


X Education- Lead Scoring Case Study

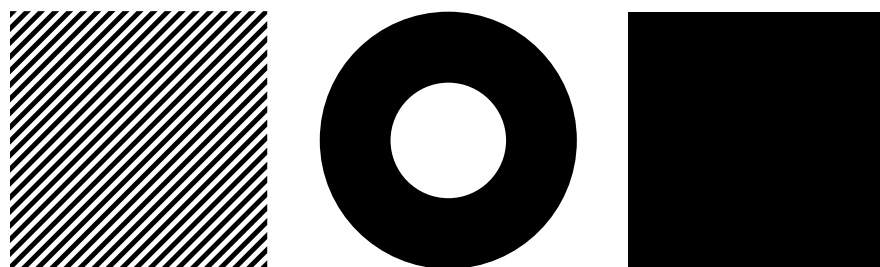
Optimizing Marketing Efforts to Target Hot
Leads and Enhance Conversion Rates for X
Education

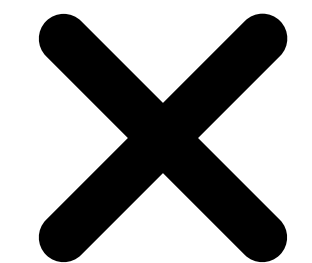
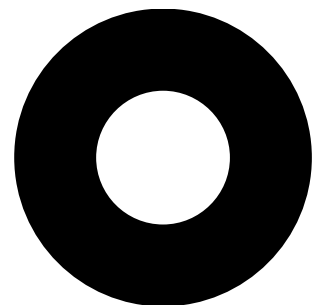
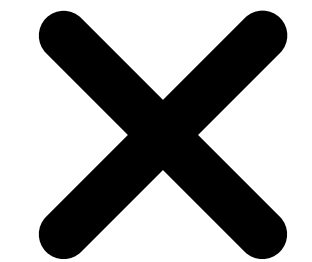
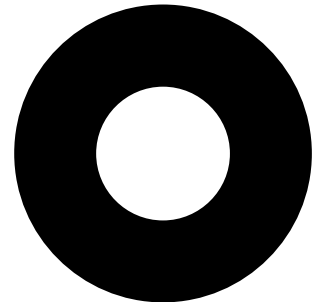
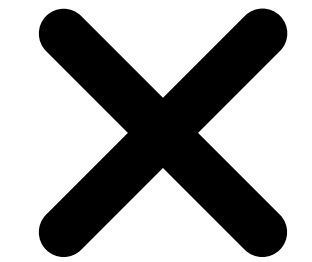


Group Members:
Samith &
Desh
Deepak

Contents

- 
- History of X Education Company
 - Research Question and Study Goals
 - Recommended Strategies for Converting Leads
 - Analytical Method
 - Data Sanitization
 - Exploratory Data Analysis
 - Data Preprocessing
 - Constructing a model using Recursive Feature Elimination (RFE) and manually refining it.
 - Assessing the performance of a model
 - Suggestions





History of X Education Company

-
- X Education is an educational company specializing in online courses tailored for industry professionals.
 - Every day, numerous professionals visit their website to explore available courses.
 - The company promotes its courses across various websites and search engines, including Google.
 - Once visitors arrive at the website, they may explore our courses, fill out course registration forms, or watch instructional videos.
 - When individuals fill out a form with their email address or phone number, they are classified as leads.
 - Once these leads are acquired, the sales team initiates contact through calls, emails, and other communication channels.
 - During this process, some leads are converted, while the majority are not.
 - The average lead conversion rate at X education stands at approximately 30%.

Research Question and Study Goals

Research Question:

- X Education generates a high volume of leads, but its lead conversion rate remains low, hovering at approximately 30%.
- X Education aims to enhance its lead conversion process by pinpointing the most promising prospects, often referred to as "Hot Leads."
- Their sales team wants to prioritize communication with a specific set of potential leads rather than making calls to everyone indiscriminately.

Study Goals:

- To assist X Education in identifying the most promising leads—those with the highest likelihood of converting into paying customers.
- The company needs us to create a model that assigns a lead score to each lead. This score should accurately reflect the likelihood of conversion, ensuring that leads with higher scores have a greater chance of converting, while those with lower scores are less likely to convert.
- The CEO has indicated that the target lead conversion rate should ideally be around 80%.

Recommended Strategies for Converting Leads

Grouping of Leads

- Leads are categorized according to their likelihood or propensity to convert.
- This leads to a concentrated group of highly interested prospects.

Enhancing Communication

- We could focus on a more targeted pool of leads, enabling us to make a greater impact through our communications.

Increase Conversion

- By focusing on hot leads that are more likely to convert, we can achieve a higher conversion rate and reach our 80% objective effectively.

To achieve our 80% conversion rate target, it's crucial to prioritize high sensitivity in identifying hot leads.

Analytical Method



Data Sanitization:

- Loading, comprehending, and refining dataset

Exploratory Data Analysis:

- Analyze imbalance through:

1. Univariate exploration
2. Bivariate examination

Data Preprocessing:

- Categorical variables represented as indicators, dividing data for training and testing, adjusting feature magnitudes

Constructing a model:

- Request for Elimination of top 15 features, manual Feature Reduction & Finalization of Model

Assessing the performance of a model:

- Matrix illustrating classification model performance.
- Choosing threshold for classification decisions.
- Determining score for potential leads.

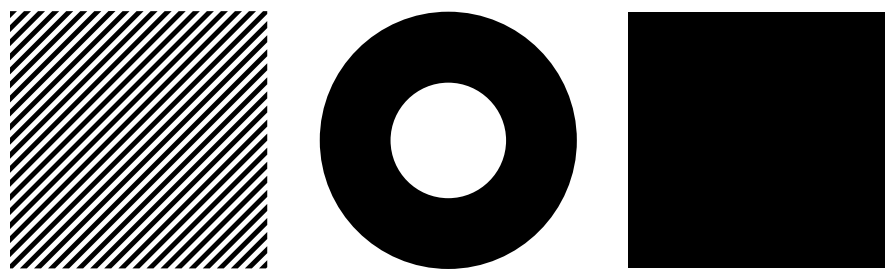
Forecasts for the Test Data:

- Compare the metrics between training and testing, assign a lead score, and identify the most important features.

Suggestions:

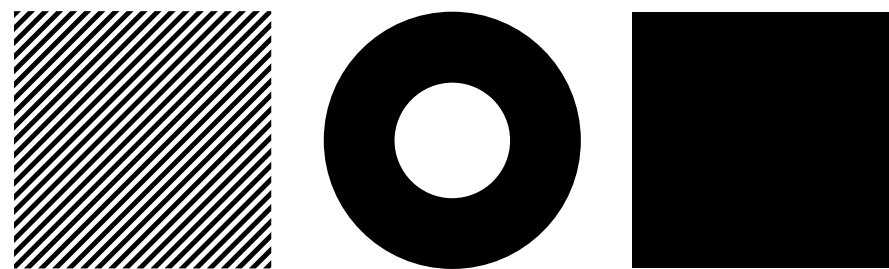
- Identify the top three features to prioritize for increasing conversions, and pinpoint areas where improvements can be made.

Data Sanitization



-
- The "Select" level indicates that certain categorical variables have null values, indicating that customers have not chosen any option from the list provided.
 - Columns containing more than 40% null values were removed/dropped.
 - Missing values in categorical columns were managed by evaluating value counts and other relevant criteria.
 - Exclude columns that do not contribute meaningful insights or value to the study objective, such as 'tags' and 'country'.
 - Imputation was employed for certain categorical variables.
 - Additional categories were introduced for certain variables.
 - The columns such as Prospect ID and Lead Number, which either have no relevance for modeling purposes or contain only one category of response, were excluded from the analysis.
 - Numerical data was imputed using the mode after verifying its distribution.

Data Sanitization

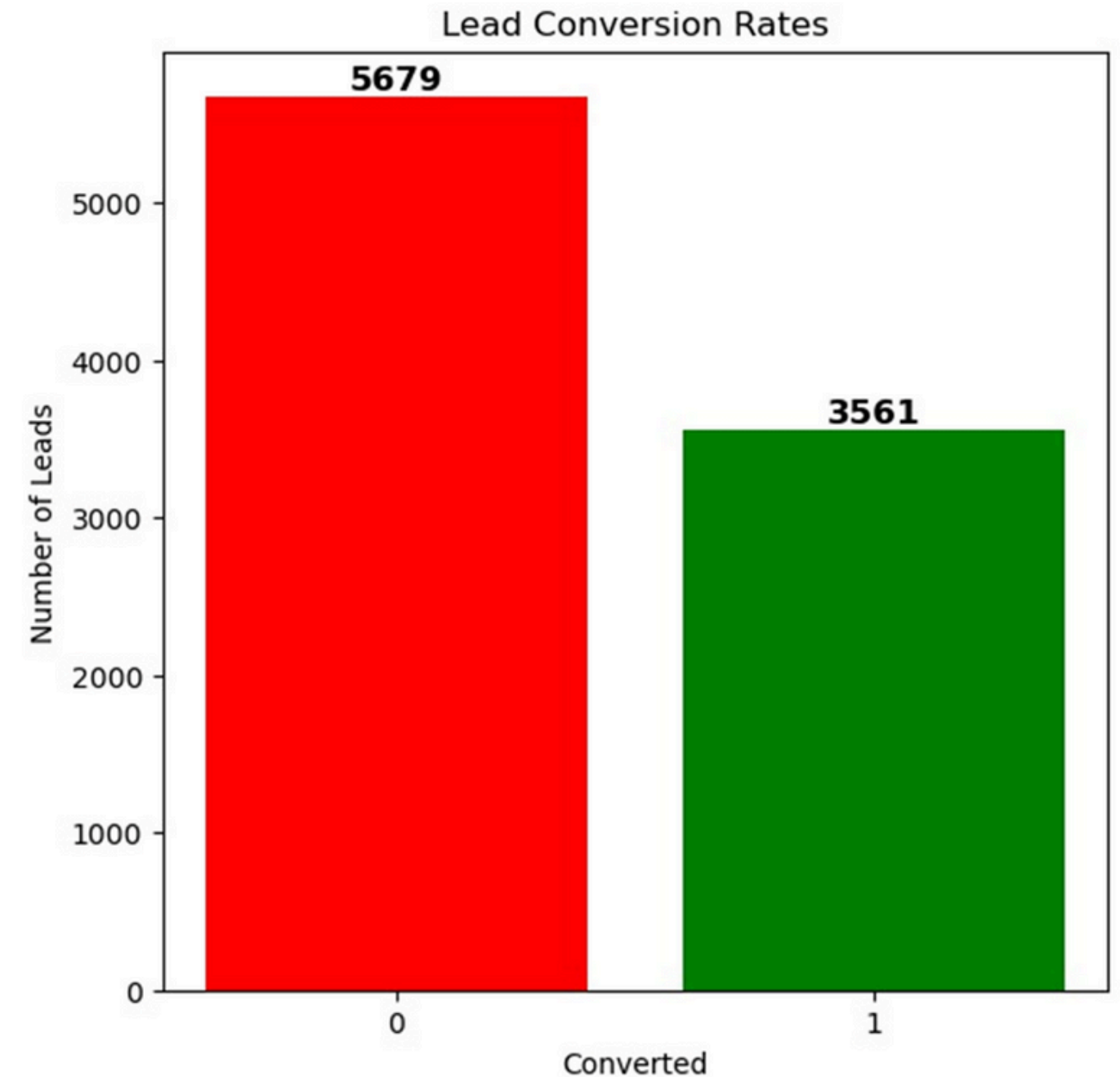


-
- Skewed category columns were identified and removed to mitigate bias in logistic regression models.
 - Outliers in TotalVisits and Page Views Per Visit were identified and subsequently capped.
 - Invalid values have been corrected, and data in certain columns, like lead source, has been standardized.
 - Low frequency values were consolidated under the category "Others".
 - Binary categorical variables were encoded.
 - Additional cleaning activities were conducted to enhance data quality and accuracy.
 - Standardizing data in columns by correcting casing styles and addressing invalid values (e.g., ensuring 'Google' is consistently formatted).

Exploratory Data Analysis

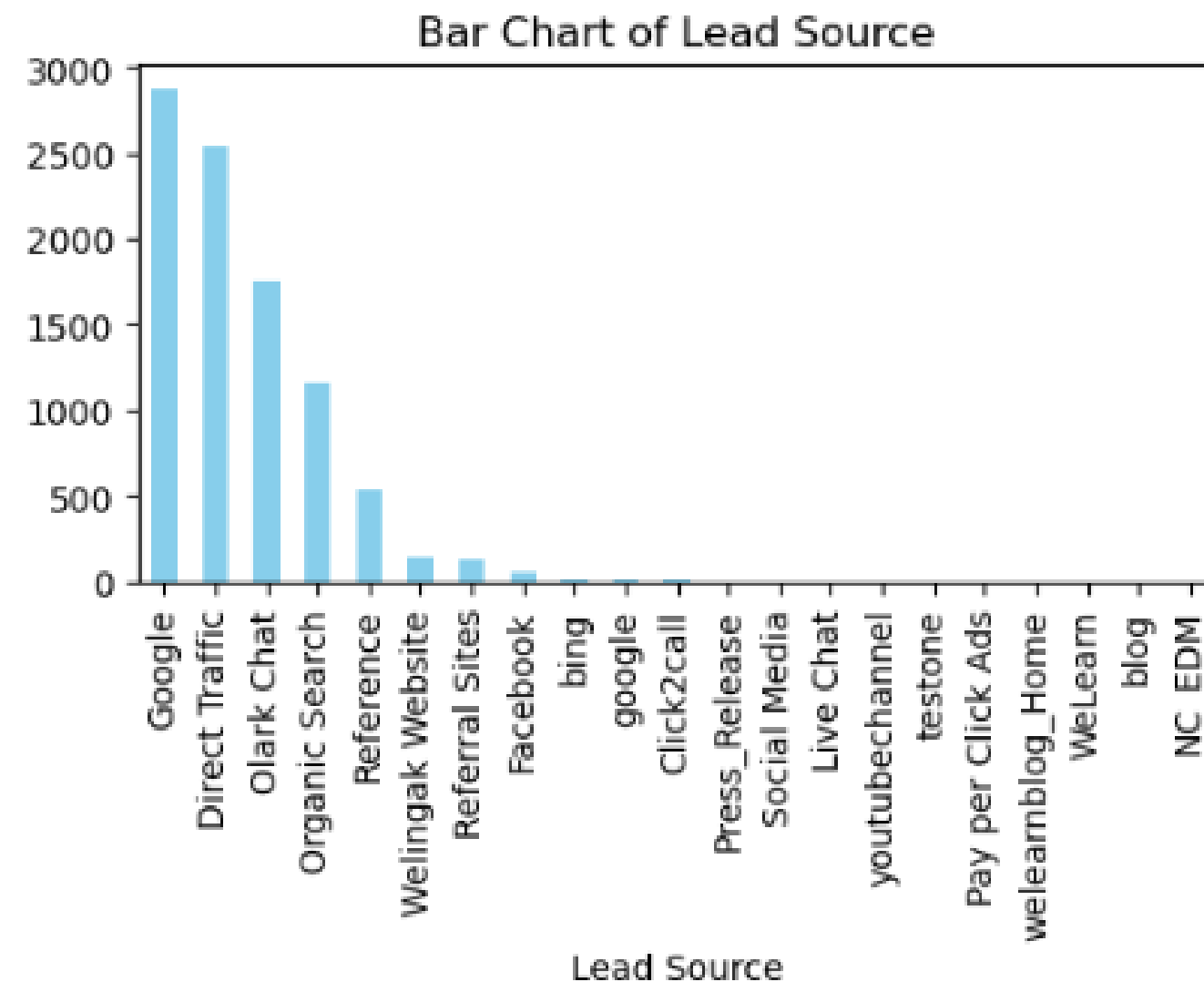
The target variable exhibits imbalance in the data analysis.

- The conversion rate stands at 3561, indicating that only a minority of individuals have transitioned into leads.
- 5679 of the people did not convert into leads.

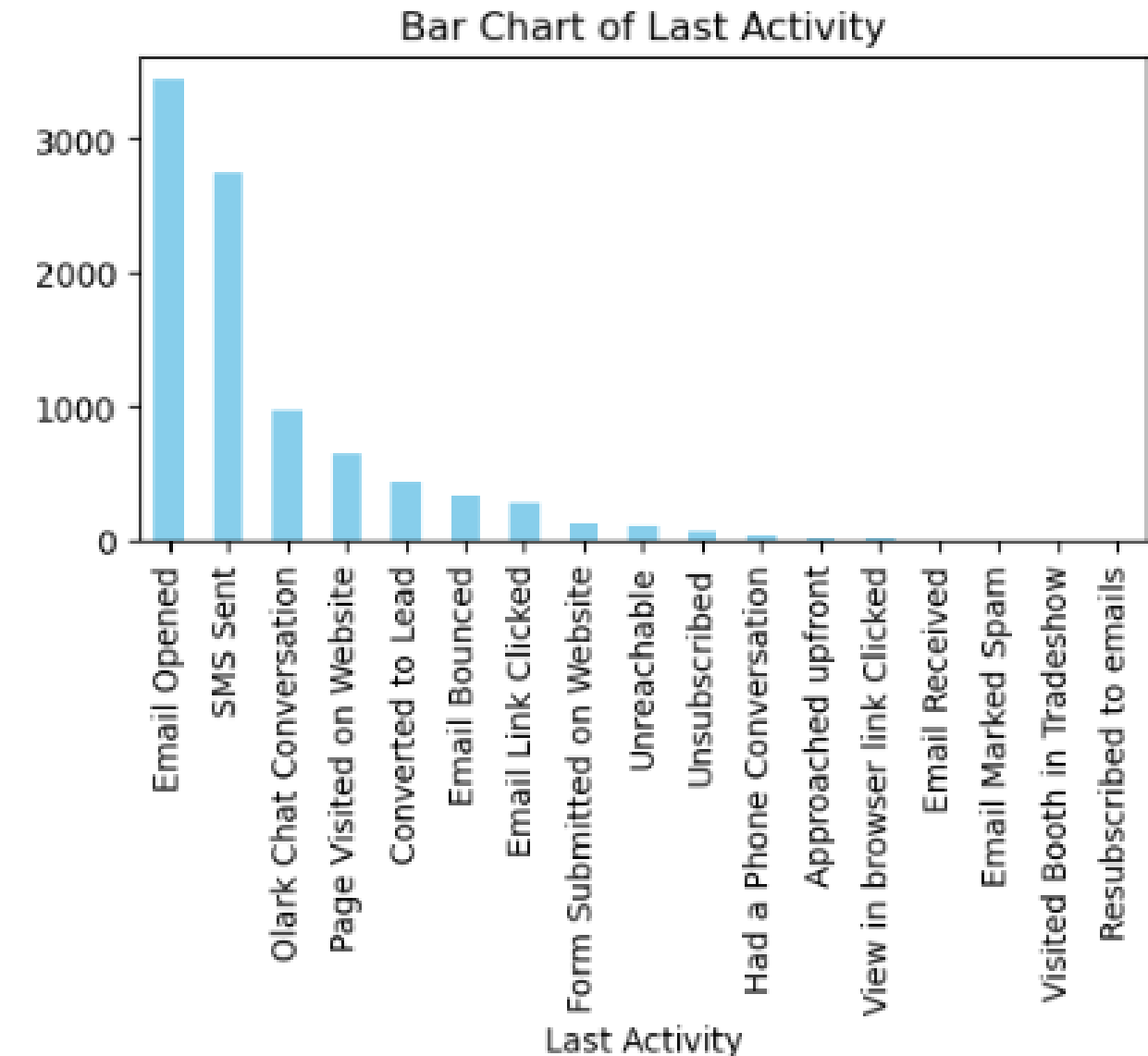


Exploratory Data Analysis

- Exploring Categorical Variables through Univariate Analysis



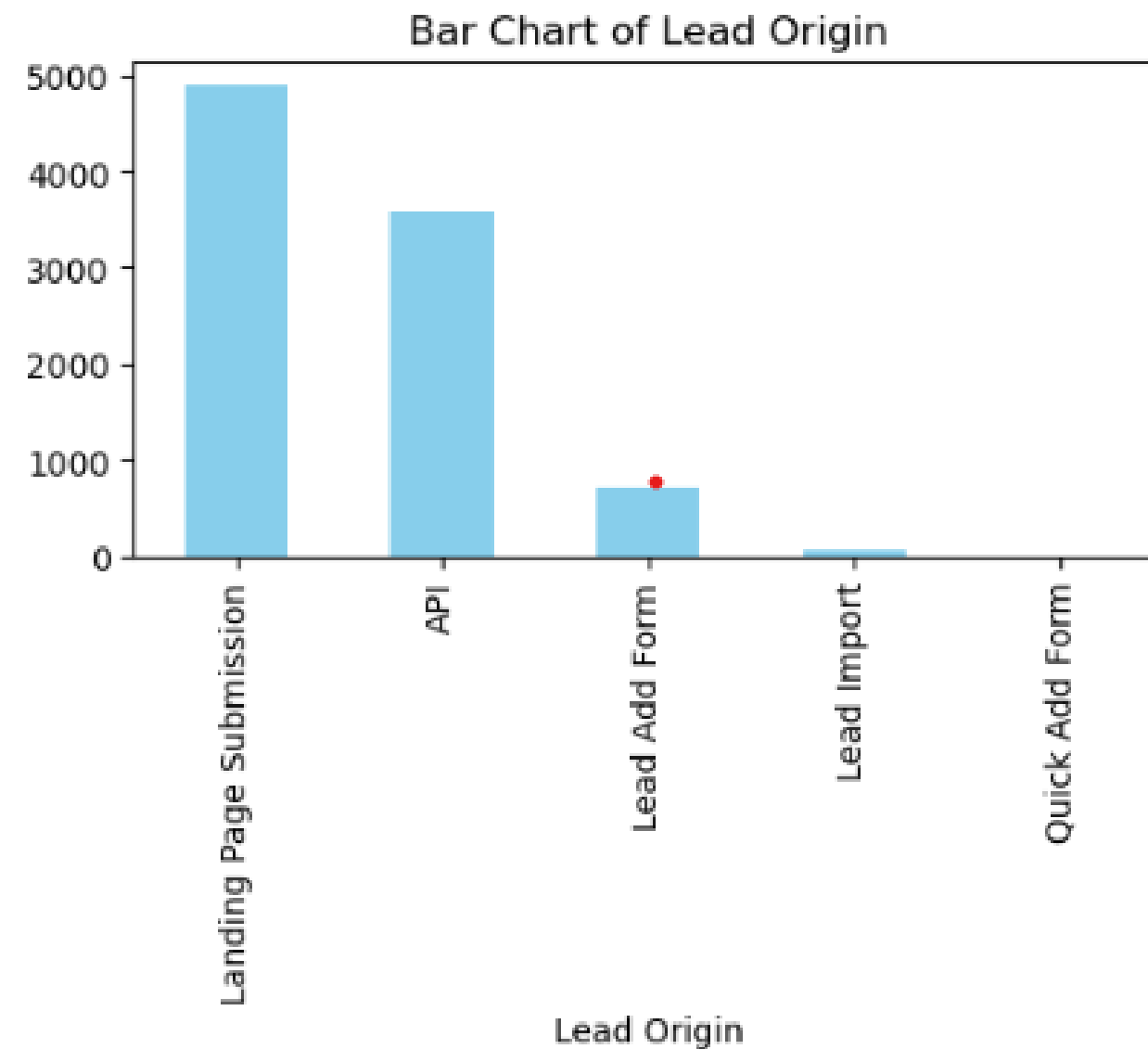
- Lead Source:** Most of the leads originate from a combination of Google search and direct traffic sources.



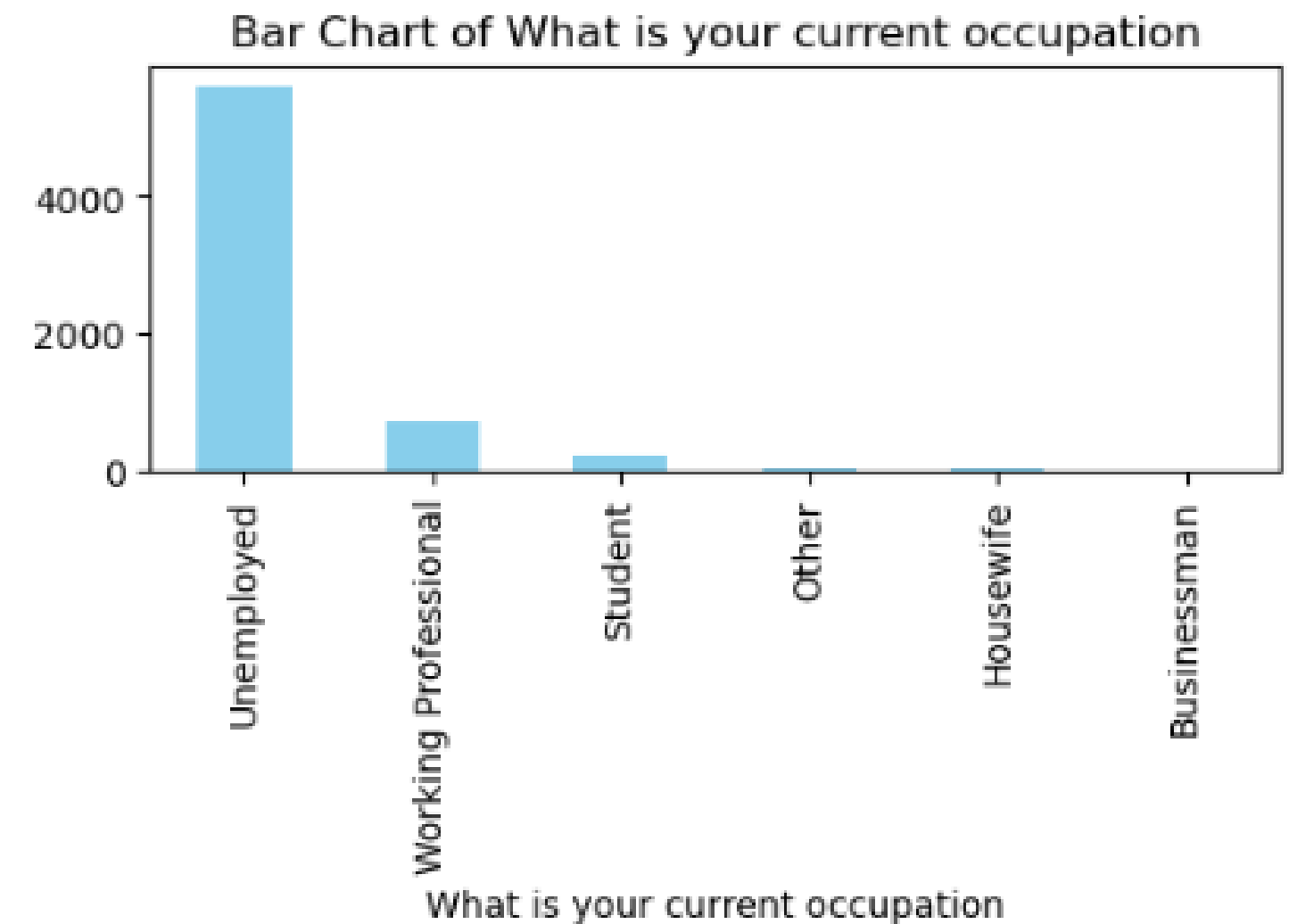
- Last Activity:** Most of the customers actively engage with SMS and Email activities, contributing significantly.

Exploratory Data Analysis

- Exploring Categorical Variables through Univariate Analysis

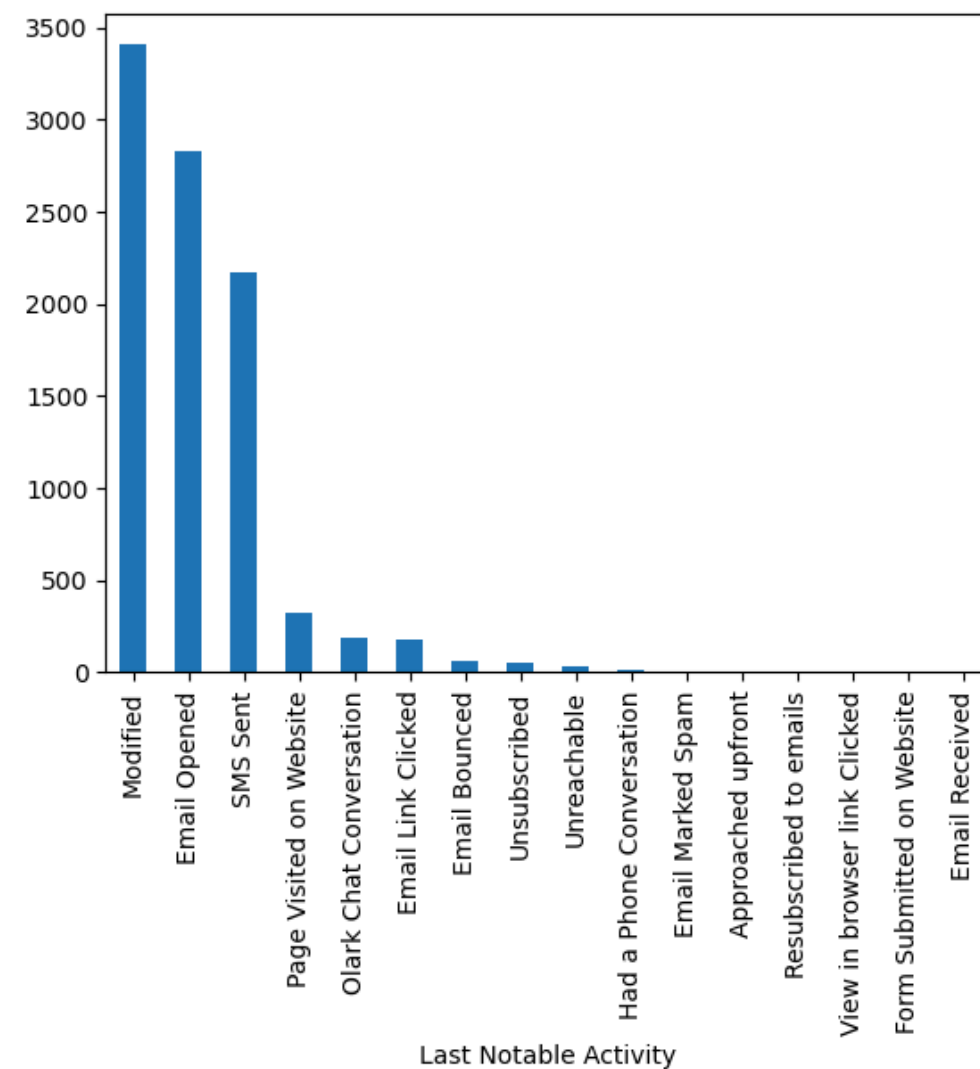


- Lead Origin:** Most of the customers were identified through 'Landing Page Submission', while others were identified through 'API'.



- Current_occupation:** The current occupation involves serving a clientele where most of the customers are unemployed.

Exploratory Data Analysis

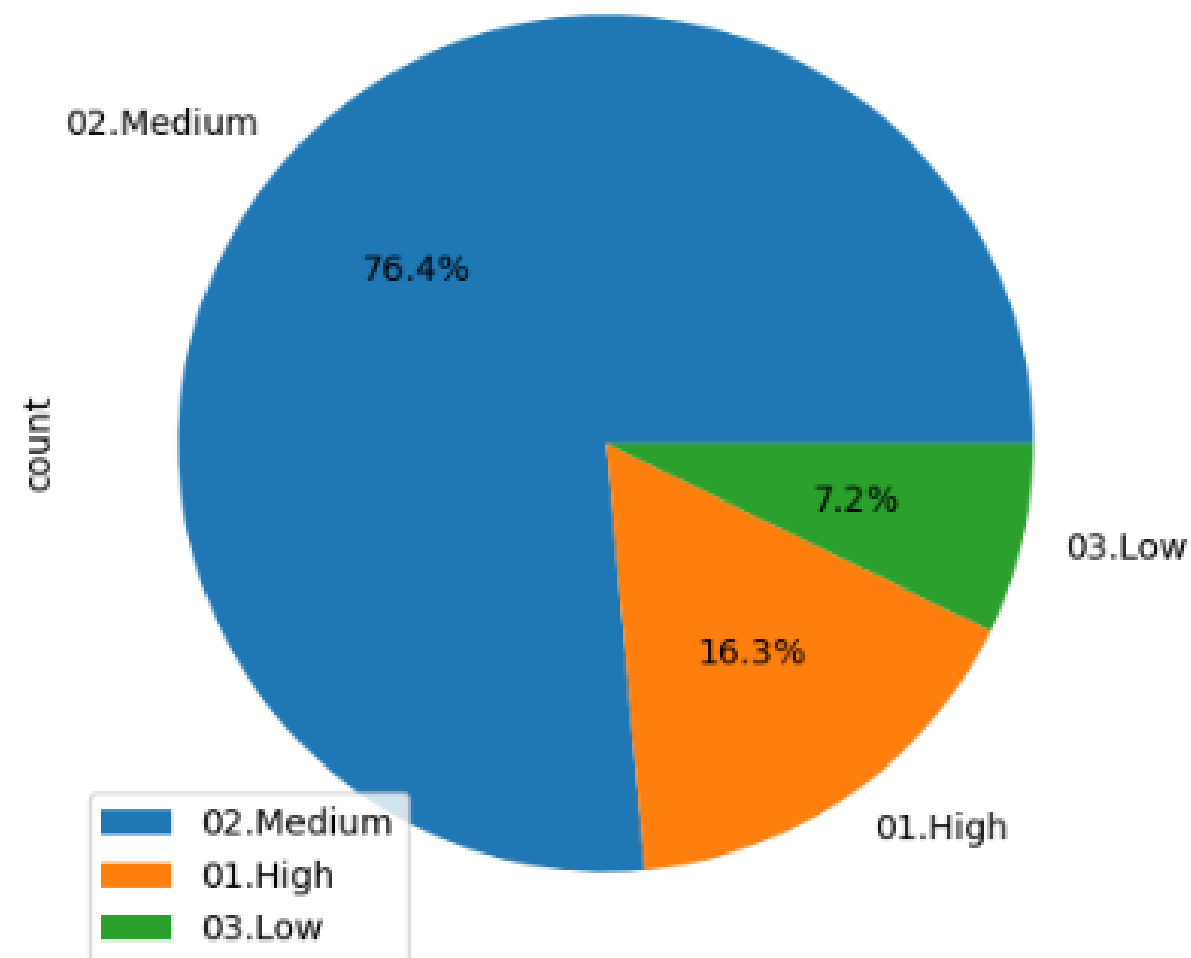


- The bar graph titled “Last Notable Activity” displays various customer engagement activities. Here are the key takeaways:
- Email Opened: The most frequent activity, indicating high customer interaction.
- SMS Sent: Second most common activity after email.
- Olark Chat Conversation: Moderate engagement through chat.
- Page Visited on Website: Indicates website visits.
- Email Link Clicked: Shows interest in email content.
- Email Bounced: Some emails failed to deliver.
- Unsubscribed: Customers opting out.
- Had a Phone Conversation: Indicates phone interactions.
- Approached Upfront: A less common activity.
- Resubscribed to Emails: Indicates re-engagement.
- This graph provides insights into customer behavior and marketing effectiveness.

Exploratory Data Analysis

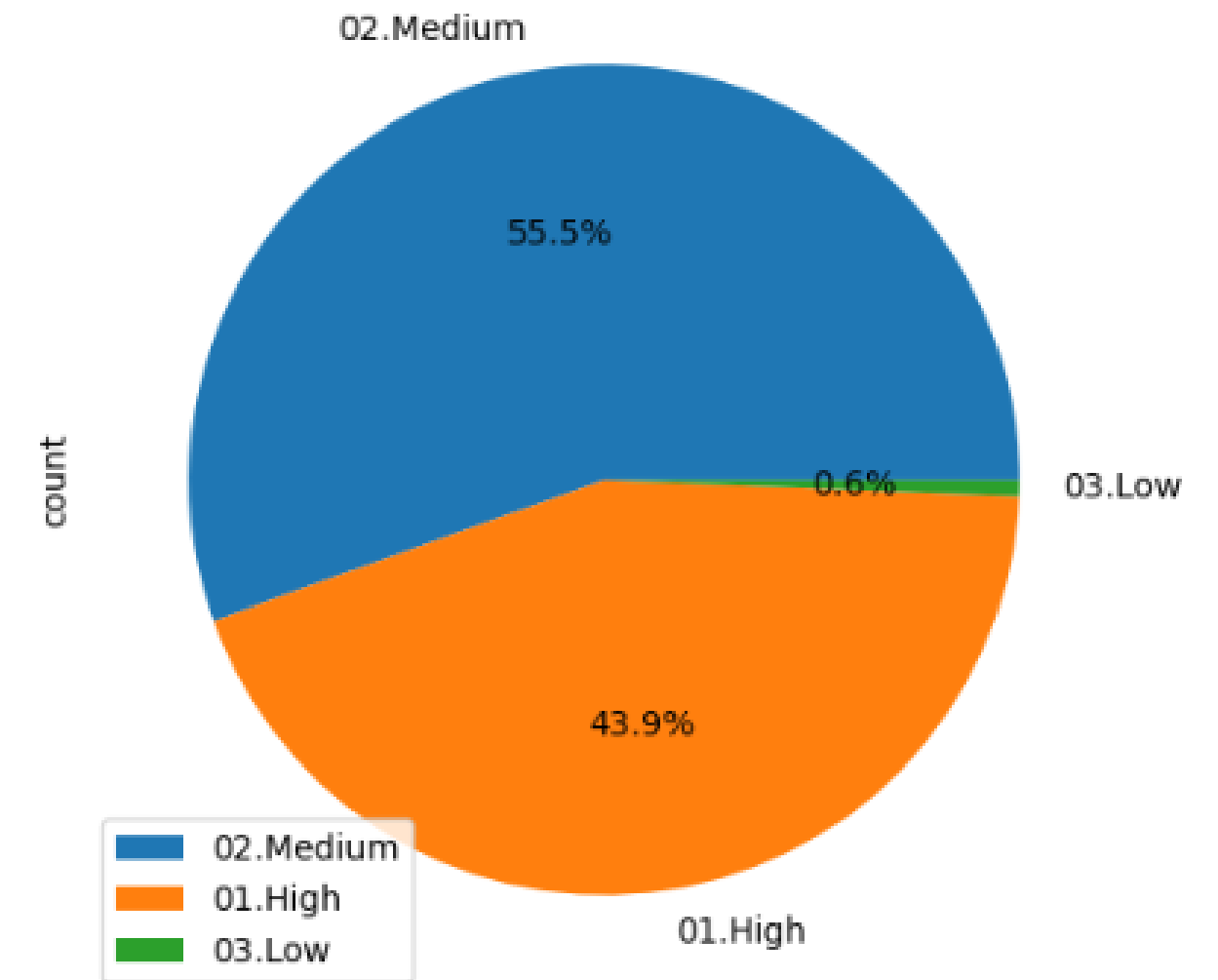
- Exploring Categorical Variables through Univariate Analysis

Pie Chart of Asymmetrique Activity Index



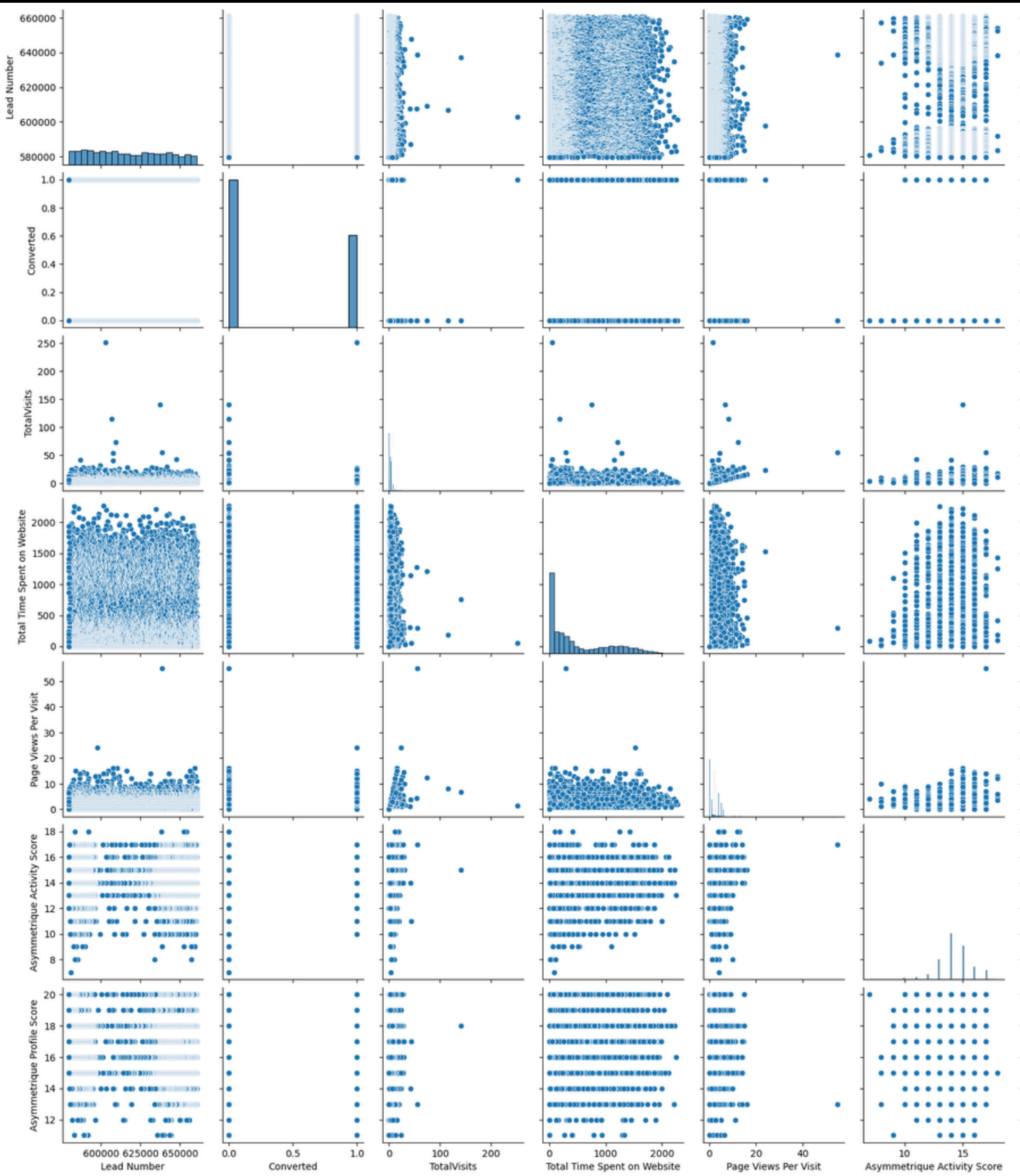
- Asymmetrique Activity Index:** The pie chart shows activity levels: Medium (76.4%), High (16.3%), Low (7.2%).

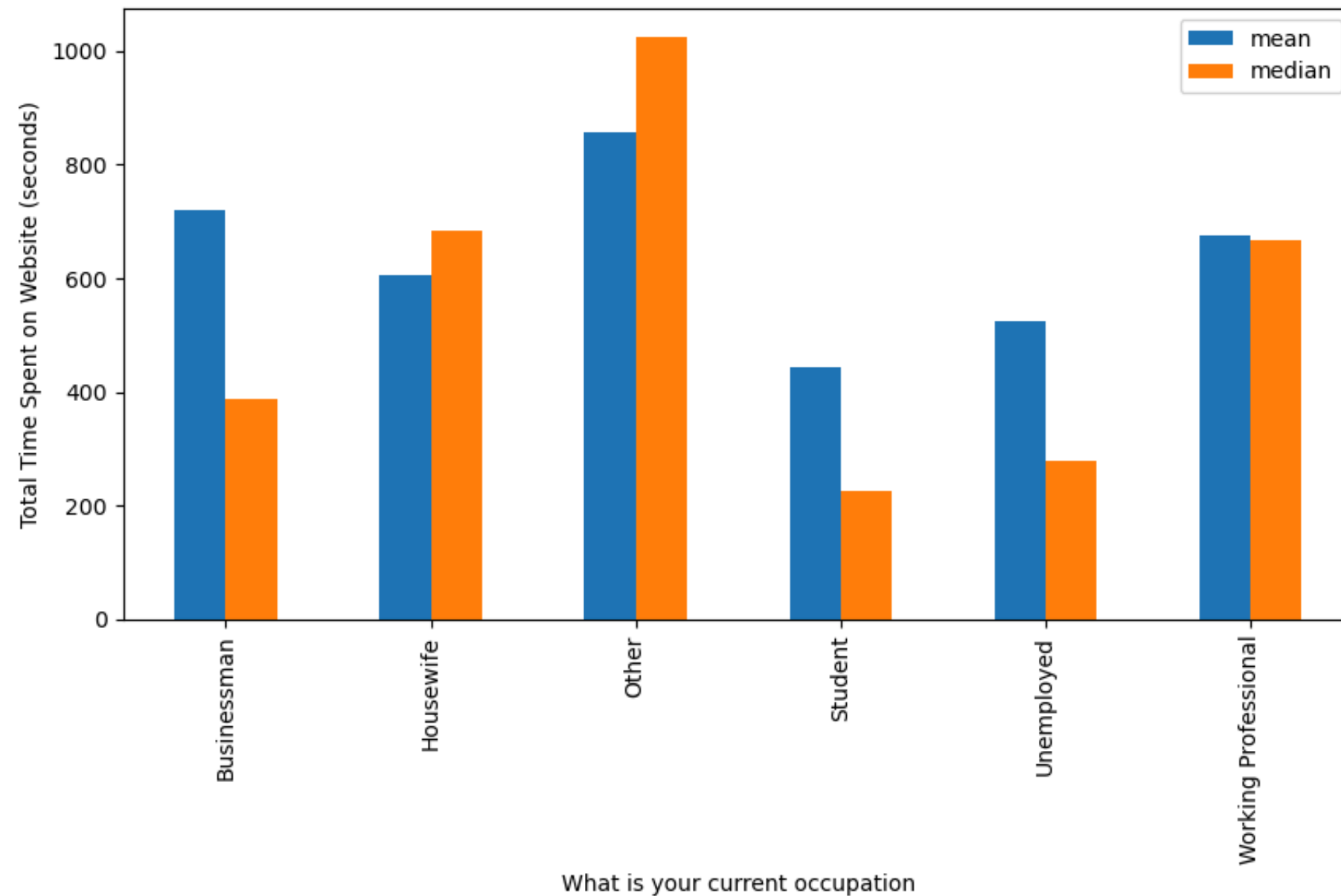
Pie Chart of Asymmetrique Profile Index



- Asymmetrique Profile Index:** The pie chart shows activity levels: Medium (55.5%), High (43.9%), and Low (0.6%).

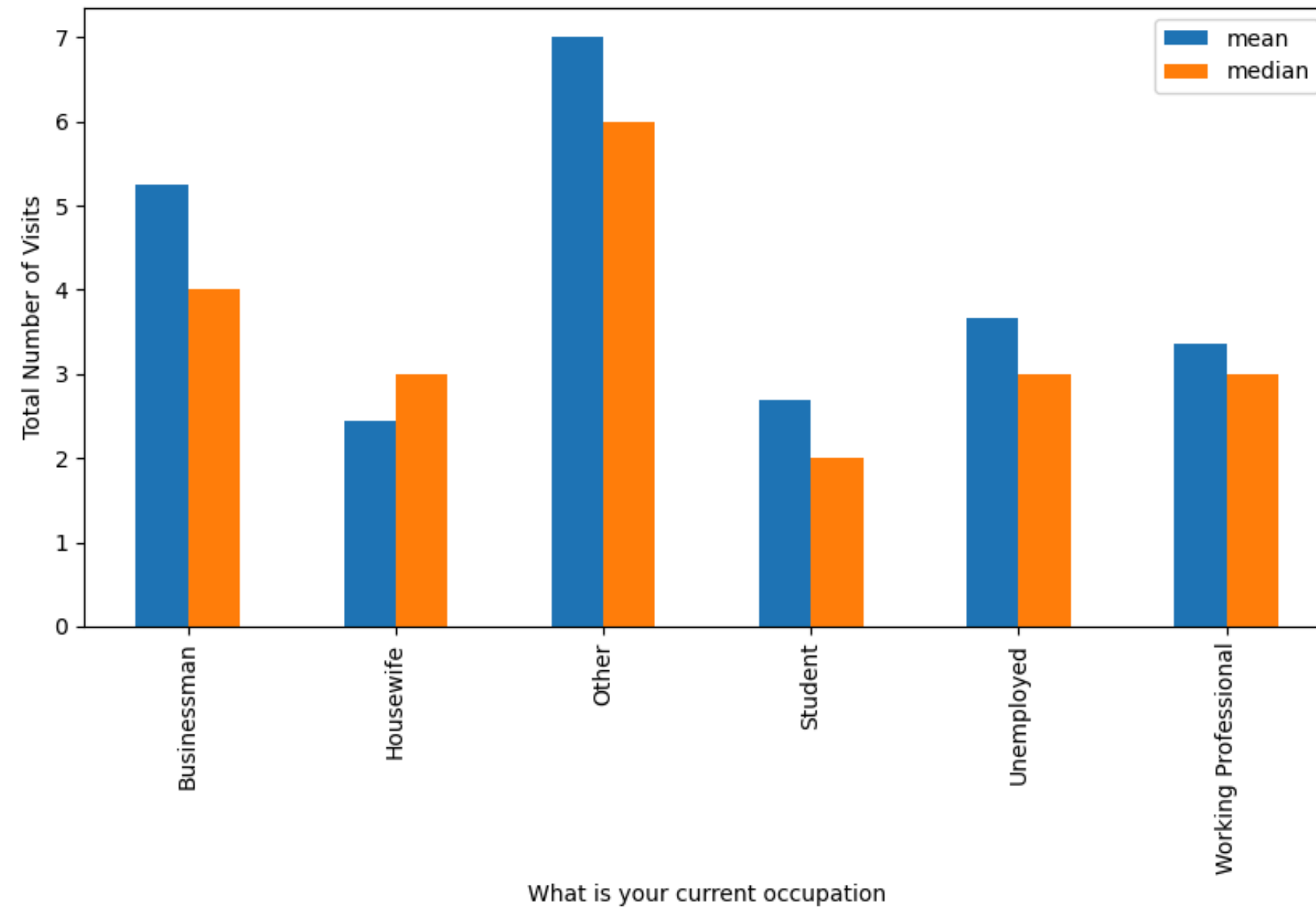
Exploratory Data Analysis-Bivariate Analysis for Numerical Variables





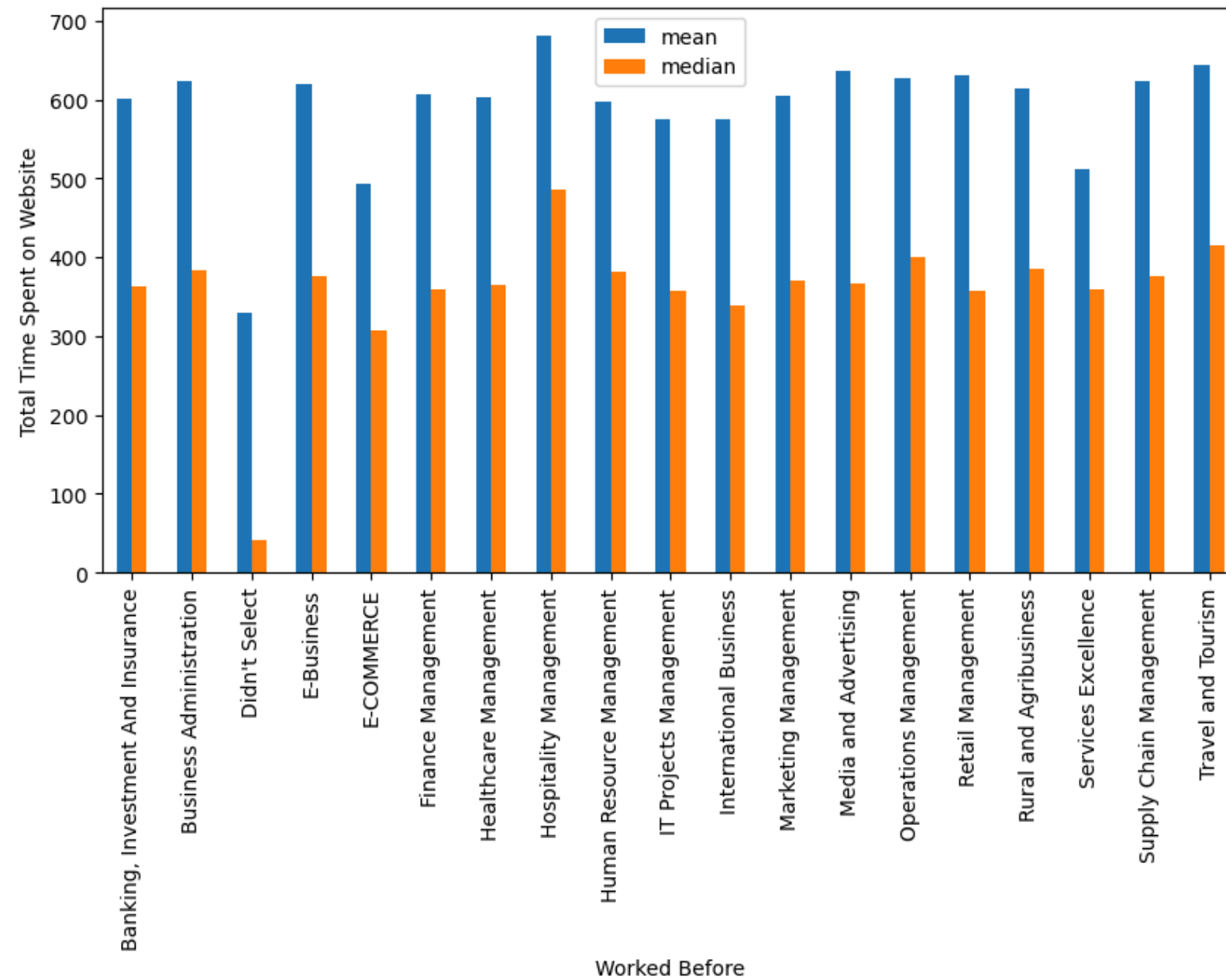
Exploratory Data Analysis

The bar graph displays the total time spent, along with the mean and median of current occupations.



Exploratory Data Analysis

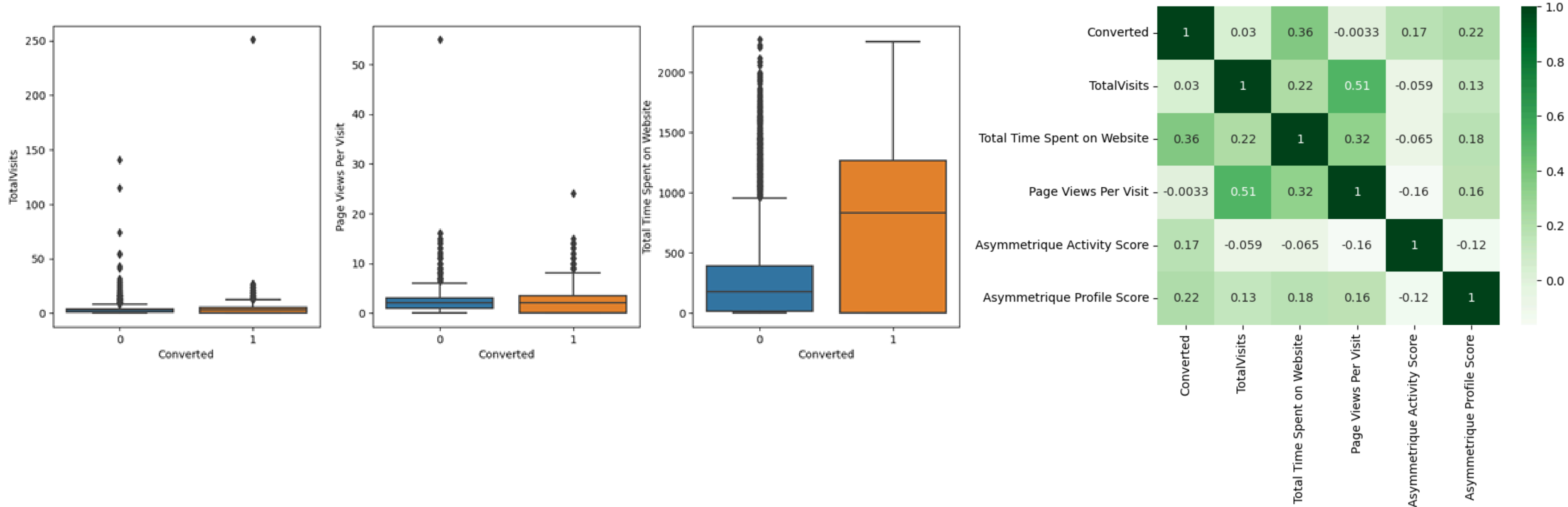
The bar graph displays total visits, as well as the mean and median of current occupations.



Exploratory Data Analysis

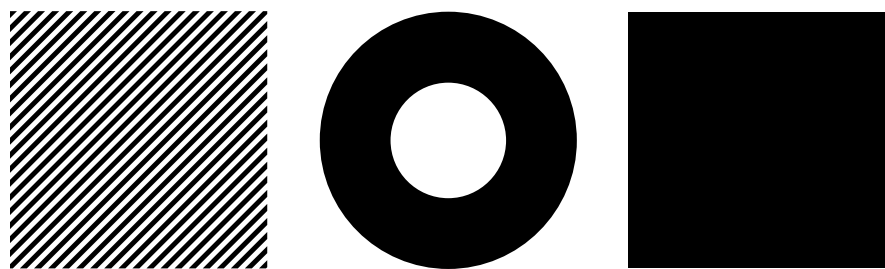
The bar graph displays the total time spent on the website, along with the mean and median of the domains they worked in previously.

Exploratory Data Analysis-Bivariate Analysis for Numerical Variables

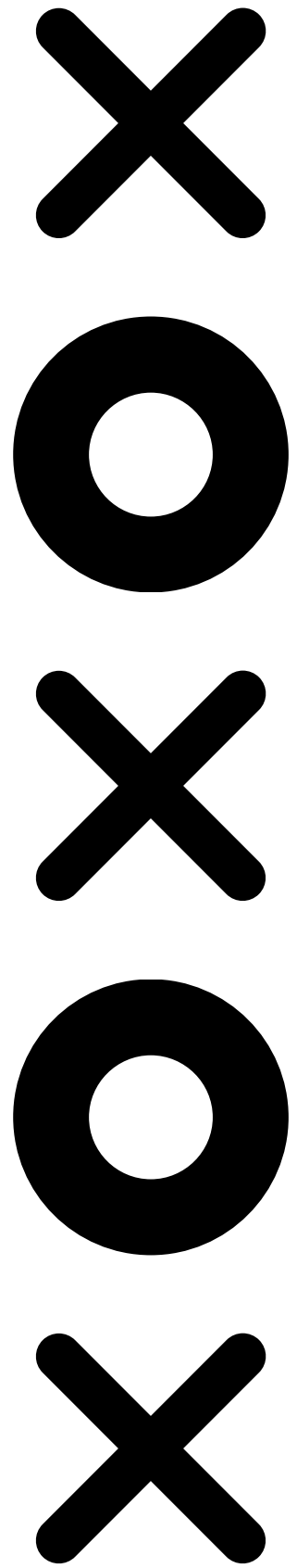


- Past leads who spend more time on the website demonstrate a higher likelihood of successful conversion compared to those who spend less time, as illustrated in the box plot.

Data Preprocessing Before Constructing The Model



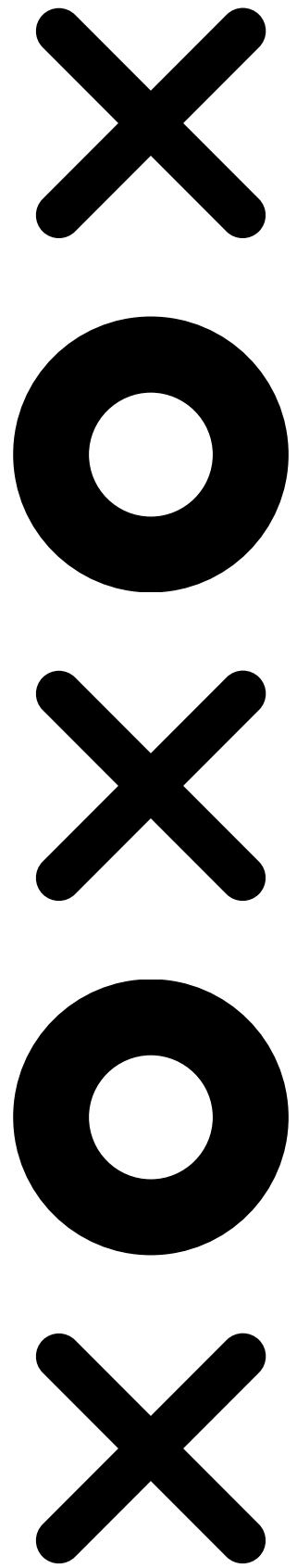
-
- The binary categorical columns have already been encoded as 1s and 0s in earlier stages.
 - I generated dummy features using one-hot encoding for categorical variables such as Lead Origin, Lead Source, Last Activity, Specialization, and Current Occupation.
 - The dataset was divided into a 70:30 ratio for the train and test sets.
 - The features were scaled using the standardization method.
 - The predictor variables highly correlated with each other (Lead Origin_Lead Import and Lead Origin_Lead Add Form) were excluded from the analysis.



Constructing The Model

Selecting the Features

- The dataset is characterized by high dimensionality and a large number of features.
- This could potentially decrease model performance and lead to increased computation time.
- Therefore, it is crucial to conduct Recursive Feature Elimination (RFE) to selectively choose the most significant columns.
- We can then fine-tune the model manually.
- Outcome of RFE: Before feature selection, there were 48 columns, and after applying RFE, the number reduced to 15 columns.



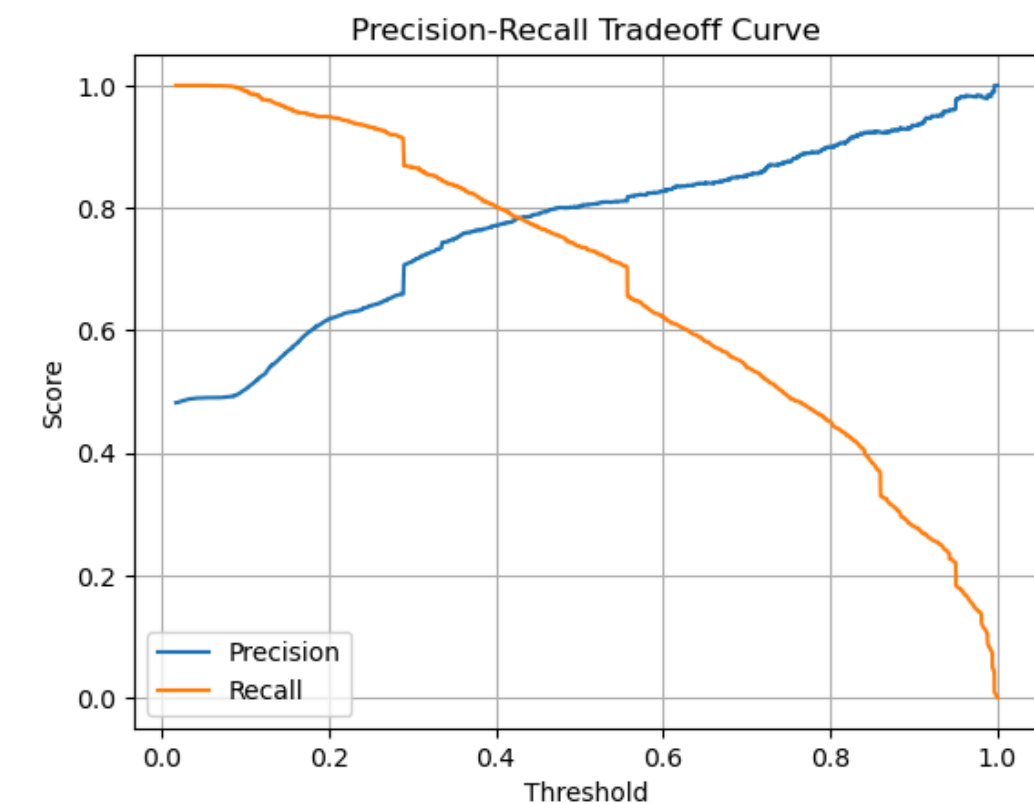
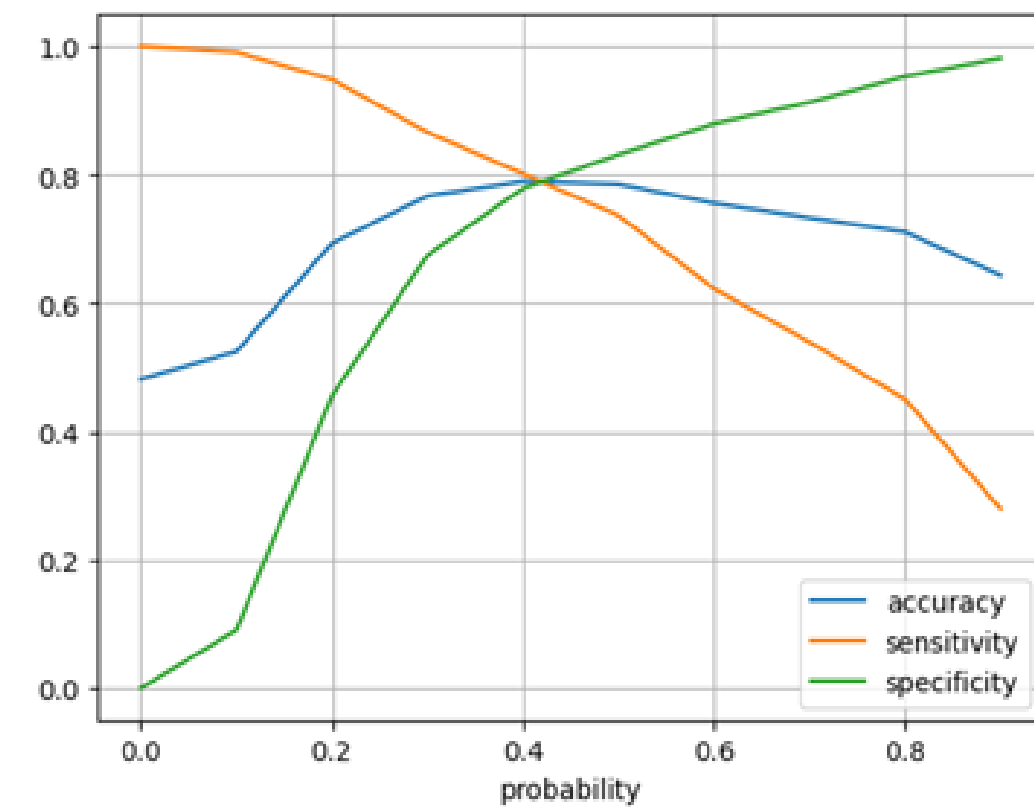
Constructing The Model

-
- The models were constructed using a manual feature reduction approach, where variables with p-values exceeding 0.05 were systematically omitted.
 - **Model 4 appears stable after four iterations:** The results indicate significant p-values below the threshold ($p < 0.05$) and no evidence of multicollinearity, as all variance inflation factors (VIFs) are less than 5.
 - Therefore, logm4 will serve as our final model, utilized for Model Evaluation to make subsequent predictions.

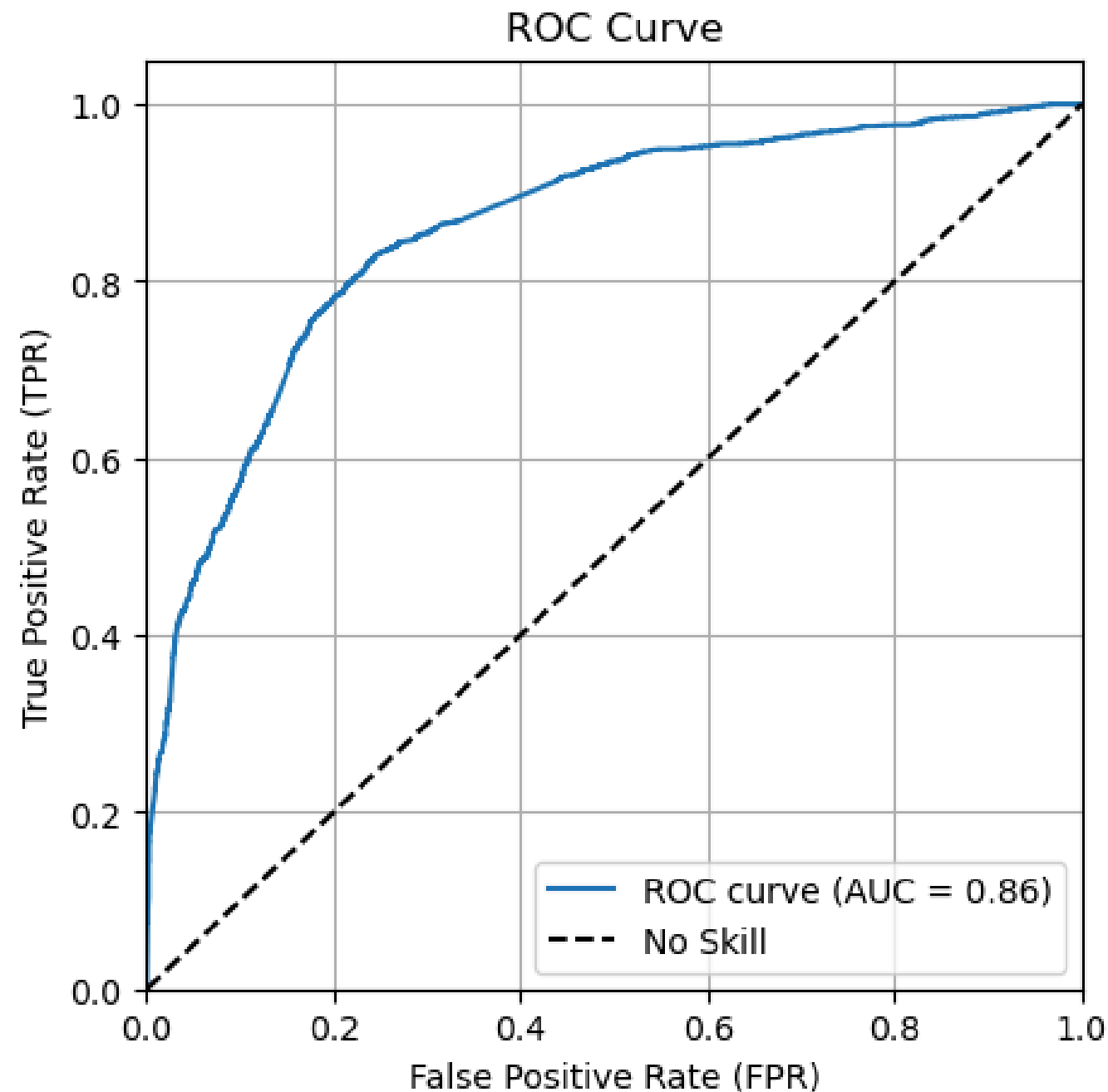
Assessing the performance of a model

Train Data Set

After evaluating metrics from both plots, it was decided to proceed with a cutoff of 0.42.



Assessing the performance of a model



ROC Curve

- The area under the ROC curve is 0.86, suggesting that the predictive model performs well.
- The curve closely hugs the top left corner of the plot, indicating a model with consistently high true positive rates and low false positive rates across all threshold values.

Assessing the performance of a model

Confusion Matrix & Metrics

- With a cutoff value set at 0.42, the model attained a sensitivity of 78.68% on the training set and 78.27% on the test set.
- Sensitivity in this context refers to the model's ability to correctly identify converting leads out of all potential leads.
- The CEO of X Education set a target sensitivity goal of approximately 80%.
- The model also attained an 79.02% accuracy rate, aligning closely with the study's goals.

Suggestions Based On Final Model



-
- Improving lead conversion is paramount for the growth and success of X Education. To achieve this goal, we have developed a regression model designed to pinpoint the most influential factors affecting lead conversion.
 - We have identified key features with the highest positive coefficients that should be prioritized in our marketing and sales efforts to maximize lead conversion:
 - Lead Source_Welingak Website (Coefficient: 5.39)
 - Lead Source_Reference (Coefficient: 2.93)
 - Current_occupation_Working Professional (Coefficient: 2.67)
 - Last Activity_SMS Sent (Coefficient: 2.05)
 - Last Activity_Others (Coefficient: 1.25)
 - Total Time Spent on Website (Coefficient: 1.05)
 - Last Activity_Email Opened (Coefficient: 0.94)
 - Lead Source_Olark Chat (Coefficient: 0.91)
 - We have also pinpointed features with negative coefficients that highlight potential areas for improvement, such as:
 - Specialization in Hospitality Management (-1.09)
 - Specialization in Other Fields (-1.20)
 - Lead Origin: Landing Page Submission (-1.26)

Suggestions Based On Final Model



To enhance our lead conversion rates

- Highlight features with positive coefficients for targeted marketing strategies.
- Formulate effective strategies to attract premium leads from top-performing sources.
- Optimize communication channels based on their impact on lead engagement.
- Craft targeted messages to resonate with working professionals.
- More budget could be allocated to the Welingak Website for advertising and related expenditures.
- Offering incentives or discounts for successful referrals that convert into leads encourages individuals to provide more references.
- Targeting working professionals aggressively is strategic due to their high conversion rates and better financial capacity to afford higher fees.

To pinpoint areas for improvement

- Examine the implications of negative coefficients in specialization offerings.
- Please review the landing page submission process for potential areas of improvement.

