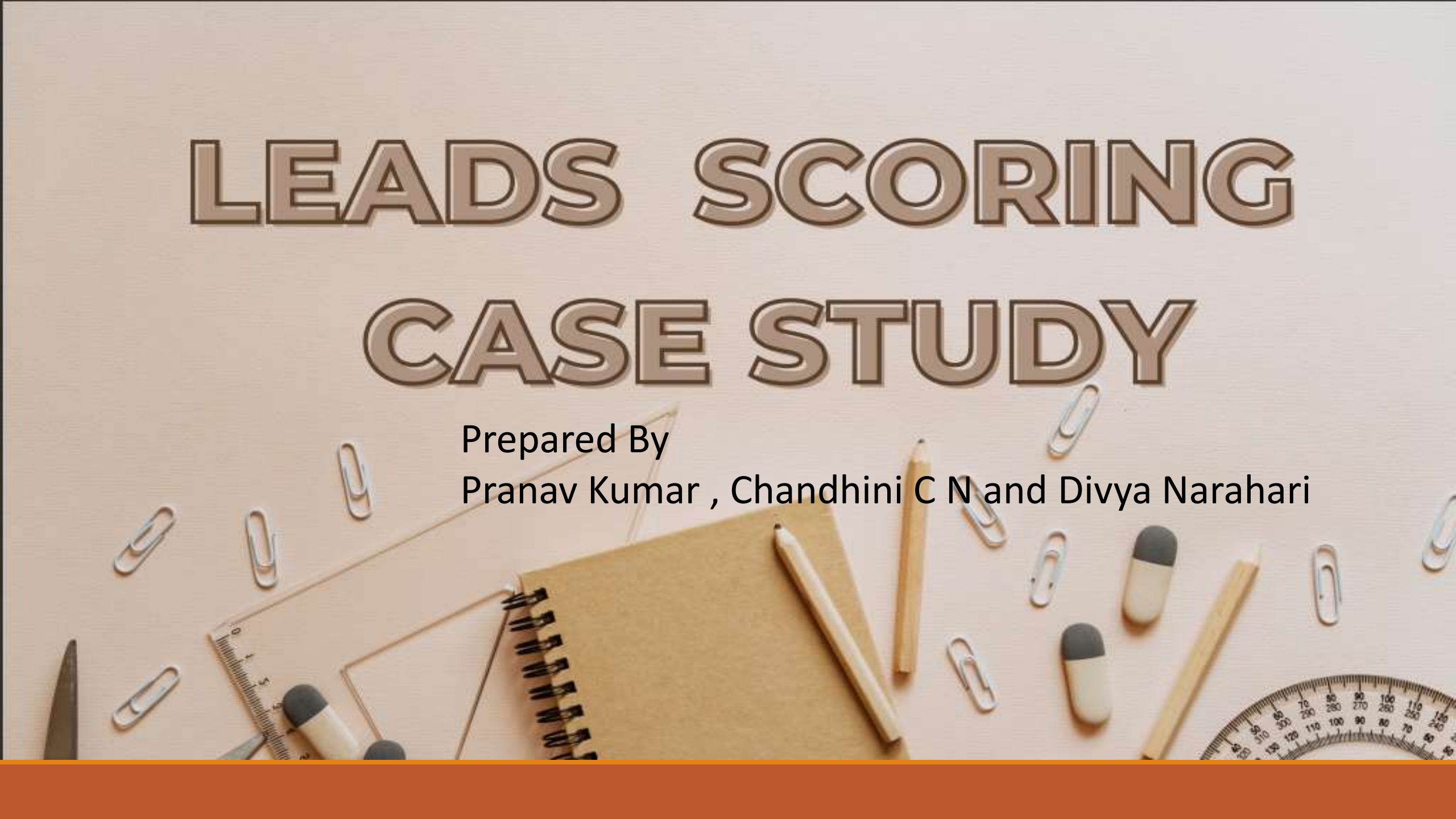
LEADS SCORING CASE STUDY

Prepared By
Pranav Kumar , Chandhini C N and Divya Narahari



Problem Statement

- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google . And the leads are being generated by making those people fill up a form providing their email address or phone number who lands in the website .
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Approximate lead conversion rate at X education is around 30%.
- Since the lead conversion rate is very poor, X Education company wants to make the process more efficient and they wishes to identify the most potential leads, also known as ‘Hot Leads’.
- Their sales team want to know these potential sets of leads , inorder to have more on focus on the potential leads and increase the conversion rate .

Objectives Of the study

Some of the objectives of this case study:

- 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. Lead score with high rate has higher conversion rate and vice versa.
- To help X education company to select the promising leads and convert them in to their clients .
- The Company requires the model should be able to adjustable to the mentioned requirement changes in the future

Steps Involved

- Data Cleaning

Loading Data Set, Understanding and Cleaning the data

- Exploratory Data Analysis

Checking imbalance , univariate , bivariate analysis

- Data Preparation

Dummy Variable creation , feature Scaling, Train-Test split

- Model Building

RFE for top 15 feature , Manual feature reduction and Finalizing the model

- Model Evaluation
- Confusion Matrix, cutoff selection

- Prediction on Test Data

Compare Train data and test dataset, identify top features

- Recommendation

Suggest the best features to identify the Hot leads in order to focus more on the areas of improvement.

Data Cleaning

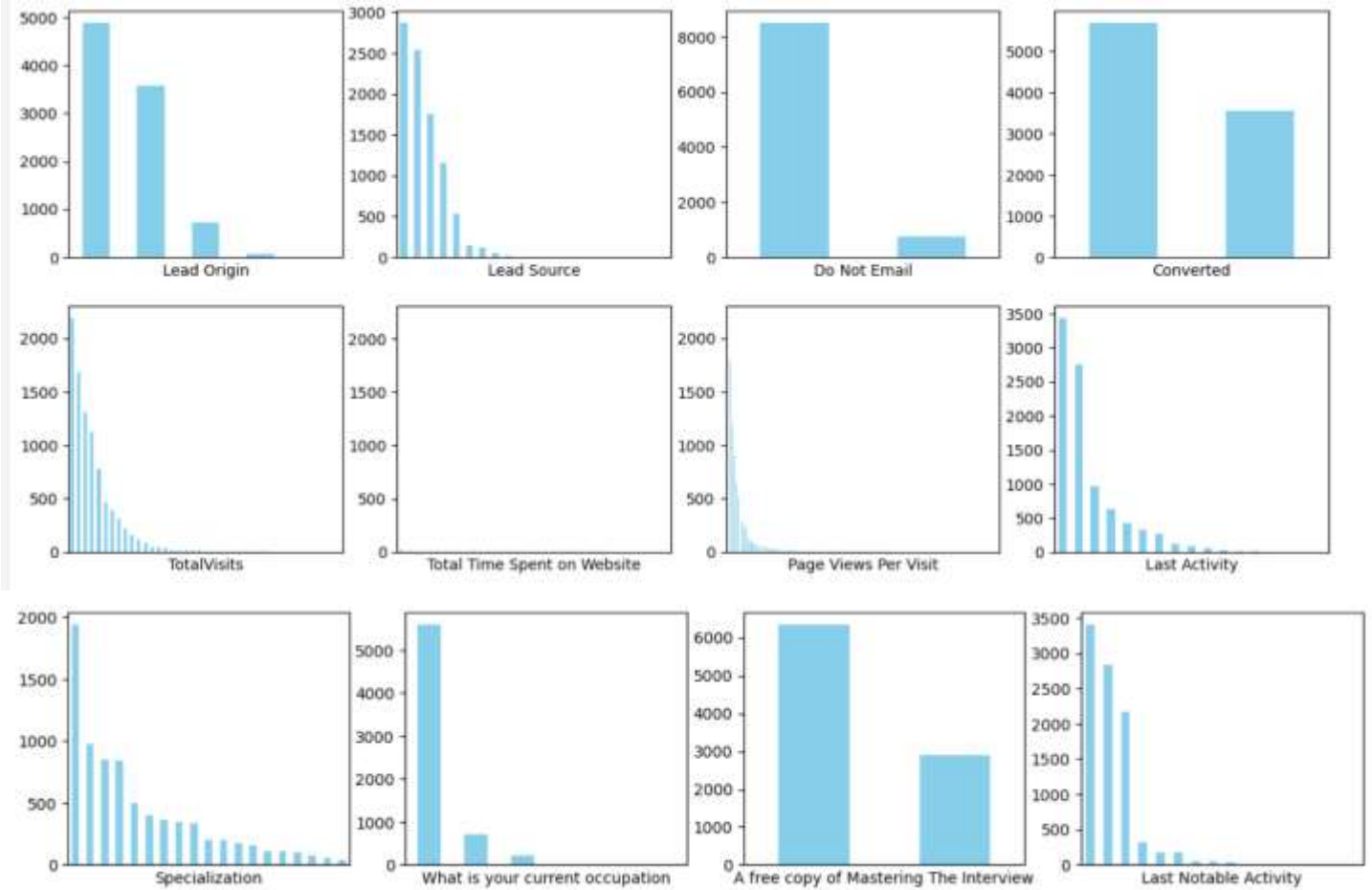
Missing Value Treatment

- Missing Value treatment has been done with the threshold cutoff as 3000.
- Based on further analysis , the we have dropped the variables which had more “Select “ options
- Also dropped the columns like [Do not call, Search, Magazine , Newspaper, Newspaper Article, Digital Advertising etc] based on checking the normality of distribution .

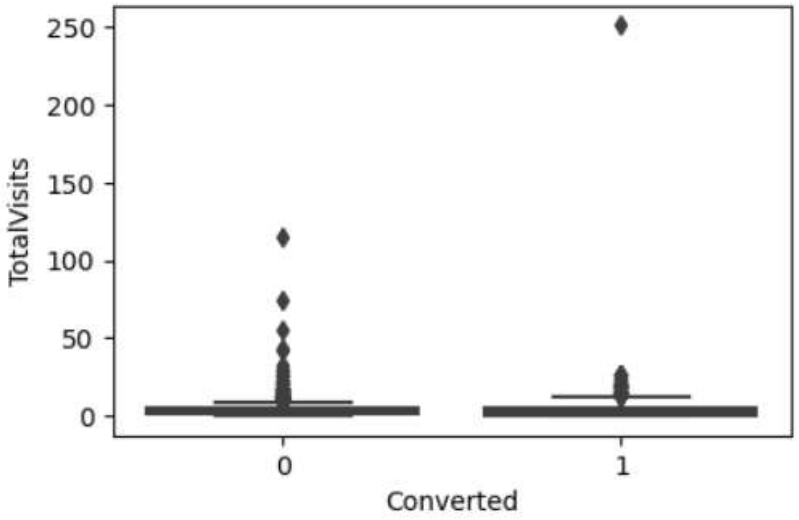
Outlier Treatment

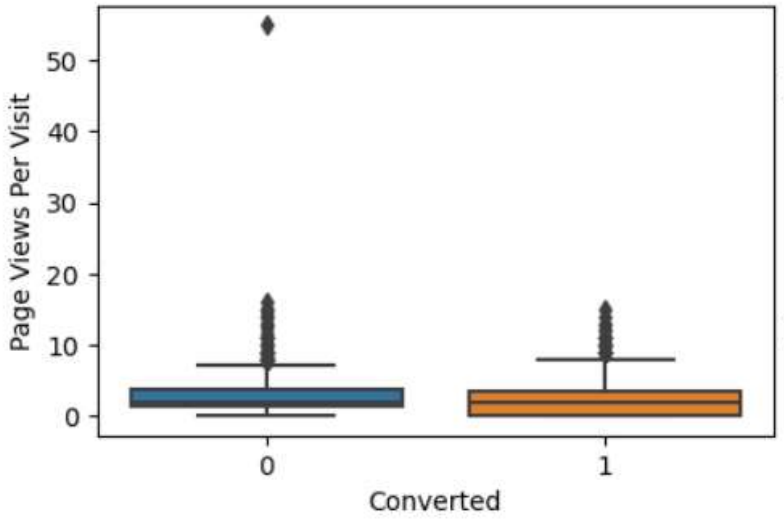
- Outliers are present in the data set in Total visit, Page views per visit, Total Time spent on visit .
- And outliers from the above columns has been cleared before building the model.

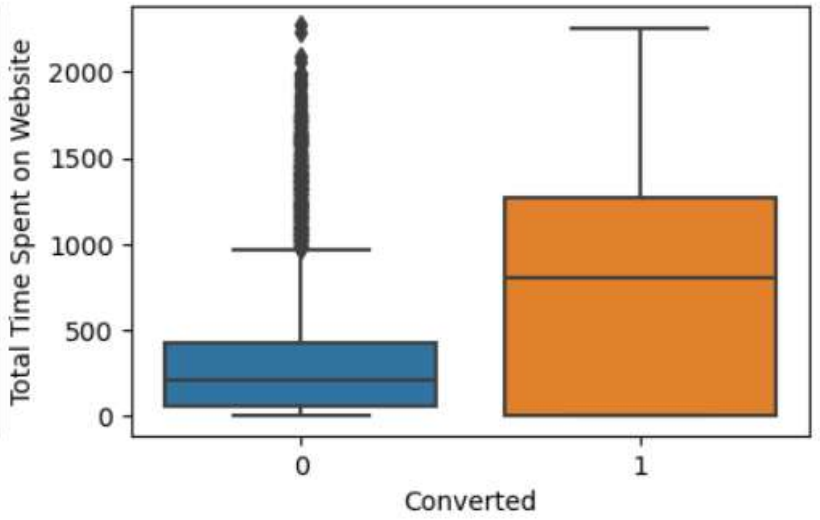
Normality of Distribution of variables



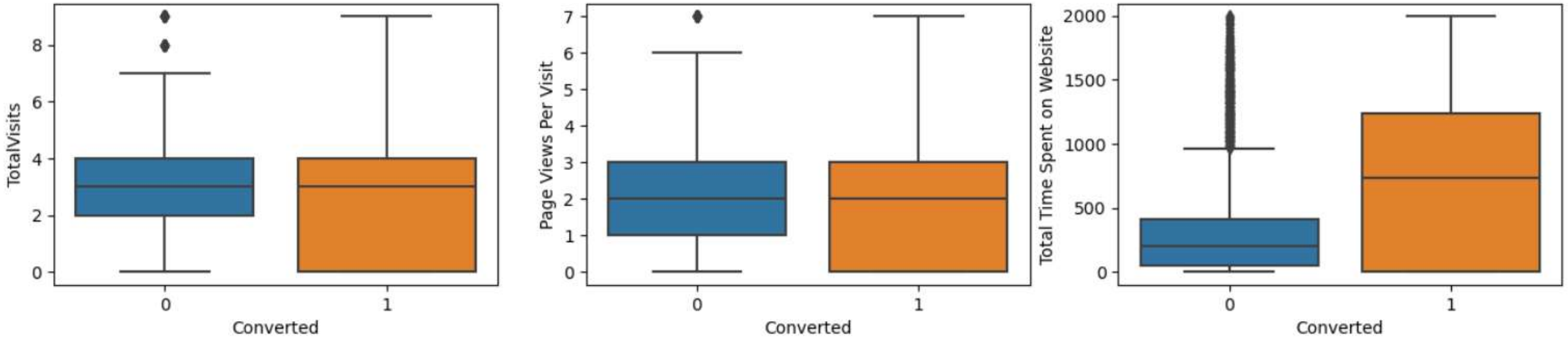
Outliers


TotalVisits
250
200
150
100
50
0
0
1
Converted

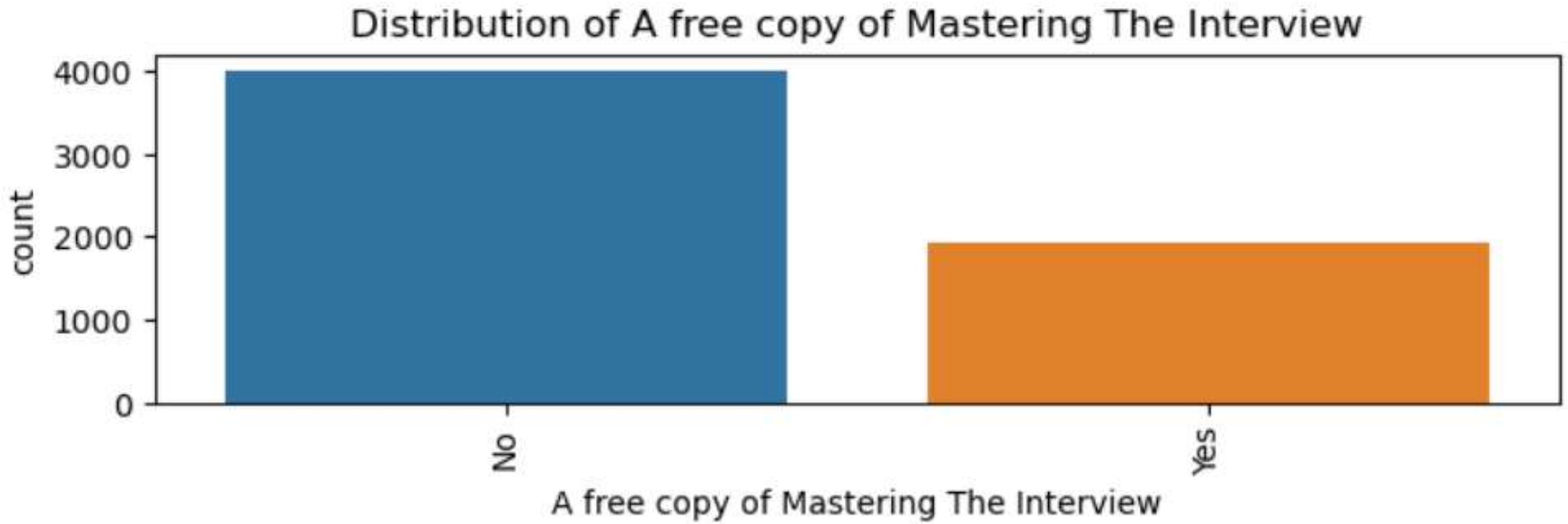

Page Views Per Visit
50
40
30
20
10
0
0
1
Converted


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

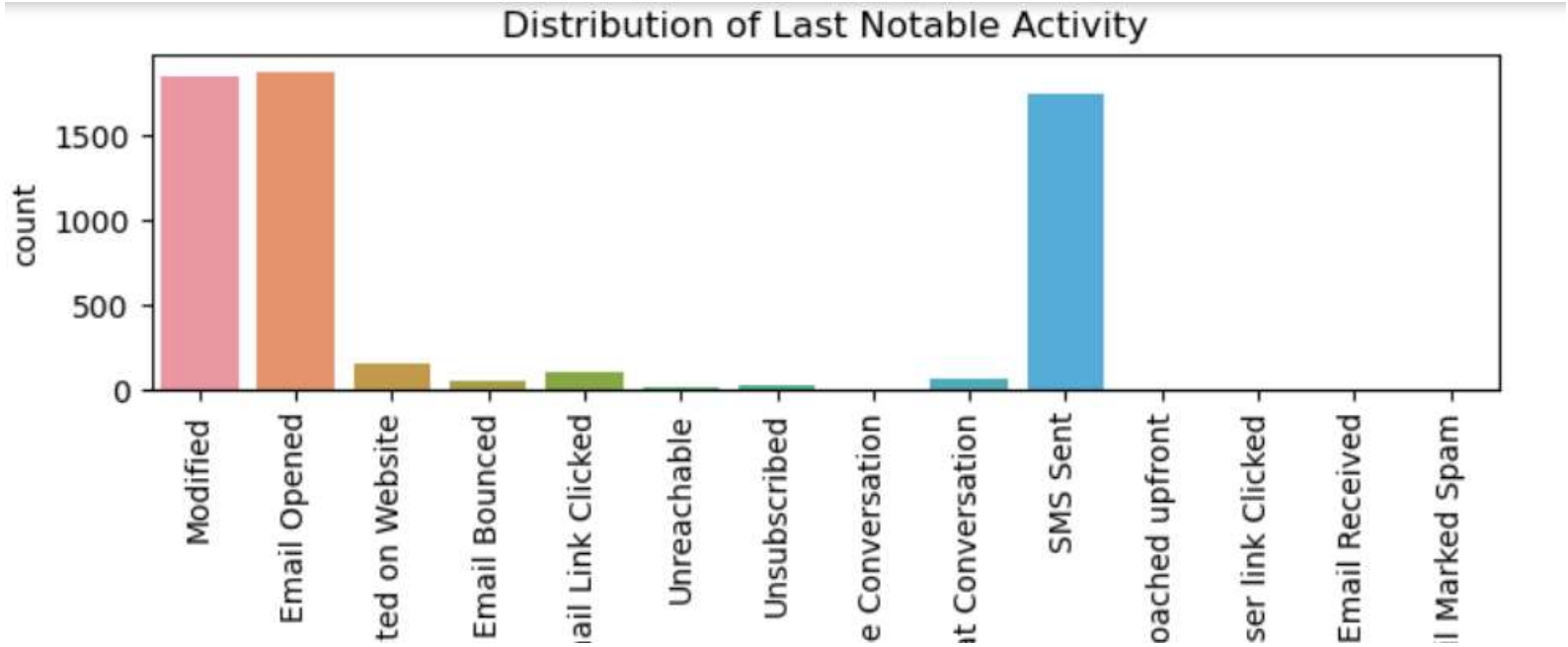
After Outlier Treatment



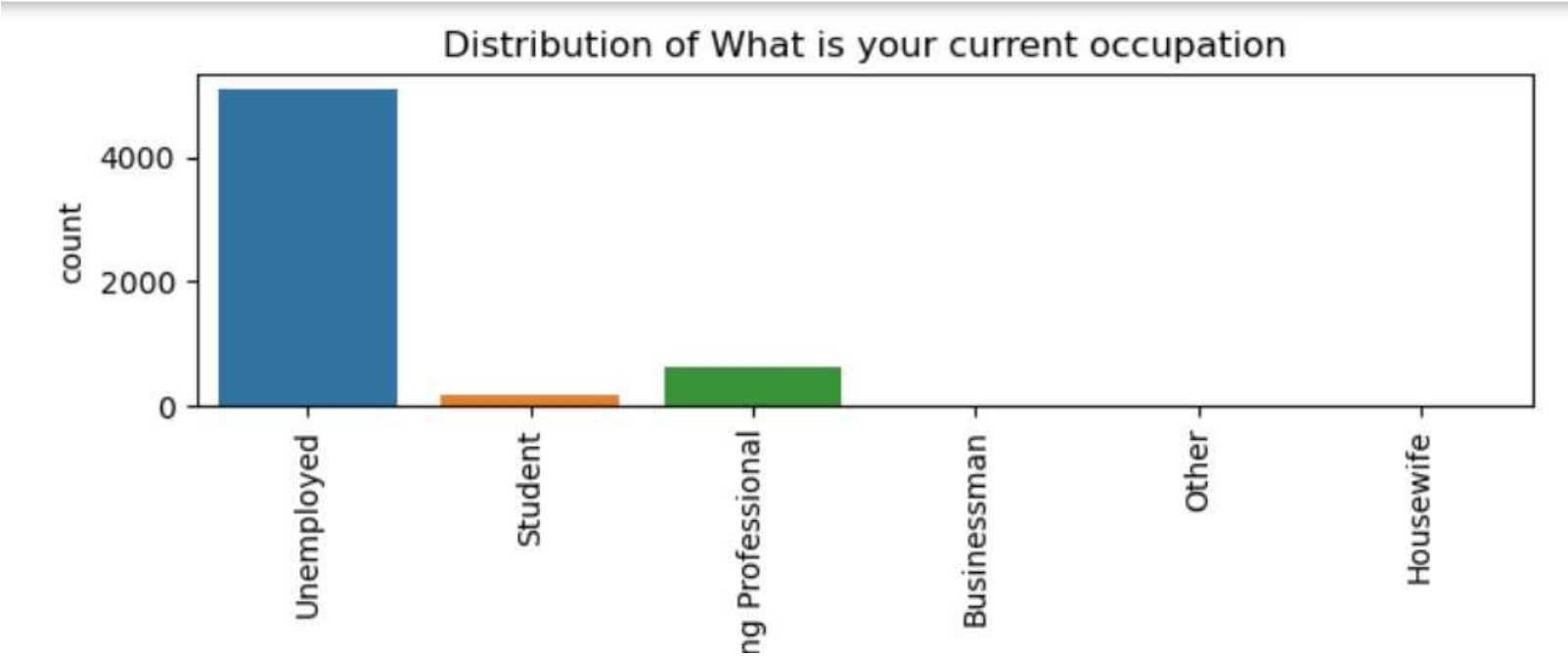
EDA (Univariate Analysis)



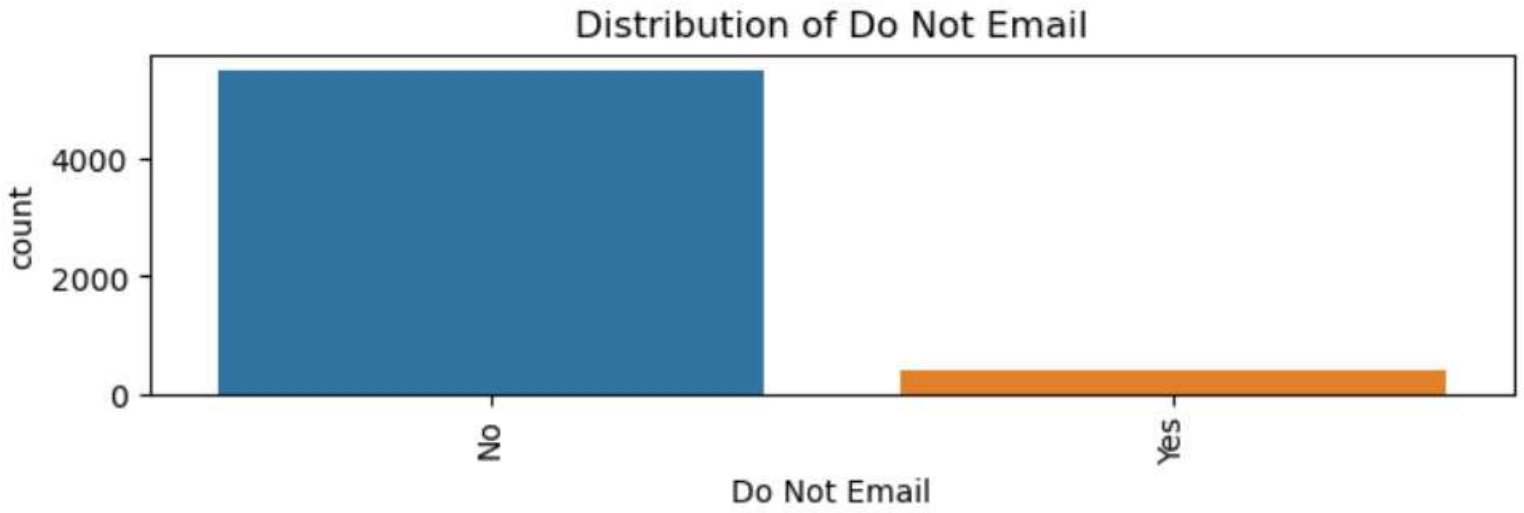
Distribution of Last Notable Activity



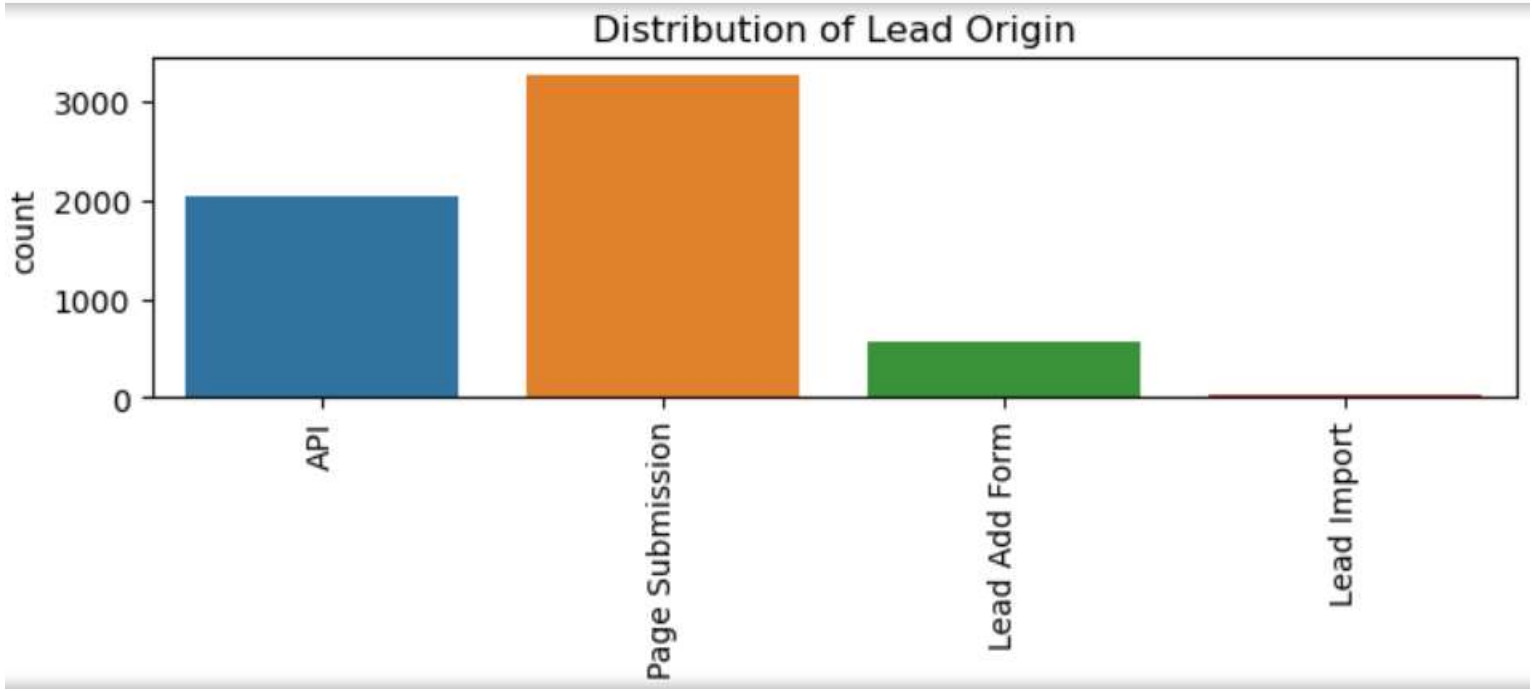
Distribution of What is your Occupation



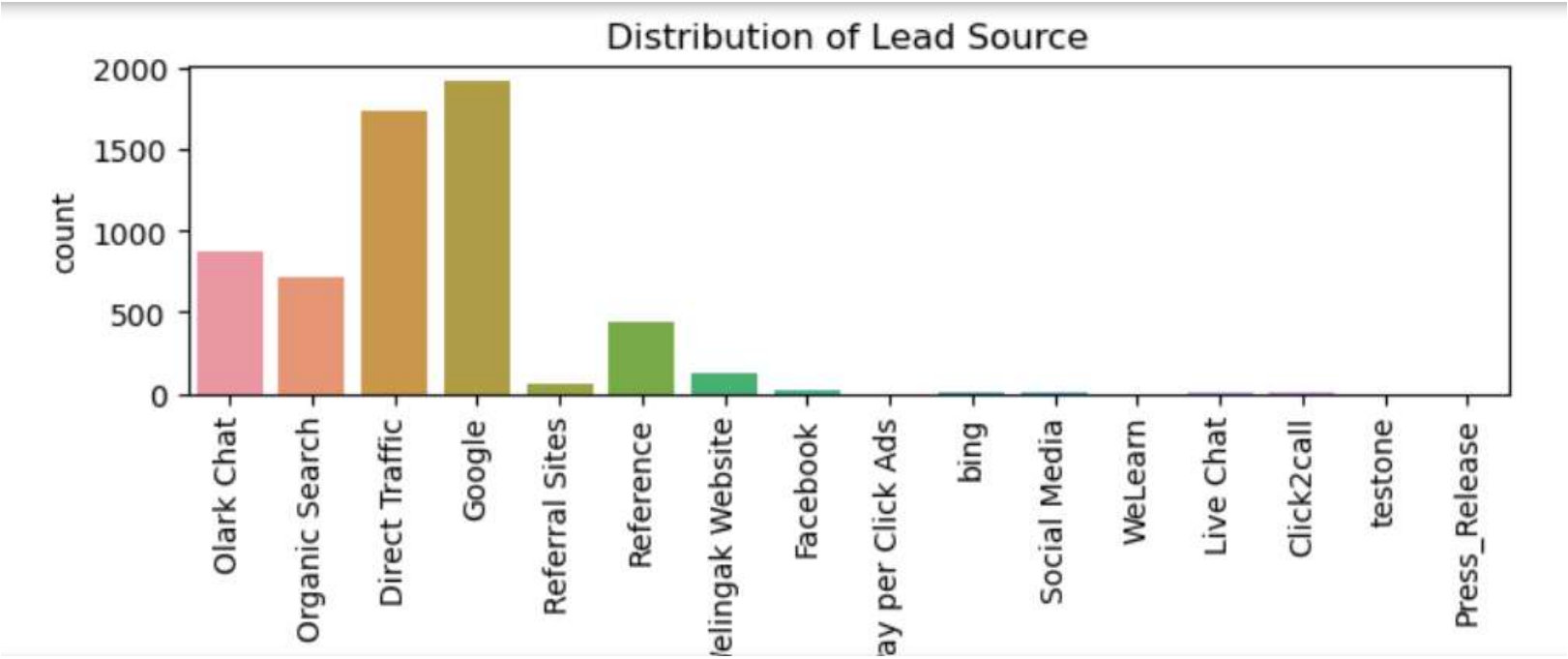
Distribution of Do Not Email



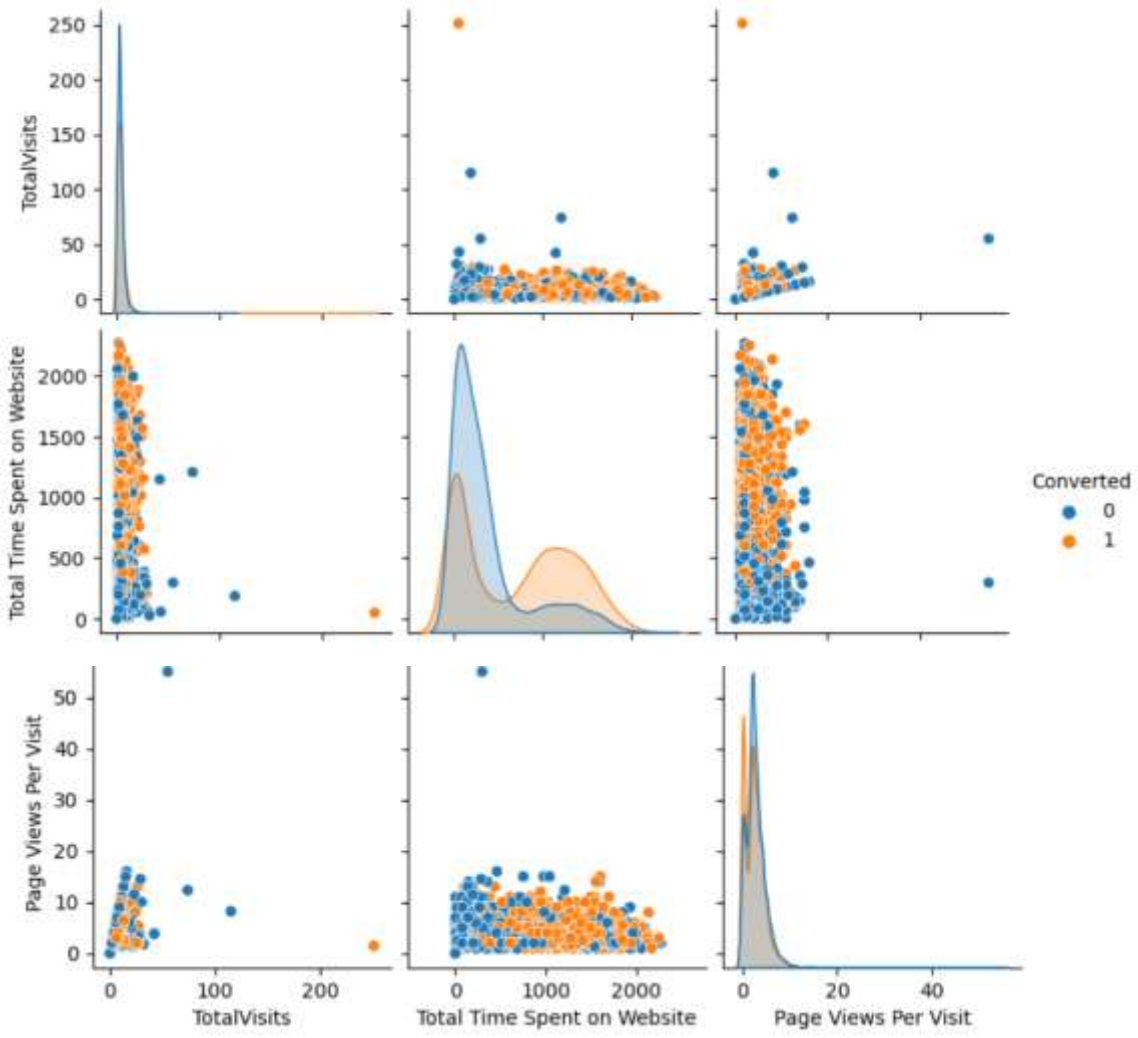
Distribution Of Lead Origin



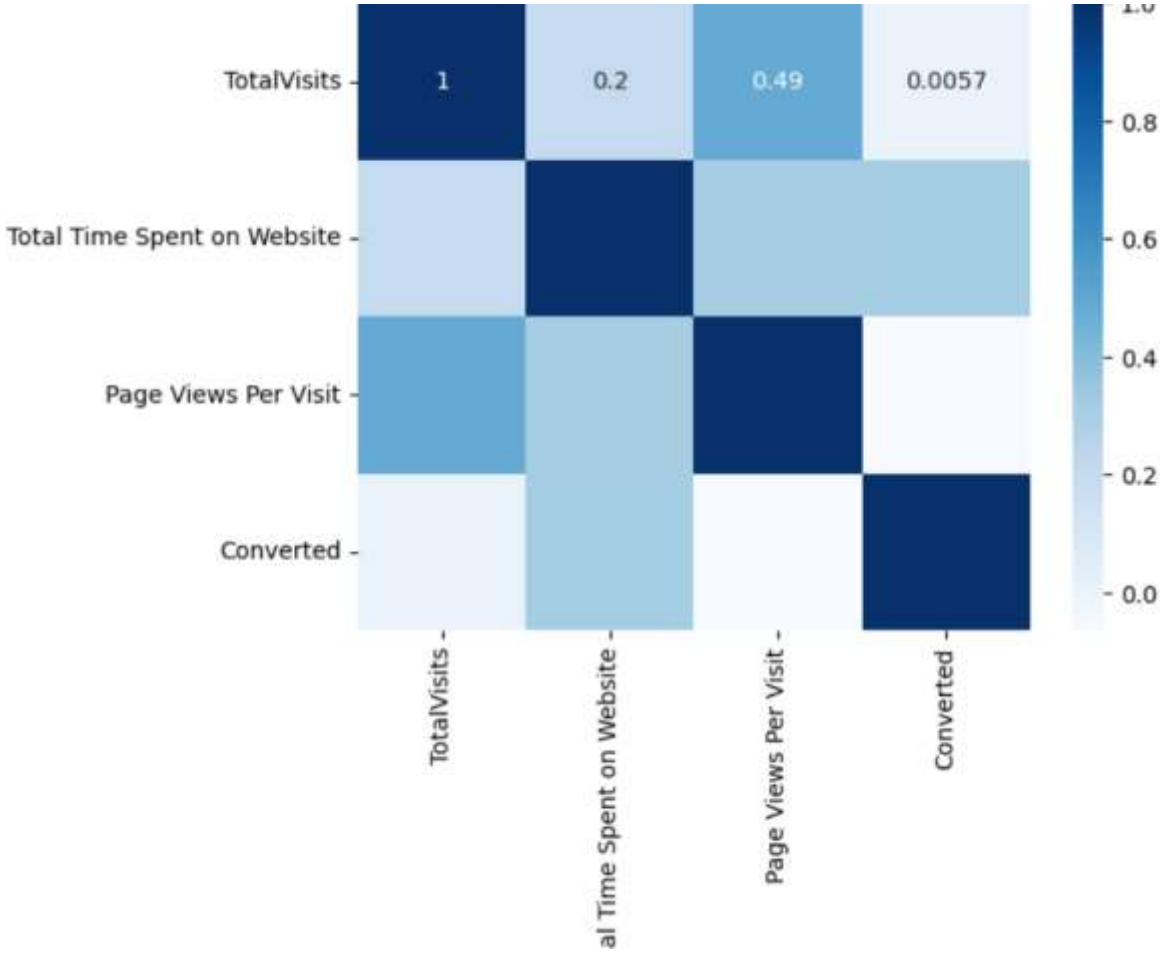
Lead Source


Distribution of Lead Source
count
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
elingak Website
Facebook
ay per Click Ads
bing
Social Media
WeLearn
Live Chat
Click2call
testone
Press_Release

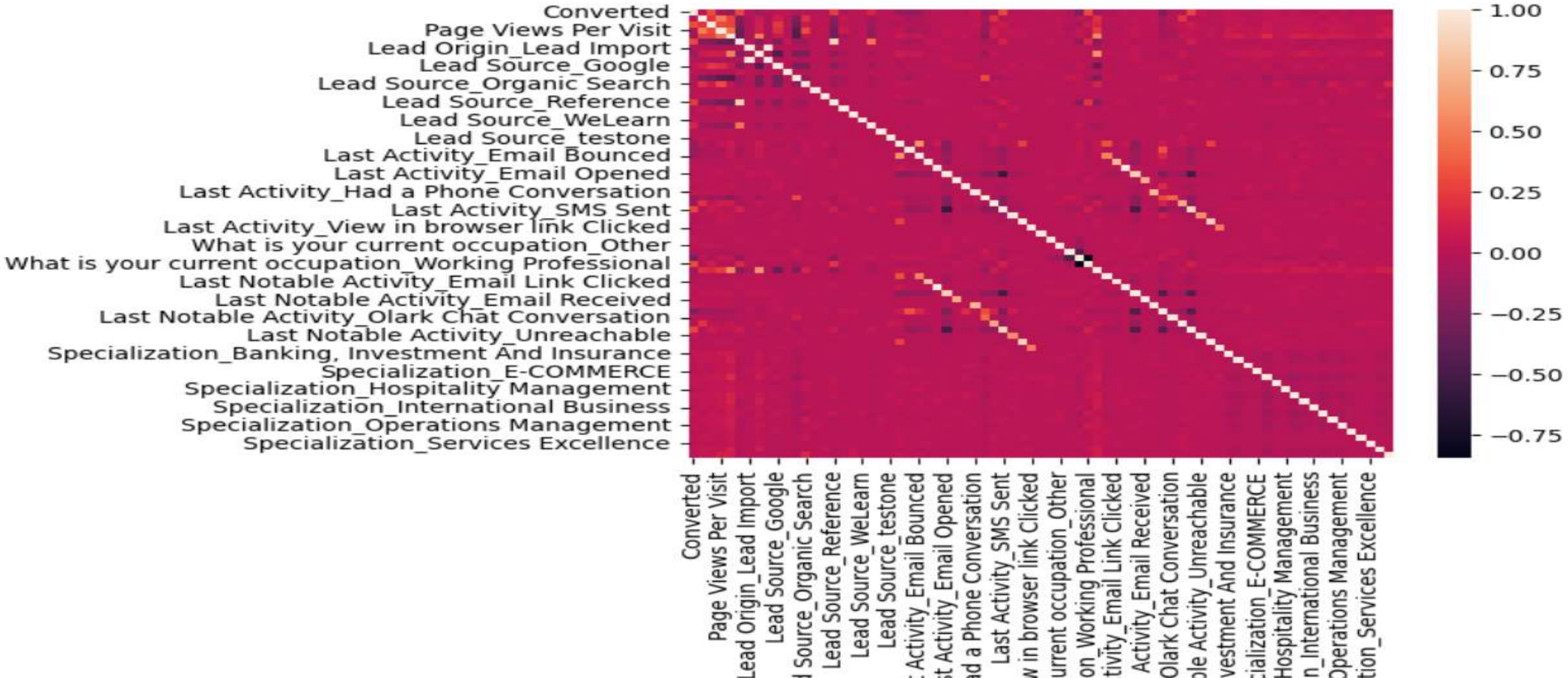
Bivariate Analysis



Heatmap


TotalVisits
Total Time Spent on Website
Page Views Per Visit
Converted
1
0.2
0.49
0.0057
0.8
0.6
0.4
0.2
0.0
TotalVisits
al Time Spent on Website
Page Views Per Visit
Converted

Correlation Matrix



Model Building

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8951	0.093	-9.595	0.000	-1.078	-0.712
Total Time Spent on Website	4.0202	0.175	22.970	0.000	3.677	4.363
Lead Origin_Lead Add Form	3.8610	0.239	16.125	0.000	3.392	4.330
Lead Source_Olark Chat	1.3871	0.124	11.231	0.000	1.145	1.629
Lead Source_Welingak Website	2.3900	1.037	2.304	0.021	0.357	4.423
Do Not Email_Yes	-1.4884	0.236	-6.295	0.000	-1.952	-1.025
Last Activity_Converted to Lead	-1.0392	0.236	-4.401	0.000	-1.502	-0.576
Last Activity_Email Bounced	-0.9967	0.432	-2.305	0.021	-1.844	-0.149
Last Activity_Email Link Clicked	-1.4877	0.262	-5.681	0.000	-2.001	-0.974
Last Activity_Olark Chat Conversation	-1.5399	0.202	-7.617	0.000	-1.936	-1.144
Last Activity_Page Visited on Website	-1.0208	0.199	-5.140	0.000	-1.410	-0.632
What is your current occupation_Working Professional	2.6142	0.208	12.591	0.000	2.207	3.021
Last Notable Activity_Email Opened	-1.0269	0.100	-10.317	0.000	-1.222	-0.832
Last Notable Activity_Modified	-1.0920	0.112	-9.741	0.000	-1.312	-0.872

	Features	VIF
12	Last Notable Activity_Modified	1.94
6	Last Activity_Email Bounced	1.71
4	Do Not Email_Yes	1.68
0	Total Time Spent on Website	1.52
1	Lead Origin_Lead Add Form	1.46
8	Last Activity_Olark Chat Conversation	1.41
11	Last Notable Activity_Email Opened	1.37
2	Lead Source_Olark Chat	1.31
3	Lead Source_Welingak Website	1.28
5	Last Activity_Converted to Lead	1.28
10	What is your current occupation_Working Profes...	1.19
9	Last Activity_Page Visited on Website	1.11
7	Last Activity_Email Link Clicked	1.05

Model Evaluation

Trained Dataset

Accuracy : 0.79

Sensitivity :0.73

Specificity :0.84

Test Dataset

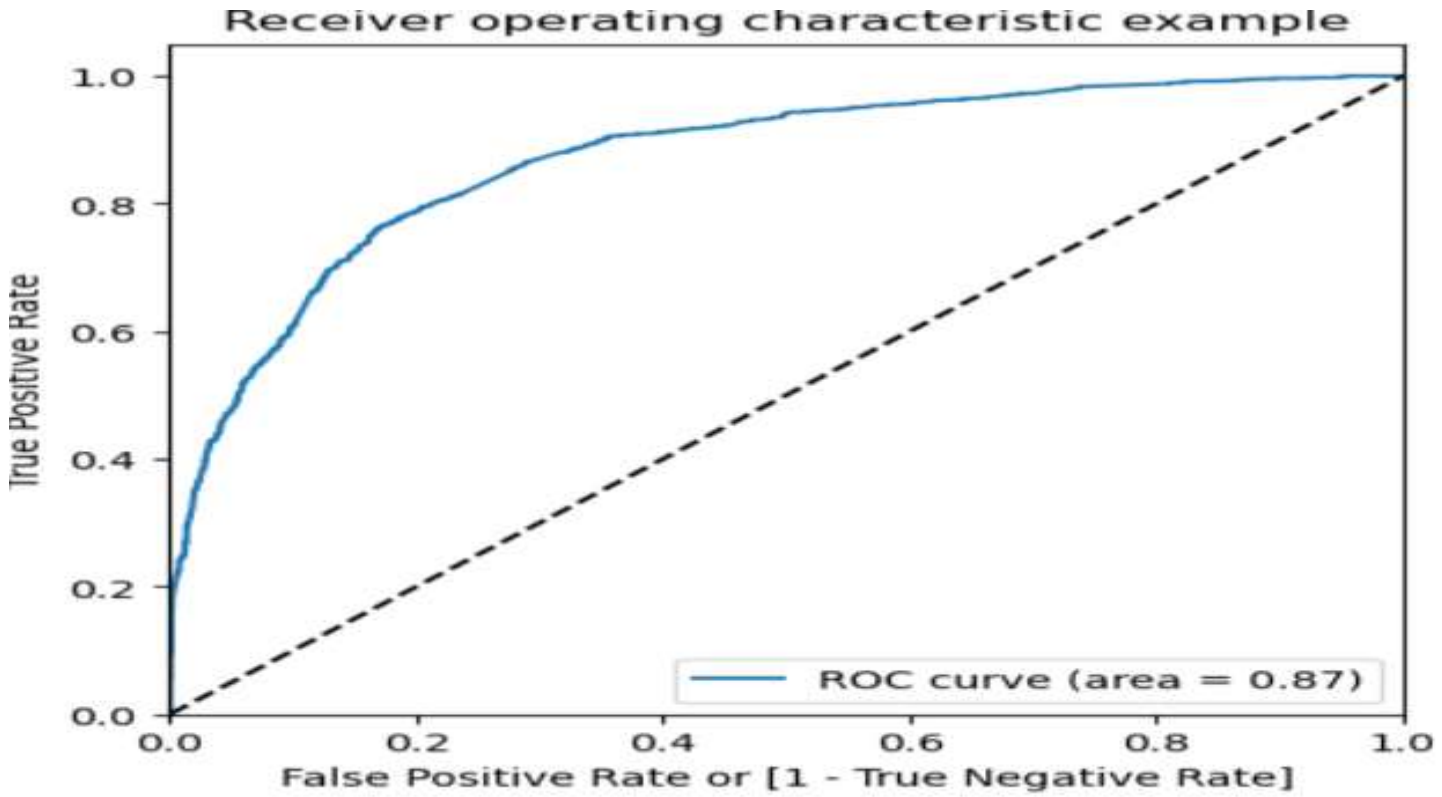
Accuracy : 0.78

Sensitivity :0.76

Specificity :0.80

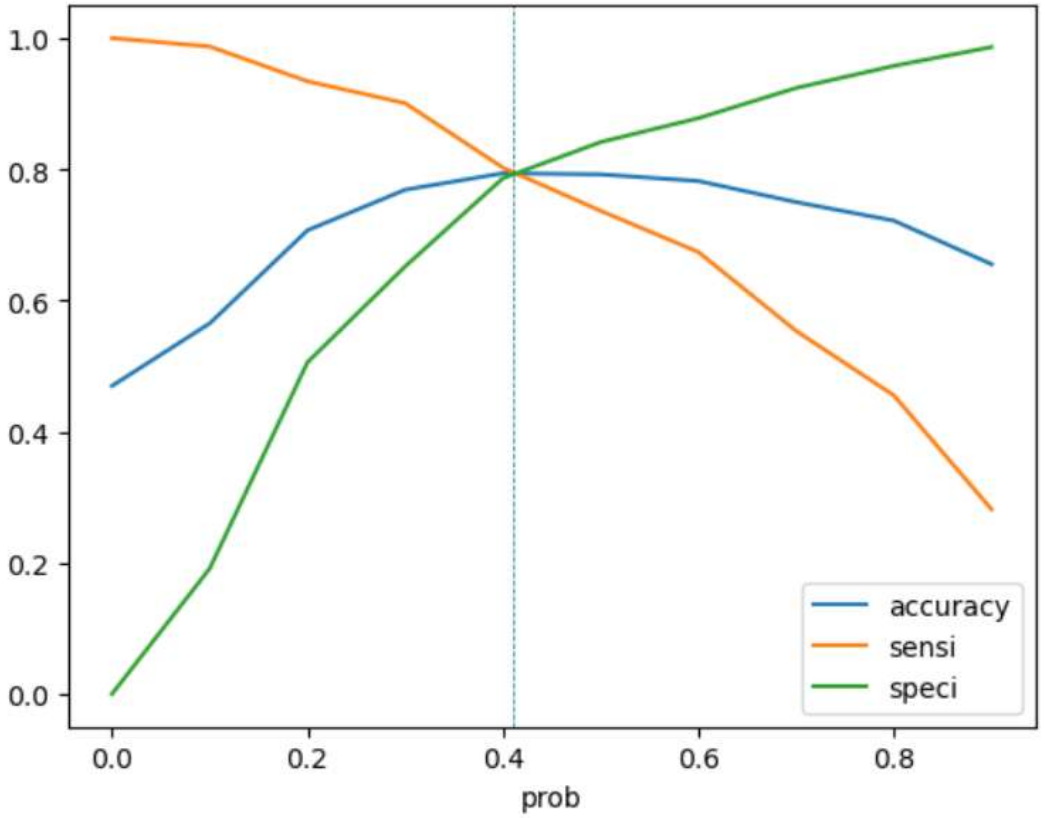
Model Evaluation

The area under Roc Curve is 0.87

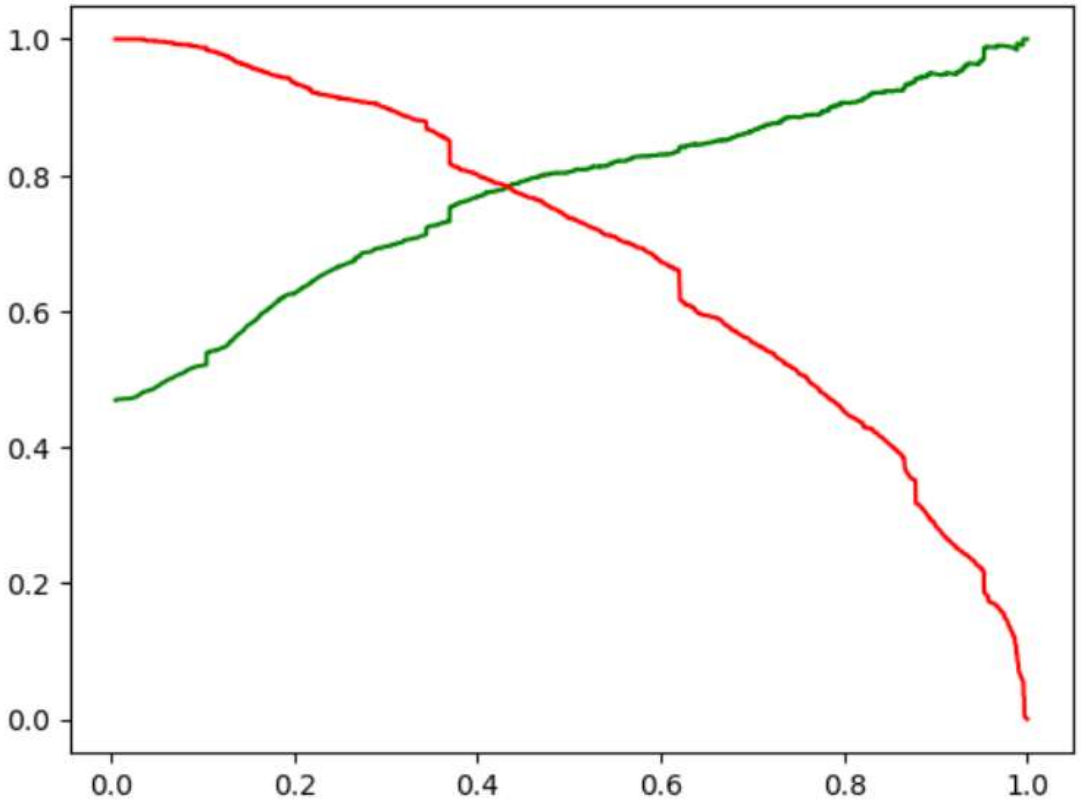


Probability Cutoff

The optimal values of 3 metrics is 0.41.



Precision and recall trade off



Conclusion

The Top 3 features that contribute highly in predicting Hot leads are listed below . X education company should concentrate more on these top features .

1)Total Time Spent on Website - 0.43

2)Lead Origin_Lead Add Form- 0.27

3) What is your current Occupation_working professional -0.11

Recommendation

- Focus on the features with positive coefficients
- Develop strategies to attract the Hot Leads
- Optimize communication channel based on the impact
- Allocate more funds for improving advertising
- Allow incentives /discounts in fee for the providing reference for converted leads.
- As areas of improvement review landing page submission

Thank You ……..!!!