



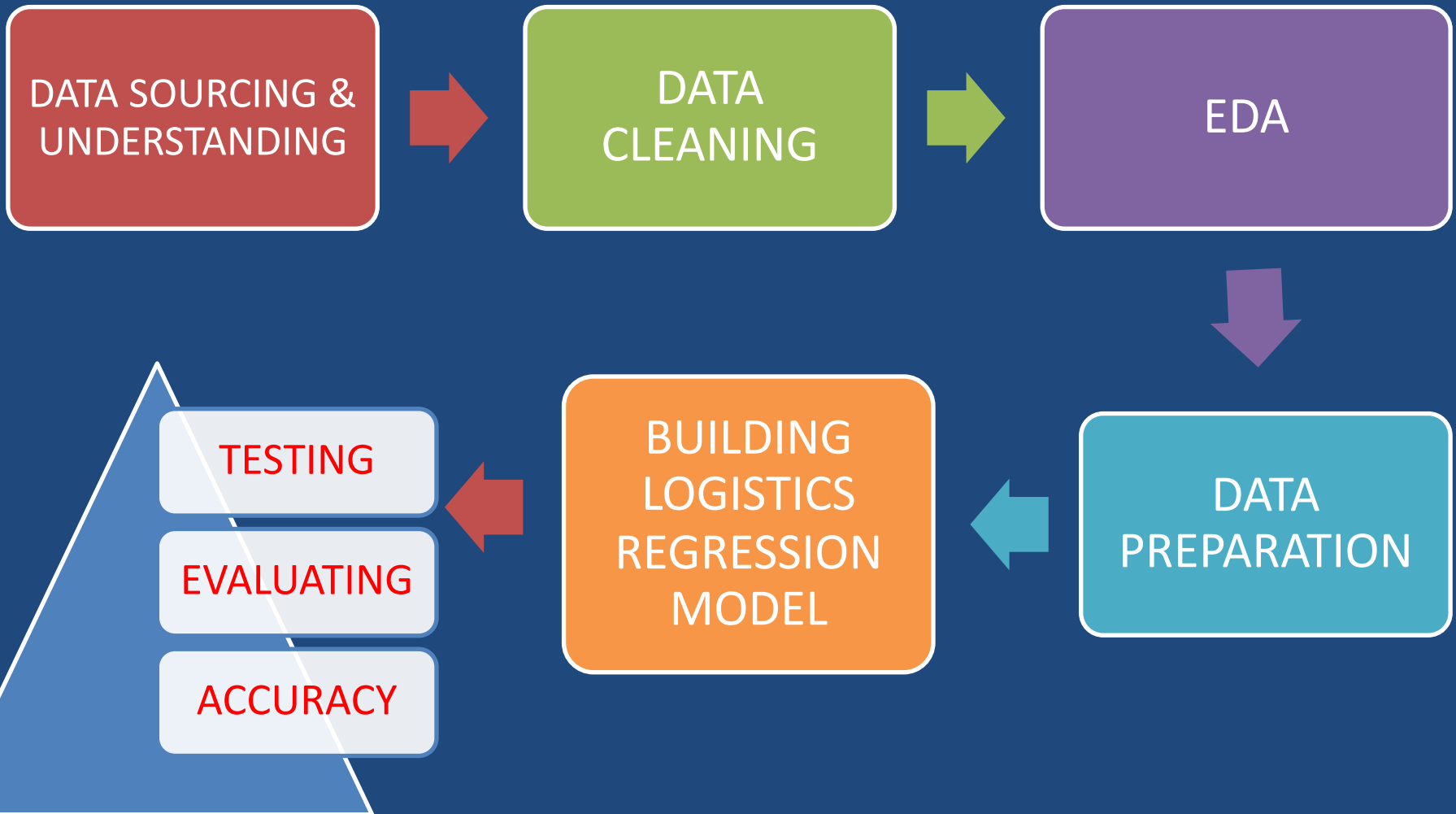
# LEAD SCORE CASE STUDY



# PROBLEM STATEMENT

- To understand most promising Leads
- To build a model which identifies Hot Leads
- To assist in company in enhancing Lead conversion rate

# APPROACH



# DATA CLEANING

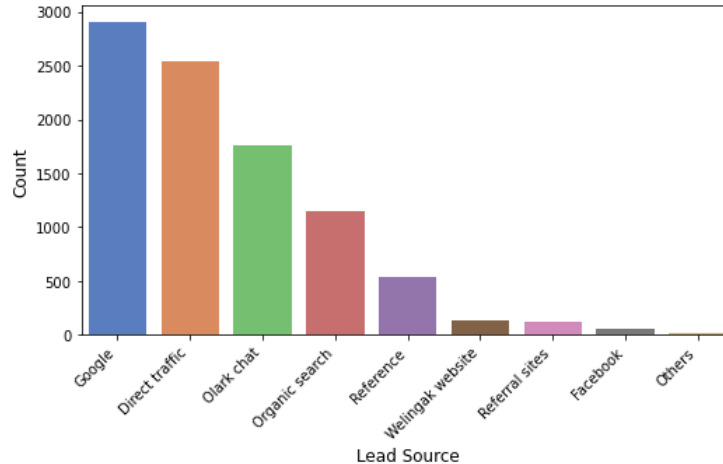
- DROPPING COLUMNS WITH MORE THAN 40% MISSING VALUES
  - 'How did you hear about X Education'
  - Lead Quality'
  - 'Lead Profile'
  - Asymmetrique Activity Index'
  - Asymmetrique Profile Index'
  - Asymmetrique Activity Score'
  - Asymmetrique Profile Score'
  
- DROPPING COLUMNS WITH HIGHLY SKEWED DATA
  - ('Do Not Call,' 'Search,' 'Newspaper Article,' 'X Education Forums,' 'Newspaper,' 'Digital Advertisement,' and 'Through Recommendations,' as they may introduce bias in logistic regression models, impacting their performance.**

# DATA CLEANING : COLUMN WISE

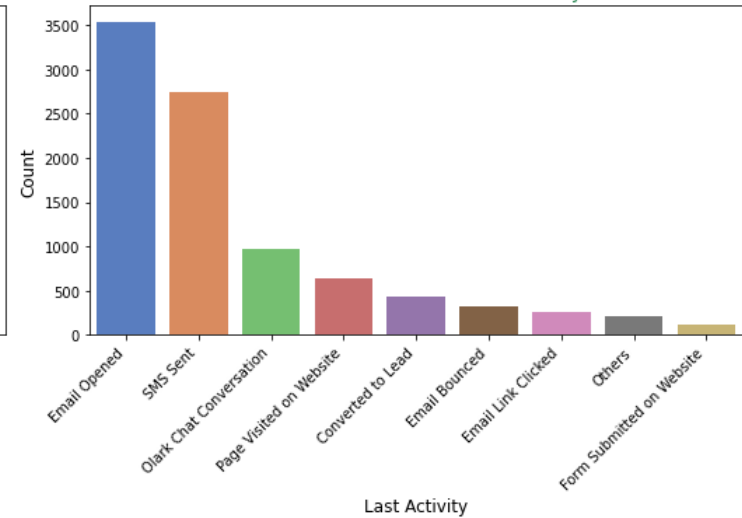
- **City:** Drop the column due to 39.71% missing values; imputing with Mumbai (57.84%) may bias the model.
- **Specialization:** Create an 'Others' category for 36.58% missing values, as specialization is evenly distributed.
- **Tags:** Drop the column with 36.29% missing values, as it won't be useful for modeling.
- **What matters most:** Drop with 29.32% missing values; 99.95% choosing 'better career prospects' is highly skewed and lacks insight.
- **Current Occupation:** Impute missing values with 'Unemployed' (29.11% missing) as it's crucial for X Education targeting unemployed individuals.
- **Country:** Drop due to 96% of customers being from India, causing significant skewness.
- **Last Activity:** Impute missing values (1.11%) with 'Email Opened,' the most frequent activity.
- **Lead Source:** Impute missing values (0.39%) with 'Google,' the most common source.

# EDA

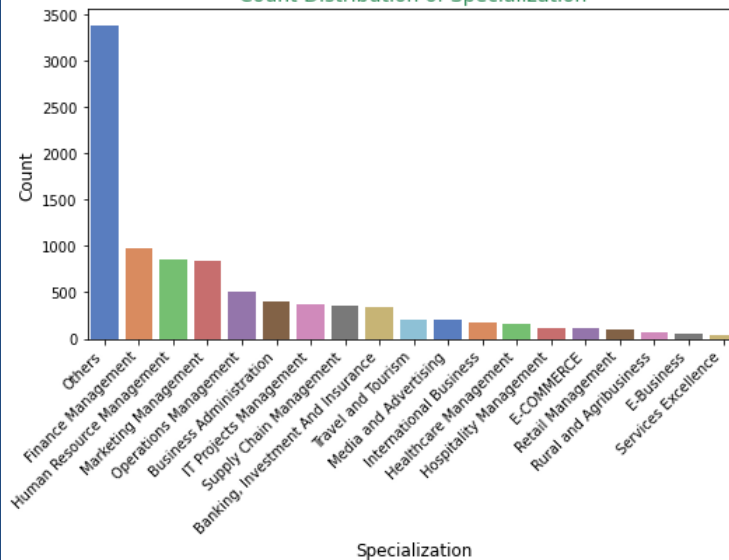
Count Distribution of Lead Source



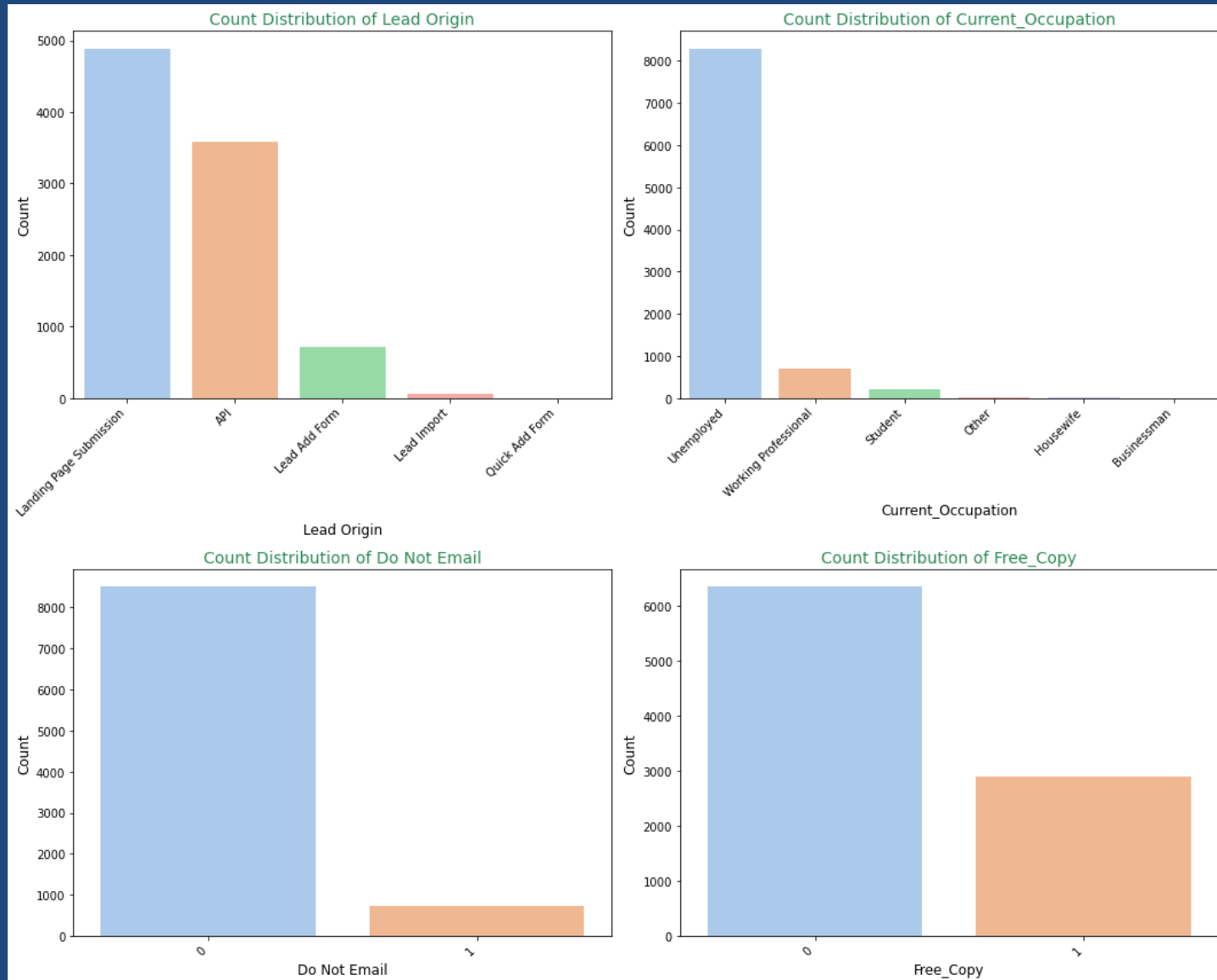
Count Distribution of Last Activity



Count Distribution of Specialization



# EDA



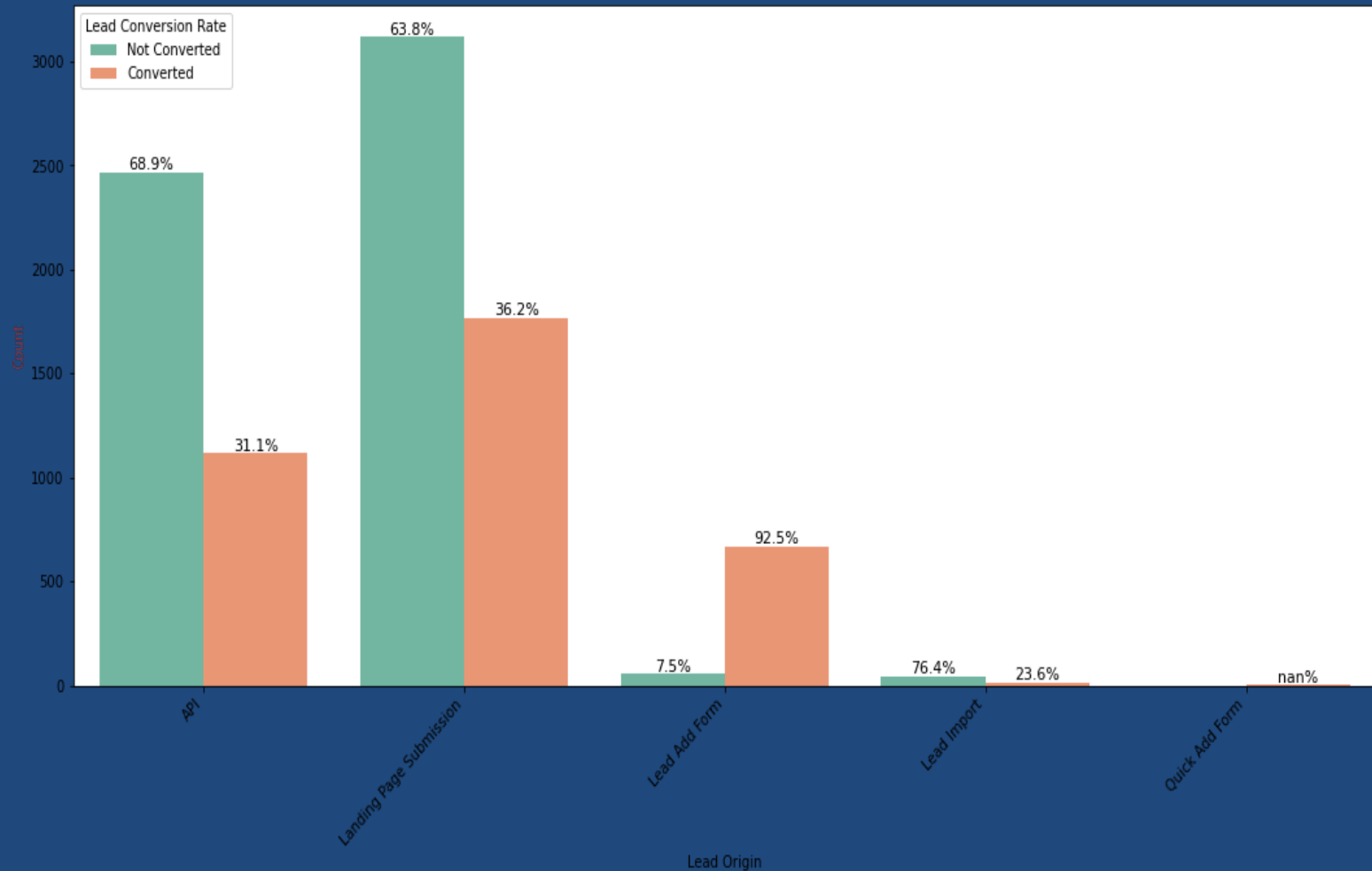
# ANALYSIS

- Most leads, 52.9%, come from 'Landing Page Submission',
- 38.7% are through 'API'
- A significant 89.7% of customers are unemployed.
- The majority, 92.1%, prefer not to receive course-related emails.
- Google is the primary lead source at 31.5%, and Direct Traffic follows at 27.5%.
- For the last activity, 38.3% opened an email, and 29.7% sent an SMS.
- In Specialization, Finance, HR, Marketing, and Operations at 10.6%, 9.2%, 9.1%, and 5.4%, respectively.

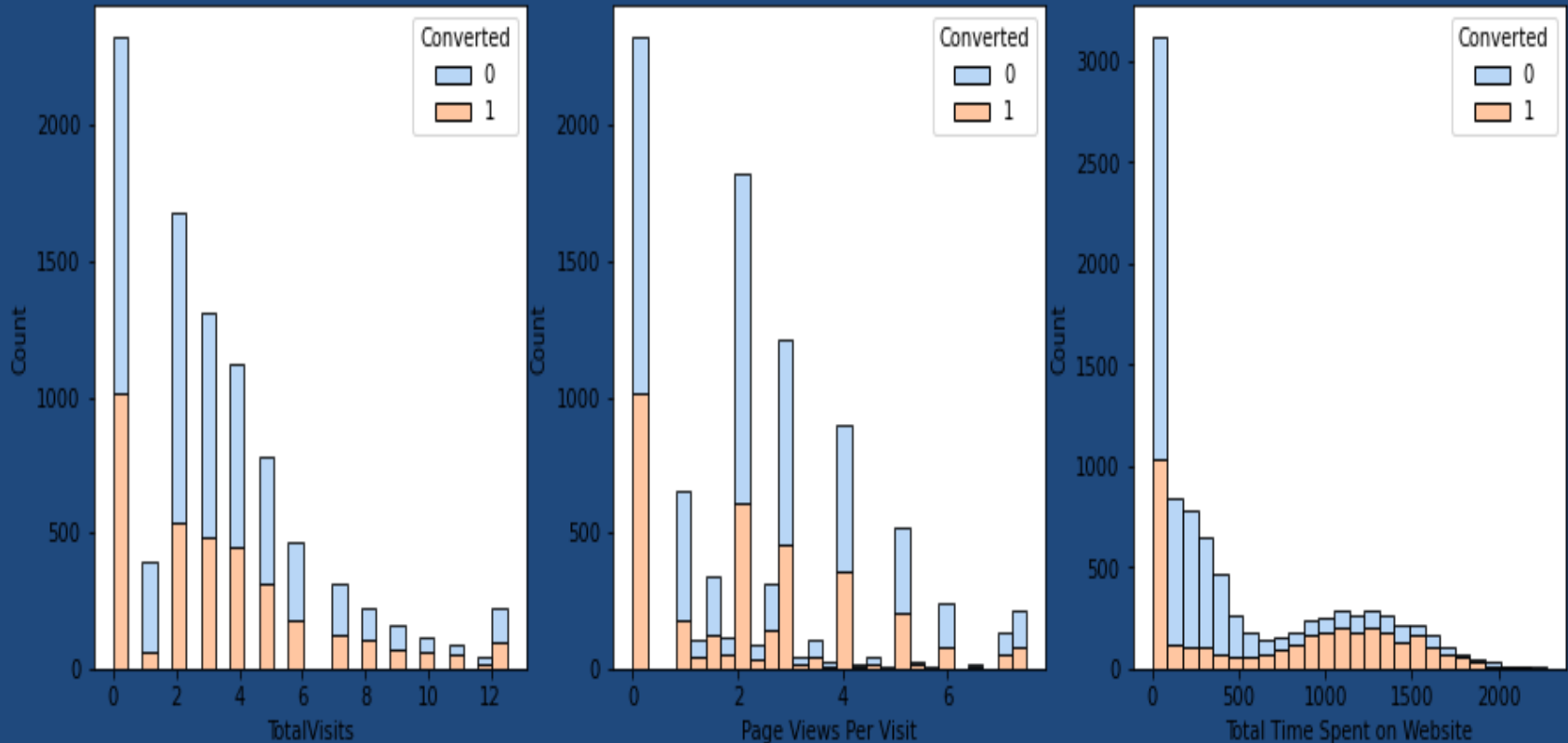


# LEAD CONVERSION RATE

Lead Conversion Rate of Lead Origin

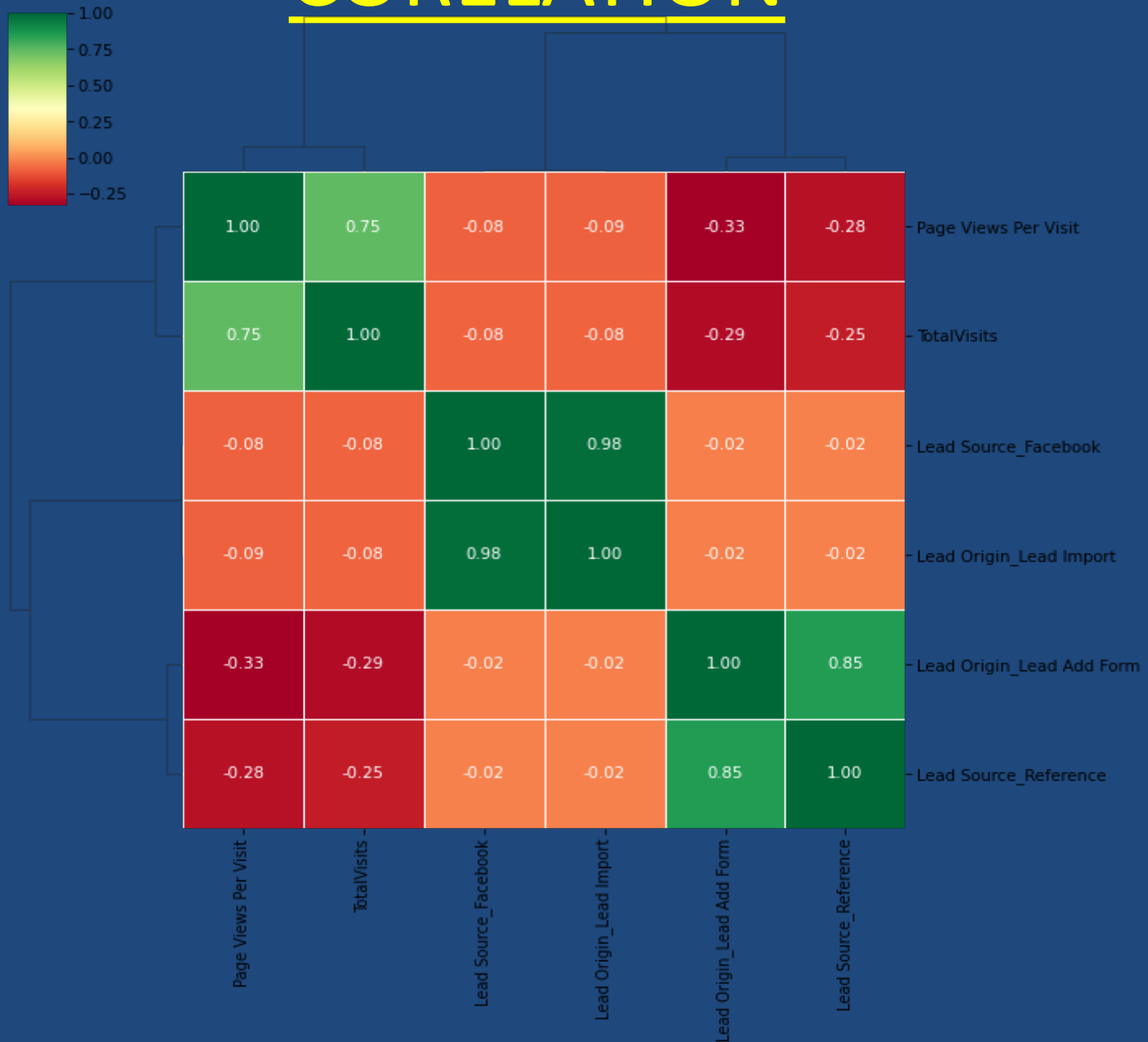


# CONVERSION RATE



**Customers who visit the website more often tend to view more pages per visit, showing a strong positive correlation between these two factors.**

# CORRELATION



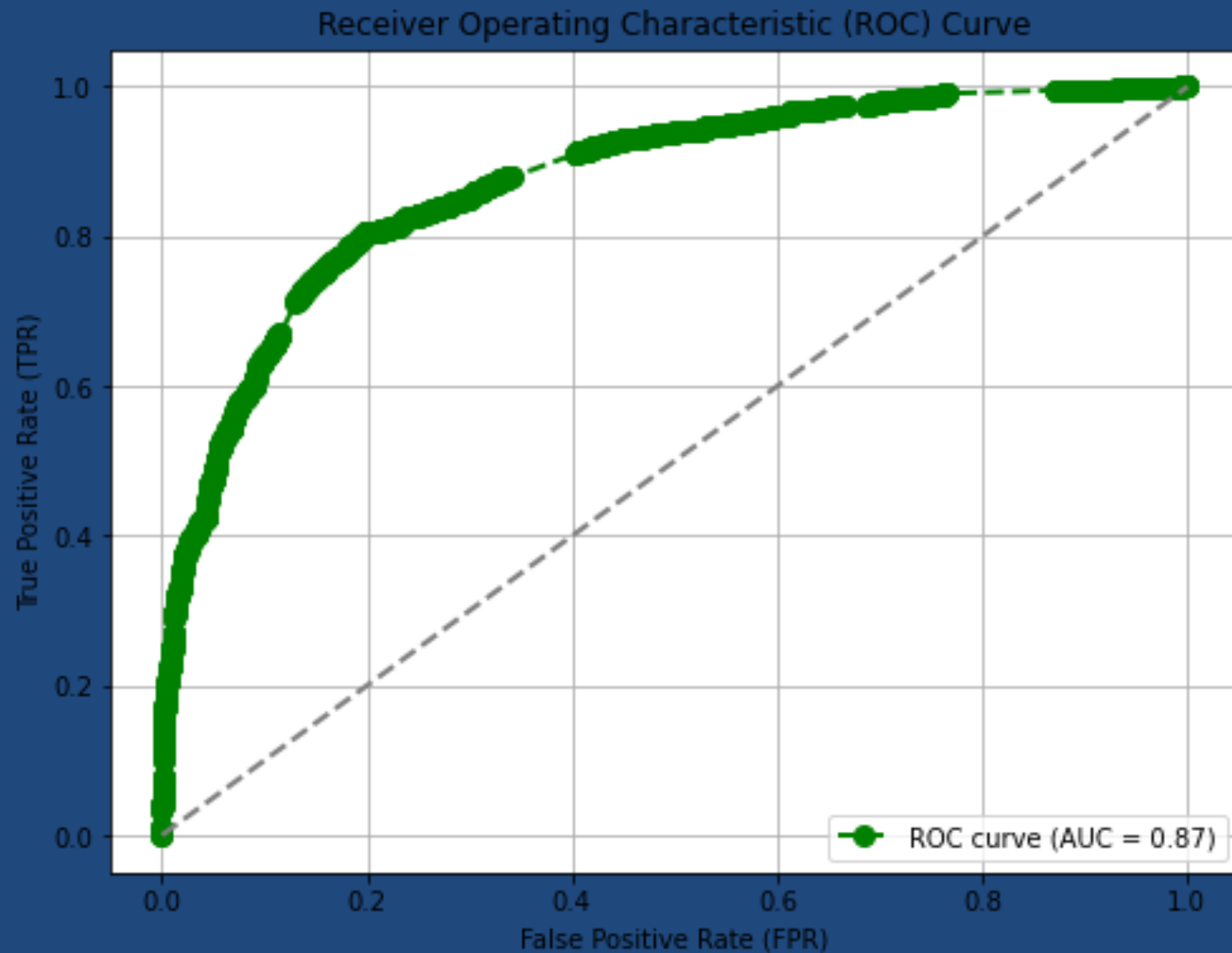
# CORRELATION

- Lead Origin\_Lead Import' has a very strong connection with 'Lead Source\_Facebook,' indicated by a correlation of 0.98.
- Similarly, 'Lead Origin\_Lead Add Form' is highly correlated with 'Lead Source\_Reference,' with a correlation coefficient of 0.85.
- This suggests a close relationship between these variables, meaning that if one is dropped, it may not substantially impact the overall model performance.

# MODEL BUILDING

- SPLITTING DATA INTO TRAIN AND TEST
- USING RFE
- BUILDING MODEL
- PREDICTIONS ON TEST SET

# ROC CURVE



# CONCLUSION

- In the training dataset, the model achieved an accuracy of 80.57%, indicating the overall correctness of predictions. The sensitivity, measuring the model's ability to correctly identify positive cases, stood at 79.72%, while the specificity, indicating the model's accuracy in identifying negative cases, was 81.08%.
- For the test dataset, the model exhibited an accuracy of 80.34%. The sensitivity for this set, denoting the model's effectiveness in correctly identifying positive instances, was 79.27%, and the specificity, representing the model's accuracy in recognizing negative instances, reached 81.04%. These metrics collectively assess the performance of the model on both the training and test datasets.

# CONCLUSION

- Certain factors, such as using the 'Lead Add Form', being a 'Working Professional', and spending more time on the website, positively influence a higher conversion rate. In particular, individuals employed or unemployed show better conversion rates. Referral leads from existing customers also have notably high conversion rates. Google and Direct Traffic channels appear promising for conversions. Additionally, leads engaged through 'SMS Sent' or 'Email Opened' tend to convert more. The most common specialization category is 'Others', followed by Finance Management, HR Management, and Marketing Management.



# RECOMMENDATIONS

- Prioritize features with high conversion rates, such as 'Lead Origin\_Lead Add Form,' 'Current\_Occupation\_Working Professional,' and 'Total Time Spent on Website,' in lead generation efforts.
- Aggressively target working professionals for higher conversion, given their higher probability and better financial situations.
- Incentivize referral leads from old customers to boost the conversion rate.
- Increase media usage frequency, like Google ads or email campaigns, for improved efficiency.
- Focus efforts on leads with 'Last Activity' as 'SMS Sent' or 'Email Opened' for higher conversion rates.
- Enhance user experience by analyzing customer behavior on the website, creating engaging content, and ensuring user-friendly navigation.