LEAD SCORING CASE STUDY

By:
DSC-41

Swapnil Kudale

and

Thulasiram Saravanan

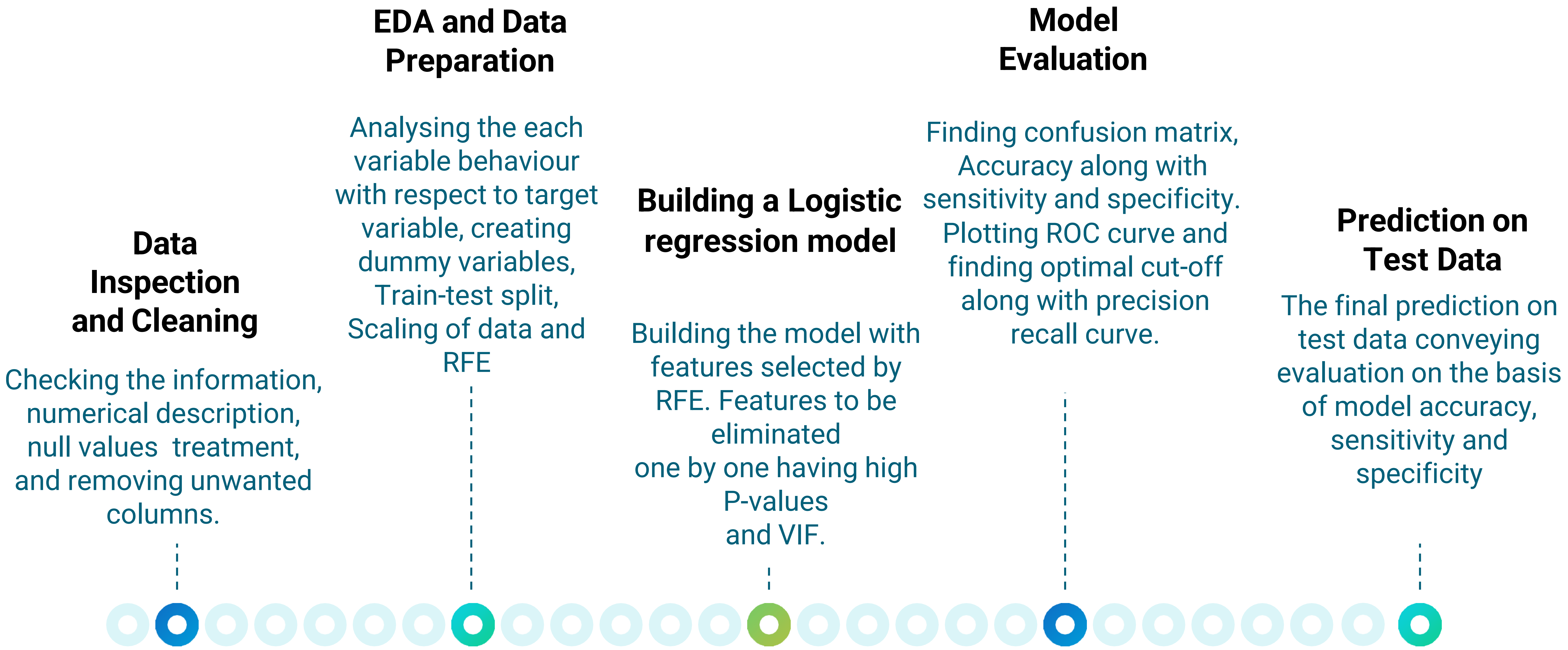
Problem Statement

- X education sells online courses to Industry Professionals.
- X education gets a lots of leads, its lead conversion rate it very poor. For example, if 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.

Business Objective

- Help X education to select the Most Promising Leads (Hot Leads)
- Build a Logistic regression model to assign a lead score value between 0 to 100 to each of the leads which can be used by the company to target Potential Leads

Analysis Approach



Data Inspection and Cleaning

Checking the information, numerical description, null values treatment, and removing unwanted columns.

EDA and Data Preparation

Analysing the each variable behaviour with respect to target variable, creating dummy variables, Train-test split, Scaling of data and RFE

Building a Logistic regression model

Building the model with features selected by RFE. Features to be eliminated
one by one having high P-values
and VIF.

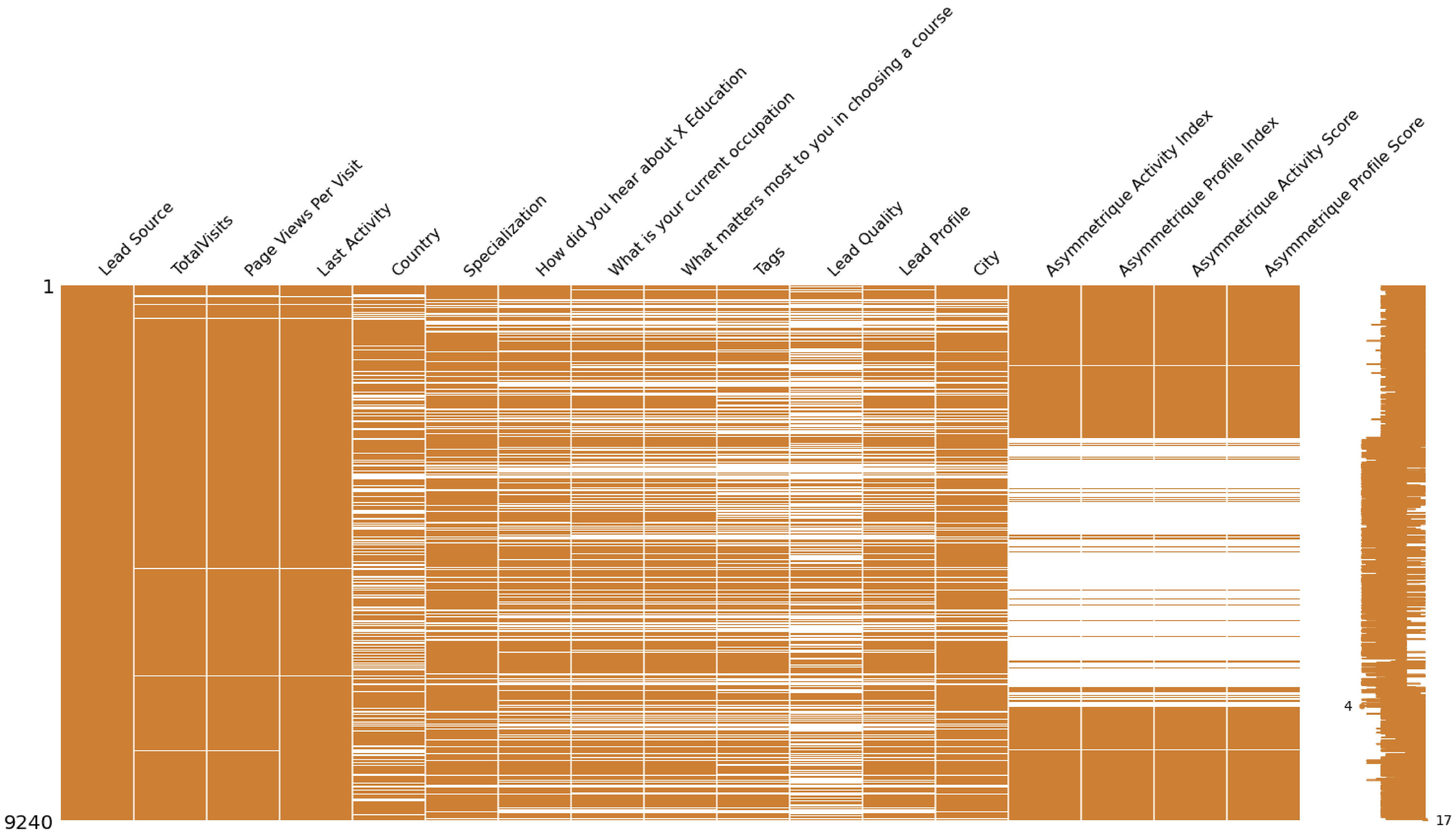
Model Evaluation

Finding confusion matrix, Accuracy along with sensitivity and specificity. Plotting ROC curve and finding optimal cut-off along with precision recall curve.

Prediction on Test Data

The final prediction on test data conveying evaluation on the basis of model accuracy, sensitivity and specificity

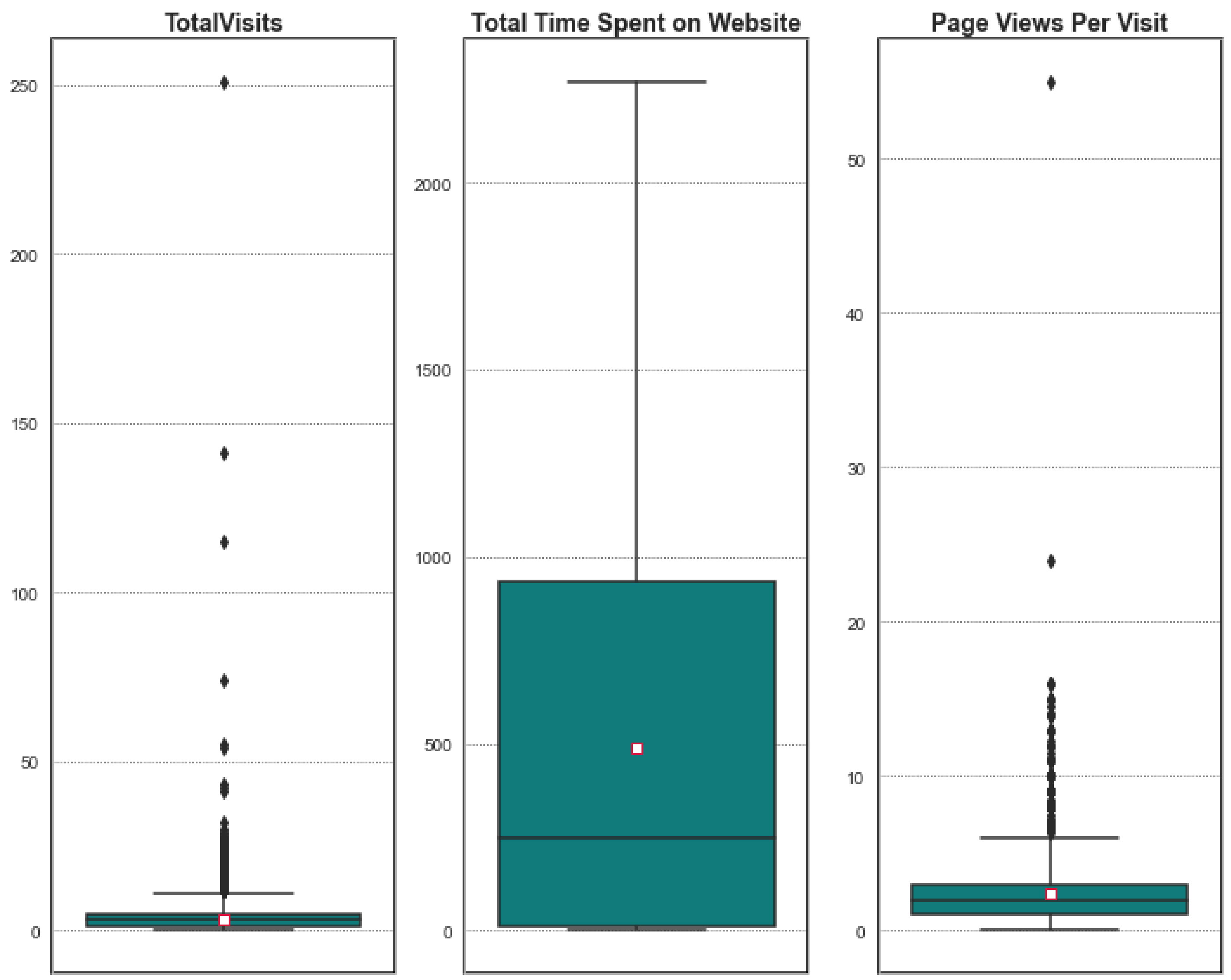
Visualizing Missing values



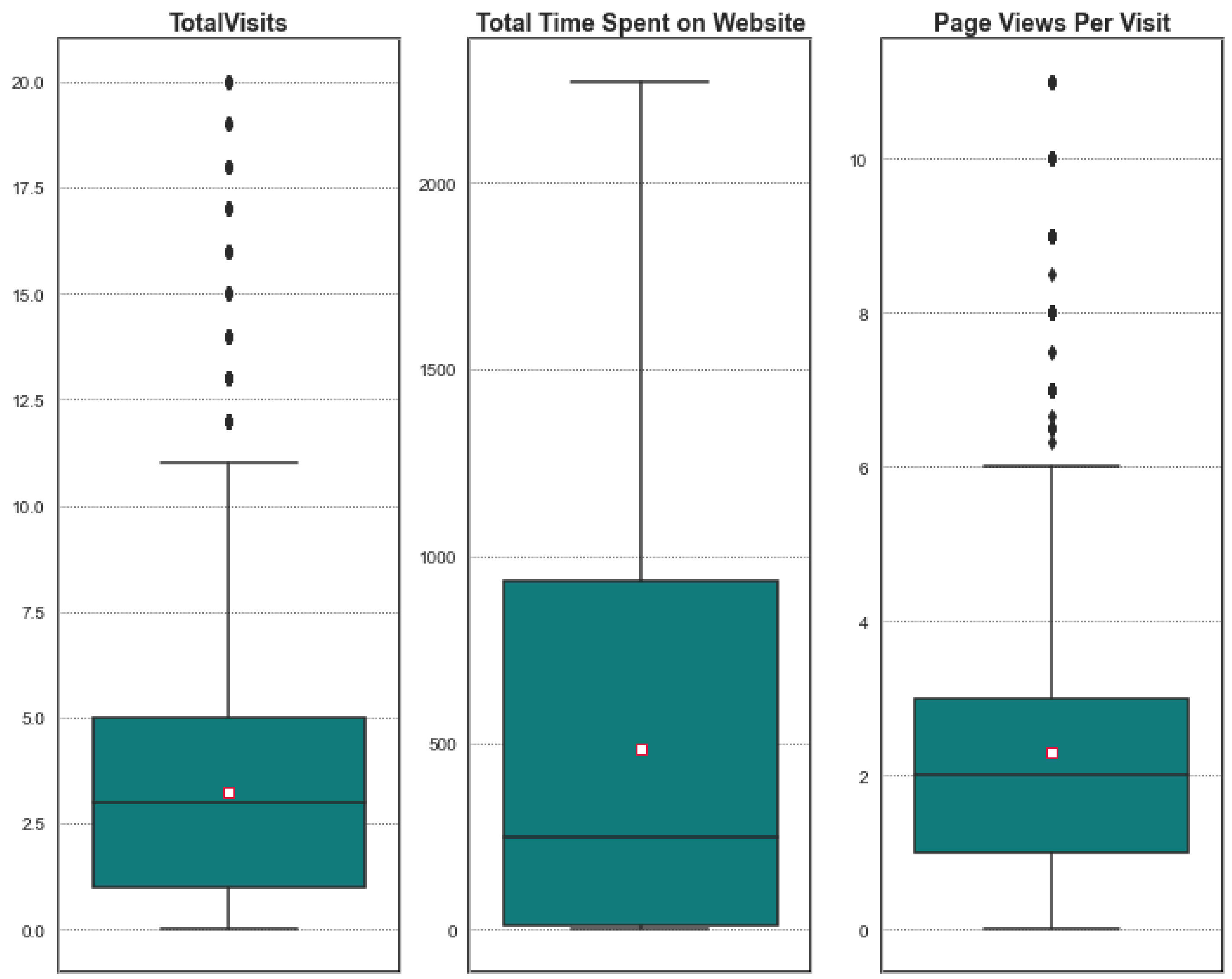
There are plenty of missing values in the last four columns
As per the plot.

These columns have been eliminated.

Outlier Treatment



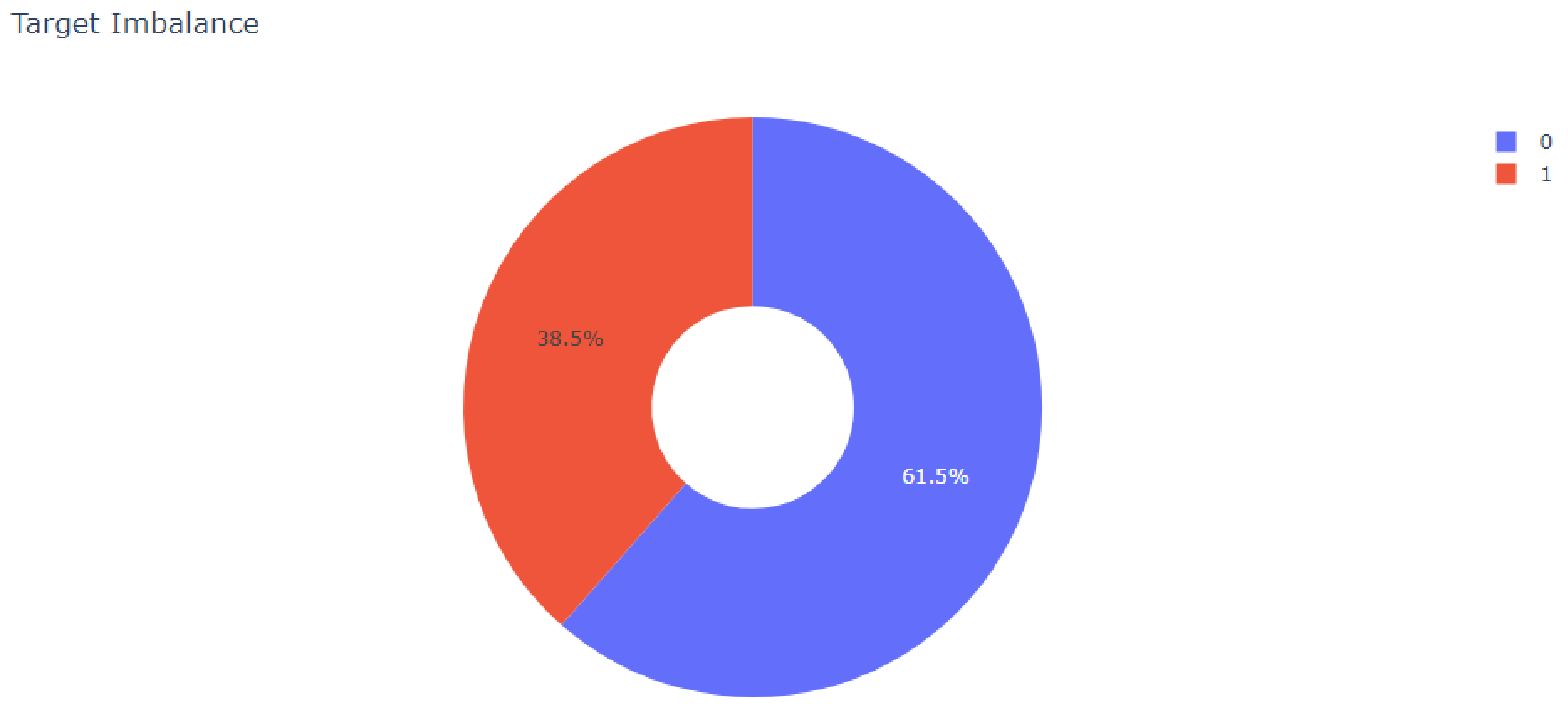
There are outliers in total visits and Pages views per visit and we need to eliminate them.



After eliminating and retaining 0.995 quantile of data we get a decent box plot with very few outlier.

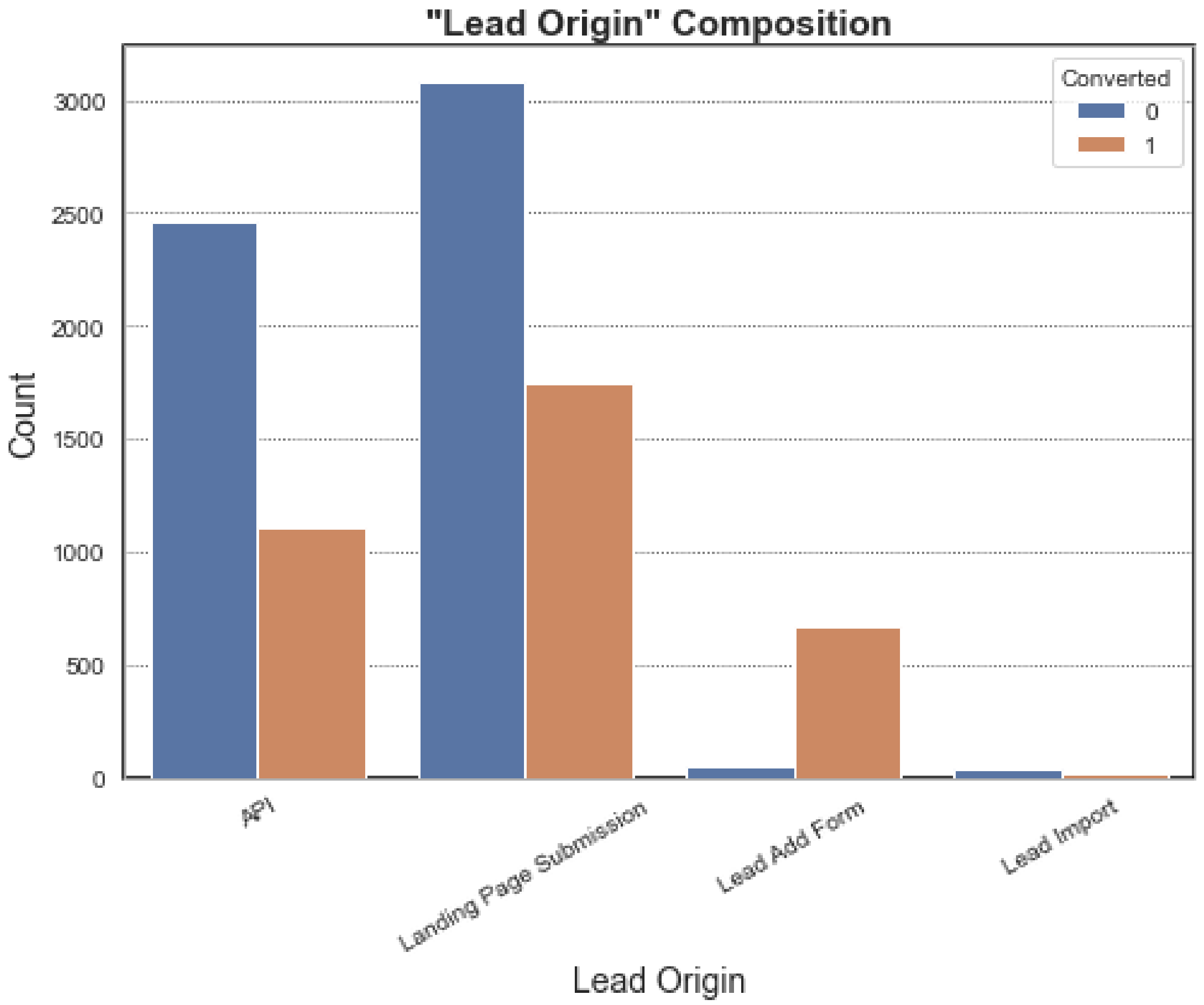
Exploratory Data Analysis

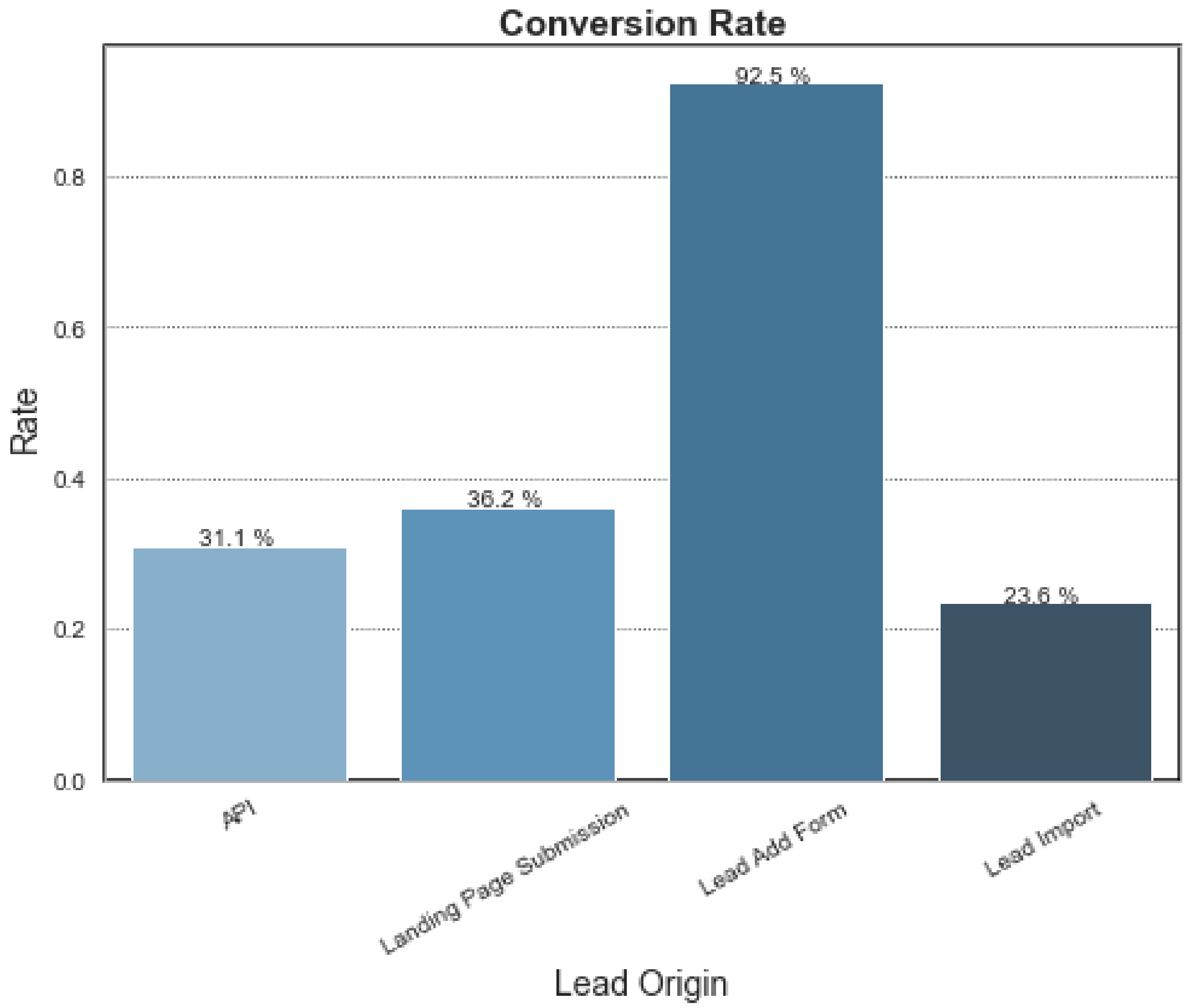
Balance Ratio Analysis of Target Variable



There is a little target imbalance of the converted vs non converted. But that is as per the problem statement that only 30% are converted and hence we can consider this as a valid case.

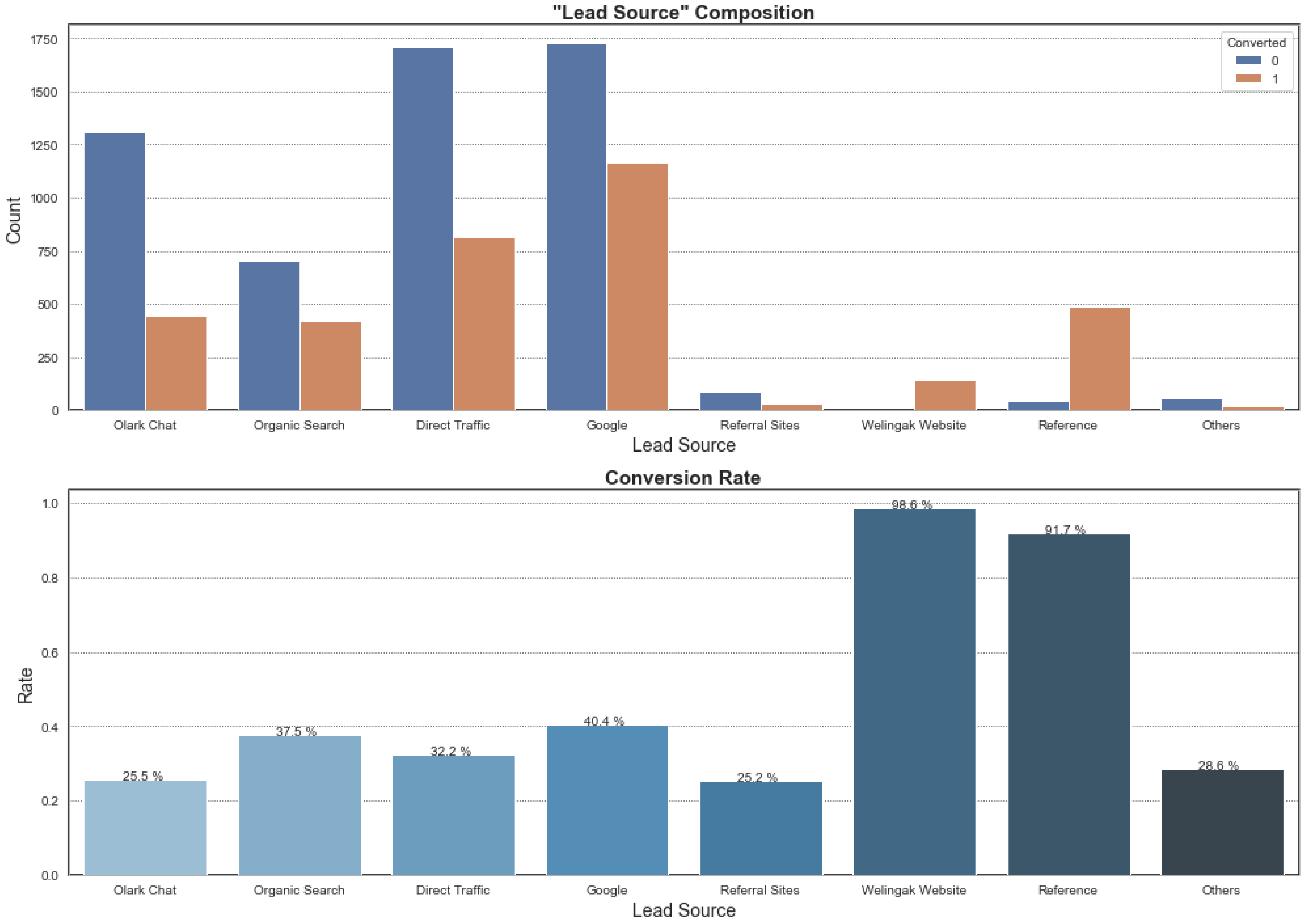
Analysis - Lead Origin


"Lead Origin" Composition


Conversion Rate

- **Observation:**
- **The majority of the leads came from submissions on the landing page, followed by API, where approximately 30% are converted.**
- **Leads from the Lead Add Form have the highest conversion rate in this category, accounting for approximately 90%.**
- **Lead imports are few in number, and the conversion rate is also low.**
- **To increase overall lead conversion rates, we must concentrate on improving lead conversion from API and Landing Page Submission origins.**
- **Even though, Lead Add Form identifies brings in less leads but the conversion rate of the leads identified by the it is very high. Company should try to bring in more leads by this method.**

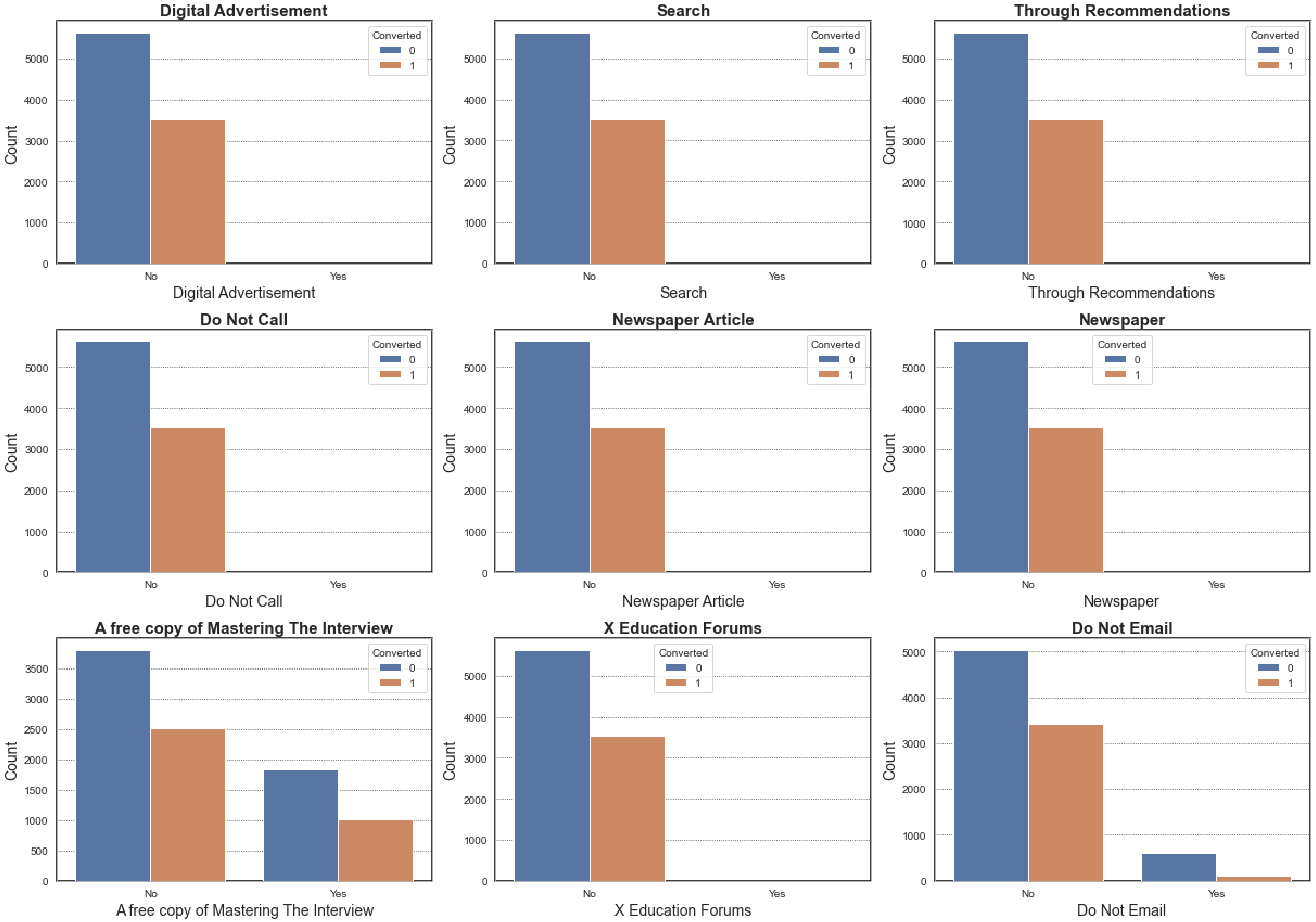
Analysis - Lead Source



Observation:

- **Most number of leads come from Google and Direct Traffic. Conversion rate of leads from direct traffic is less than overall conversion rate and the same for Google is slightly more than overall average.**
- **A very high percentage of leads from welingak website and References have converted. The company should invest more resources into acquiring leads from these sources**

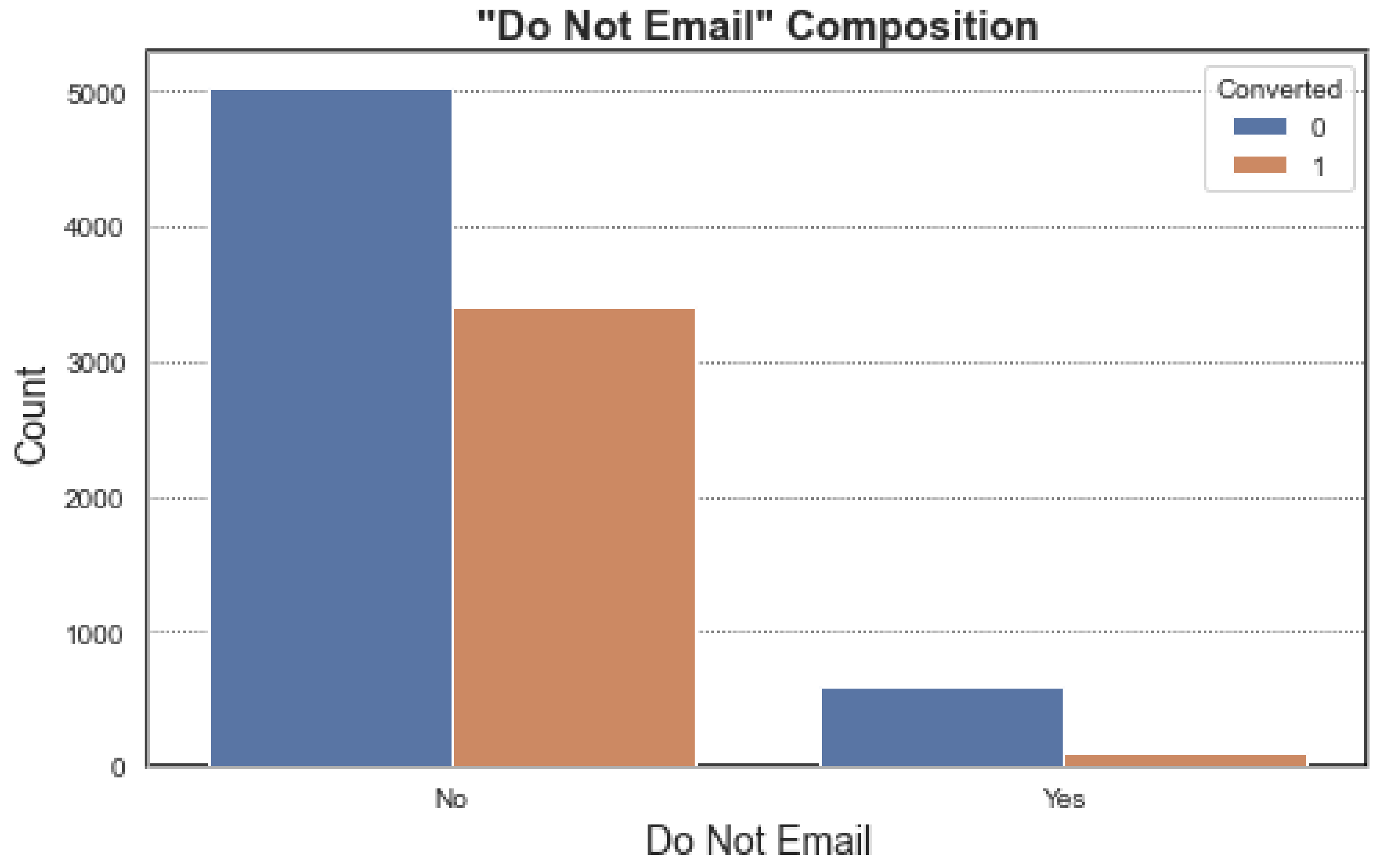
Visualizing Binary categorical variables

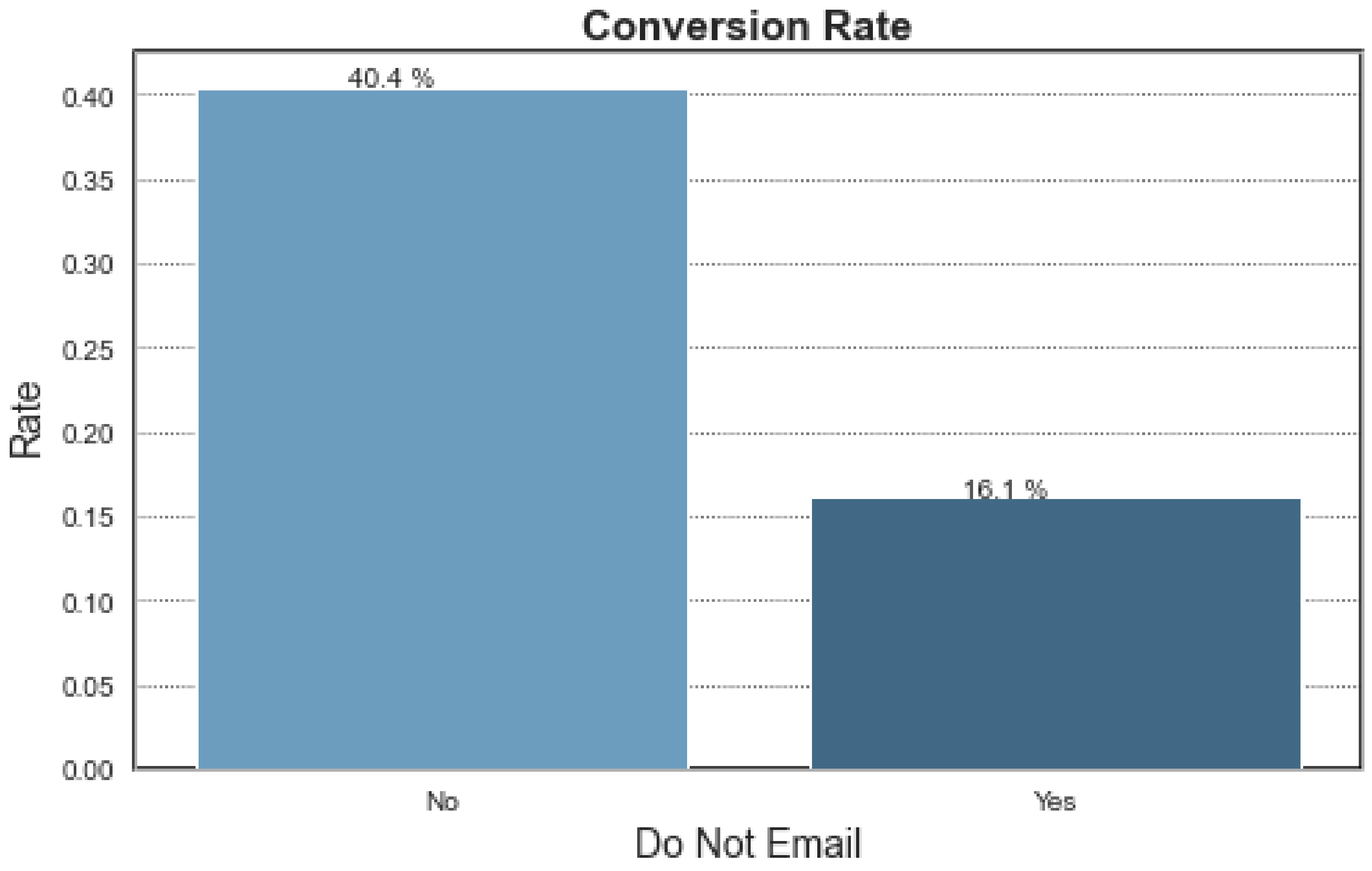


Observation:

- **It looks like for many of the variable here, one of the level is highly dominant and the other has almost no contribution.**

Analysis - Do Not Email

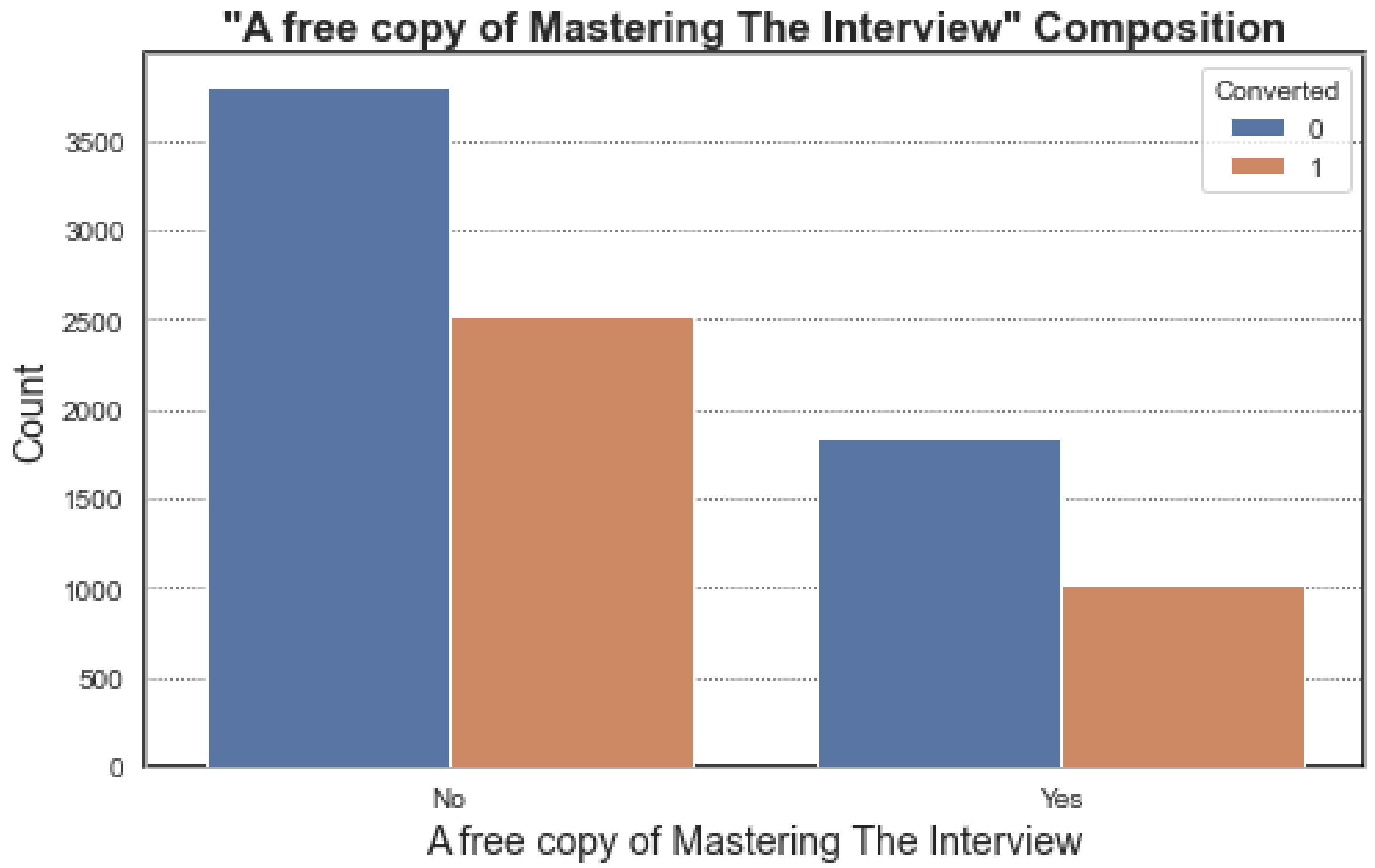




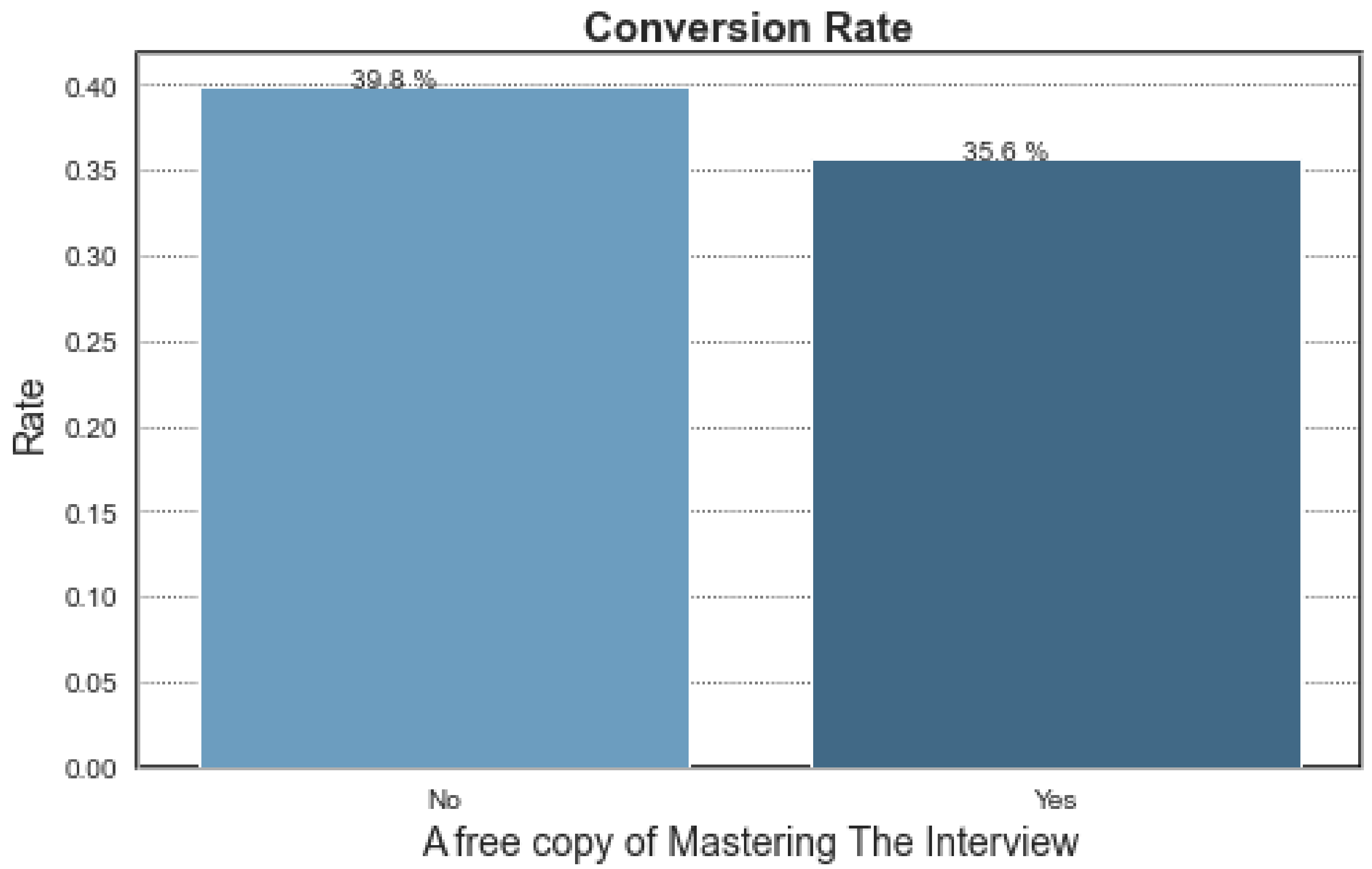
Observation:

- **Majority of the people want Email (~80%)**
- **People who have opted to receive Email has higher rate of conversion (40%)**

Analysis - A free copy of Mastering The Interview



"A free copy of Mastering The Interview" Composition

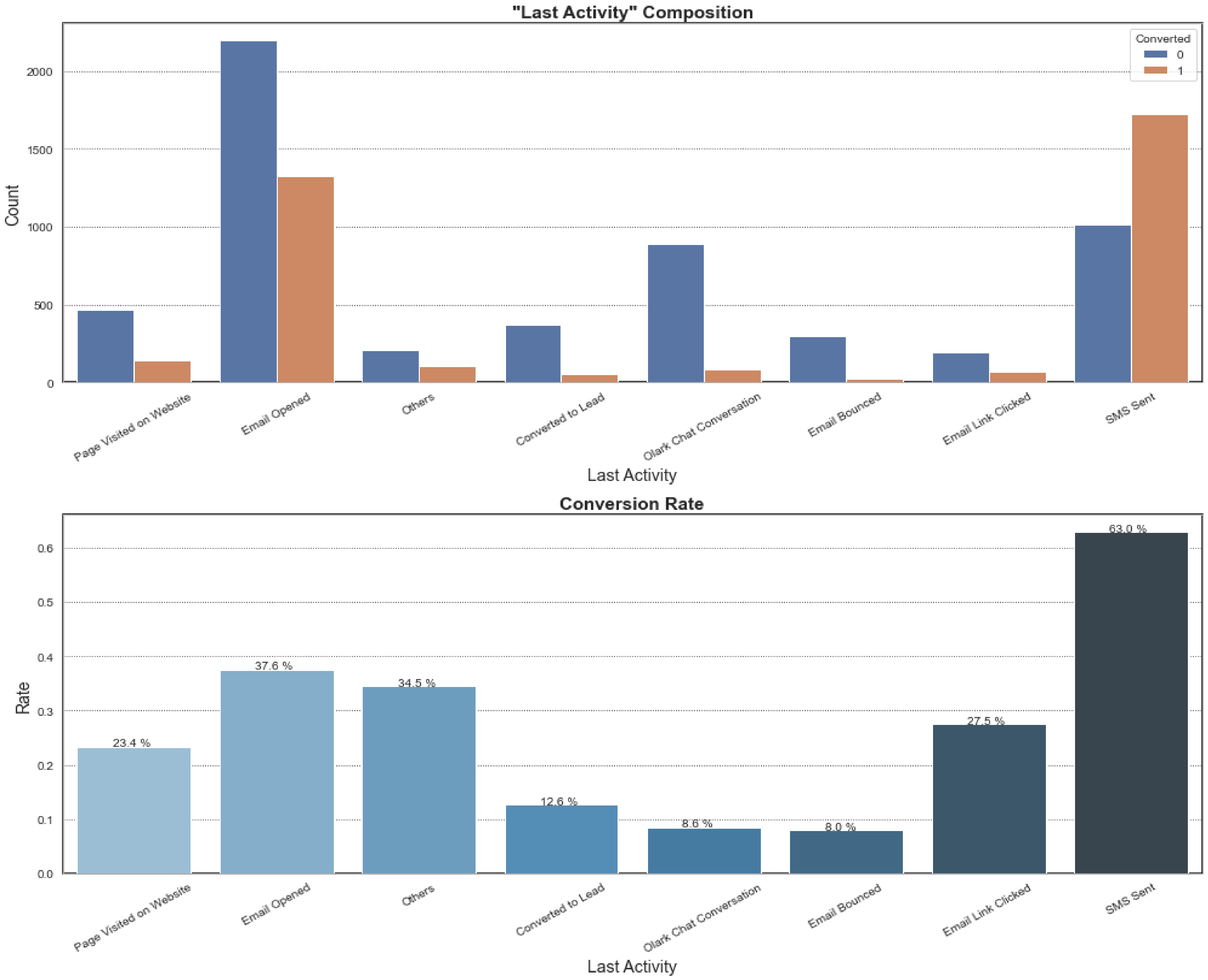


Conversion Rate

Observation:

- **Distributing Free-Copy of Mastering Interview doesn't seem affect the conversion as the conversion rate is almost same.**

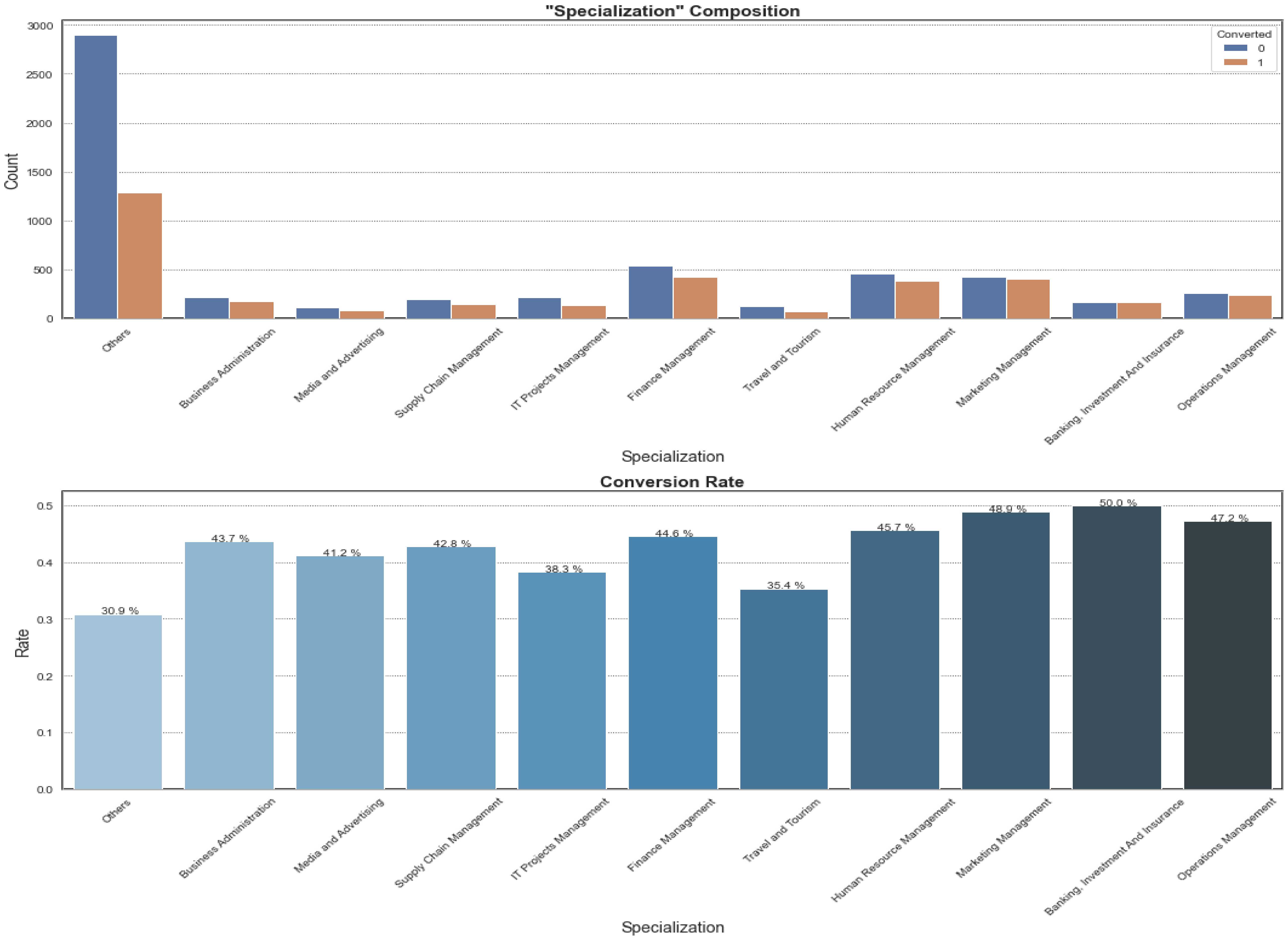
Analysis - Last Activity



Observation:

- **As their most recent activity, the majority of the leads have opened their email.**
- **Lead conversion is very high (70 percent) after combining smaller Last Activity types as Other Activities.**
- **The conversion rate for leads with the most recent activity as SMS Sent is nearly 60%.**

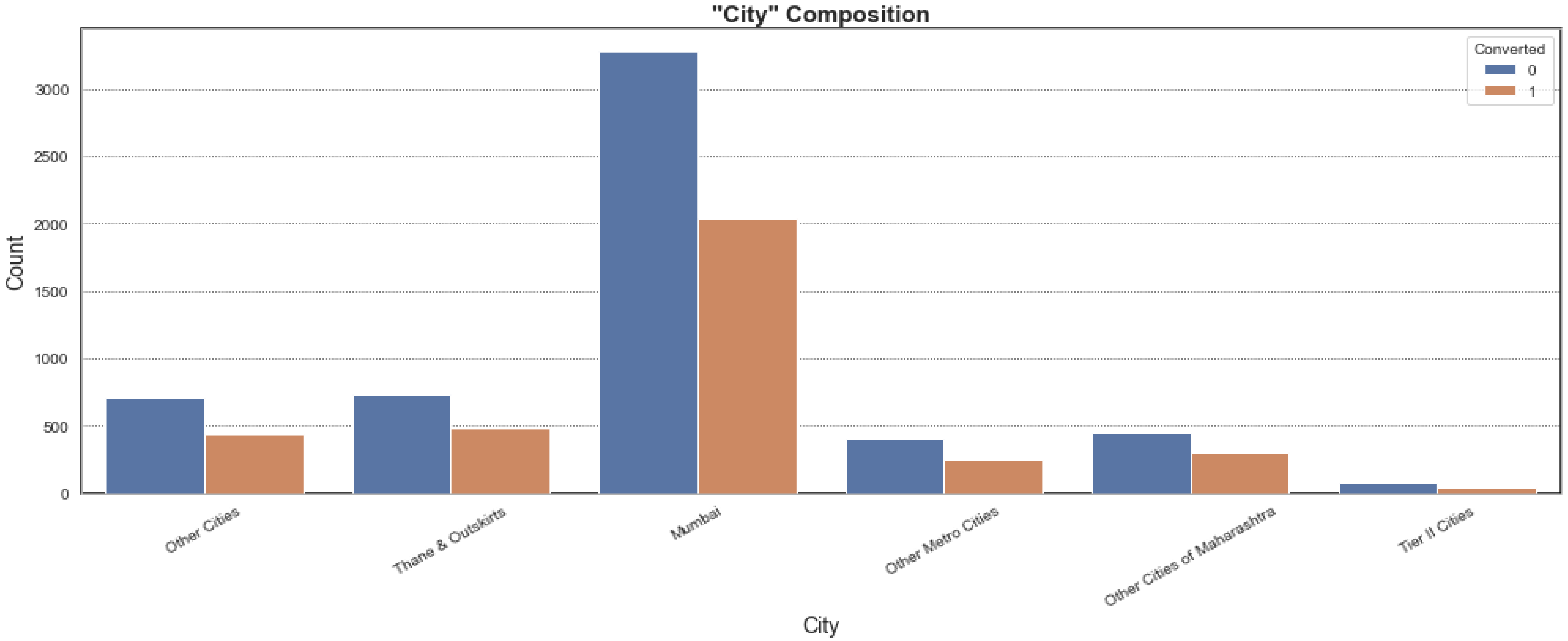
Analysis - Specialization

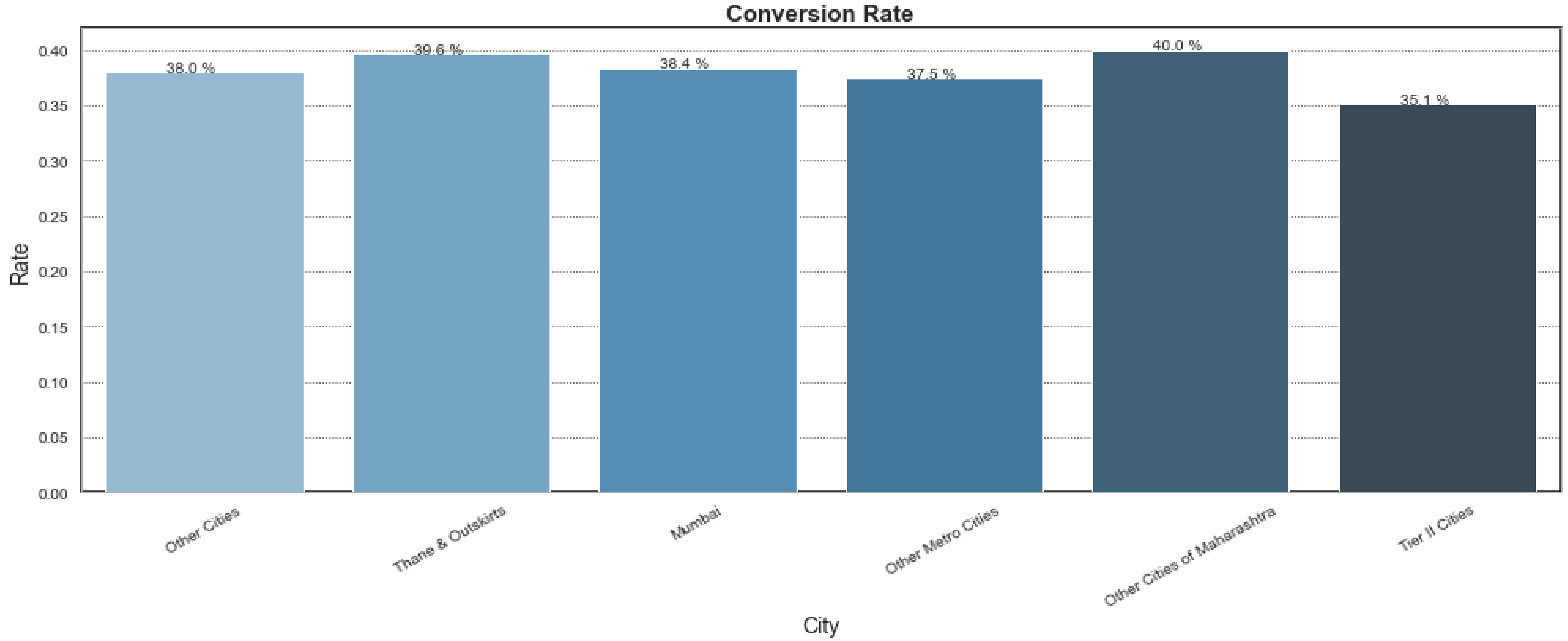


Observation:

- **Most of the leads have not mentioned a specialization and around 28% of those converted**
- **Leads with Banking Investment and insurance and Marketing Management - Over 45% Converted**

Analysis - City





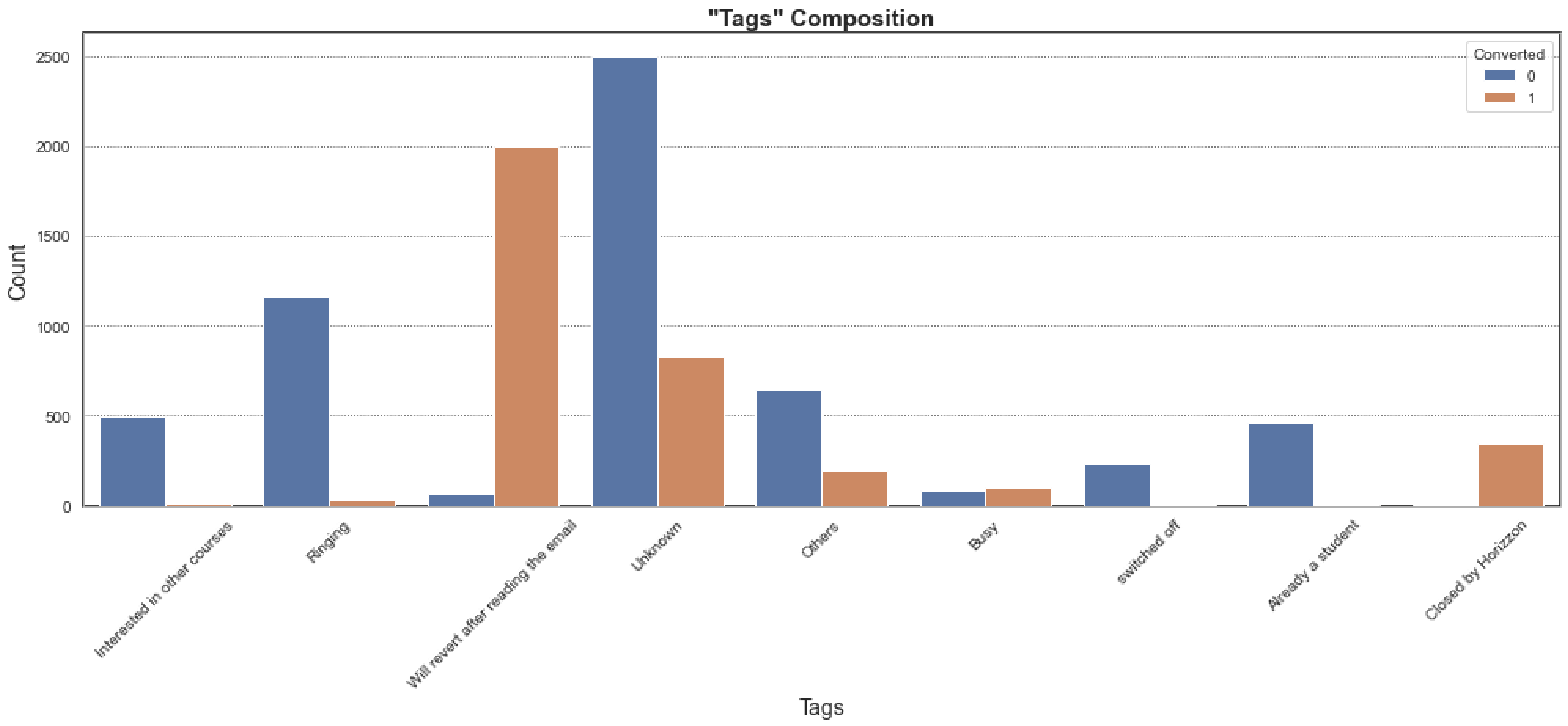
Observation:

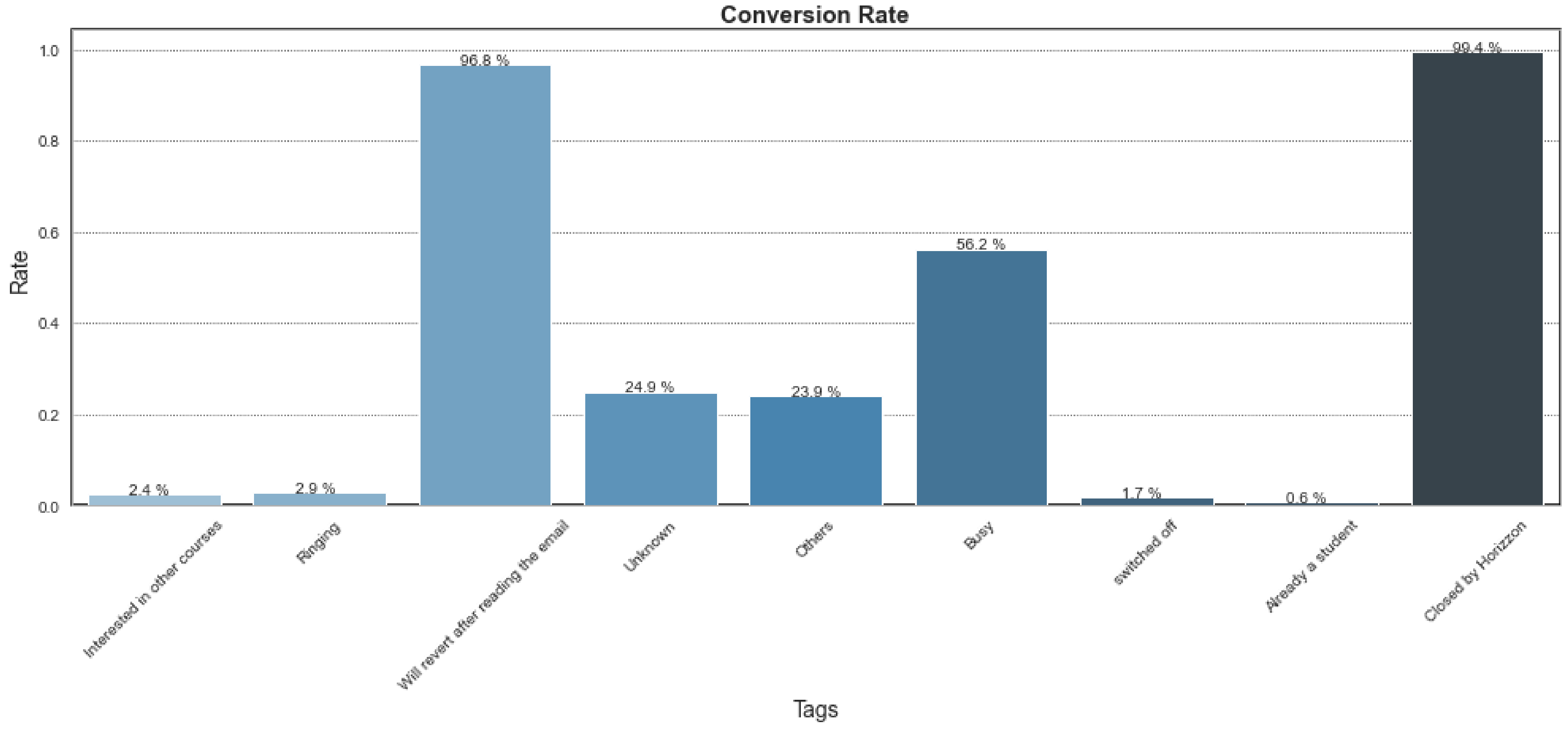
- **A huge proportion of leads acquired are from Mumbai. Conversion rates for all the cities is close to the overall average, 38.5 %**

Analysis of Tags

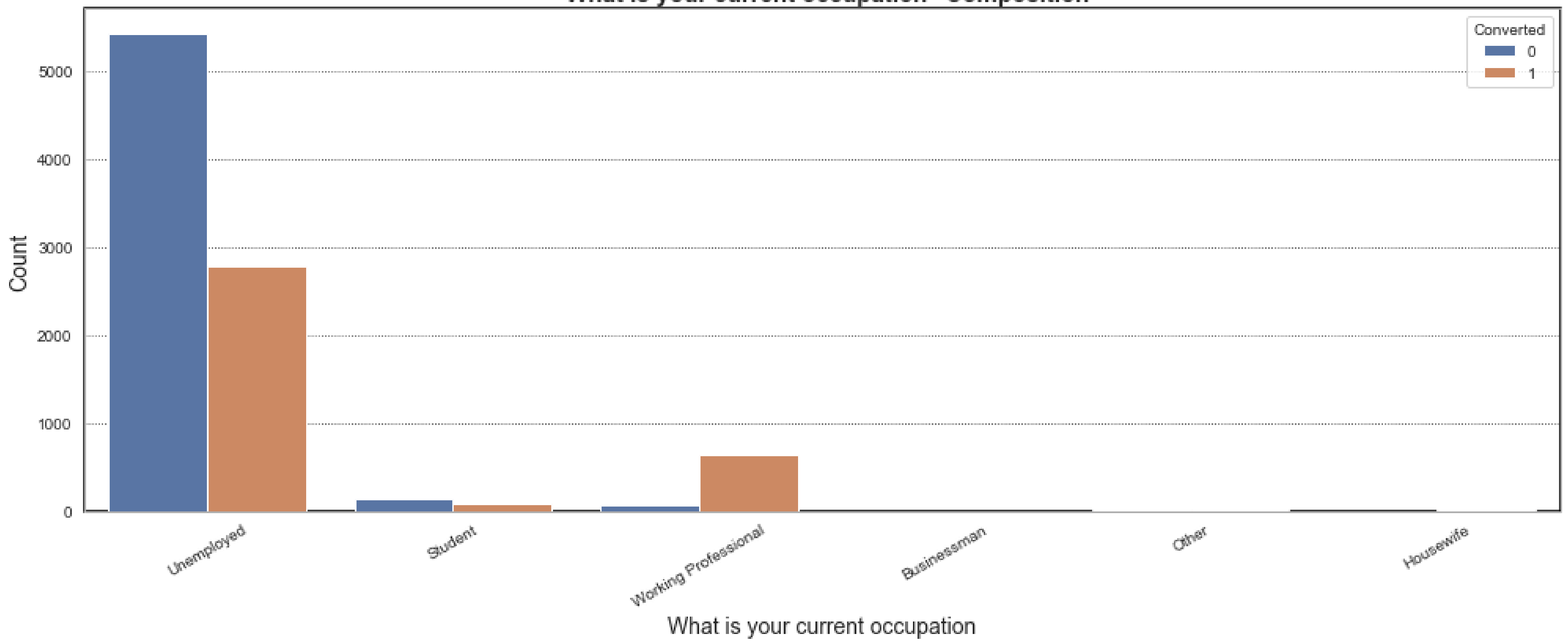
Observation:

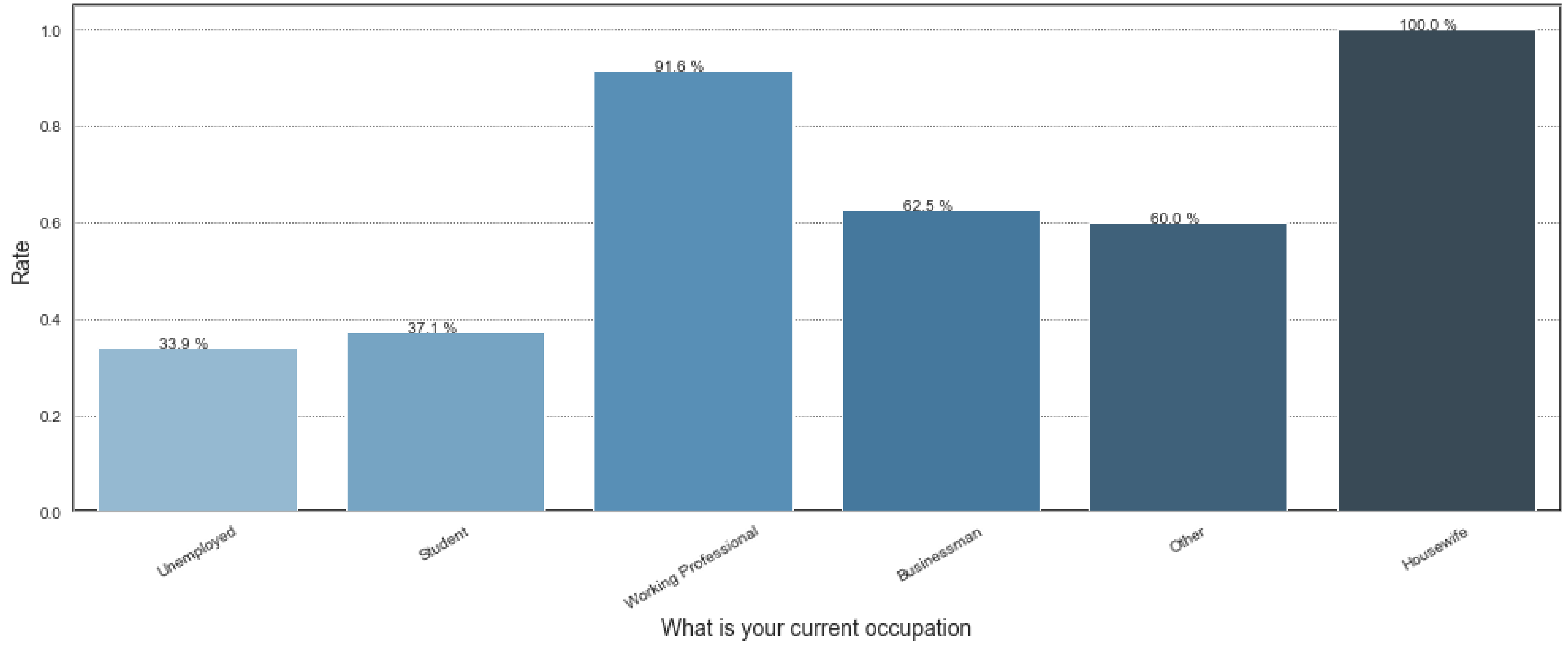
- Leads with tags/current status, 'Will revert after reading the email' have a very high likelihood of converting. This group has high potential leads.
- People with tags, 'Already a Student', 'Interested in other courses', 'Ringing' have very low conversion rate. The company should spend less resources on people in this group.





Analysis - Current Occupation

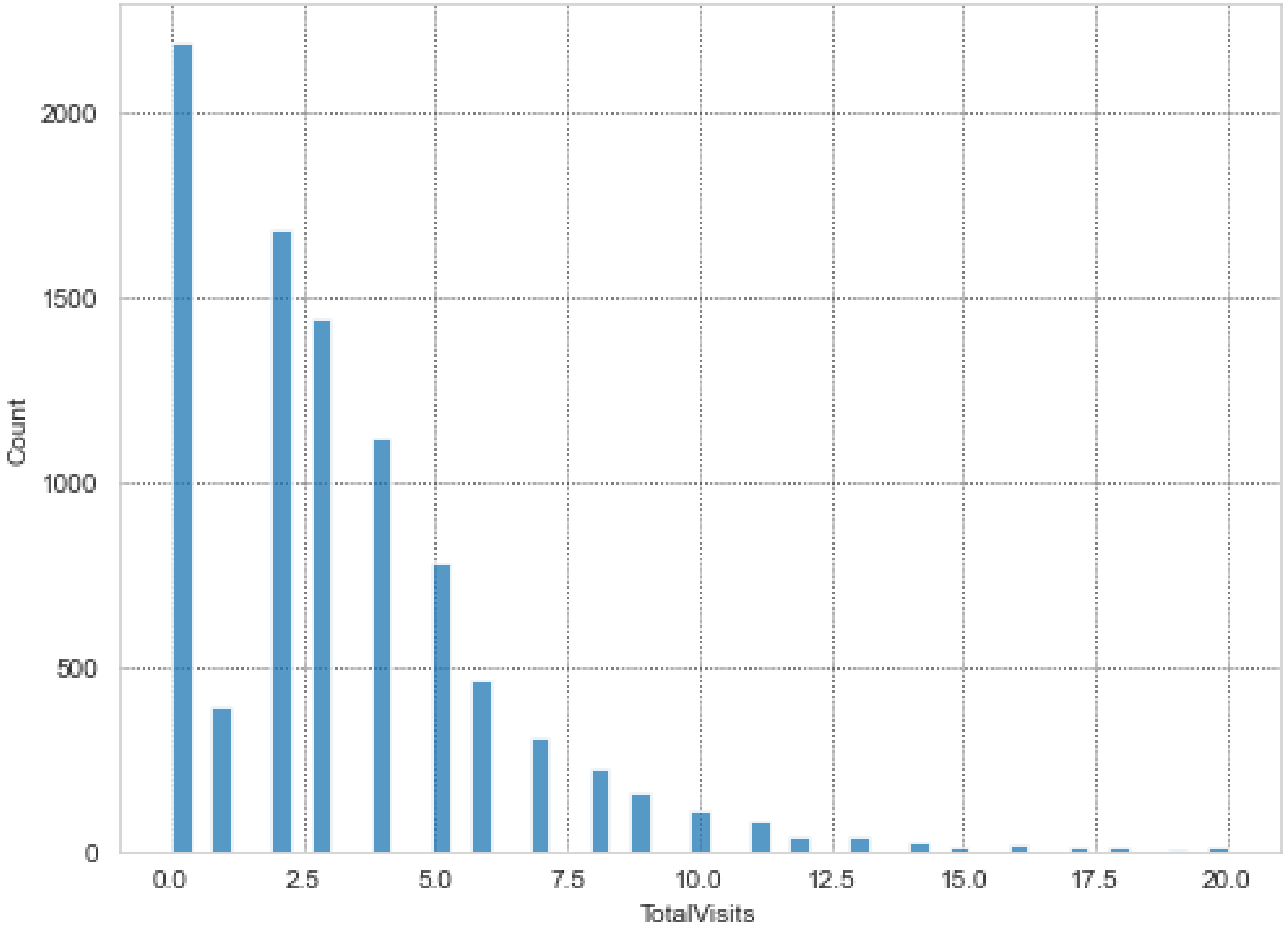

"What is your current occupation" Composition

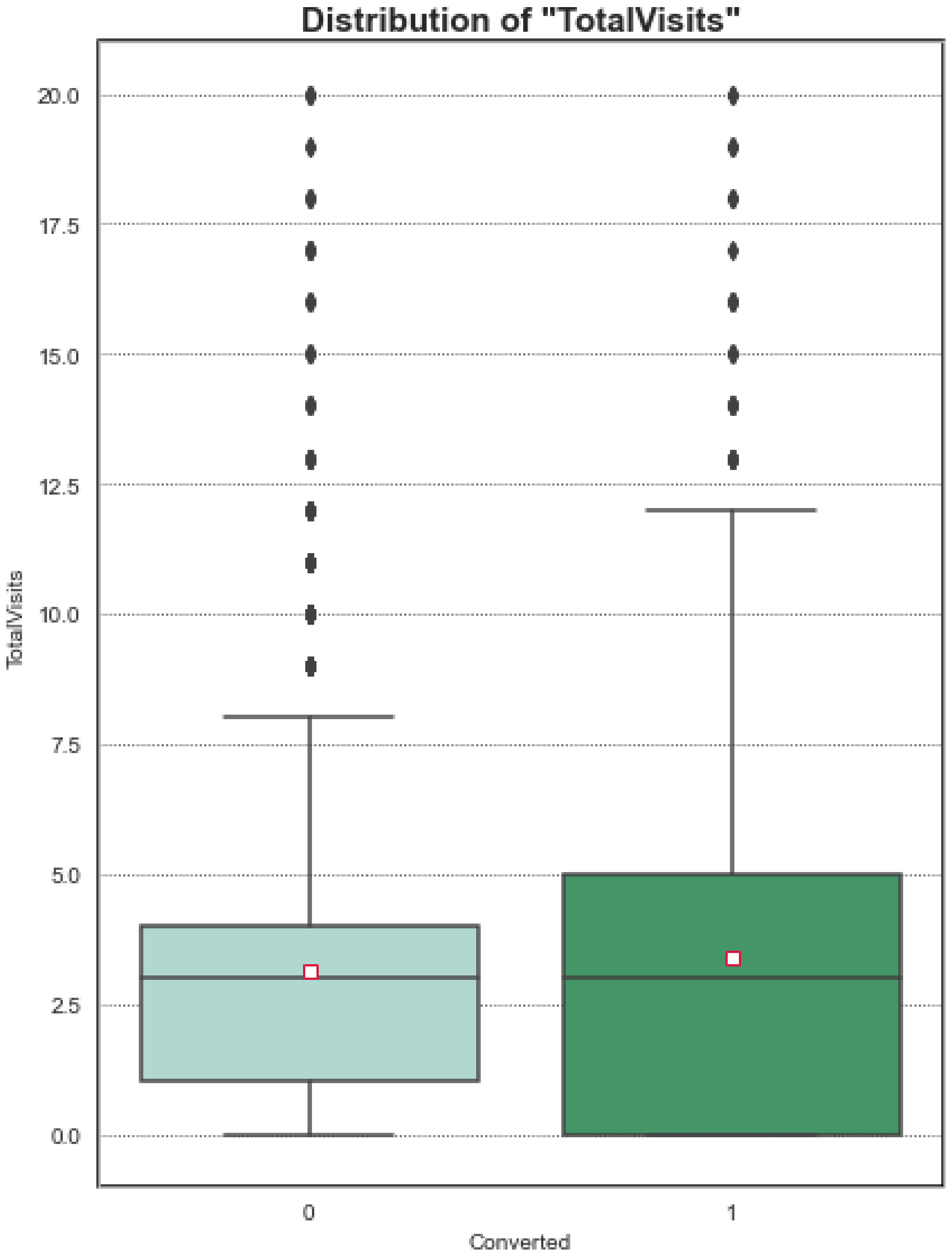

Conversion Rate

Observation:

- **Housewives are less in numbers, but have 100% conversion rate.**
- **Working professionals, Businessmen and Other have high conversion rate.**
- **Leads with Unemployed occupation is highest in number, but the conversion rate is low (~40%).**

Analysis – Total Visits

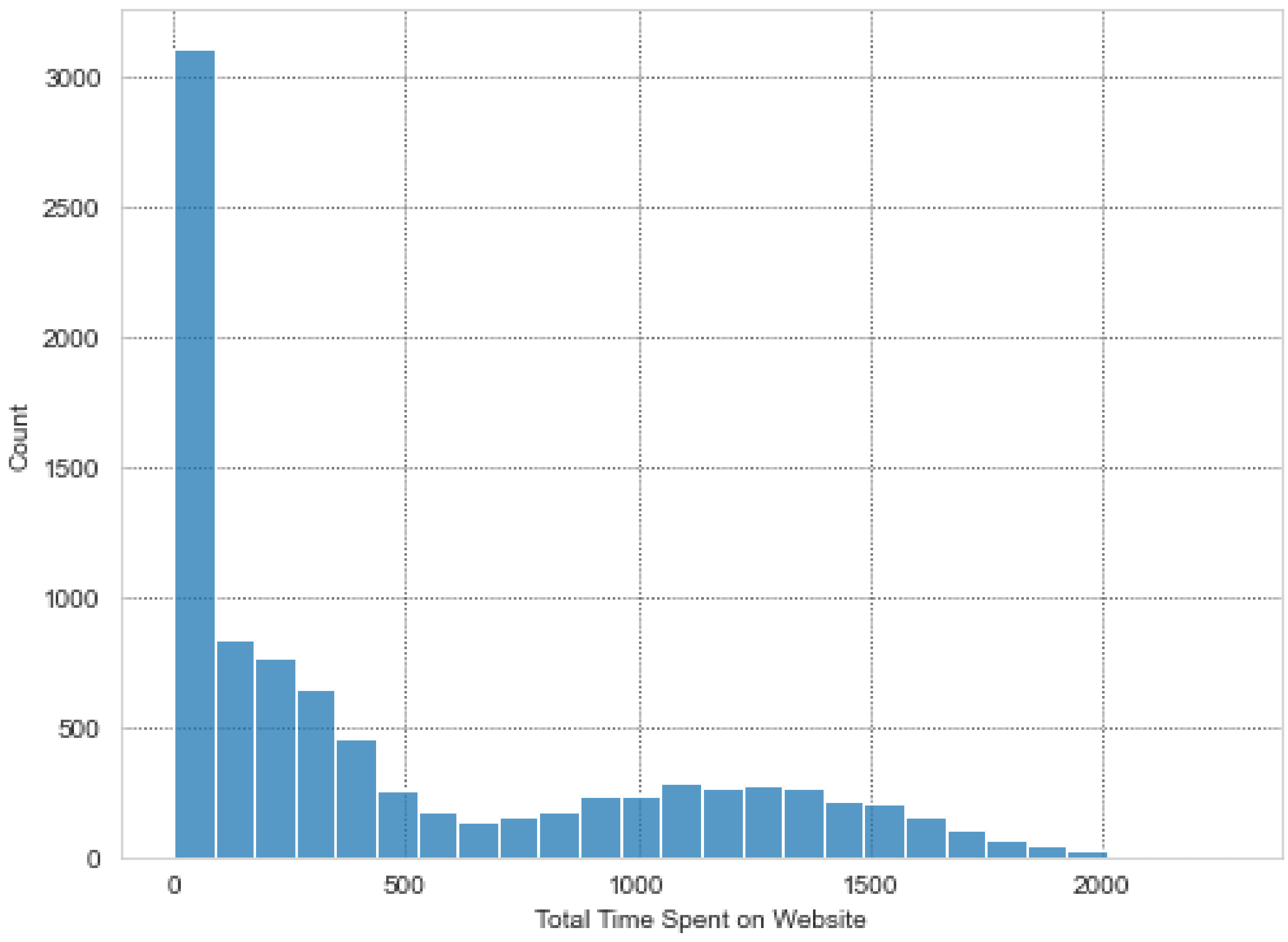


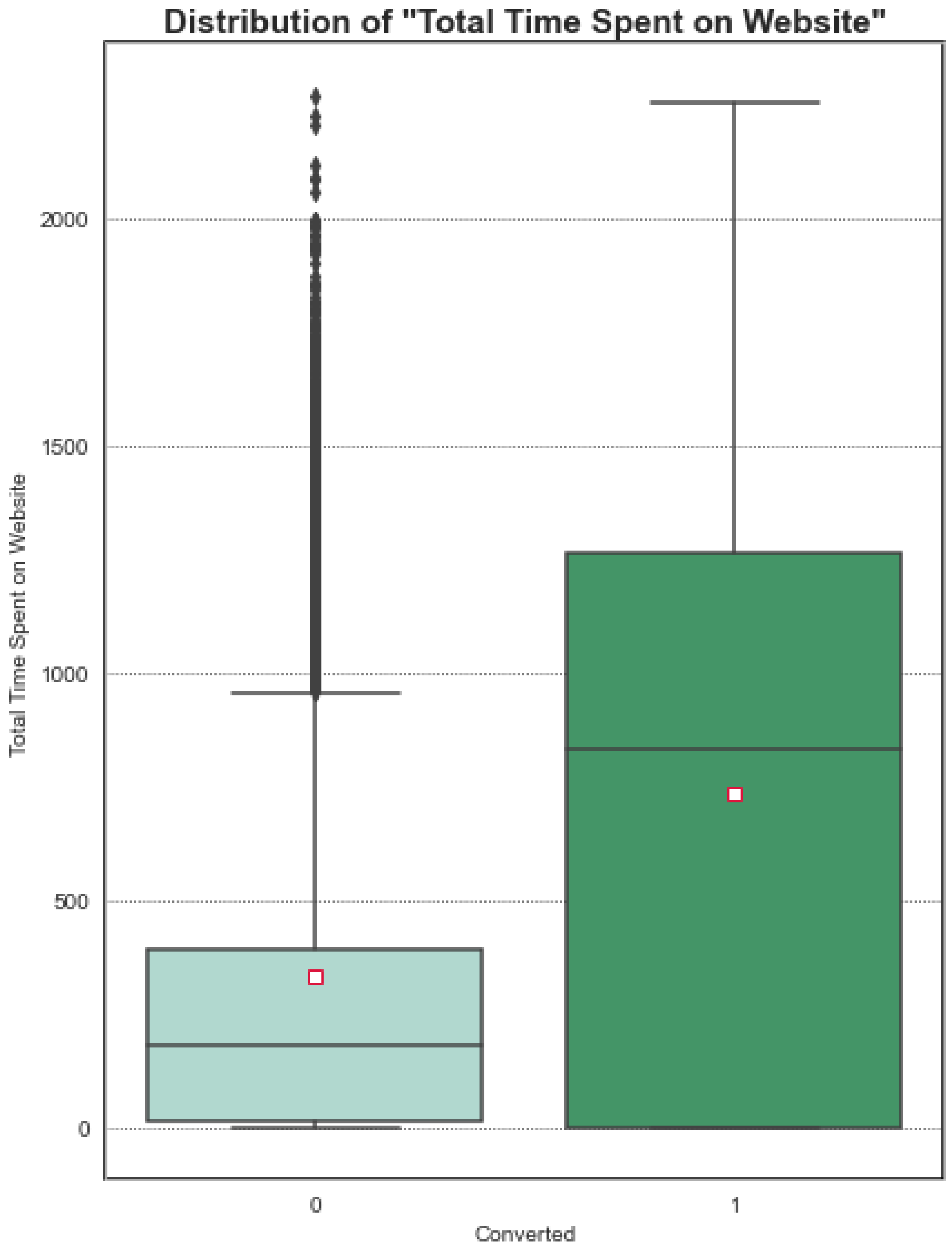


Observation:

- **The median of the converted is very little high for the Total Visits.**
- **Maximum Total visits to the website for majority of people is 7 only,**

Analysis – Total Time Spent on Website

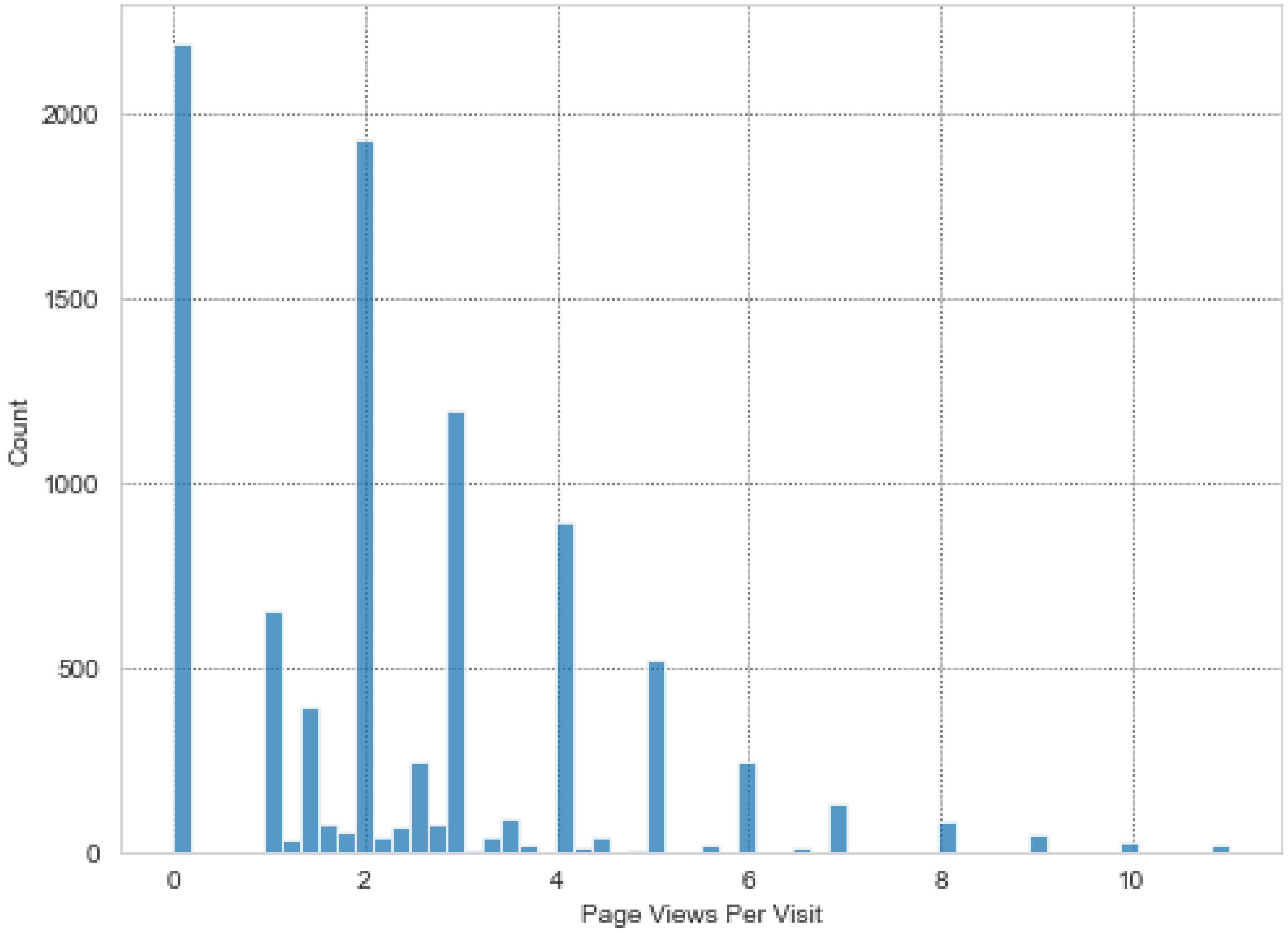


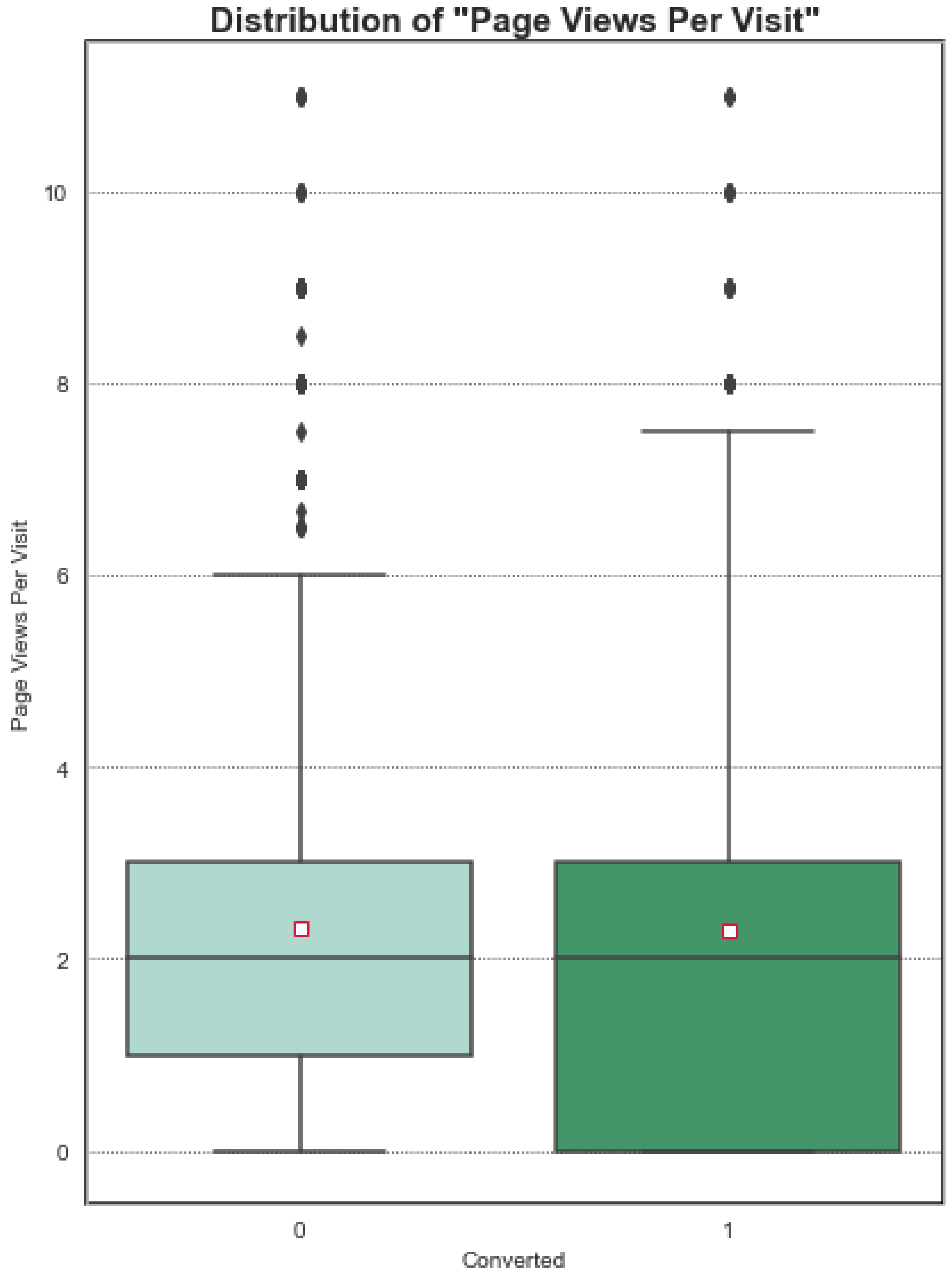


Observation:

- **Many people do not log in into Website as such.**
- **But for those who log in and view its contents the conversion rate is higher.**

Analysis – Page Views Per Visit

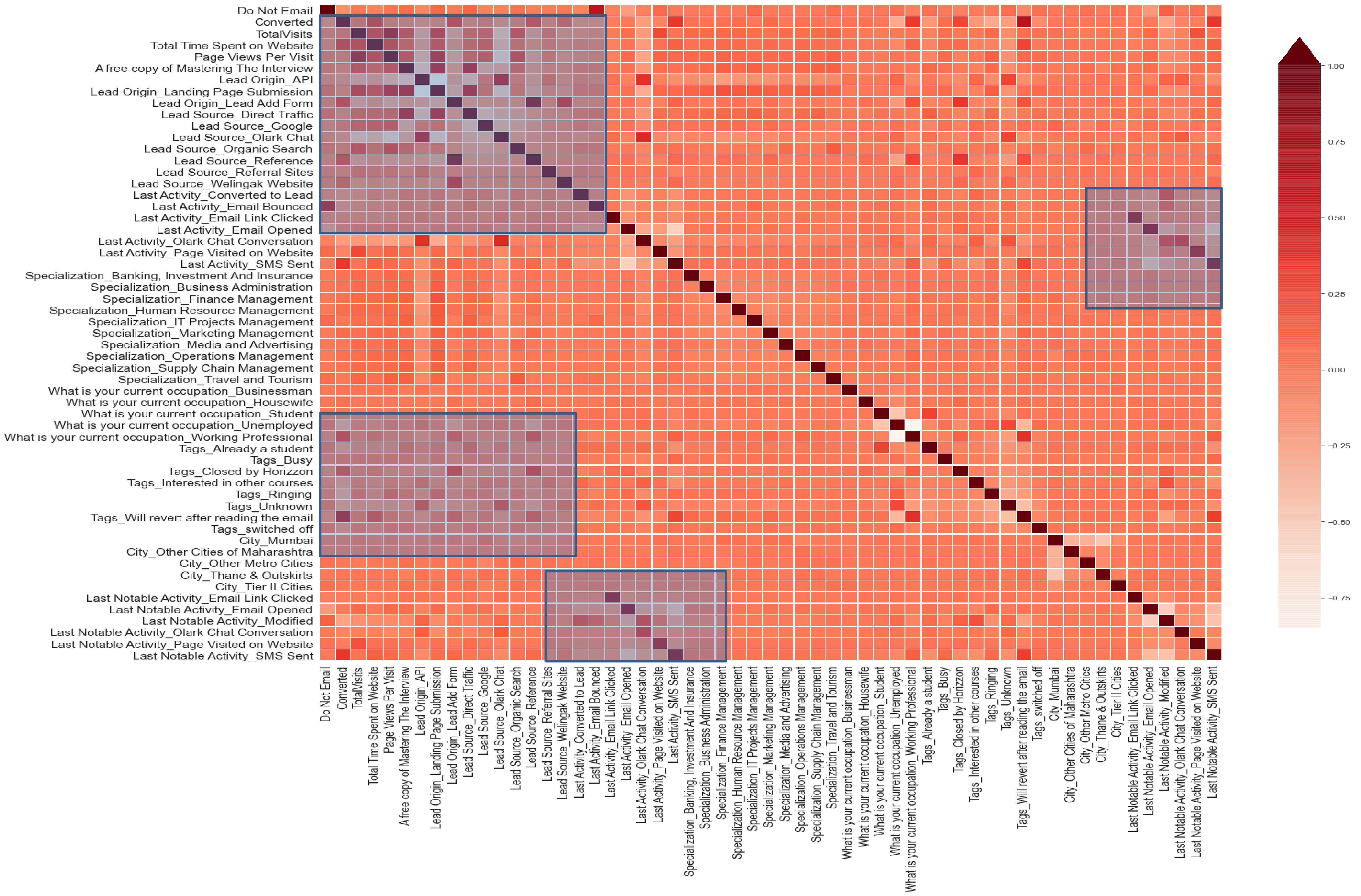




Observation:

- **The Median of the conversion rate is same.**
- **People viewing more than 4 pages per visit is very low.**

Analysis- Multivariate

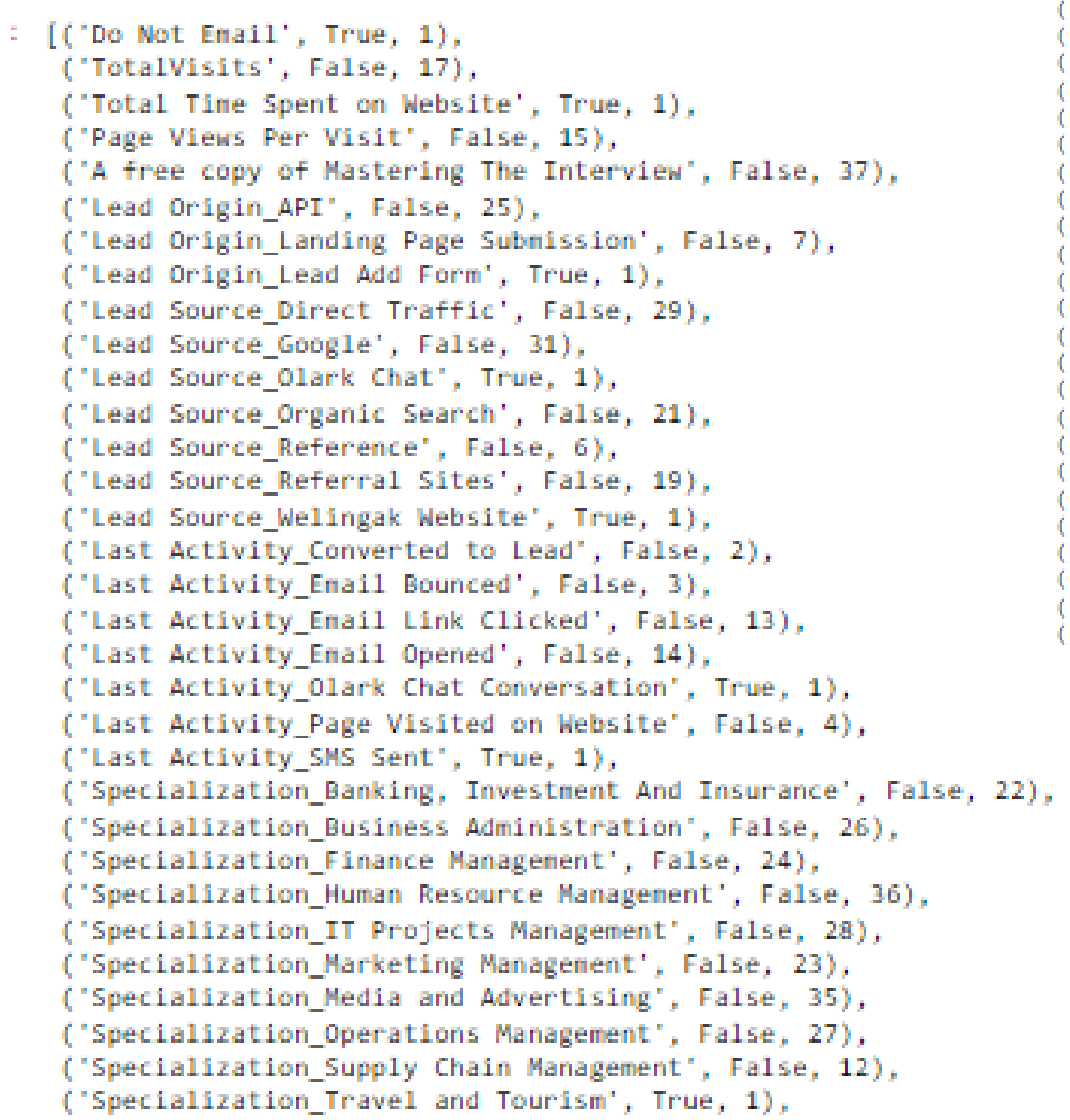


Observation:

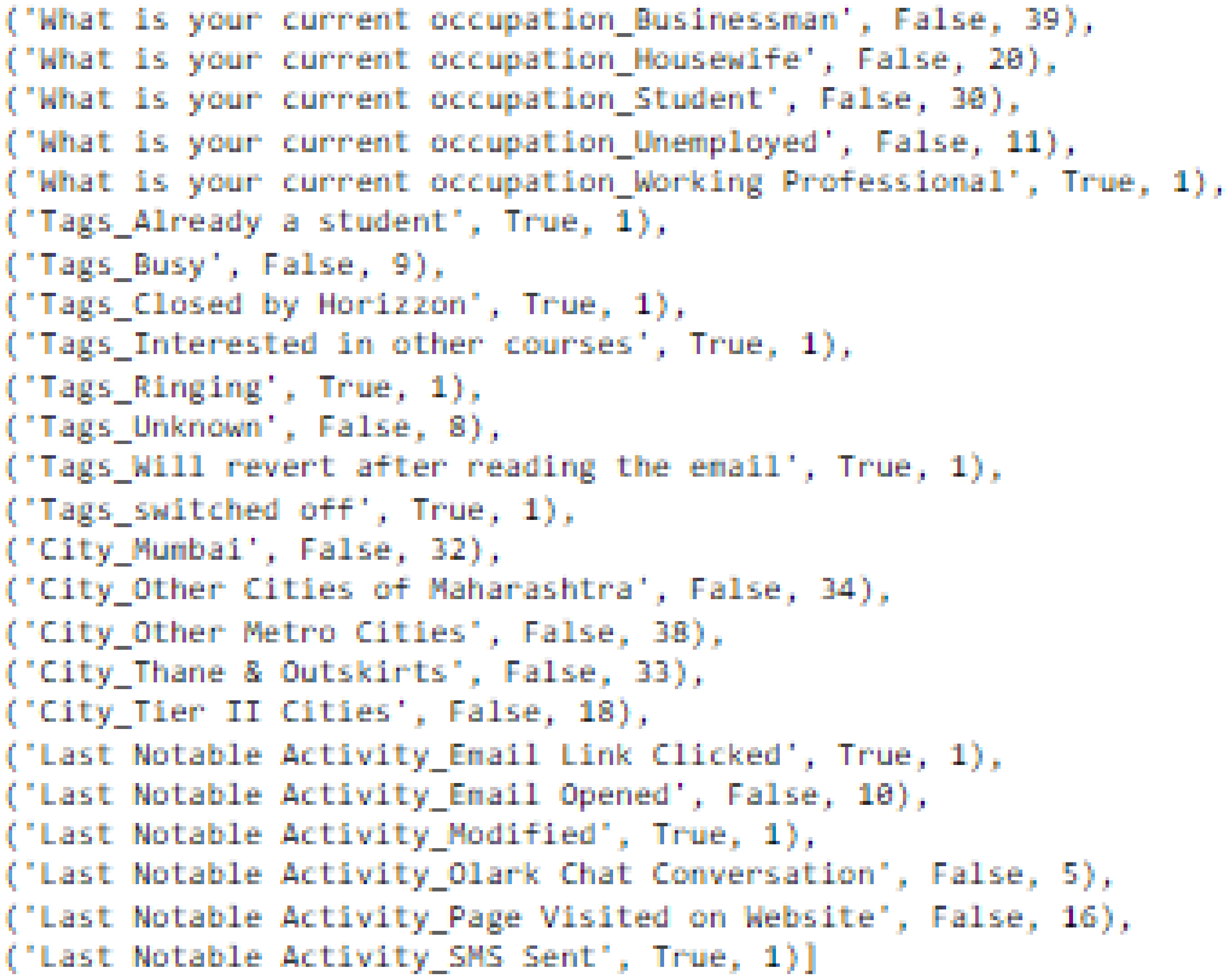
- **The Highlighted areas have a high correlation of variables while the remaining areas have medium or low correlation of variables.**

Model Building

Feature Selection



```
: [('Do Not Email', True, 1),
   ('TotalVisits', False, 17),
   ('Total Time Spent on Website', True, 1),
   ('Page Views Per Visit', False, 15),
   ('A free copy of Mastering The Interview', False, 37),
   ('Lead Origin_API', False, 25),
   ('Lead Origin_Landing Page Submission', False, 7),
   ('Lead Origin_Lead Add Form', True, 1),
   ('Lead Source_Direct Traffic', False, 29),
   ('Lead Source_Google', False, 31),
   ('Lead Source_Olark Chat', True, 1),
   ('Lead Source_Organic Search', False, 21),
   ('Lead Source_Reference', False, 6),
   ('Lead Source_Referral Sites', False, 19),
   ('Lead Source_Welingak Website', True, 1),
   ('Last Activity_Converted to Lead', False, 2),
   ('Last Activity_Email Bounced', False, 3),
   ('Last Activity_Email Link Clicked', False, 13),
   ('Last Activity_Email Opened', False, 14),
   ('Last Activity_Olark Chat Conversation', True, 1),
   ('Last Activity_Page Visited on Website', False, 4),
   ('Last Activity_SMS Sent', True, 1),
   ('Specialization_Banking, Investment And Insurance', False, 22),
   ('Specialization_Business Administration', False, 26),
   ('Specialization_Finance Management', False, 24),
   ('Specialization_Human Resource Management', False, 36),
   ('Specialization_IT Projects Management', False, 28),
   ('Specialization_Marketing Management', False, 23),
   ('Specialization_Media and Advertising', False, 35),
   ('Specialization_Operations Management', False, 27),
   ('Specialization_Supply Chain Management', False, 12),
   ('Specialization_Travel and Tourism', True, 1),
```



```
('What is your current occupation_Businessman', False, 39),
('What is your current occupation_Housewife', False, 20),
('What is your current occupation_Student', False, 30),
('What is your current occupation_Unemployed', False, 11),
('What is your current occupation_Working Professional', True, 1),
('Tags_Already a student', True, 1),
('Tags_Busy', False, 9),
('Tags_Closed by Horizzon', True, 1),
('Tags_Interested in other courses', True, 1),
('Tags_Ringing', True, 1),
('Tags_Unknown', False, 8),
('Tags_Will revert after reading the email', True, 1),
('Tags_switched off', True, 1),
('City_Mumbai', False, 32),
('City_Other Cities of Maharashtra', False, 34),
('City_Other Metro Cities', False, 38),
('City_Thane & Outskirts', False, 33),
('City_Tier II Cities', False, 18),
('Last Notable Activity_Email Link Clicked', True, 1),
('Last Notable Activity_Email Opened', False, 10),
('Last Notable Activity_Modified', True, 1),
('Last Notable Activity_Olark Chat Conversation', False, 5),
('Last Notable Activity_Page Visited on Website', False, 16),
('Last Notable Activity_SMS Sent', True, 1)]
```

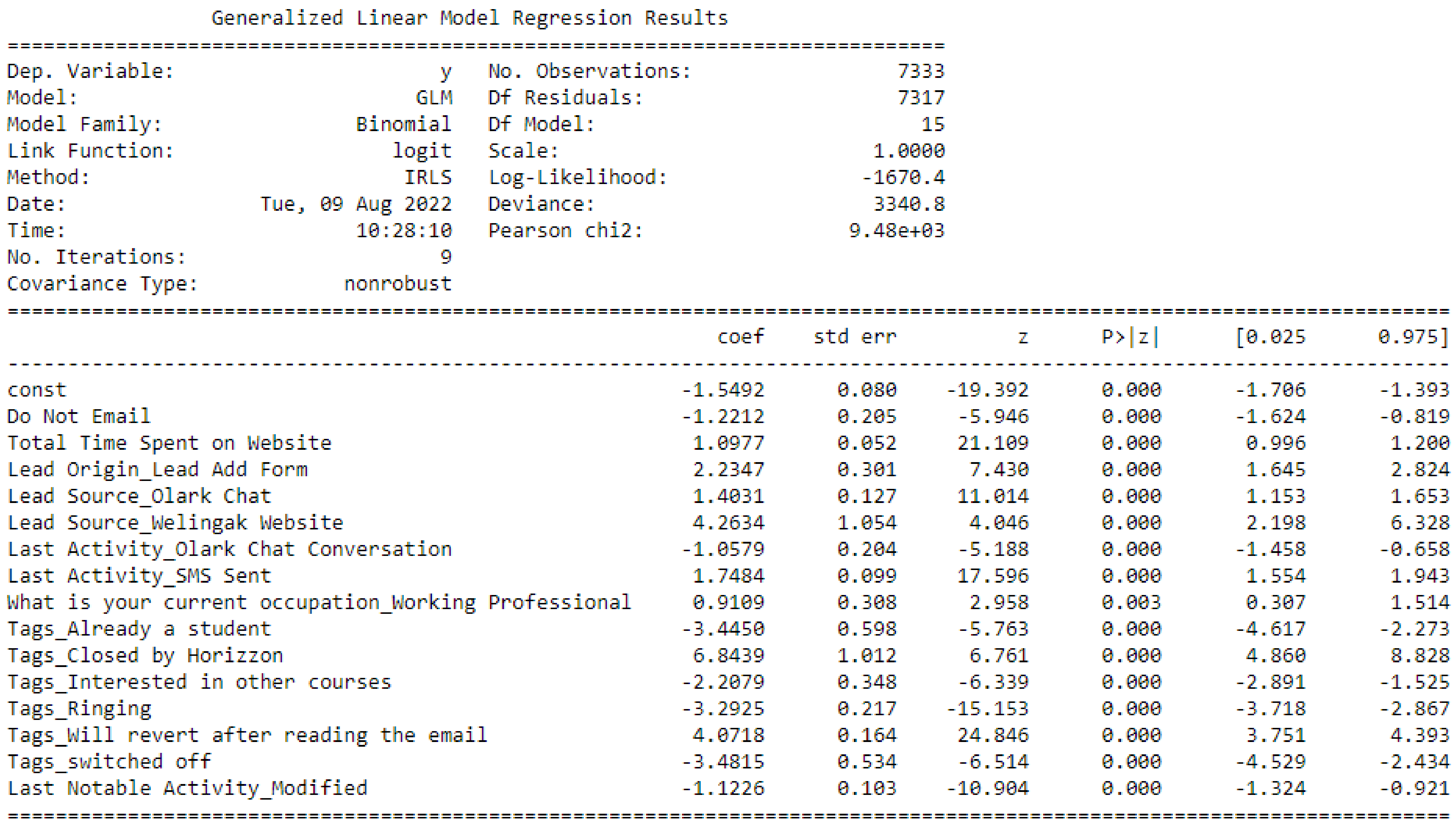
Observation:

- **Selecting the Top 18 Features from the total features using RFE method**

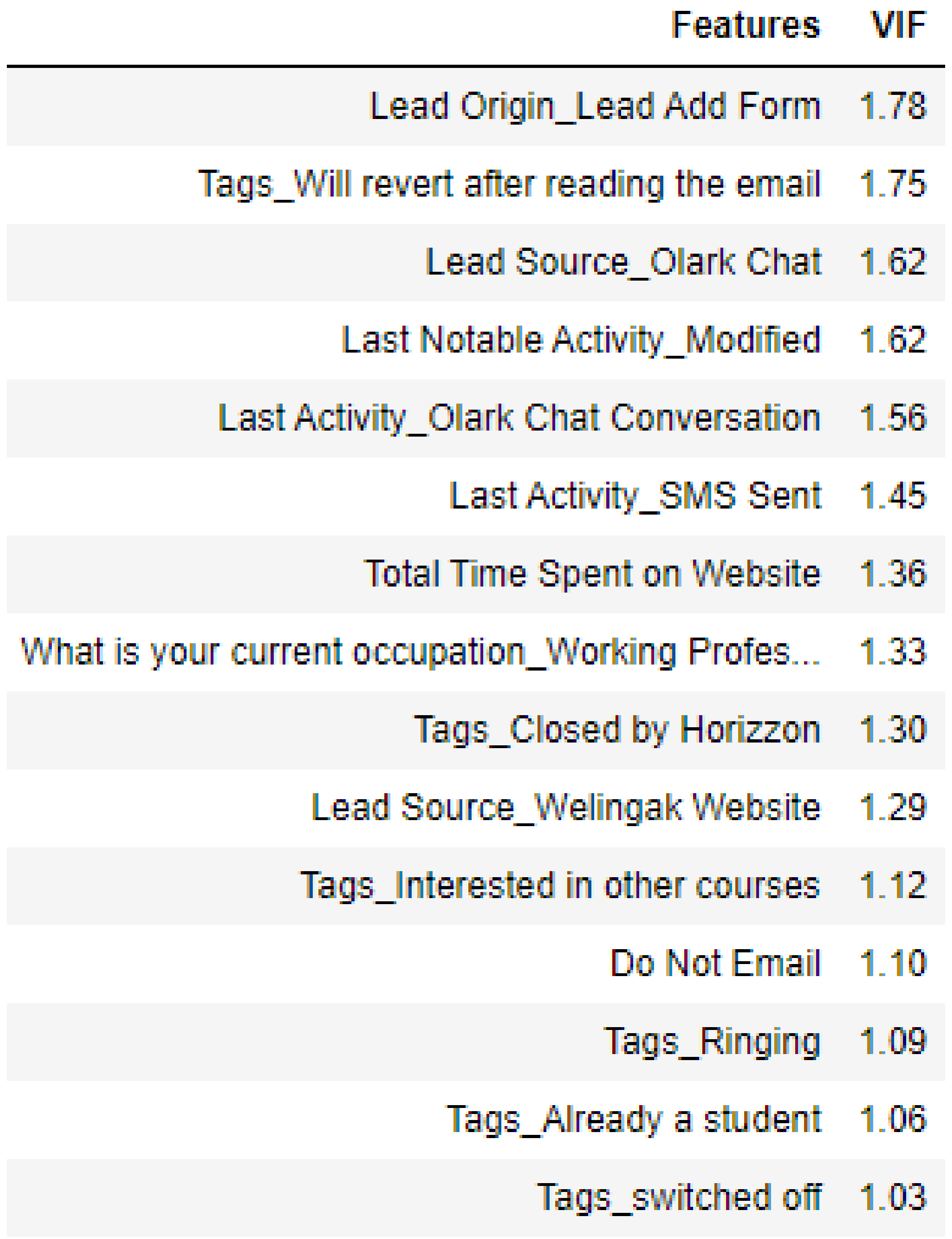
Feature Selection

1. Do Not Email
2. Total Time Spent on Website
3. Lead Origin_Lead Add Form
4. 'Lead Source_Olark Chat
5. 'Lead Source_Welingak Website
6. 'Last Activity_Olark Chat Conversation
7. 'Last Activity_SMS Sent
8. Specialization_Travel and Tourism
9. What is your current occupation_Working Professional
10. Tags_Already a student
11. Tags_Closed by Horizzon
12. Tags_Interested in other courses
13. Tags_Ringing
14. Tags_Will revert after reading the email
15. Tags_switched off
16. Last Notable Activity_Email Link Clicked
17. Last Notable Activity_Modified
18. Last Notable Activity_SMS Sent

Final Model Summary



```
                 Generalized Linear Model Regression Results                  
==============================================================================
Dep. Variable:                      y   No. Observations:                 7333
Model:                            GLM   Df Residuals:                     7317
Model Family:                Binomial   Df Model:                           15
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1670.4
Date:                Tue, 09 Aug 2022   Deviance:                       3340.8
Time:                        10:28:10   Pearson chi2:                 9.48e+03
No. Iterations:                     9                                         
Covariance Type:            nonrobust                                         
========================================================================================================================
                                                           coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                               -1.5492      0.080    -19.392      0.000      -1.706      -1.393
Do Not Email                                        -1.2212      0.205     -5.946      0.000      -1.624      -0.819
Total Time Spent on Website                          1.0977      0.052     21.109      0.000       0.996       1.200
Lead Origin_Lead Add Form                            2.2347      0.301      7.430      0.000       1.645       2.824
Lead Source_Olark Chat                               1.4031      0.127     11.014      0.000       1.153       1.653
Lead Source_Welingak Website                         4.2634      1.054      4.046      0.000       2.198       6.328
Last Activity_Olark Chat Conversation               -1.0579      0.204     -5.188      0.000      -1.458      -0.658
Last Activity_SMS Sent                               1.7484      0.099     17.596      0.000       1.554       1.943
What is your current occupation_Working Professional 0.9109      0.308      2.958      0.003       0.307       1.514
Tags_Already a student                              -3.4450      0.598     -5.763      0.000      -4.617      -2.273
Tags_Closed by Horizzon                              6.8439      1.012      6.761      0.000       4.860       8.828
Tags_Interested in other courses                    -2.2079      0.348     -6.339      0.000      -2.891      -1.525
Tags_Ringing                                        -3.2925      0.217    -15.153      0.000      -3.718      -2.867
Tags_Will revert after reading the email             4.0718      0.164     24.846      0.000       3.751       4.393
Tags_switched off                                   -3.4815      0.534     -6.514      0.000      -4.529      -2.434
Last Notable Activity_Modified                      -1.1226      0.103    -10.904      0.000      -1.324      -0.921
========================================================================================================================
```

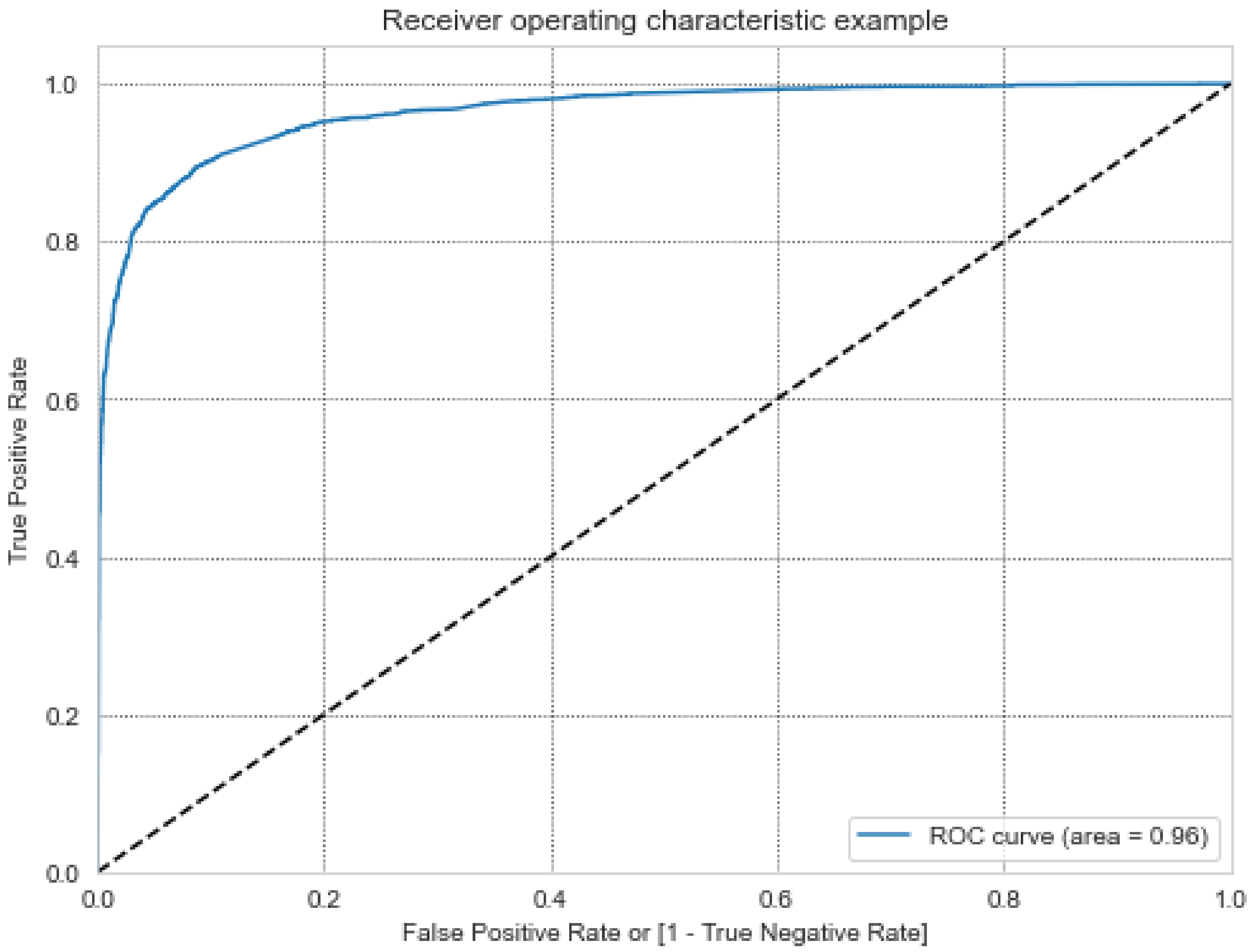


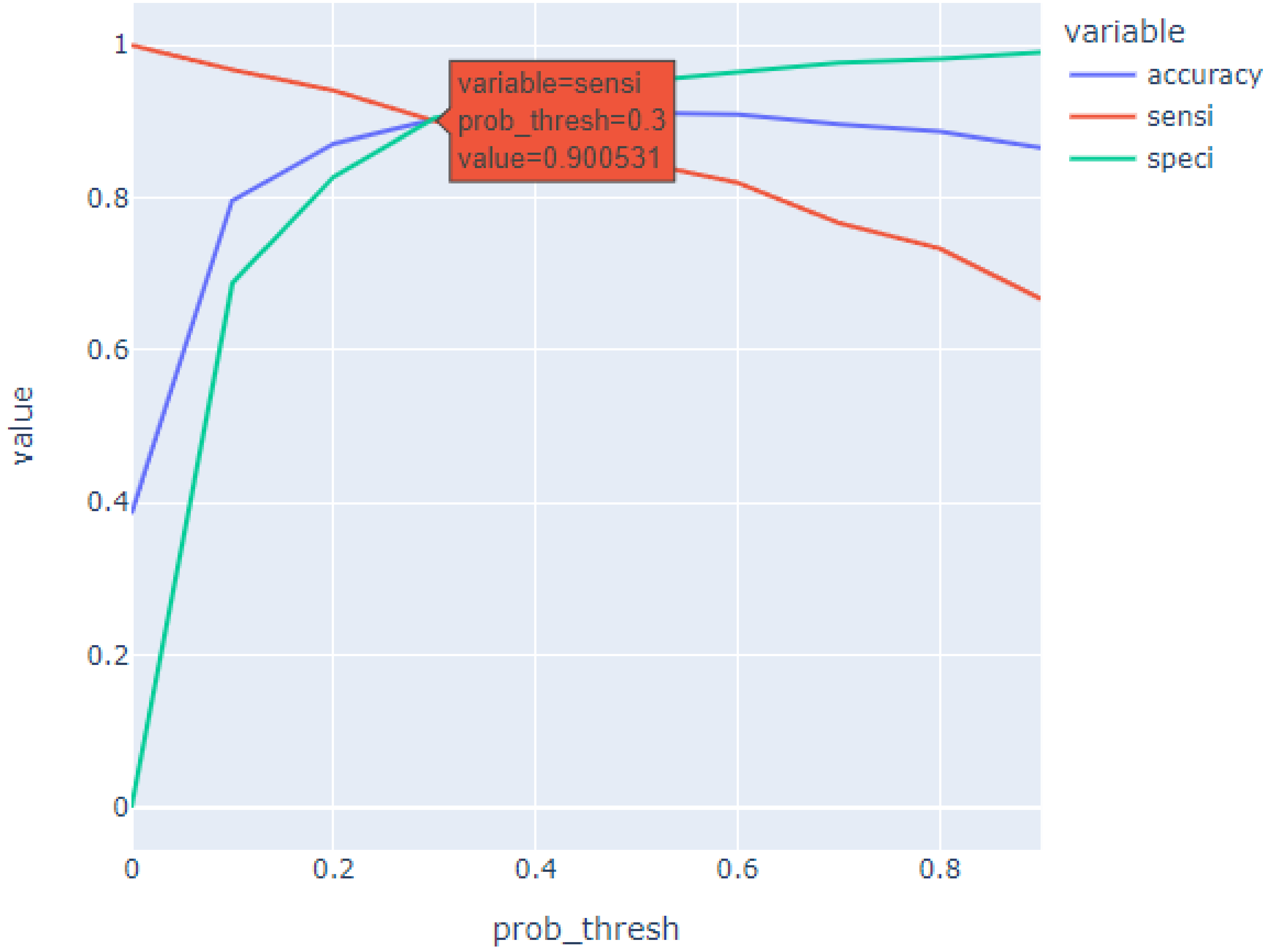
Features	VIF
Lead Origin_Lead Add Form	1.78
Tags_Will revert after reading the email	1.75
Lead Source_Olark Chat	1.62
Last Notable Activity_Modified	1.62
Last Activity_Olark Chat Conversation	1.56
Last Activity_SMS Sent	1.45
Total Time Spent on Website	1.36
What is your current occupation_Working Profes...	1.33
Tags_Closed by Horizzon	1.30
Lead Source_Welingak Website	1.29
Tags_Interested in other courses	1.12
Do Not Email	1.10
Tags_Ringing	1.09
Tags_Already a student	1.06
Tags_switched off	1.03

Observation:

- **The final model looks good. p-values corresponding to all the variables is very low, which means all the features in final model are significant.**
- **VIFs are are < 2 for all the final set of features which means very low multicollinearity.**

Optimal cut-off



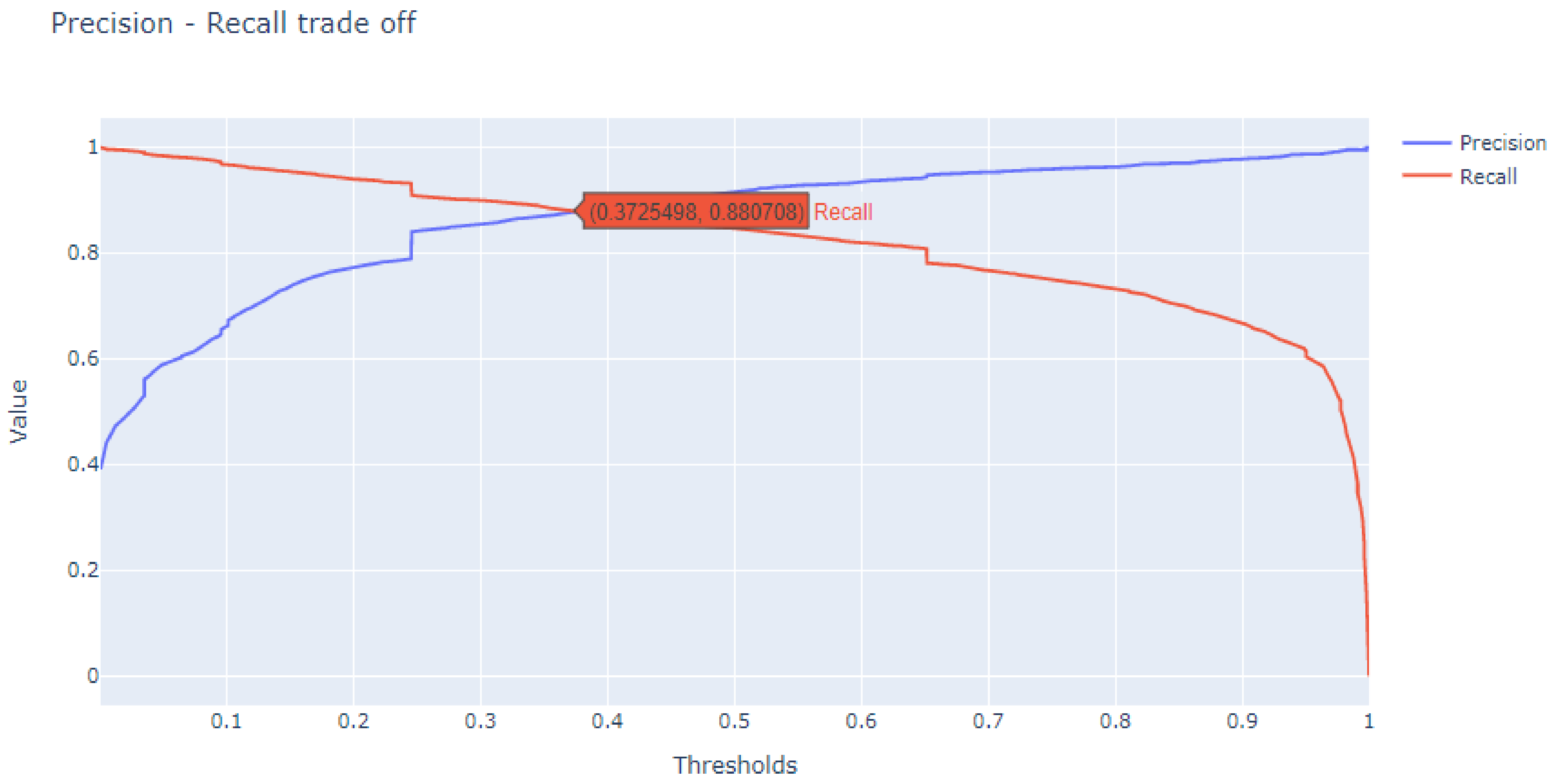


Optimal cut-off is 0.3 as per the point of contact of sensitivity, specificity and accuracy

Precision – Recall Cut off

Observation:

- **The Point of intersection of Precision and Recall is 0.372 approximately.**
- **We want to identify as many positives(leads that will convert) as possible but the CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. So this would mean if the company were to classify only leads predicted as positives(leads that will convert) as hot leads, the precision score has to be above 0.8.**
- **So, we can push the threshold a little lower, say 0.27. Lets look at confusion matrix with this threshold.**



Model Evaluation

Train Data

Optimal Cut off 0.3

0.27 Precision – Recall Trade-off

Accuracy	90%
Sensitivity	90.05%
Specificity	90.41%
AUC	96%
F1 Score	90%
Recall	90.41%
Precision	84.74%
Positive predictive value	84.73%
Negative predictive value	93.72%

Test Data

Optimal Cut off 0.3

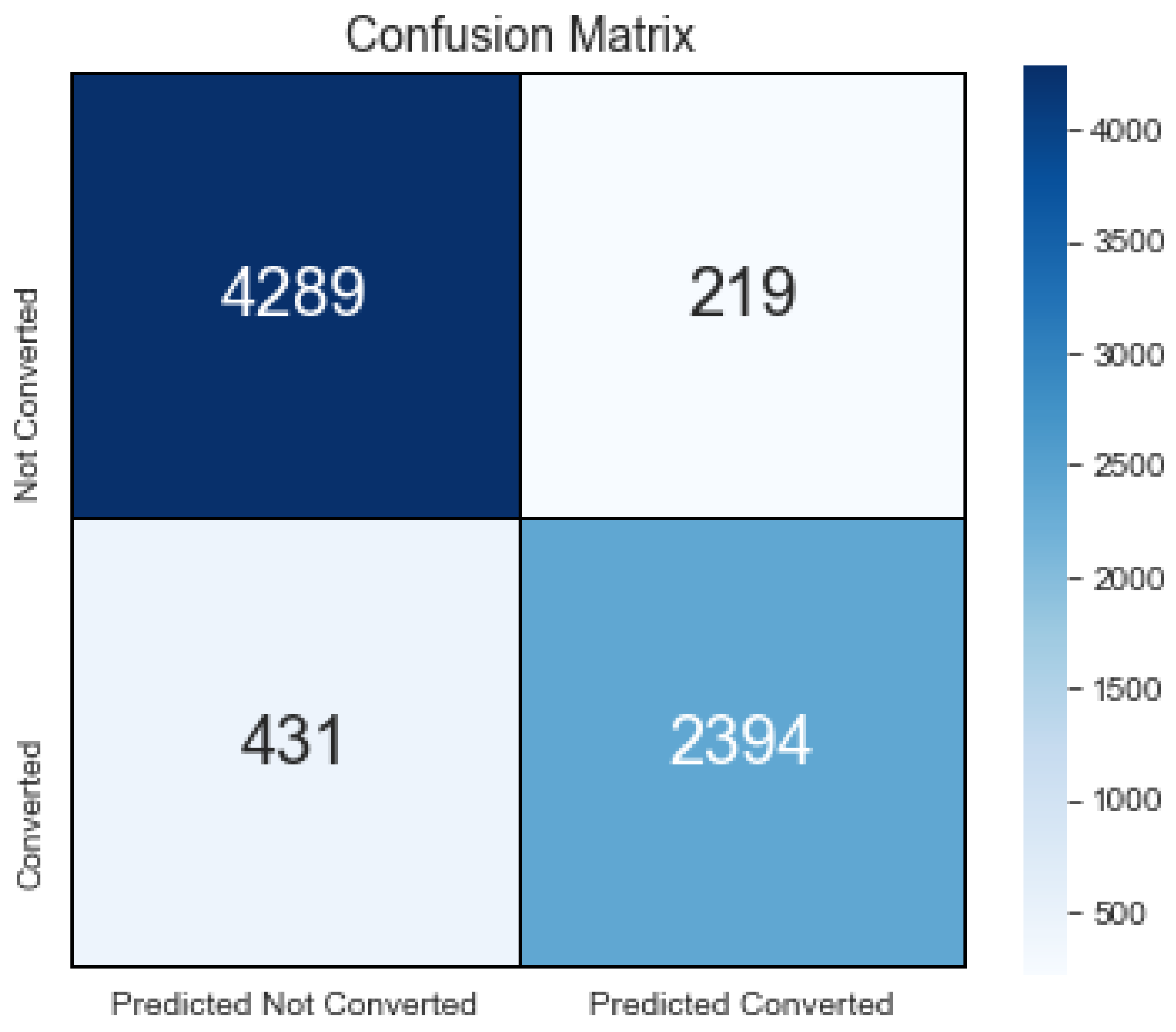
0.27 Precision – Recall Trade-off

Accuracy	Sensitivity	Specificity	AUC	F1 Score
90.13%	90.70%	89.71%	96%	90%

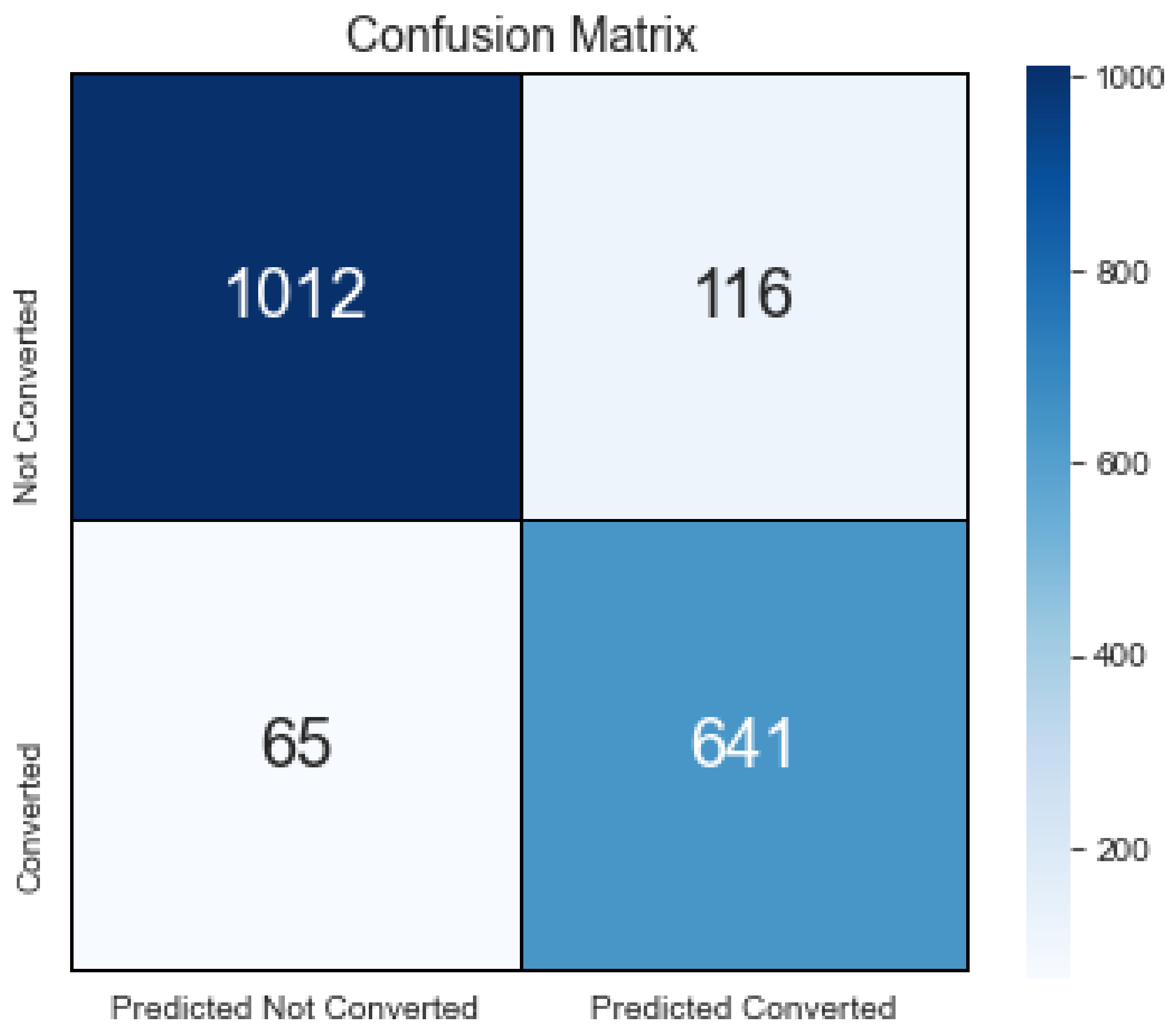
Recall	Precision	Positive predictive value	Negative predictive value
90.72%	84.68%	84.67%	93.96%

Confusion Matrix

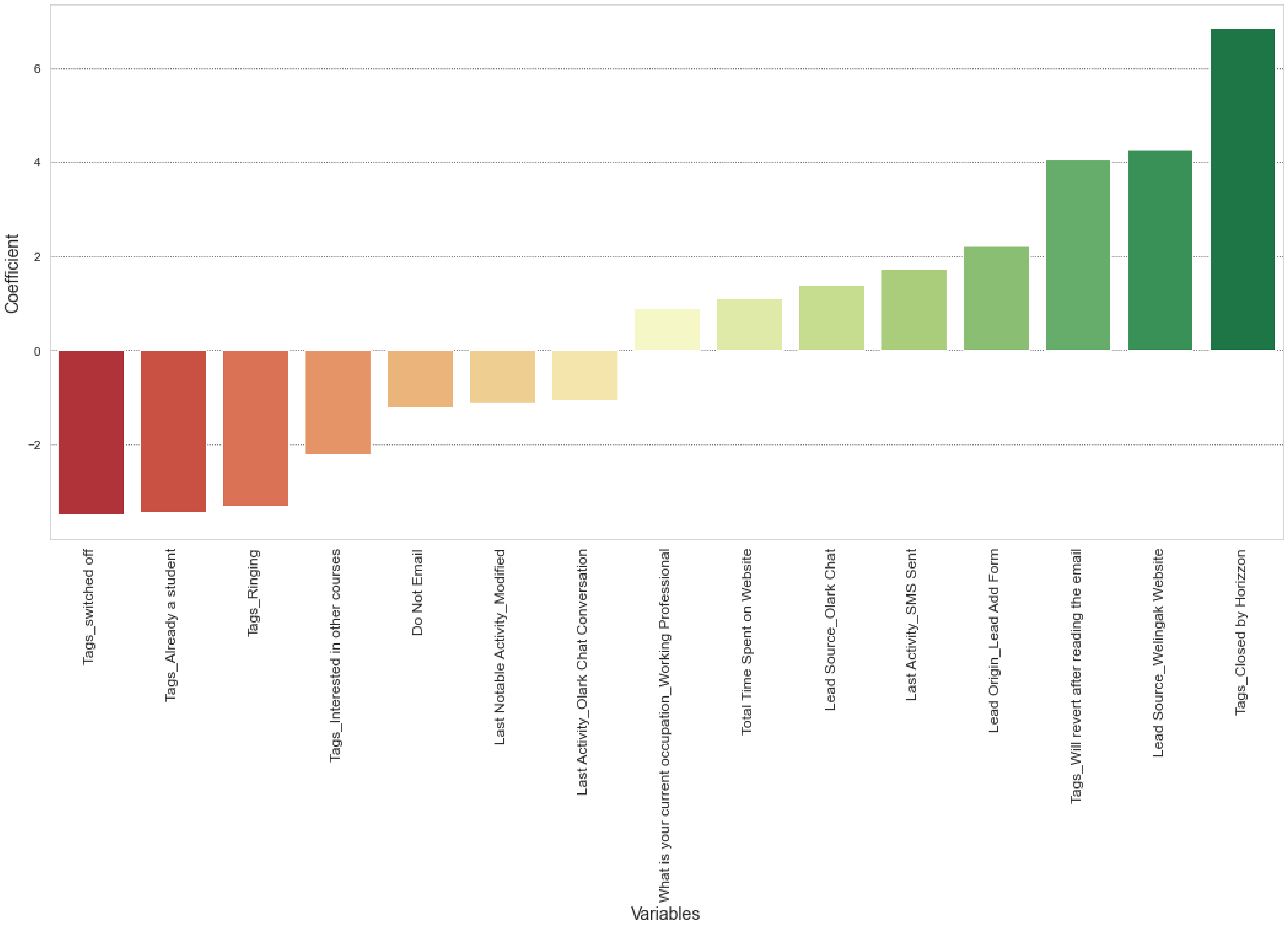
Train Dataset Prediction Confusion Matrix



Test Dataset Prediction Confusion Matrix



Variable vs Its Coefficient plot



Feature Importance

1. 'Tags_Closed by Horizzon': If this variable is True or 1, then the log-odds go up by 6.84.
2. 'Lead Source_Welingak Website': If the Lead source is Welingak website then the log odds increase by 4.26.
3. 'Tags_Will revert after reading the email': If the current status / tag is 'Will revert after reading the email' then the log odds increase by 4.07.
4. 'Tags_switched off': If the current status / tag is 'switched off', then the log odds decrease by 3.48.
5. 'Tags_Already a student': If the current status / tag is 'Already a student', then the log odds decrease by 3.44.
6. 'Tags_Ringing': If the current status / tag is 'Ringing', then the log odds decrease by 3.29.

Recommendation

- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and Google leads and generate more leads from reference and welingak website.

- Lead conversion rate, can be improved by focusing more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form

- Though Google is the highest source to get leads, the lead conversion through Google is low comparatively.

- Focus on Working Professional which has high conversion

- Website should be made more engaging to make leads spend more time

- Improve the Olark Chat service since this is affecting the conversion negatively