Lead Scoring Case Study

Presentation

Business Objective

- The primary objective is to enhance the rate at which potential leads turn into paying customers from 30% to a goal of 80%.
- Develop a scoring system (0 to 100) to help focus on leads more likely to become customers.
- Create a flexible model that can handle changes and new challenges the company may face.
- Guide the sales team to concentrate on leads with higher scores, making their efforts more effective.
- Provide solutions for additional problems the company is facing based on insights from the model, ensuring a comprehensive approach.

Approach

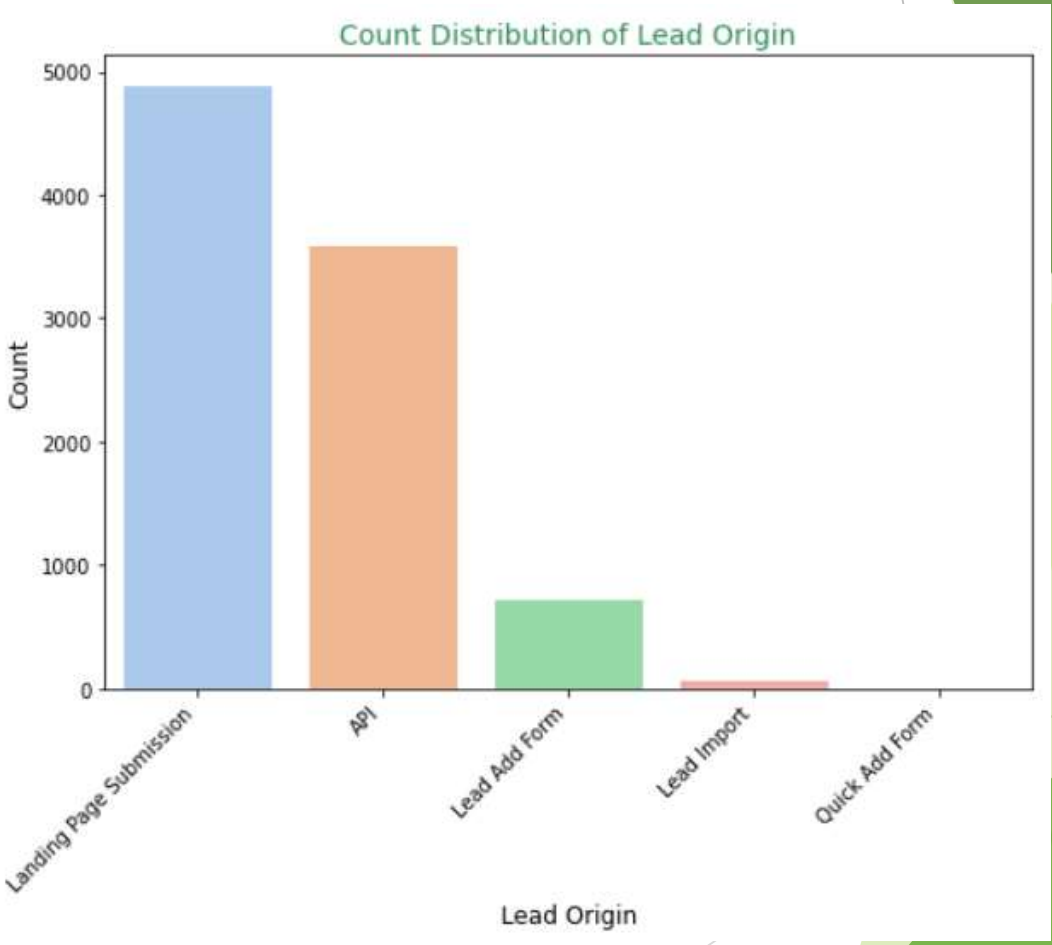
- Data Understanding
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Model Building and Evaluation
- Making Predictions on Test Data Set

Data Cleaning

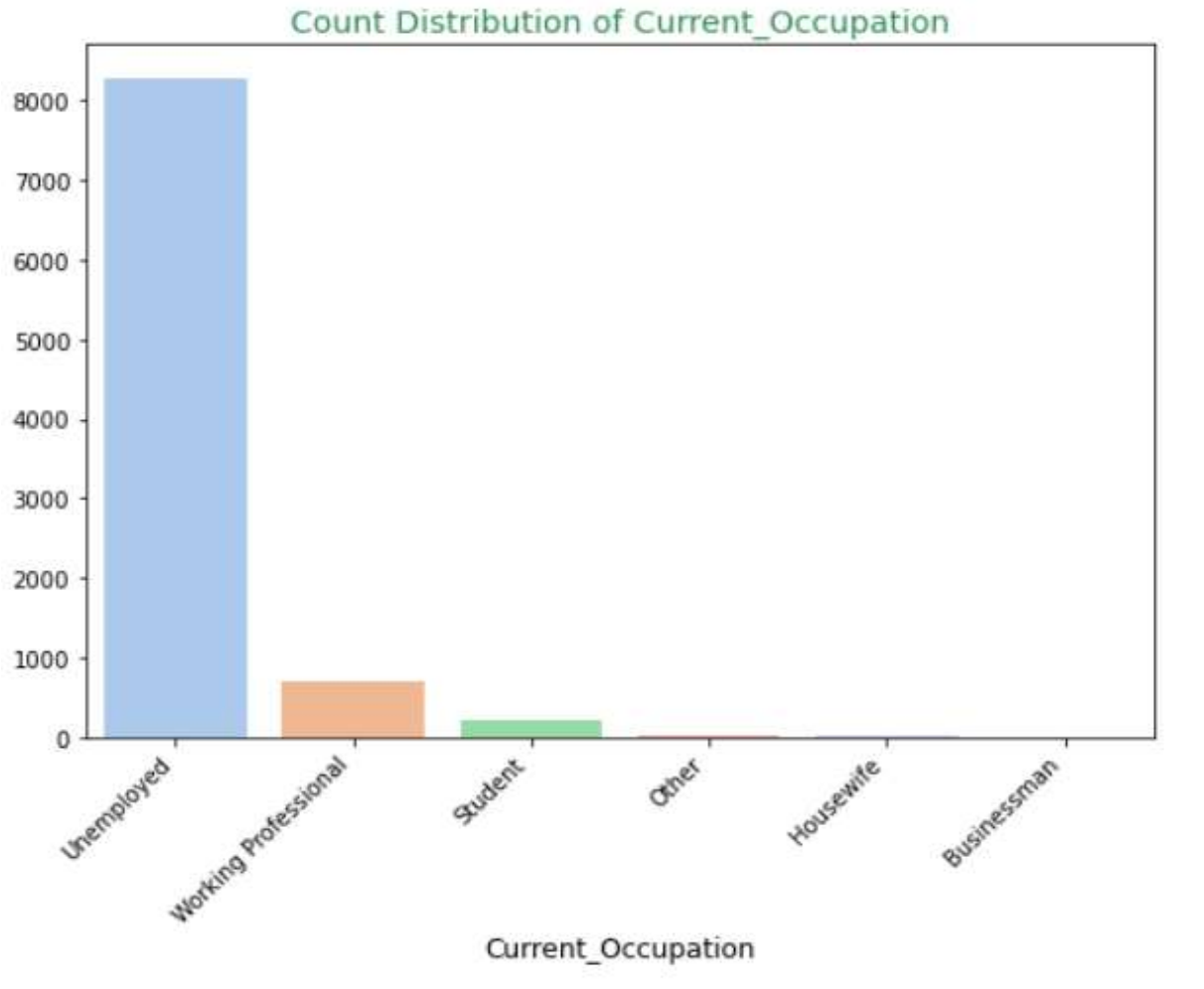
- Setting a threshold of 40% to eliminate columns with null values
- Filling Missing values with Median or Mode
- Missing values in categorical columns are handled based on value counts
- Columns that are of no use for modeling are dropped
- Skewed category columns are checked and dropped to avoid biasing
- Low-frequency values are grouped as “Others”

EDA – Univariate Analysis

Most of the customers were identified through ‘Landing Page Submission’ as the lead origin, followed by API

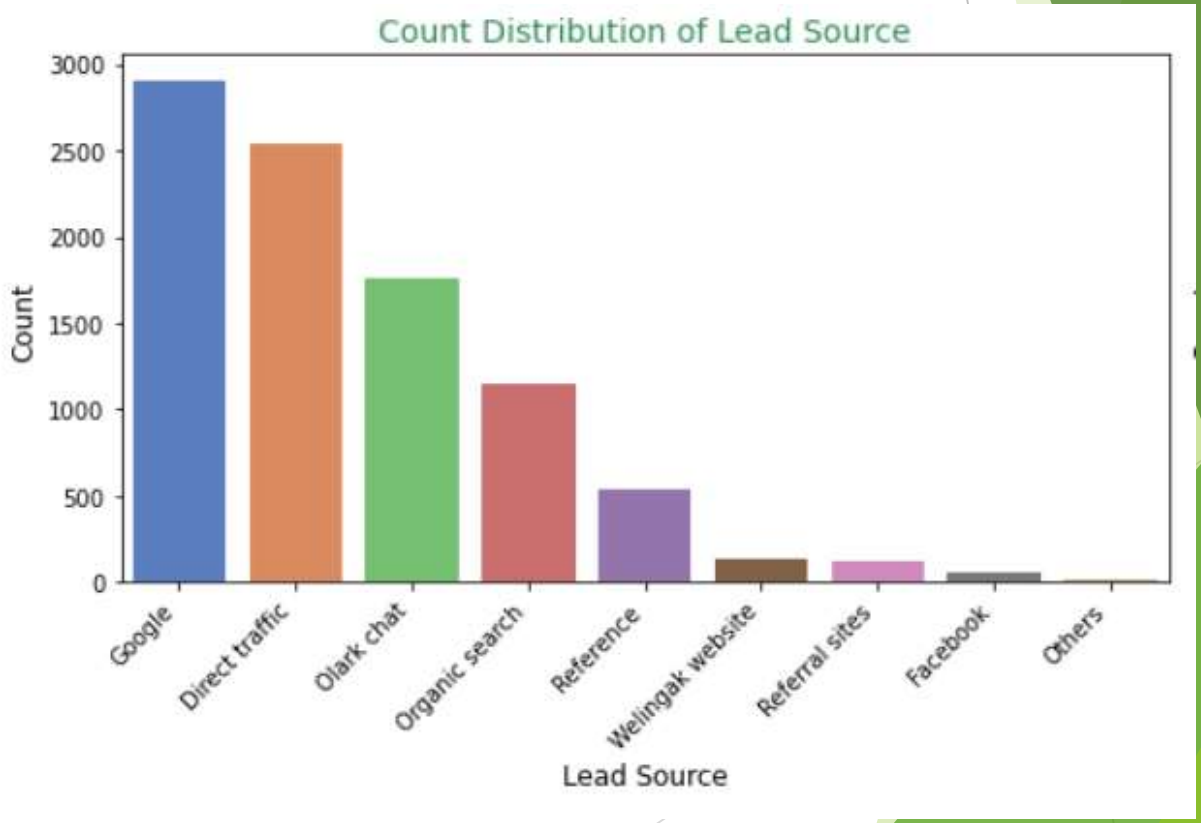


A large percentage of customers are unemployed according to the current occupation information

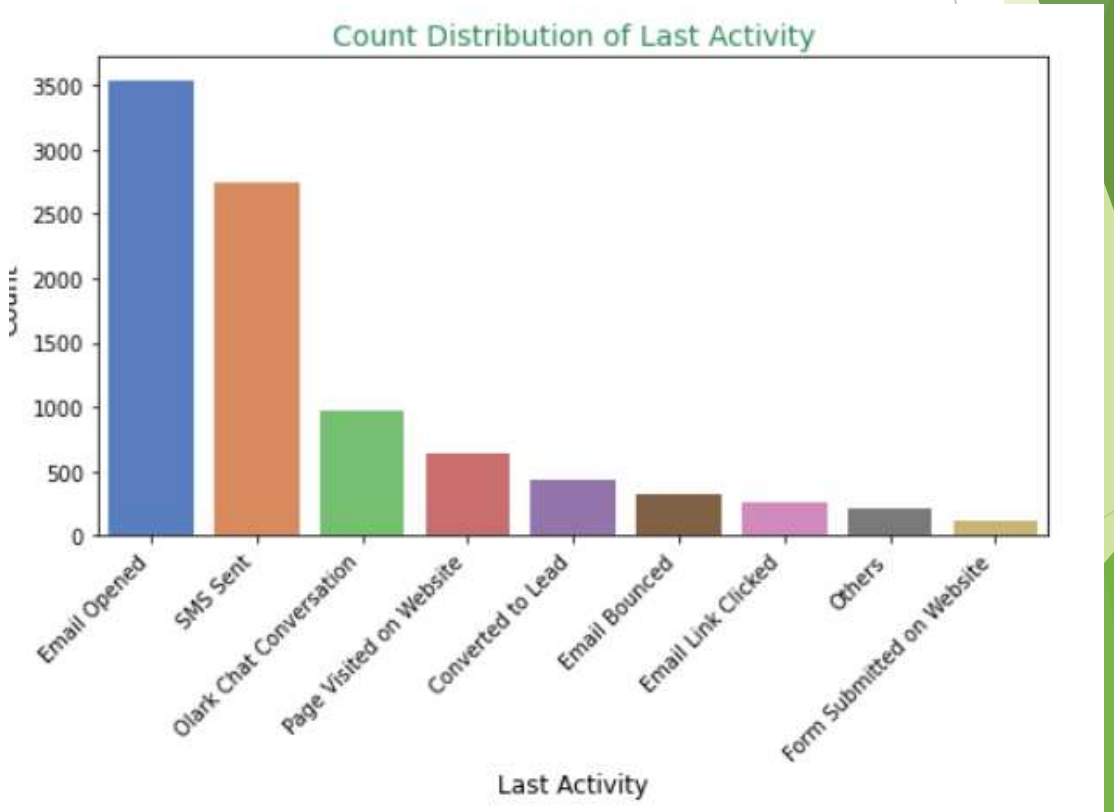


The primary lead source is **Google**

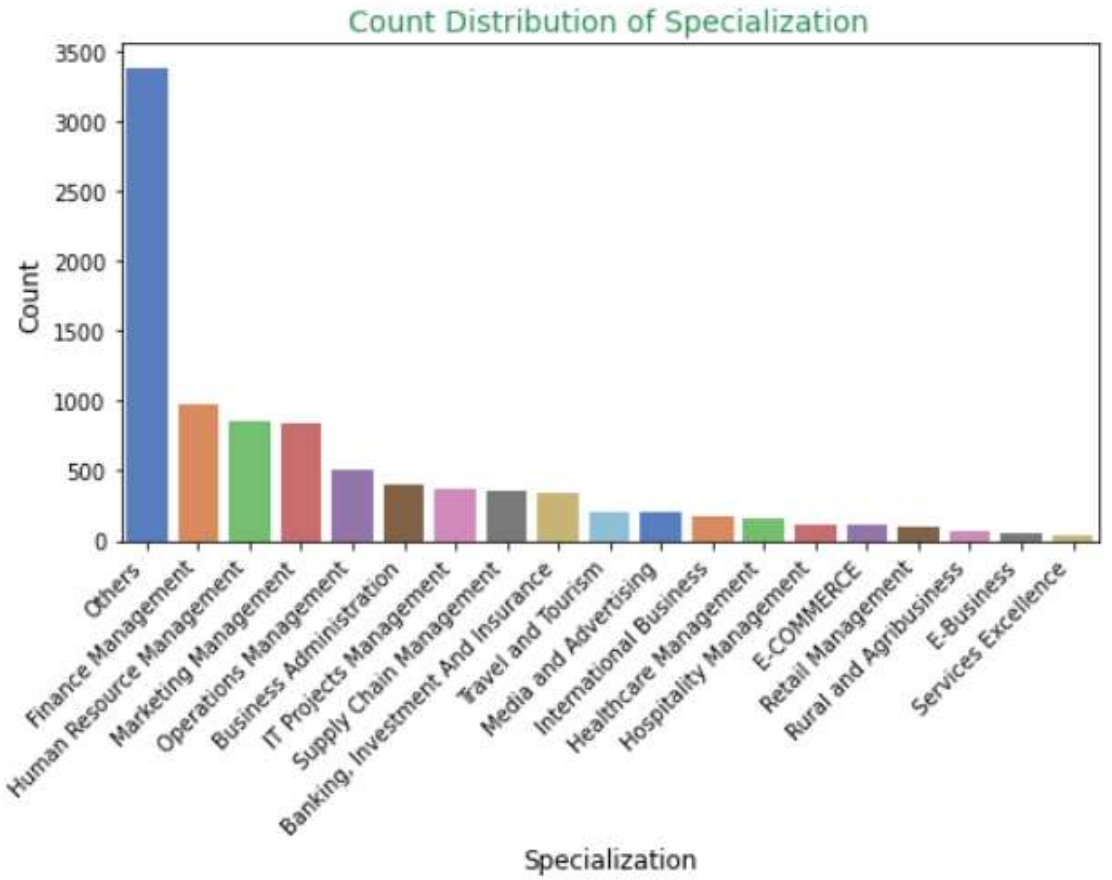
followed by **Direct Traffic**



Email is the most common last activity, with most of customers having opened an email followed by customers having sent an SMS.

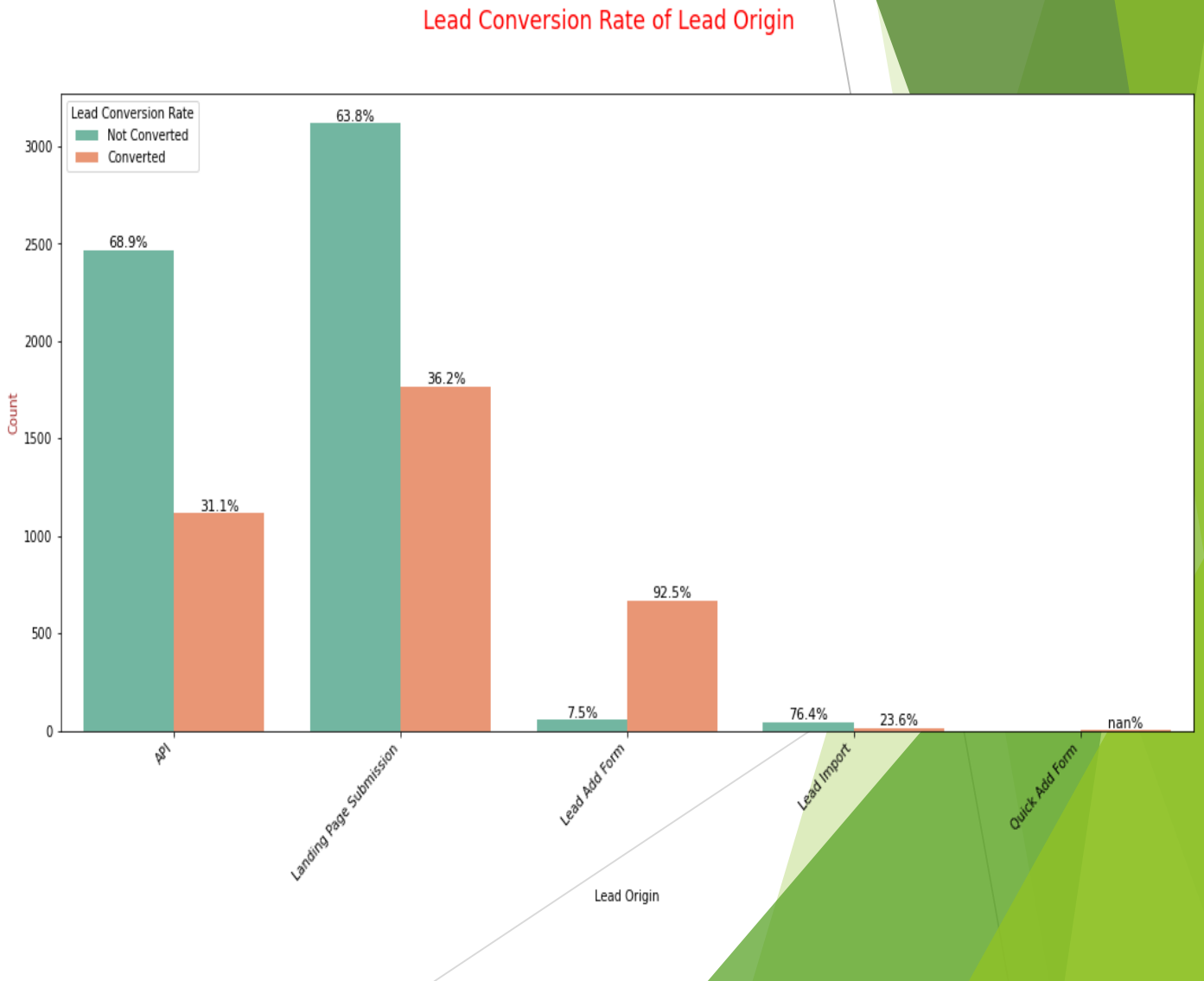


The ‘Others’ specialization category is the most common among customers, followed by Finance Management, HR Management, Marketing Management, and Operations Management

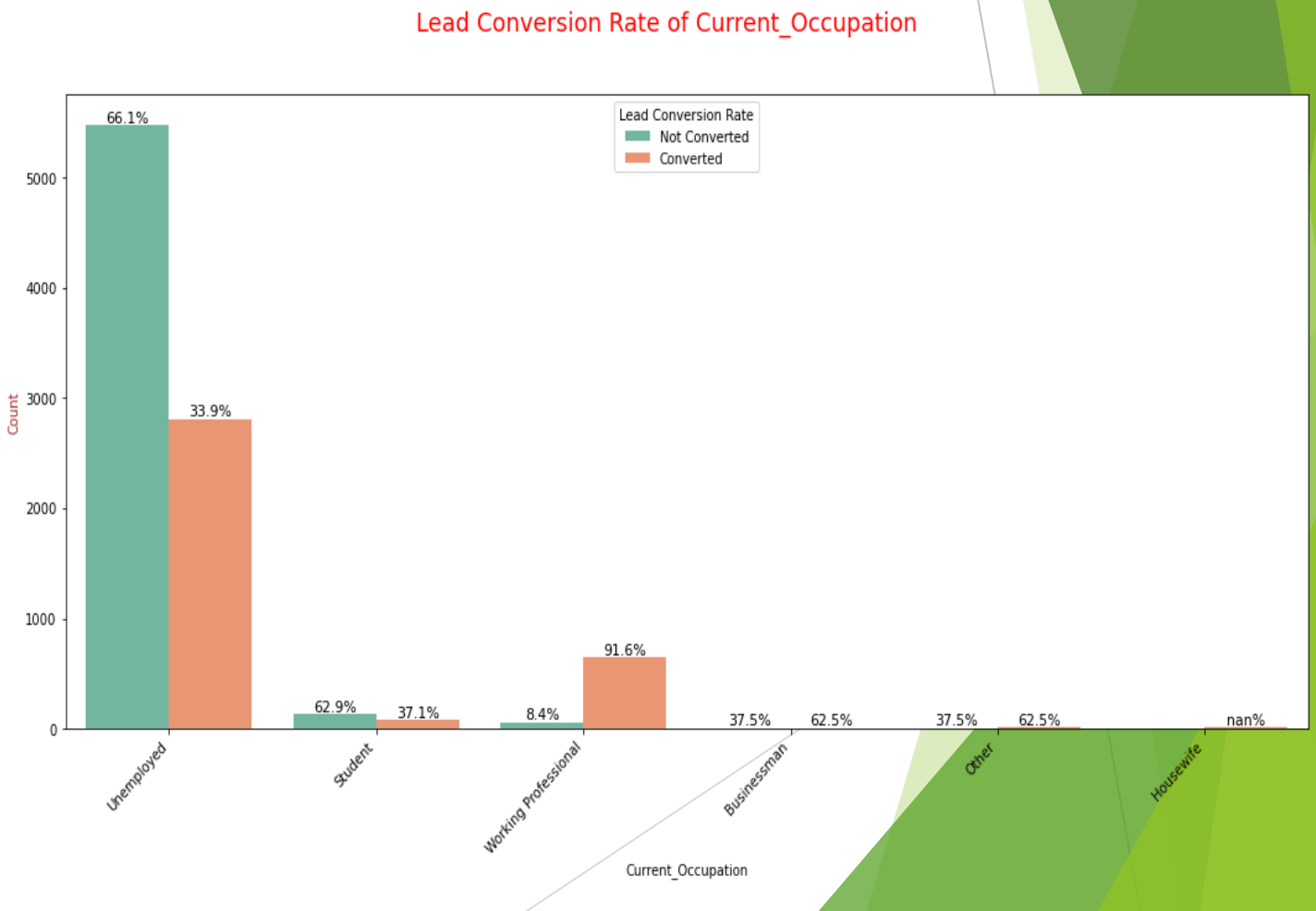


EDA – Bivariate Analysis

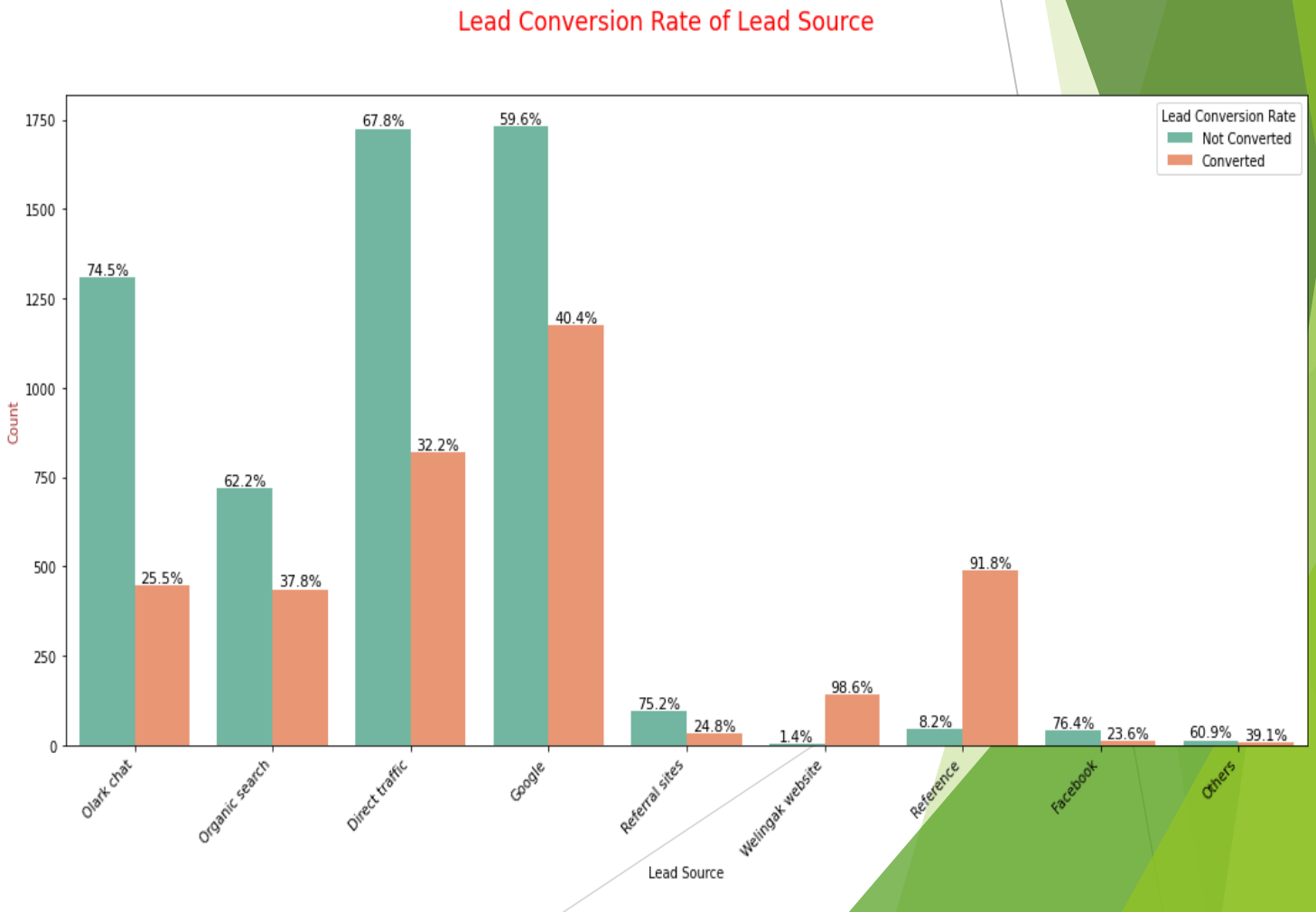
The most effective lead origin is 'Landing Page Submission' with a Lead Conversion Rate (LCR) of 36.2%, followed by 'API' at 31.1%.



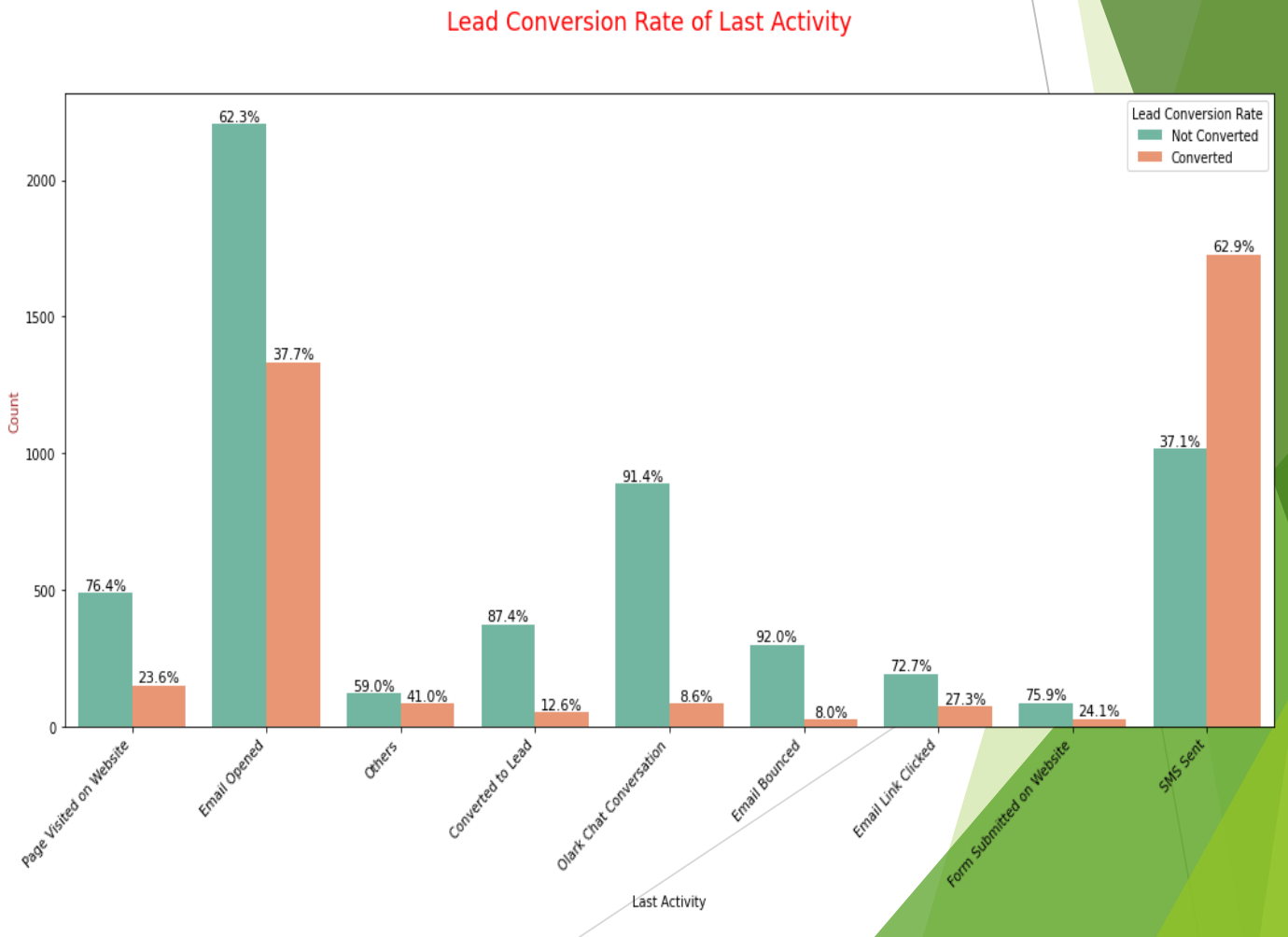
Working professionals exhibit a significantly higher Lead Conversion Rate (LCR) of 91.6% in contrast to unemployed individuals, who have an LCR of 33.9%.



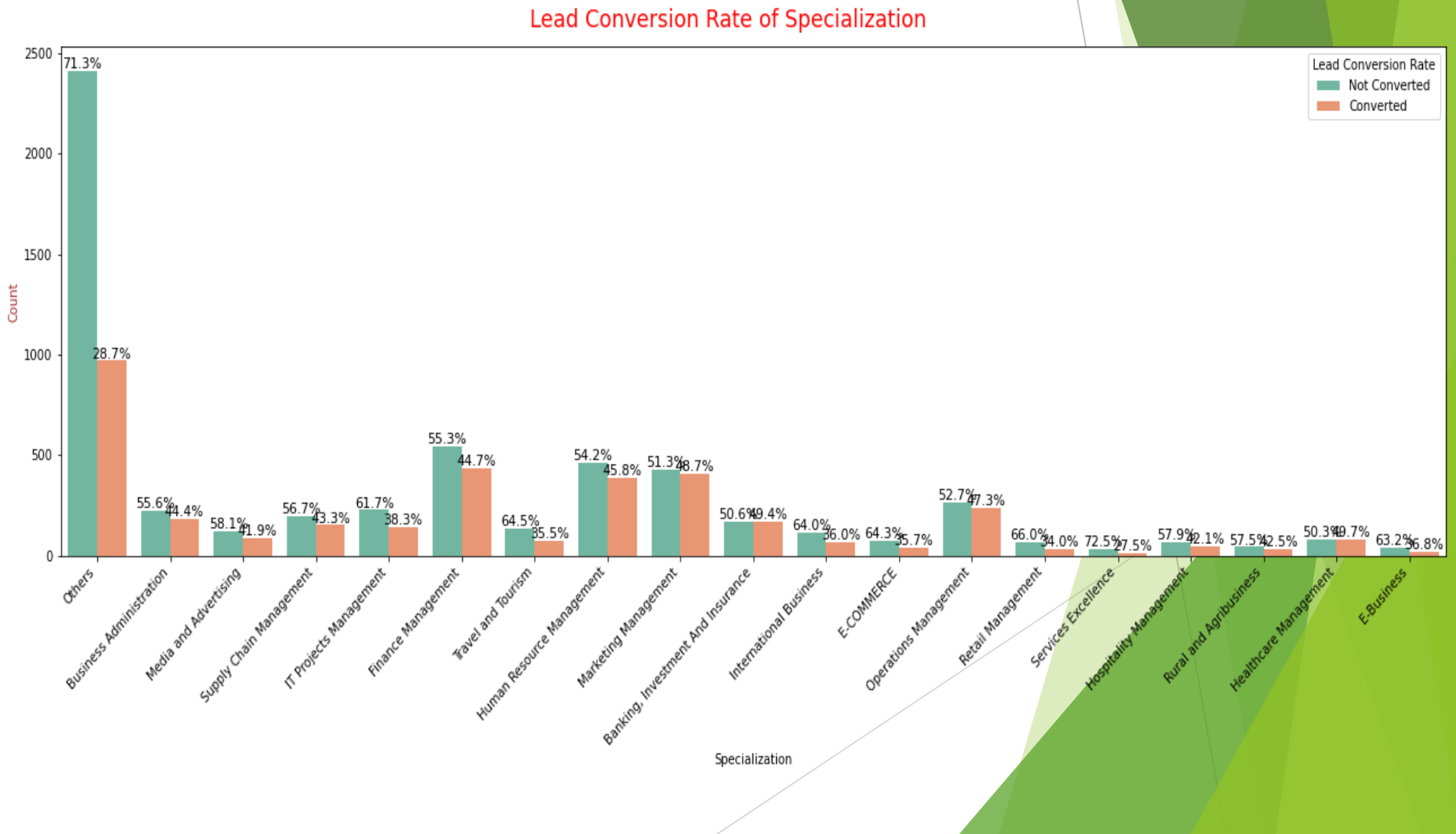
Among the lead sources, Google stands out as the most effective, boasting a Lead Conversion Rate (LCR) of 40.4%. Direct Traffic follows with an LCR of 32.2%, and Organic Search contributes to a smaller portion, holding an LCR of 37.8%. Notably, Reference has the highest LCR at 91.8%, but it constitutes only 5.8% of the total customers through this lead source.



SMS Sent and Email Opened emerge as the most impactful Last Activity types, exhibiting Lead Conversion Rates (LCRs) of 62.9% and 37.7%, respectively.

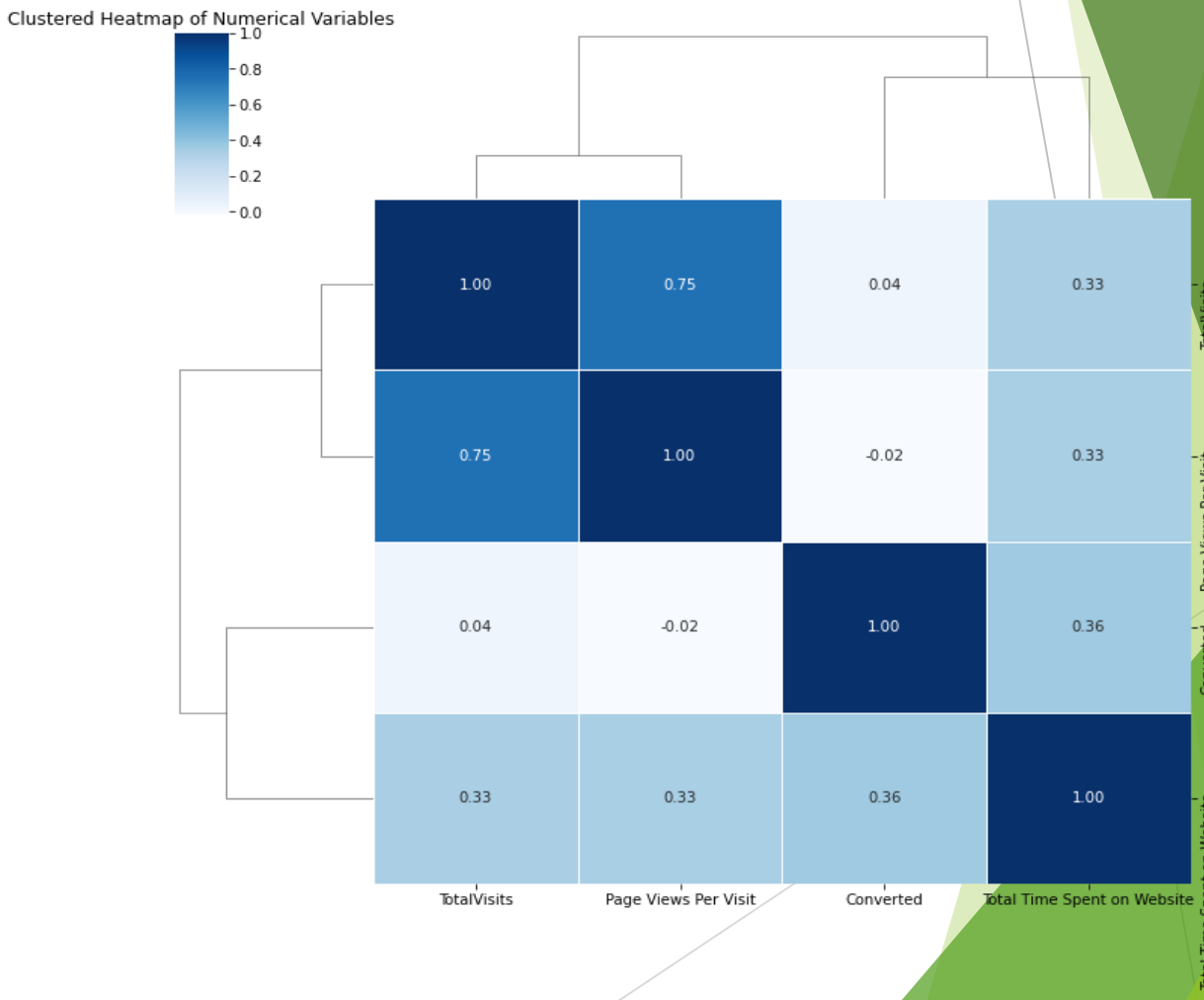


Popular specializations like Marketing, HR, Finance, and Operations Management exhibit strong customer interest with good Lead Conversion Rates.



Correlation Analysis

Frequent website visitors tend to view more pages per visit, showing a strong positive correlation between 'Total Visits' and 'Page Views per Visit'.



Data Preparation

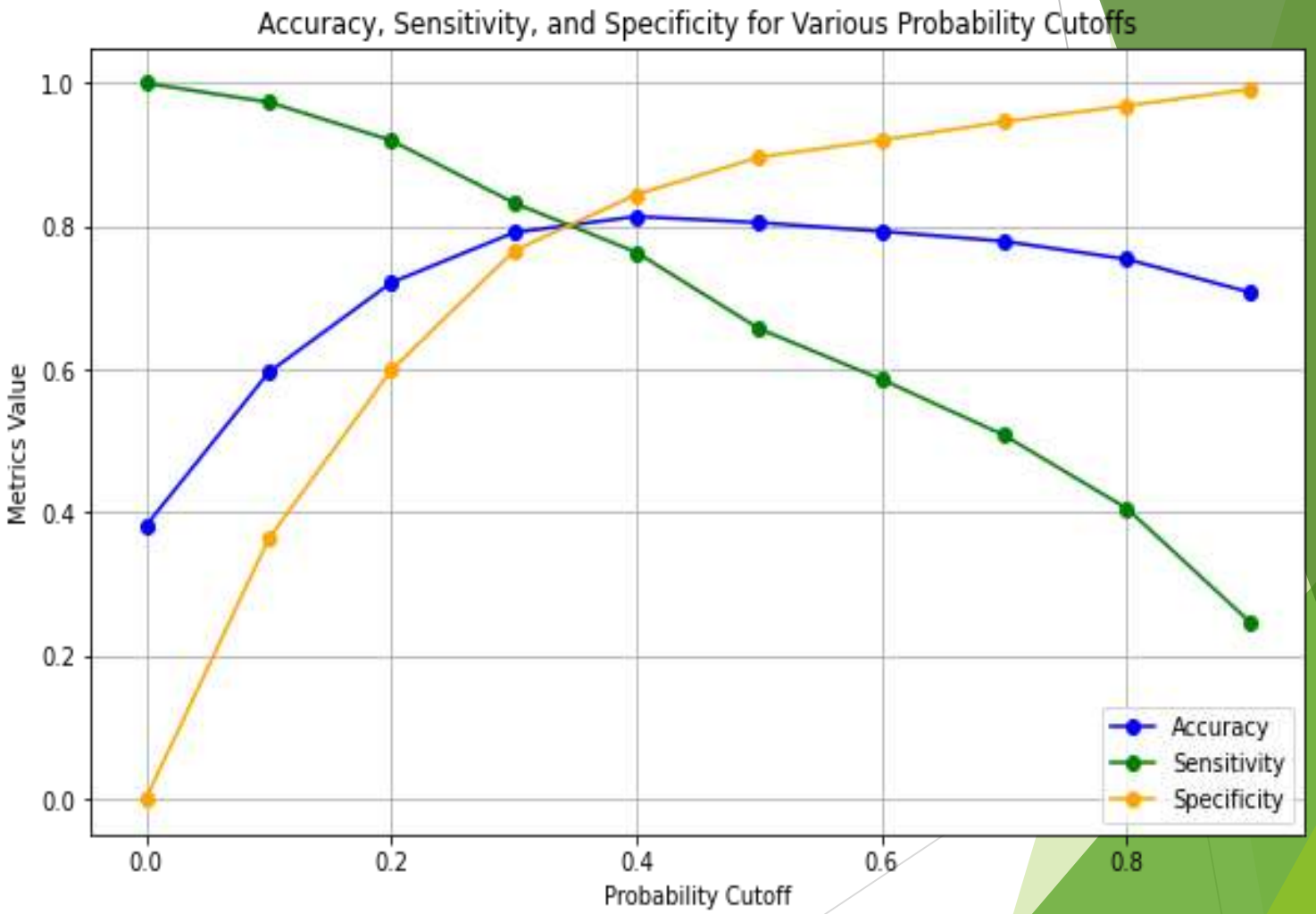
- Binary categorical columns were converted to 1 and 0 for logistic regression compatibility.
- Dummy features were created using one-hot encoding for categorical variables like Lead Origin, Lead Source, Last Activity, Specialization, and Current_Occupation.
- The dataset was split into 70% training and 30% testing for model training and evaluation.
- Standardization was applied for feature scaling to ensure consistent scales.
- Correlated predictor variables were removed to prevent multicollinearity.

Model Building

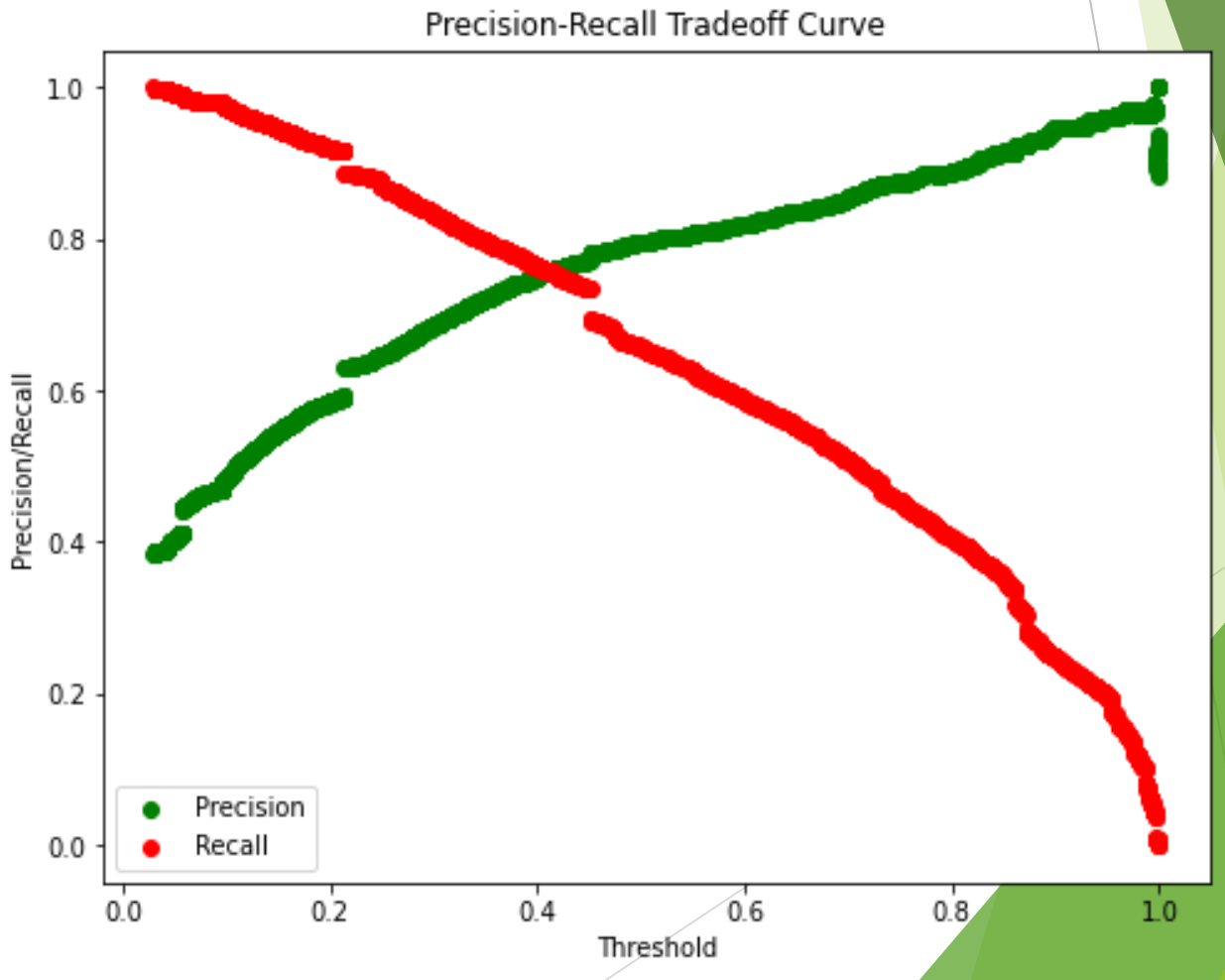
- The dataset initially had numerous features, leading to potential performance issues and longer computation times.
- Recursive Feature Elimination (RFE) was applied to select essential columns.
- Post-RFE, the dataset was reduced from 48 to 15 columns.
- Logistic Regression Model 1 served as a basic model.
- Logistic Regression Models 2 and 3 underwent manual feature reduction by removing variables with p-values above 0.05.
- Logistic Regression Model 4, after four iterations, achieved stability with significant p-values (≤ 0.05) and no multicollinearity (VIFs < 5).
- Logistic Regression Model 4 (LRMod 4) emerged as the final model for evaluation and predictions.

Model Evaluation

Analyzing the curve, it is recommended to set the cutoff probability at around 0.35 for optimal classification.

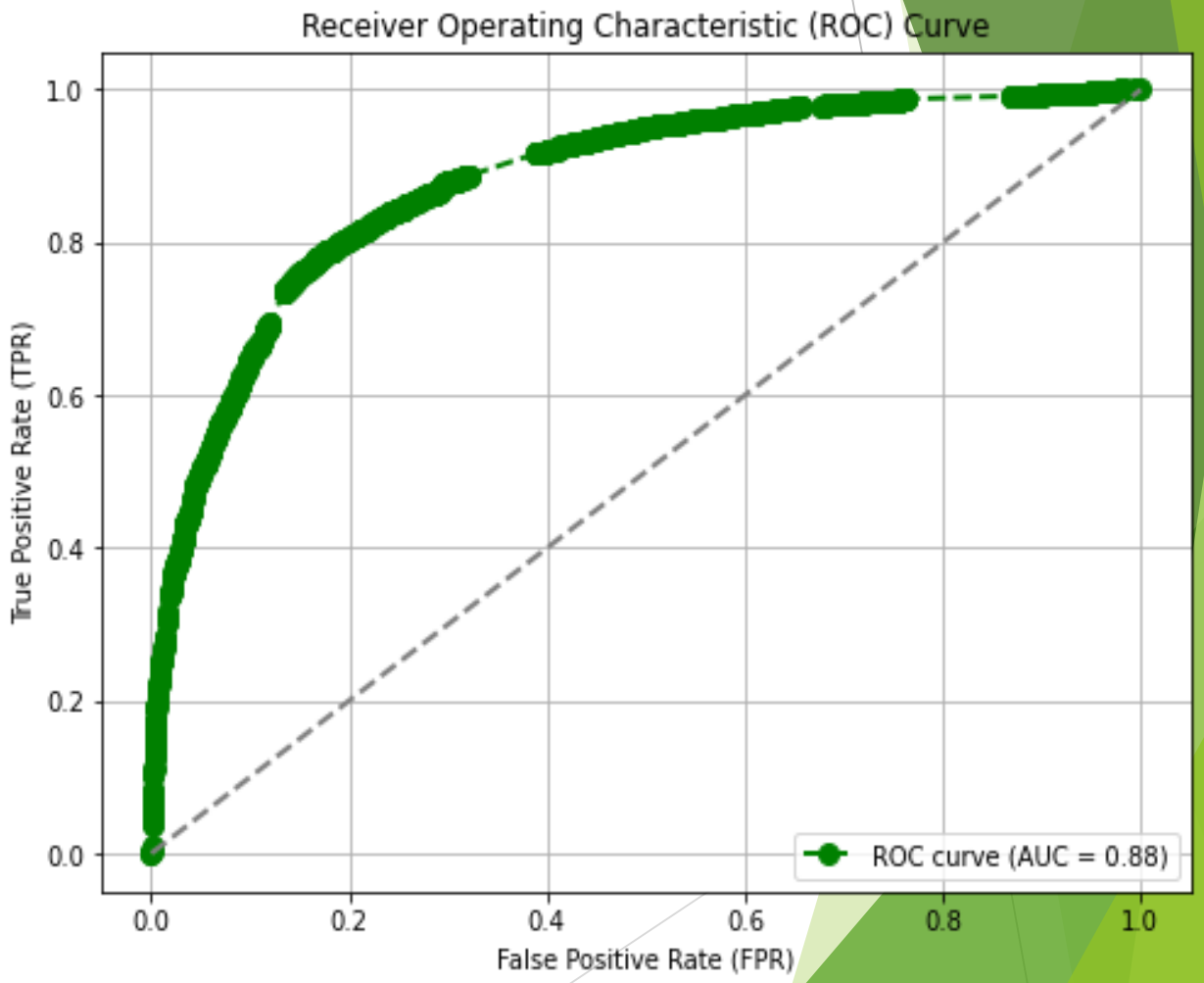


The precision-recall curve suggests that a threshold of 0.4 achieves a favorable equilibrium between precision and recall



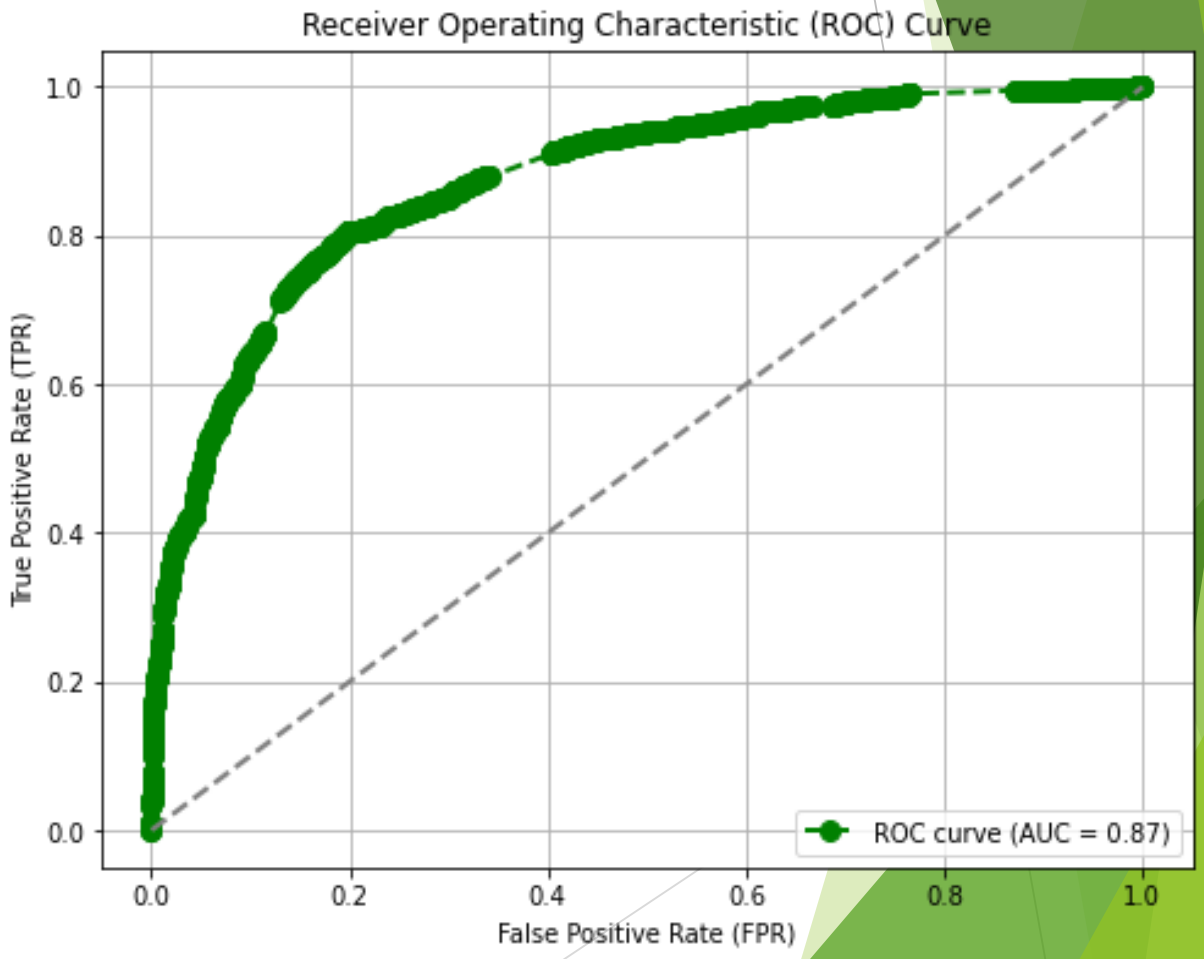
Making Predictions – ROC Curve Train Data Set

The model's Area under ROC curve is 0.88, signifying strong predictive ability, as evidenced by the curve positioned close to the top-left corner, indicating high true positive and low false positive rates across thresholds.



Making Predictions – ROC Curve Test Data Set

The model's Area under ROC curve is 0.87, demonstrating strong predictive capability, with the curve positioned close to the top-left corner, indicating high true positive and low false positive rates across thresholds.



Conclusion

- Effective factors for a good conversion rate include 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional', and 'Total Time Spent'.
- Higher conversion rates are observed among working professionals and unemployed customers.
- Referral leads from old customers show a significantly higher conversion rate.
- Promising conversion rates are observed for leads from Google and Direct Traffic channels.
- Leads with 'Last Activity' as 'SMS Sent' or 'Email Opened' tend to have higher conversion rates.
- The most common specialization category is 'Others', followed by Finance Management, HR Management, and Marketing Management.

Recommendations

- Utilize features with high conversion rates, such as 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional', and 'Total Time Spent on Website', for effective lead generation.
- Aggressively target working professionals for higher conversion, considering their better financial situations.
- Incentivize referral leads from old customers with discounts or rewards to boost the conversion rate.
- Increase media usage frequency, like Google ads or email campaigns, to save time and improve conversion rates.
- Focus on leads with 'Last Activity' as 'SMS Sent' or 'Email Opened' for more targeted efforts and higher conversion rates.
- Enhance the user experience by analyzing the behavior of customers spending more time on the website, and create engaging content and user-friendly navigation.
- Tailor course offerings and marketing campaigns based on popular specializations like Marketing Management and HR Management to attract and retain customers in those fields.

Thank You