

Lead Scoring Case Study

PRESENTED BY:

Deepanjana Roy





Strategy

- **About X Education Company**
- **Problem Statement**
- **Steps take to solve Problems**
- **Load and explore the dataset**
- **Clean the Dataset**
- **Performing Eda(Exploratory Data Analysis)**
- **Prepare the data for model building**
- **Build the logistic regression model**
- **Evaluate the model's performance**
- **Checking Precision and Recall**
- **Making predictions on the test dataset**
- **Determining the important features**
- **Final Observations**
- **Recommendation**

About X Education Company

X Education, a prominent provider of online courses for industry professionals, has a growing online presence, attracting numerous leads daily through platforms like Google. These leads are individuals who express interest by interacting with the website, often by submitting contact forms. Despite the high volume of leads, the company faces a significant challenge: a relatively low conversion rate of around 30%. With limited time and resources, the sales team struggles to engage all leads effectively.

To enhance efficiency and improve conversion rates, X Education aims to prioritize high-potential leads, referred to as "Hot Leads." The objective is to identify these leads—those with the greatest likelihood of becoming paying customers—so that the sales team can focus their efforts where they matter most. The CEO has set an ambitious goal of achieving an 80% conversion rate, and the company plans to use a data-driven approach to meet this target.

Problem Statement

X Education has provided a dataset of past leads with various attributes that may influence conversion rates. The objective is to build a logistic regression model that assigns a lead score between 0 and 100, indicating the likelihood of conversion for each lead. A higher score suggests a greater probability of conversion, allowing the sales team to prioritize these leads. Additionally, the model should be adaptable to future requirements and evolving business needs.

Through this analysis, X Education hopes to boost its lead conversion rate by identifying "Hot Leads" and enabling the sales team to focus on the most promising opportunities, ultimately improving the company's overall sales performance.

Steps take to solve Problems

Import necessary libraries

Import all required libraries such as pandas, numpy, seaborn, matplotlib, and sklearn for data manipulation, visualization, and modeling.

Load and explore the dataset

Load the dataset and perform an initial exploration to understand its structure, including the number of rows, columns, and data types.

Clean the dataset

Handle missing values, remove duplicates, and correct inconsistencies to ensure the dataset is accurate and ready for analysis.

Performing EDA

Analyze key features, visualize distributions, and identify relationships between variables to gain insights and detect patterns.

Prepare the data for model building

Convert categorical variables, standardize numerical features, and split the data into training and test sets.

Build the logistic regression model

:Train a logistic regression model on the prepared data to predict lead conversion probabilities.

Evaluate the model's performance

Assess the model using metrics such as accuracy, precision, recall, and ROC-AUC score to gauge its effectiveness.

Check precision and recall

Evaluate precision and recall to understand the model's ability to correctly identify positive conversions and minimize false positives.

Make predictions on the test set

Generate final predictions for the test set to assess the model's performance on unseen data.

Determine the important features

Identify the most influential features in the model that drive lead conversions by analyzing feature coefficients or importance scores.

Final observations

Summarize key findings and provide actionable recommendations based on the model's results and insights derived from the analysis.


Recommendation

For increasing conversion Rate.

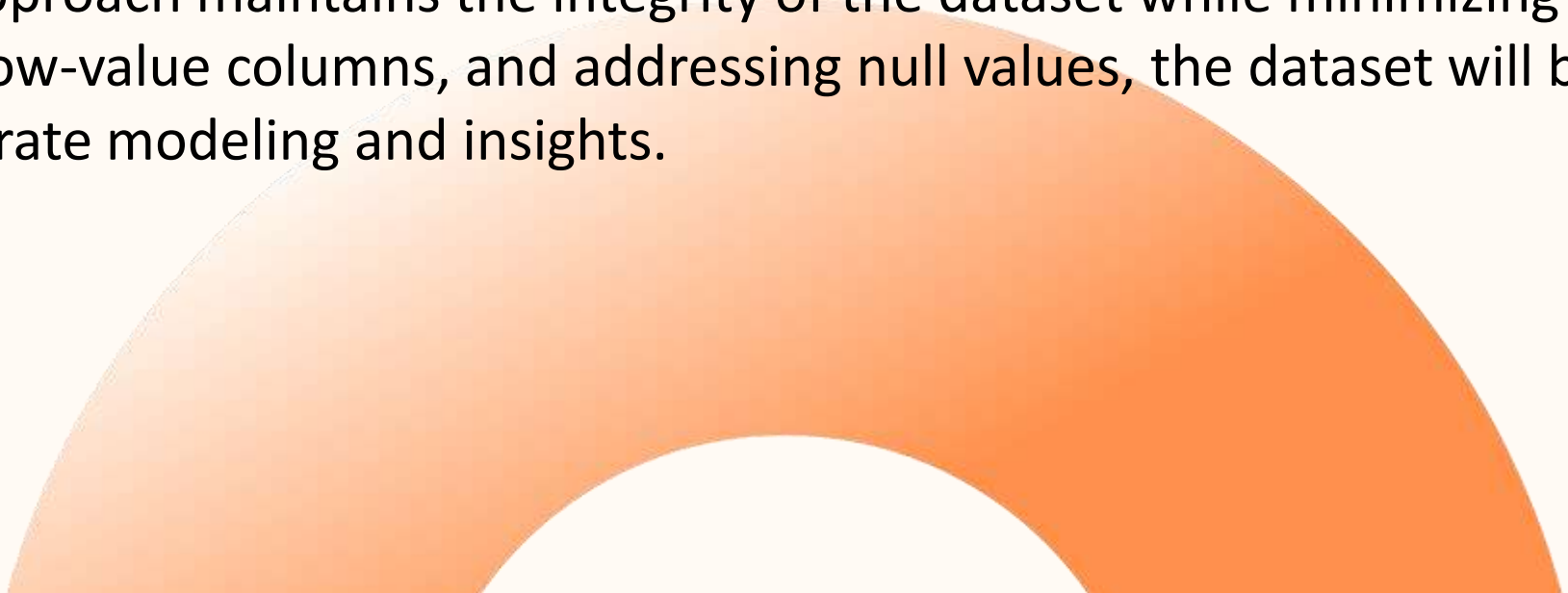
Import some necessary libraries

- Pandas for data manipulation and analysis, providing data structures like DataFrames that make it easy to work with structured data.
- NumPy for numerical computations in Python, offering support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions.
- Seaborn & Matplotlib for interactive visualizations
- Warnings for control warning messages, used here to suppress unnecessary warnings during the execution of the code for cleaner output.

Load and explore the dataset

- Begin by importing the datasets into the environment for analysis.
 - Conduct an initial review of the datasets to gain insights into their structure and content.
 - Examine the distribution, central tendency, and variability of the data. This analysis helps identify potential issues such as missing values or skewness in the data.
 - Assess the data types of each variable to ensure they are appropriate for analysis and modeling.
 - Inspect all column names in the datasets for any leading or trailing spaces and remove them using the strip function, ensuring consistency in naming.
- 

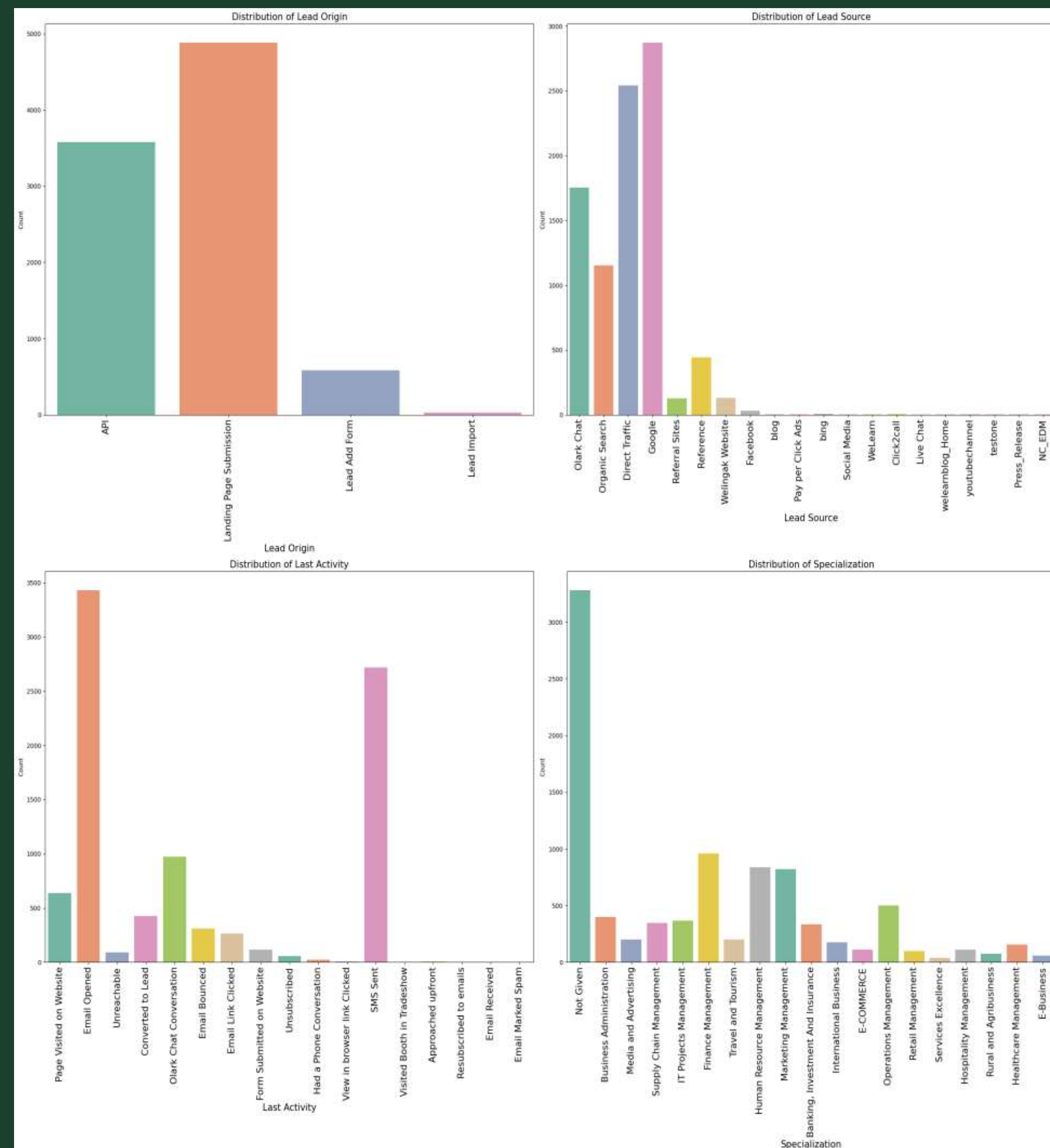
Clean the Dataset

- Begin by calculating the percentage of null values in each column to understand the extent of missing data, which will inform subsequent data cleaning steps.
 - Certain columns, such as "Specialization," "How did you hear about X Education," "Lead Profile," and "City," contain a high frequency of the value "Select." These values will be converted to NaN to improve data integrity.
 - Any columns with 40% or more missing values will be removed from the dataset, as they do not provide sufficient information for analysis.
 - Columns like "Update me on Supply Chain Content," "Get updates on DM Content," and "I agree to pay the amount through cheque" contain predominantly the value "No." Such columns do not contribute valuable insights and should be dropped.
 - The "Country" column shows that 6492 out of 9240 entries are from India, with an additional 2461 null values. This redundancy (approximately 96% of the data) indicates that the column can be removed to streamline the analysis.
 - Upon examining the data, several columns—such as "Do Not Call," "Magazine," and "Newspaper Article"—primarily consist of a single value ("No"). Their lack of variability makes them statistically insignificant and unhelpful for analysis, warranting their removal.
 - For the columns "Specialization" and "What is your current occupation," any remaining null values will be replaced with the placeholder "Not Given." This approach maintains the integrity of the dataset while minimizing missing information.
 - After removing redundant and low-value columns, and addressing null values, the dataset will be cleaner and more suitable for analysis, enabling more accurate modeling and insights.
- 

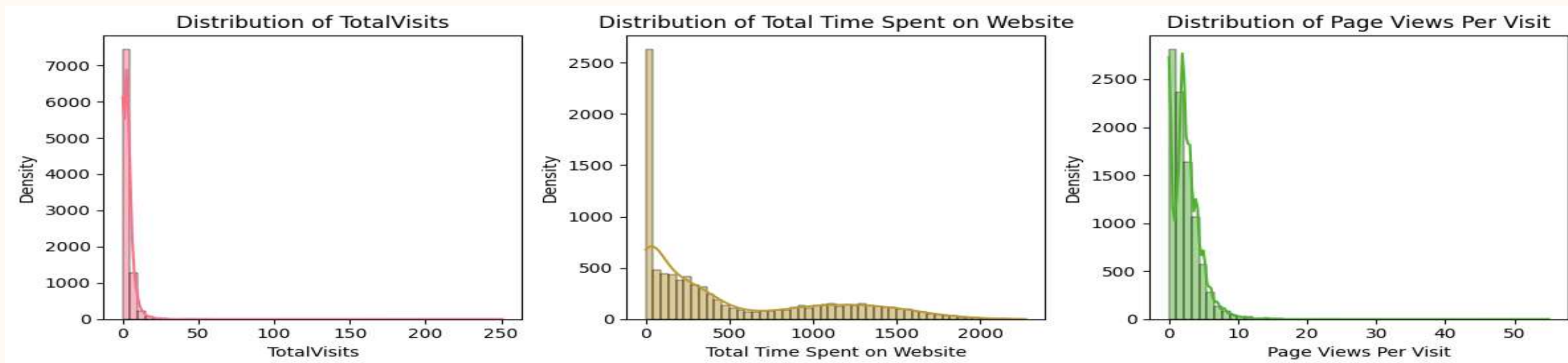
Performing EDA

Univariate Analysis

- The Lead Origin analysis reveals that the majority of leads are generated through the Landing Page Submission channel, highlighting its effectiveness in attracting potential customers.
- Google emerges as the most prominent lead source, suggesting a significant portion of leads come from Google searches or advertisements.
- The most frequent Last Activity recorded is Email Opened, indicating strong engagement with the company's email campaigns.
- A considerable number of leads have 'Not Provided' in the Specialization field, implying many individuals did not select a specific specialization during form submission.
- The Unemployed category has the highest representation in terms of occupations, suggesting that individuals seeking to enhance their employability or upskill are more inclined to explore X Education's offerings.
- The majority of leads have declined the offer to receive a free copy of 'Mastering The Interview', indicating limited interest in this incentive.
- Similar to the last activity, Email Opened is the most common Last Notable Activity, further reflecting email engagement.

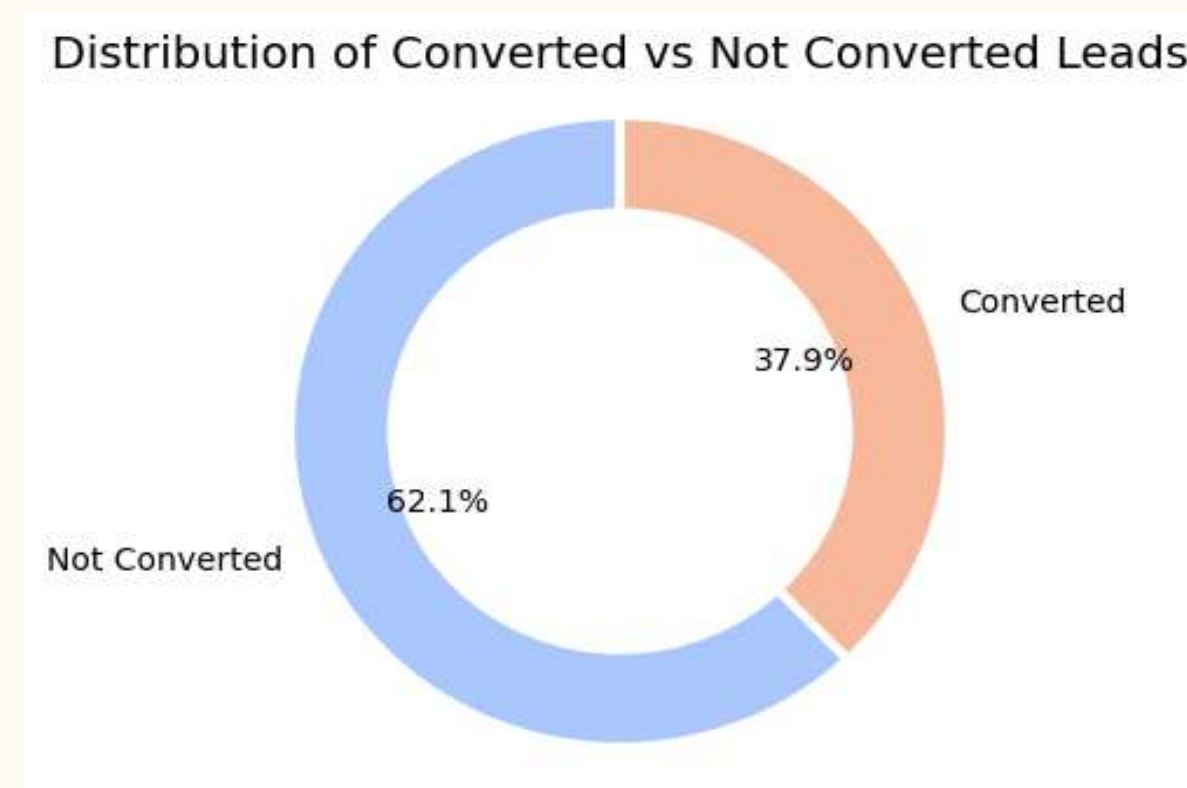


Numerical Columns

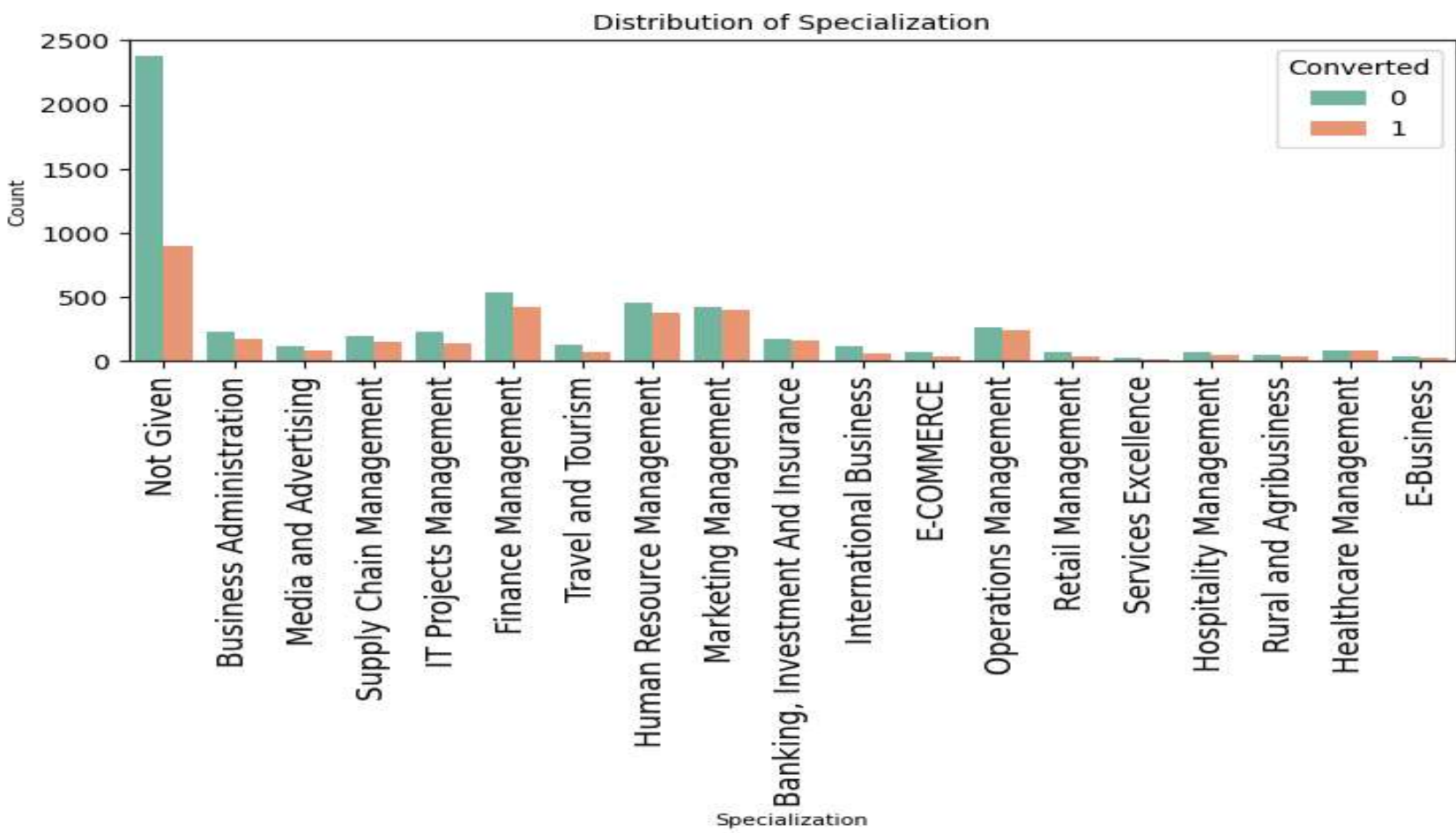
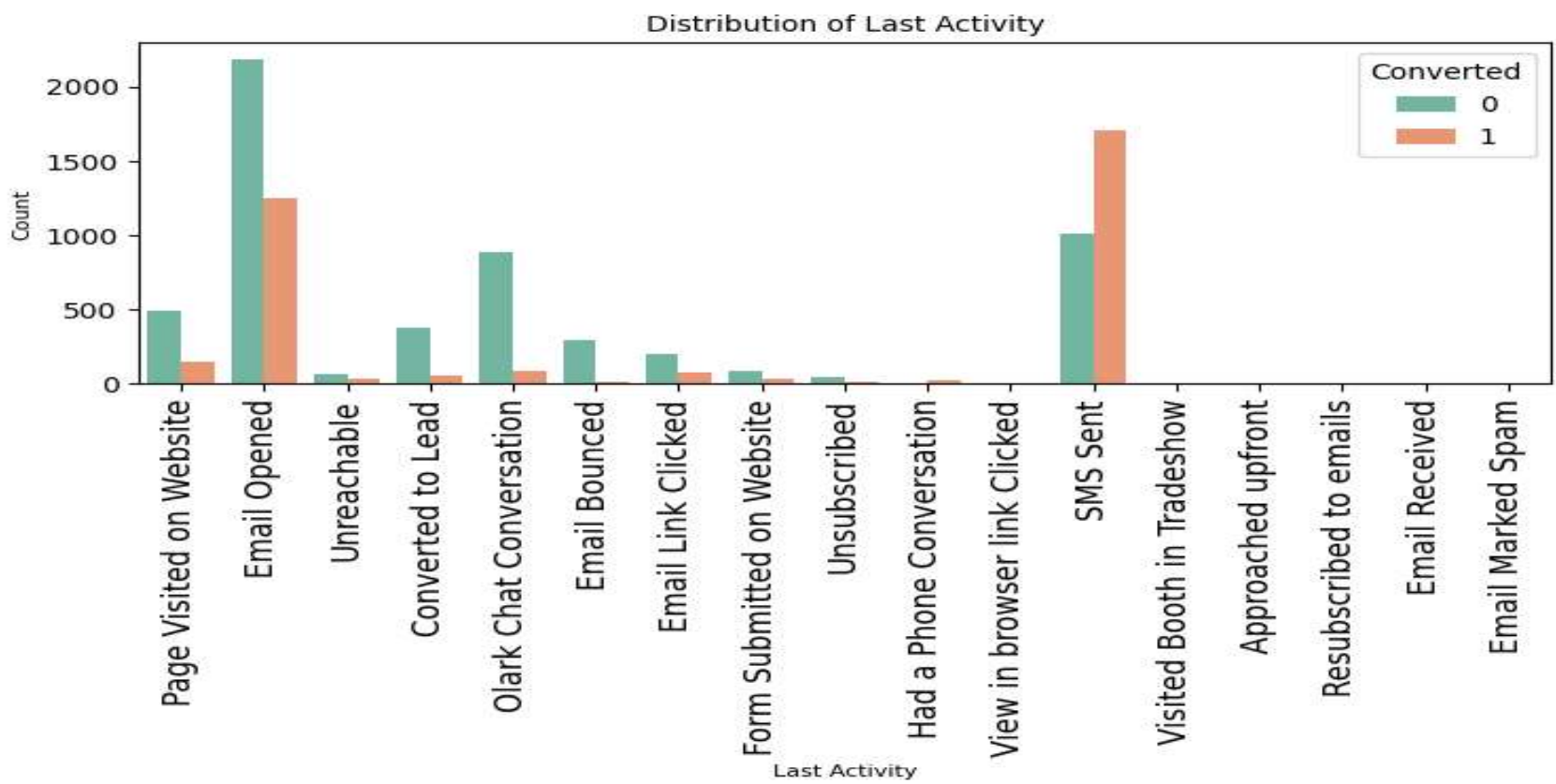
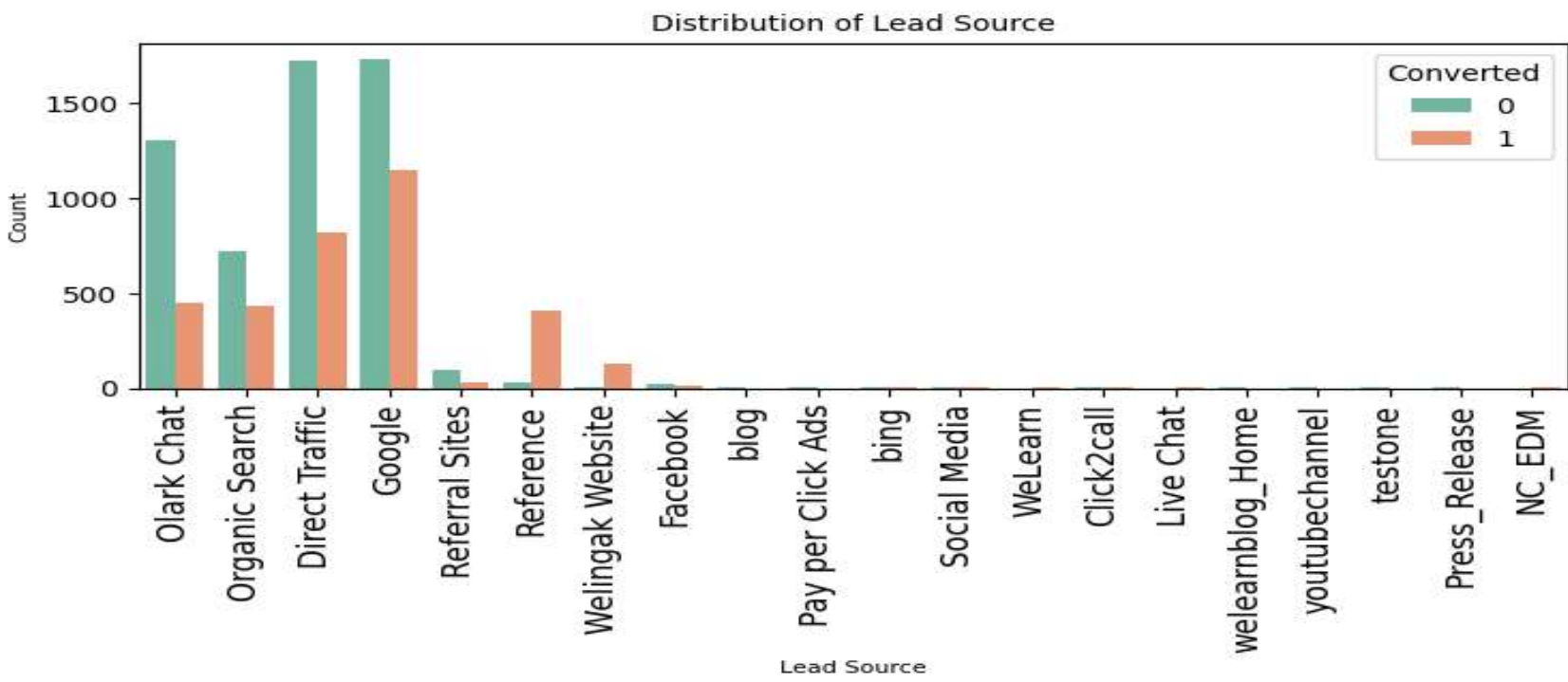
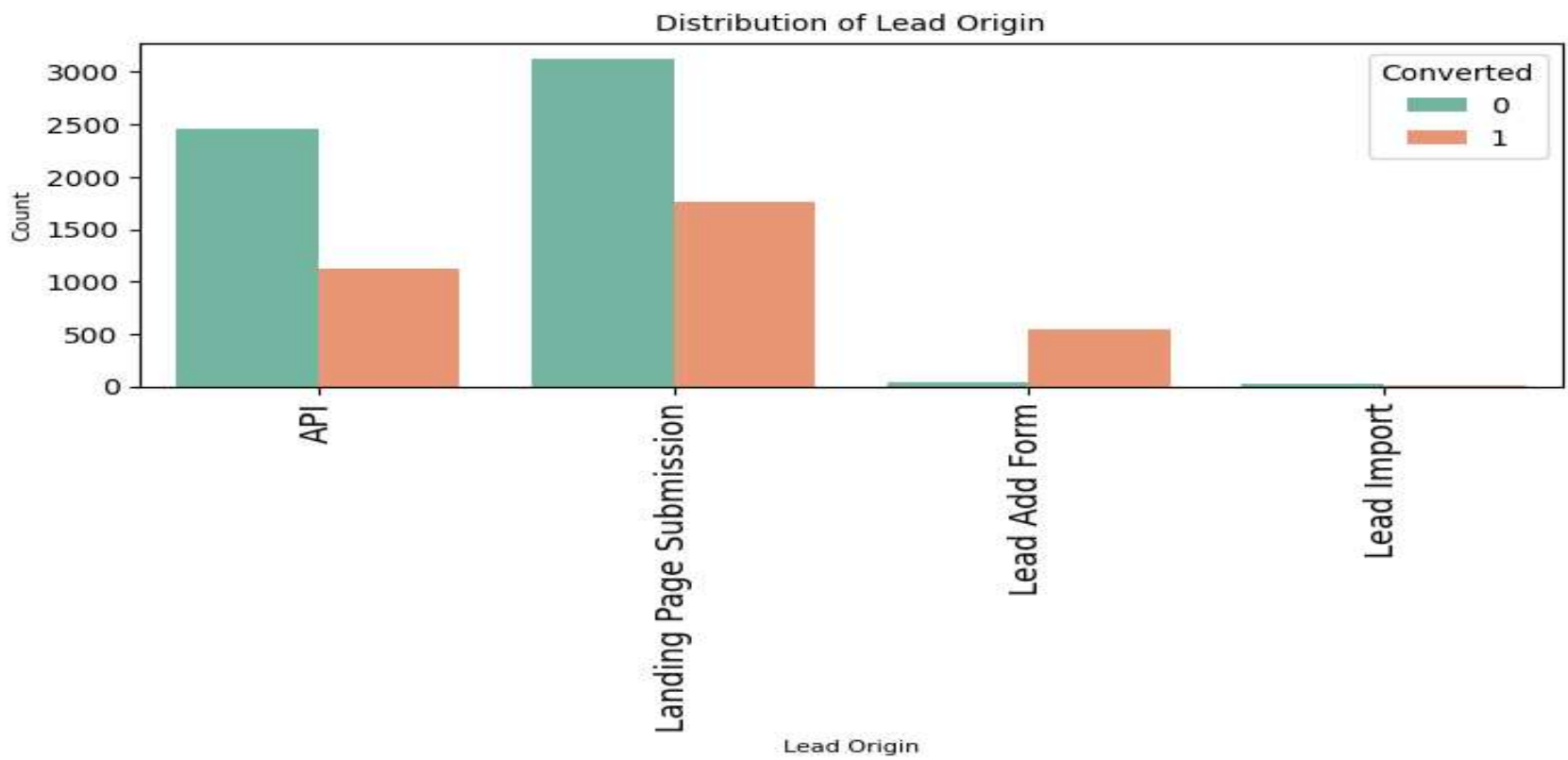


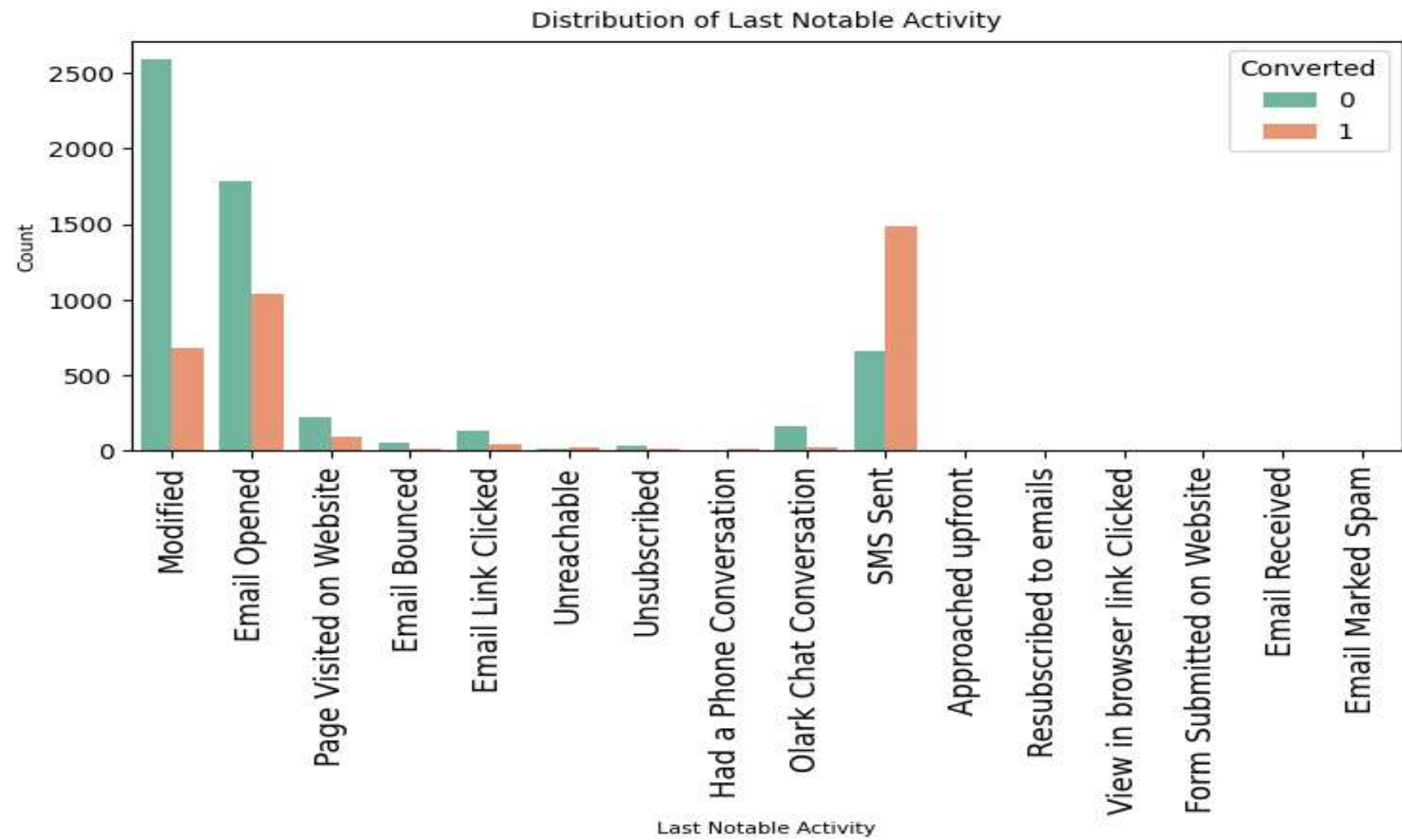
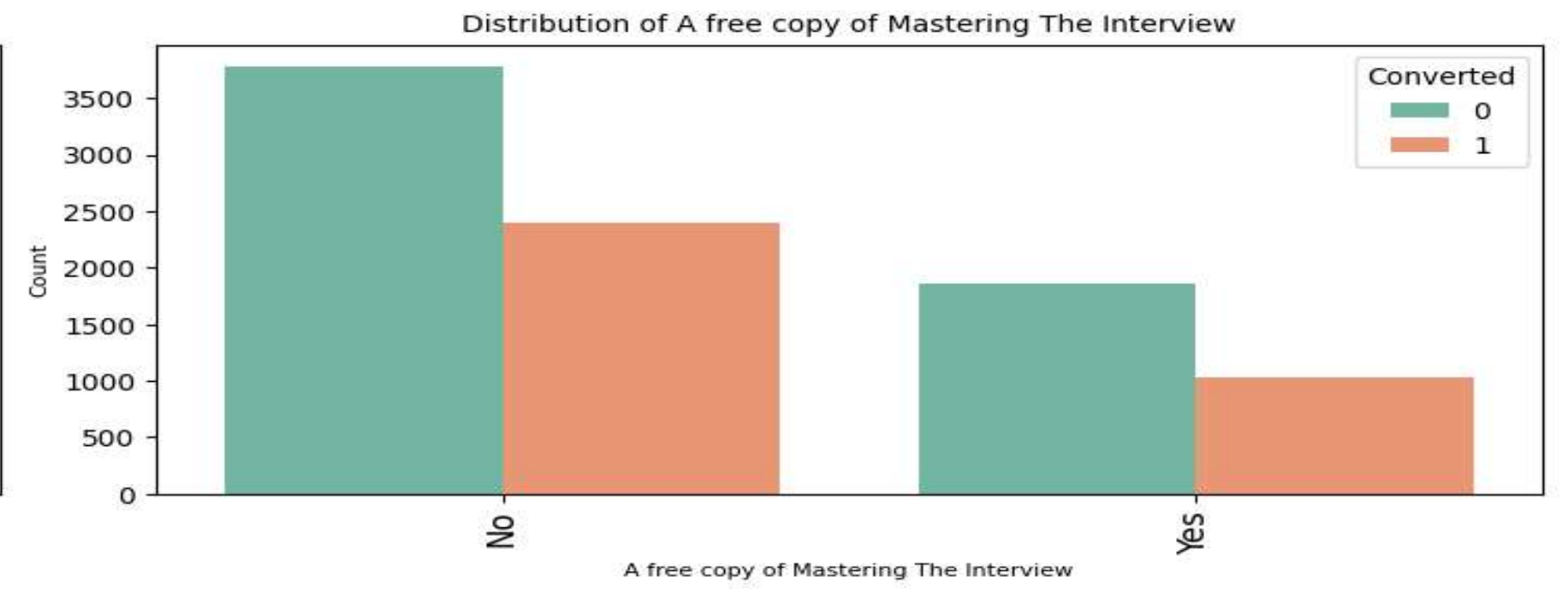
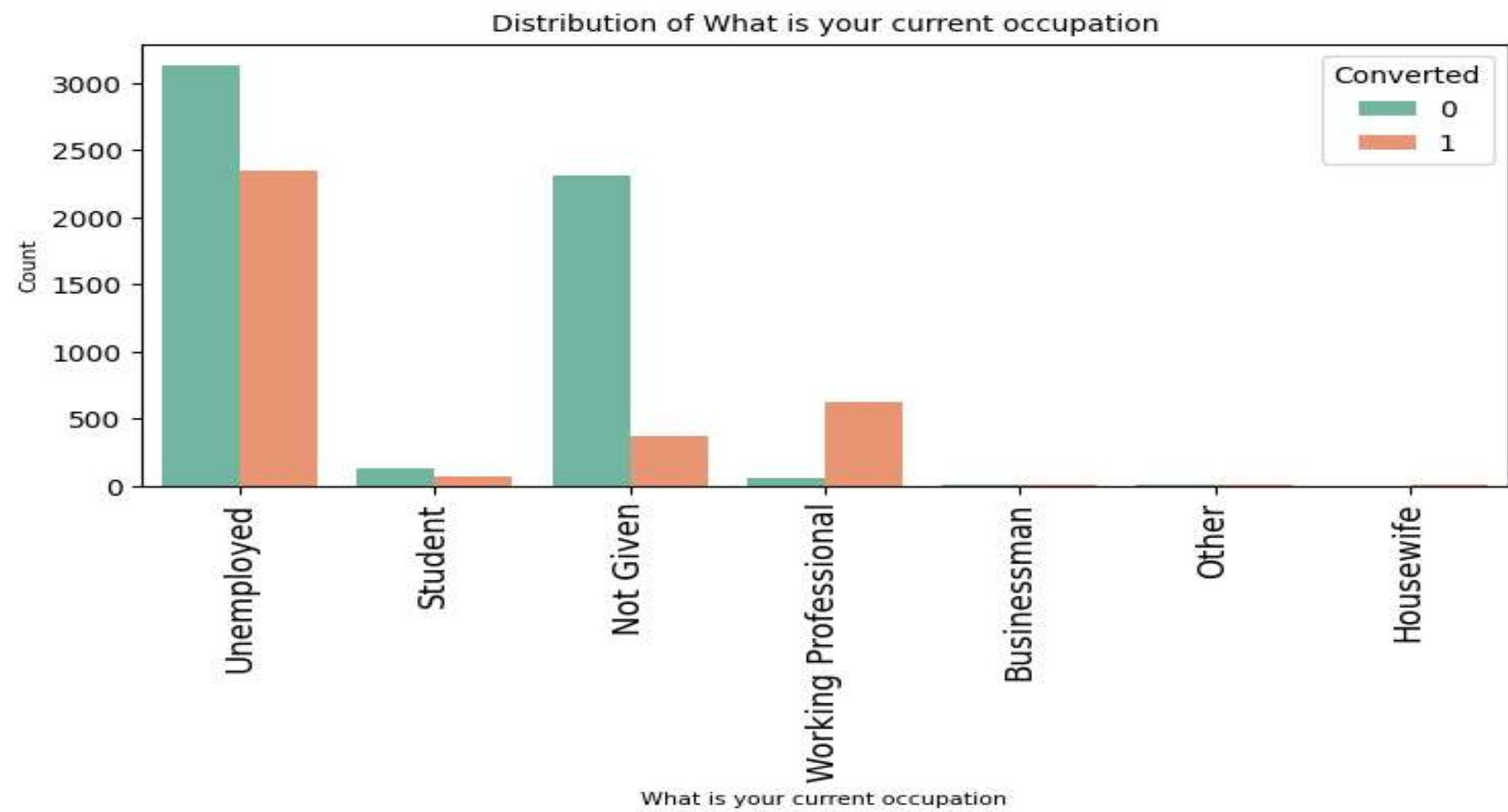
Target Variable

As we can see in the donut chart that conversion rate is only 37.9%

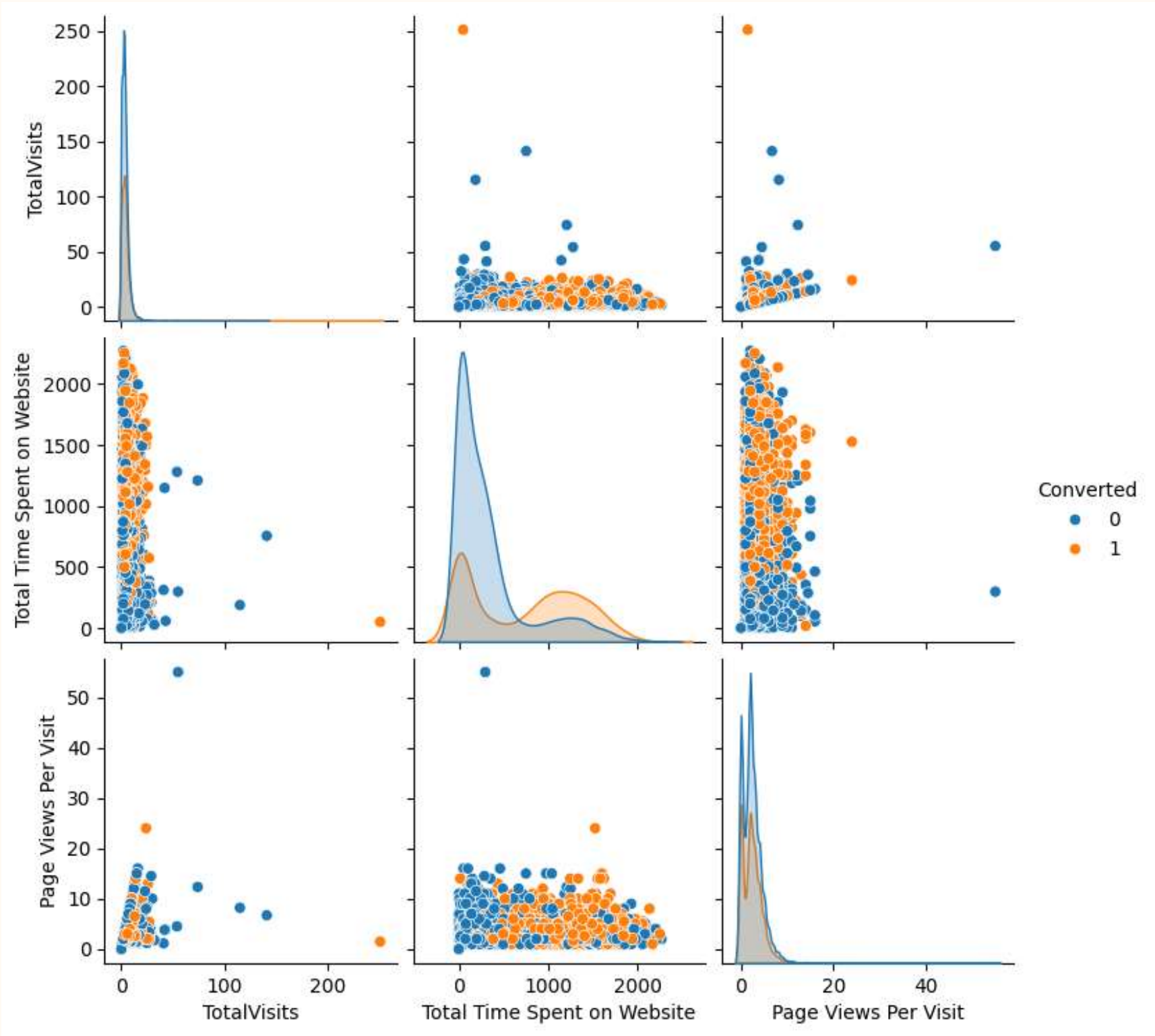
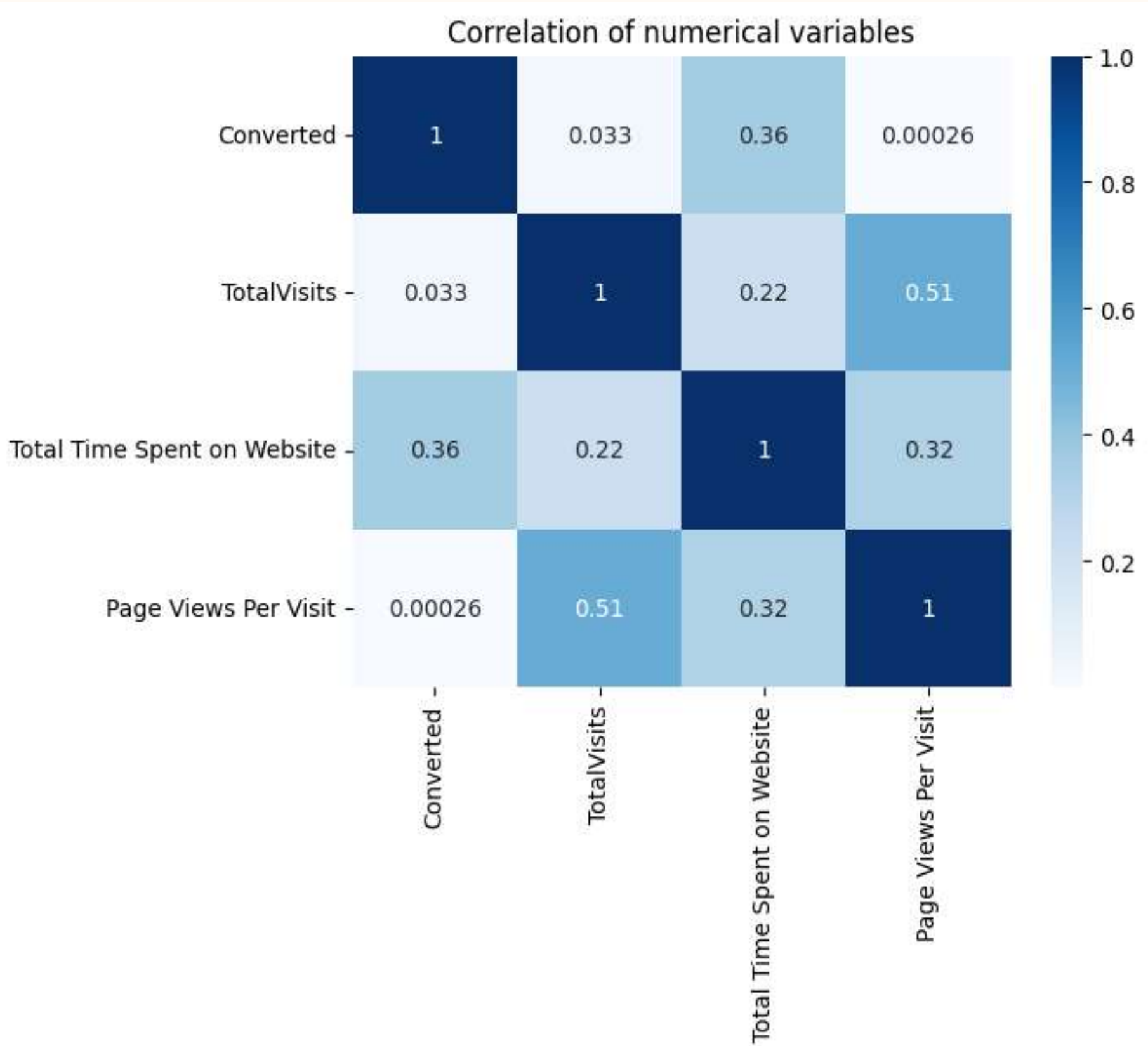


Bivariate Analysis

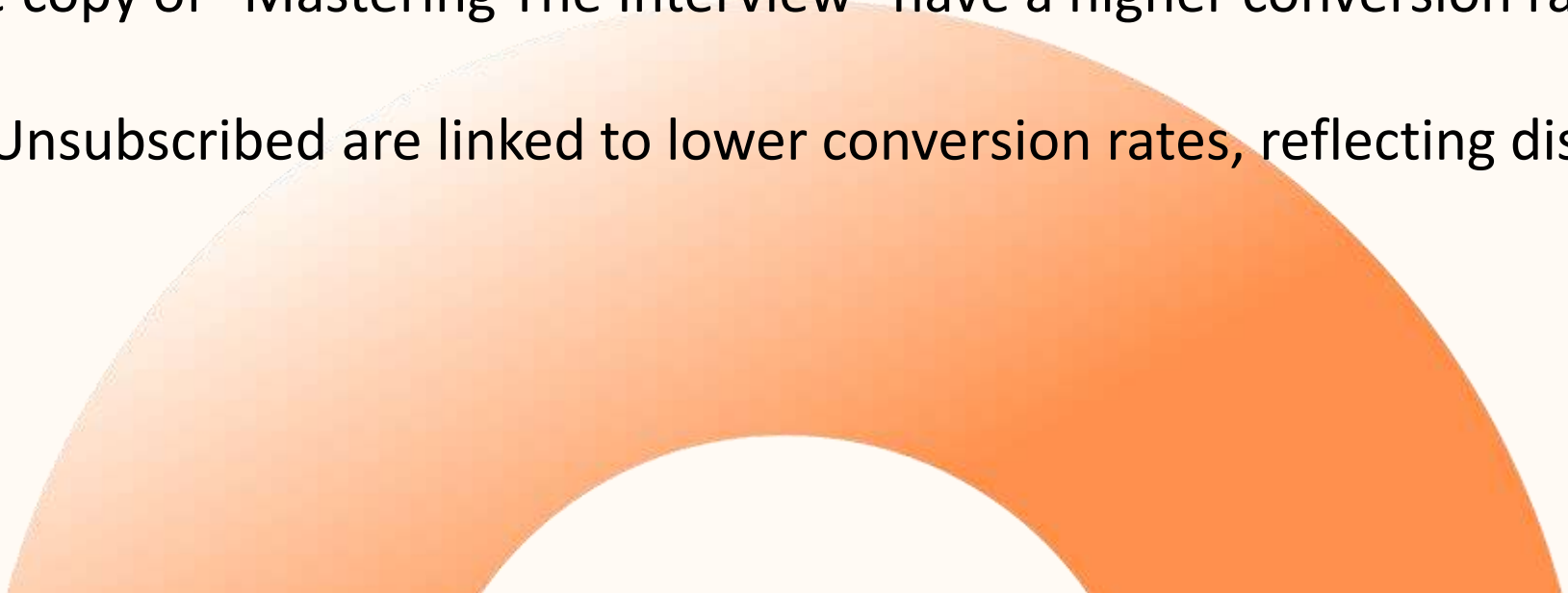




Multivariate Analysis



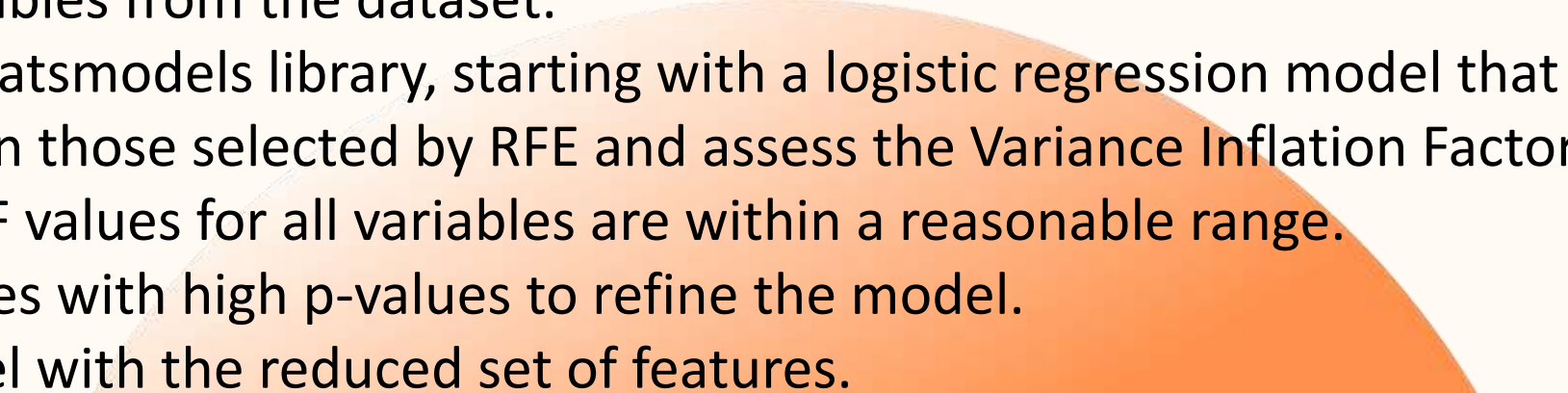
Observation from EDA

- Landing Page Submission attracts many leads but has only moderate conversion rates, indicating lower lead quality.
 - Lead Add Form has a strong conversion rate, highlighting the acquisition of higher-quality leads.
 - API and Lead Import channels have lower conversion rates, suggesting less effective lead acquisition.
 - Google, Direct Traffic, and Olark Chat are highly effective with above-average conversion rates, making them key lead sources.
 - Facebook, Pay-per-Click Ads, and Social Media generate many leads but have lower conversion rates, possibly due to weaker audience targeting.
 - Email Opened and SMS Sent activities correlate with higher conversions, emphasizing the impact of proactive communication.
 - Page Visited on Website generates significant traffic but does not correlate with high conversions, indicating low engagement.
 - Working Professionals have higher conversion rates compared to Students or Unemployed, making them a key target group.
 - Leads with specializations like Finance Management or Marketing Management are more likely to convert than those without a listed specialization.
 - Leads that did not request a free copy of "Mastering The Interview" have a higher conversion rate, indicating more serious buyers.
 - Activities like Email Bounced or Unsubscribed are linked to lower conversion rates, reflecting disengagement.
- 

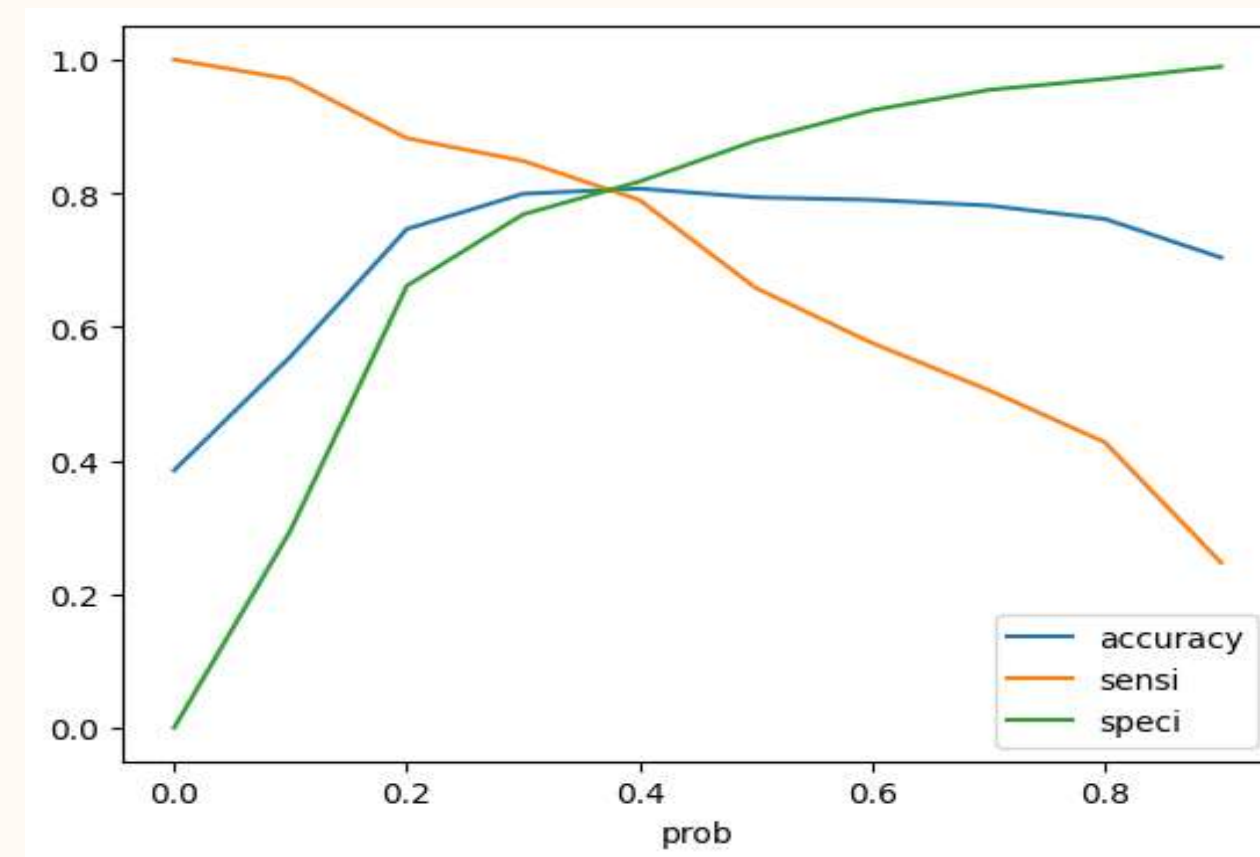
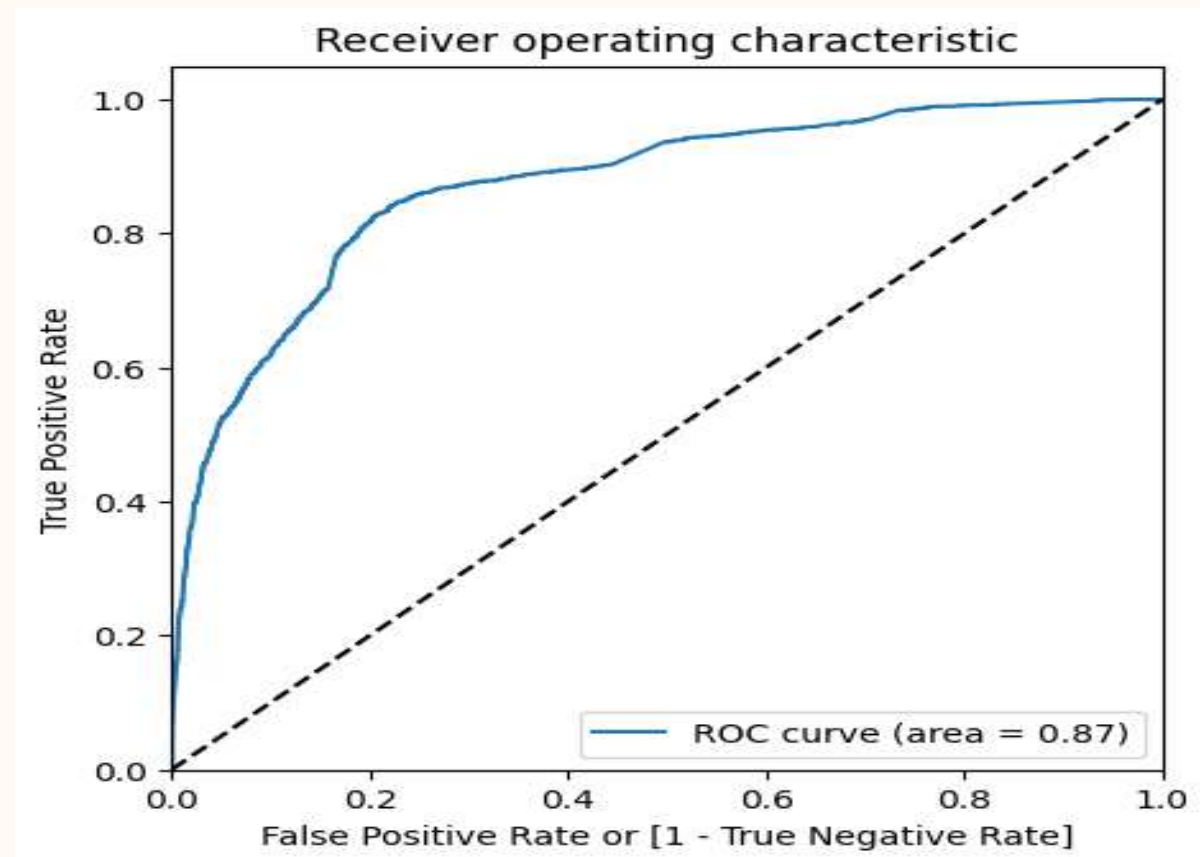
Prepare the data for model building

- Create dummy variables for categorical features to convert them into numerical form.
- Drop the original categorical columns after creating dummy variables to avoid redundancy.
- Separate the dataset into dependent (target) and independent (predictor) variables.
- Split the data into training and testing sets for model building and evaluation.
- Scale the numeric variables that have varying scales to standardize the data.
- Examine the correlation matrix to check for multicollinearity among the features.

Build the logistic regression model

- The dataset contains many variables that are challenging to manage, so a practical approach is to use Recursive Feature Elimination (RFE) to select a smaller set of important features.
 - Initially, a model built with the sklearn library achieved an accuracy of only 48%, indicating that using all variables results in low performance.
 - RFE, or Recursive Feature Elimination, is a technique in machine learning for selecting the most significant features for model development.
 - The goal is to select 15 key variables from the dataset.
 - Next, build a model using the statsmodels library, starting with a logistic regression model that includes all variables.
 - Create model variables based on those selected by RFE and assess the Variance Inflation Factor (VIF) to check for multicollinearity, noting that VIF values for all variables are within a reasonable range.
 - Subsequently, eliminate variables with high p-values to refine the model.
 - Finally, construct the final model with the reduced set of features.
- 

Evaluate the model's performance



Observations

- The area under the ROC curve is 0.87.
- The graph indicates that the optimal cut-off point is at 0.35.

Confusion Matrix

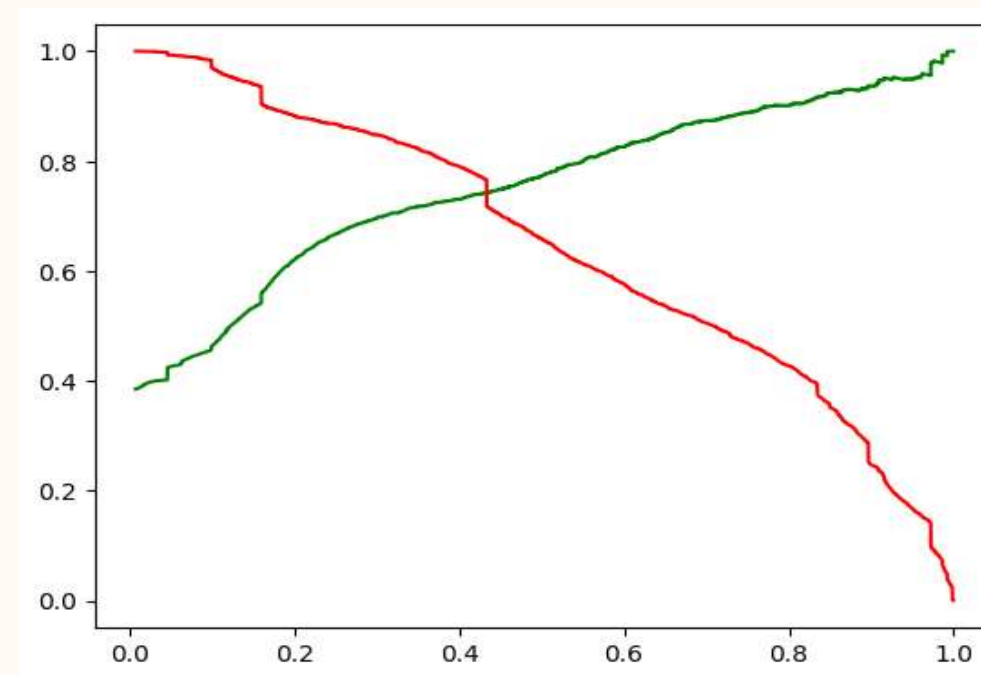
Predicted/ Actual	Not Converted	Converted
Not Converted	3113	792
Converted	432	2014

- Accuracy - 81%
- sensitivity - 82%
- specificity - 80%
- false positive rate - 20%
- Positive predictive value - 72%
- Negative predictive value - 88%

Checking Precision and Recall

Predicted/ Actual	Not Converted	Converted
Not Converted	3431	474
Converted	836	1610

- Precision - 77%
- Recall - 66%



- This graph shows an optimal cutoff of 0.42 based on precision and recall

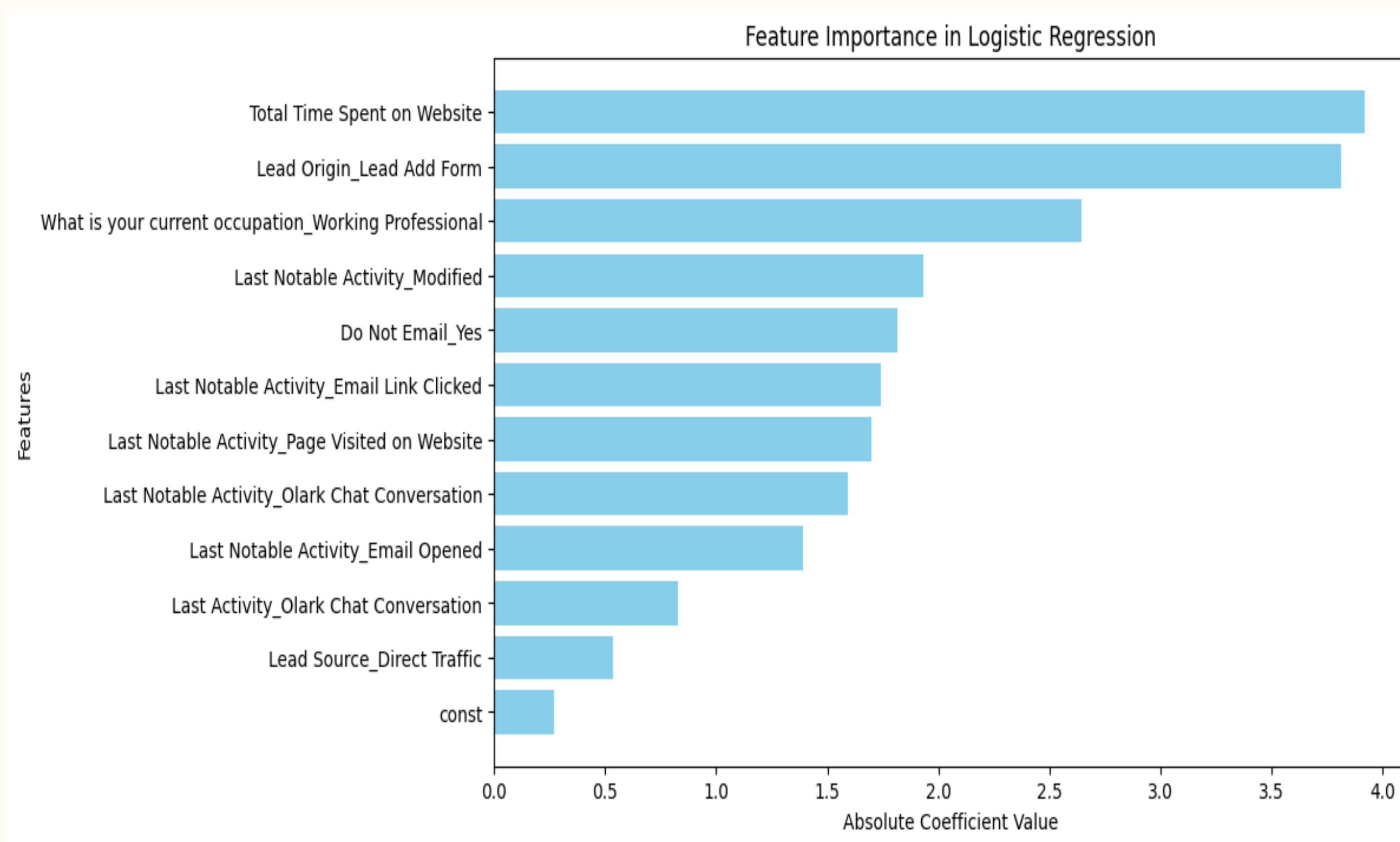
Making predictions on the test set

Predicted/ Actual	Not Converted	Converted
Not Converted	1377	357
Converted	191	798


- Accuracy - 80%
- sensitivity - 80%
- specificity - 79%

Determining the important features

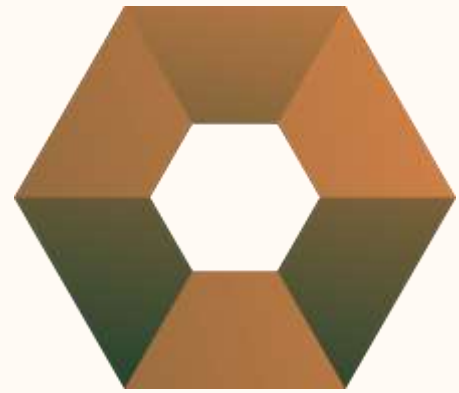
Features with larger absolute values of coefficients have a stronger impact on the target variable.



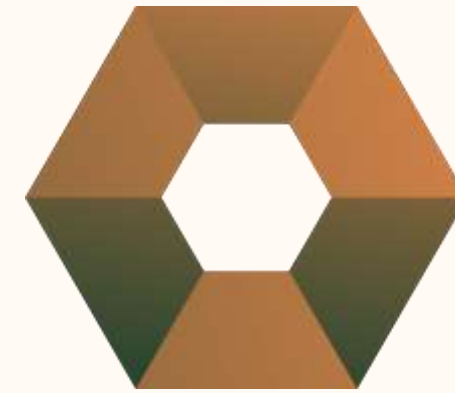
Final Observation

- Total Time Spent on Website strongly boosts lead conversion, highlighting the importance of engagement.
 - Lead Add Form Origin leads have a higher conversion likelihood.
 - Working Professionals are more likely to convert than other occupations.
 - Last Notable Activity Modified negatively impacts conversion.
 - Opting out of emails reduces conversion rates.
 - Email Link Clicked and Page Visited on Website alone don't strongly predict conversion.
 - Olark Chat Conversation negatively affects conversion.
 - Direct Traffic leads convert at a lower rate than other sources.
- 

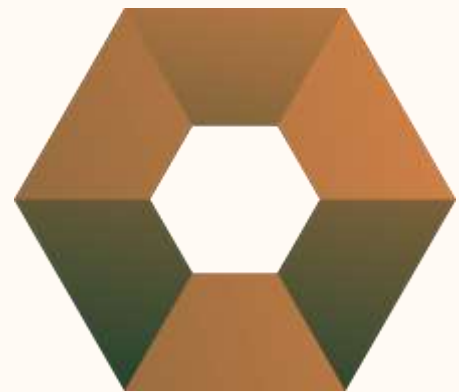
Recommendations



Increase website engagement:
Encourage leads to spend more time on the website, as higher engagement correlates with better conversion rates.



Target working professionals, as they convert at a higher rate.



Leverage Lead Add Forms
Focus on generating more leads through Lead Add Forms, as they are associated with higher conversion likelihood.

Thank you !

