LEAD SCORE CASE STUDY

Submitted By:

Prateek Jain

Ketan Makde

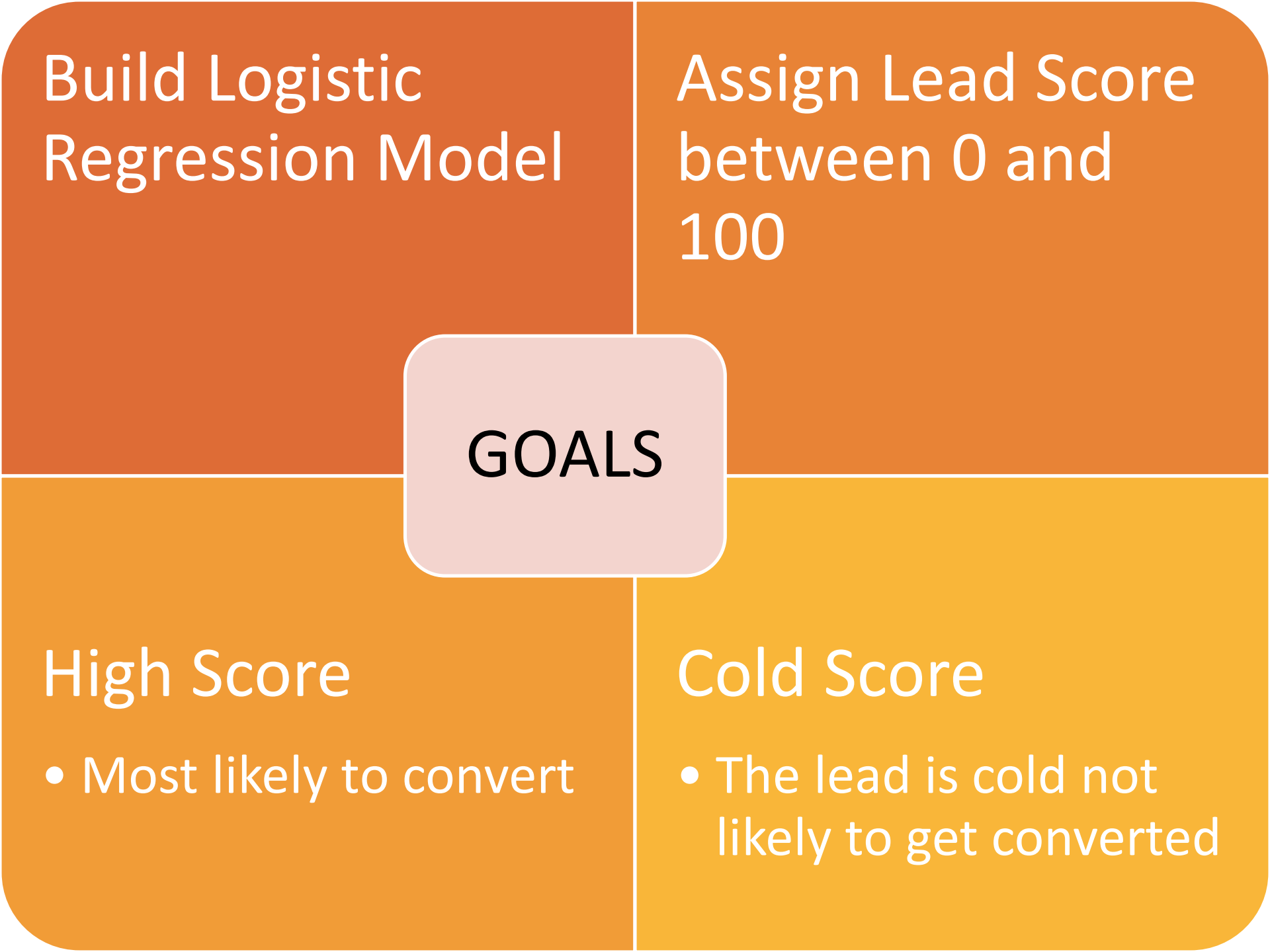
LEAD SCORE CASE STUDY

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company gets lead from website and search engines. 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%.

PROBLEM STATEMENT

- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.


Build Logistic Regression Model
Assign Lead Score between 0 and 100
GOALS
High Score
• Most likely to convert
Cold Score
• The lead is cold not likely to get converted


STEPS INVOLVED

Importing

- Libraries and Dataset

Exploratory Data Analysis

- Cleaning
- Understanding Data
- Visual Representation of Data

Data Preperation for model

- Creating Dummies
- Checking for data imbalance
- Splitting dataset
- Scaling

Model Building

- Feature Selection using RFE
- Building model using statsmodels

Model Evaluation

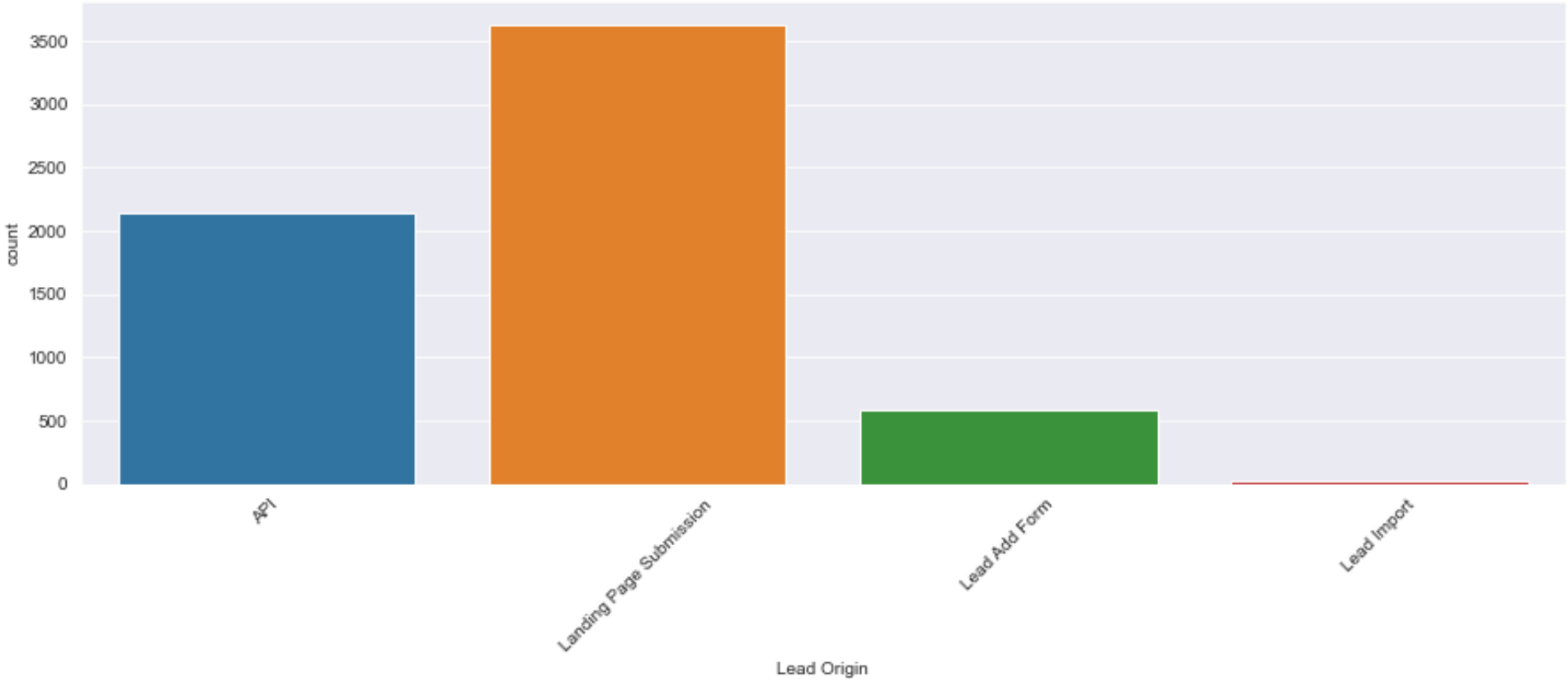
- Initial Predictions based on 0.5 probability
- Accuracy Score, Precision Score, Recall Score
- Plotting ROC curve
- Finding optimal cutoff point
- Precision and Recall Trade Off

Making predictions on test dataset

- Scaling numeric variables
- Prediction based .42 threshold
- Evaluation Matrix

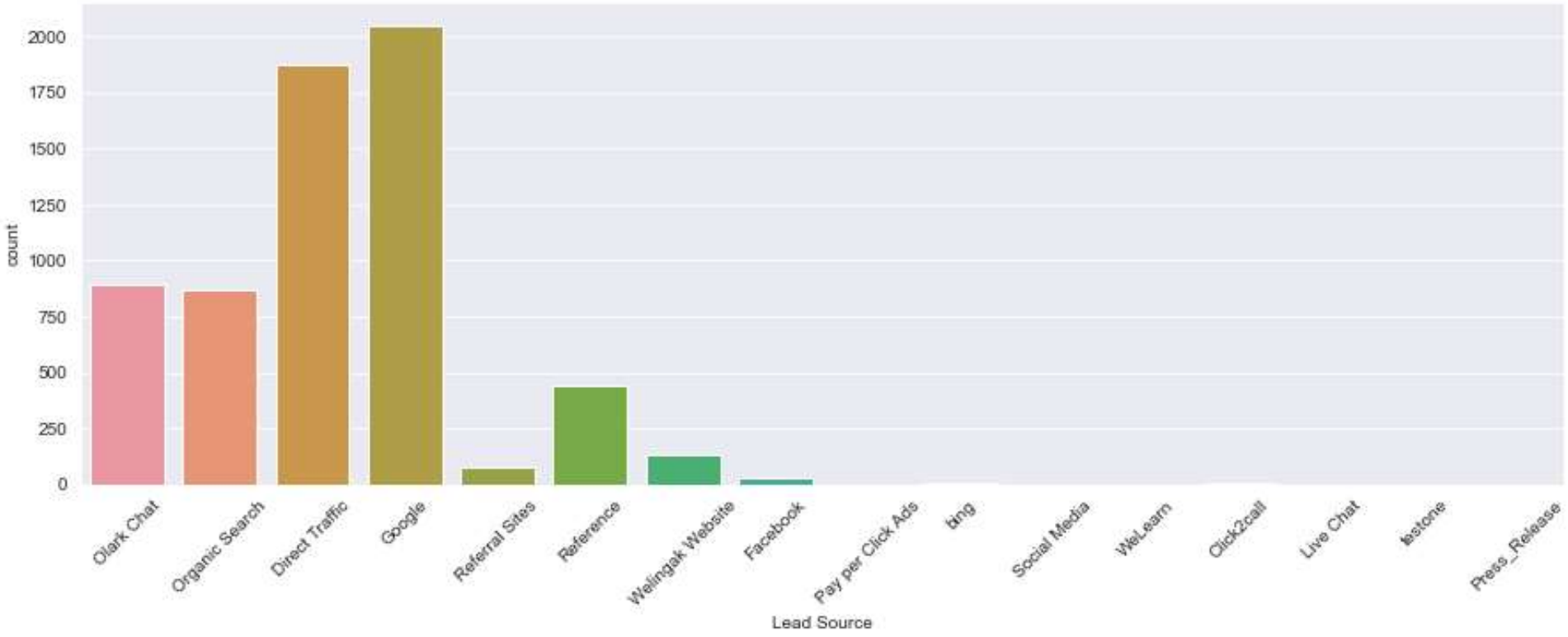

INSIGHTS

LEAD ORIGIN



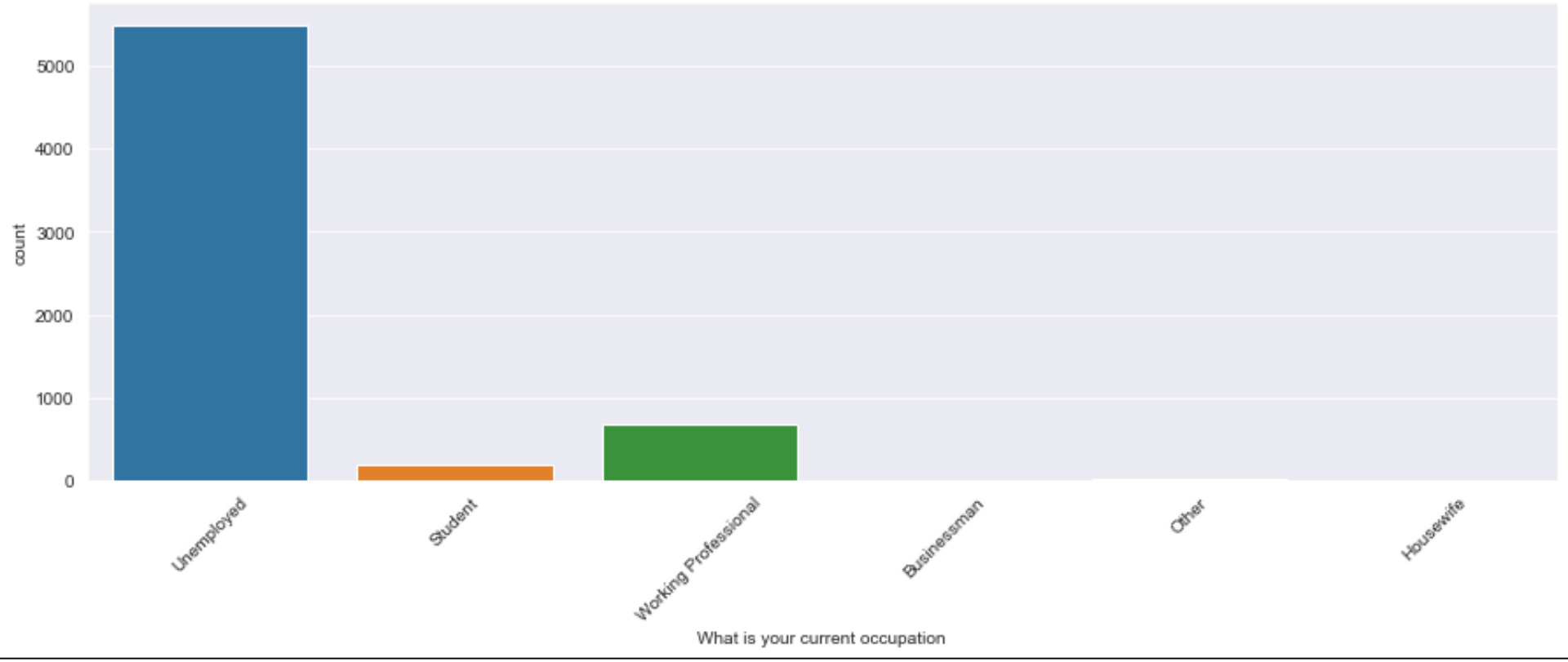
- **Maximum number of professional comes to the company's website page by clicking on the company's links ads on various website this is called landing page**
- **A good number of lead comes from API**

LEAD SOURCE



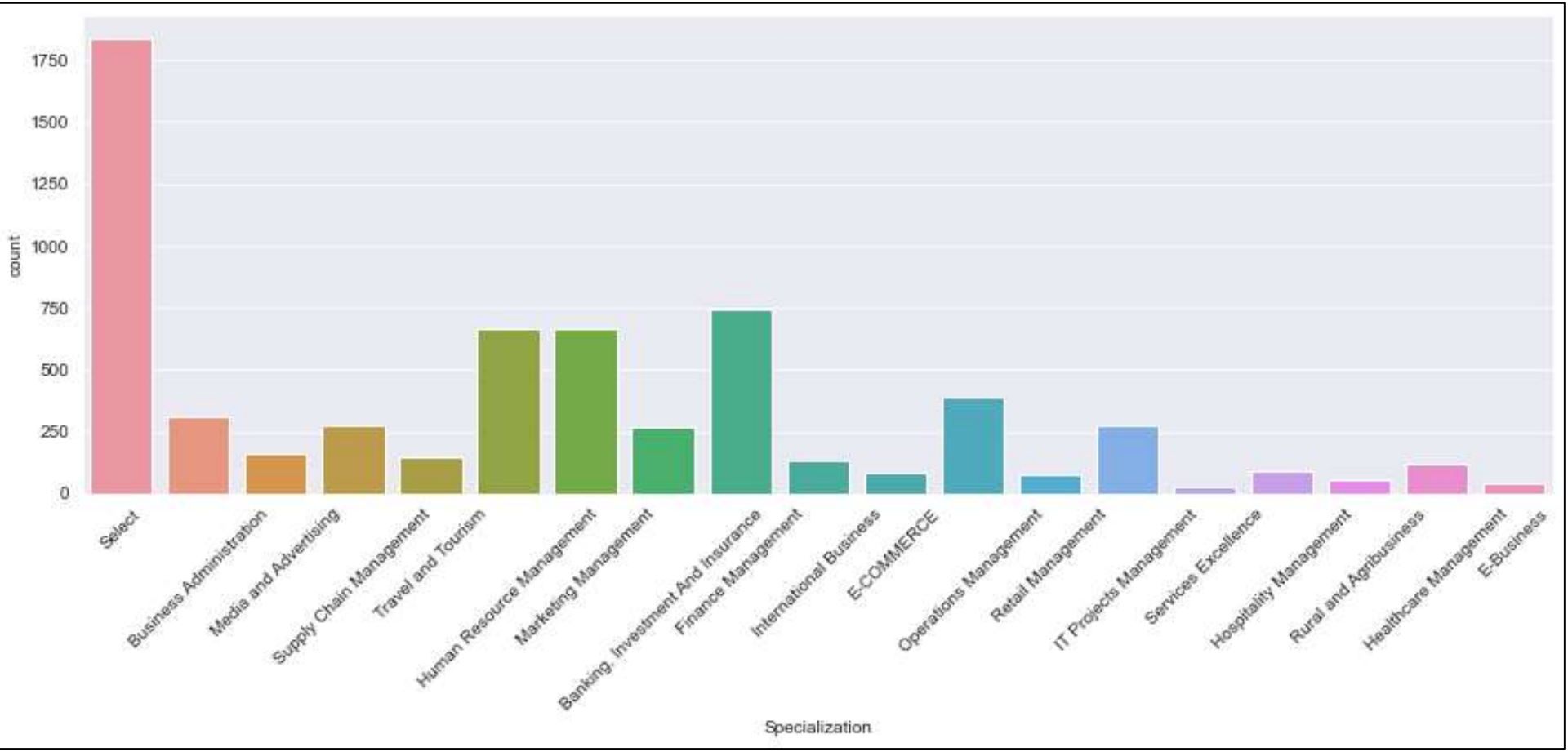
- **Maximum number of professionals comes to the company's website through Google**
- **Most of the professionals comes directly to the company's website and search for the courses**

OCCUPATION STATUS



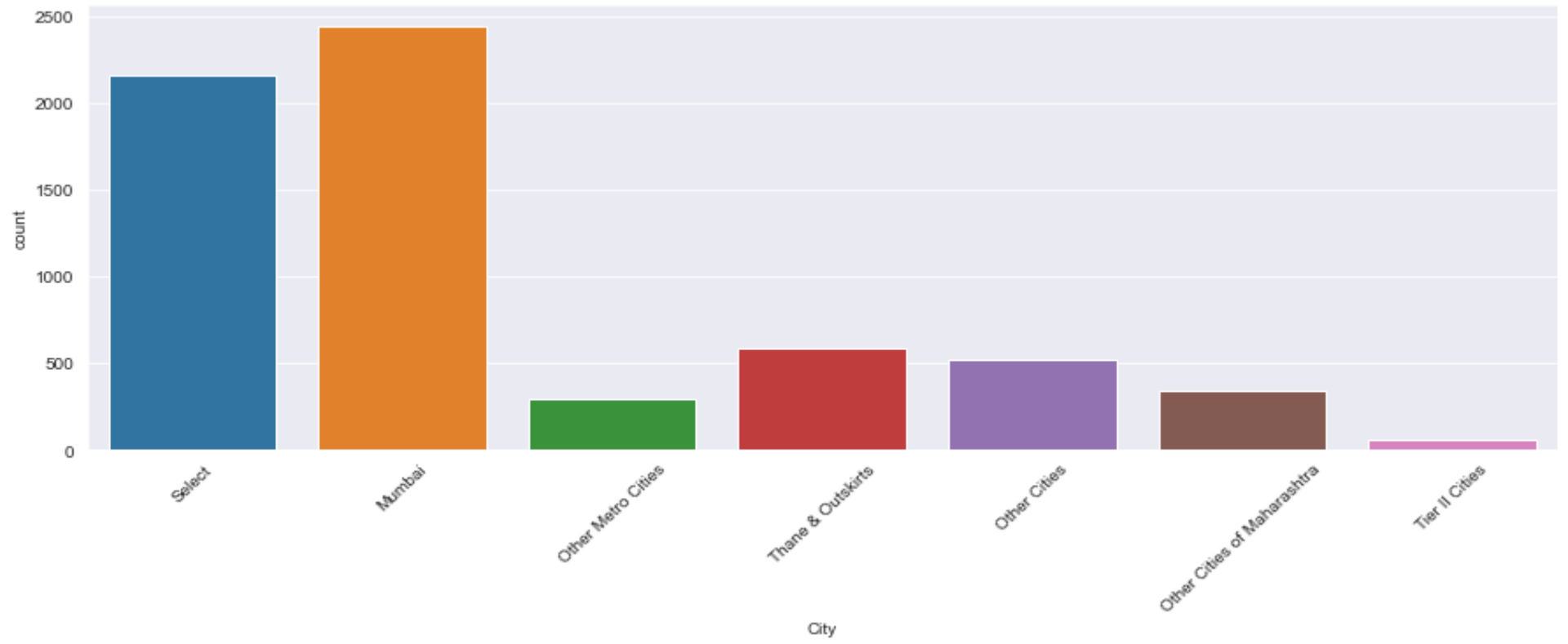
- **Maximum number of people who filled the form are unemployed**
- **Working professionals and students have less proportion in the population but working professionals are good leads that can be converted**

SPECIALIZATION



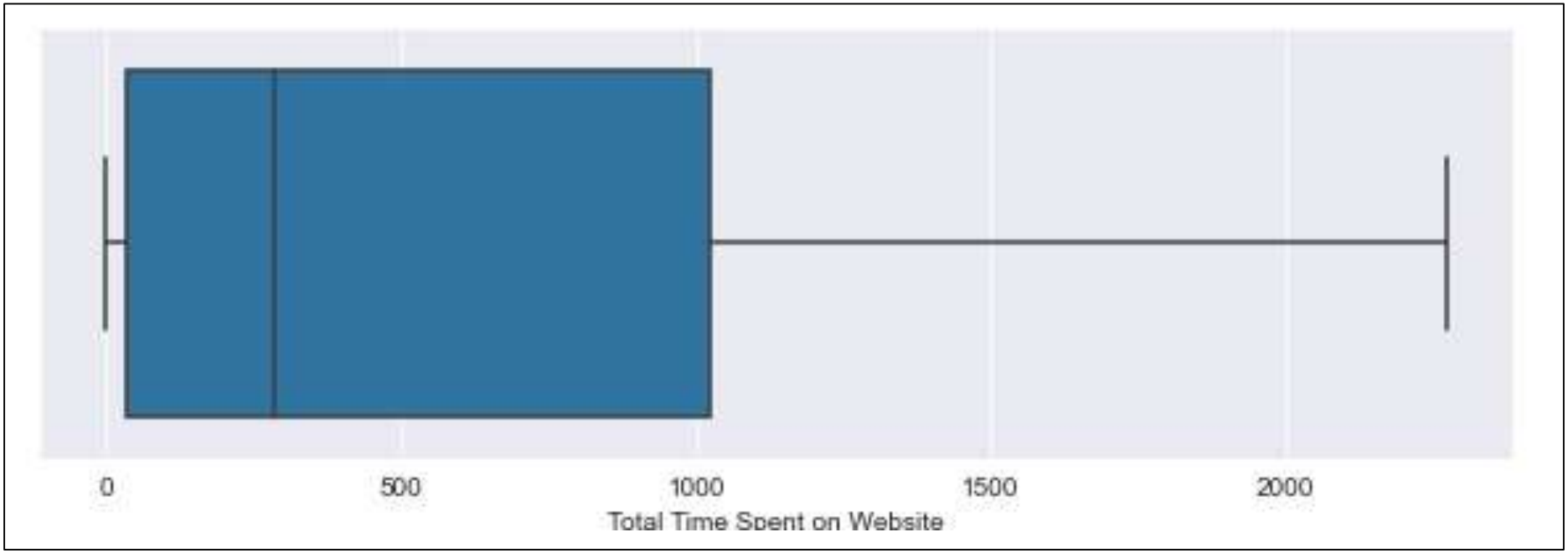
- **Maximum number of professional does not want to show their specialization**
- **Most of the working professional from Banking, investment, and insurance sector**
- **Significant number of working professional are from Travel industry & HR management**

CITY



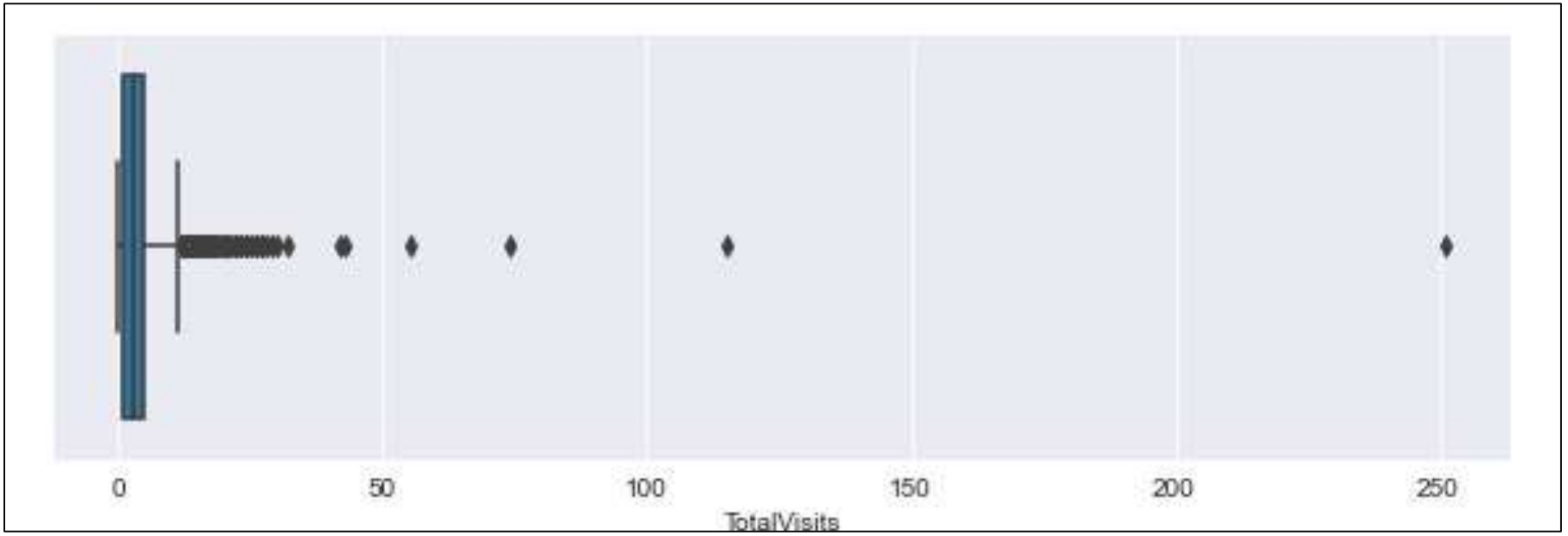
- **Most of the working professionals belongs to Mumbai, Thane and Outskirts**

Total Time Spent on Website



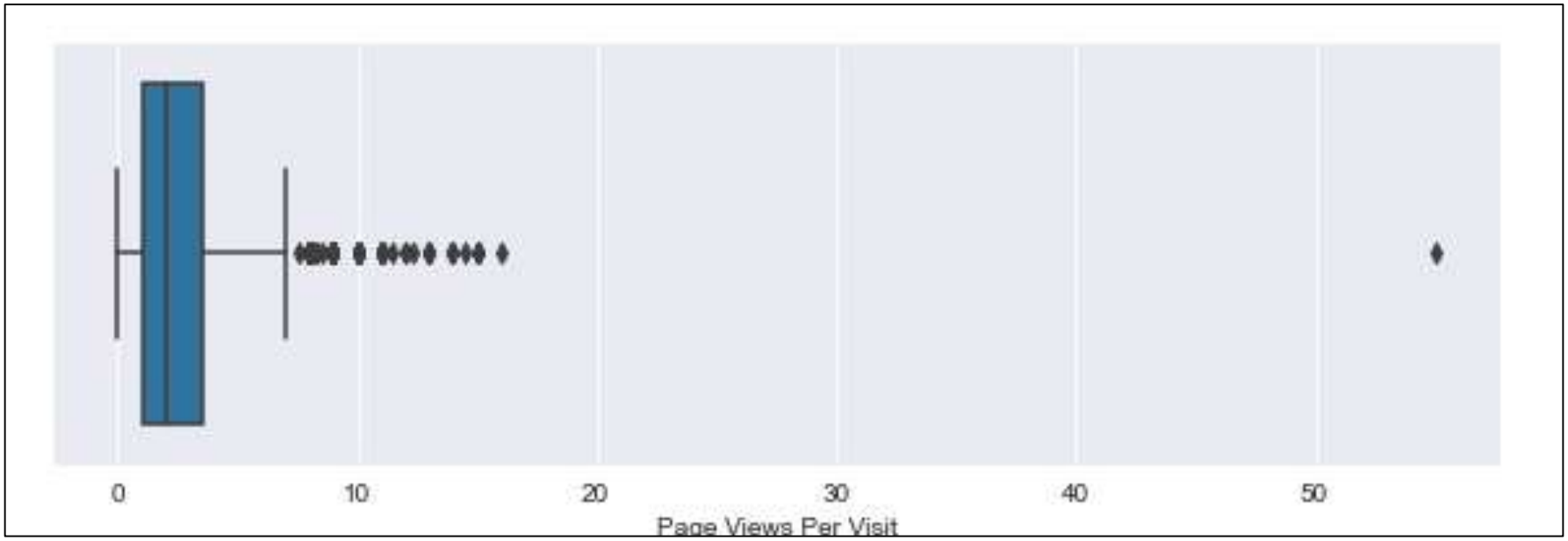
- **The median and average time spend by customer on the website is 287 and 535.22 units respectively.**

Total Visits



- **On an average each customer makes 4 visits to the website.**

Page Views Per Visit



- **On an average each customer records 4 page views per visit**

MODEL BUILDING

1

Model

- Logistics Regression Model

2

Feature Selection

- Recursive Feature Elimination technique
- 20 Variable selected

3

Final Model

- Based on statistical modelling we select 13 variable for lead prediction

4

Optimum Cutoff Probability

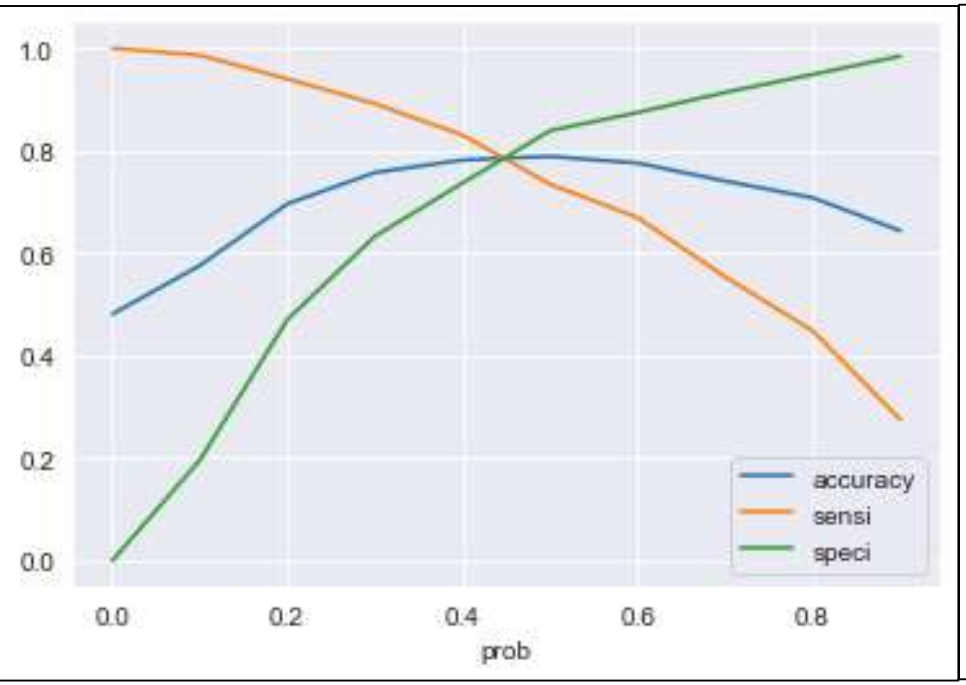
- Accuracy, sensitivity, specificity intersection
- 0.45 probability

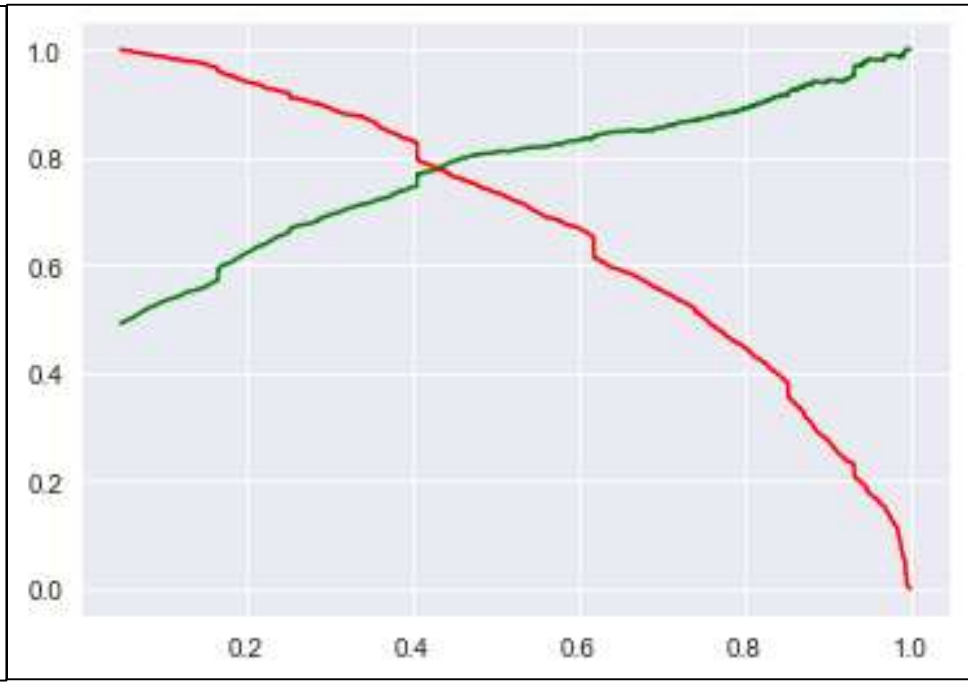
LEAD PREDICTORS

1. Lead Origin
2. Lead Source
3. Time Spent on Website
4. Current Occupation
5. Specialization of candidates
6. Last Activity
7. Last Notable Activity

DUMMY VARIABLES LEAD PREDICTORS

Finding Optimum Probability





- **On the basis of trade-off between Accuracy, Sensitivity and Specificity, we can consider initial probability for conversion as 0.42.**

METRICS (in %)

Parameter	Train Dataset	Test Dataset
Accuracy	78.65	80.28
Precision	77.40	78.97
Recall	78.55	80.43
F1 - Score	77.97	79.69

Conclusions

- The future lead conversion can be predicted based on the logistic regression model.
- Based on the evaluation matric, the obtained model predicts potential leads by 77.6%, with an Accuracy of 79.86%.
- The variable which will used for future lead prediction are:
    - Do Not Email
    - Lead Origin_Landing Page Submission
    - Lead Source_Reference
    - Last Activity_Email Opened
    - Last Activity_SMS Sent
    - LNA_Modified
    - LNA_Olark Chat Conversation
    - Total Time Spent on Website
    - Lead Source_Olark Chat
    - Lead Source_Welingak Website
    - Last Activity_Had a Phone Conversation
    - WIYCO_Working Professional
    - LNA_Unreachable

THANK YOU