Lead Score Case Study

Presented by:

Vishal Rai

Kiran Panchavati

Problem Statement and Business Objectives:

Problem Statement:

- An education company X Education sells online courses to industry professionals.
- 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%.

Business Objectives:

- 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.
- We need to build a machine learning model that assigns a lead score to each lead that’s been generated on their website and make a prediction on the chances of such leads getting converted.

Strategy Employed:

- Data Sourcing for analysis.
- Data cleaning and preparation.

  - Handling missing values.

  - Handling outliers in numerical attributes.

  - Checking Data imbalance in categorical attributes.

  - Imputing missing values ,if necessary.

  - Dropping less significant attributes.
- Exploratory Data Analysis.

  - Univariate Analysis

  - Bivariate Analysis

  - Multivariate Analysis
- Data Preparation and modelling.

  - Train-Test Split

  - Scaling Variables(Standardization)

  - Creation of Dummy variable
- Feature selection(RFE).
- Model Building(Classification technique-Logistic Regression).
- Model Evaluation.
- Making Predictions.

Data Wrangling and Manipulation

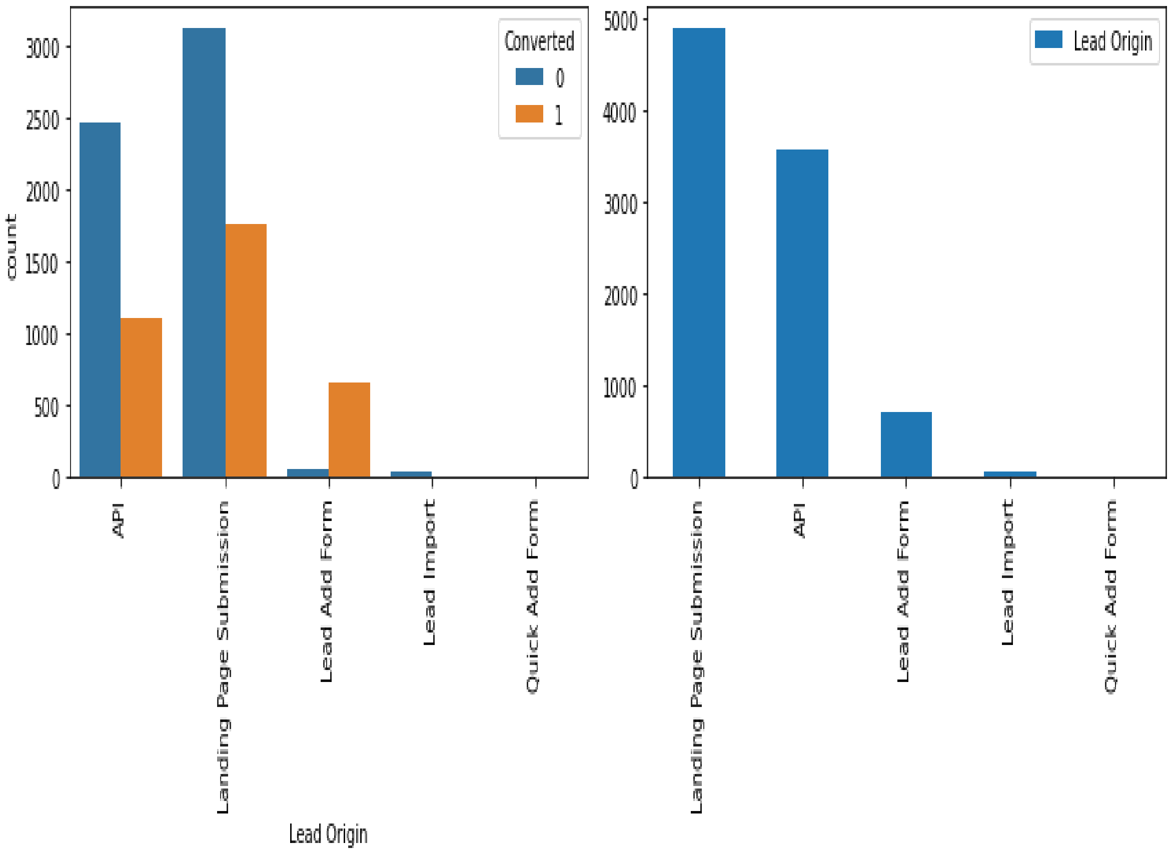
The Dataset had 37 attributes and 9240 entries with 7 Numerical attributes and 30 Categorical attributes.

To check data imbalance, attributes such as "I agree to pay the amount through cheque, Get updates on DM Content, Update me on Supply Chain Content, Receive More Updates About Our Courses, Through Recommendations, Digital Advertisement, Newspaper, X Education Forums, Newspaper Article, What matters most to you in choosing a course, Search, Country, Do Not Call, Magazine" were removed.

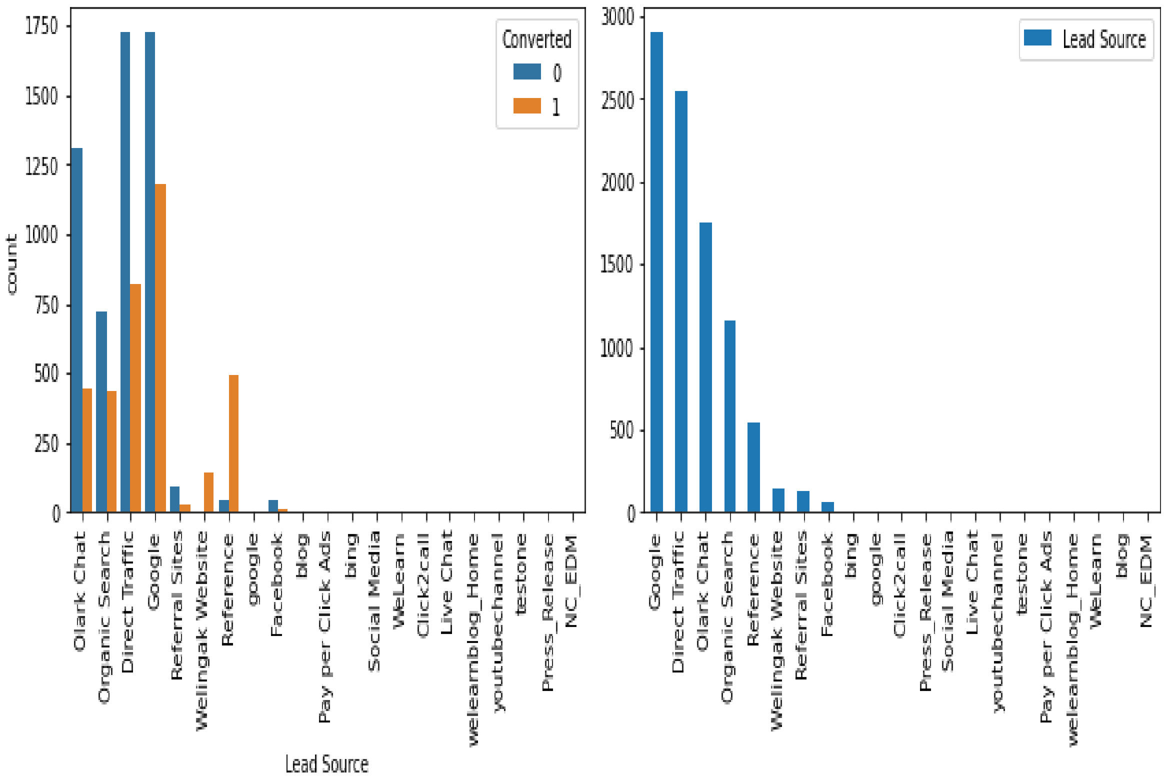
Due to presence of high missing values we also removed attributes such as "How did you hear about X Education, Lead Quality, Lead Profile, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score'. These attributes were removed as the missing values were over 45% of the total entries.

Dropped attributes such as "Lead Number" and "Prospect ID" as they were deemed unnecessary for the analysis.

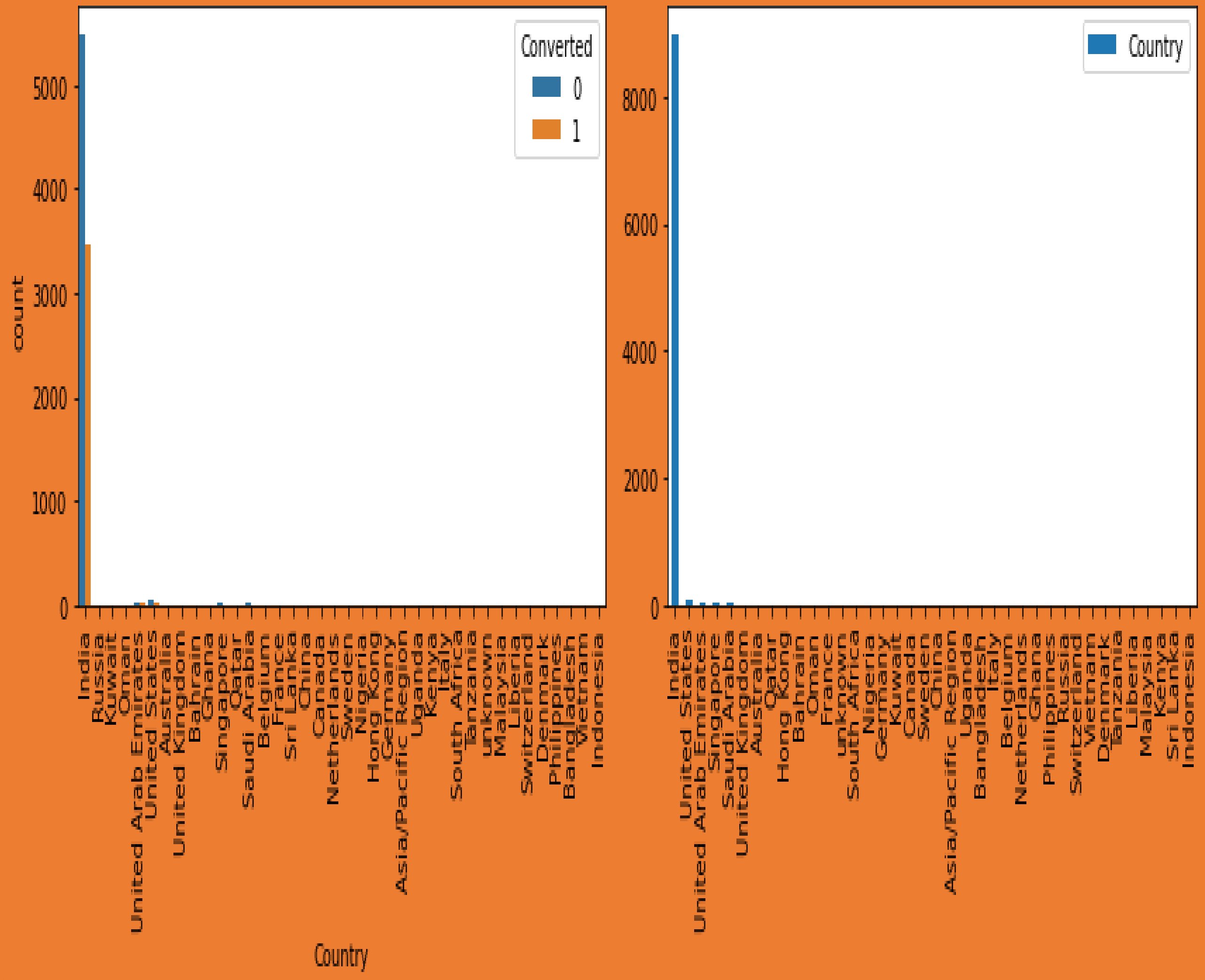
Exploratory Data Analysis

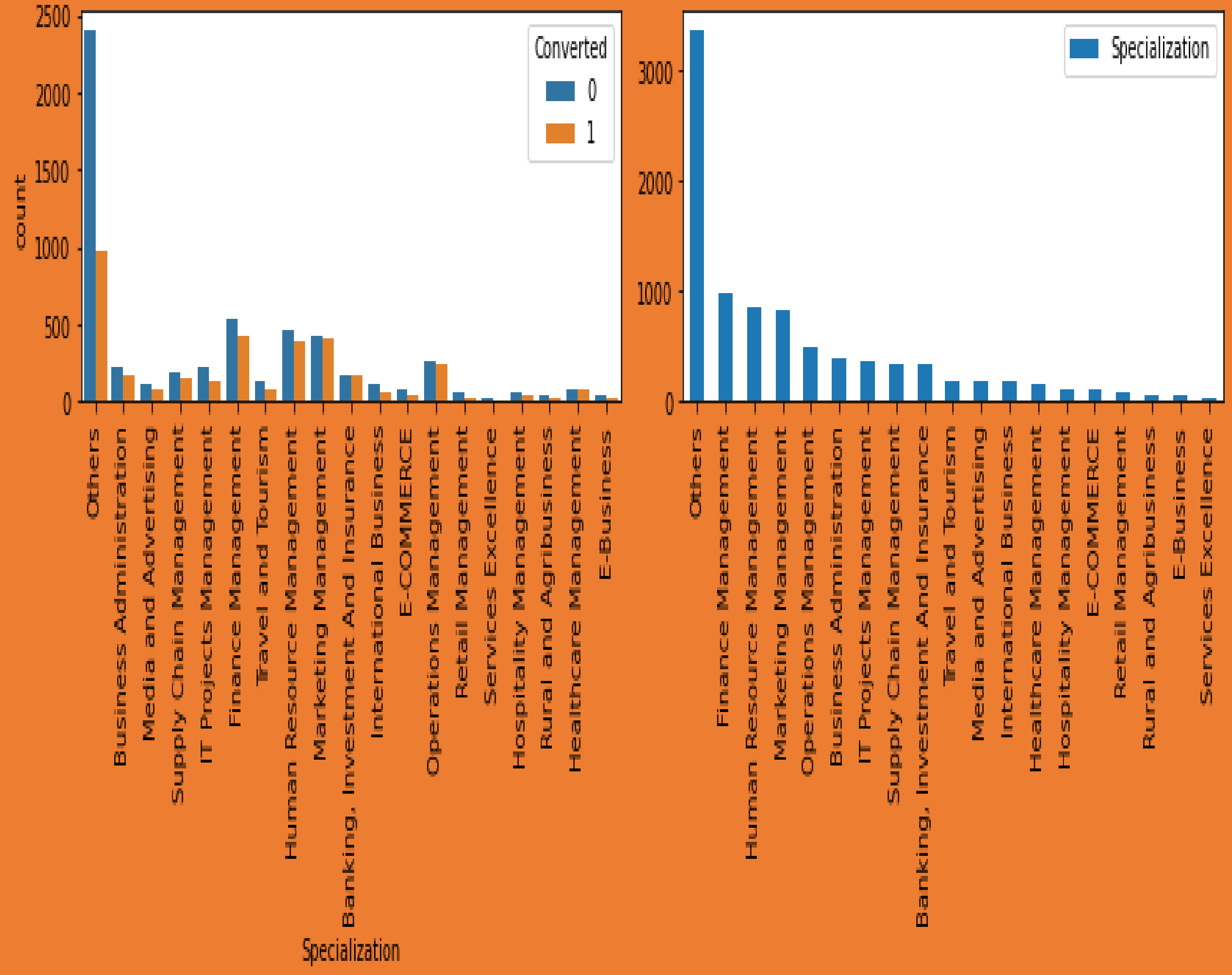


- The maximum leads are originated on "Landing Page Submission".
- Maximum leads are converted that has their origin as "Landing Page Submission", "Lead Add Form" or "API".



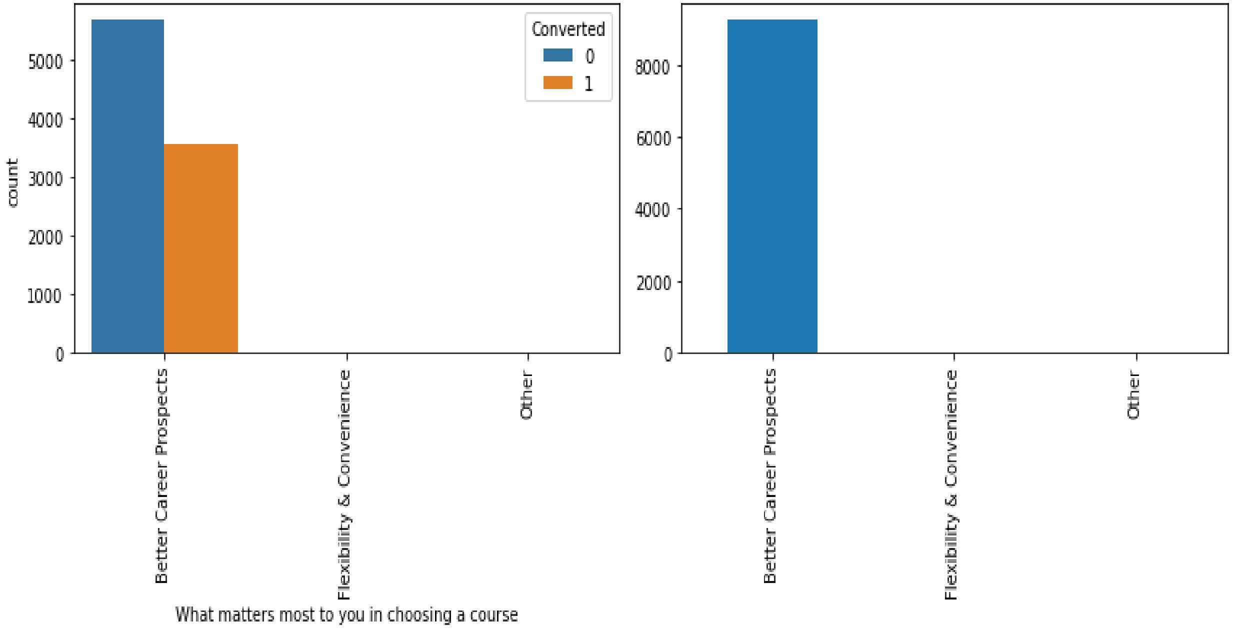
- The maximum leads are sourced from "Google" or has "Direct Traffic" as their source.
- Maximum leads are converted that has their source as "Google", "Direct Traffic" or "Reference".

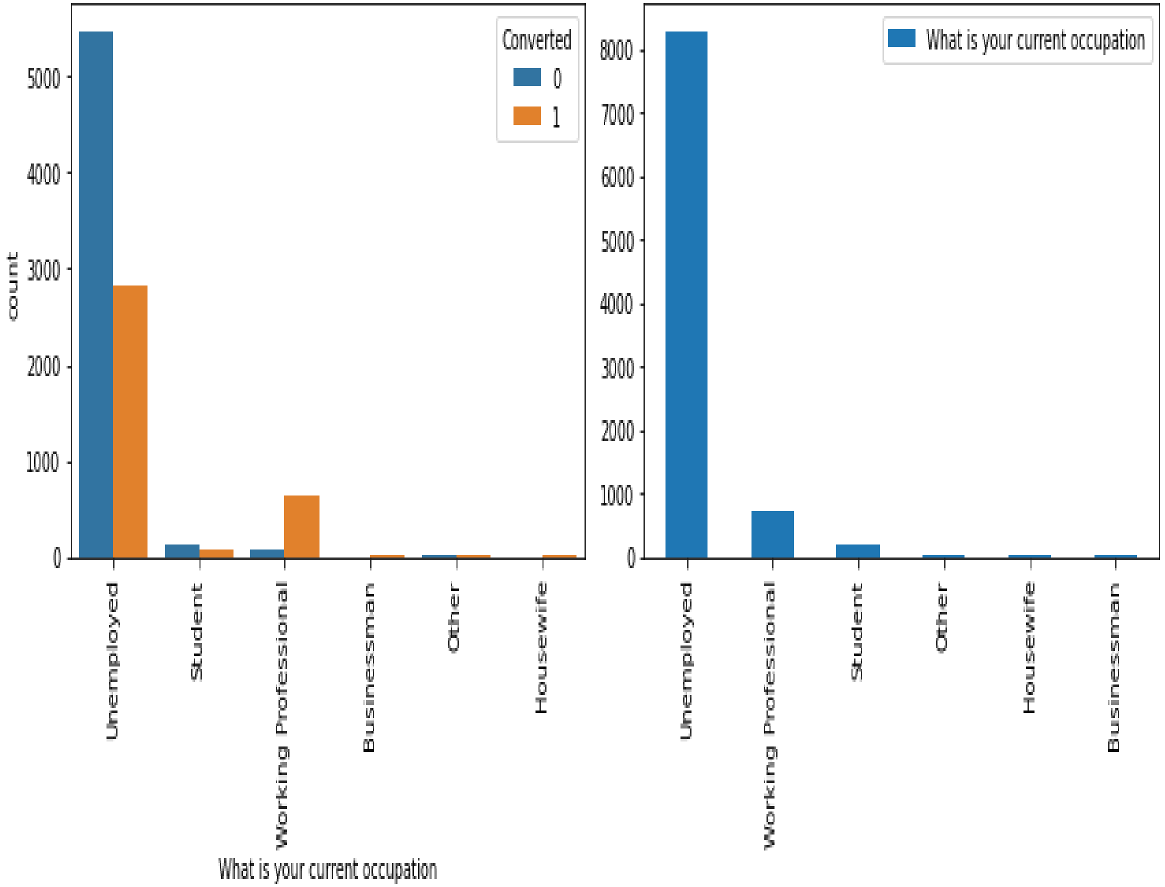


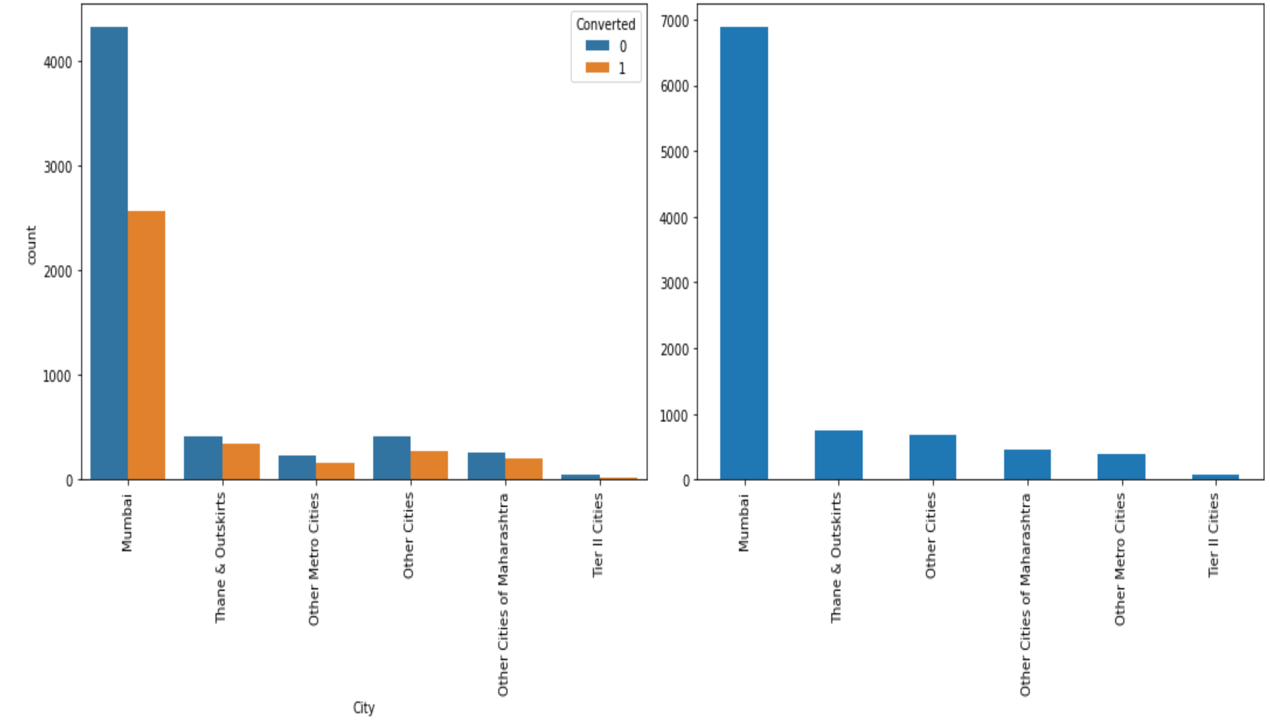


- Most leads generated are from India followed by USA.
- Most leads conversion also happens from India followed by Russia.

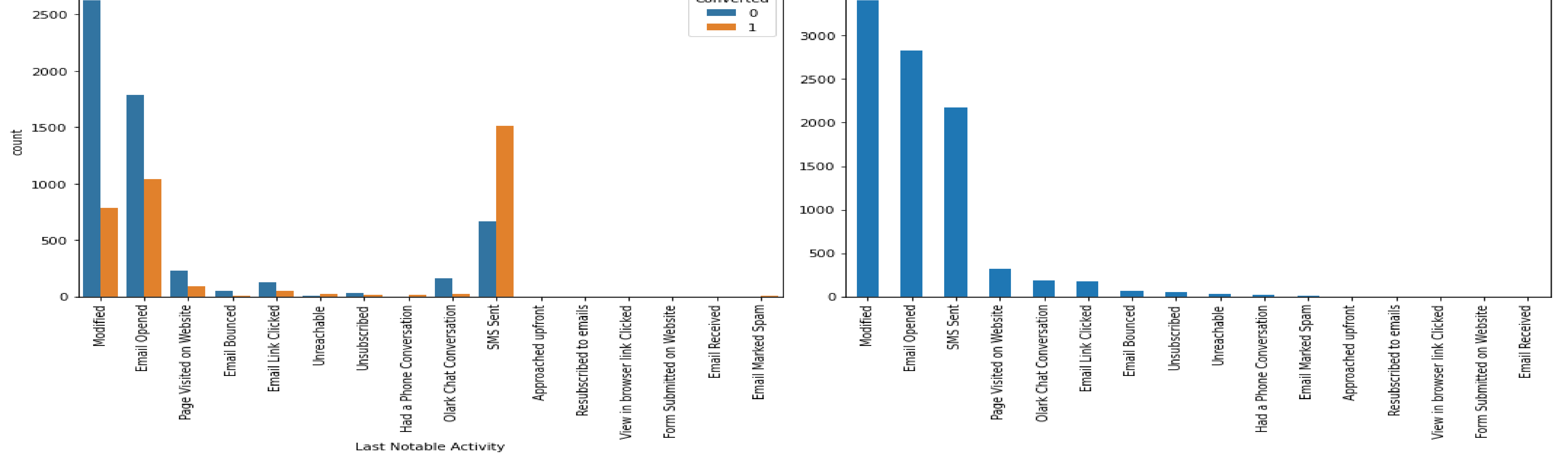
Maximum leads generated and converted have selected management as their specialization.

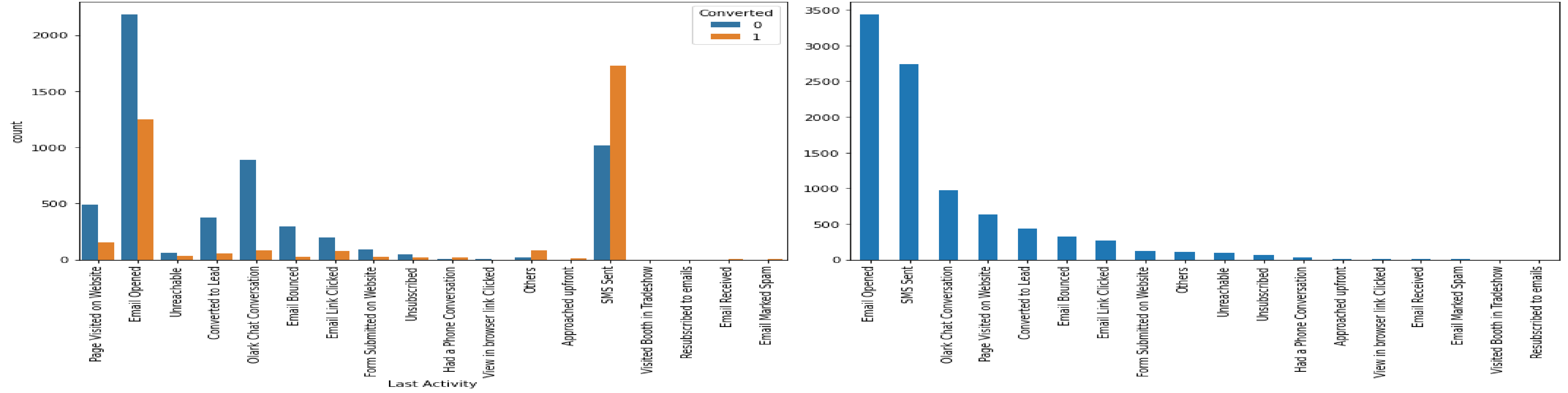


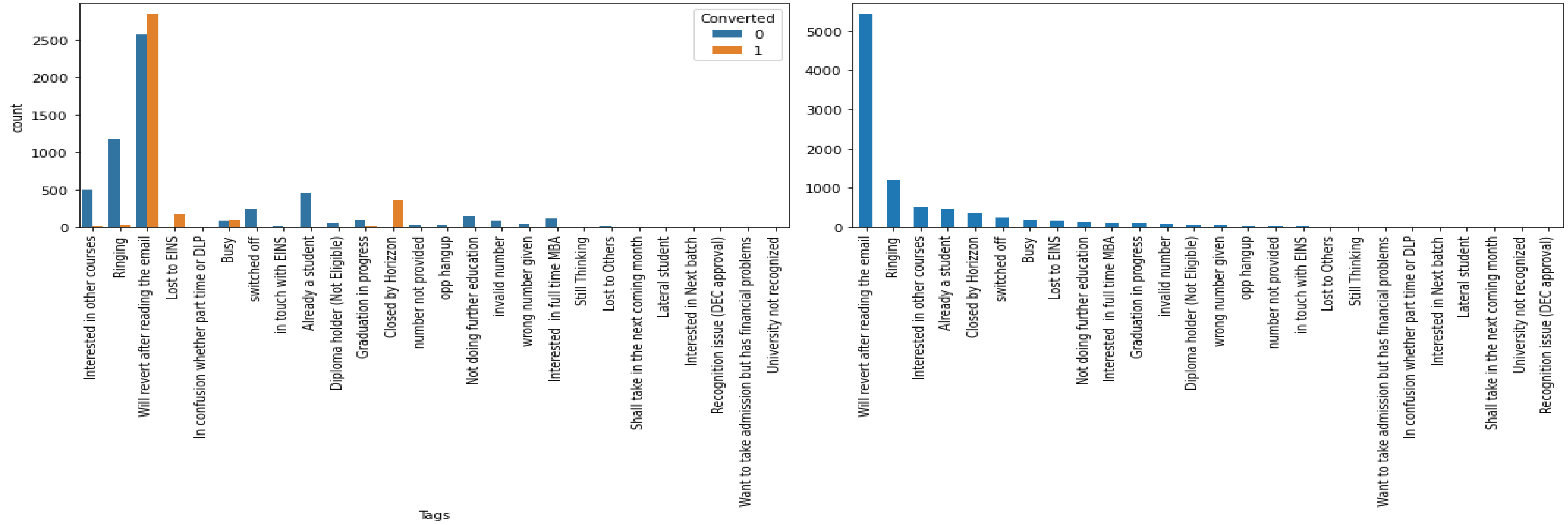


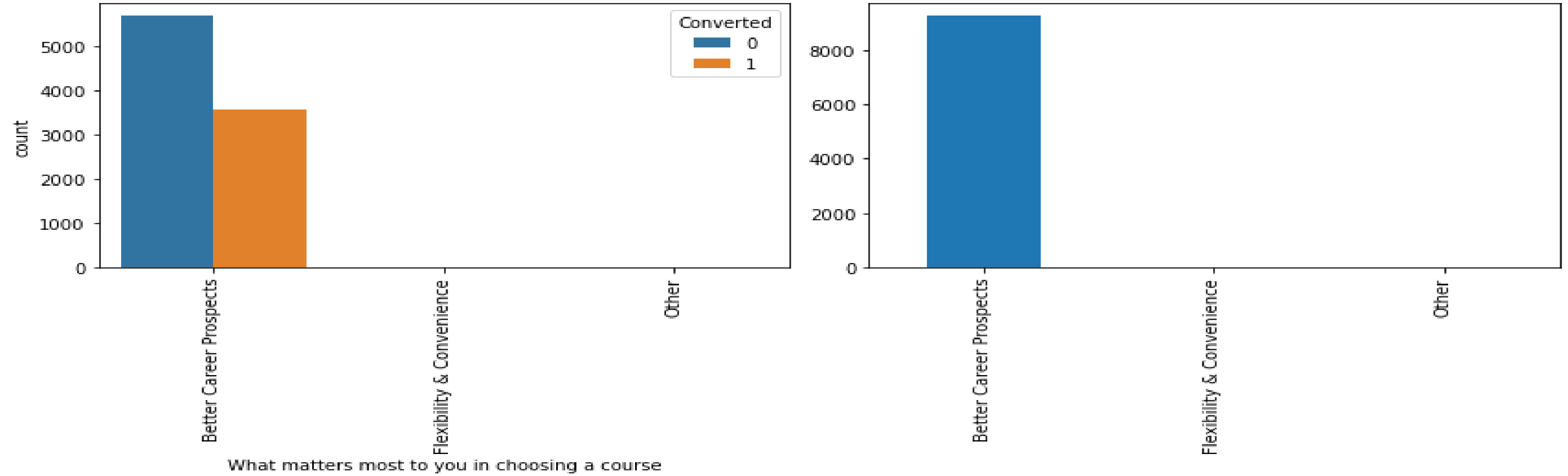


- **Most leads want a better career prospect for which they are considering joining the course.**
- **Maximum leads generated are unemployed and conversion rate of leads is also highest for those who are either 'unemployed' or 'working professionals'.**
- **Maximum leads generated and converted are from Mumbai city.**


Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity
3000
2500
2000
1500
1000
500
0
Modified
Email Opened
SMS Sent
Page Visited on Website
Olark Chat Conversation
Email Link Clicked
Email Bounced
Unsubscribed
Unreachable
Had a Phone Conversation
Email Marked Spam
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received


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
Others
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity
3500
3000
2500
2000
1500
1000
500
0
Email Opened
SMS Sent
Olark Chat Conversation
Page Visited on Website
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Others
Unreachable
Unsubscribed
Had a Phone Conversation
Approached upfront
View in browser link Clicked
Email Received
Email Marked Spam
Visited Booth in Tradeshow
Resubscribed to emails

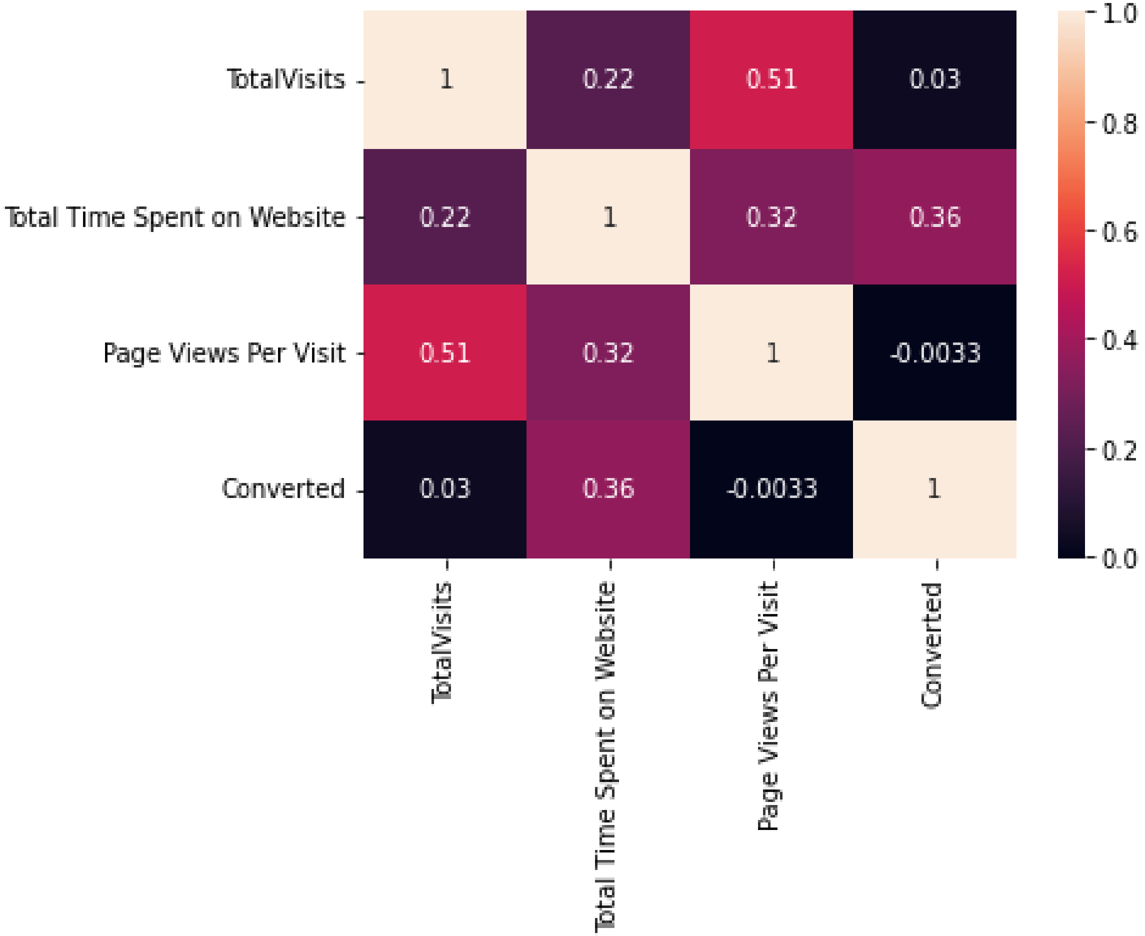


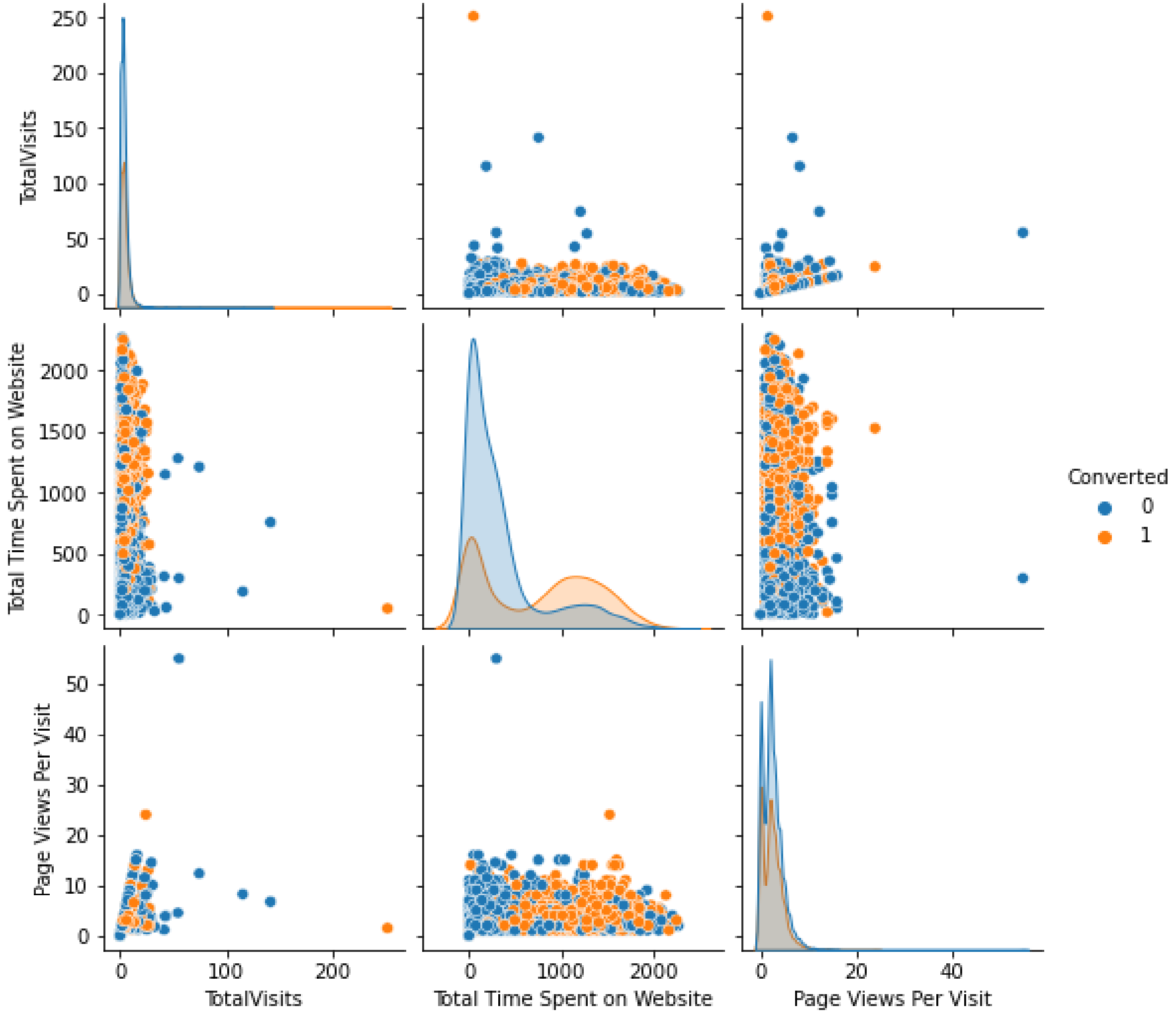


- **Most leads are generated when the selection in the Tags attribute reads "Will revert after reading the mail and also maximum conversion occur for those who would either revert after reading the mail or who were closed by horizzon.**

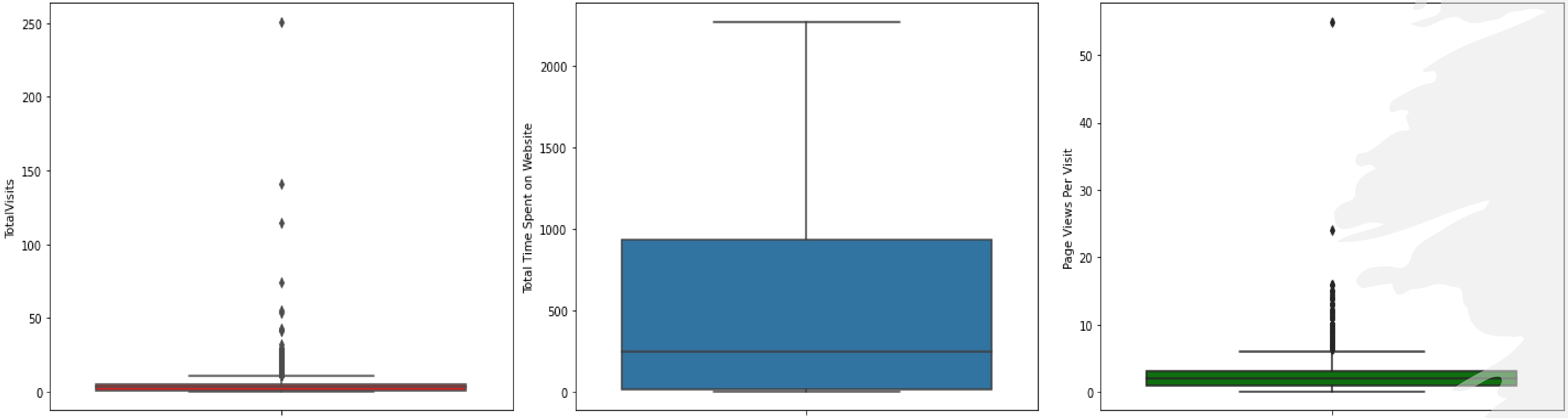
- **Most leads generated want a better career prospect and those who make this selection are highly likely to be converted.**

Numerical Variable Analysis:

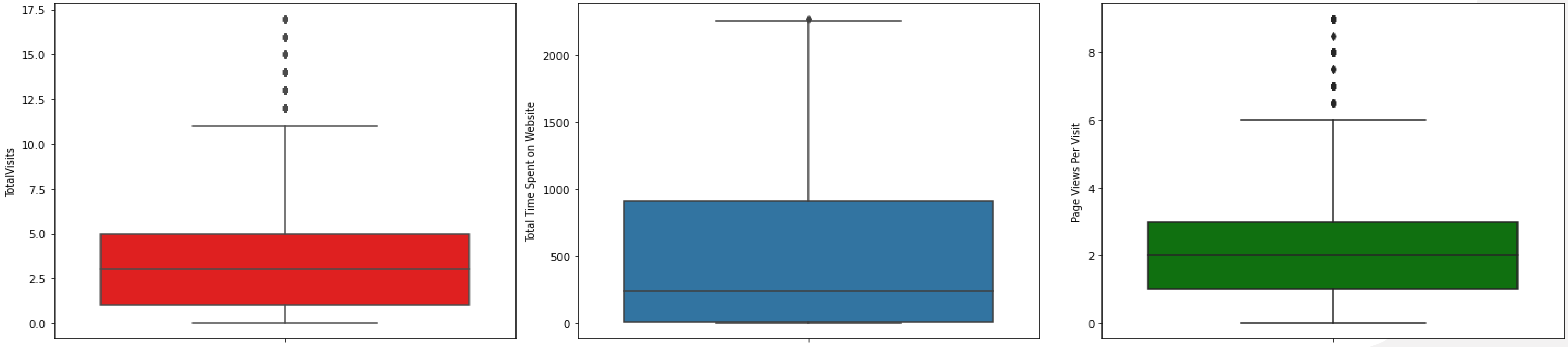


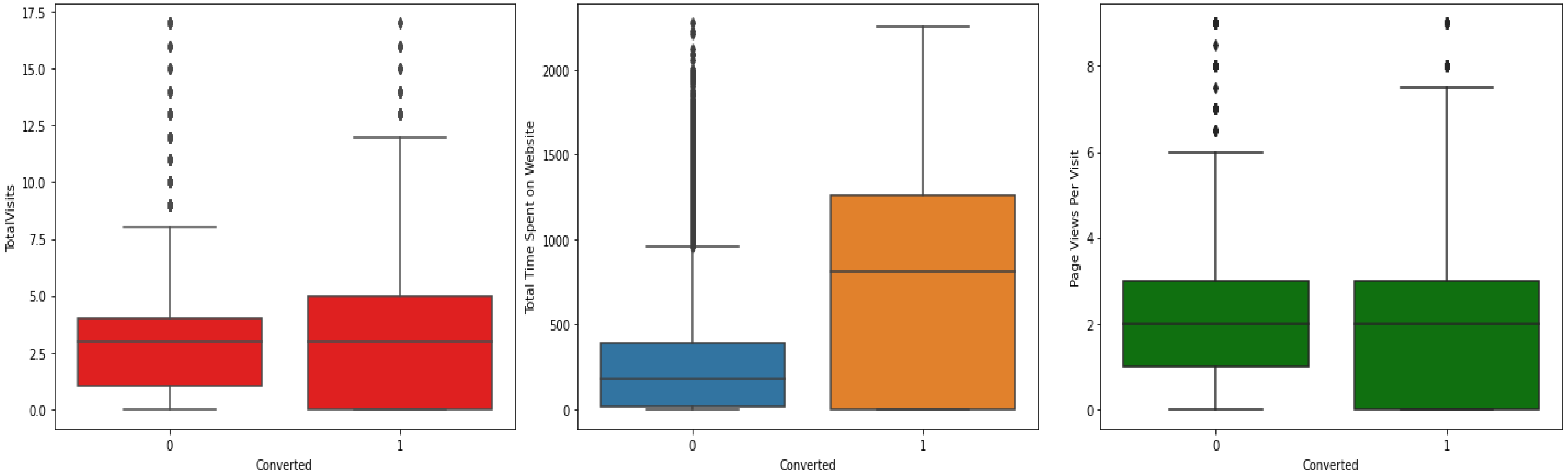


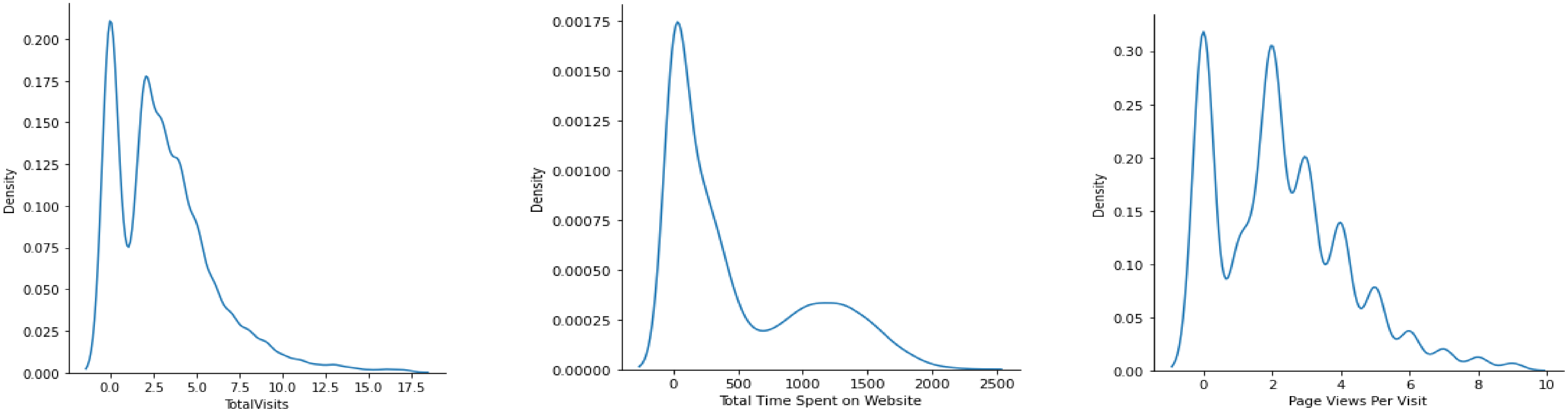
Before Outlier Treatment:



After Outlier Treatment:




TotalVisits
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
1
Converted
Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted
Page Views Per Visit
8
6
4
2
0
0
1
Converted


Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
TotalVisits
Density
0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
0
500
1000
1500
2000
2500
Total Time Spent on Website
Density
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
Page Views Per Visit

Data Modelling and Model Building

- Creating dummy variables for our categorical attributes.
- Splitting the data into train and test datasets to run our classification algorithm.
- Scaling the numerical attributes to ensure that the dataset becomes unitless.
- Important attributes were selected using Recursive Feature Elimination(RFE).
- The following attributes were selected for our algorithm:

  Total Time Spent on Website

  Origin_Lead Add Form

  lastactivity_Email Bounced

  lastactivity_Olark Chat Conversation

  Occupation_Working Professional

  Tags_Busy

  Tags_Closed by Horizzon

  Tags_Lost to EINS

  Tags_Ringing

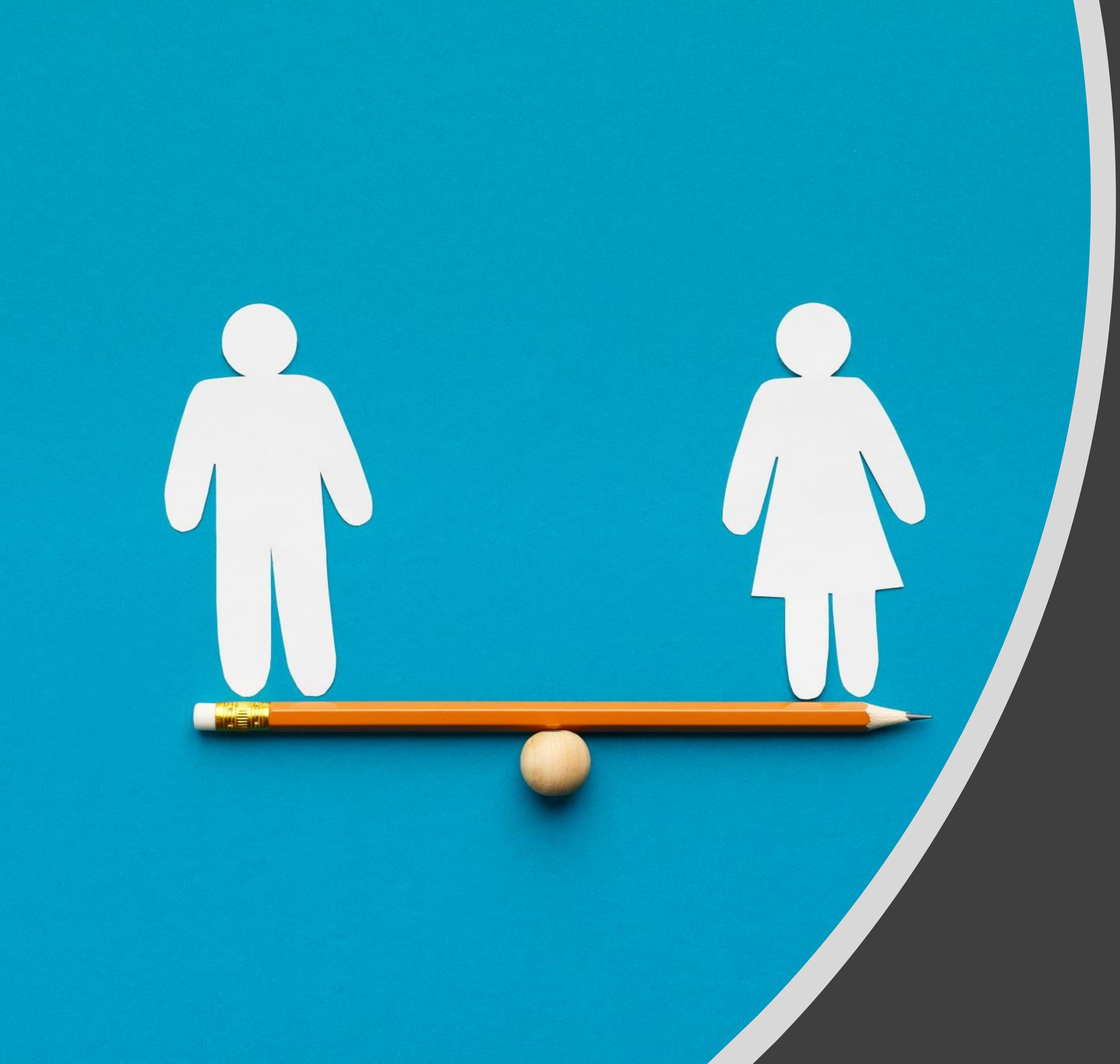
  Tags_Will revert after reading the email

  Tags_switched off

  Last_Notable_Activity_Email Bounced

  Last_Notable_Activity_SMS Sent

- In three iterations of our classification algorithm , we achieved a model with no statistically insignificant attributes and multicollinearity.



Model Performance And Evaluation

Training Dataset(decision boundary = 0.5)

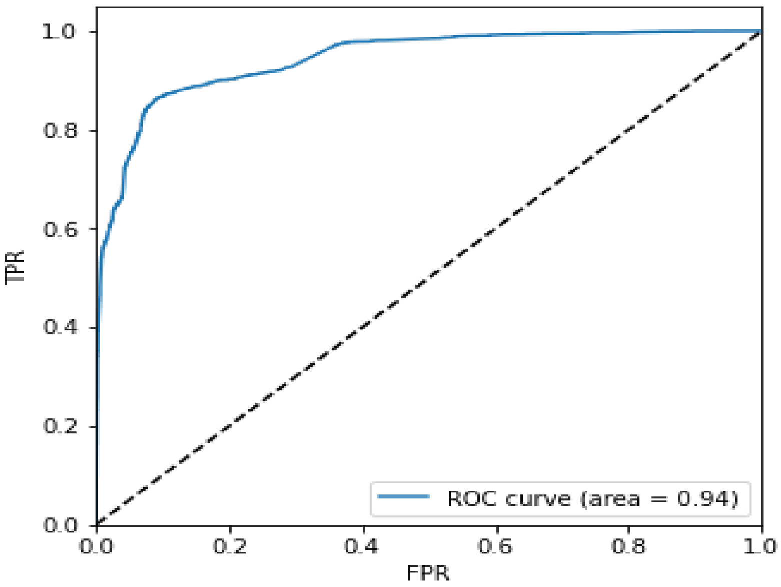
accuracy score: 0.89
precision: 0.87
recall: 0.84
sensitivity: 0.84
specificity: 0.93
Positive_prediction_rate: 0.87
negative_prediction_rate: 0.90
False_positive_rate: 0.07

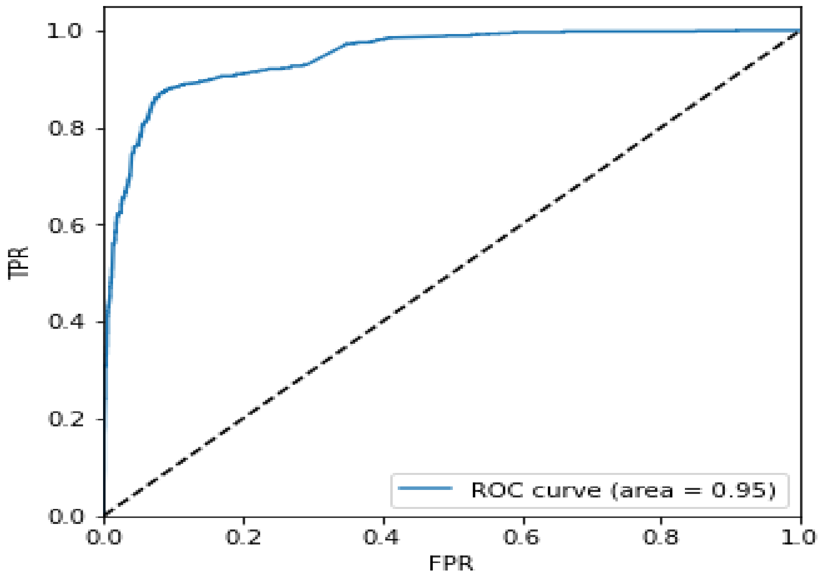
Training Dataset(Decision boundary = 0.3)

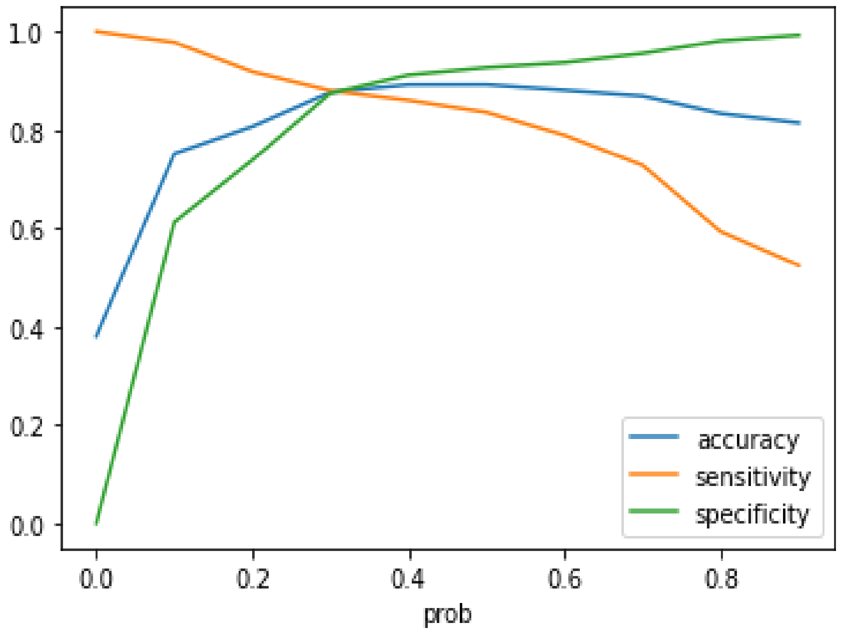
accuracy score: 0.88
precision: 0.81
recall: 0.88
sensitivity: 0.88
specificity: 0.87
Positive_prediction_rate: 0.81
negative_prediction_rate: 0.92
False_positive_rate: 0.13

Test Dataset

accuracy score: 0.88
precision: 0.81
recall: 0.89
sensitivity: 0.89
specificity: 0.87
Positive_prediction_power:0.81
negative_prediction_power:0.93
False_positive_rate: 0.13







The AUC (Area Under Curve) value achieved was 0.94 on the training dataset and we want this to be as close to 1 as possible. The figure on the right signifies the Sensitivity-Specificity tradeoff and the interesection of the three lines gives us the optimal threshold or decision boundary value.

Conclusion:

- After inspecting the Sensitivity-Specificity Tradeoff, the threshold or decision boundary value of 0.3 was selected to make predictions on the test dataset.
- The top attributes(in order of importance) that impacted the possibility of a lead getting converted are as follows:
  - Tags : Closed By Horizzon
  - Tags: Lost to EINS
  - Lead Origin: Lead Add Form
- The sensitivity and specificity score of our model is 89% and 87% respectively and our model seems to predict the conversion rate really well and can give the CEO confidence in making strategic decisions to maximize profits.