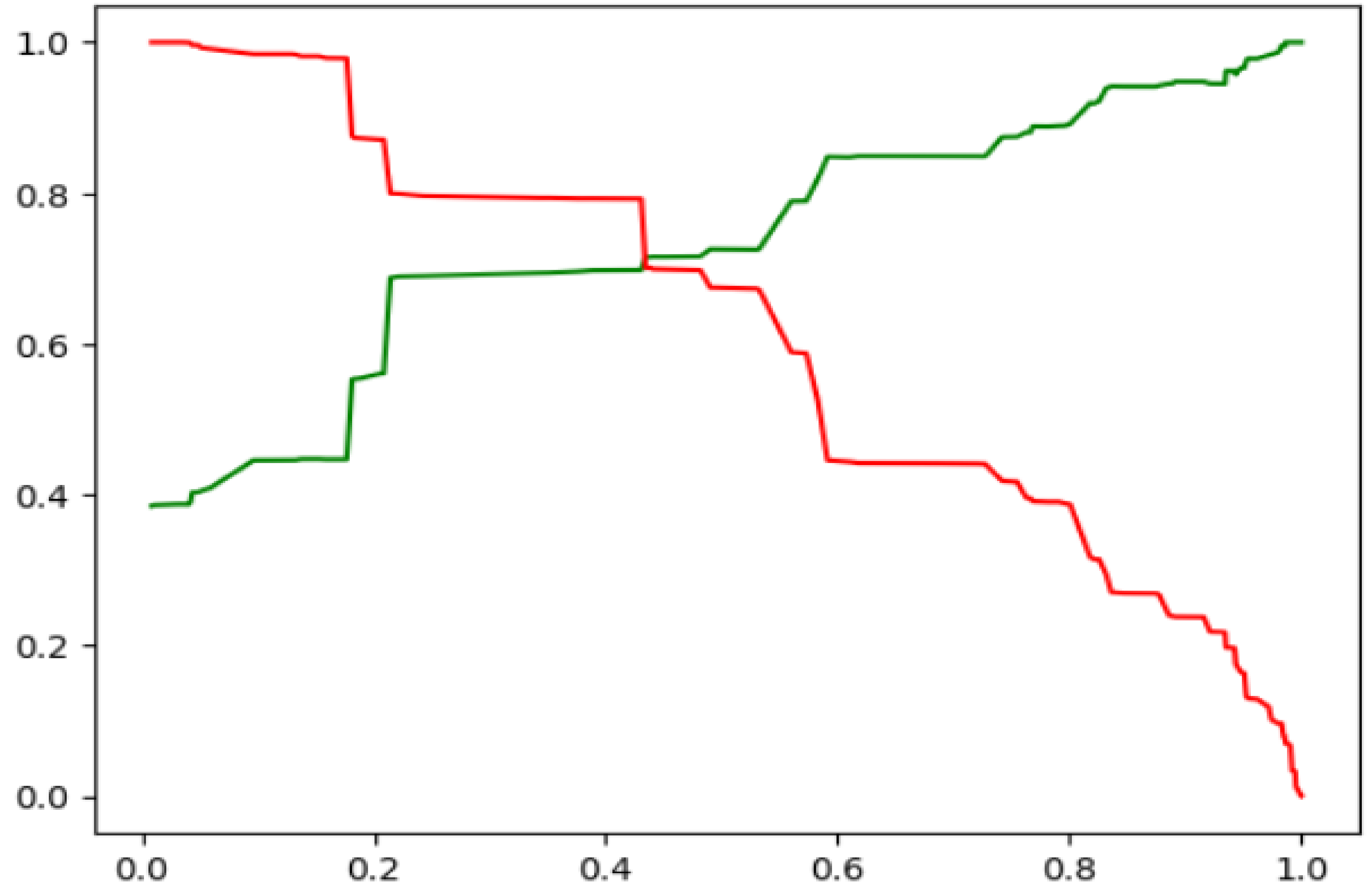
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



accuracy sensitivity and specificity for various probabilities



trade-off curve between precision and recall



Results :Comparing the values obtained for Train & Test:

Train Data : Accuracy : 79 % Sensitivity : 79 % Specificity : 70 %

Test Data : Accuracy : 78 % Sensitivity : 72 % Specificity : 81 %

# Thank you

Submitted by  - Godwin Neal
Email : gnneal96@gmail.com