# LEAD SCORE CASE STUDY

GROUP MEMBERS:

AATISH KAUSHAL

HEMAVATHI A.B

# PROBLEM STATEMENT

- X education sells online courses to industry professionals.
- X education gets a lot of leads, but its lead conversion rate is very poor.
- 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.

**Business Objective:**

- X Education wants to identify the most promising leads
- With that they want to create a Model which identify hot leads
- Deployment of the model for the future use.

# SOLUTION APPROACH

**Data cleaning and manipulation:**

- Identify and manage duplicate entries.
- Address NAN or missing values and option “SELECT”.
- Remove columns with significant missing data that are not essential for analysis.
- Impute missing values when necessary.
- Detect and manage outliers.

**EDA:**

Univariate Data Analysis: Analyze value counts and variable distributions.

Bivariate Data Analysis: Examine correlation coefficients and patterns between variables.

**Feature Scaling & Dummy Variables and encoding of the data**

**Classification technique: logistic regression used for the model making and prediction**

**Validation of the model**

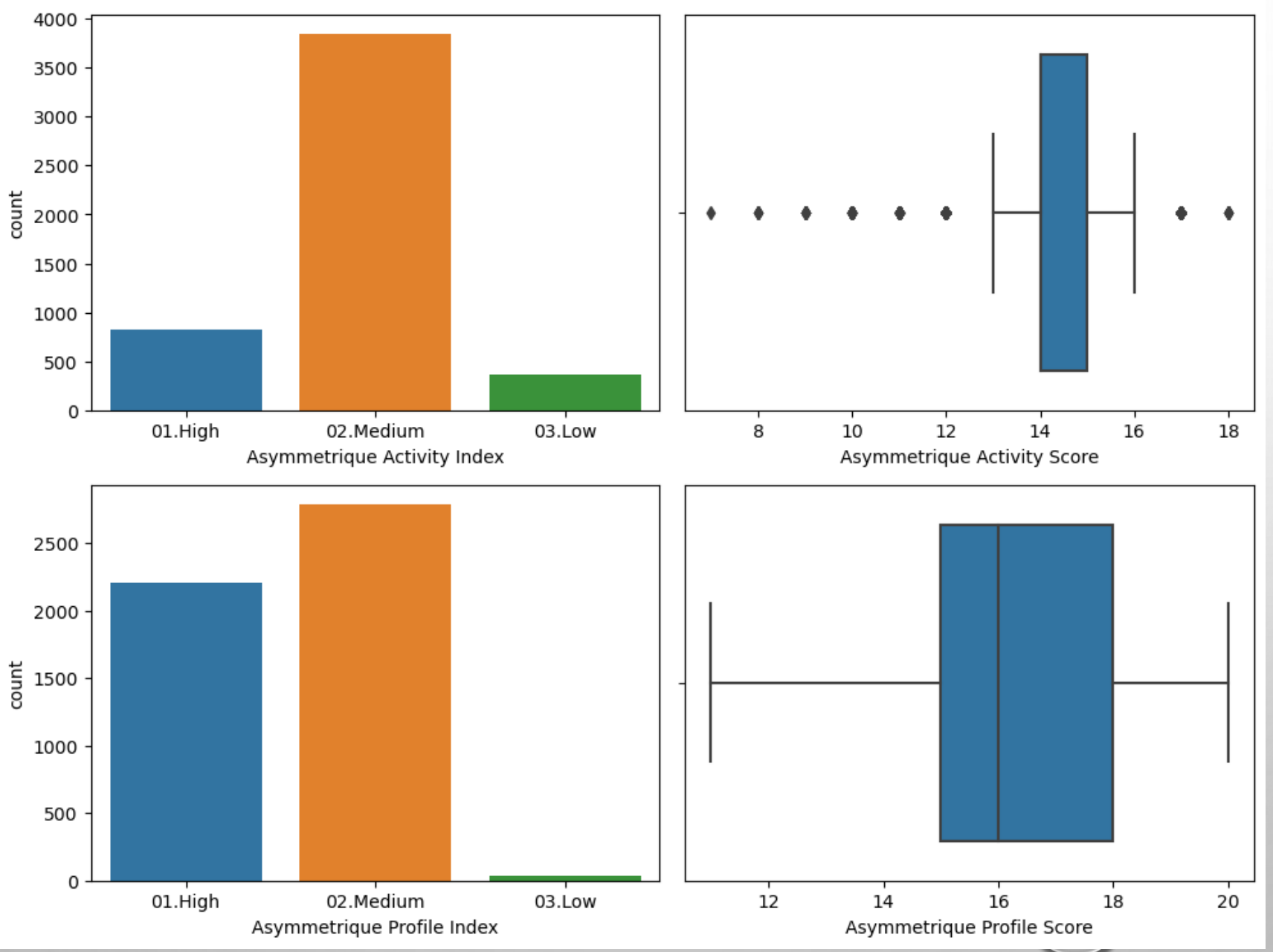
**Model presentation**

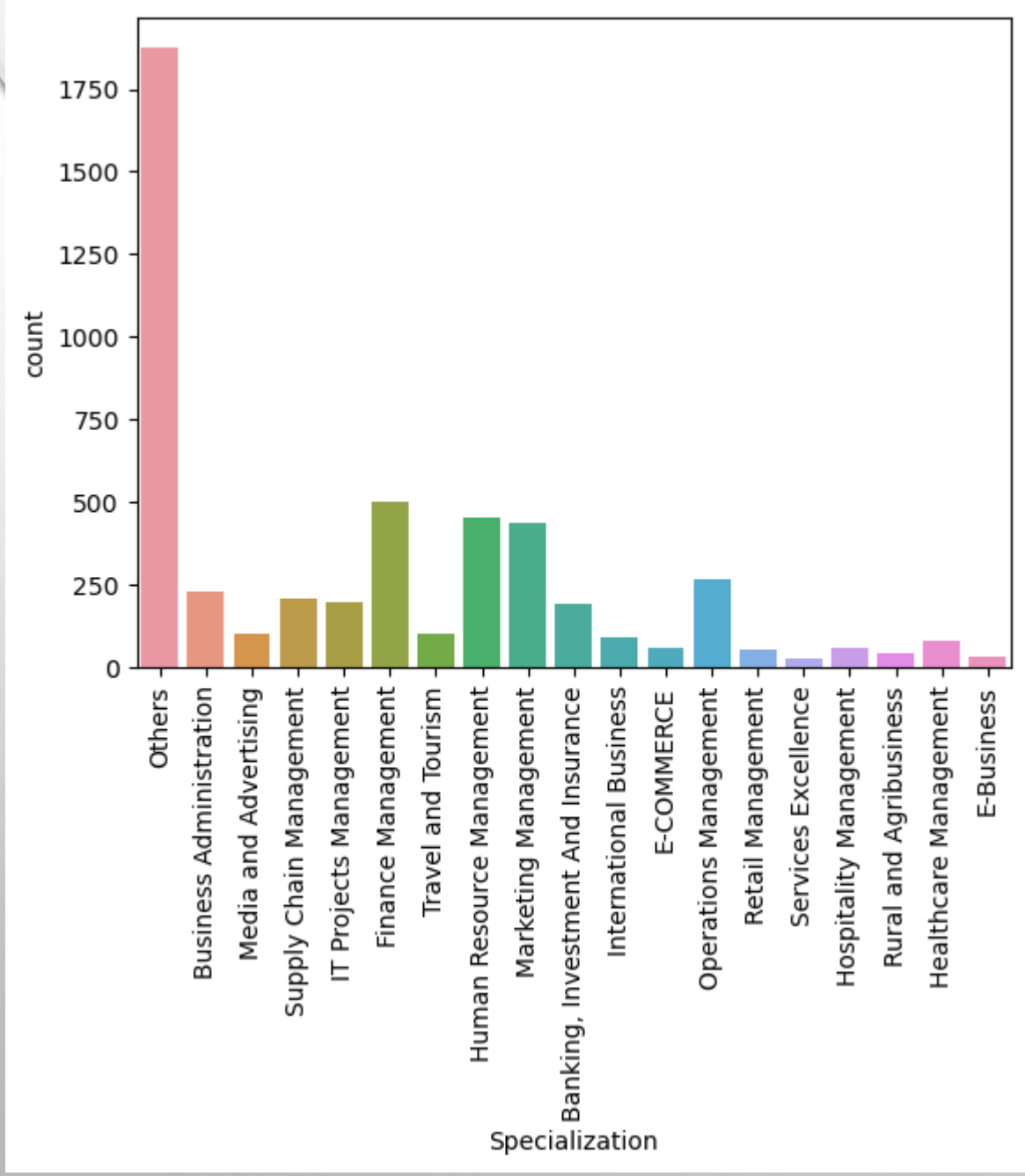
**Conclusions and recommendations**

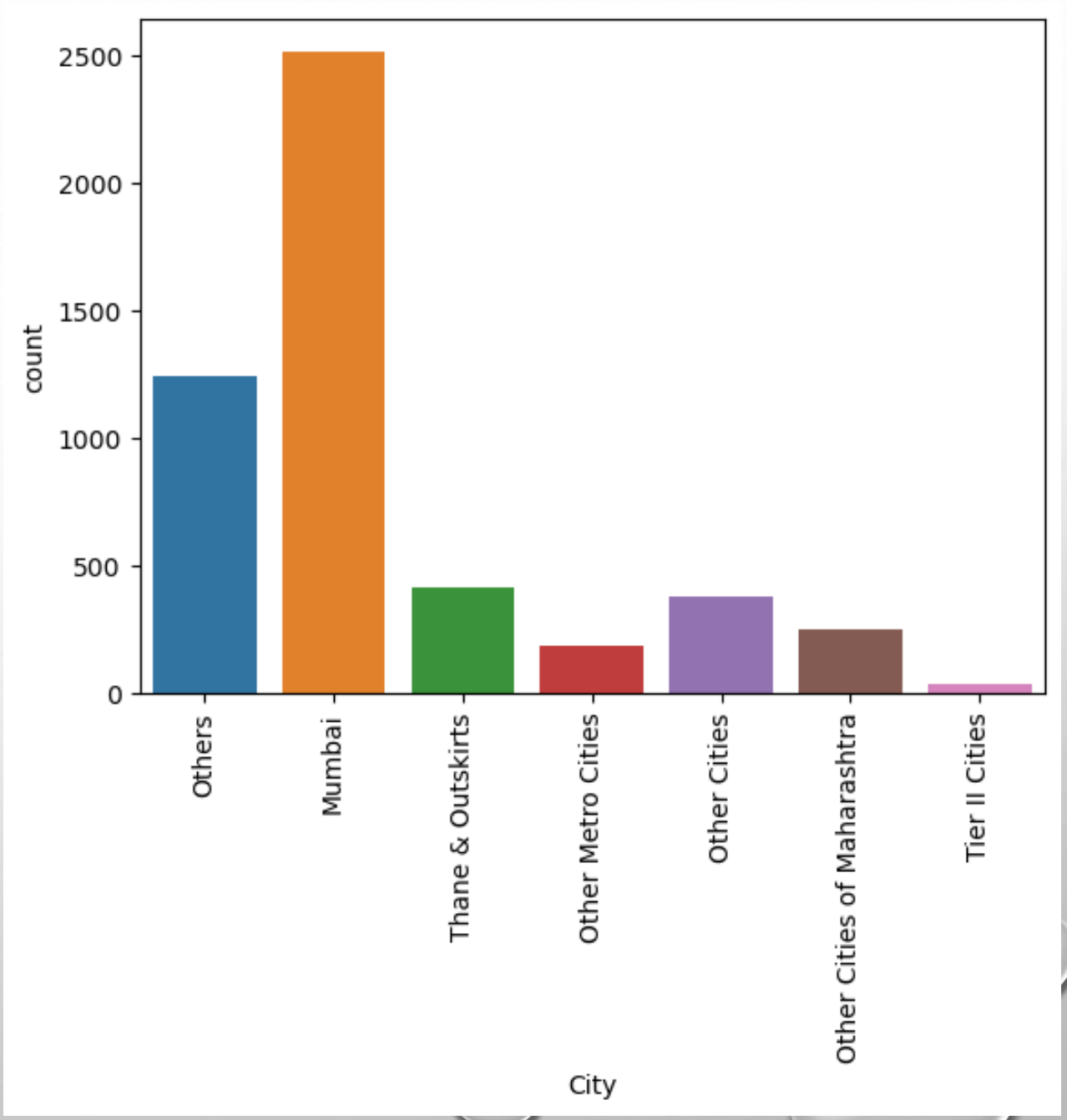
# DATA MANIPULATION

- TOTAL NUMBER OF ROWS = 9240, TOTAL NUMBER OF COLUMNS =37.
- MANY COLUMNS HAD SINGLE VALUE FEATURE “NO” LIKE INDICATING WHETHER THE CUSTOMER HAD SEEN THE AD IN ANY OF THE LISTED ITEMS FOR EXAMPLE “MAGAZINE” AND FOR , “RECEIVE MORE UPDATES ABOUT OUR COURSES”, “UPDATE ME ON SUPPLY CHAIN CONTENT”, “I AGREE TO PAY THE AMOUNT THROUGH CHEQUE” ETC.
- AN INDEX AND SCORE ASSIGNED TO EACH CUSTOMER BASED ON THEIR ACTIVITY AND THEIR PROFILE COLUMNS WERE DROPPED AS THEY SHOWED MORE THAN >45% NULL ROWS AND TOO MUCH OF VARIATIONS.
- CHOOSE MEDIAN AND MODE STATISTICS TO IMPUTE FOR MISSING ROWS AND REPLACED NAN AND SELECT WITH ‘’OTHERS”
- IN THE ‘’TAGS” COLUMN KEPT CONSIDERABLE LAST ACTIVITIES AS IS AND CLUBBED ALL OTHERS TO "OTHER_ACTIVITY"
- REMOVING THE “PROSPECT ID” AND “LEAD NUMBER” WHICH IS NOT NECESSARY FOR THE ANALYSIS.
- AFTER CHECKING FOR THE VALUE COUNTS FOR SOME OF THE OBJECT TYPE VARIABLES, WE FOUND SOME OF THE FEATURES WHICH HAD NOT ENOUGH VARIANCE, IS DROPPED, THE FEATURES ARE: “DO NOT CALL”, “WHAT MATTERS MOST TO YOU IN CHOOSING COURSE”, “SEARCH”, “NEWSPAPER ARTICLE”, “X EDUCATION FORUMS”, “NEWSPAPER”, “DIGITAL ADVERTISEMENT” ETC.
- DROPPING THE COLUMNS LEAD ORIGIN', 'LEAD SOURCE', 'LAST ACTIVITY', 'SPECIALIZATION','WHAT IS YOUR CURRENT OCCUPATION','TAGS','LEAD QUALITY','HOW DID YOU HEAR ABOUT X EDUCATION','LEAD PROFILE','CITY','LAST NOTABLE ACTIVITY

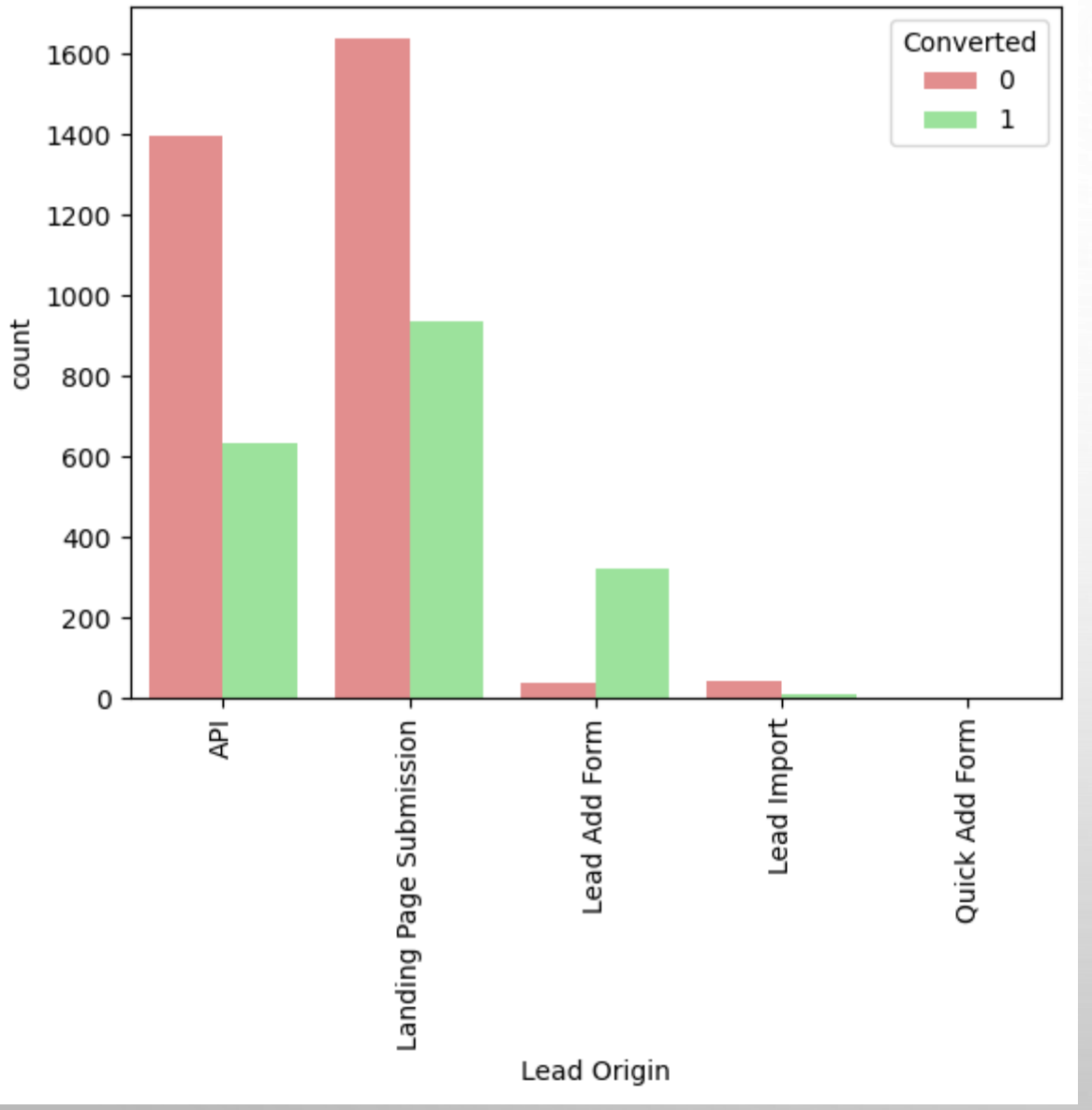
# EDA

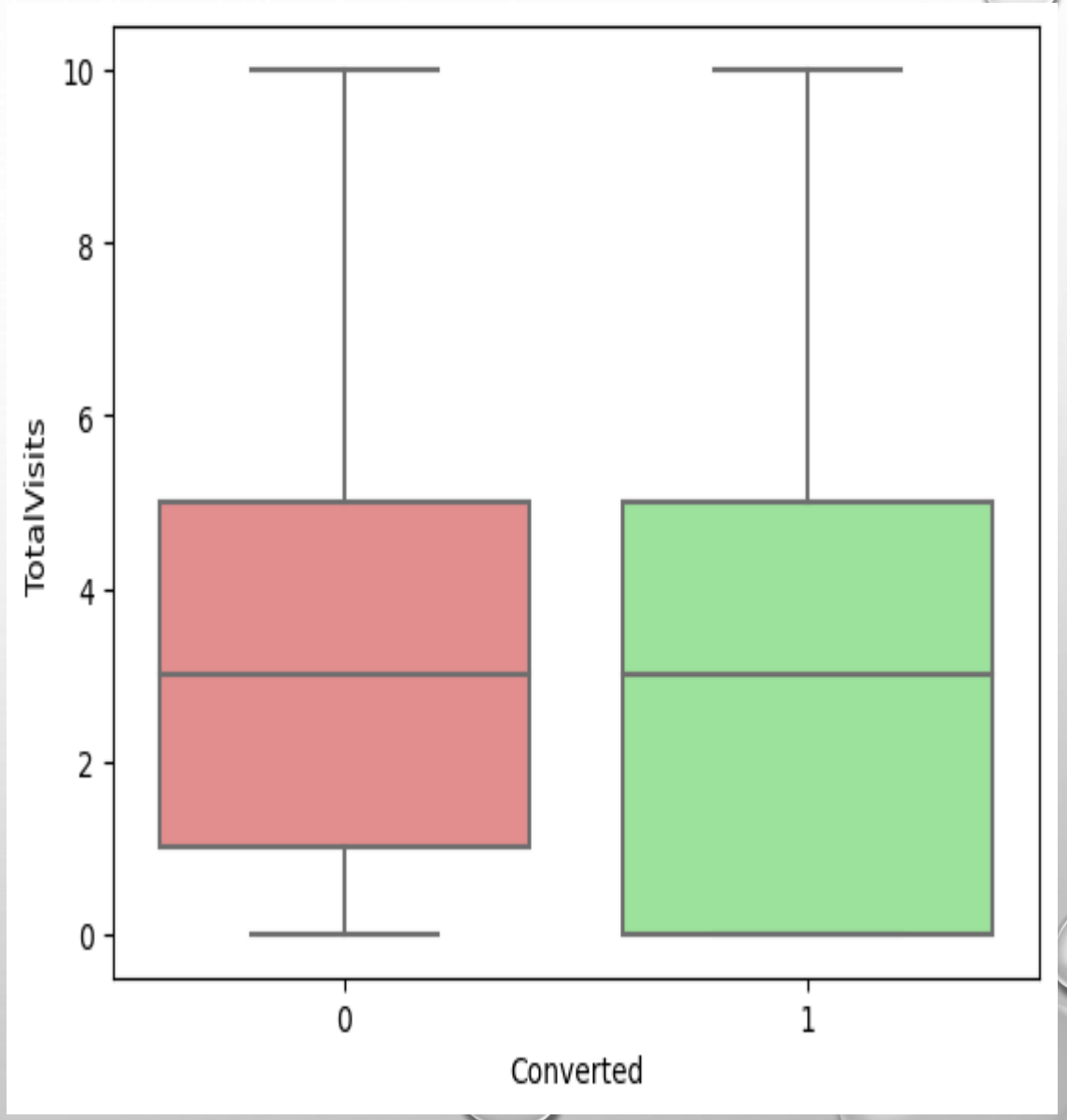

4000
3500
3000
2500
2000
1500
1000
500
0
count
01.High
02.Medium
03.Low
Asymmetrique Activity Index
8
10
12
14
16
18
Asymmetrique Activity Score
2500
2000
1500
1000
500
0
count
01.High
02.Medium
03.Low
Asymmetrique Profile Index
12
14
16
18
20
Asymmetrique Profile Score

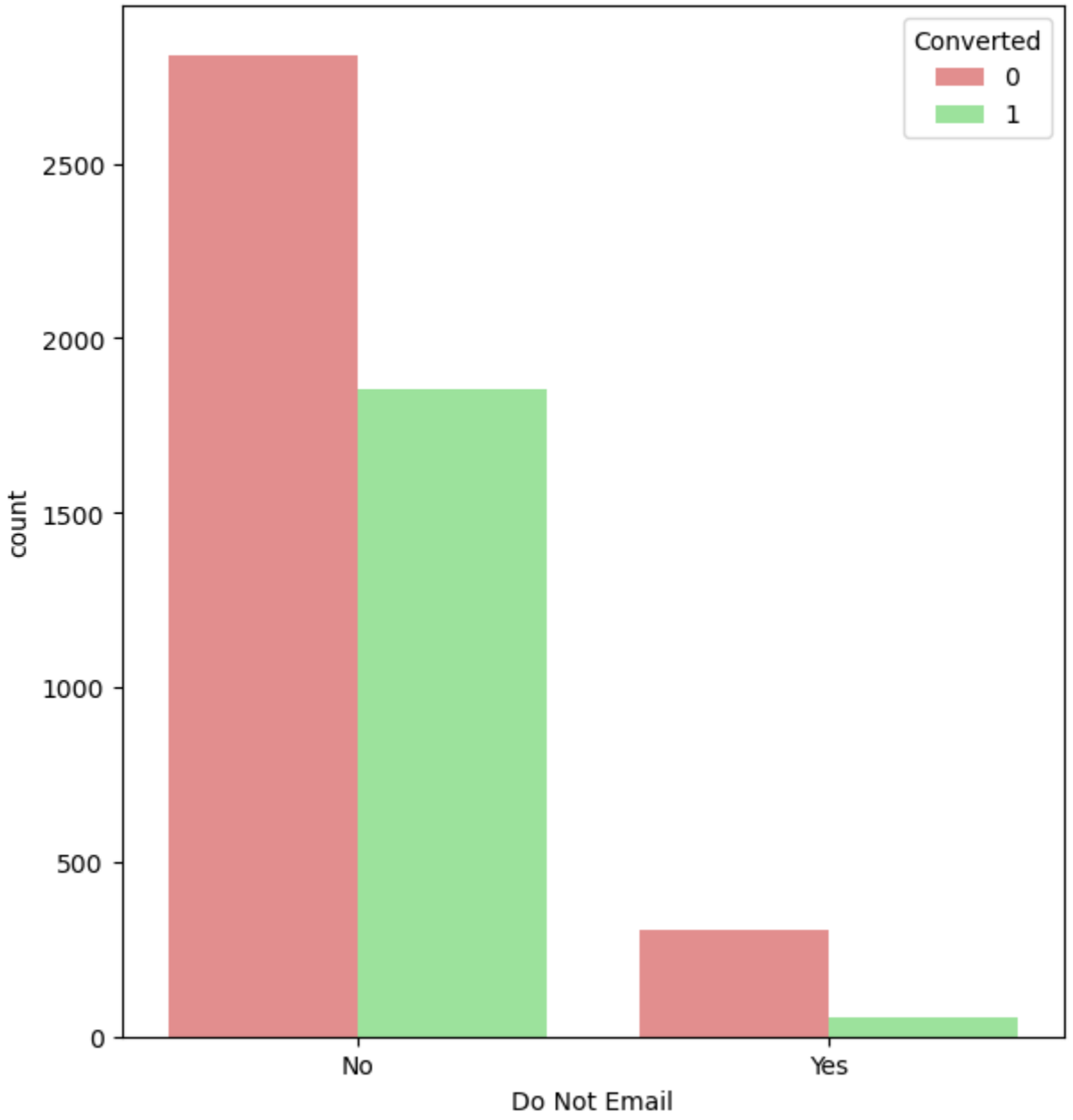

count
1750
1500
1250
1000
750
500
250
0
Others
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

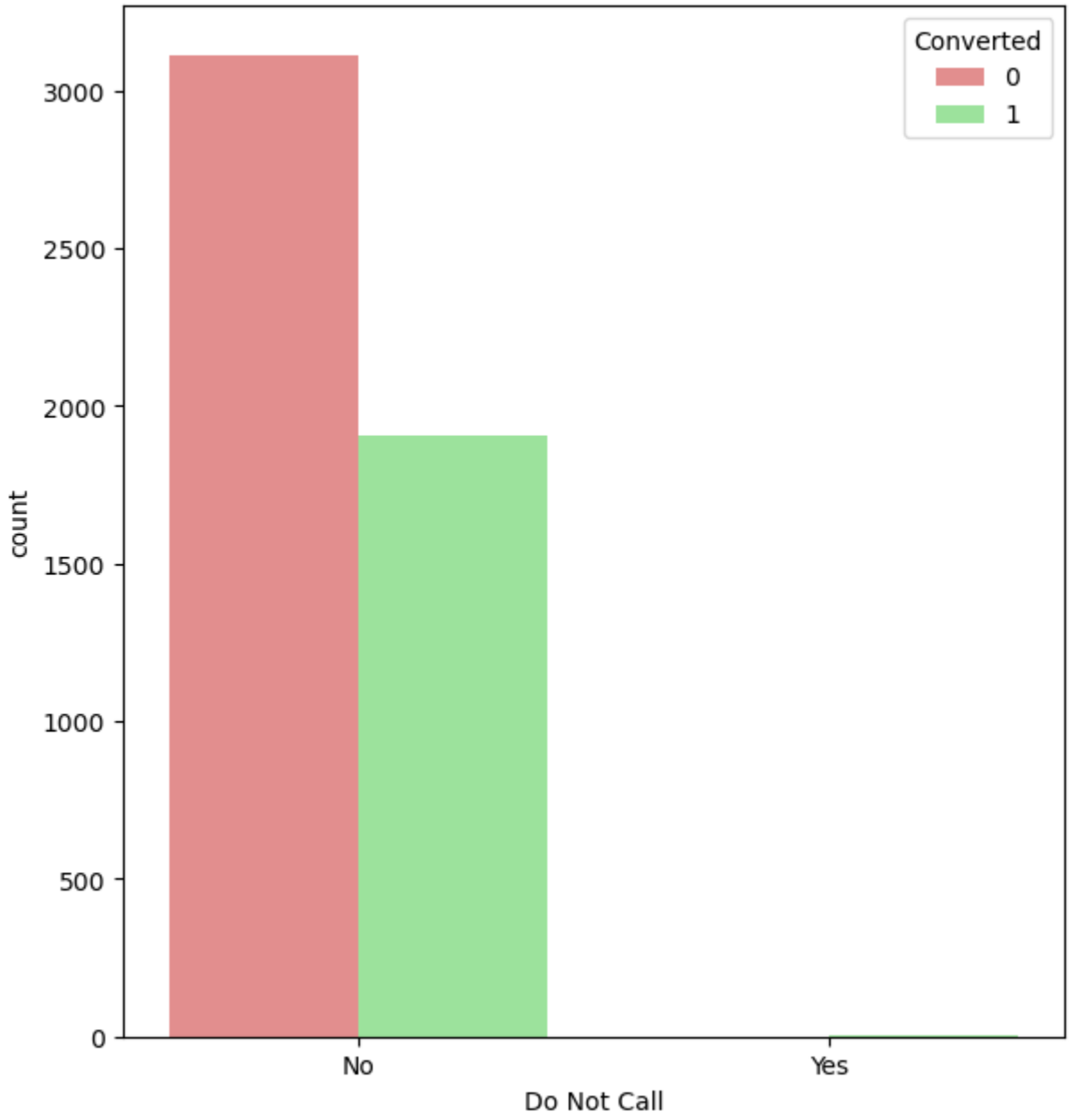

count
2500
2000
1500
1000
500
0
Others
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City

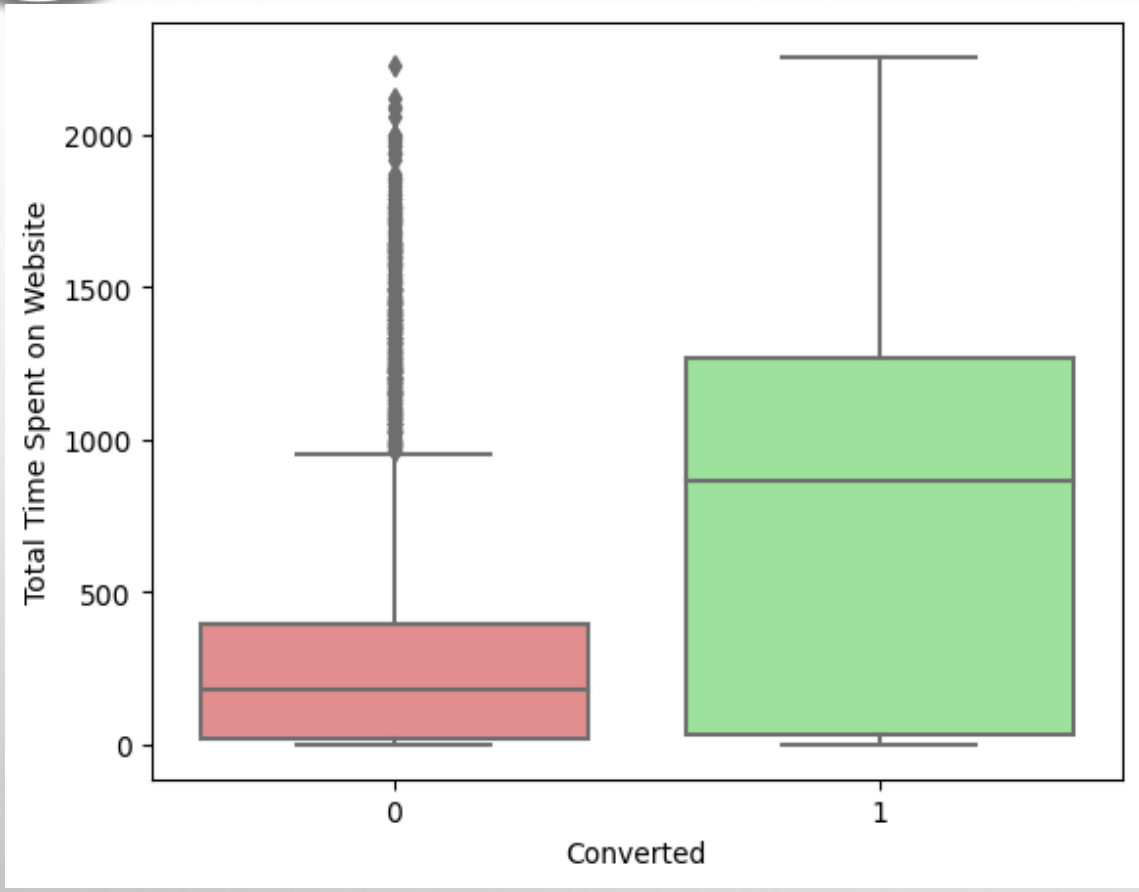
# CATEGORICAL VARIABLE RELATION

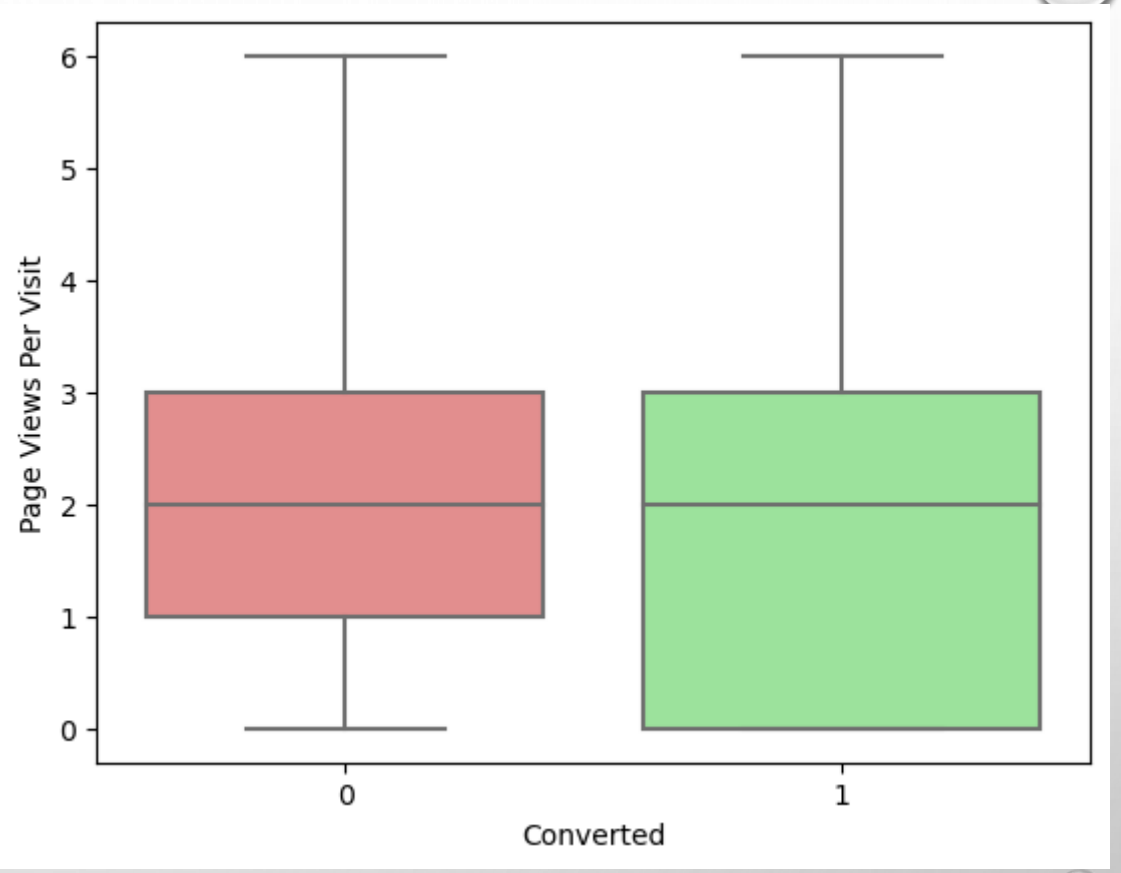



Converted
0
1
2500
2000
1500
1000
500
0
count
No
Yes
Do Not Email


Converted
0
1
3000
2500
2000
1500
1000
500
0
count
No
Yes
Do Not Call


Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted


Page Views Per Visit
6
5
4
3
2
1
0
0
1
Converted

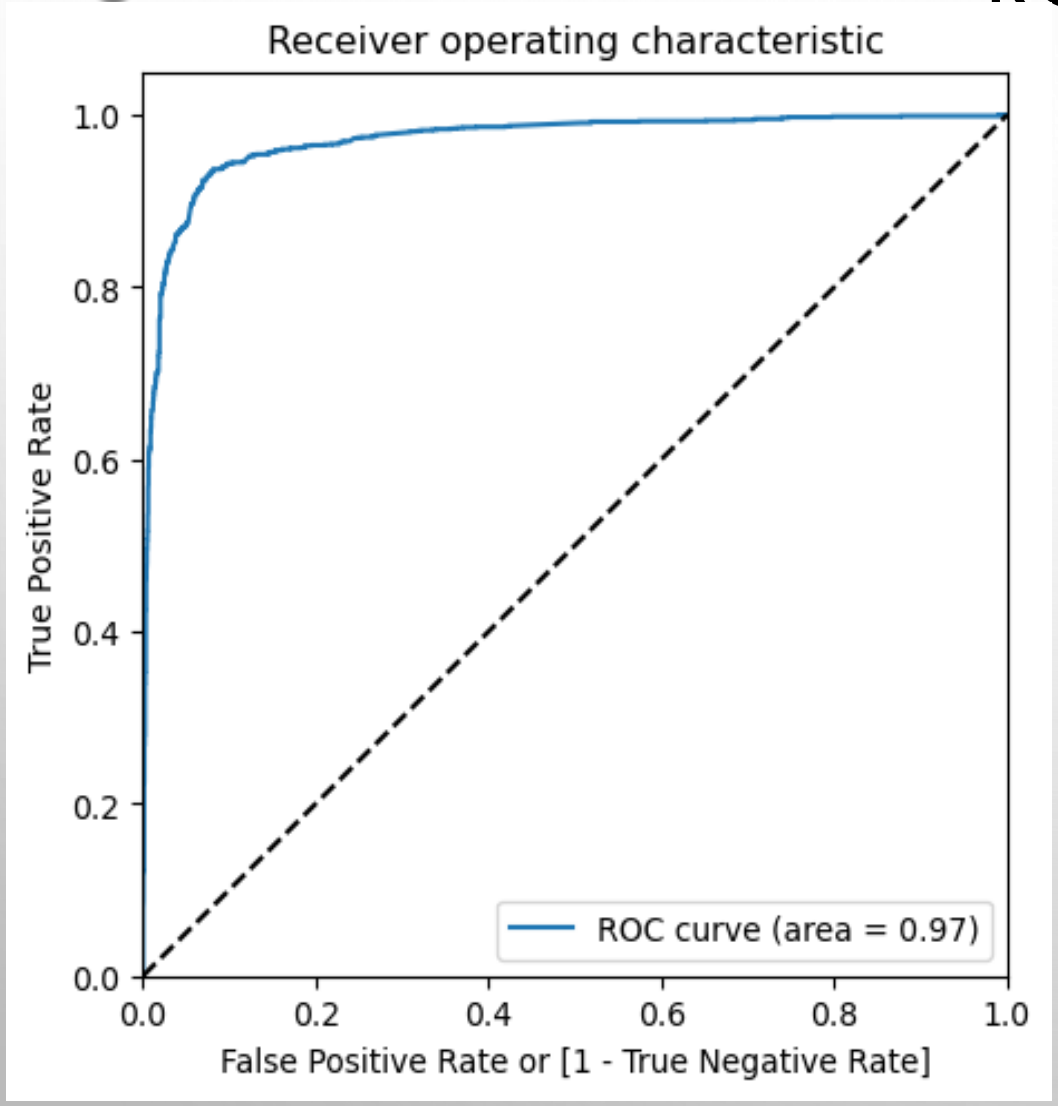
# DATA CONVERSION

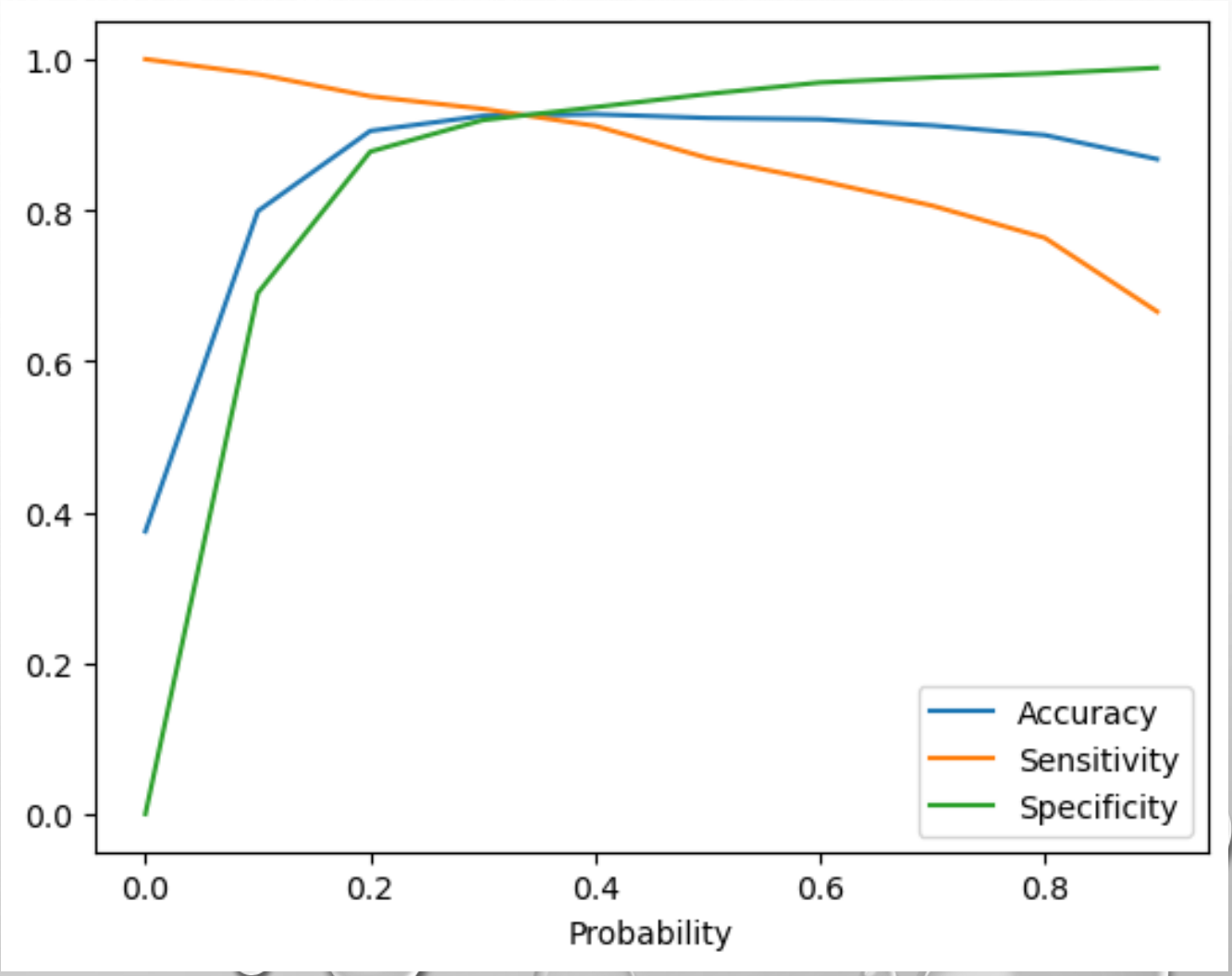
- Numerical variables are normalised using StandardScaler().
- Dummy variables are created for object type variables.
- Total rows for analysis: 5022.
- Total columns for analysis: 86.

# MODEL BUILDING

- SPLITTING THE DATA INTO TRAINING AND TESTING SETS
- THE FIRST BASIC STEP FOR REGRESSION IS PERFORMING A TRAIN-TEST SPLIT, WE HAVE CHOSEN 80:20 RATIO.
- USE RFE FOR FEATURE SELECTION
- RUNNING RFE WITH 20 VARIABLES AS OUTPUT
- BUILDING MODEL BY REMOVING THE VARIABLE WHOSE P- VALUE IS GREATER THAN 0.05 AND VIF VALUE IS GREATER THAN 5
- PREDICTIONS ON TEST DATA SET
- OVERALL ACCURACY 92%

# ROC CURVE


Receiver operating characteristic
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.97)
0.0
0.2
0.4
0.6
0.8
1.0


Probability
Accuracy
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0

# CONCLUSION

**THE KEY FACTORS INFLUENCING POTENTIAL BUYERS, IN DESCENDING ORDER OF IMPORTANCE, ARE:**

- TOTAL TIME SPENT ON THE WEBSITE
- TOTAL NUMBER OF VISITS
- LEAD SOURCE, SPECIFICALLY:
    - GOOGLE
    - DIRECT TRAFFIC
    - ORGANIC SEARCH
    - WELINGAK WEBSITE
- LAST ACTIVITY, SPECIFICALLY:
    - SMS
    - OLARK CHAT CONVERSATION
- LEAD ORIGIN BEING LEAD ADD FORMAT
- CURRENT OCCUPATION AS A WORKING PROFESSIONAL BY FOCUSING ON THESE FACTORS, X EDUCATION CAN SIGNIFICANTLY INCREASE THEIR CHANCES OF CONVERTING POTENTIAL BUYERS INTO CUSTOMERS, THEREBY BOOSTING THEIR COURSE SALES.