Lead Scoring Case Study using Logistic Regression

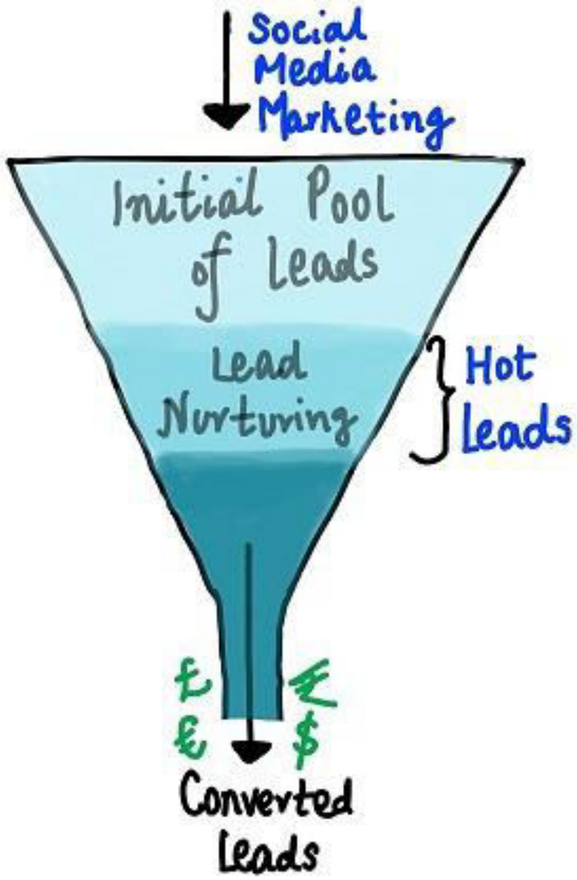
Submitted by:
Arshad Shaikh

Contents

- Problem Statement
- Business Objective
- Methodology
- EDA
- Model Building
- Model Evaluation
- Conclusion

Problem Statement

- 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 markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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%.

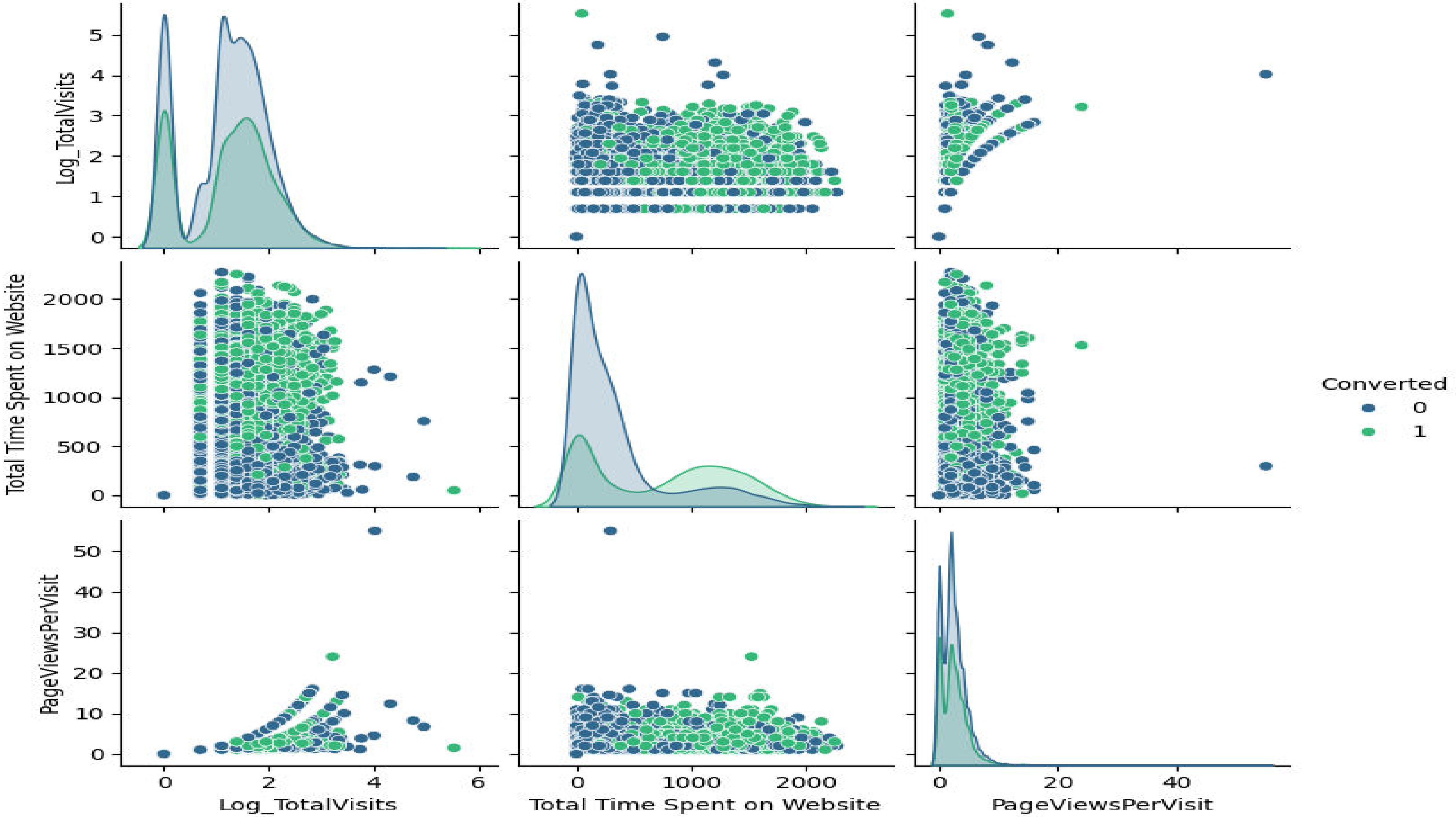


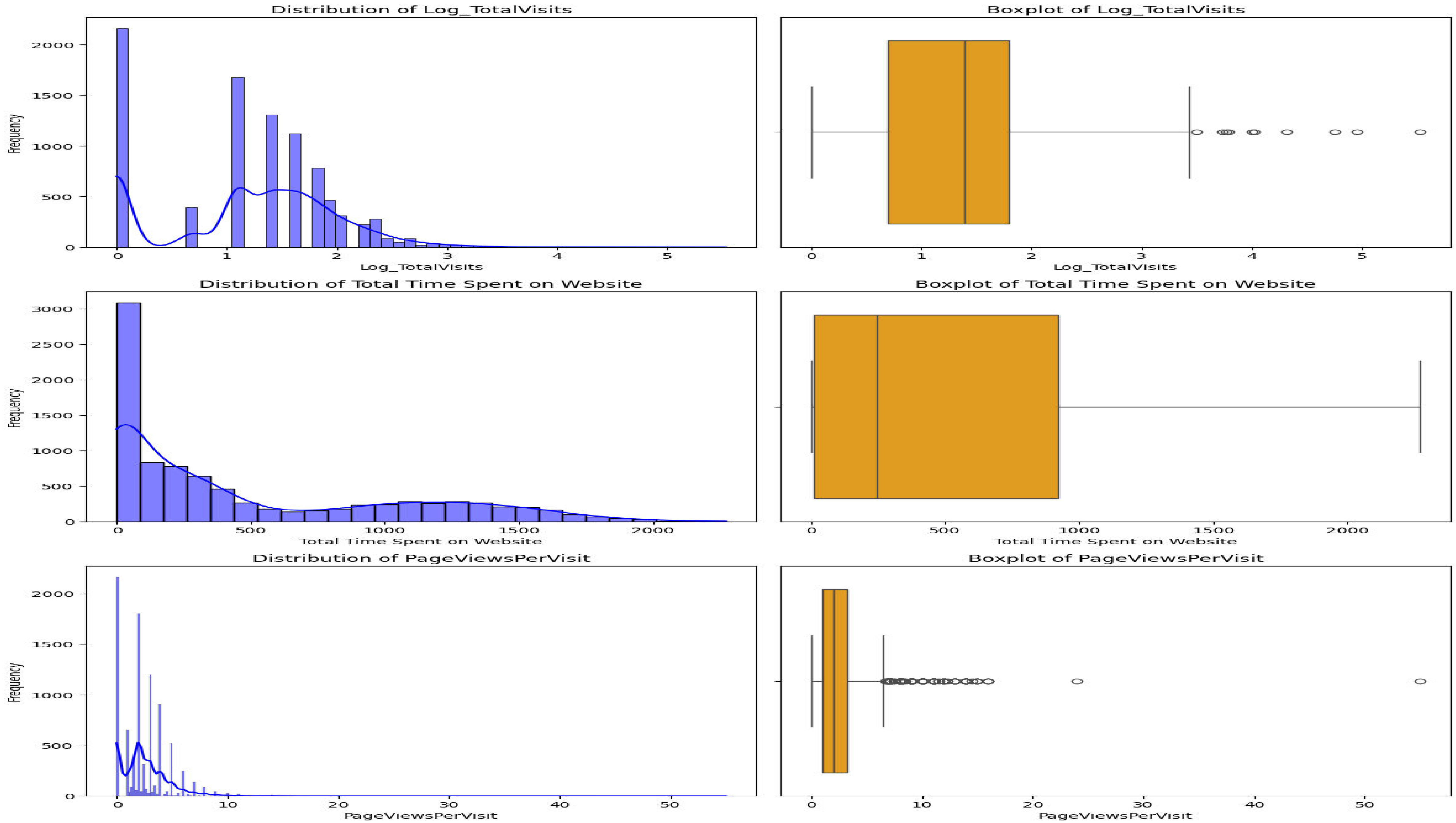
Business Objectives

- Develop a Predictive Model: Create a model to accurately predict lead conversion probability based on relevant factors.
- Assign Lead Scores: Assign a score to each lead, indicating their likelihood of conversion.
- Prioritize High-Potential Leads: Focus on leads with high scores for efficient resource allocation.
- Optimize Sales and Marketing Efforts: Align sales and marketing strategies to maximize conversion from high-potential leads.

Methodology

- Data Collection: Gather lead information such as demographics, website engagements, and conversion rates.
- Clean the data, handle missing values, and convert category variables to numerical representations.
- EDA: Visualize patterns, correlations, and predictions.
- Feature engineering entails developing new features or interaction terms to improve model performance.
- Model Selection: Select appropriate methods such as logistic regression, decision trees, or gradient-boosting.
- Model Training: Develop the model, assess performance using metrics (accuracy, precision, recall), and assure generalizability using cross-validation.


Log_TotalVisits
Total Time Spent on Website
PageViewsPerVisit
0
1
2
3
4
5
0
500
1000
1500
2000
0
10
20
30
40
50
0
2
4
6
0
1000
2000
0
20
40
Converted
0
1



Distribution of Log_TotalVisits

Boxplot of Log_TotalVisits

Distribution of Total Time Spent on Website

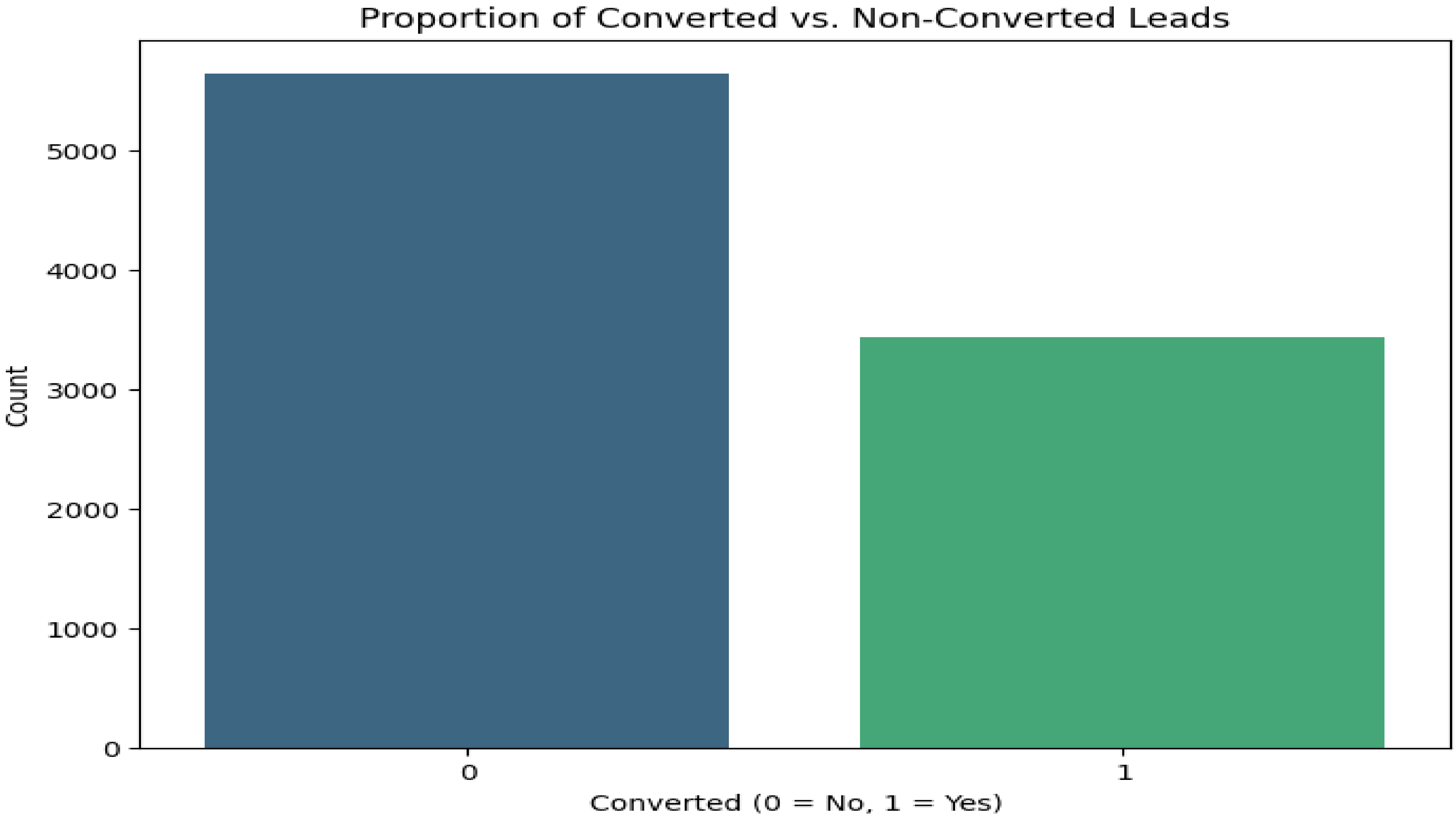
Boxplot of Total Time Spent on Website

Distribution of PageViewsPerVisit

Boxplot of PageViewsPerVisit

Summary

- Log_TotalVisits has a moderate positive skew (0.61), indicating that most visitors have similar visit counts, with only a few having more.
- The total time spent on the website (0.95 skew) indicates that most visitors spend a medium amount of time there, but some are quite involved.
- The Page Views Per Visit (1.25 skew) metric illustrates that while most visitors just view a few pages, others interact deeply.
- Website Engagement: A segment of highly engaged users could be useful for conversions.
- Marketing and Content: Insights can help you target engaged users and optimize content based on popular pages.



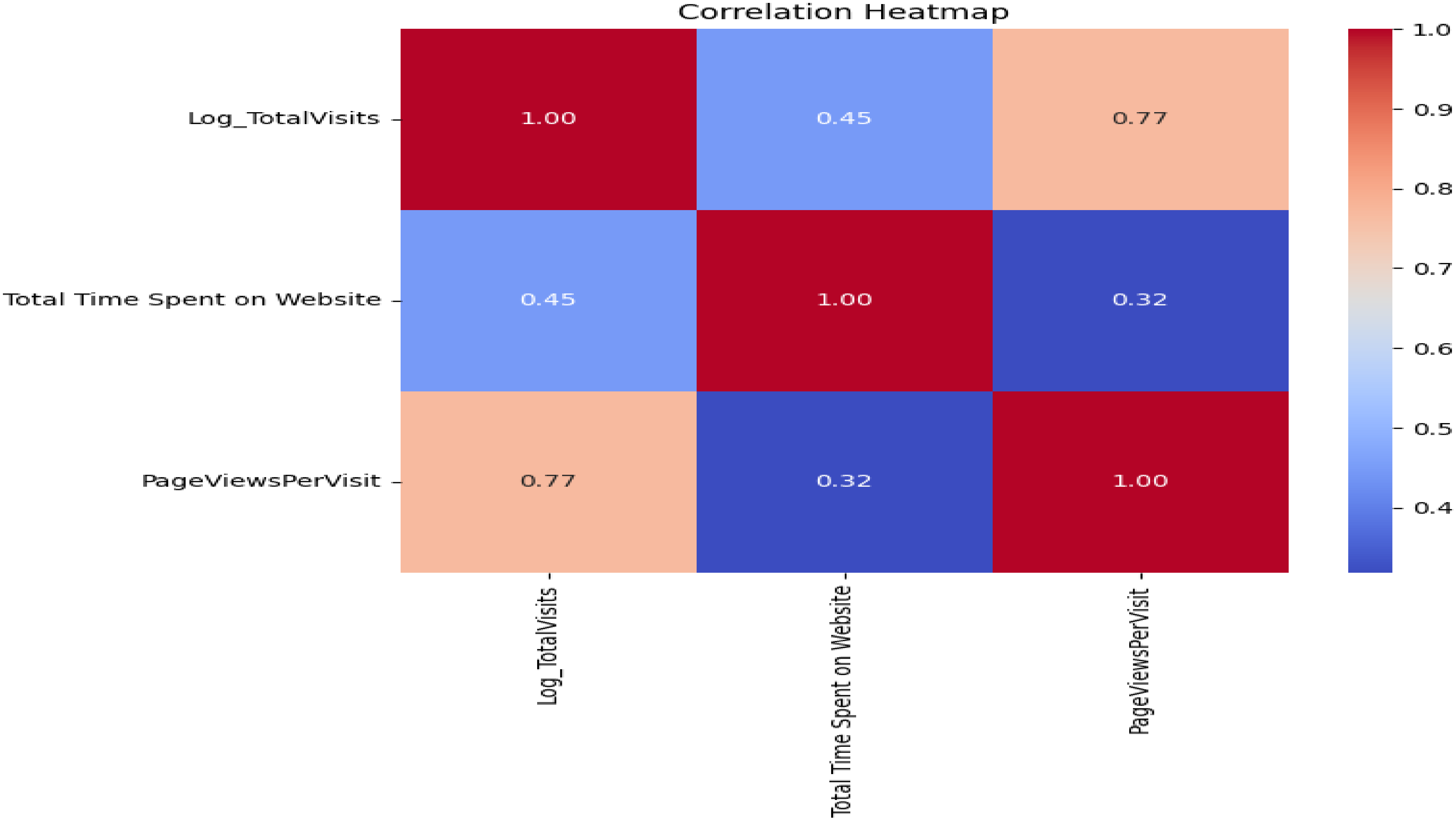
Proportion of Converted vs. Non-Converted Leads

Summary

Imbalance: The results show that the dataset has an imbalance in the target variable (Converted). There are more leads who did not convert (0) compared to those who converted (1). This imbalance is important to consider when building a machine learning model as it might affect the model's performance

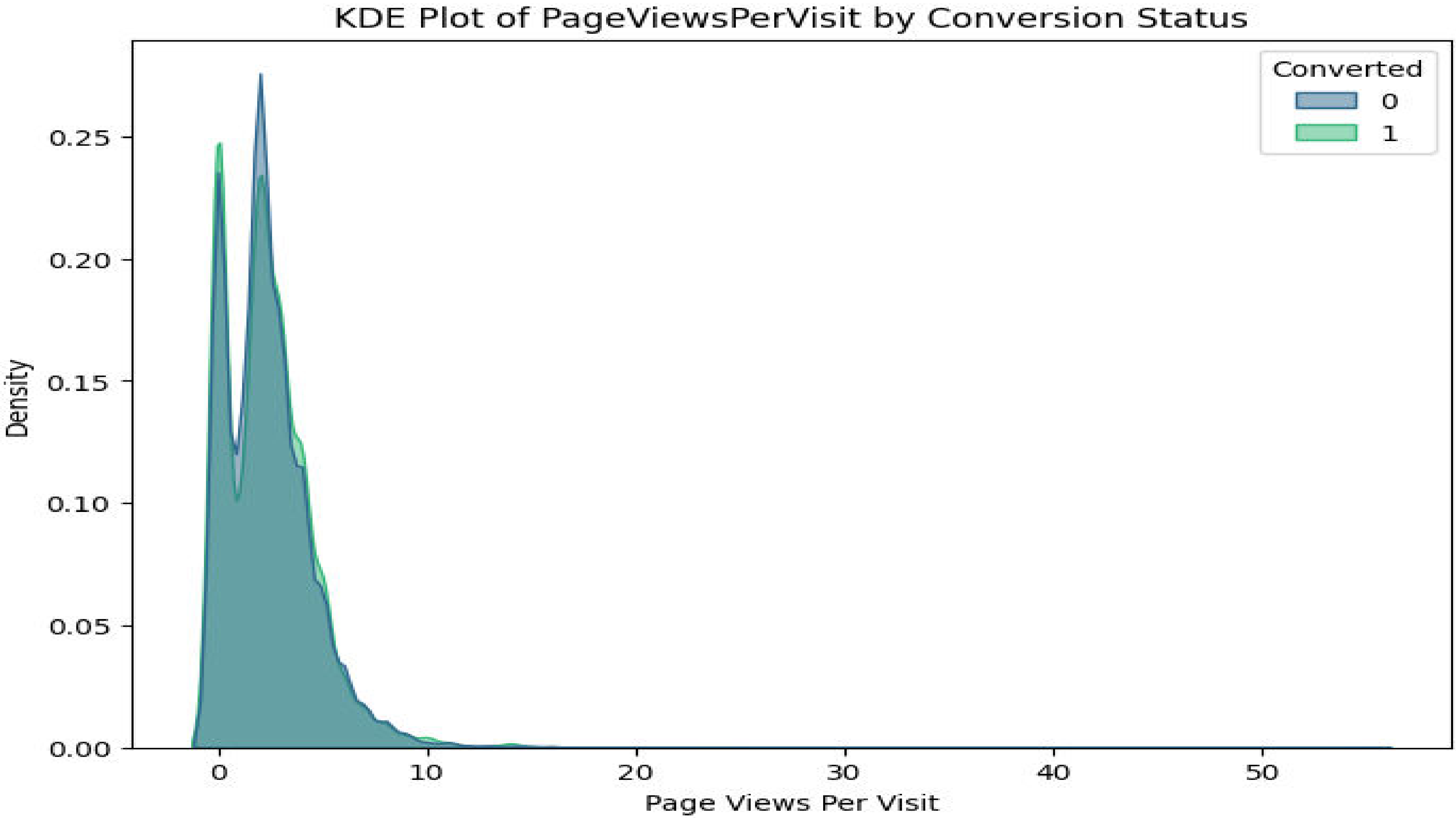
Conversion Rate: The conversion rate is approximately 37.5%. This means that about 37.5% of the leads generated actually converted into paying customers. This is lower than the CEO's target of 80%.

Focus on Improvement:** The company's current lead conversion process needs improvement to reach the desired conversion rate. The task is to identify and target potential "hot leads" more effectively.


Correlation Heatmap
Log_TotalVisits
Total Time Spent on Website
PageViewsPerVisit
1.00
0.45
0.77
0.45
1.00
0.32
0.77
0.32
1.00
Log_TotalVisits
Total Time Spent on Website
PageViewsPerVisit
1.0
0.9
0.8
0.7
0.6
0.5
0.4

Summary

- Log_TotalVisits and PageViewsPerVisit have a strong positive association (0.77), showing that visitors who log more total visits also see more pages per visit.
- Log_TotalVisits and Total Time Spent on Website show a moderate positive association (0.45), indicating that as the number of visits to the website increases, so does the amount of time spent on it.
- Total Time Spent on Website and PageViewsPerVisit have a weaker positive correlation (0.32), indicating that, while there is a relationship, time spent on the site does not greatly influence the number of pages viewed per visit.


KDE Plot of PageViewsPerVisit by Conversion Status
Converted
0
1
Density
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
Page Views Per Visit

Summary

- Right-Skewed: Both groups have a right-skewed distribution, with most customers visiting only a few pages throughout their visit.
- Converted clients typically have more page views per visit.
- Both categories peak at zero, indicating that many consumers have few page views.
- Slow Tail Off: A few clients browse a large number of pages.
- Conclusion: More page views per visit lead to a greater conversion rate.

Model Building

We utilized the RFE method to generate the logistic regression model.

RFE approach:

RFE is a feature selection approach that recursively removes the least important features based on the model's performance, while retaining the most important ones.

Process:

- The model (e.g., logistic regression) is first trained on all features.
- The least important elements are eliminated one at a time, according to their importance or coefficient values.
- This technique is performed recursively until the desired number of features has been chosen.

Goal:

to improve model performance and prevent overfitting by deleting unnecessary or redundant features, and then select the most influential predictors for the target variable.

Model Evaluation

Train Data

Accuracy: 80.38%

Sensitivity: 68.27 %

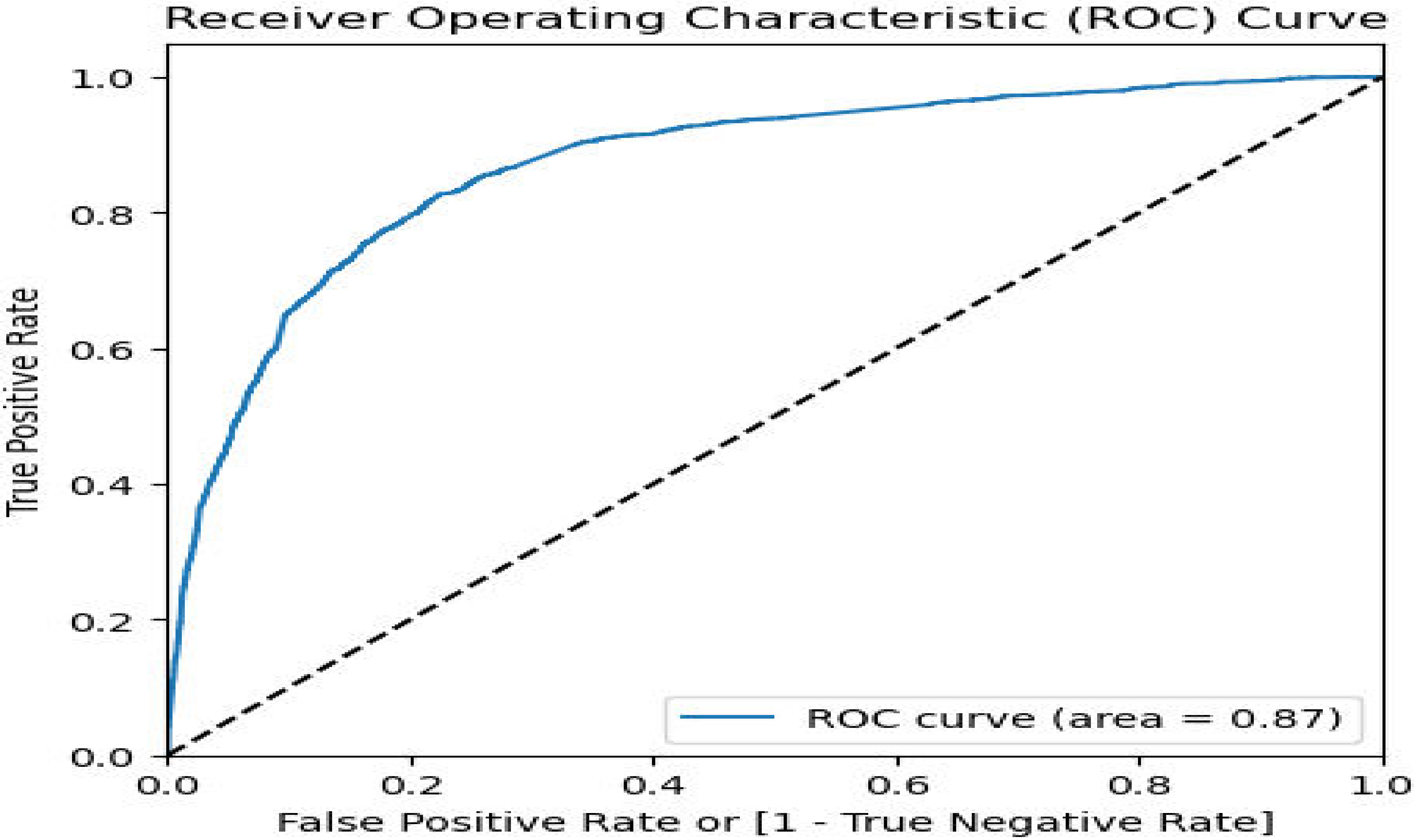
Specificity: 87.96%

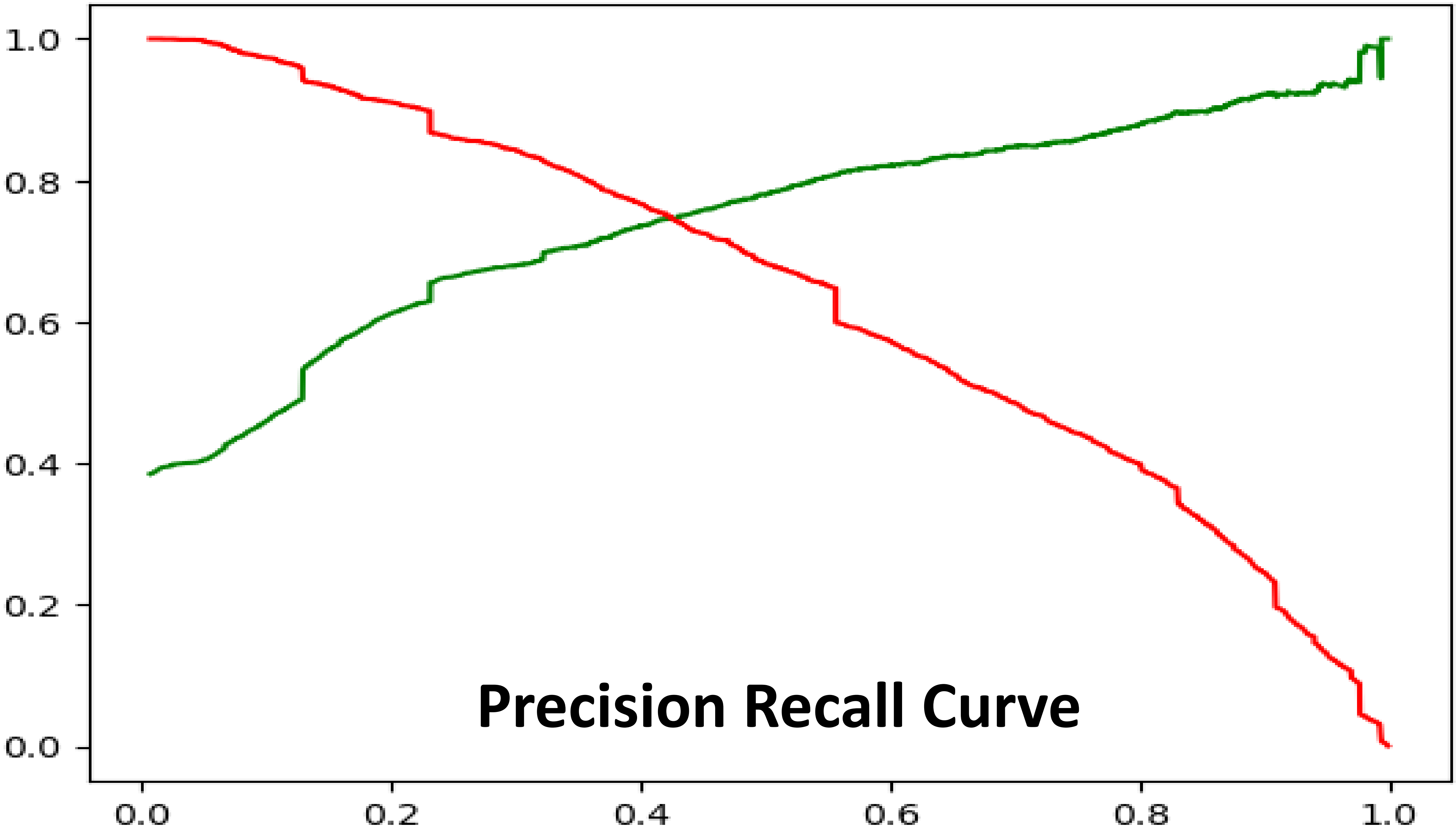
Test Data

Accuracy: 80.34%

Sensitivity: 68.52 %

Specificity: 88.04%


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.87)



Precision Recall Curve

Final Feature List

- Log_TotalVisits
- Total Time Spent on Website
- Last Notable Activity_Email Opened
- Last Notable Activity_Modified
- LeadSource_Olark Chat
- Last Notable Activity_Page Visited on Website
- occupation_Working Professional
- Lead Origin_Lead Add Form
- Do Not Email_Yes
- Last Notable Activity_Olark Chat Conversation
- Last Notable Activity_Email Link Clicked
- Lead Origin_Lead Import

Conclusion

- Focus on High Engagement Leads: Leads who have more overall visits and time spent on the website are more likely to convert, so they should be prioritized in marketing initiatives.
- Improve Communication Channels: LeadSource_Olark Chat and Last Notable Activity_Email Opened are highly predictive of conversion. Strengthening these channels could result in improved conversion rates.
- Segment "Do Not Email" Leads: Leads with Do Not Email_Yes have lower conversion rates. Consider alternate re-engagement strategies, such as content offers or targeted promotions.
- Optimize Website Content: Pages that keep users engaged for longer periods of time and generate more page views per visit should be analyze
- Strengthen Lead Generation: Sources such as Lead Origin_Lead Add Form are significant conversion indicators. Improve these lead channels to generate more high-quality prospects.
- Use Activity Insights: Last Notable Activity factors can help discover crucial touchpoints. Customize follow-ups based on these findings to increase conversions.
- Refine email campaigns: Email engagements, particularly link clicks and opens, have a major impact on conversions. Improve email methods to increase engagement and results and optimized for conversion.

Thank you