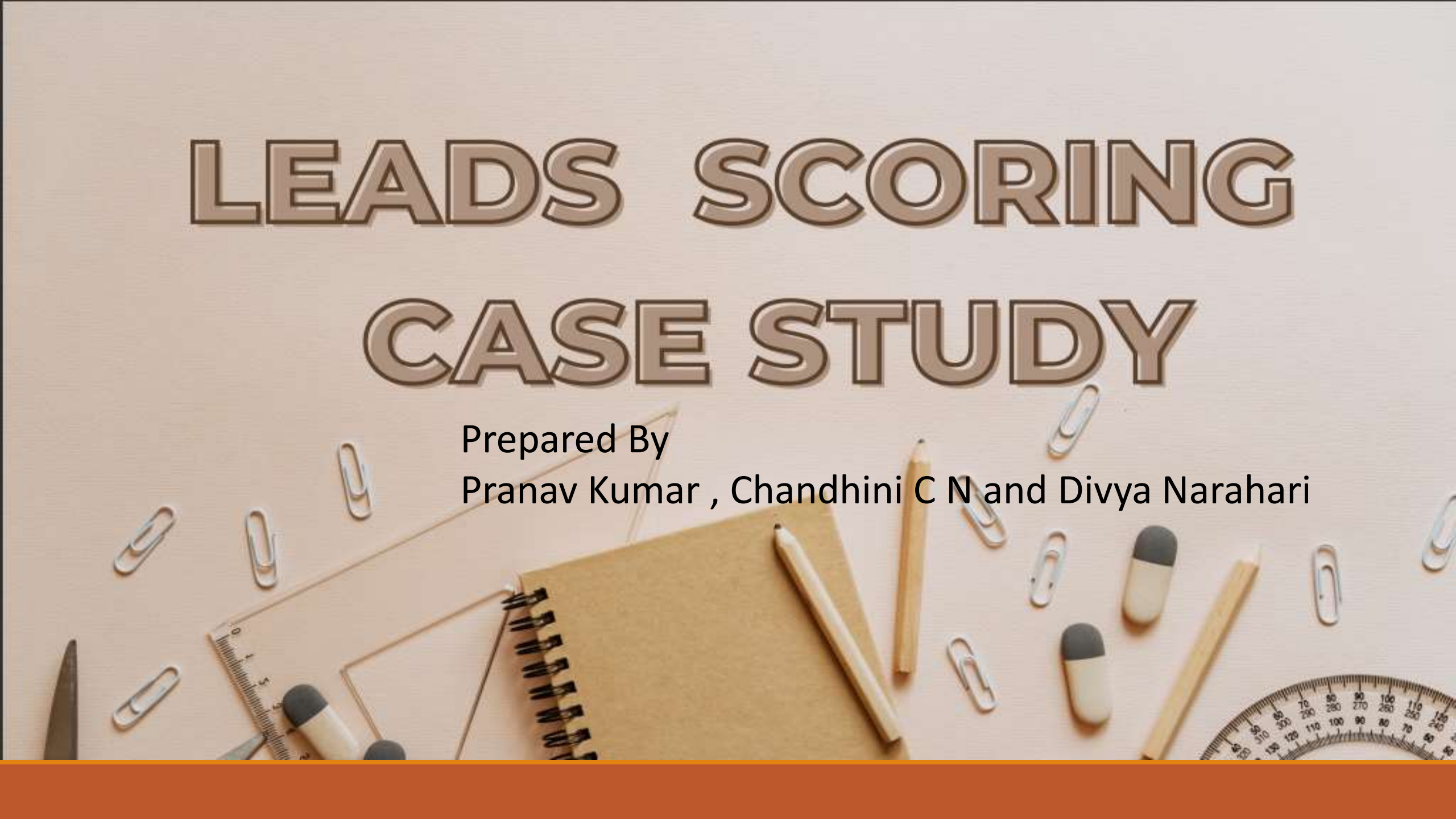
# LEADS SCORING CASE STUDY

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 , assigning lead score

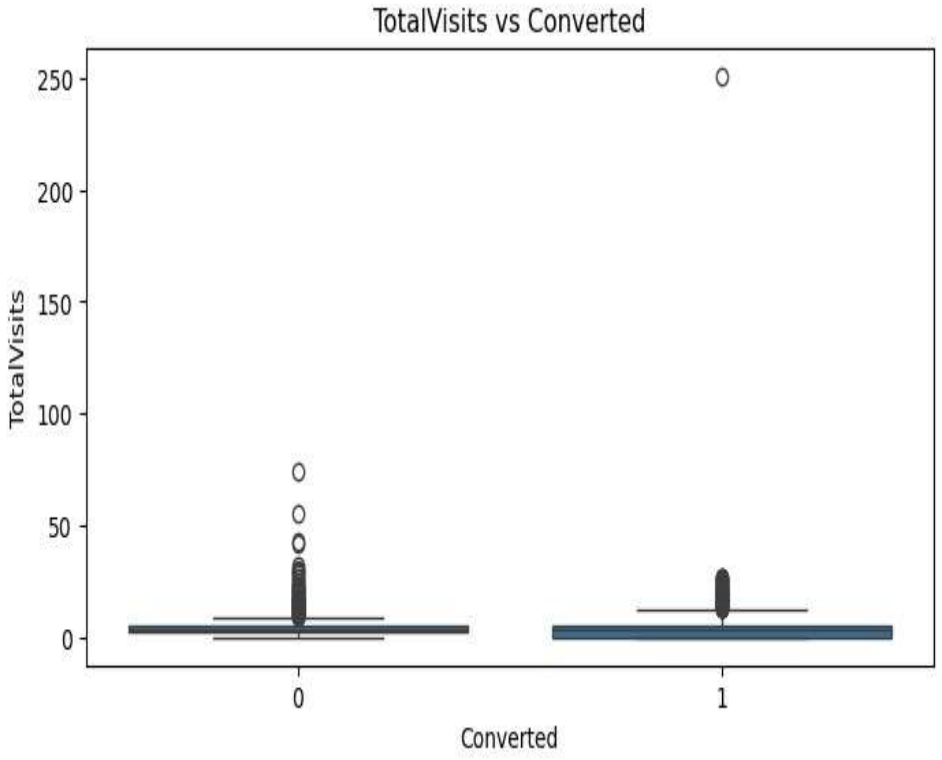
- Prediction on Test Data

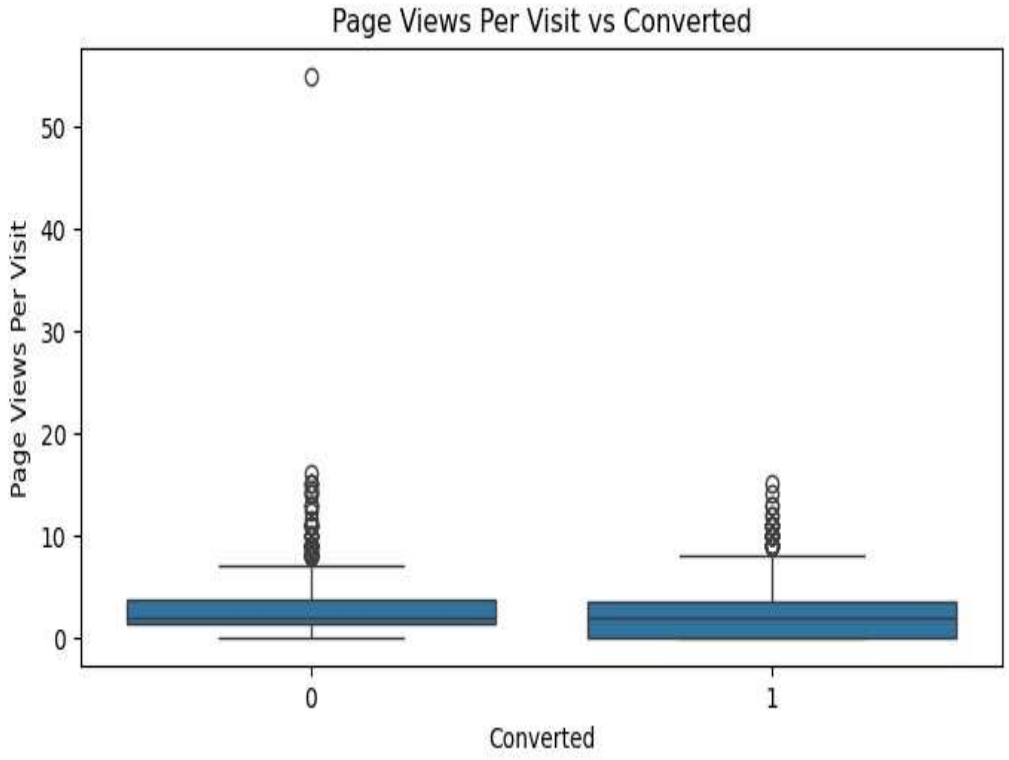
Compare Train data and test dataset, Assign lead score and identify top features

- Recommendation

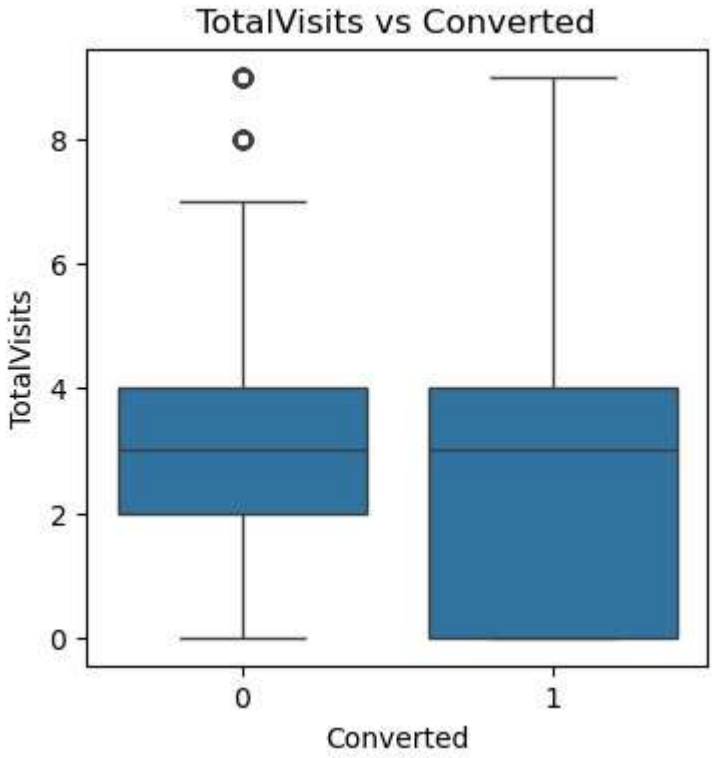
Suggest the best features to identify the Hot leads in order to focus more on the ares of improvement.

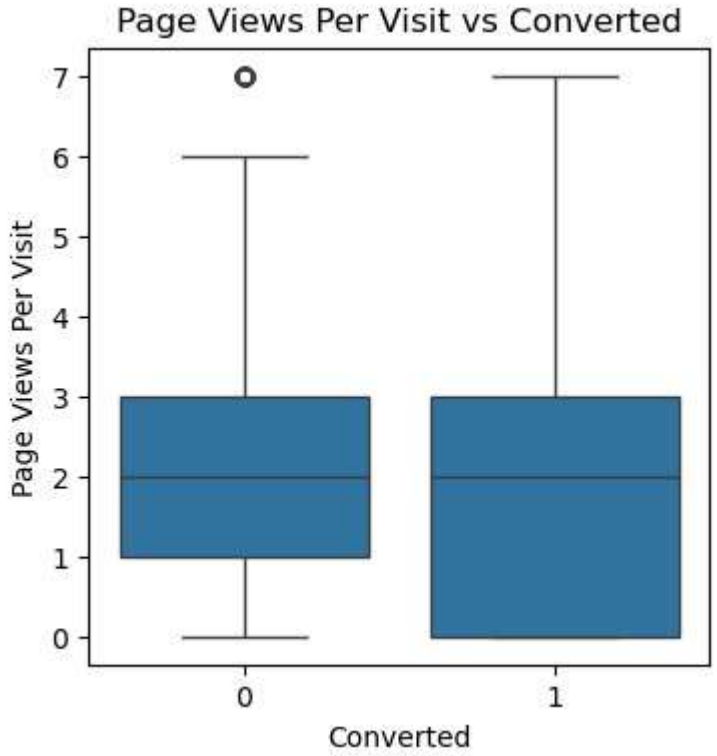
# Outliers


TotalVisits vs Converted
TotalVisits
250
200
150
100
50
0
0
1
Converted


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

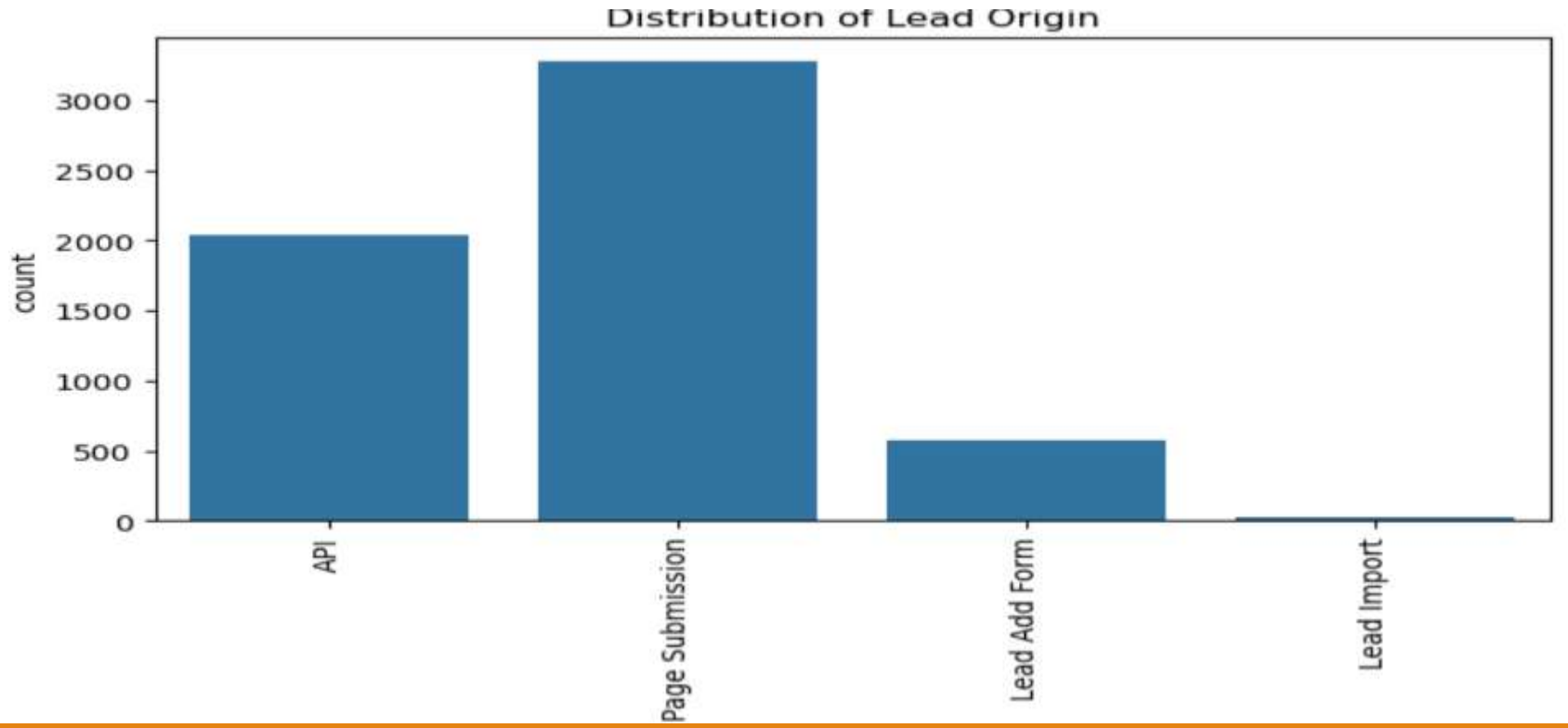
# After Outlier Treatment



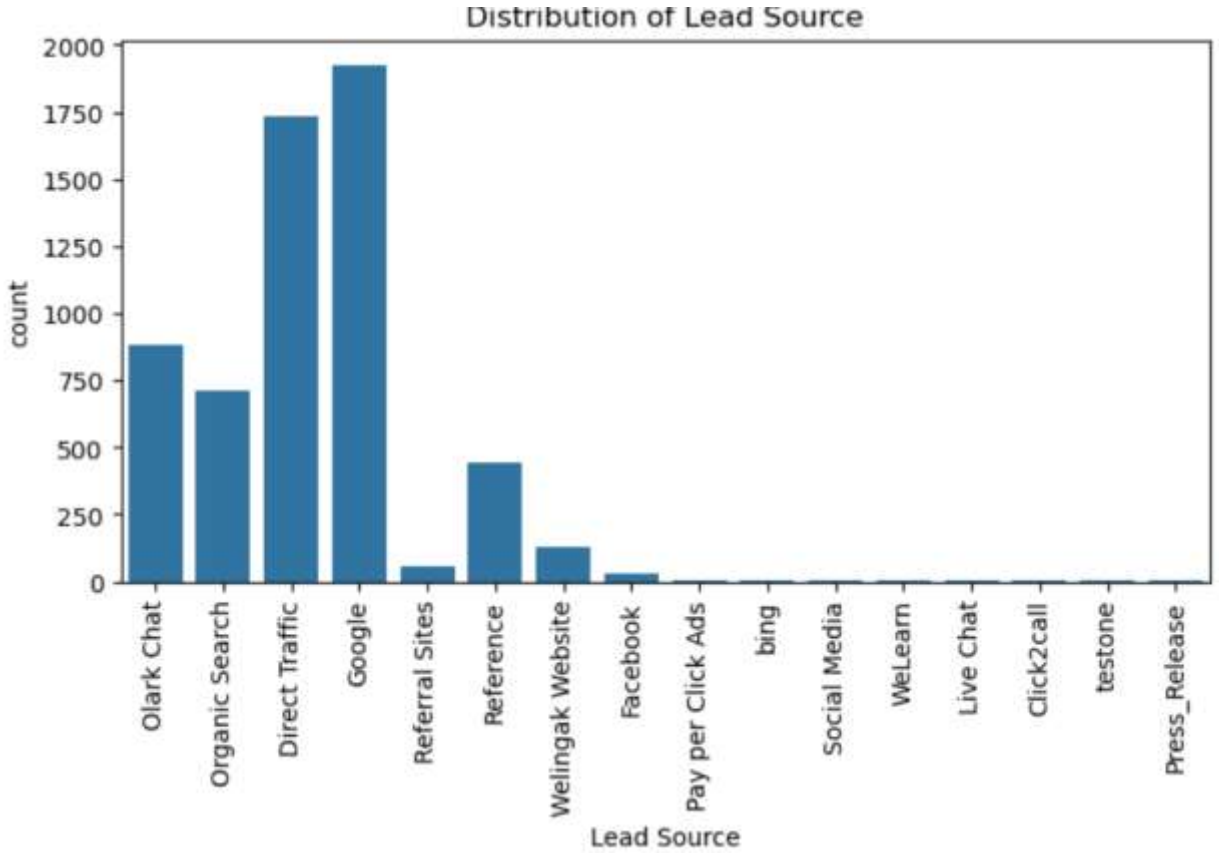



# EDA( Univariate Analysis)

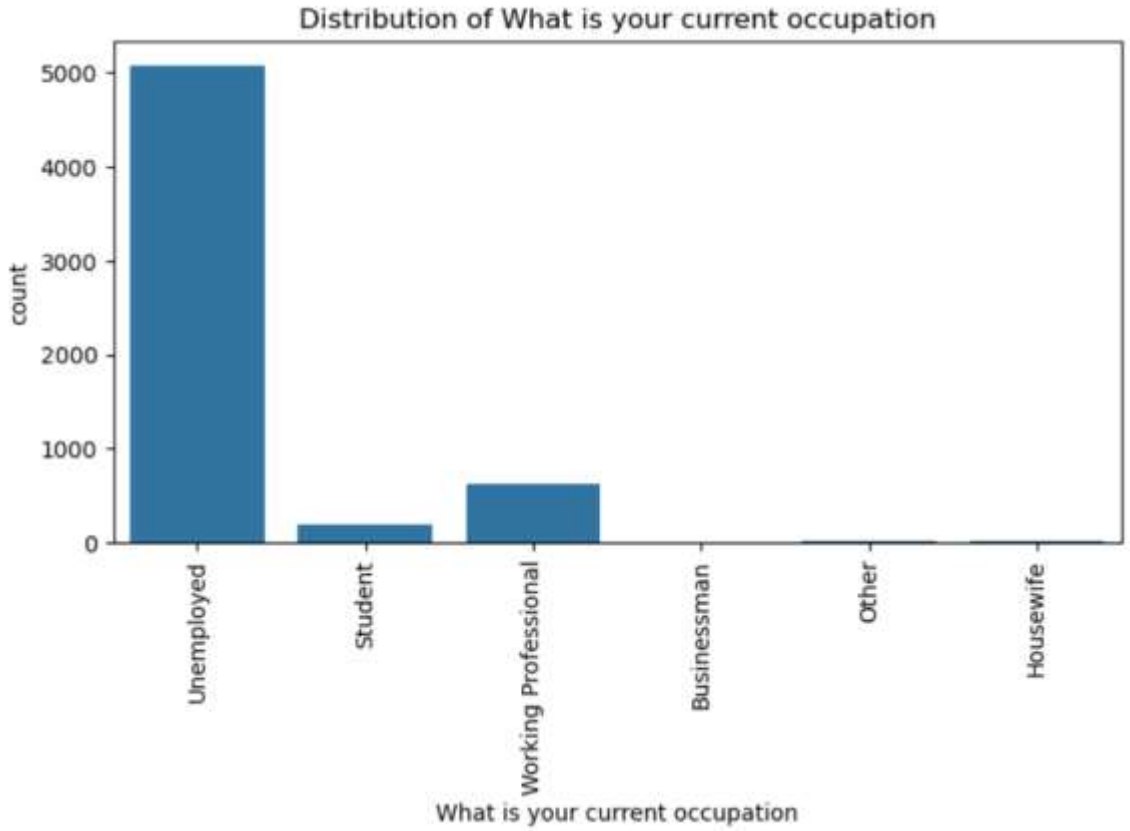
Lead Origin



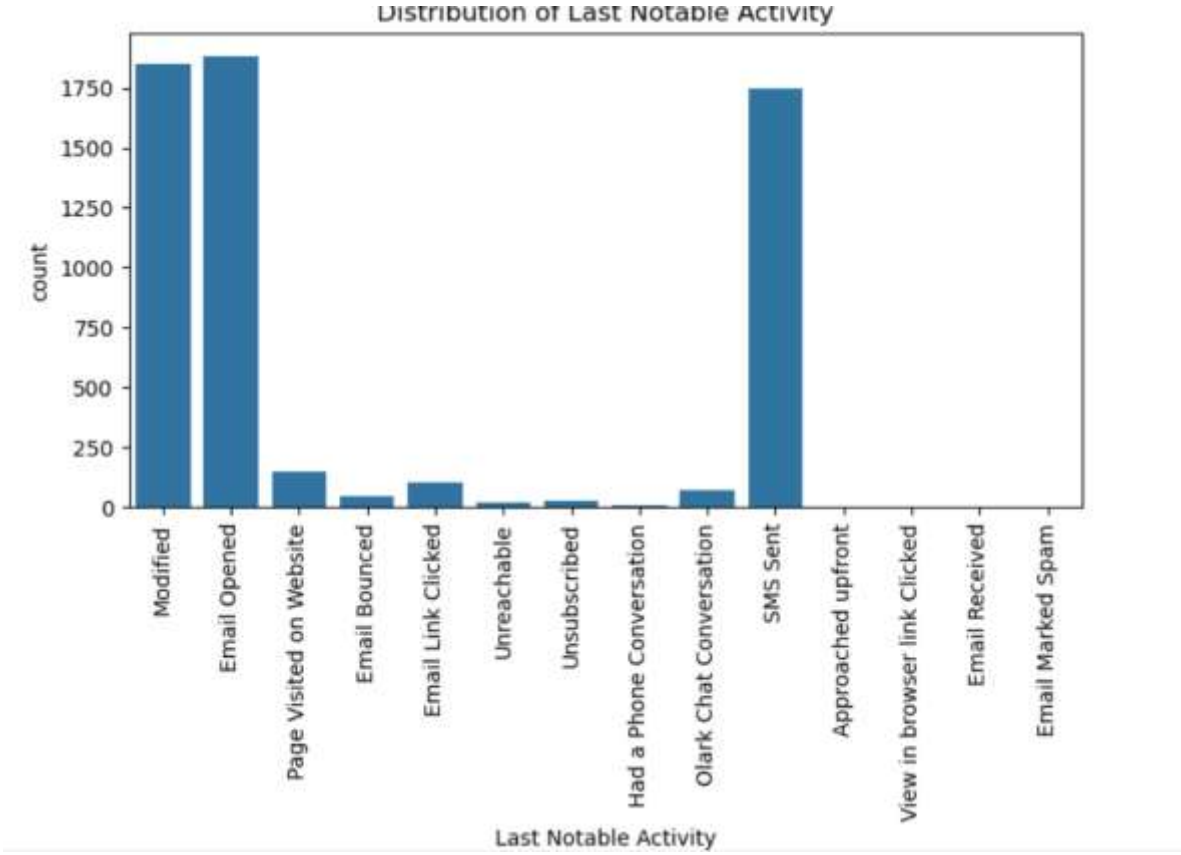
# Lead source


Distribution of Lead Source
count
2000
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Pay per Click Ads
bing
Social Media
WeLearn
Live Chat
Click2call
testone
Press_Release
Lead Source

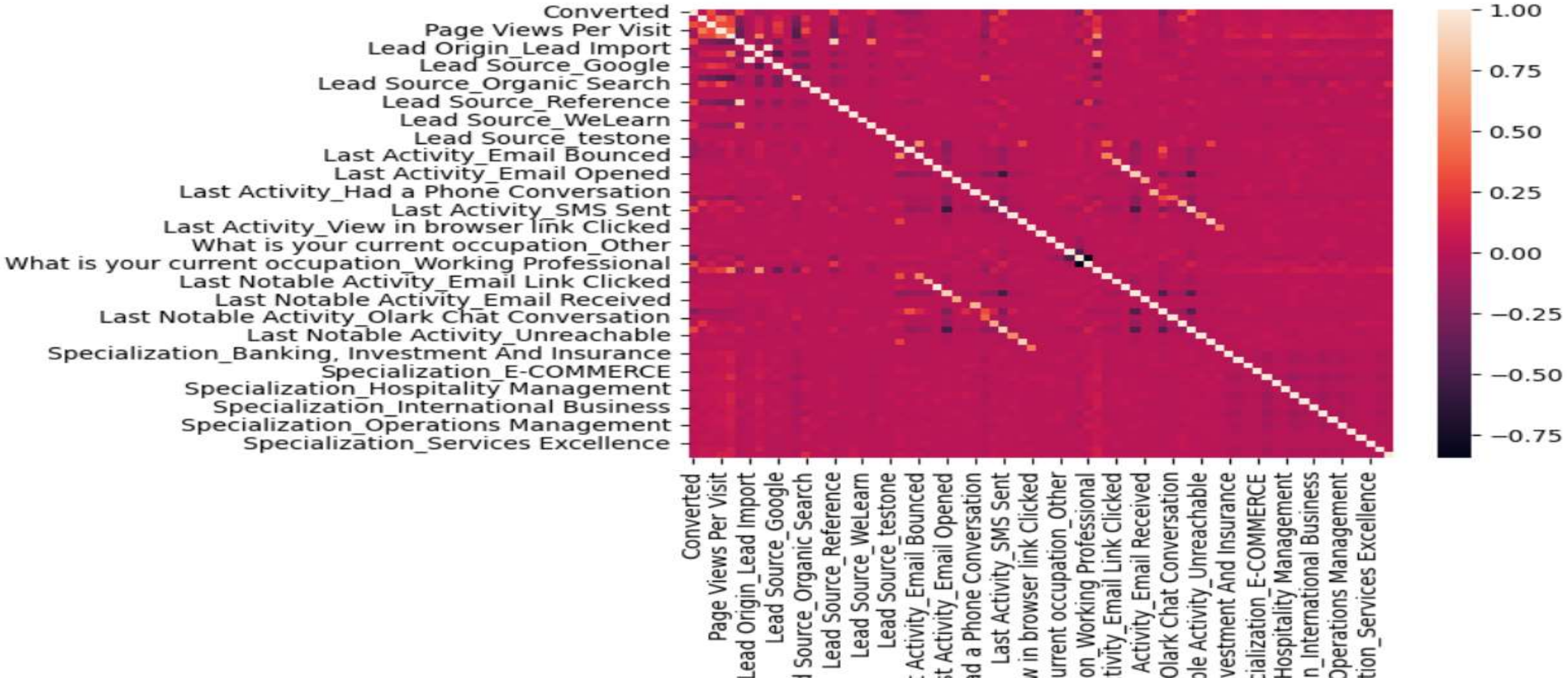
# Current occupation



# Last Noble activity



# Multivariate Analysis(Heatmap)



# Selected Model logm3 P-Values

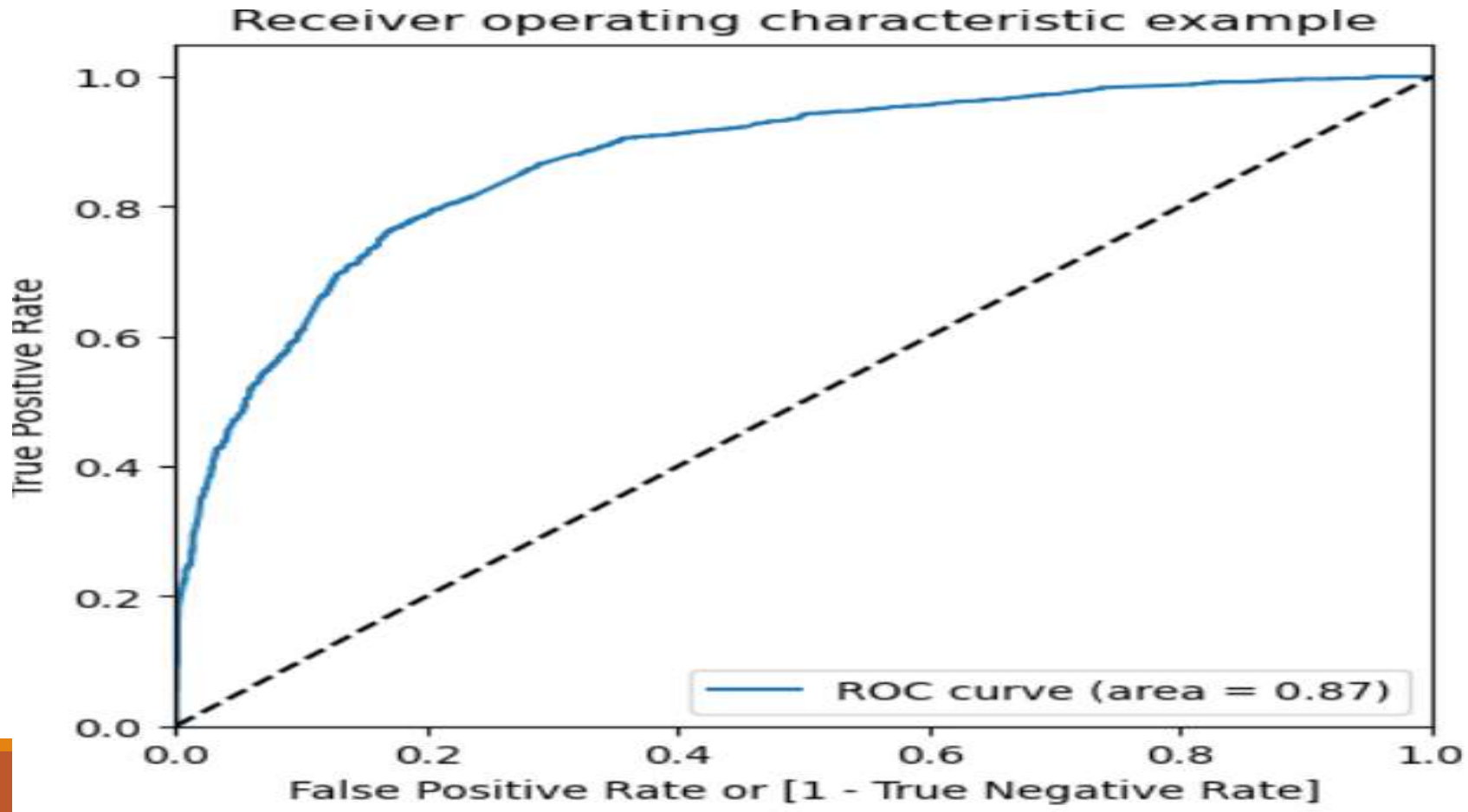
| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -0.9037 | 0.094 | -9.659 | 0.000 | -1.087 | -0.720 |
| **Total Time Spent on Website** | 4.0283 | 0.175 | 22.992 | 0.000 | 3.685 | 4.372 |
| **Lead Origin_Lead Add Form** | 3.8985 | 0.240 | 16.243 | 0.000 | 3.428 | 4.369 |
| **Lead Source_Olark Chat** | 1.3930 | 0.124 | 11.276 | 0.000 | 1.151 | 1.635 |
| **Lead Source_Welingak Website** | 2.4083 | 1.039 | 2.319 | 0.020 | 0.373 | 4.444 |
| **Do Not Email_Yes** | -1.7690 | 0.207 | -8.551 | 0.000 | -2.175 | -1.364 |
| **Last Activity_Converted to Lead** | -0.9915 | 0.236 | -4.199 | 0.000 | -1.454 | -0.529 |
| **Last Activity_Email Link Clicked** | -1.4749 | 0.263 | -5.616 | 0.000 | -1.990 | -0.960 |
| **Last Activity_Had a Phone Conversation** | 1.8241 | 0.838 | 2.178 | 0.029 | 0.183 | 3.466 |
| **Last Activity_Olark Chat Conversation** | -1.5043 | 0.202 | -7.435 | 0.000 | -1.901 | -1.108 |
| **Last Activity_Page Visited on Website** | -0.9847 | 0.199 | -4.940 | 0.000 | -1.375 | -0.594 |
| **What is your current occupation_Working Professional** | 2.6154 | 0.208 | 12.584 | 0.000 | 2.208 | 3.023 |
| **Last Notable Activity_Email Opened** | -1.0228 | 0.100 | -10.266 | 0.000 | -1.218 | -0.828 |
| **Last Notable Activity_Modified** | -1.1341 | 0.111 | -10.193 | 0.000 | -1.352 | -0.916 |

# VIF of Selected Model

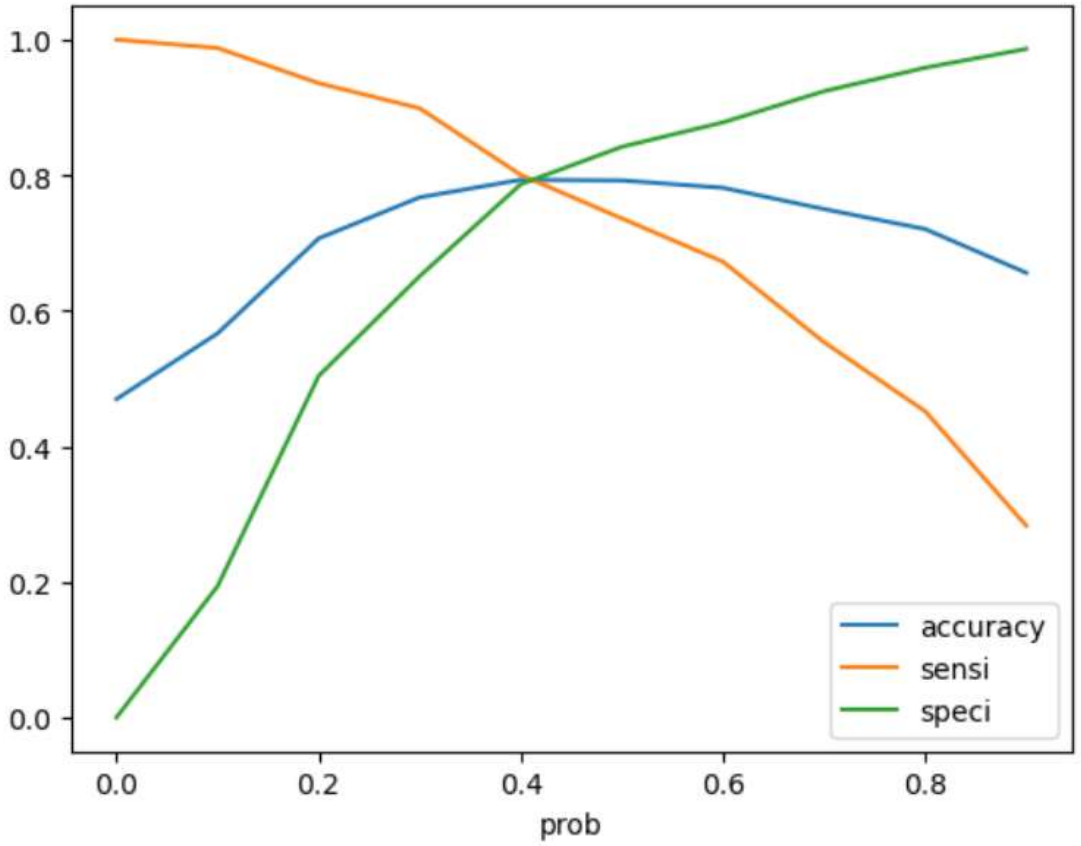
| | Features | VIF |
|---|---|---|
| 12 | Last Notable Activity_Modified | 1.88 |
| 0 | Total Time Spent on Website | 1.52 |
| 1 | Lead Origin_Lead Add Form | 1.46 |
| 8 | Last Activity_Olark Chat Conversation | 1.39 |
| 11 | Last Notable Activity_Email Opened | 1.37 |
| 2 | Lead Source_Olark Chat | 1.31 |
| 3 | Lead Source_Welingak Website | 1.28 |
| 5 | Last Activity_Converted to Lead | 1.27 |
| 10 | What is your current occupation_Working Profes... | 1.20 |
| 9 | Last Activity_Page Visited on Website | 1.10 |
| 4 | Do Not Email_Yes | 1.07 |
| 6 | Last Activity_Email Link Clicked | 1.04 |
| 7 | Last Activity_Had a Phone Conversation | 1.01 |

# Model Evaluation

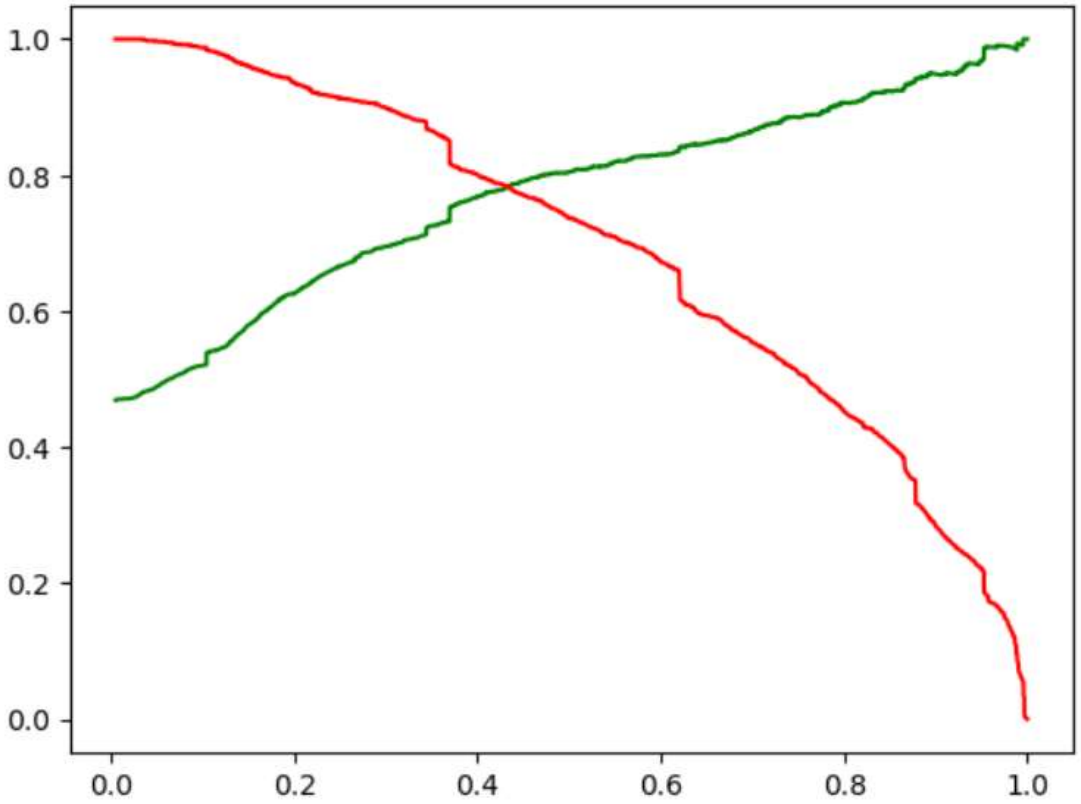
The area under Roc Curve is 0.86



# Probability Cutoff



# 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

2)Lead Origin_Lead Add Form

3)Lead Source_Olark Chat

The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead

1)Last Notable Activity : Modified

2)Lead Origin: Lead Add Form

3)Last Activity: Olark Chat Conversation

Overall Accuracy of test and train data is 0.78

# 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 ……..!!!*