# CASE STUDY:
# LEAD SCORING

Group Members :

1.Satyam Singh

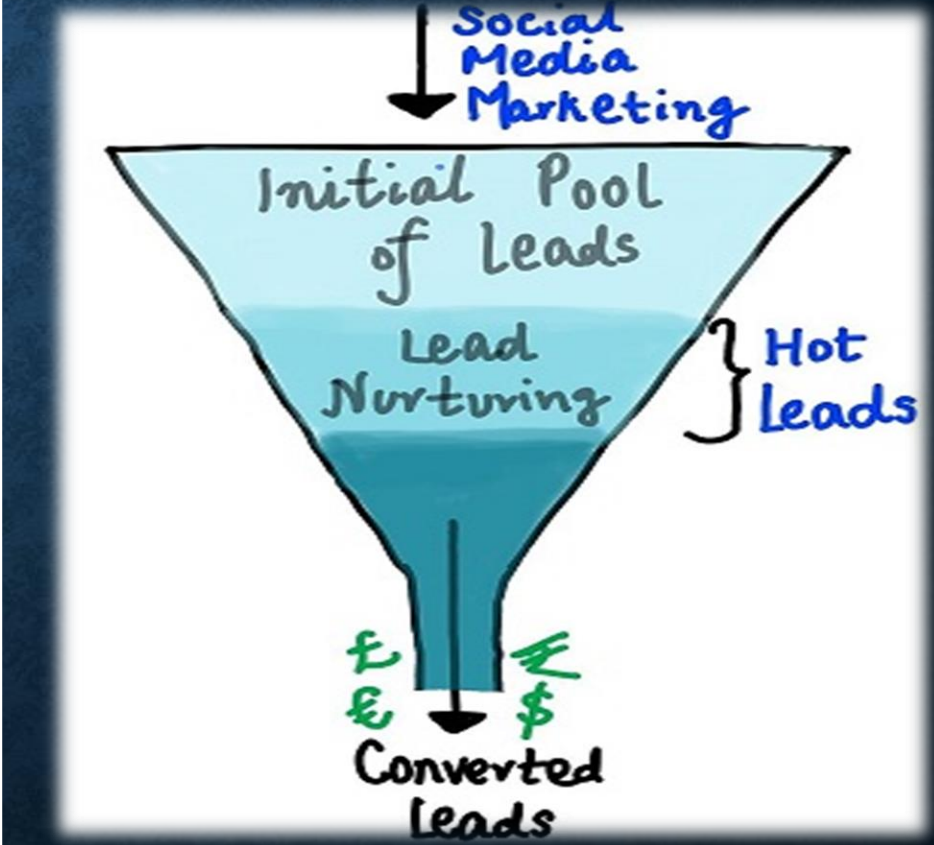
2. Somnath Nemani

3.Pushyamitra Sharma

# PROBLEM STATEMENT

- X Education Sells Online Courses To Industry Professionals And Many Professionals Browse Through Their Website For Courses. When These People Fill Up A Form And Provide Their Email Address And Their Phone Number They Are Classified As A Lead.
- The Lead Conversion Rate At X Education Is Around 30% ( If They Acquire 100 Leads In A Day Only 30 Would Get Converted) Which Clearly Is A Very Poor Rate.
- The Company Wishes To Identify The Most Potential Leads Which Are " Hot Leads".
- If They Identify This Set Of Leads, The Lead Conversion Rate Would Go Up As The Sales Team Would Focus On Communicating With The Potential Leads Rather Than Making Calls To Everyone

A Typical Lead Conversion Process Can Be Represented Using The Following Funnel:



# BUSINESS OBJECTIVE

X Education is keen on pinpointing the most promising leads, emphasizing those with a high probability of transitioning into paying customers. Our assignment involves crafting a model to allocate lead scores, strategically attributing higher scores to customers with an elevated conversion likelihood, aligning with the CEO's target conversion rate of around 80%. It's imperative that the model is designed for long-term applicability, ensuring its effectiveness in future scenarios.

# METHODOLOGY USED

**Library Import:**

Initiate the process by importing crucial libraries like Pandas, NumPy, and others essential for data analysis. Additionally, load the provided dataset to facilitate subsequent steps in the analysis.

**Data Cleaning and Manipulation:**

**Missing Values Assessment**: Conduct a thorough examination of the dataset to identify the presence and extent of missing values in various columns.

**Null Percentage Calculation**: Calculate the percentage of null values for each column, providing insights into the overall data completeness.

**Dropping Irrelevant Columns**: Evaluate columns with a substantial number of missing values, opting to eliminate those that do not contribute significantly to the analytical objectives.

# METHODOLOGY USED

**Imputation of Missing Values**: Where applicable, perform imputation of missing values to enhance the completeness of the dataset. This ensures a more robust foundation for subsequent analyses.

**Outlier Detection and Handling:** Scrutinize the dataset for outliers—data points significantly deviating from the norm. Implement appropriate strategies to handle these outliers, maintaining the integrity and reliability of the dataset for subsequent analytical procedures.

# METHODOLOGY USED

**Exploratory Data Analysis (EDA)**

## Univariate Analysis

**Value Count**: Explore the frequency distribution of individual variables, gaining insights into the occurrence of unique values within each attribute.

**Distribution of Variables**: Analyze the distribution patterns of individual variables, providing a comprehensive understanding of the range and spread of each variable.

## Bivariate Analysis

**Correlation Coefficients**: Investigate the relationships between pairs of variables by calculating correlation coefficients. This quantifies the strength and direction of linear associations.

**Patterns Between Variables**: Explore patterns, dependencies, and interactions between variables, unraveling insights into how changes in one variable relate to changes in another.

*This two-fold approach to Exploratory Data Analysis aims to uncover intrinsic patterns, relationships, and characteristics within the dataset, setting the stage for informed decision-making and further analysis.*

# METHODOLOGY USED

## Data Preparation

**Creating Dummy Variables for Categorical Variables:** Generate dummy variables for categorical attributes, converting them into a format suitable for machine learning models. Exclude the first dummy variable to avoid multicollinearity issues.

**Test-Train Split**: Divide the dataset into training and testing sets to evaluate model performance. This ensures that the model is trained on one subset and tested on another, providing a reliable assessment of its generalizability.

**Feature Scaling**: Normalize or standardize the features to ensure that variables with different scales do not unduly influence the machine learning model. This step enhances the model's performance and convergence during training.

# METHODOLOGY USED

**Model Building**

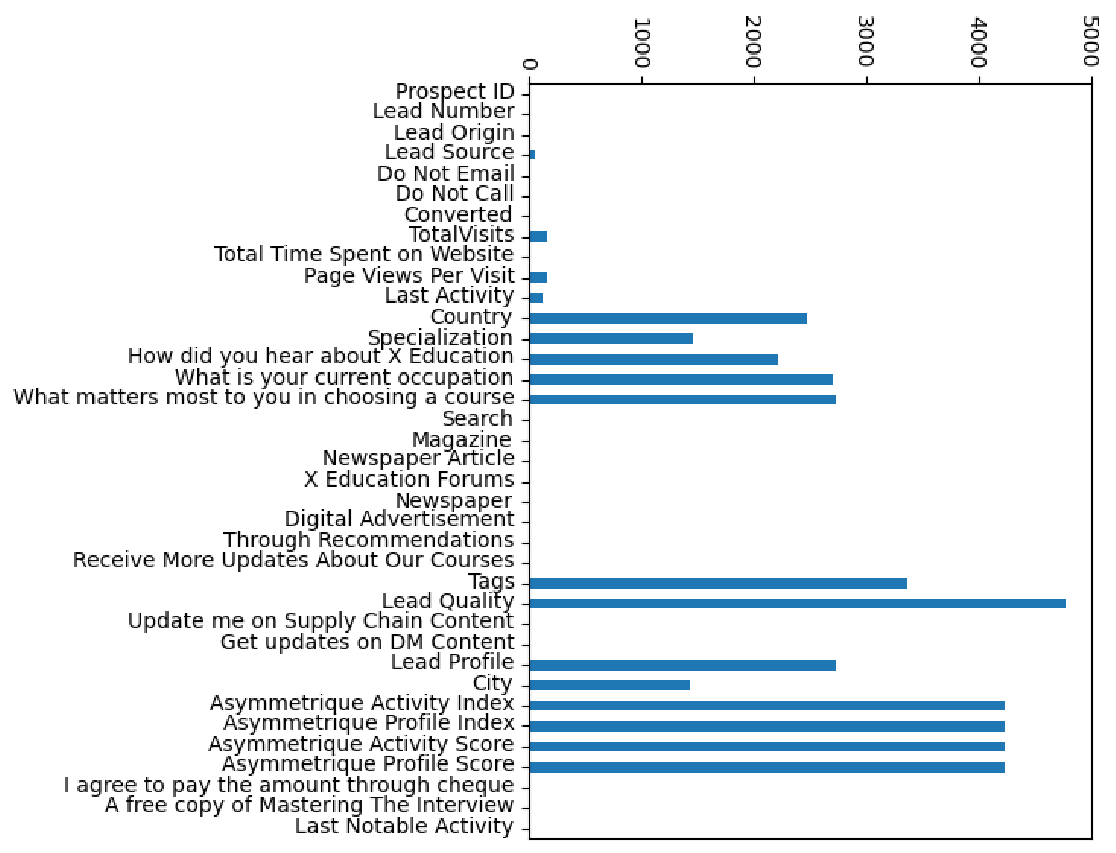
**Logistic Regression:** Employ Logistic Regression for model construction and prediction, particularly suited for binary classification tasks.

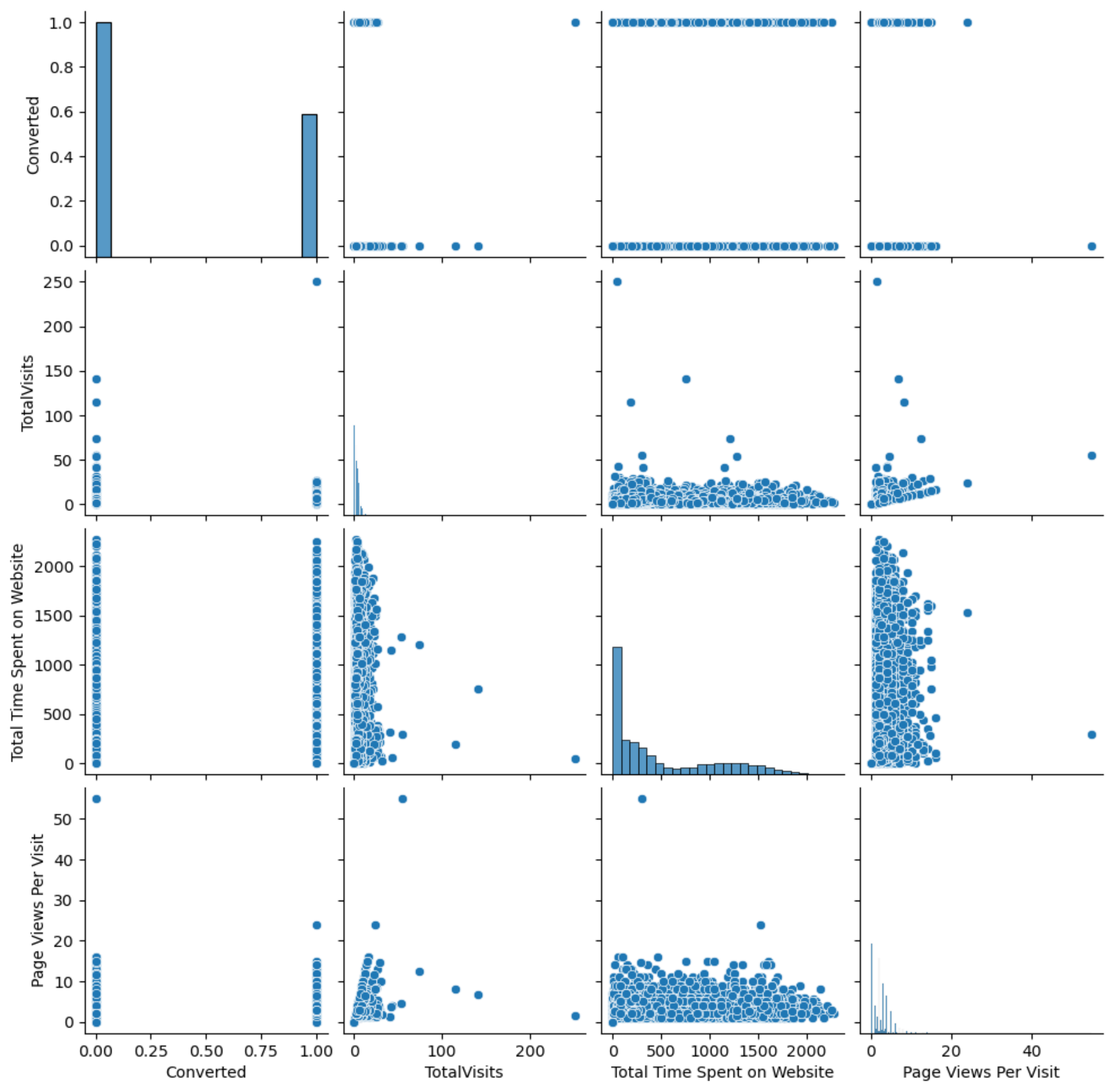
**Validation of the Model**: Assess the model's performance using appropriate validation techniques to ensure its accuracy and reliability in making predictions.

**Model Presentation**: Present the finalized model, highlighting key insights, coefficients, and any pertinent information that aids in understanding its predictive capabilities.

**Conclusions and Recommendations**: Summarize findings, draw conclusions from the analysis, and provide actionable recommendations based on the model's insights to inform decision-making.

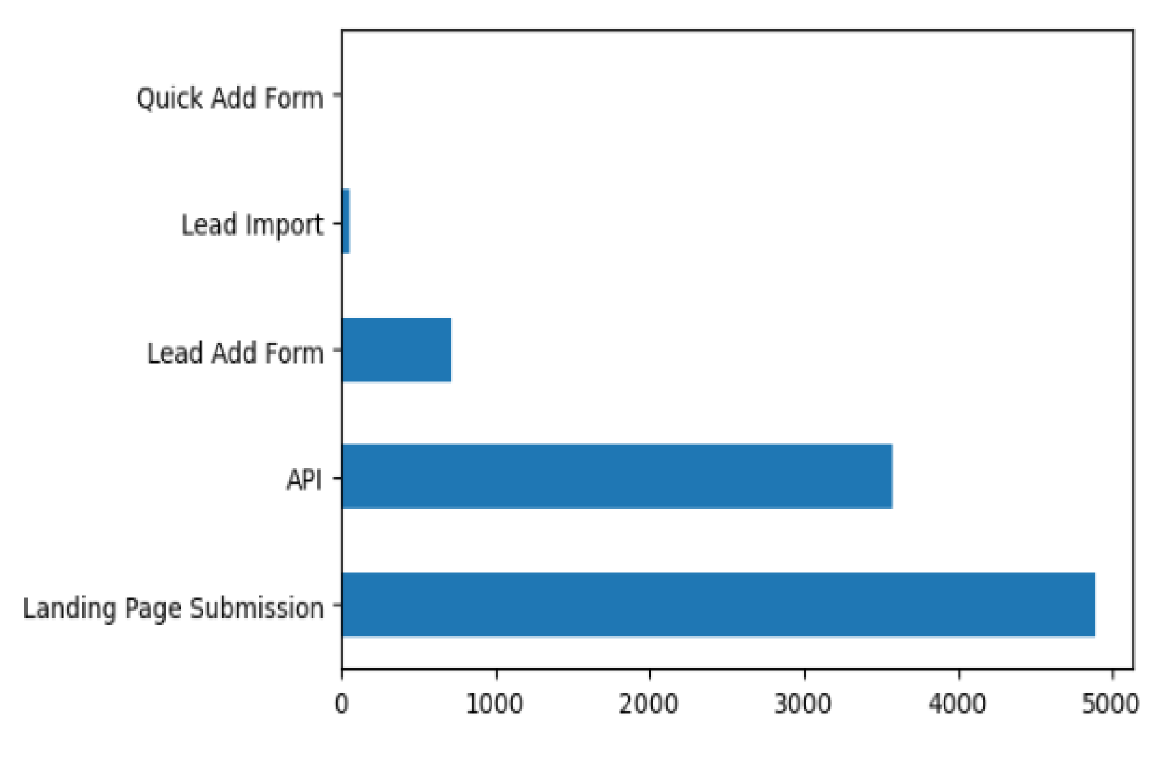
# EXPLORATORY DATA ANALYSIS



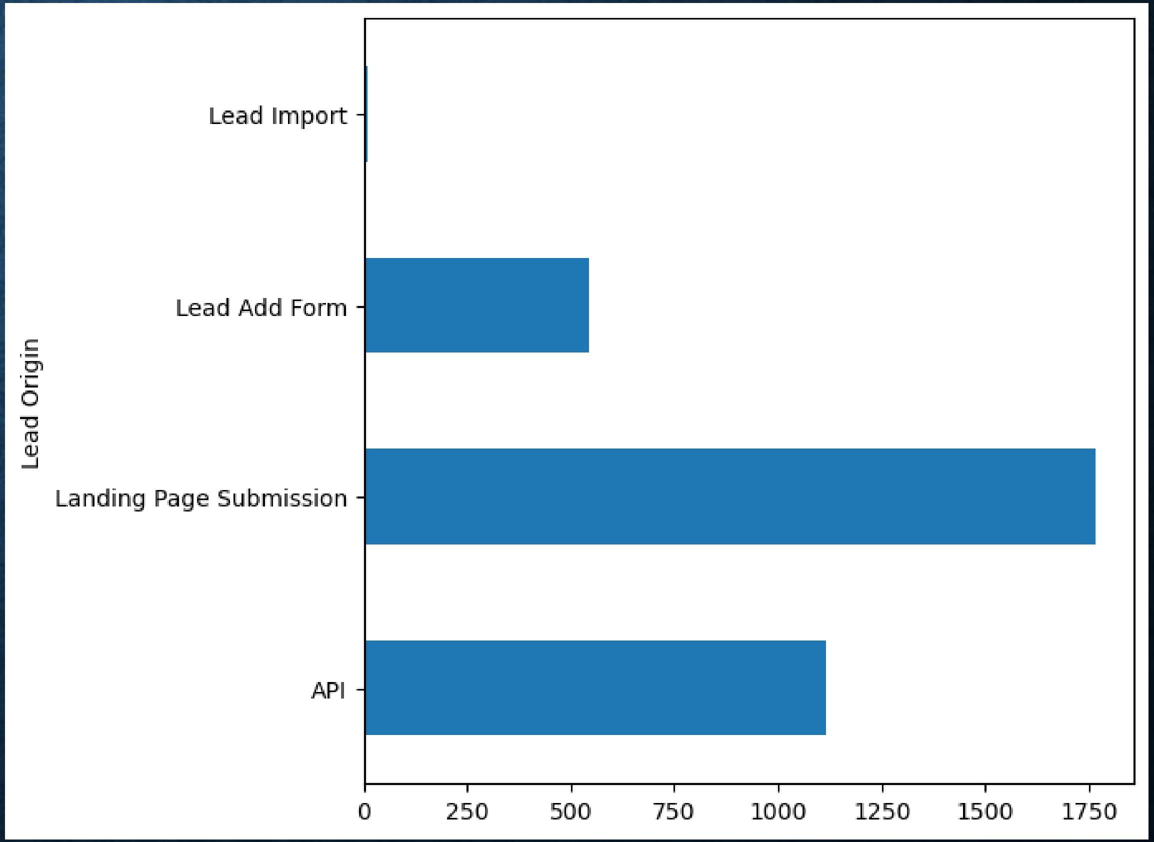


# EXPLORATORY DATA ANALYSIS

Number of people getting from the source

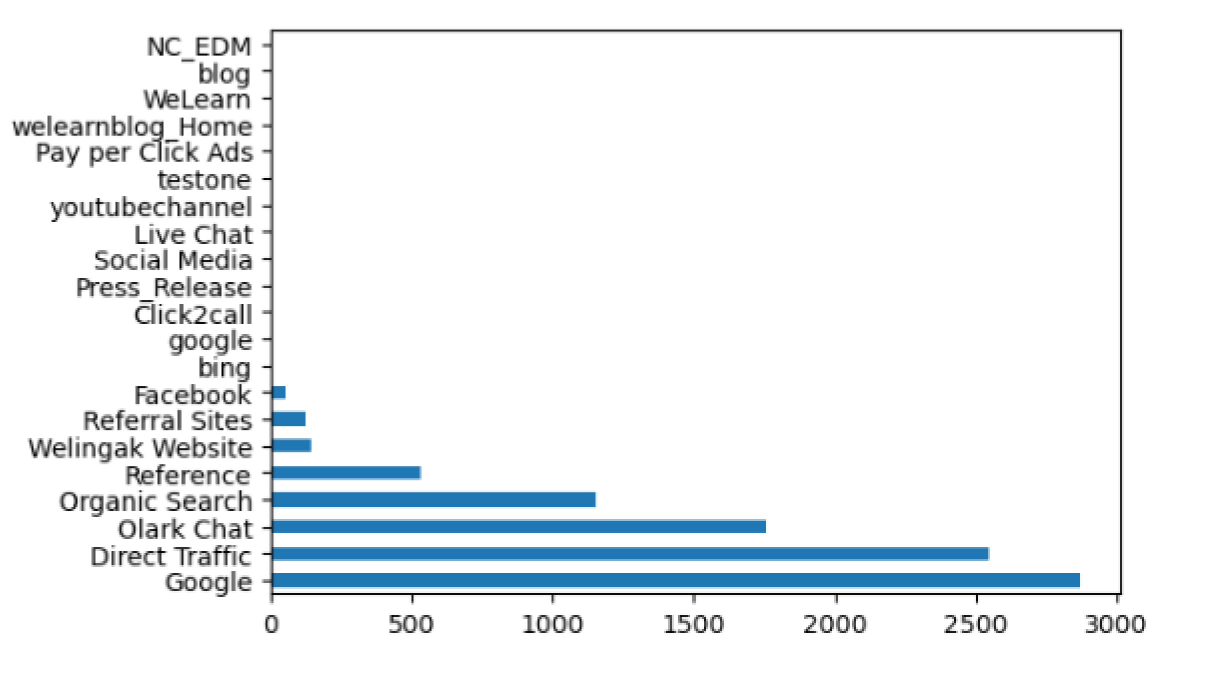


Number of people getting converted from source

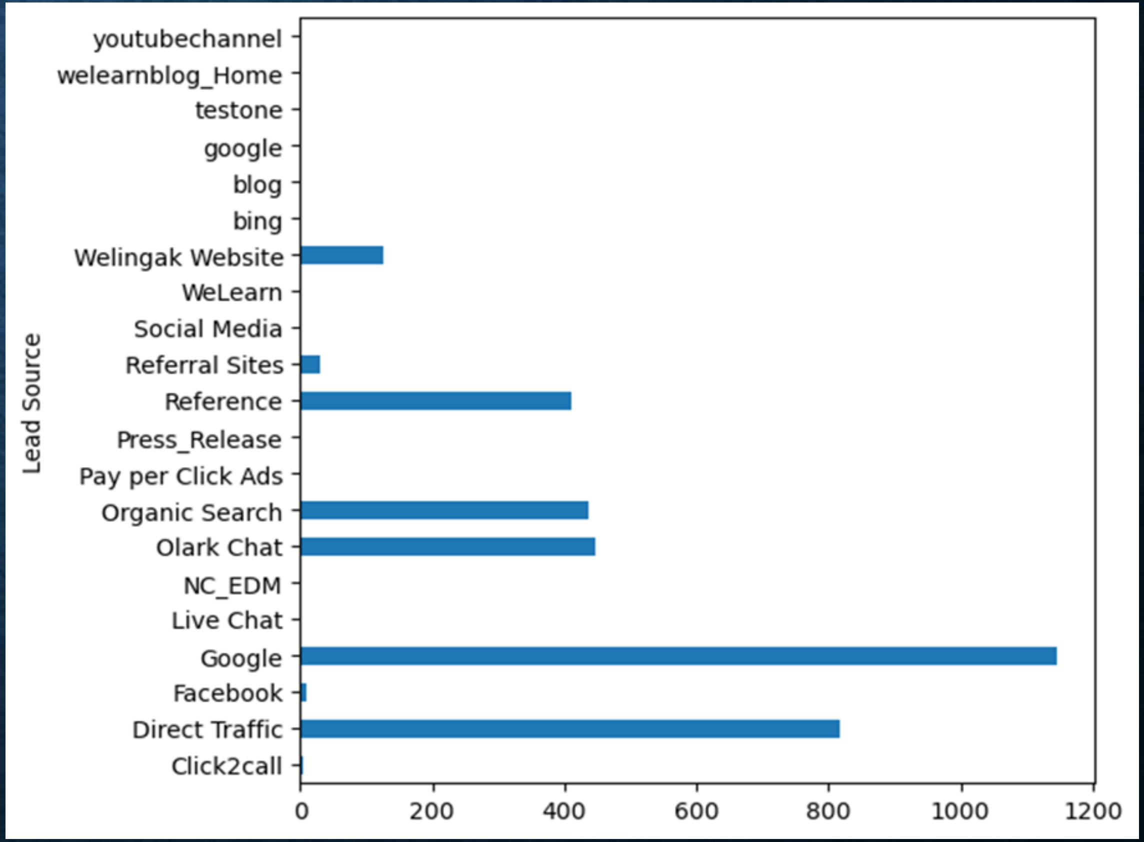


# EXPLORATORY DATA ANALYSIS

Number of people getting from the source

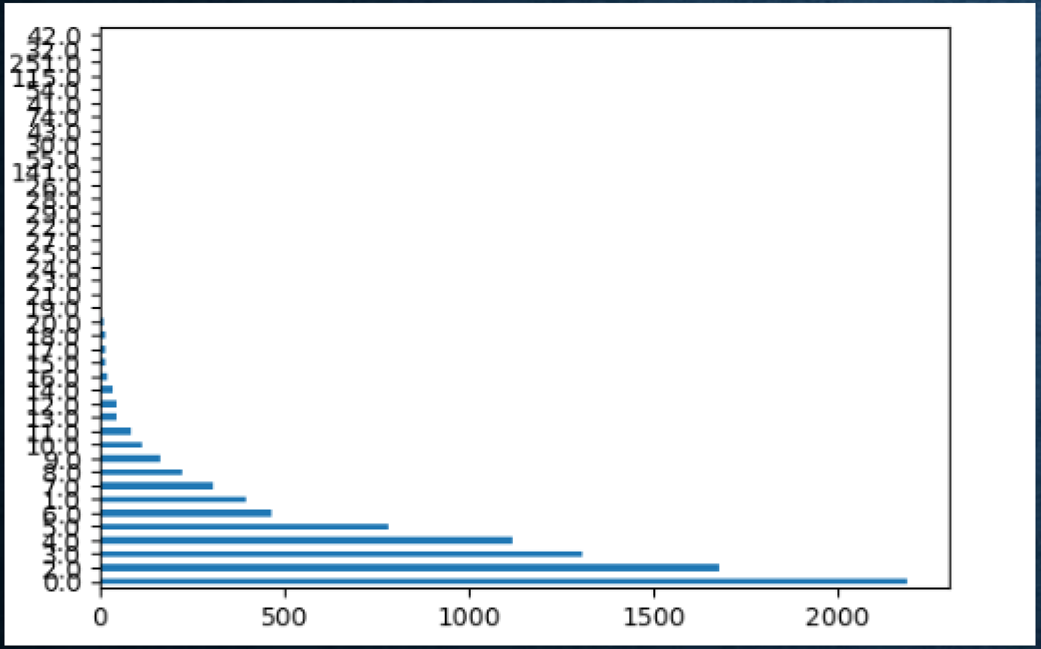


Number of people getting converted from source

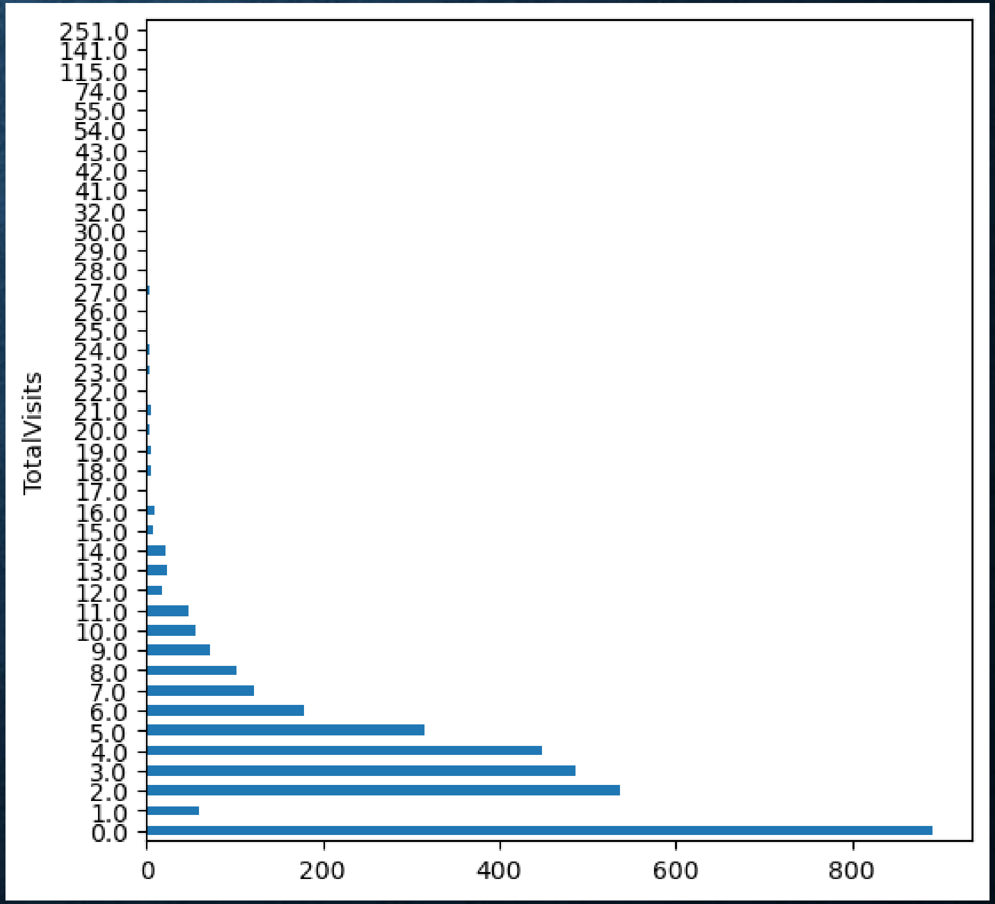


# EXPLORATORY DATA ANALYSIS

Number of people getting from the source



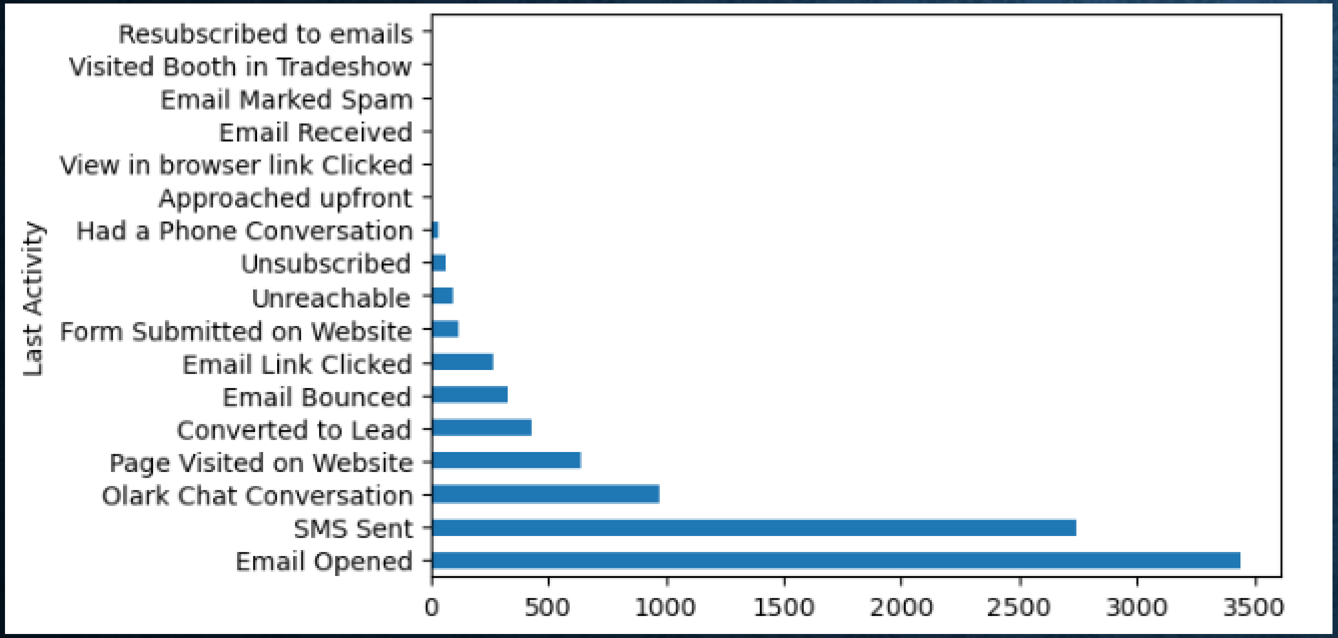
Number of people getting converted from source

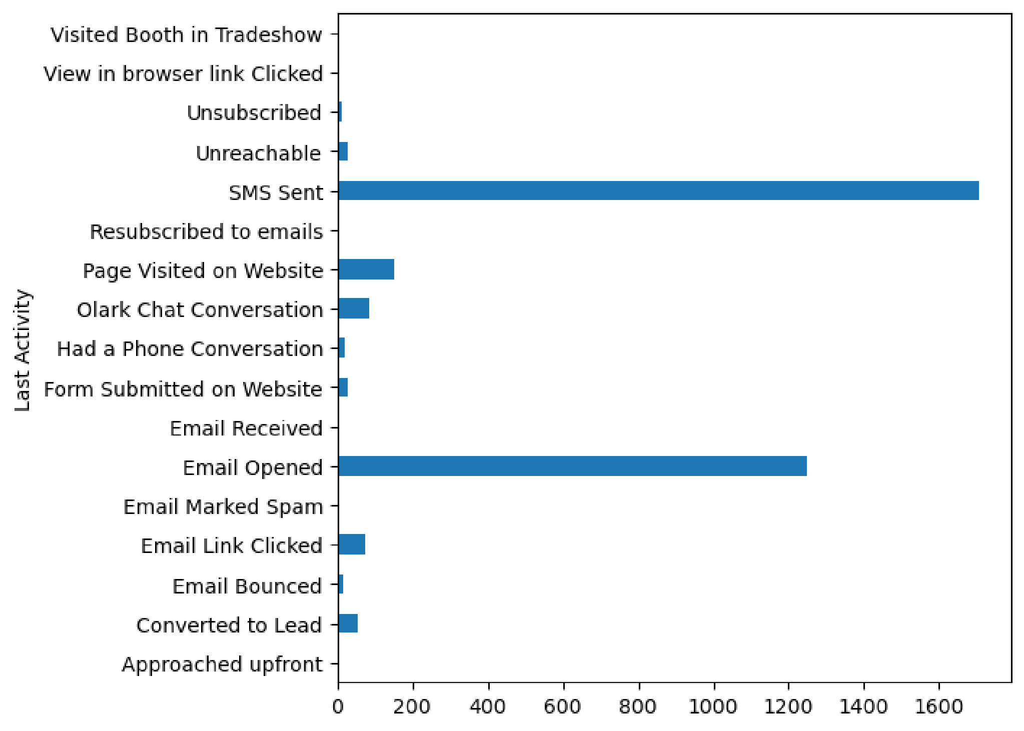


# EXPLORATORY DATA ANALYSIS

Number of people getting from the source

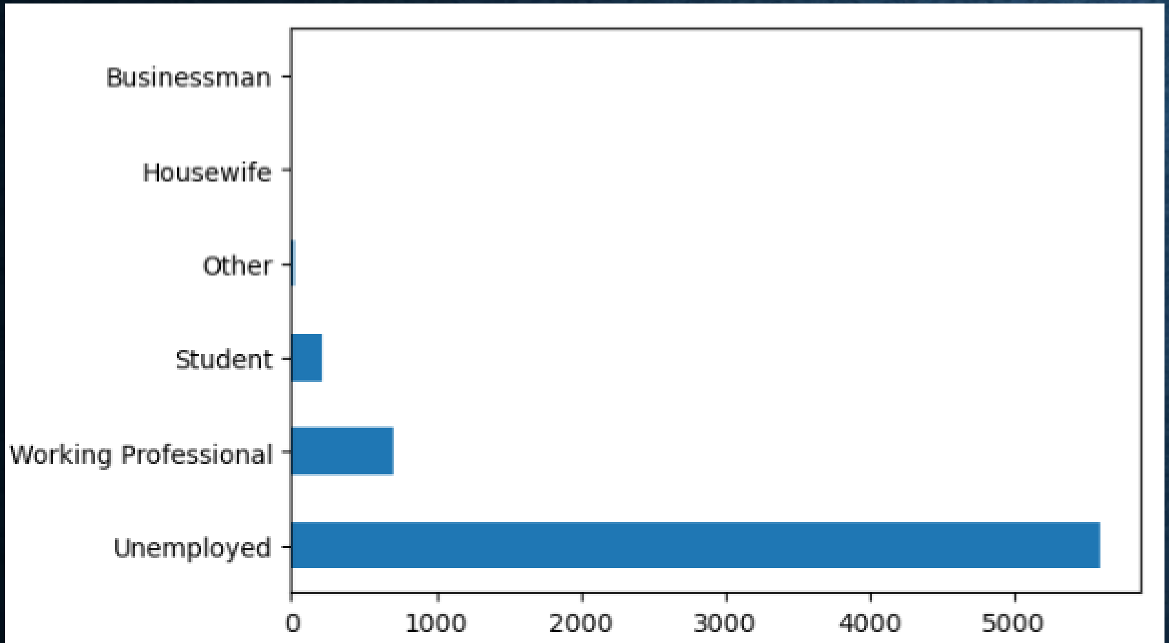
Number of people getting converted from source



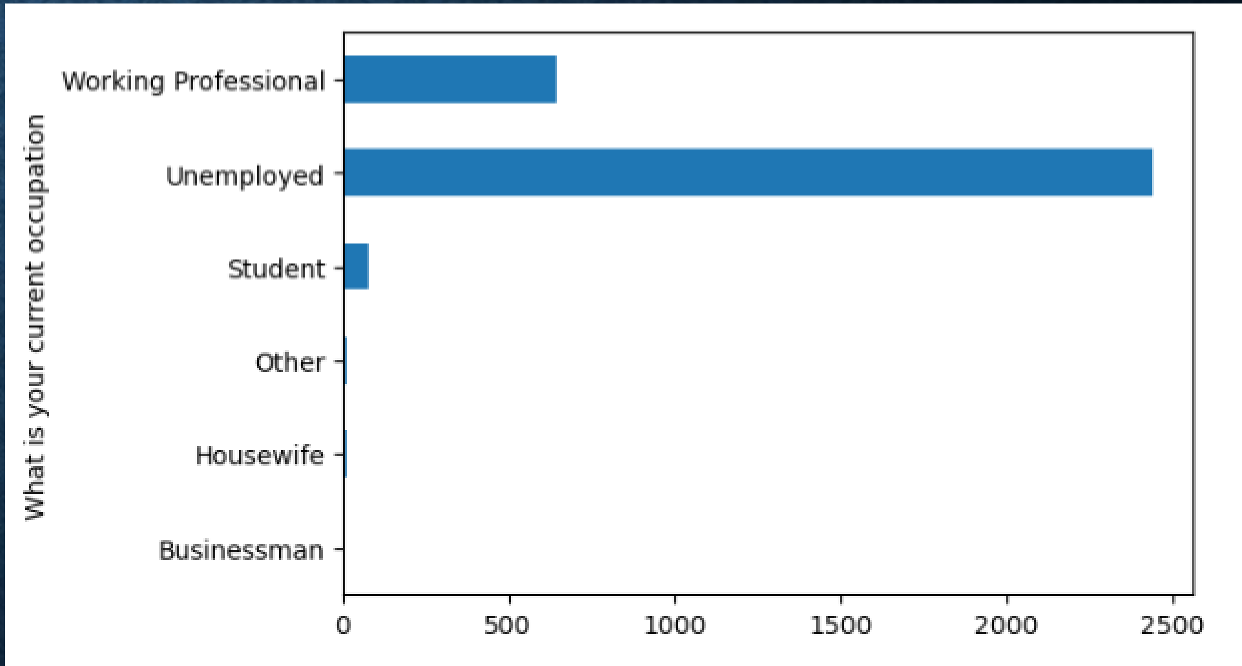


# EXPLORATORY DATA ANALYSIS

Number of people getting from the source

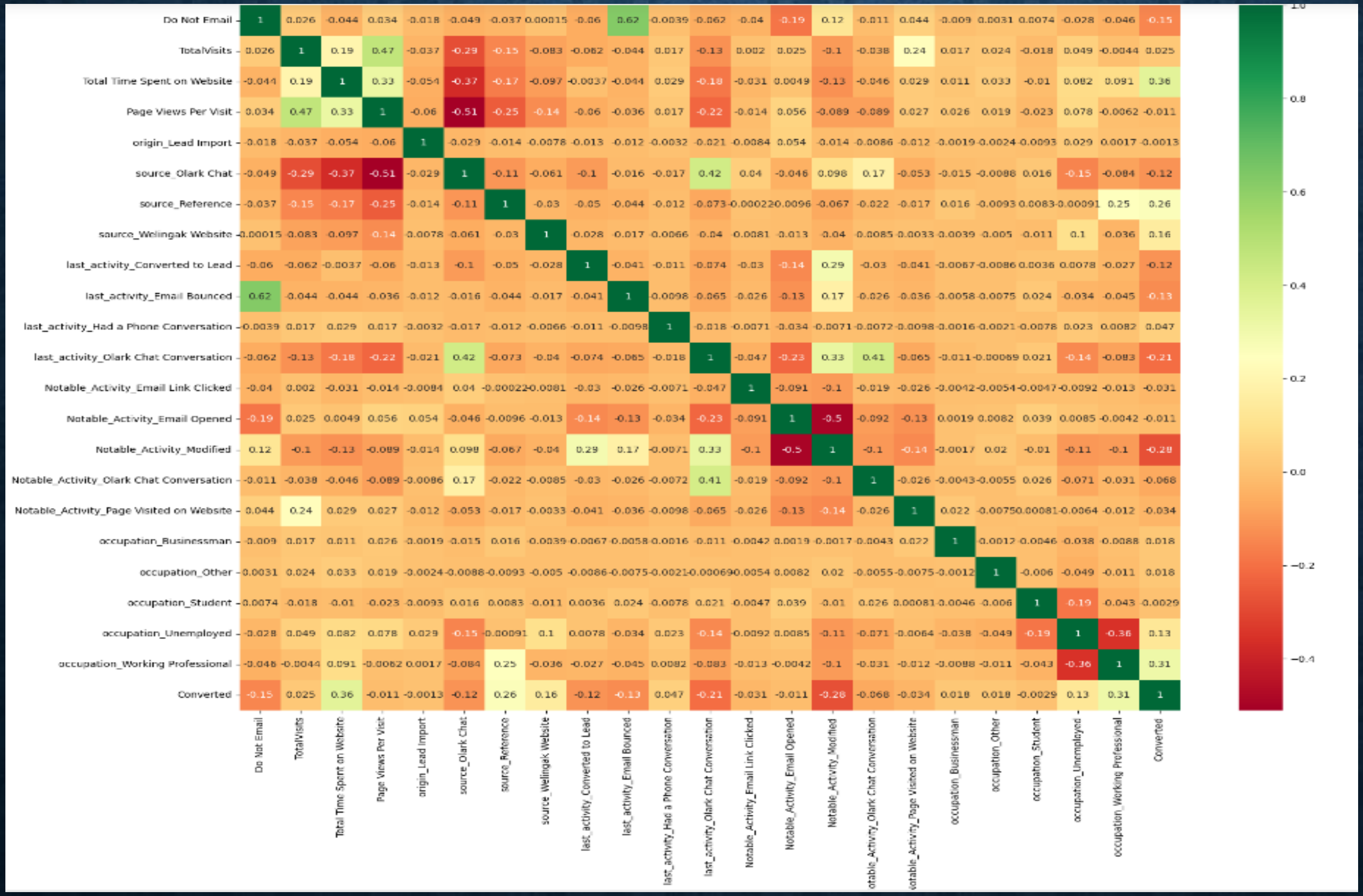


Number of people getting converted from source



# EXPLORATORY DATA ANALYSIS

Contribution of the final variables



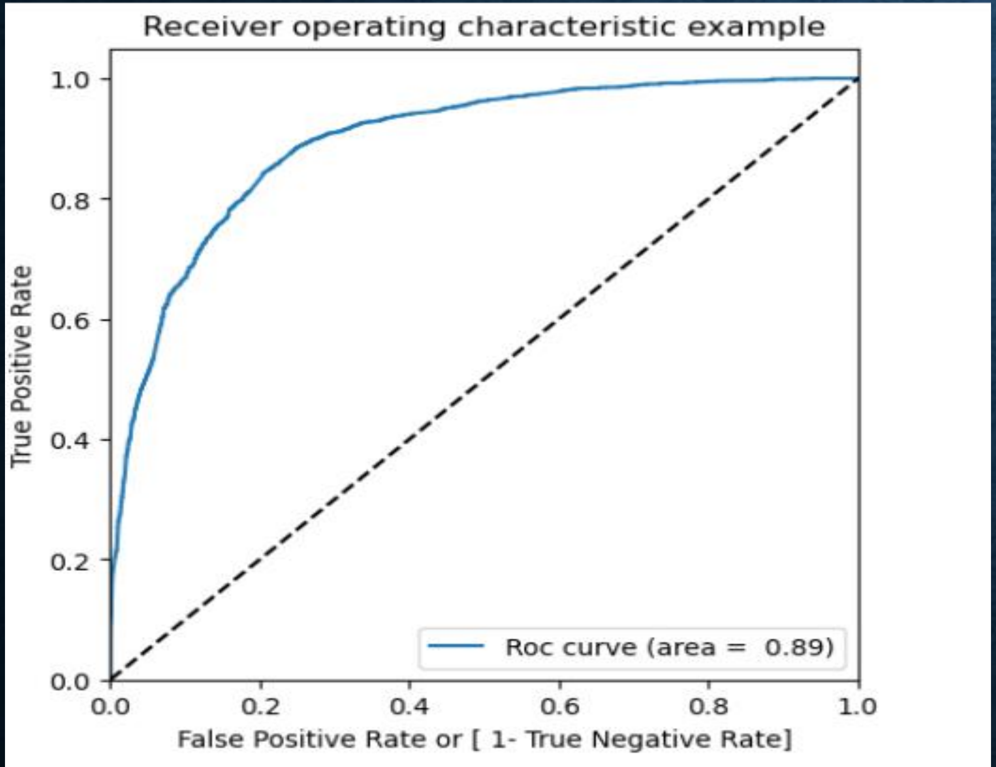
| | Do Not Email | TotalVisits | Total Time Spent on Website | Page Views Per Visit | origin_Lead Import | source_Olark Chat | source_Reference | source_Welingak Website | last_activity_Converted to Lead | last_activity_Email Bounced | last_activity_Had a Phone Conversation | last_activity_Olark Chat Conversation | Notable_Activity_Email Link Clicked | Notable_Activity_Email Opened | Notable_Activity_Modified | otable_Activity_Olark Chat Conversation | otable_Activity_Page Visited on Website | occupation_Businessman | occupation_Other | occupation_Student | occupation_Unemployed | occupation_Working Professional | Converted |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Do Not Email | 1 | 0.026 | -0.044 | 0.034 | -0.018 | -0.049 | -0.037 | 0.00015 | -0.06 | 0.62 | -0.0039 | -0.062 | -0.04 | -0.19 | 0.12 | -0.011 | 0.044 | -0.009 | 0.0031 | 0.0074 | -0.028 | -0.046 | -0.15 |
| TotalVisits | 0.026 | 1 | 0.19 | 0.47 | -0.037 | -0.29 | -0.15 | -0.083 | -0.062 | -0.044 | 0.017 | -0.13 | 0.002 | 0.025 | -0.1 | -0.038 | 0.24 | 0.017 | 0.024 | -0.018 | 0.049 | -0.0044 | 0.025 |
| Total Time Spent on Website | -0.044 | 0.19 | 1 | 0.33 | -0.054 | -0.37 | -0.17 | -0.097 | -0.0037 | -0.044 | 0.029 | -0.18 | -0.031 | 0.0049 | -0.13 | -0.046 | 0.029 | 0.011 | 0.033 | -0.01 | 0.082 | 0.091 | 0.36 |
| Page Views Per Visit | 0.034 | 0.47 | 0.33 | 1 | -0.06 | -0.51 | -0.25 | -0.14 | -0.06 | -0.036 | 0.017 | -0.22 | -0.014 | 0.056 | -0.089 | -0.089 | 0.027 | 0.026 | 0.019 | -0.023 | 0.078 | -0.0062 | -0.011 |
| origin_Lead Import | -0.018 | -0.037 | -0.054 | -0.06 | 1 | -0.029 | -0.014 | -0.0078 | -0.013 | -0.012 | -0.0032 | -0.021 | -0.0084 | 0.054 | -0.014 | -0.0086 | -0.012 | -0.0019 | -0.0024 | -0.0093 | 0.029 | 0.0017 | -0.0013 |
| source_Olark Chat | -0.049 | -0.29 | -0.37 | -0.51 | -0.029 | 1 | -0.11 | -0.061 | -0.1 | -0.016 | -0.017 | 0.42 | 0.04 | -0.046 | 0.098 | 0.17 | -0.053 | -0.015 | -0.0088 | 0.016 | -0.15 | -0.084 | -0.12 |
| source_Reference | -0.037 | -0.15 | -0.17 | -0.25 | -0.014 | -0.11 | 1 | -0.03 | -0.05 | -0.044 | -0.012 | -0.073 | -0.00022 | -0.0096 | -0.067 | -0.022 | -0.017 | 0.016 | -0.0093 | 0.0083 | -0.00091 | 0.25 | 0.26 |
| source_Welingak Website | 0.00015 | -0.083 | -0.097 | -0.14 | -0.0078 | -0.061 | -0.03 | 1 | -0.028 | -0.017 | -0.0066 | -0.04 | -0.0081 | -0.013 | -0.04 | -0.0085 | -0.0033 | -0.0039 | -0.005 | -0.011 | 0.1 | -0.036 | 0.16 |
| last_activity_Converted to Lead | -0.06 | -0.062 | -0.0037 | -0.06 | -0.013 | -0.1 | -0.05 | -0.028 | 1 | -0.041 | -0.011 | -0.074 | -0.03 | -0.14 | 0.29 | -0.03 | -0.041 | -0.0067 | -0.0086 | 0.0036 | 0.0078 | -0.027 | -0.12 |
| last_activity_Email Bounced | 0.62 | -0.044 | -0.044 | -0.036 | -0.012 | -0.016 | -0.044 | -0.017 | -0.041 | 1 | -0.0098 | -0.065 | -0.026 | -0.13 | 0.17 | -0.026 | -0.036 | -0.0058 | -0.0075 | 0.024 | -0.034 | -0.045 | -0.13 |
| last_activity_Had a Phone Conversation | -0.0039 | 0.017 | 0.029 | 0.017 | -0.0032 | -0.017 | -0.012 | -0.0066 | -0.011 | -0.0098 | 1 | -0.018 | -0.0071 | -0.034 | -0.0071 | -0.0072 | -0.0098 | -0.0016 | -0.0021 | -0.0078 | 0.023 | 0.0082 | 0.047 |
| last_activity_Olark Chat Conversation | -0.062 | -0.13 | -0.18 | -0.22 | -0.021 | 0.42 | -0.073 | -0.04 | -0.074 | -0.065 | -0.018 | 1 | -0.047 | -0.23 | 0.33 | 0.41 | -0.065 | -0.011 | -0.00069 | 0.021 | -0.14 | -0.083 | -0.21 |
| Notable_Activity_Email Link Clicked | -0.04 | 0.002 | -0.031 | -0.014 | -0.0084 | 0.04 | -0.00022 | -0.0081 | -0.03 | -0.026 | -0.0071 | -0.047 | 1 | -0.091 | -0.1 | -0.019 | -0.026 | -0.0042 | -0.0054 | -0.0047 | -0.0092 | -0.013 | -0.031 |
| Notable_Activity_Email Opened | -0.19 | 0.025 | 0.0049 | 0.056 | 0.054 | -0.046 | -0.0096 | -0.013 | -0.14 | -0.13 | -0.034 | -0.23 | -0.091 | 1 | -0.5 | -0.092 | -0.13 | 0.0019 | 0.0082 | 0.039 | 0.0085 | -0.0042 | -0.011 |
| Notable_Activity_Modified | 0.12 | -0.1 | -0.13 | -0.089 | -0.014 | 0.098 | -0.067 | -0.04 | 0.29 | 0.17 | -0.0071 | 0.33 | -0.1 | -0.5 | 1 | -0.1 | -0.14 | -0.0017 | 0.02 | -0.01 | -0.11 | -0.1 | -0.28 |
| Notable_Activity_Olark Chat Conversation | -0.011 | -0.038 | -0.046 | -0.089 | -0.0086 | 0.17 | -0.022 | -0.0085 | -0.03 | -0.026 | -0.0072 | 0.41 | -0.019 | -0.092 | -0.1 | 1 | -0.026 | -0.0043 | -0.0055 | 0.026 | -0.071 | -0.031 | -0.068 |
| Notable_Activity_Page Visited on Website | 0.044 | 0.24 | 0.029 | 0.027 | -0.012 | -0.053 | -0.017 | -0.0033 | -0.041 | -0.036 | -0.0098 | -0.065 | -0.026 | -0.13 | -0.14 | -0.026 | 1 | 0.022 | -0.0075 | 0.00081 | -0.0064 | -0.012 | -0.034 |
| occupation_Businessman | -0.009 | 0.017 | 0.011 | 0.026 | -0.0019 | -0.015 | 0.016 | -0.0039 | -0.0067 | -0.0058 | -0.0016 | -0.011 | -0.0042 | 0.0019 | -0.0017 | -0.0043 | 0.022 | 1 | -0.0012 | -0.0046 | -0.038 | -0.0088 | 0.018 |
| occupation_Other | 0.0031 | 0.024 | 0.033 | 0.019 | -0.0024 | -0.0088 | -0.0093 | -0.005 | -0.0086 | -0.0075 | -0.0021 | -0.00069 | -0.0054 | 0.0082 | 0.02 | -0.0055 | -0.0075 | -0.0012 | 1 | -0.006 | -0.049 | -0.011 | 0.018 |
| occupation_Student | 0.0074 | -0.018 | -0.01 | -0.023 | -0.0093 | 0.016 | 0.0083 | -0.011 | 0.0036 | 0.024 | -0.0078 | 0.021 | -0.0047 | 0.039 | -0.01 | 0.026 | 0.00081 | -0.0046 | -0.006 | 1 | -0.19 | -0.043 | -0.0029 |
| occupation_Unemployed | -0.028 | 0.049 | 0.082 | 0.078 | 0.029 | -0.15 | -0.00091 | 0.1 | 0.0078 | -0.034 | 0.023 | -0.14 | -0.0092 | 0.0085 | -0.11 | -0.071 | -0.0064 | -0.038 | -0.049 | -0.19 | 1 | -0.36 | 0.13 |
| occupation_Working Professional | -0.046 | -0.0044 | 0.091 | -0.0062 | 0.0017 | -0.084 | 0.25 | -0.036 | -0.027 | -0.045 | 0.0082 | -0.083 | -0.013 | -0.0042 | -0.1 | -0.031 | -0.012 | -0.0088 | -0.011 | -0.043 | -0.36 | 1 | 0.31 |
| Converted | -0.15 | 0.025 | 0.36 | -0.011 | -0.0013 | -0.12 | 0.26 | 0.16 | -0.12 | -0.13 | 0.047 | -0.21 | -0.031 | -0.011 | -0.28 | -0.068 | -0.034 | 0.018 | 0.018 | -0.0029 | 0.13 | 0.31 | 1 |

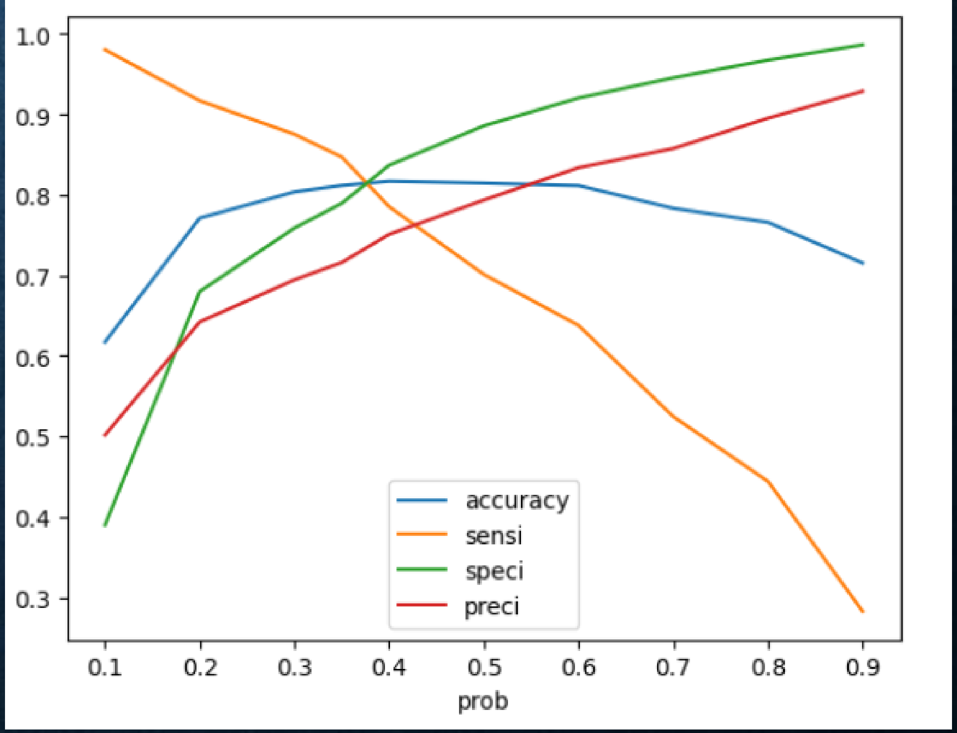
# RESULTS

1. Split the data into training and testing sets using a 70:30 ratio.
2. Employed RFE for feature selection, yielding 25 variables as output.
3. Built the model by eliminating variables with high VIF values.
4. Finalized the model with 12 variables based on a p-value threshold of 0.
5. Conducted predictions on the test dataset.
6. Achieved an overall accuracy of 81%.

# METHODOLOGY USED / RESULTS

1. Constructed an ROC curve (Receiver Operating Characteristic curve) for model evaluation.
2. Determined the optimal cutoff point using Accuracy, Sensitivity and Specificity.
3. The area under the ROC curve (AUC) is 0.89, indicating a strong discriminative ability of the model in distinguishing between classes.





# CONCLUSION

1. The model demonstrates adaptability to future company requirements.
2. Key variables influencing potential buyers include:
   1. Total time spent on the website.
   2. Lead source from olark chat ,Reference and Wellingak website.
   3. Current occupation being a working professional.
   4. Lead origin through landing page submission.