LEAD SCORE CASE STUDY

LOGISTIC REGRESSION MODEL FOR LEAD CONVERSION

Santhosh Balaji

Jyoti Singh

Bhanu Teja

Deep Sheth

OBJECTIVE OF THE ANALYSIS

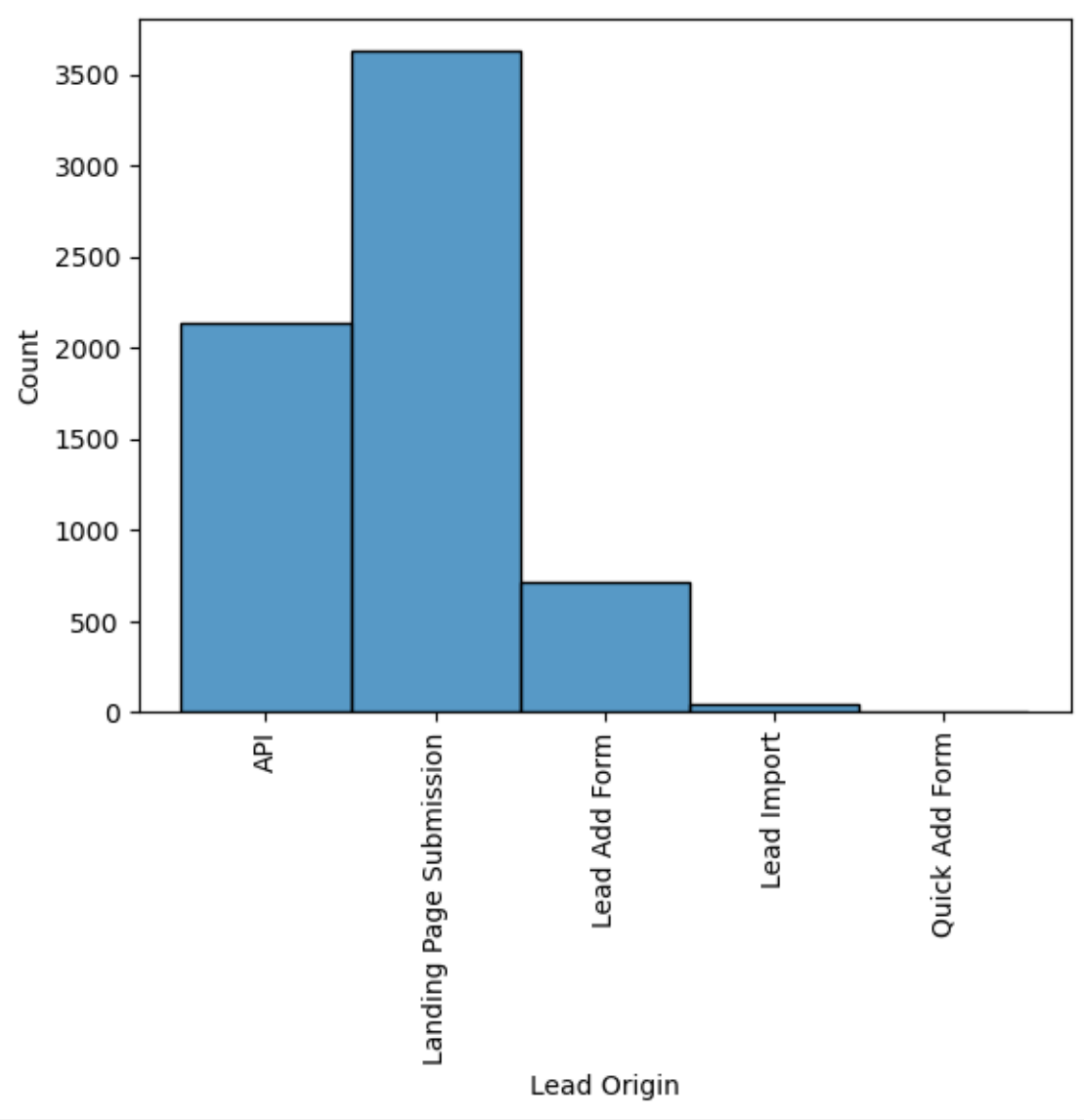
- **Key Point:**
  Analyzing factors contributing to lead conversion and optimize efforts for better conversion rates.
- **Objective:**
  - Identify significant features influencing conversion.
  - Develop and validate a logistic regression model.
  - Evaluate performance with accuracy, sensitivity, and specificity.

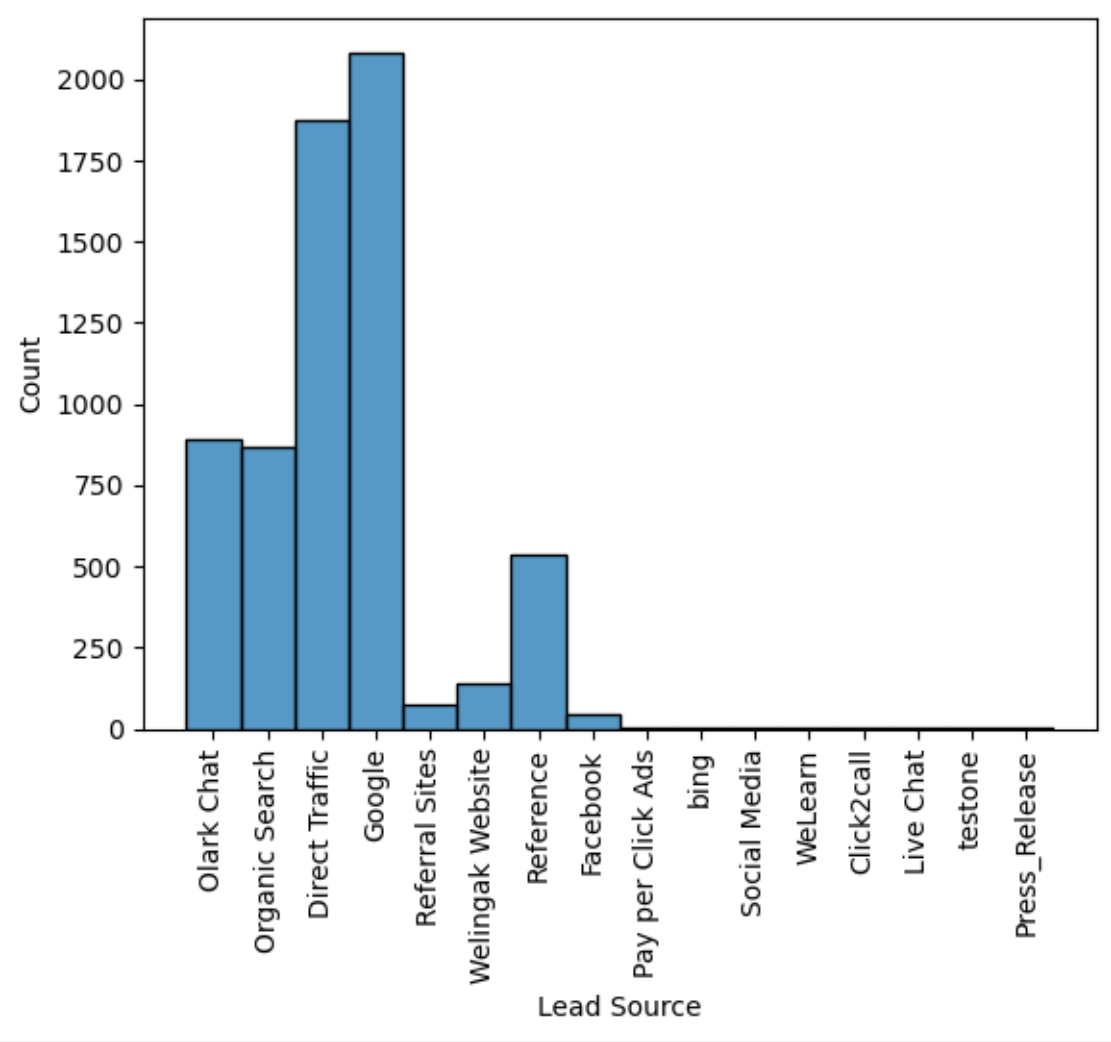
DATA OVERVIEW

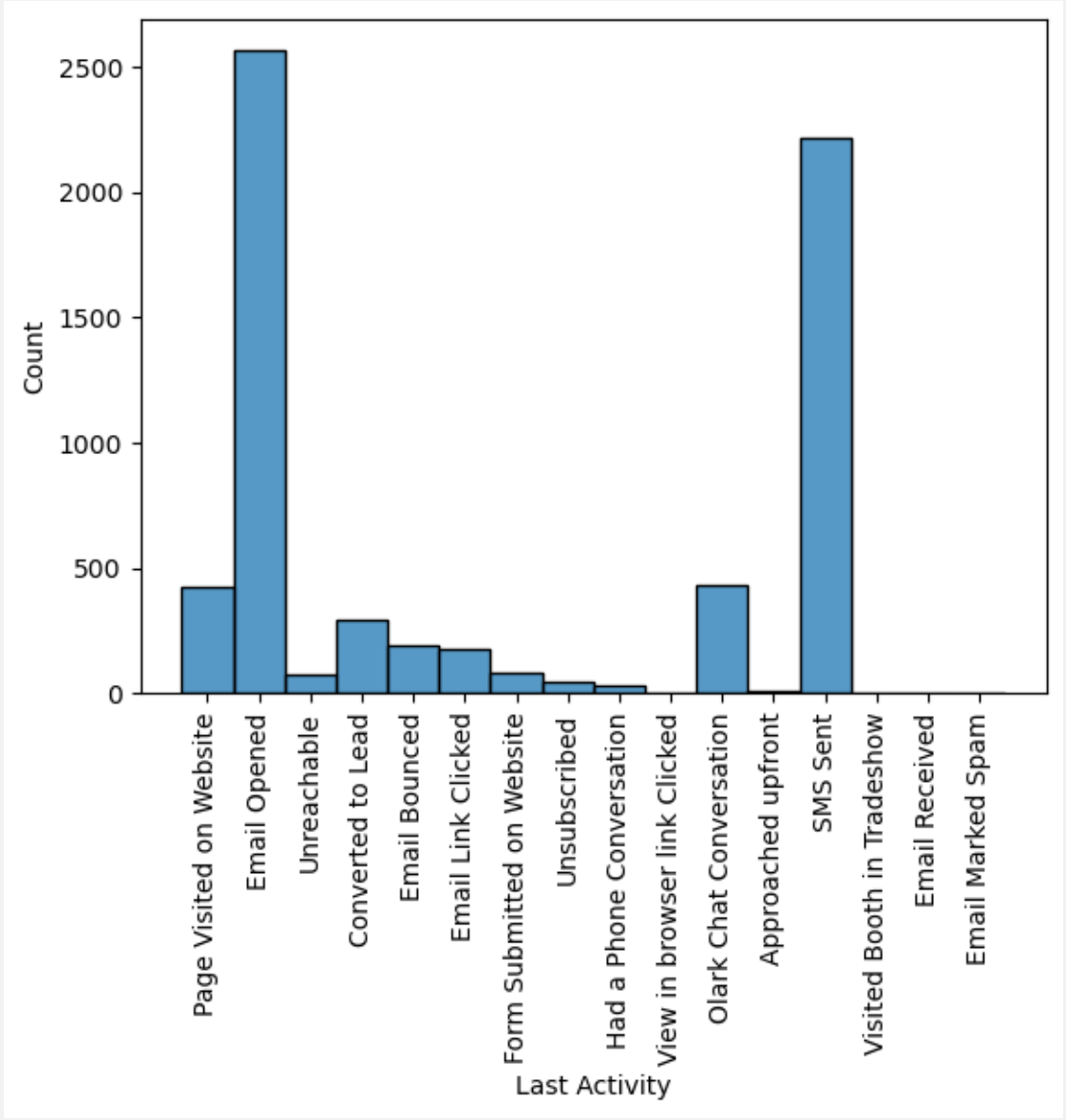
Dataset Details:

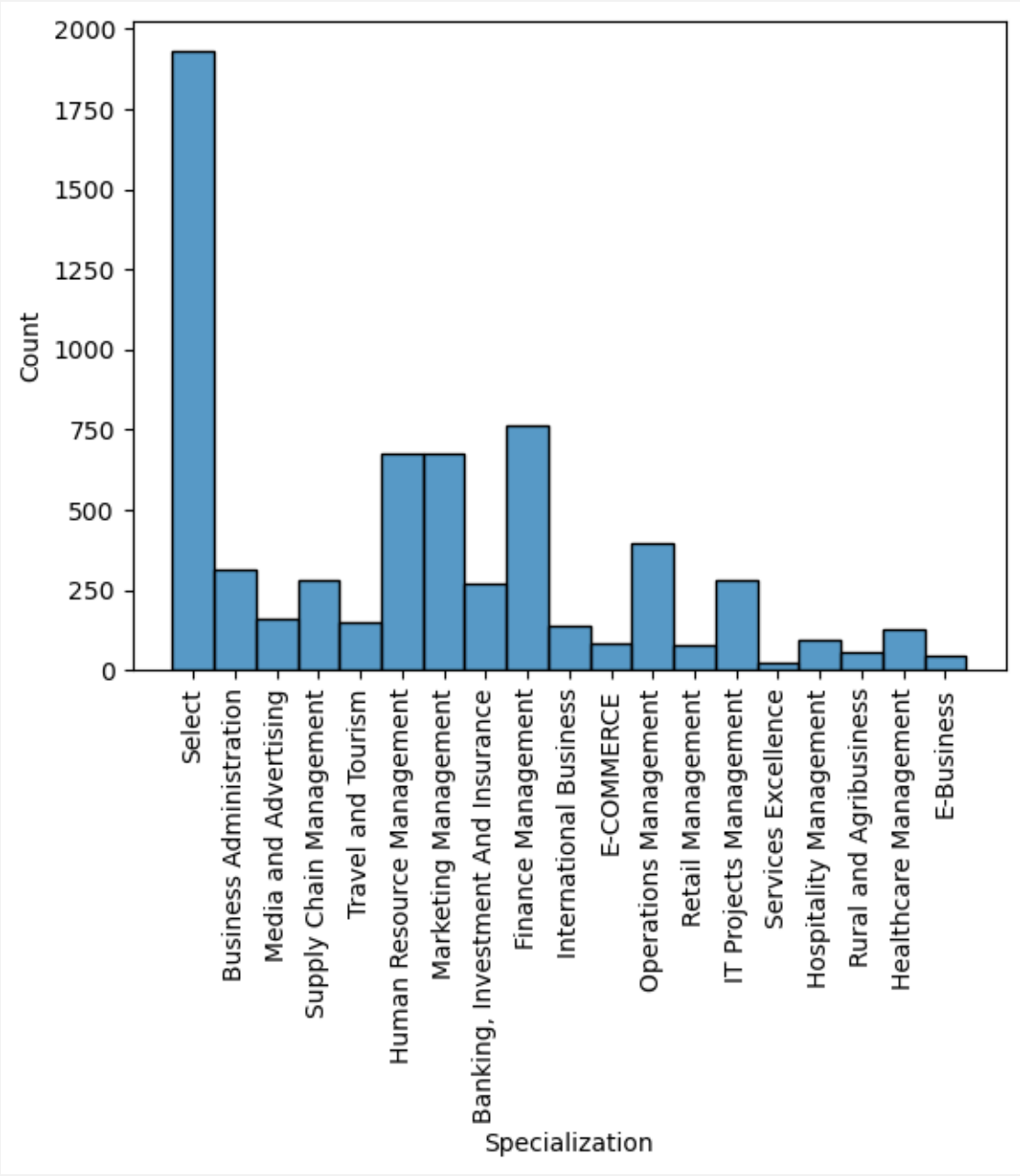
- **Total Data points:** 9240
- **Target Variable** : Converted
- **Key Features:**
  - Total Visits
  - Total Time Spent on Website
  - Lead Origin Lead Add Form
  - Last Notable Activity Unreachable
  - Lead Source Welingak Website

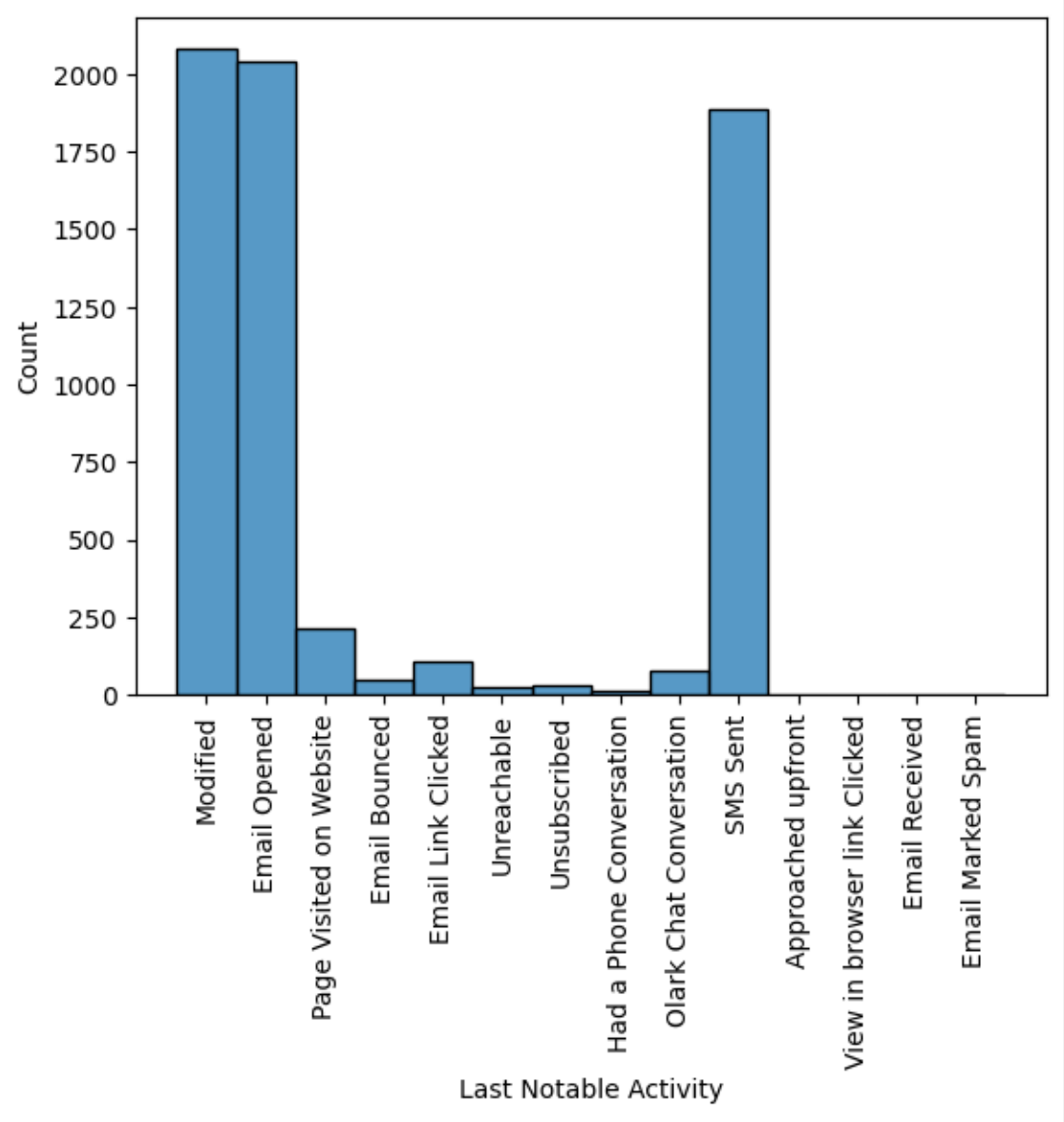
UNIVARIATE ANALYSIS (HISTOGRAM)

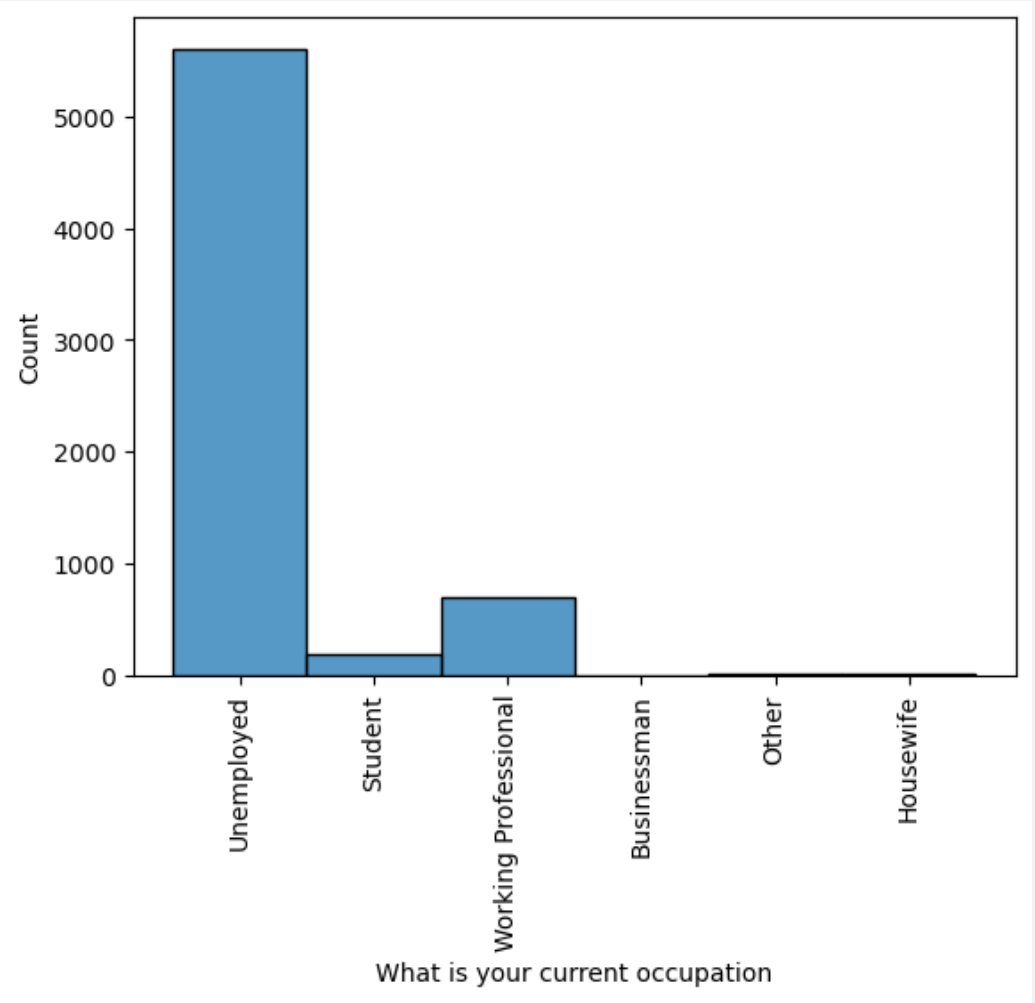




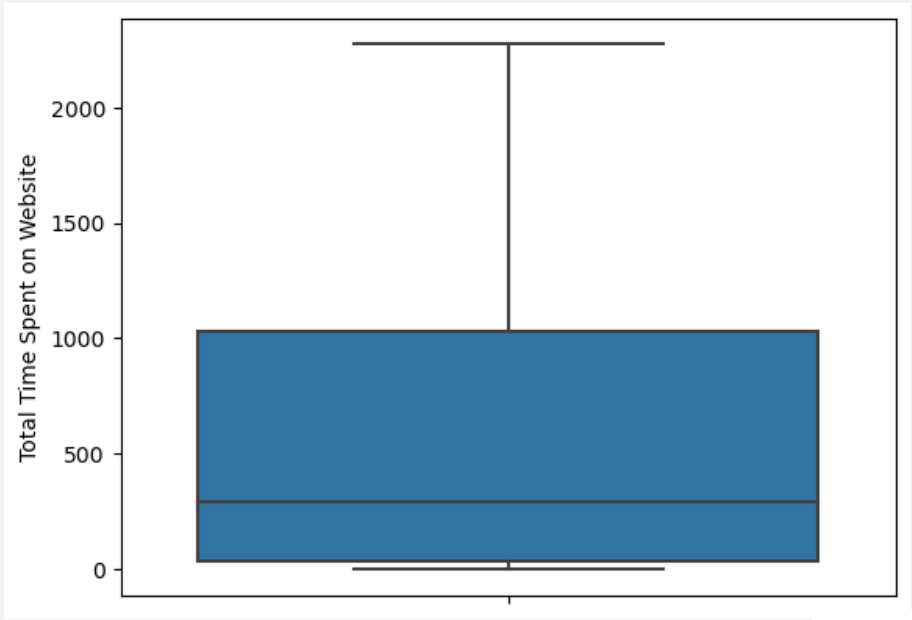

Count
2500
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Olark Chat Conversation
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Email Received
Email Marked Spam
Last Activity

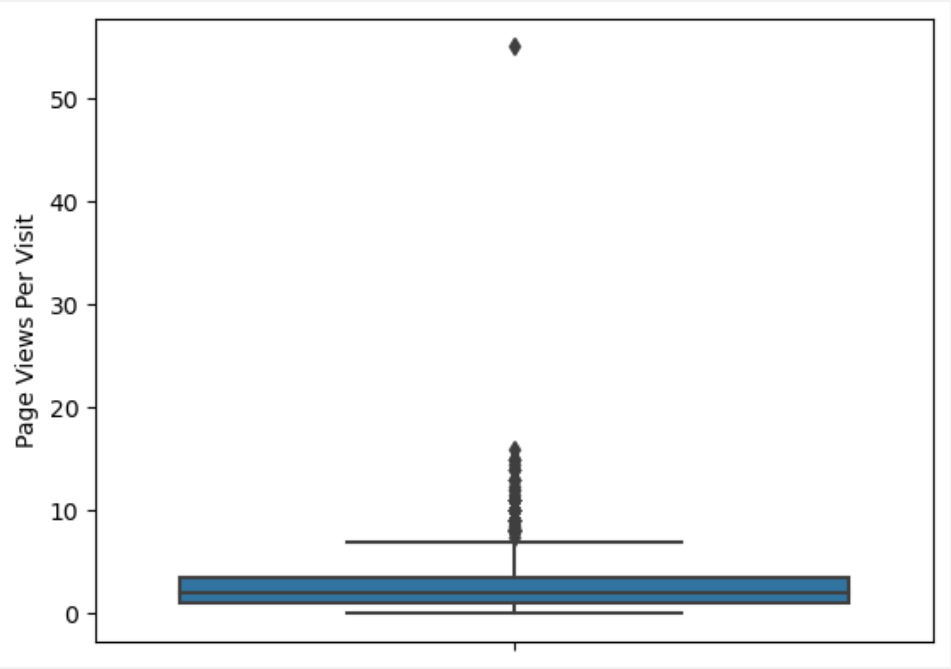

Count
2000
1750
1500
1250
1000
750
500
250
0
Select
Business Administration
Media and Advertising
Supply Chain Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
Finance Management
International Business
E-COMMERCE
Operations Management
Retail Management
IT Projects Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization

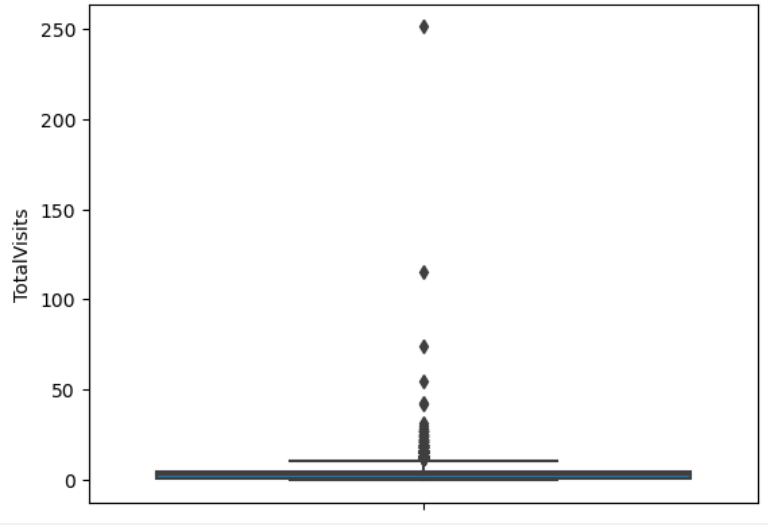

Count
2000
1750
1500
1250
1000
750
500
250
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
View in browser link Clicked
Email Received
Email Marked Spam
Last Notable Activity


Count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

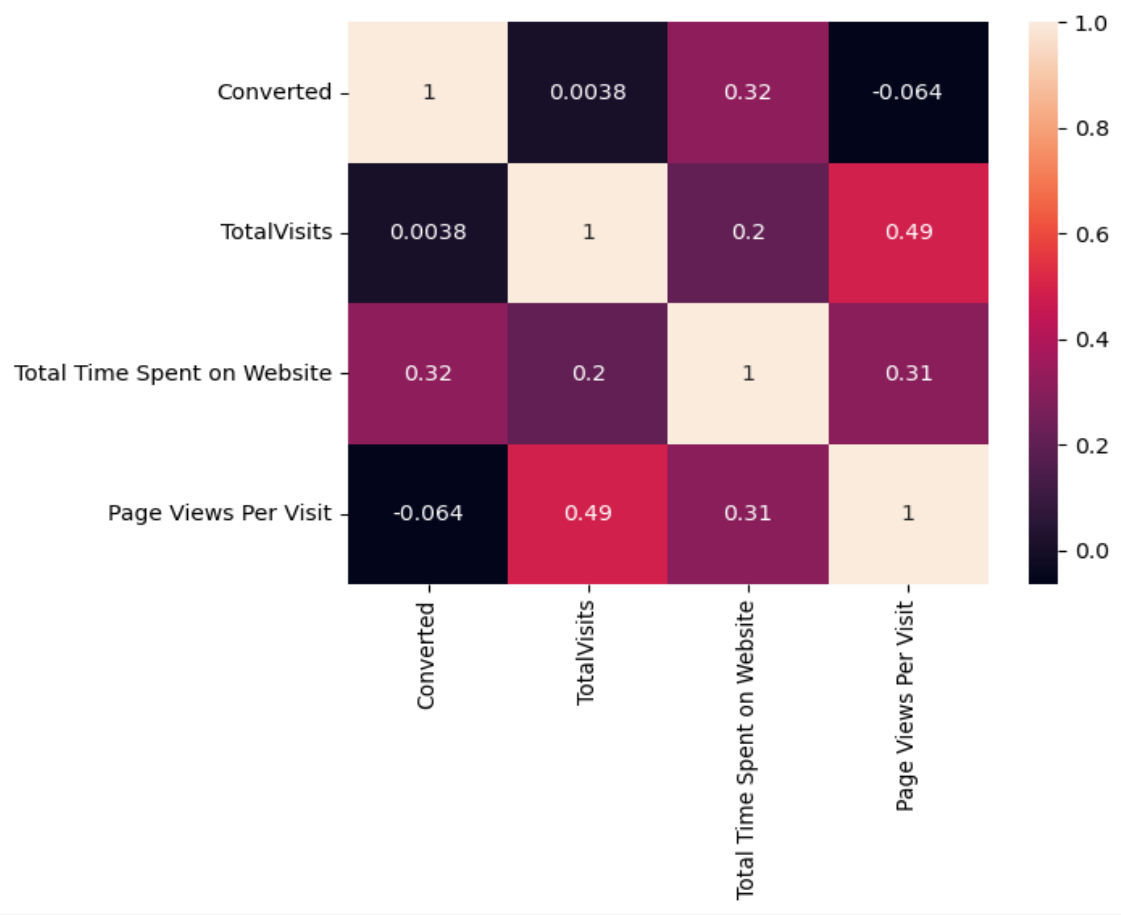
BIVARIATE ANALYSIS - BOXPLOT FOR OUTLIER ANALYSIS







MULTIVARIATE ANALYSIS (HEATMAP)



From basic correlation using Heatmap, it is clear that the one who spent more time on website were highly converted or taken the course because there is high correlation between features 'Total Time Spent on Website' and 'Converted'

LOGISTIC REGRESSION MODEL

- **Model Selection:** Logistic Regression
- Target Variable: Converted

Variables and Coefficients:

- TotalVisits: **10.48**
- Total Time Spent on Website: **4.39**
- Lead Source - Welingak Website: **2.92**
- Last Notable Activity - Unreachable: **2.85**

Multicollinearity Check:

- Variance Inflation Factor (VIF) < 5 for all variables.

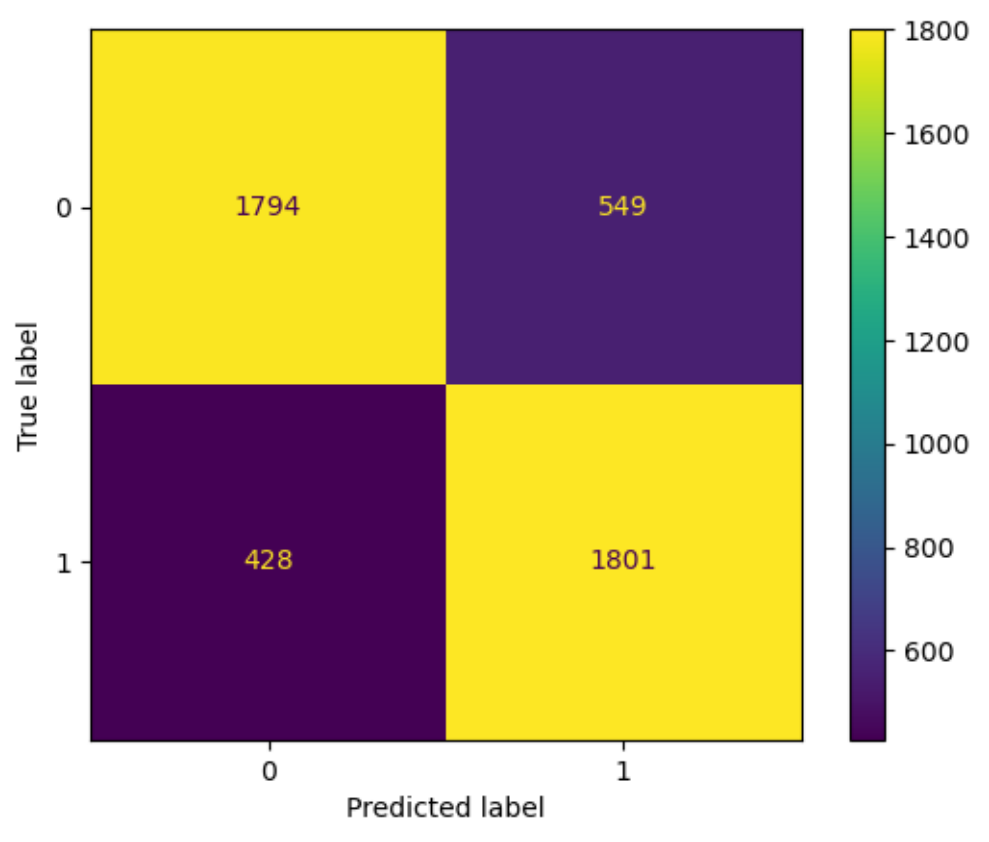
FEATURE SELECTION AND IMPORTANCE

Top 3 Numerical Features (Continuous):

1.TotalVisits
2.Total Time Spent on Website
3.Page Views Per Visit

Top Categorical Features:
1.Lead Source - Welingak Website
2.Lead Origin - Lead Add Form
3.Last Notable Activity - Unreachable

CONFUSION MATRIX ON TRAIN DATA



The Confusion Matrix created has four different quadrants:

•True Negative (Top-Left Quadrant) – (0,0)
➔ Not converted are correctly predicted

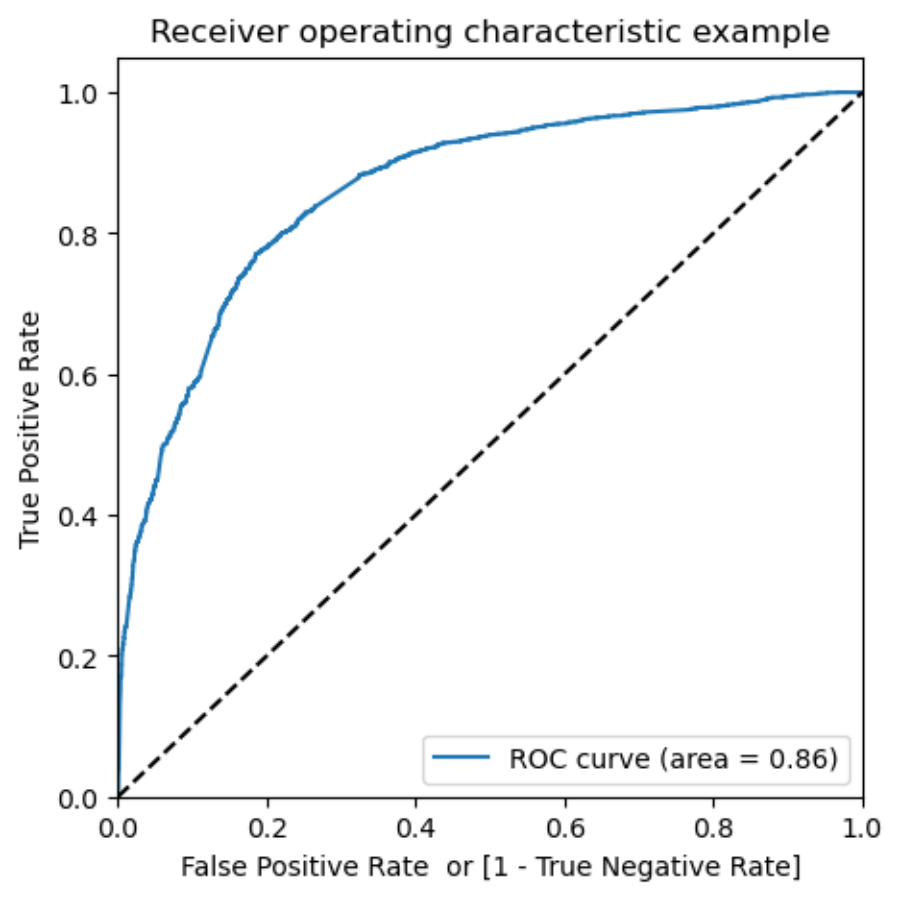
•False Positive (Top-Right Quadrant) – (0,1)
➔ Not converted are falsely predicted as converted

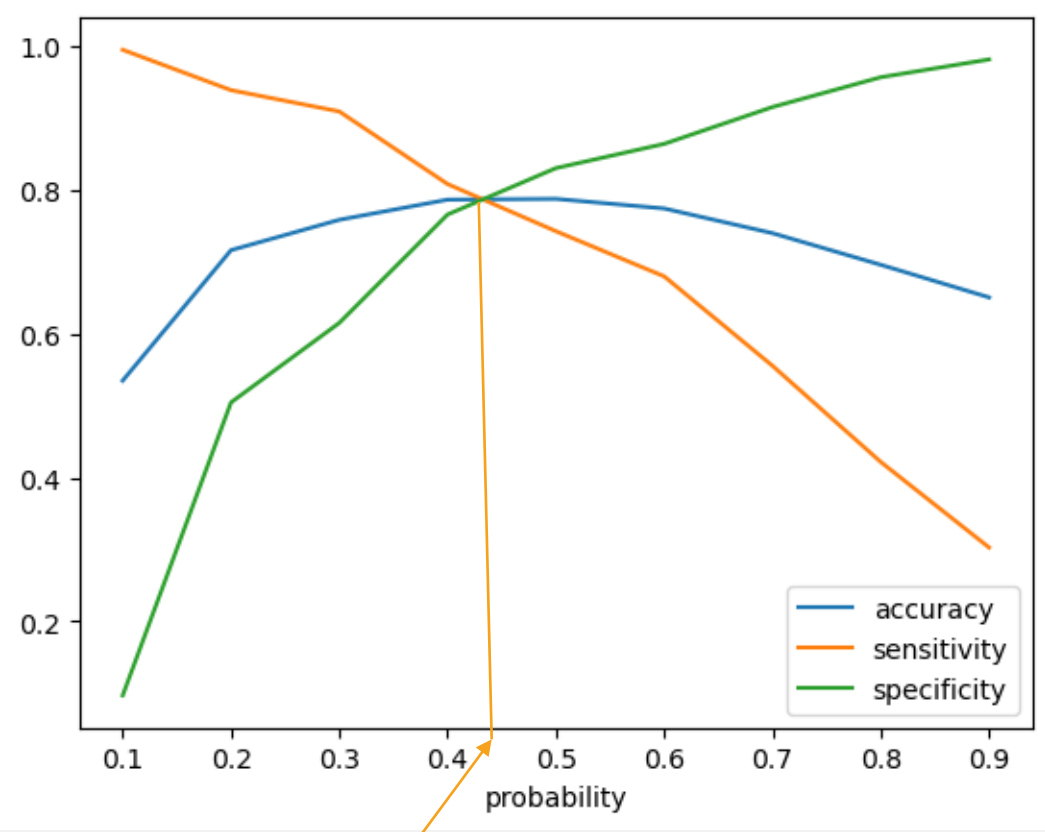
•False Negative (Bottom-Left Quadrant) – (1,0)
➔ converted are falsely predicted as not converted

•True Positive (Bottom-Right Quadrant) – (1,1)
➔ converted are correctly predicted

True means that the values were accurately predicted,
False means that there was an error or wrong prediction.

ROC CURVE & PROBABILITY METRIC PLOT





0.44

METRICS AND CUTOFF

The metrics accuracy, sensitivity, and specificity curves are meet at prabability cutoff 0.44.

- **Optimal Probability Cutoff:** 0.44

Evaluation Metrics:

- **Accuracy:**
  - Training Data - 0.790 (79%)
  - Test Data - 0.797 (79.7%)
- **Sensitivity :** Ensures more positives are captured.
  - Training Data - 0.787 (78.7%)
  - Test Data - 0.783 (78.3%)
- **Specificity :** Balances false positives and false negatives.
  - Training Data - 0.794 (79.4%)
  - Test Data - 0.812 (81.2%)

KEY INSIGHTS

1. **Visits Matter:**
   More website visits increase the likelihood of conversion.
2. **Time on Website:**
   The time spent on website correlates significantly with lead conversions.
3. **Top Sources:**
   Focus on **Welingak Website** and **Lead Add Form** for better engagement.

RECOMMENDATIONS

- Increase user engagement by optimizing website experience.
- Target leads that spend more time and revisit the website.
- Focus marketing efforts on **Lead Add Form** and **Welingak Website** sources.
- Monitor "Unreachable" leads for additional follow-ups.

CONCLUSION

- **Final Model Status:**
  - There are 12 variables are significant for building the model(p-value < 0.05).
  - If the company wants more sensitivity reduce the cutoff probability and wants more specificity increase the cutoff.
  - No multicollinearity found between predictor variables.
- **Action Plans:**
  - Deploy the model for real-time prediction.
  - Optimize probability cutoffs for specific business goals.