LEAD SCORING CASE STUDY

By:

Ajit M

Problem Statement

- X education sells online courses to Industry Professionals.
- X education gets a lots of leads,
- But lead conversion rate it very poor. (For example, if they acquire 100 leads in a day , only about 30 of them are converted)
- 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.

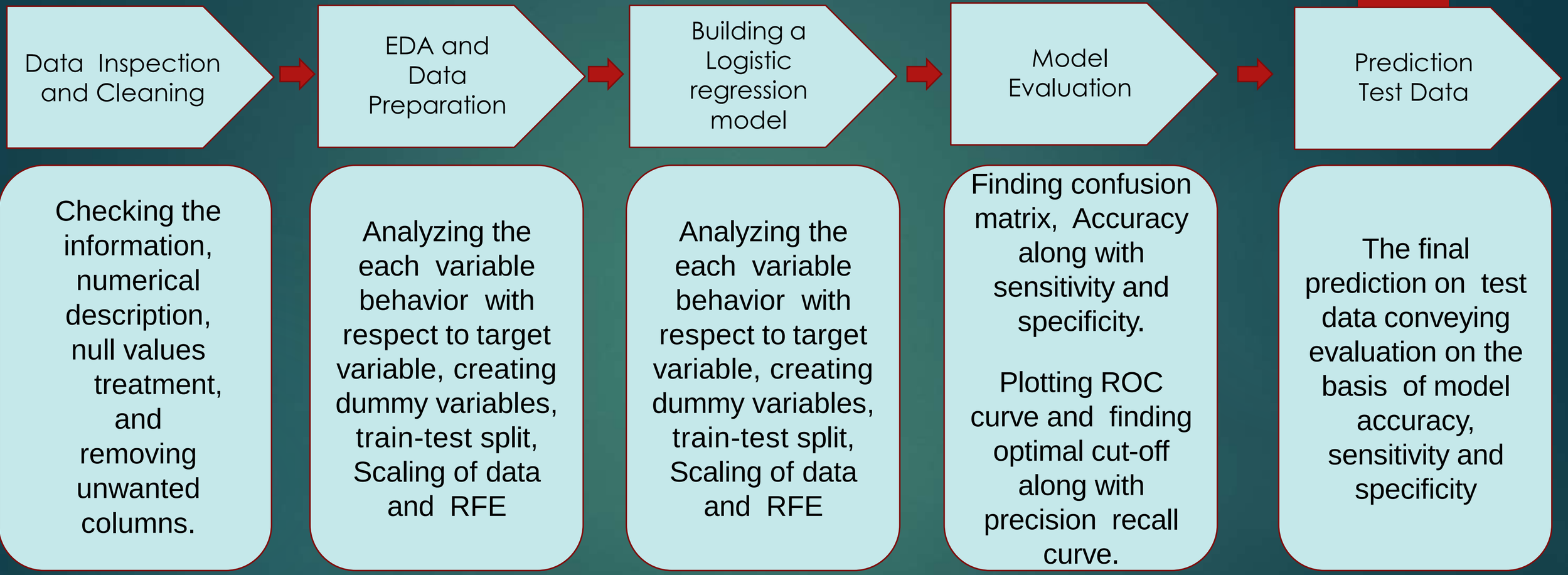
Data Given

- Dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted

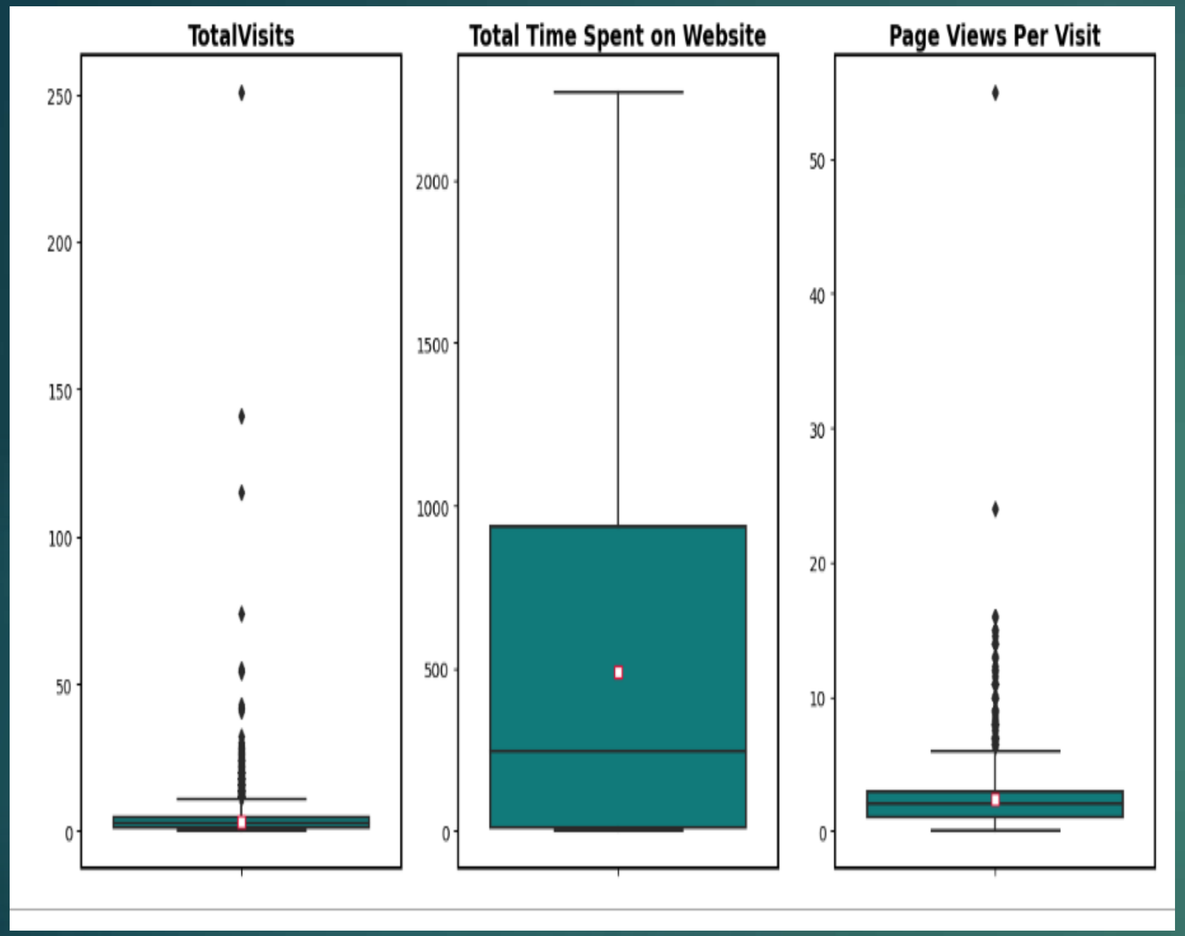
Business Objective

- To Help X education to select the “Hot Leads"
- 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.

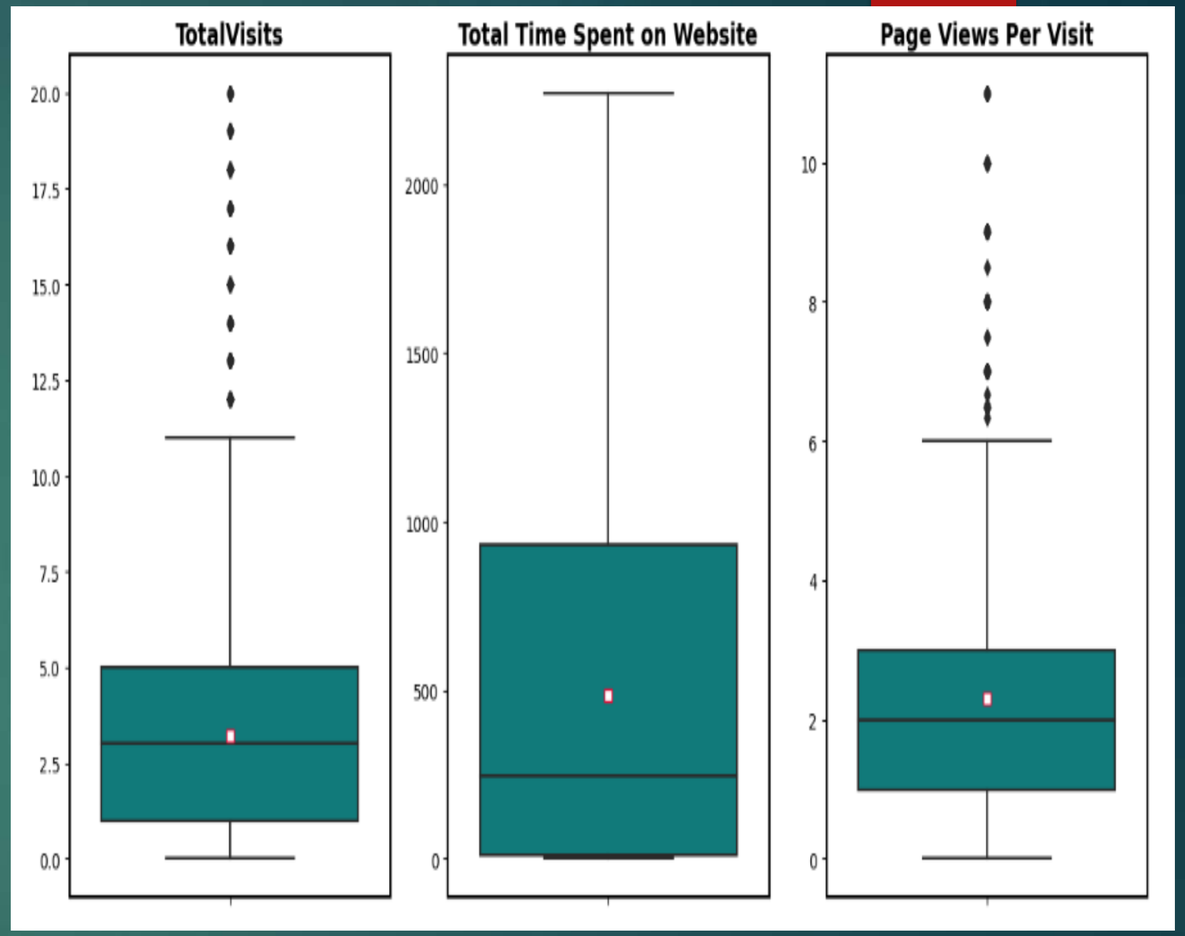
Analysis Process



Outlier Study



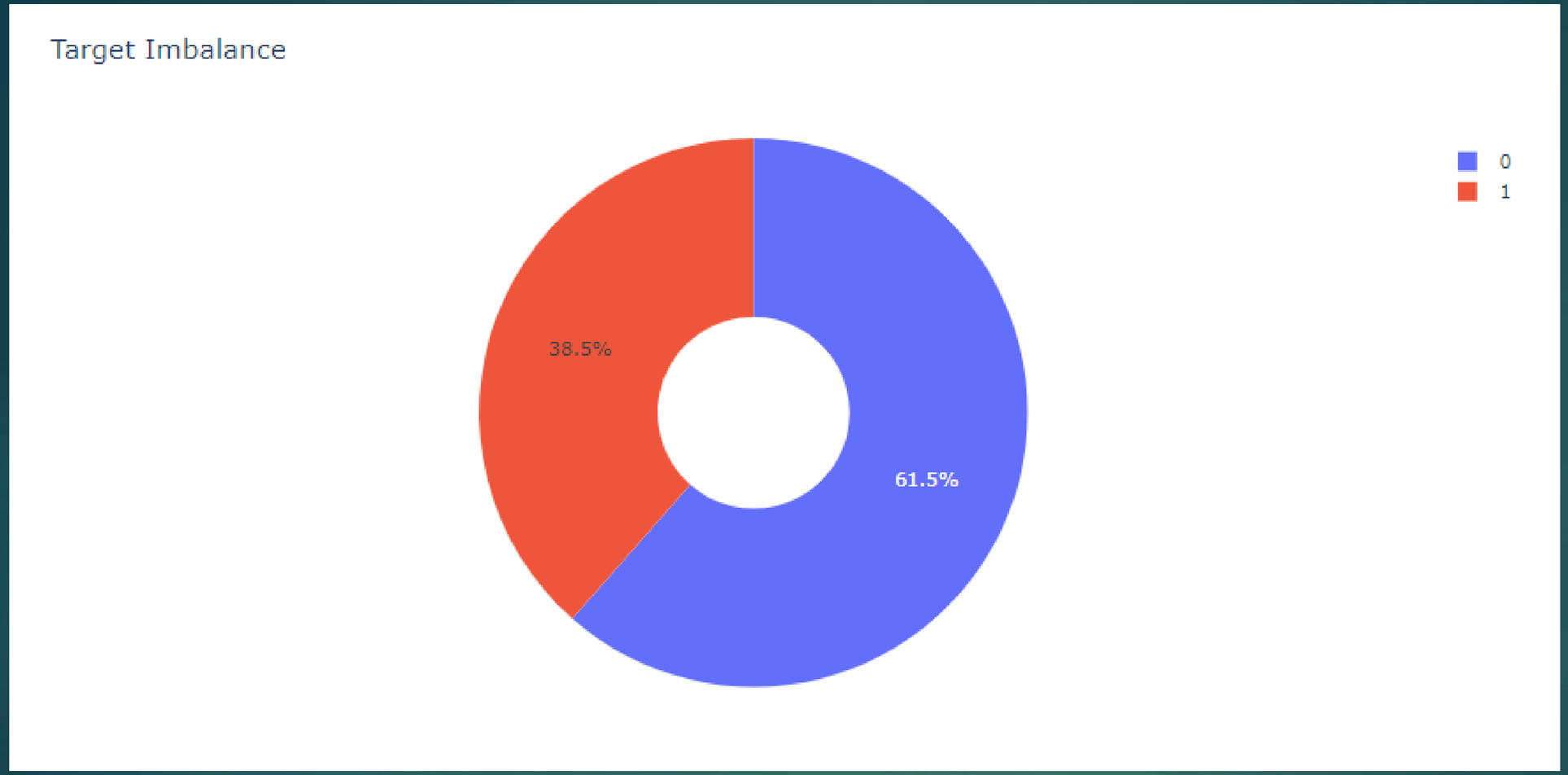
Conclusion : BEFORE ELIMINATION
outliers in ‘Total visits' and 'Pages views per visit‘ have more Outliers
we need to eliminate them.



Conclusion : AFTER ELIMINATION
The result is decent box plot with few outlier.
Retaining 0.995 quantile of data

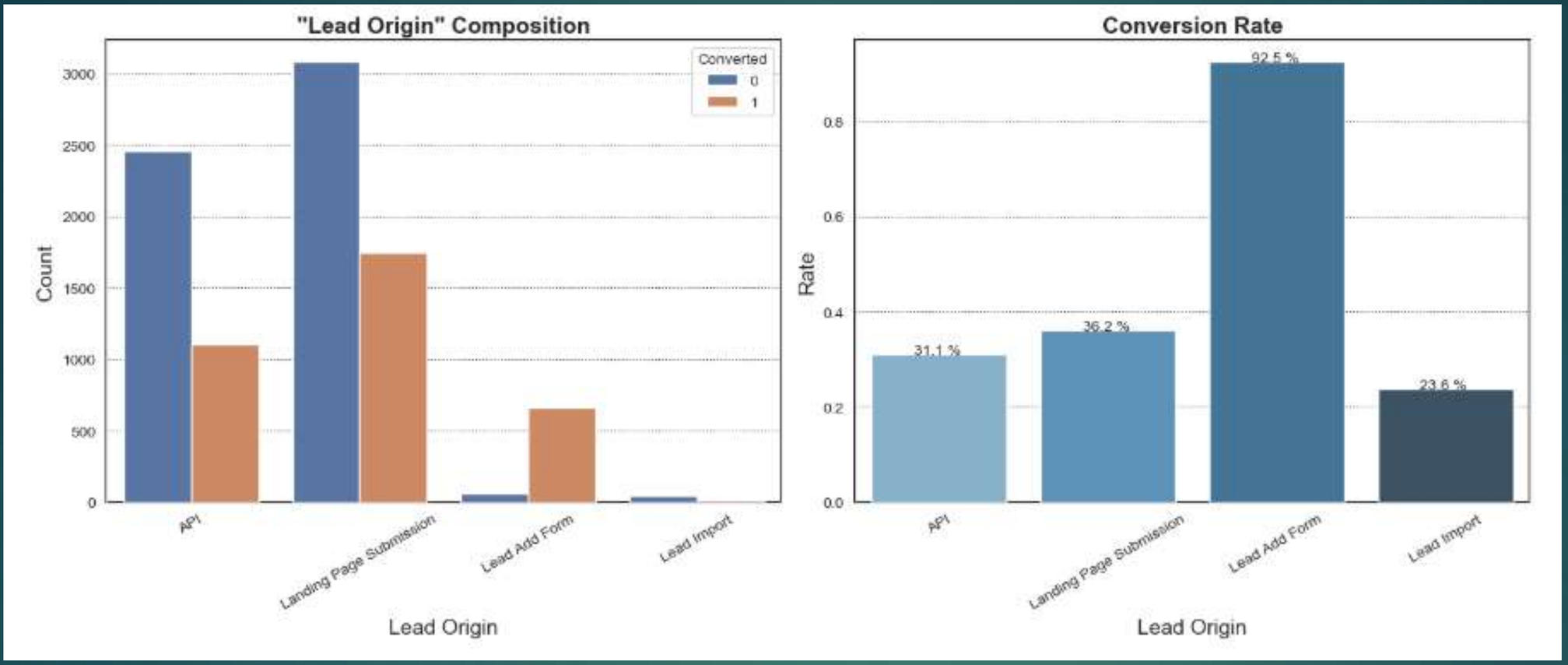
Exploratory Data Analysis

EDA - Balance Ratio Analysis of Target Variable



Conclusion :
In mentioned Chart found, 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.

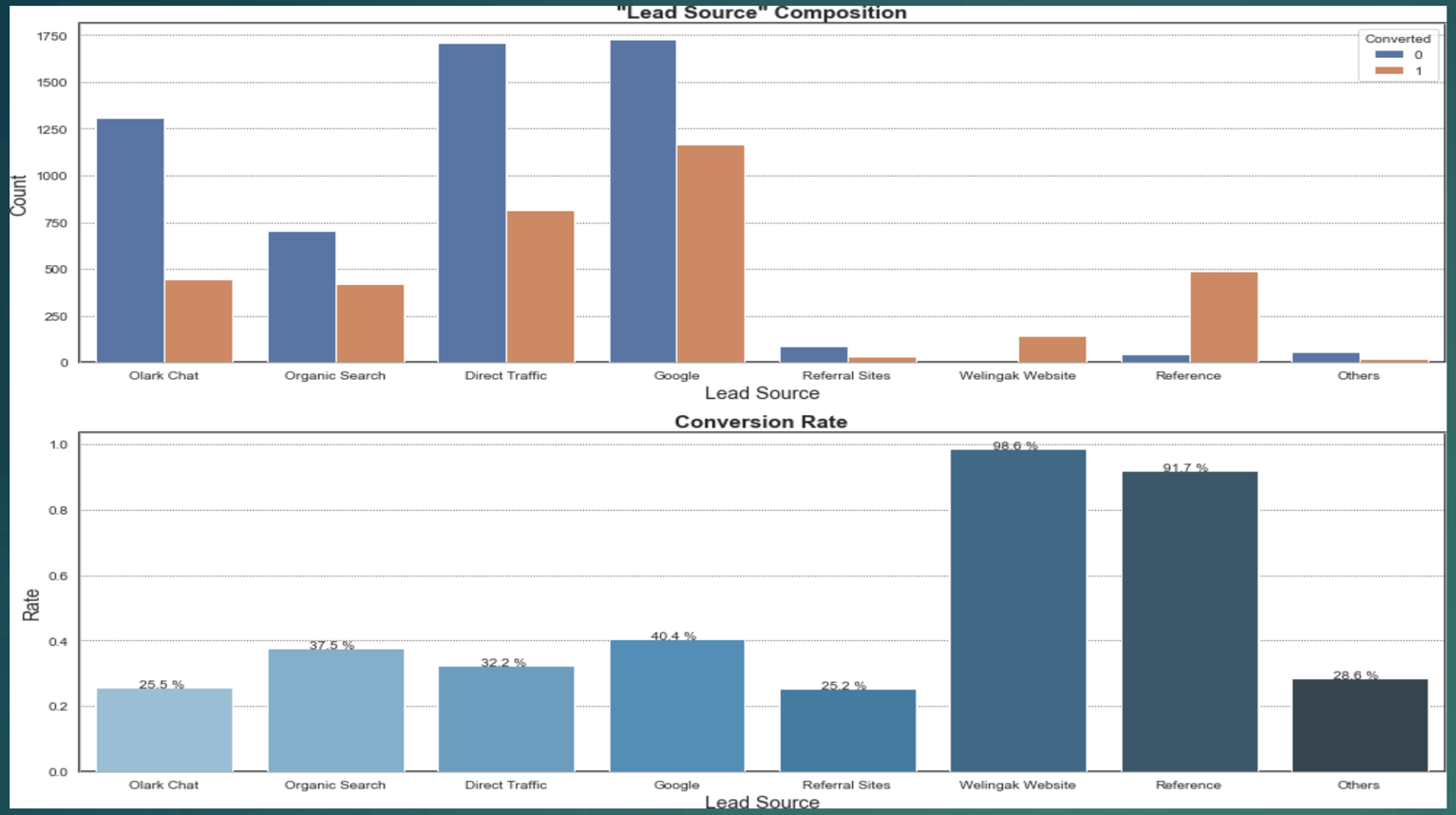
EDA - Lead Origin



Conclusion :

- The majority of the leads came 'landing page submission' followed by API
- Leads from the 'Lead Add Form' have the highest conversion rate in this category, (90%)
- 'Lead Import' are few in numberr, and the conversion rate is also low (23.6%)
- 'Lead Add Form' identifies is less leads but the conversion rate of the leads identified is very high. Company should try to bring in more leads by this method.

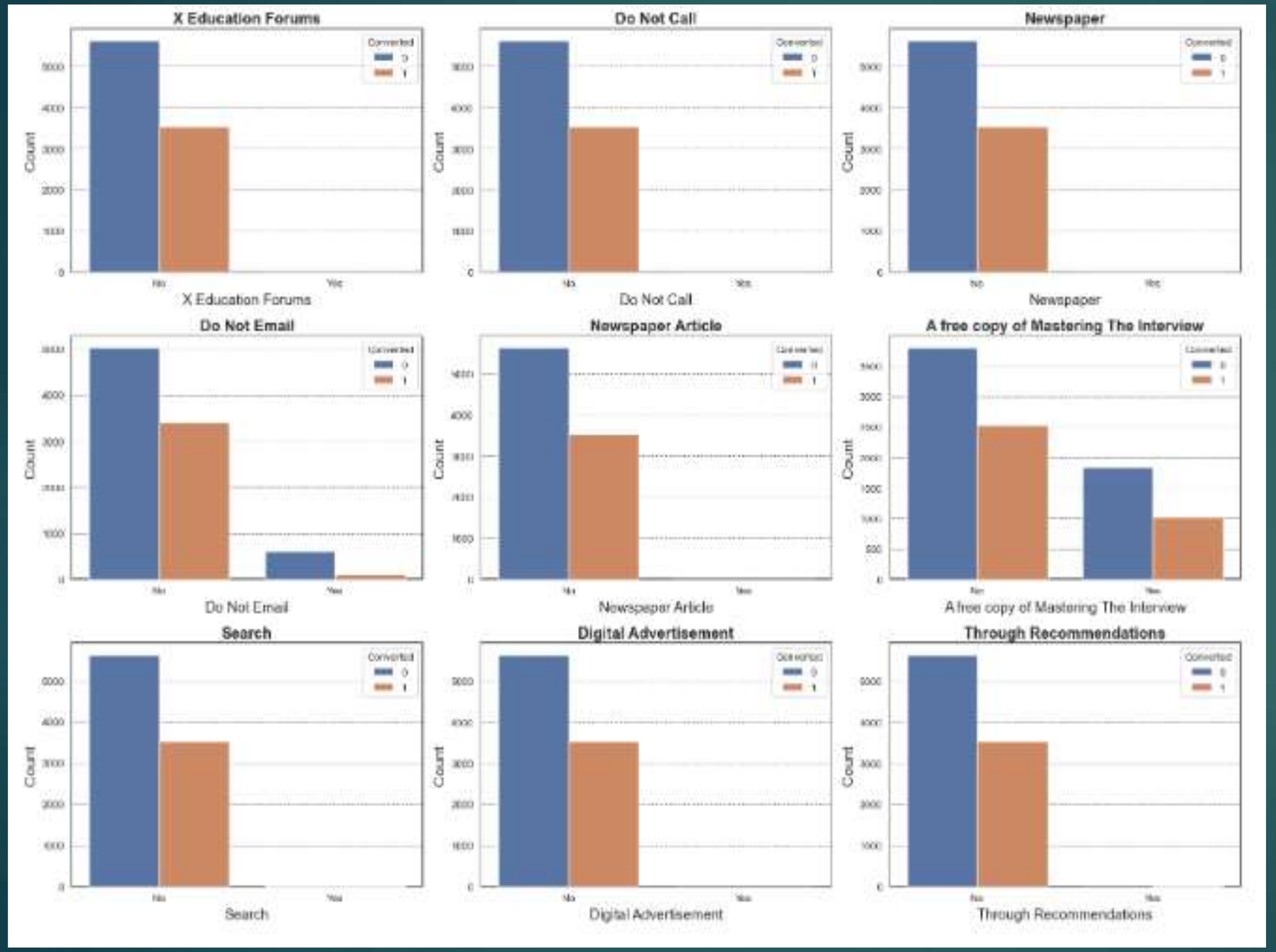
EDA - Lead Source



Conclusion :

- Majority of leads come from ‘Google’ and ‘Direct Traffic’
- Conversion rate of leads from ‘Direct traffic’ is less than overall conversion rate and same for Google is slightly more than overall average.
- Conversion Rate of ‘Welingak website’ and ‘References’ are more. The company should invest more resources into acquiring leads from these sources

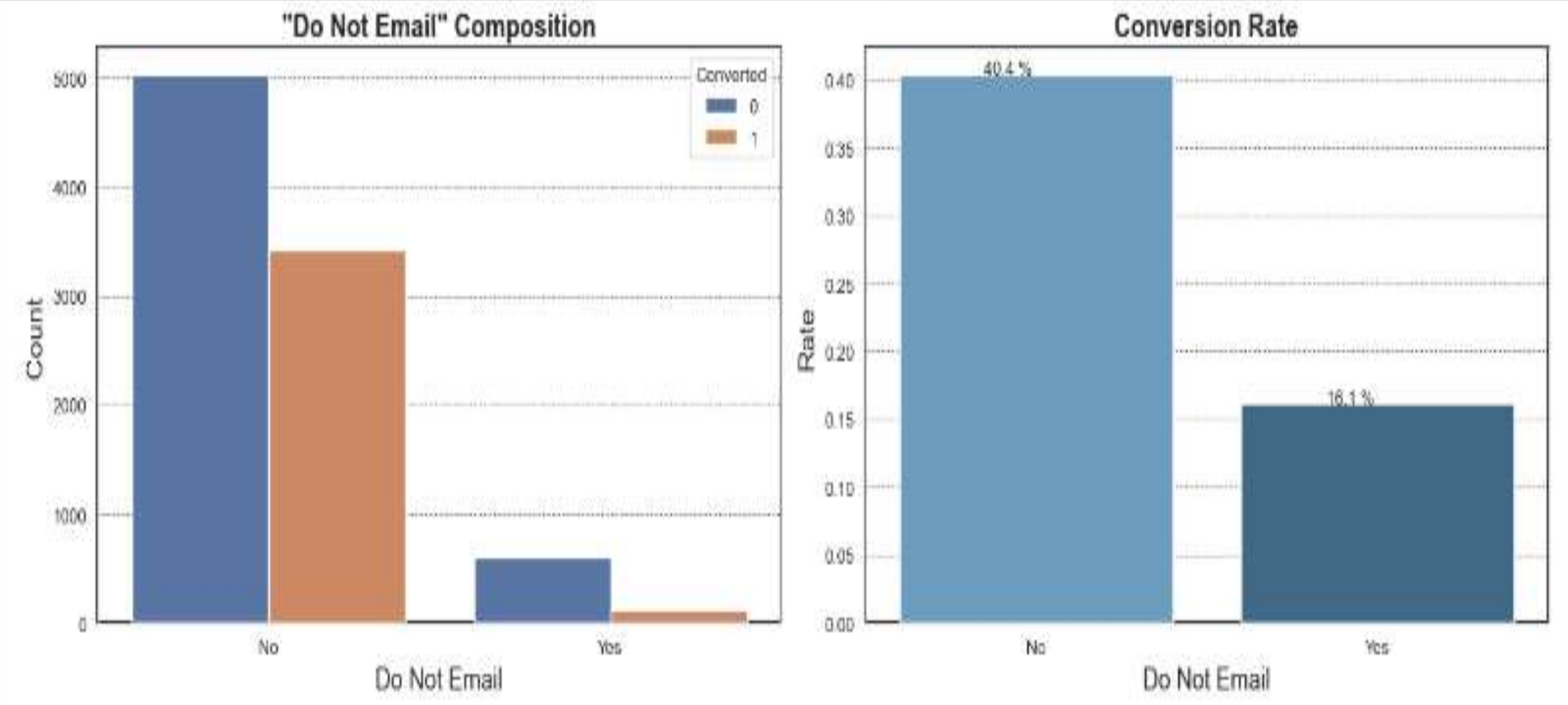
EDA - Visualizing Binary Categorical Variable



Conclusion :

- Chart defines that out of two parameter – ‘Not Convert’ dominant in all variable

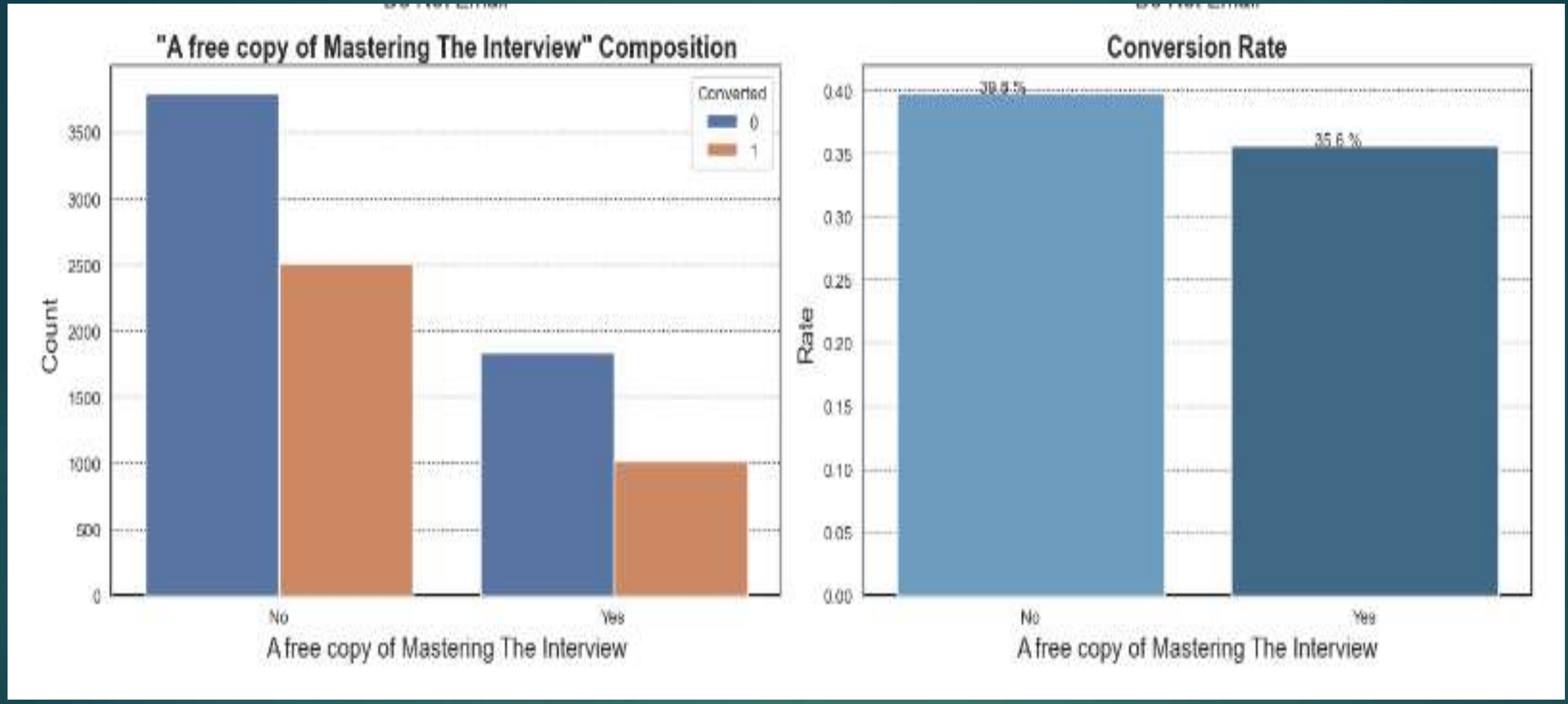
EDA Analysis – 'Do Not Email'



Conclusion :

- Majority of the people want Email (Do Not Email – No) (~80%)
- People who have opted to receive Email has Higher rate of conversion (40%)

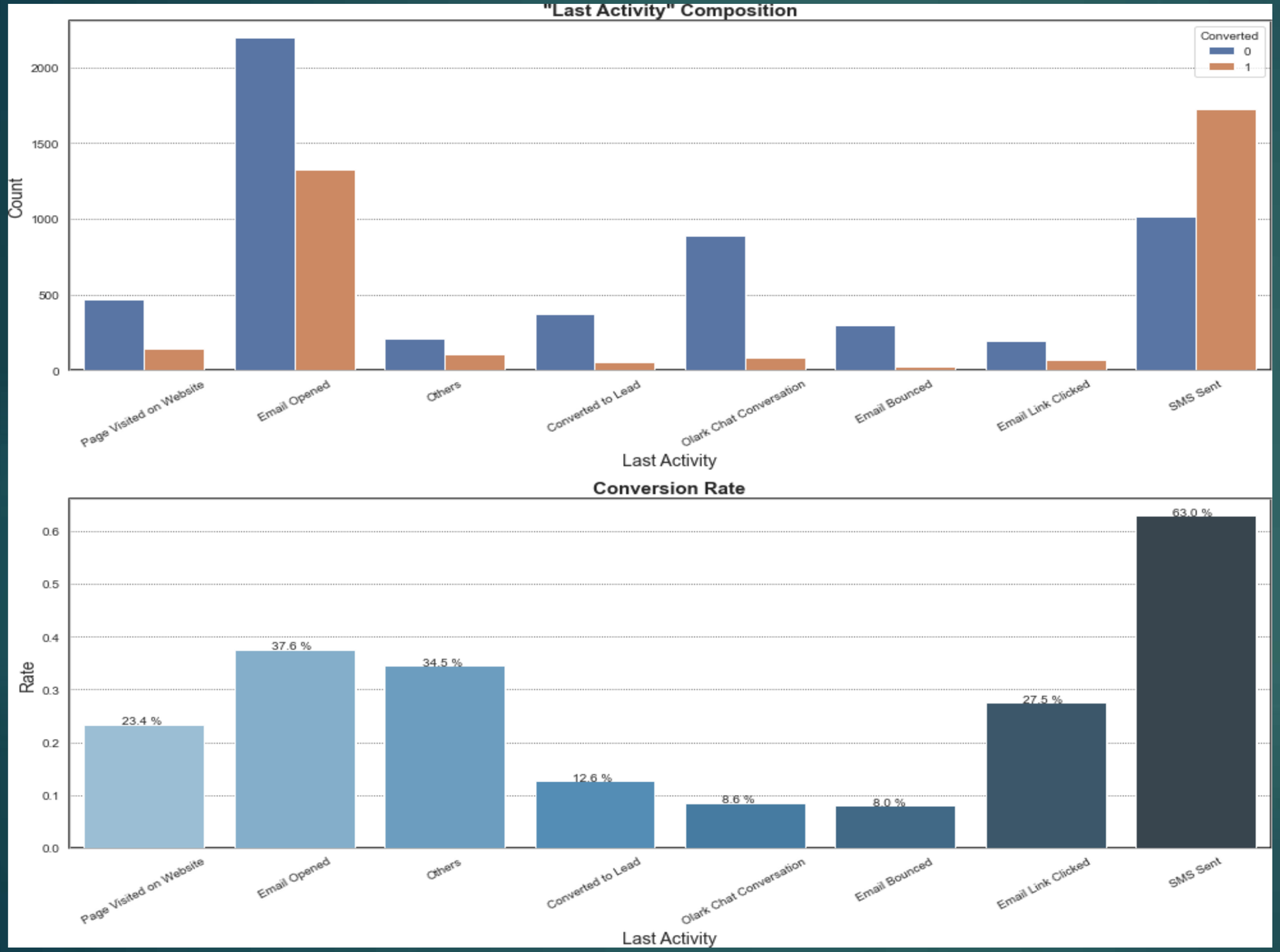
EDA -A Free Copy of Mastering Interview



Conclusion :

- Distributing ‘Free-Copy of Mastering Interview’ has same conversion rate .
- It does not affect the Conversion Rate

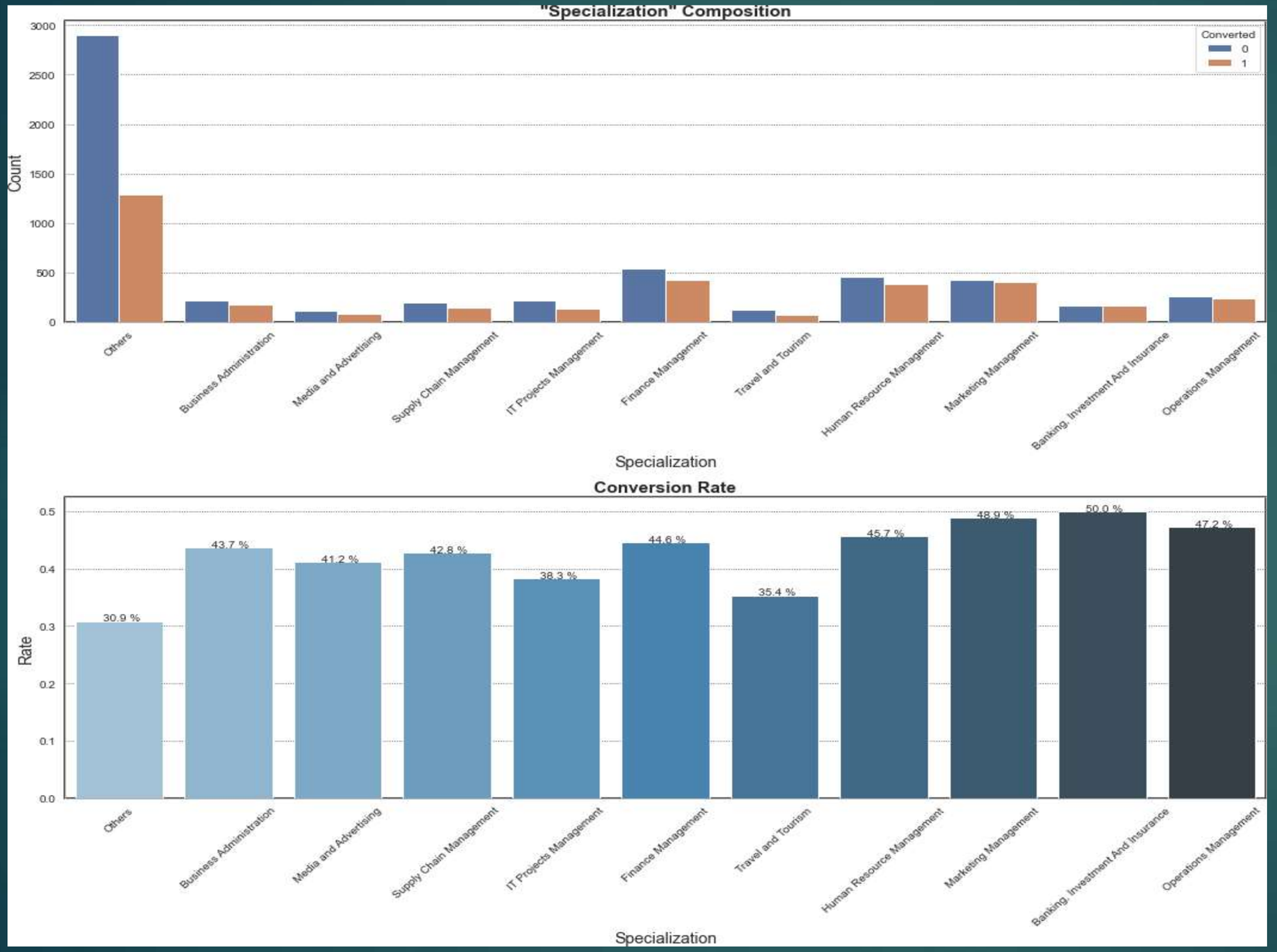
EDA Analysis - Last Activity



Conclusion :

- In Last Activity – 'E-Mail Opened has more leads.
- Conversion rate for leads with the most recent activity as 'SMS Sent' is nearly 60%.

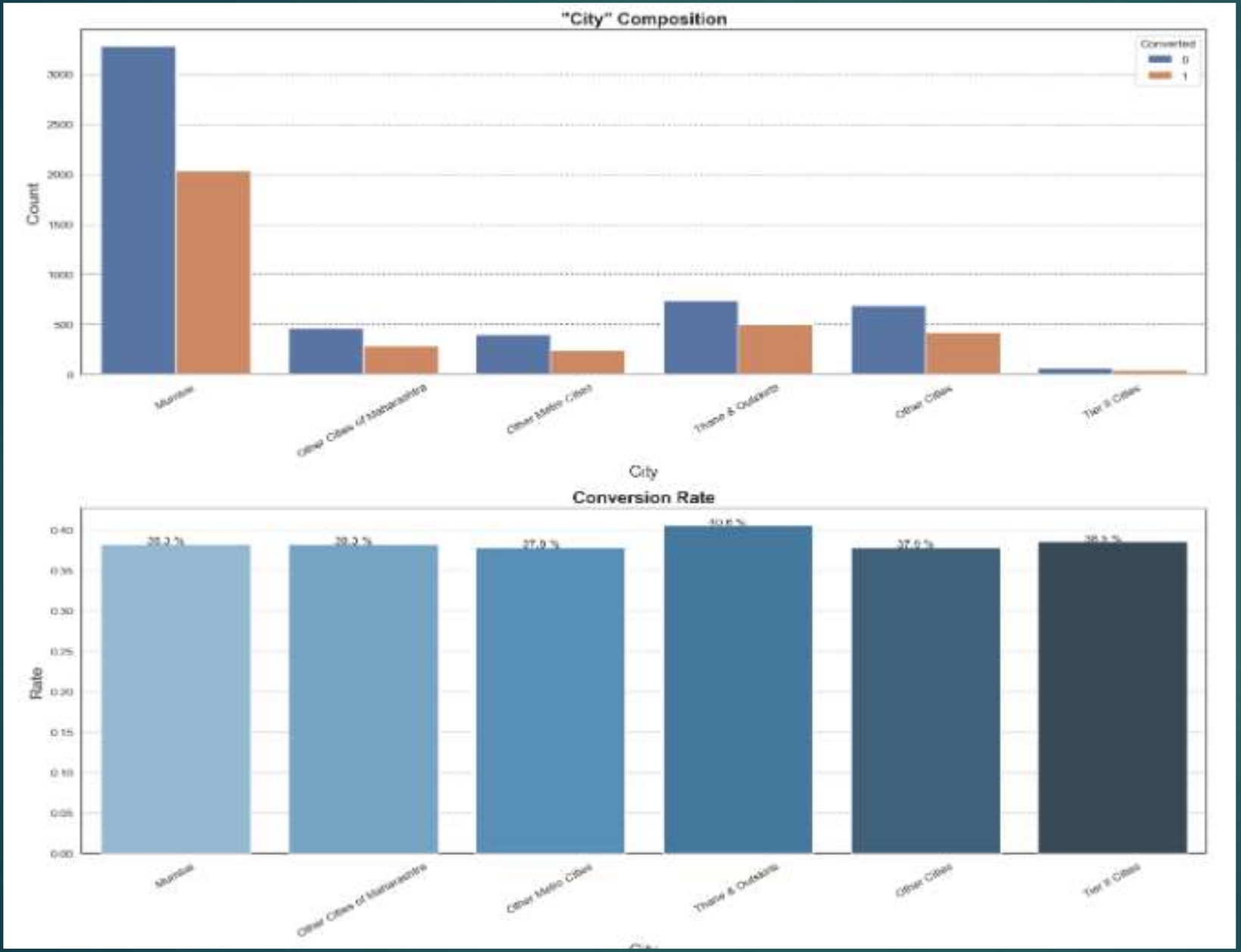
EDA Analysis - Specialization



Conclusion :

- 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 50% Converted
- Followed by ‘ Marketing Management ‘ with 48.9 % for conversion Rate .

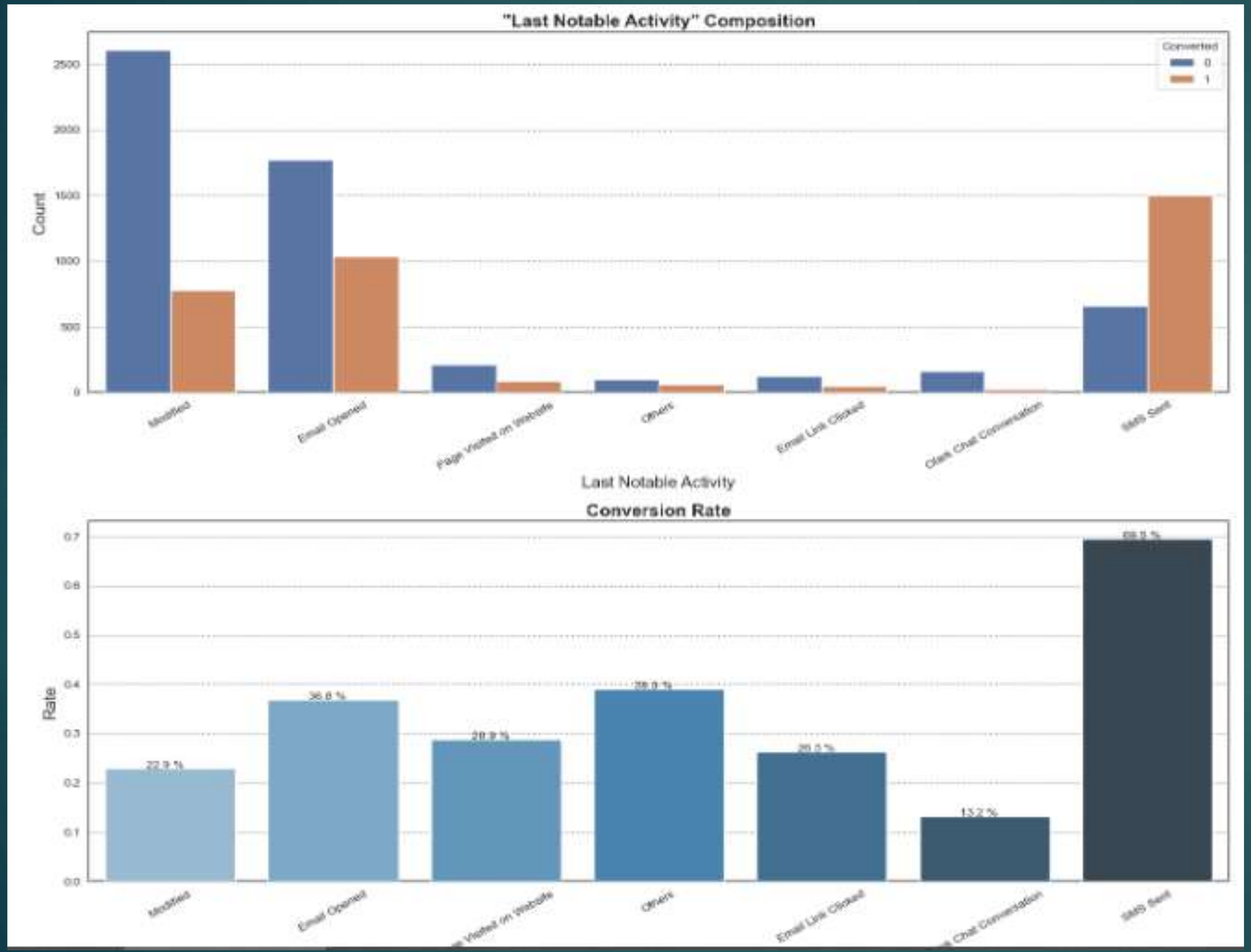
EDA Analysis - City



Conclusion :

- Majority of leads acquired are from Mumbai.
- Highest Conversion rates are from ‘Thane & Outskirts’ with 40 %

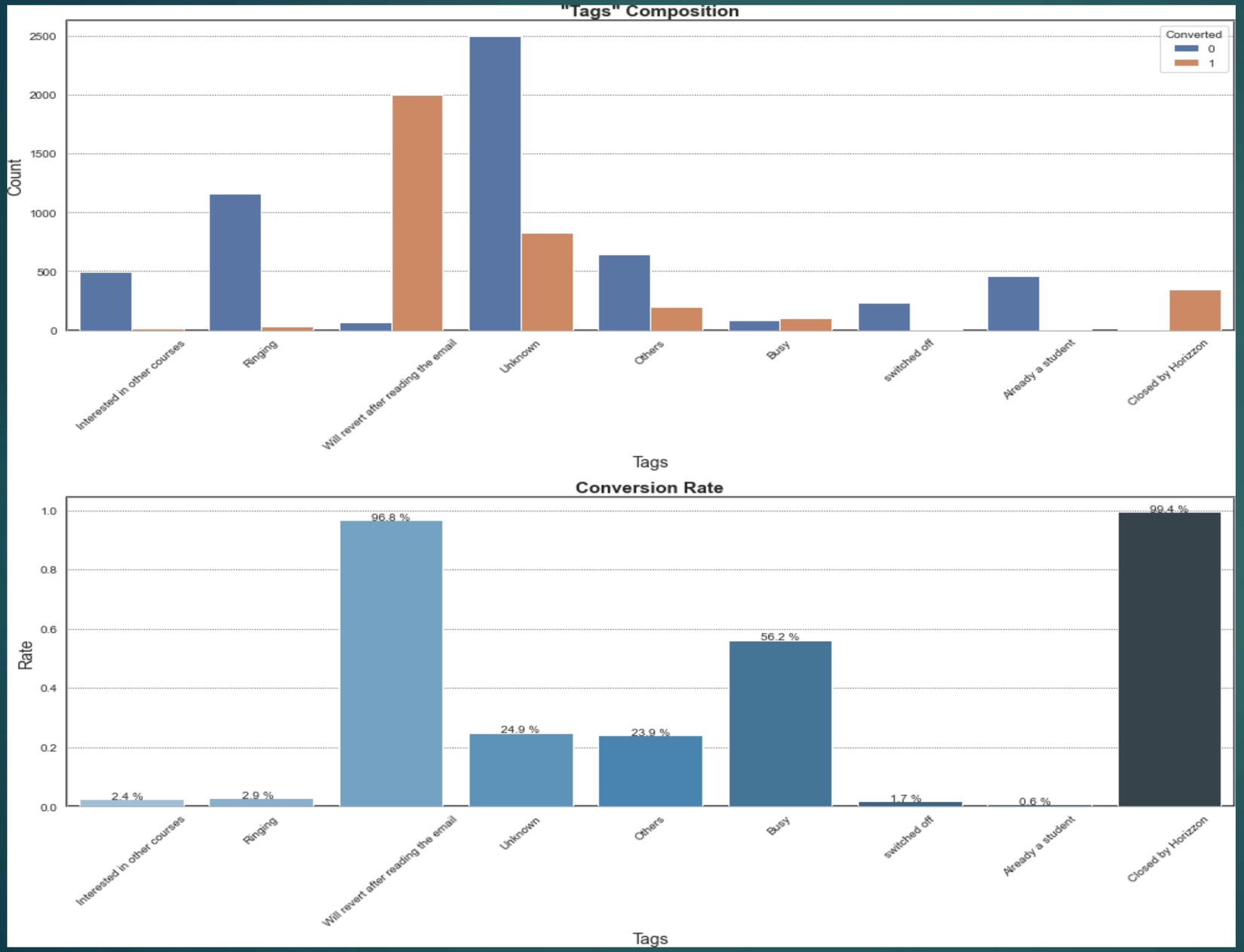
EDA Analysis – Last Notable Activity



Conclusion :

- Majority of leads acquired are from ‘Modified’
- Highest Conversion rates are from ‘SMS Sent’ with 69.5 %

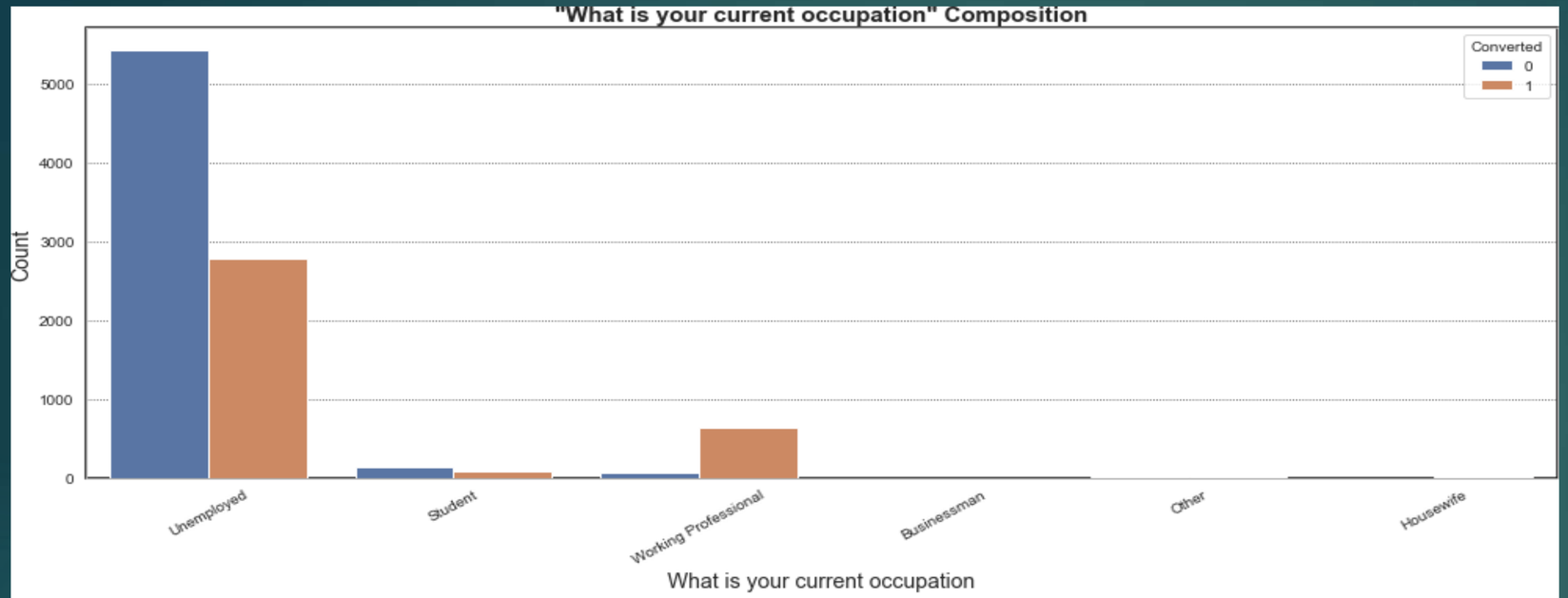
EDA Analysis – Tags

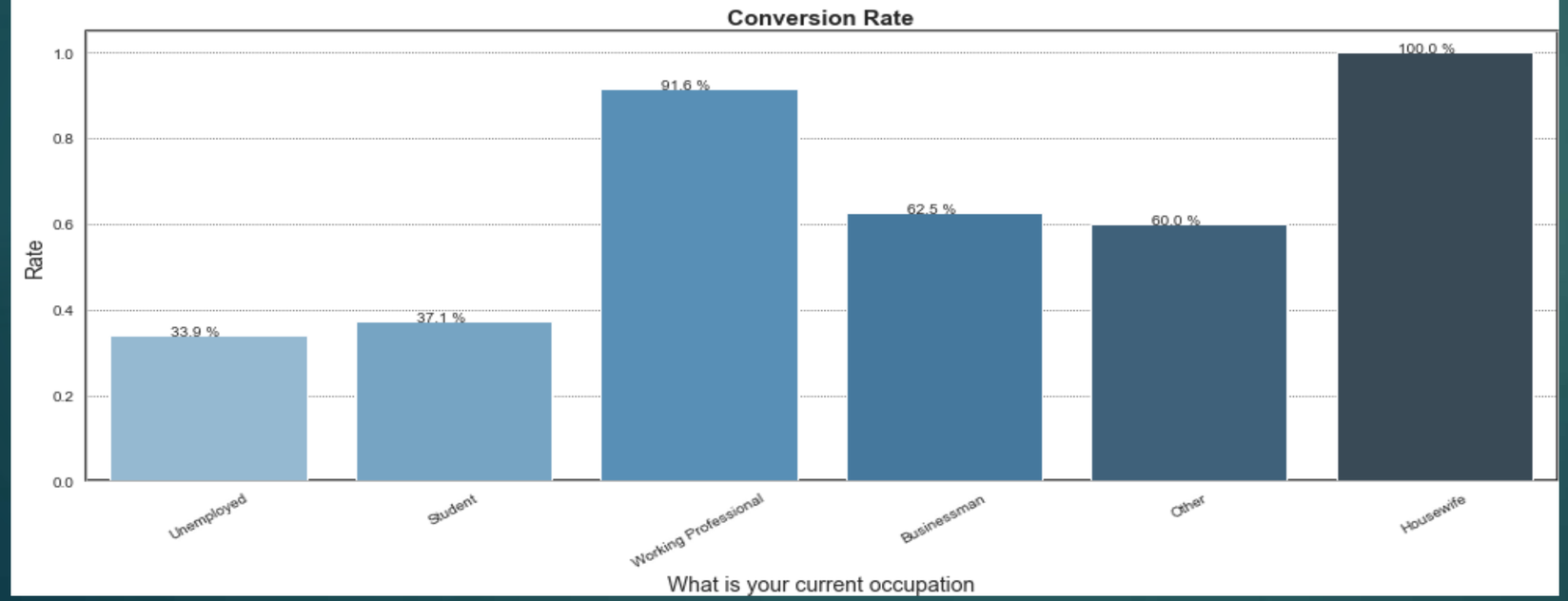


Conclusion :

- Leads with tags/current status, 'Will revert after reading the email' have a very high likelihood of converting.
- 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.

EDA Analysis – Current Occupation

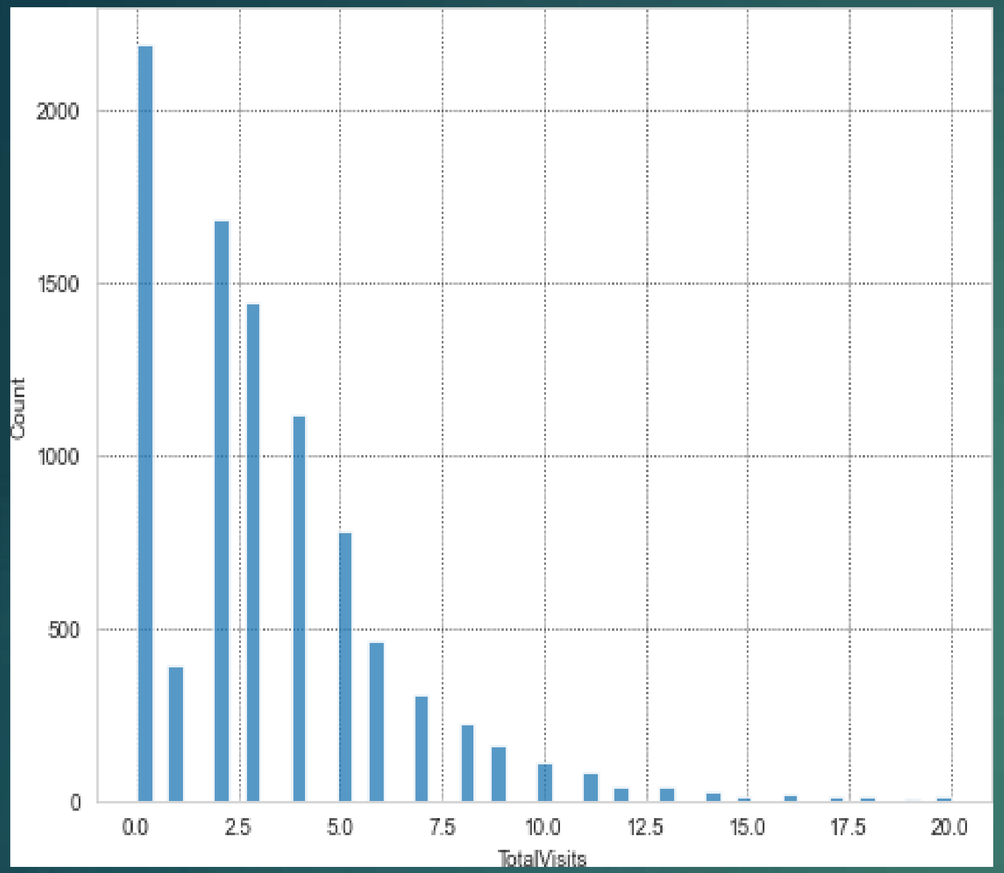


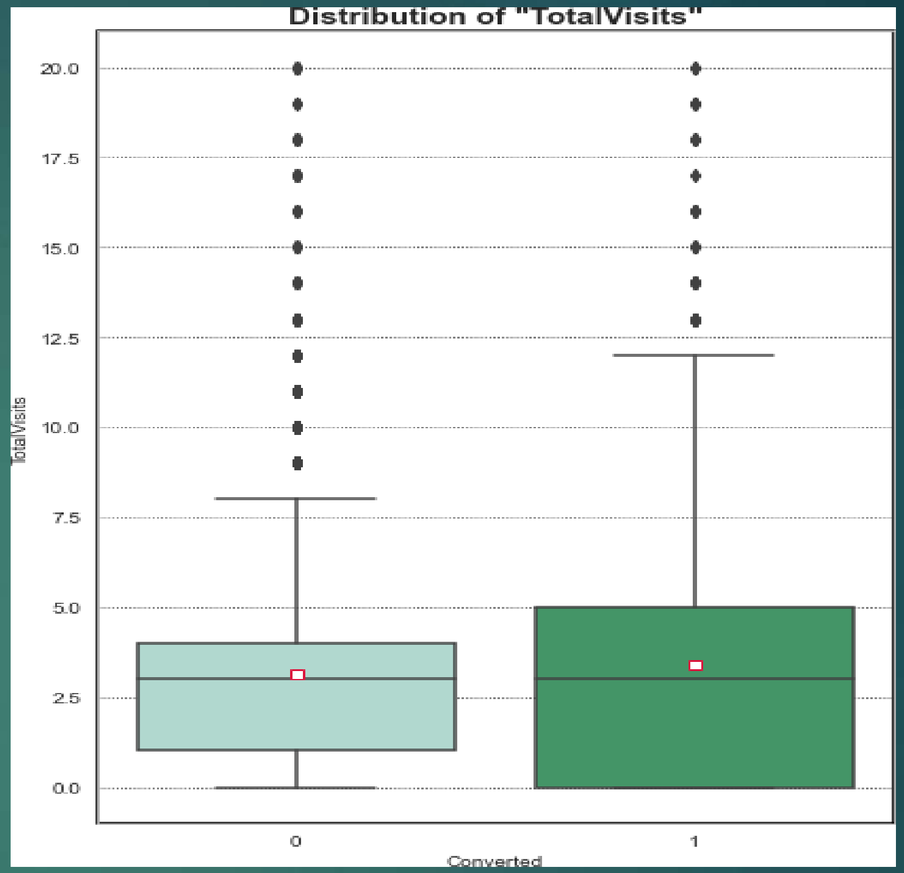


Conclusion :

- ‘Unemployed’ has the majority of leads.
- Housewives are less in numbers, but have 100% conversion rate.
- Followed by Working professionals, Businessmen and Other have high conversion rate.

EDA Analysis – Total Visit

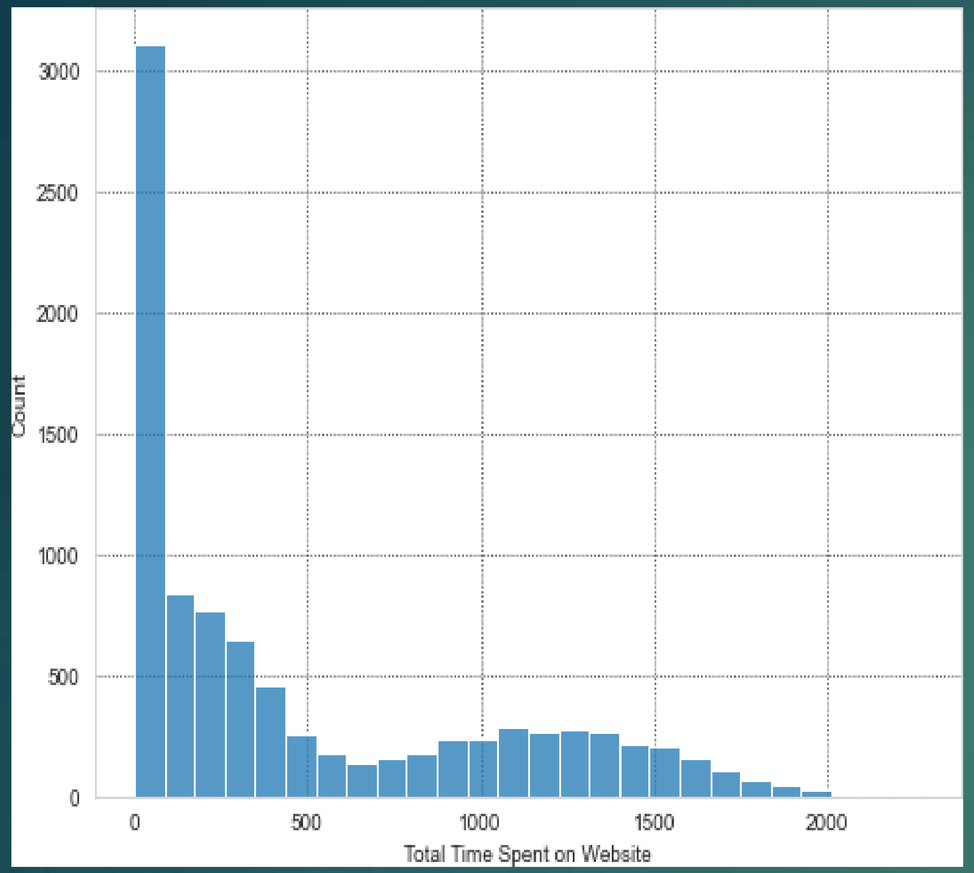


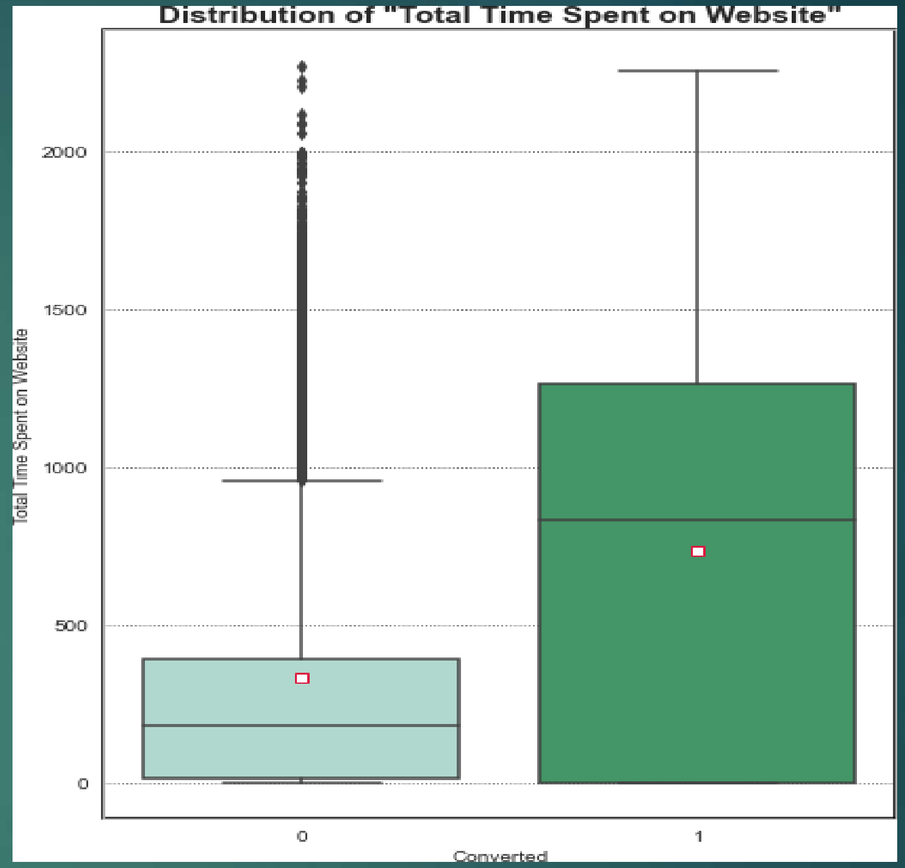


Conclusion :

- The median of the converted is very little high for the total Visits.
- Maximum l'otal visits to the website for majority of people is 7 only,

EDA Analysis – Total Time Spent on Website

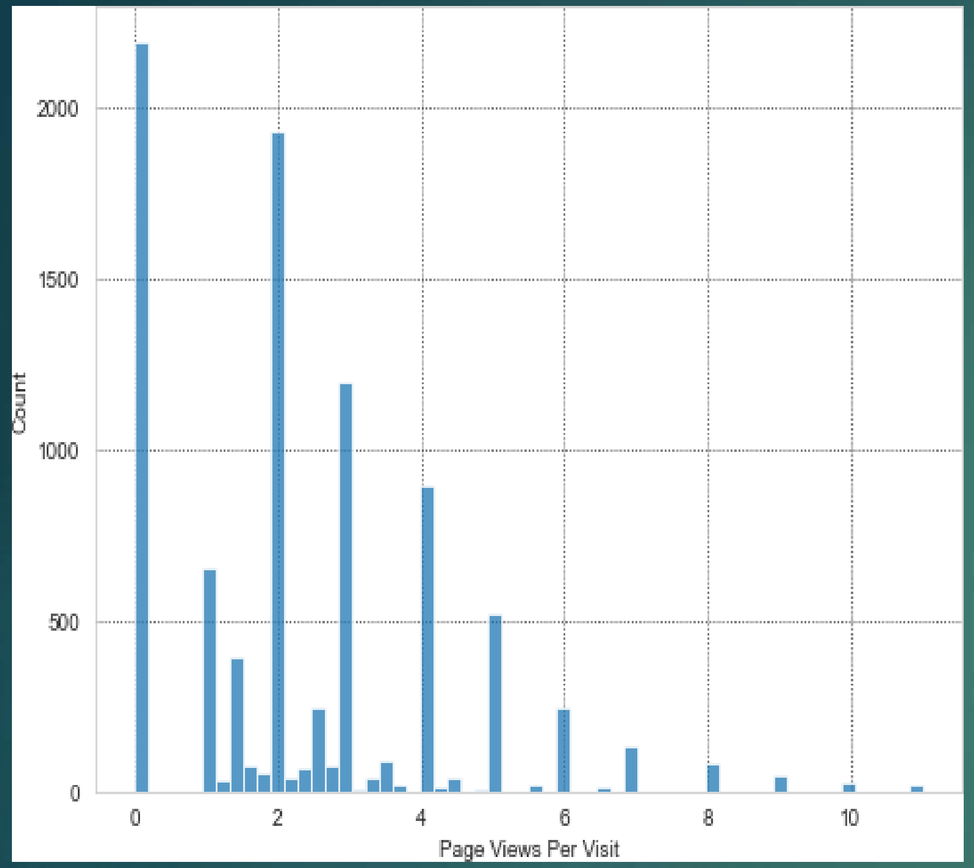


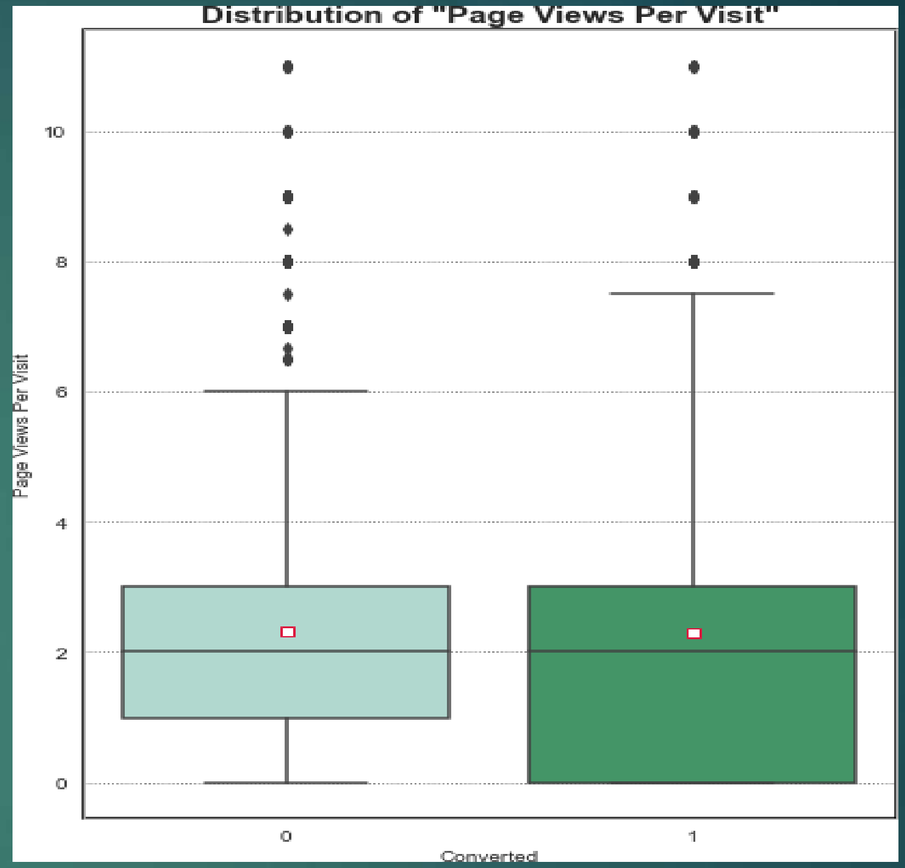


Conclusion :

- 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.

EDA Analysis – Page Views per Visit

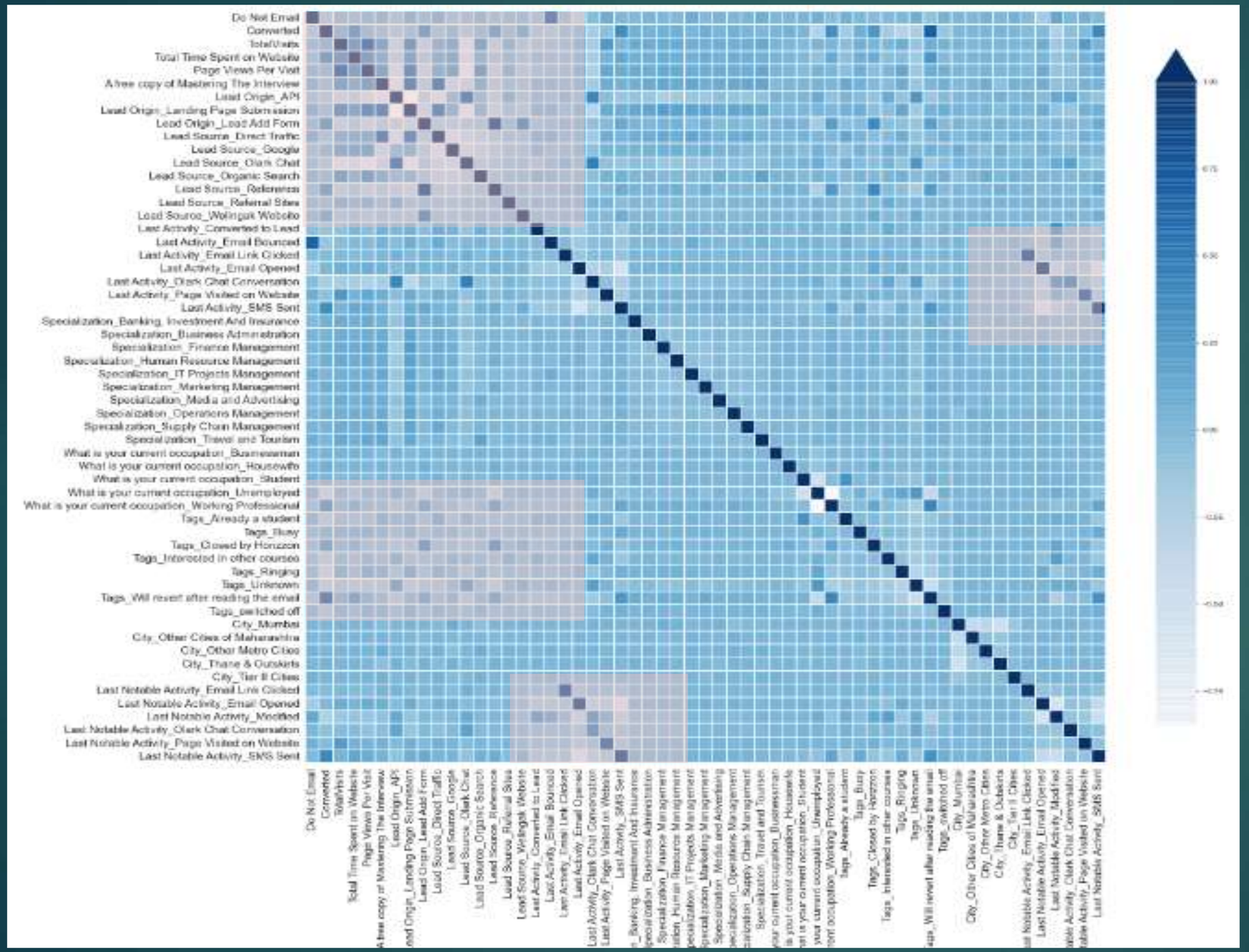




Conclusion :

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

EDA Analysis – Multivariate



Conclusion :

- The Highlighted areas (Red Highlighted) have a high correlation of variables
- While other areas has low (or) Medium Correlation

Model Building – Logistic Regression

Feature Selection

```
[('Do Not Email', True, 1),
 ('TotalVisits', False, 18),
 ('Total Time Spent on Website', True, 1),
 ('Page Views Per Visit', False, 16),
 ('A free copy of Mastering The Interview', False, 38),
 ('Lead Origin_API', False, 28),
 ('Lead Origin_Landing Page Submission', False, 8),
 ('Lead Origin_Lead Add Form', True, 1),
 ('Lead Source_Direct Traffic', False, 30),
 ('Lead Source_Google', False, 34),
 ('Lead Source_Olark Chat', True, 1),
 ('Lead Source_Organic Search', False, 22),
 ('Lead Source_Reference', False, 7),
 ('Lead Source_Referral Sites', False, 20),
 ('Lead Source_Welingak Website', True, 1),
 ('Last Activity_Converted to Lead', False, 3),
 ('Last Activity_Email Bounced', False, 4),
 ('Last Activity_Email Link Clicked', False, 14),
 ('Last Activity_Email Opened', False, 15),
 ('Last Activity_Olark Chat Conversation', True, 1),
 ('Last Activity_Page Visited on Website', False, 5),
 ('Last Activity_SMS Sent', True, 1),
 ('Specialization_Banking, Investment And Insurance', False, 23),
 ('Specialization_Business Administration', False, 27),
 ('Specialization_Finance Management', False, 25),
 ('Specialization_Human Resource Management', False, 37),
 ('Specialization_IT Projects Management', False, 29),
 ('Specialization_Marketing Management', False, 24),
 ('Specialization_Media and Advertising', False, 36),
 ('Specialization_Operations Management', False, 26),
 ('Specialization_Supply Chain Management', False, 13),
 ('Specialization_Travel and Tourism', True, 1),
 ('What is your current occupation_Businessman', False, 39),
 ('What is your current occupation_Housewife', False, 21),
 ('What is your current occupation_Student', False, 33),
 ('What is your current occupation_Unemployed', False, 12),
 ('What is your current occupation_Working Professional', True, 1),
 ('Tags_Already a student', True, 1),
 ('Tags_Busy', False, 10),
 ('Tags_Closed by Horizzon', True, 1),
 ('Tags_Interested in other courses', True, 1),
 ('Tags_Ringing', True, 1),
 ('Tags_Unknown', False, 9),
 ('Tags_Will revert after reading the email', True, 1),
 ('Tags_switched off', True, 1),
 ('City_Mumbai', False, 32),
 ('City_Other Cities of Maharashtra', False, 31),
 ('City_Other Metro Cities', False, 19),
 ('City_Thane & Outskirts', False, 35),
 ('City_Tier II Cities', False, 2),
 ('Last Notable Activity_Email Link Clicked', True, 1),
 ('Last Notable Activity_Email Opened', False, 11),
 ('Last Notable Activity_Modified', True, 1),
 ('Last Notable Activity_Olark Chat Conversation', False, 6),
 ('Last Notable Activity_Page Visited on Website', False, 17),
 ('Last Notable Activity_SMS Sent', True, 1)]
```

Conclusion :

- Selecting the Top 18 features from the total features using RIE 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. ‘Tages _ Closed by Horizzon'
12. ‘Tages _ Interested in other courses'
13. ‘ Tages _ Ringing ‘
14. ‘Tages _ Will revert after reading the email'
15. ‘Tages _ 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:	Thu, 14 Dec 2023	Deviance:	3340.8
Time:	10:35:34	Pearson chi2:	9.48e+03
No. Iterations:	9	Pseudo R-squ. (CS):	0.5842
Covariance Type:	nonrobust		

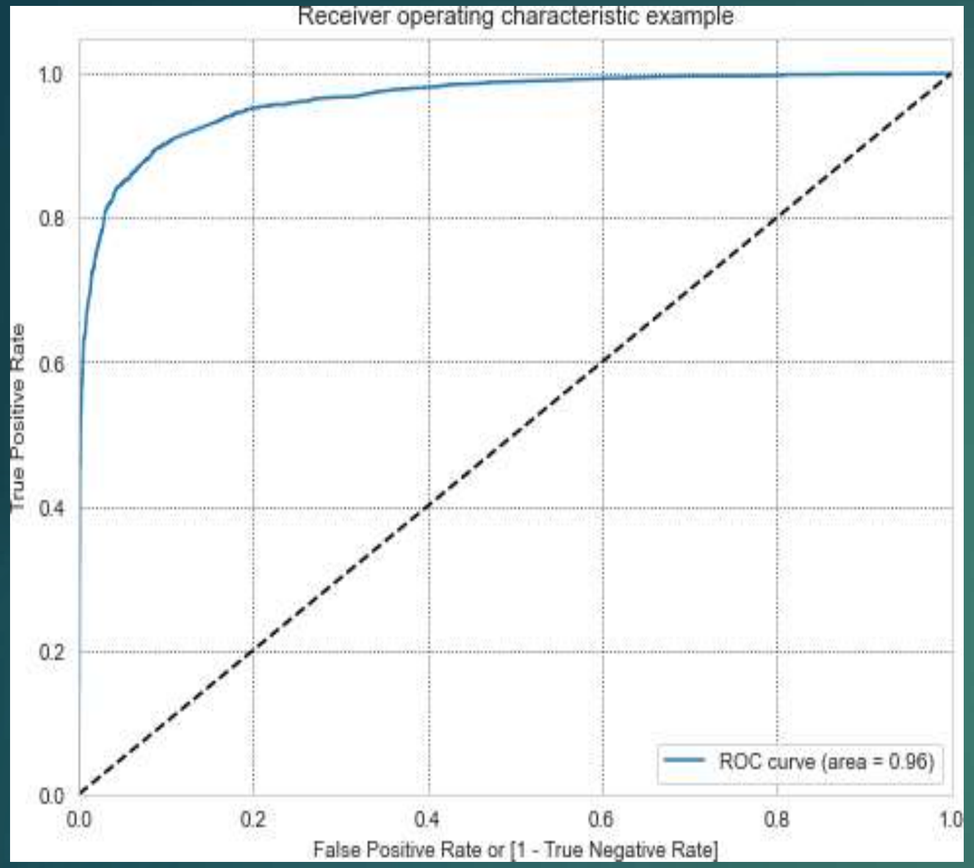
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5492	0.080	-19.392	0.000	-1.706	-1.393
Do Not Email	-1.2212	0.205	-5.946	0.000	-1.624	-0.819
Total Time Spent on Website	1.0977	0.052	21.109	0.000	0.996	1.200
Lead Origin_Lead Add Form	2.2347	0.301	7.430	0.000	1.645	2.824
Lead Source_Olark Chat	1.4031	0.127	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

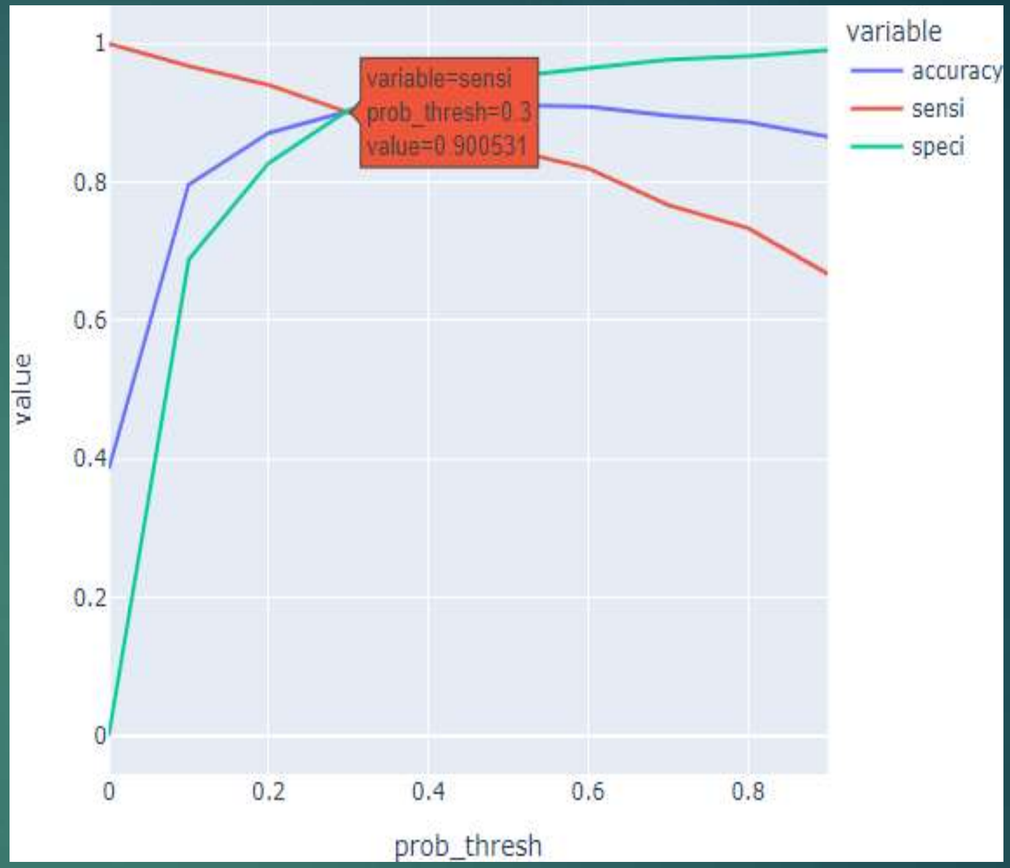
	Features	VIF
2	Lead Origin_Lead Add Form	1.78
12	Tags_Will revert after reading the email	1.75
3	Lead Source_Olark Chat	1.62
14	Last Notable Activity_Modified	1.62
5	Last Activity_Olark Chat Conversation	1.56
6	Last Activity_SMS Sent	1.45
1	Total Time Spent on Website	1.36
7	What is your current occupation_Working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

Conclusion :

- 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.
- VIF are < 2 for all the final set of features which means very low multicollinearity.

Optimal Cut-off

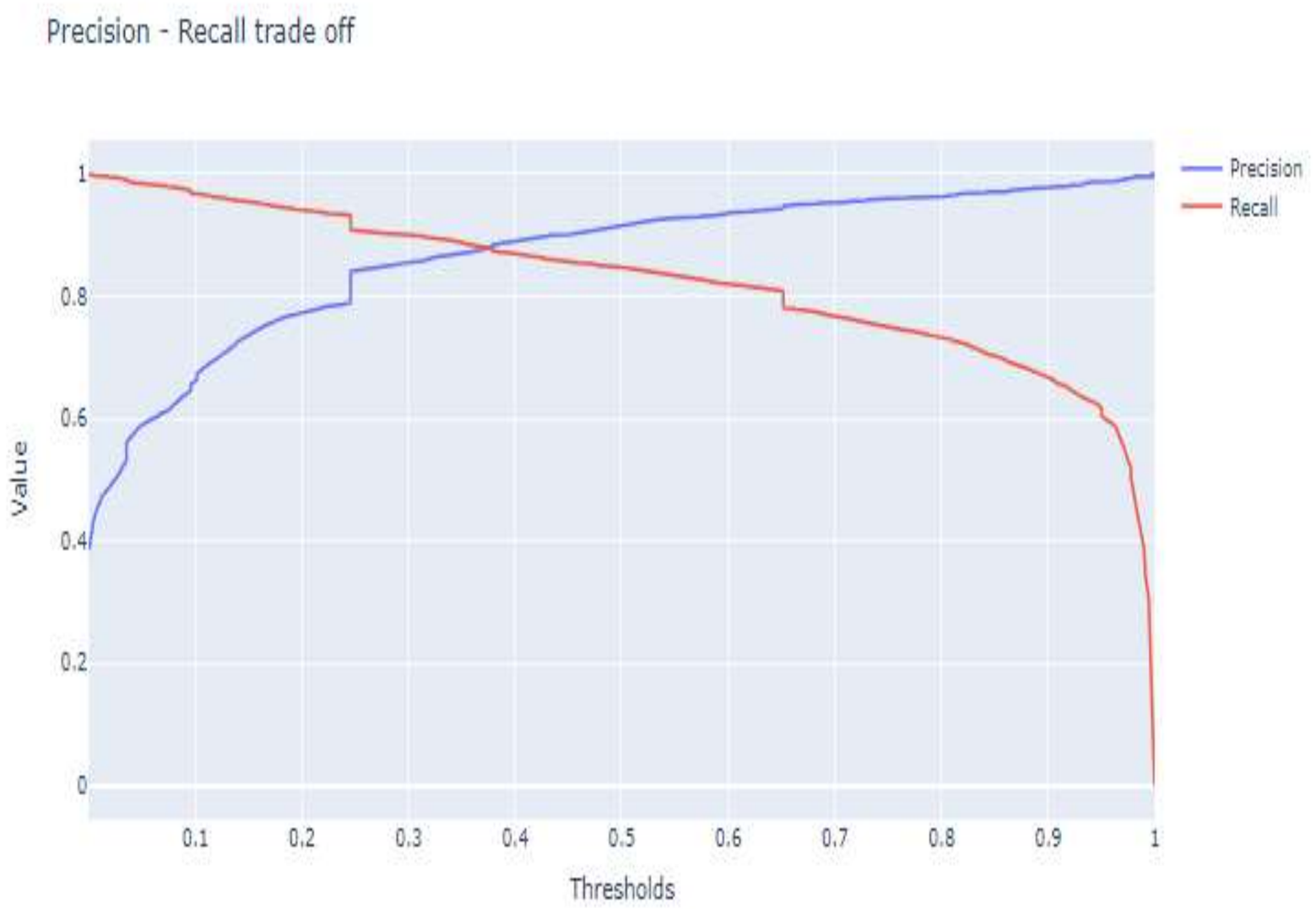




Conclusion :

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

Precision & Recall Trade Cut off



Conclusion :

- 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
- So, we can push the threshold a little lower, say 0.27.

Model Evaluation

Train Data

Optimal Cut Off – 0.3

Precision Recall Trade Off – 0.27

Parameter	Percentage
Accuracy	90 %
Sensitivity	90.05 %
Specificity	90.41 %
AUC	96 %
F1 Score	90 %
Recall	90.41 %
Precision	84.74 %
Positive Predicative Value	84.73 %
Negative Predicative Value	93.72 %

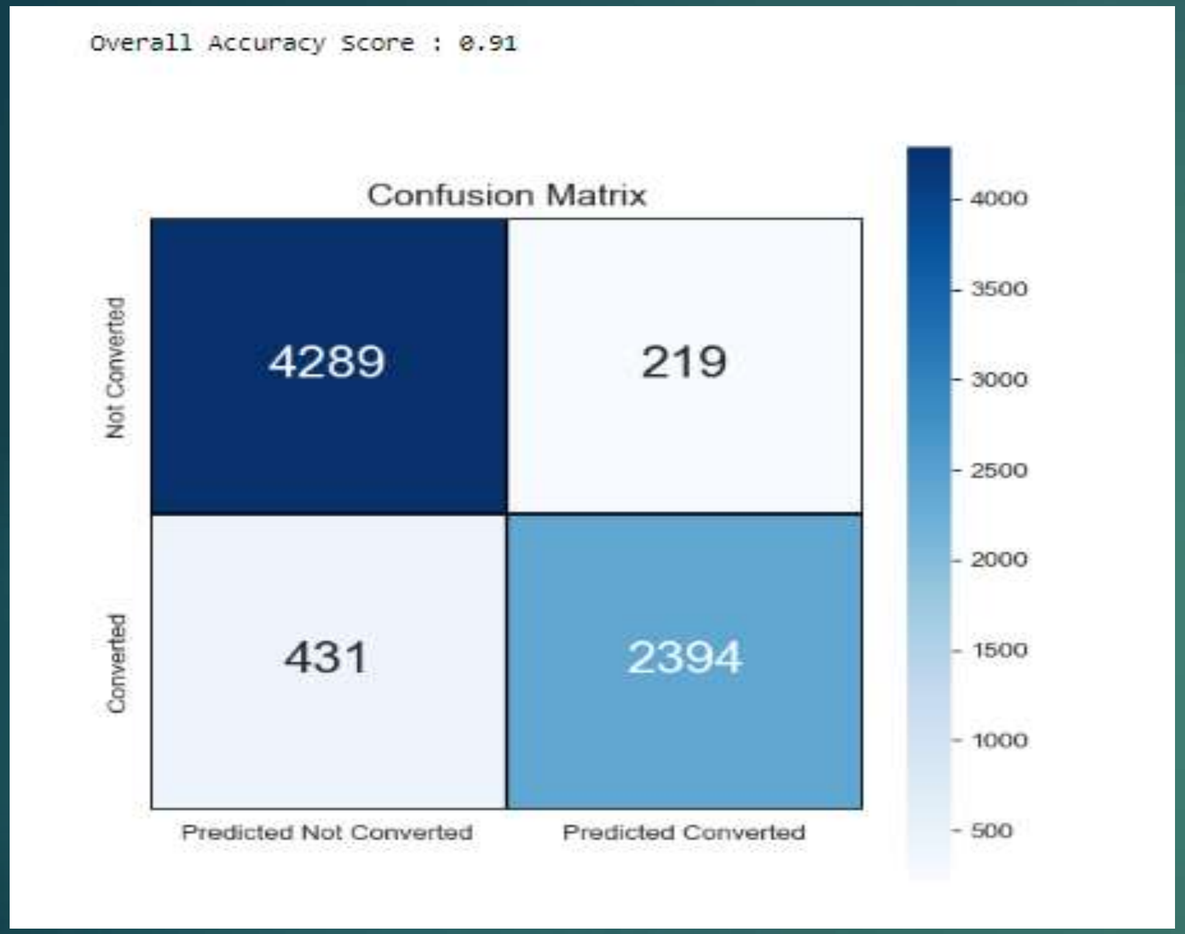
Test Data

Optimal Cut Off – 0.3

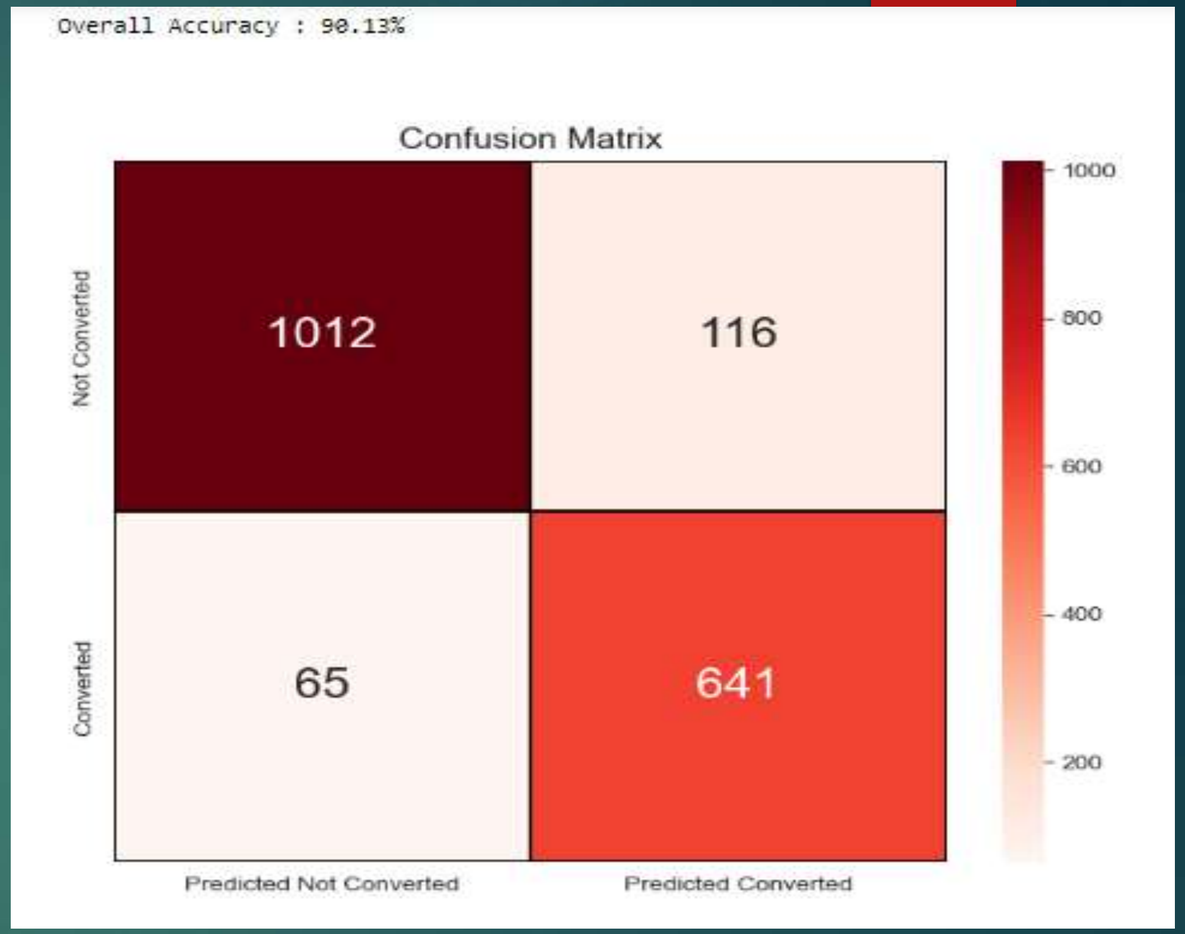
Precision Recall Trade Off – 0.27

Parameter	Percentage
Accuracy	90.13 %
Sensitivity	90.70 %
Specificity	89.71 %
AUC	96 %
F1 Score	90 %
Recall	90.72 %
Precision	84.68 %
Positive Predicative Value	84.67 %
Negative Predicative Value	93.96 %

Confusion Matrix

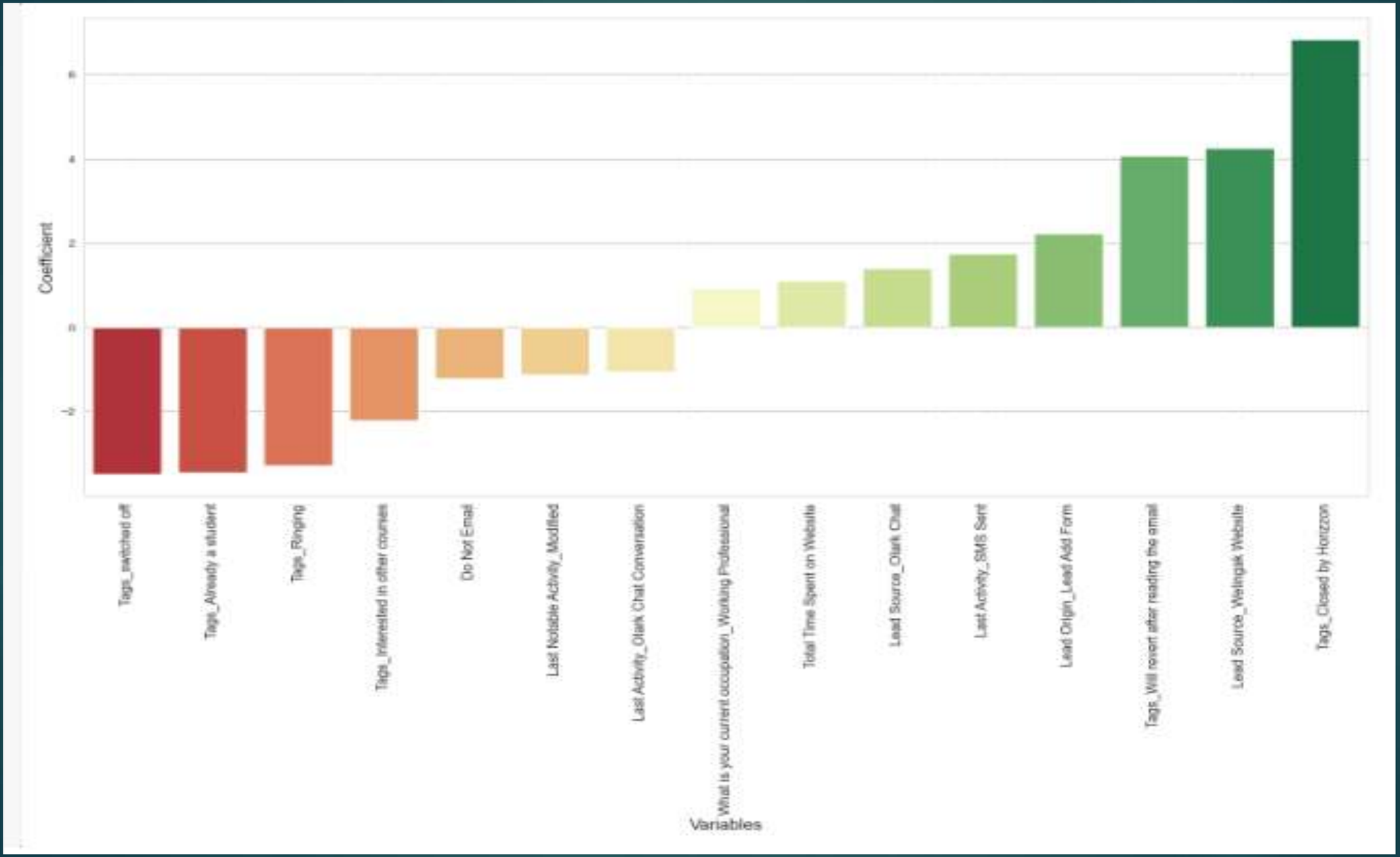


Train Dataset Predication : Confusion Matrix



Test Dataset Predication : Confusion Matrix

Variable vs Coefficient plot



Feature Importance

-. 'Tags _ Ringing': If the current status / tag is 'Ringing', then the log odds decrease by 3.29.
-. 'Tags _ Already a student': If the current status / tag is 'Already a student', then the log odds decrease by 3.44.
-. 'Tags _ switched off': If the current status / tag is 'switched off', then the log odds decrease by 3.48.
-. '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.
-. 'Lead Source _ Welingak Website': If the Lead source is Welingak website then the log odds increase by 4.26.
-. 'Tags _ Closed by Horizzon': If this variable is True or 1, then the log-odds go up by 6.84.

Recommendation

-. 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

-. '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

-. 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.