Lead Scoring Project

Prepared By -
Purvi Gupta

Praful B

PROBLEM STATEMENT

X Education needs a lead scoring model to identify high-potential leads, improve sales efficiency.

CEO's Expectation: The CEO has set a target of 80% conversion rate, emphasizing the importance of this project.

Need for Lead Prioritization: The company needs a system to identify and prioritize "hot leads" – those most likely to convert – to improve sales efficiency.

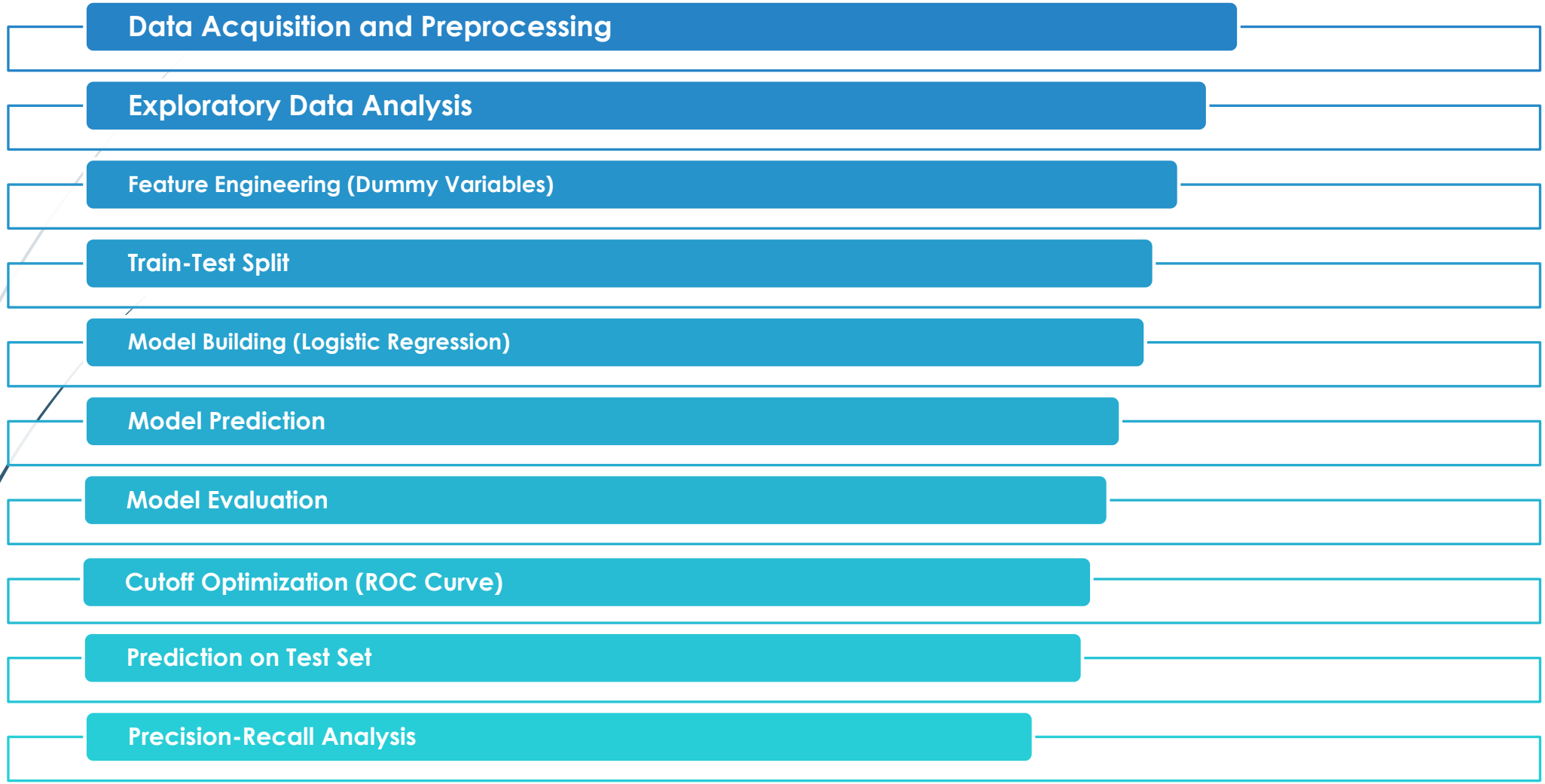
Lead Conversion Rate: X Education faces a challenge with a low lead conversion rate of approximately 30%, despite generating a significant number of leads.

Inefficient Sales Process: The current sales process involves contacting all leads, resulting in wasted effort and resources on less promising prospects.

BUSINESS GOALS

- **Business Impact:** Contribute to achieving the company's target conversion rate of 80% by enabling focused sales efforts on high-potential leads.
- **Develop a Lead Scoring Model:** Build a logistic regression model to assign a lead score between 0 and 100 to each lead. This score will indicate the likelihood of conversion, with higher scores representing "hot leads" and lower scores representing "cold leads.“
- **Provide Actionable Recommendations:** Based on the developed model and analysis, provide clear and actionable recommendations to X Education for optimizing their lead conversion process. This will include how to use the lead scores effectively.
- **Data-Driven Insights:** Leverage data analysis to identify key factors influencing lead conversion and incorporate these insights into the lead scoring model.

Steps Followed in Data Analysis



Data Acquisition & Data Understanding

- Dataset Overview:
  - Contains ~9000 leads with various attributes
  - Features include: Demographics, Source of lead, Activity on website
  - Target variable: Converted (0/1)
  - Multiple categorical and numerical variables

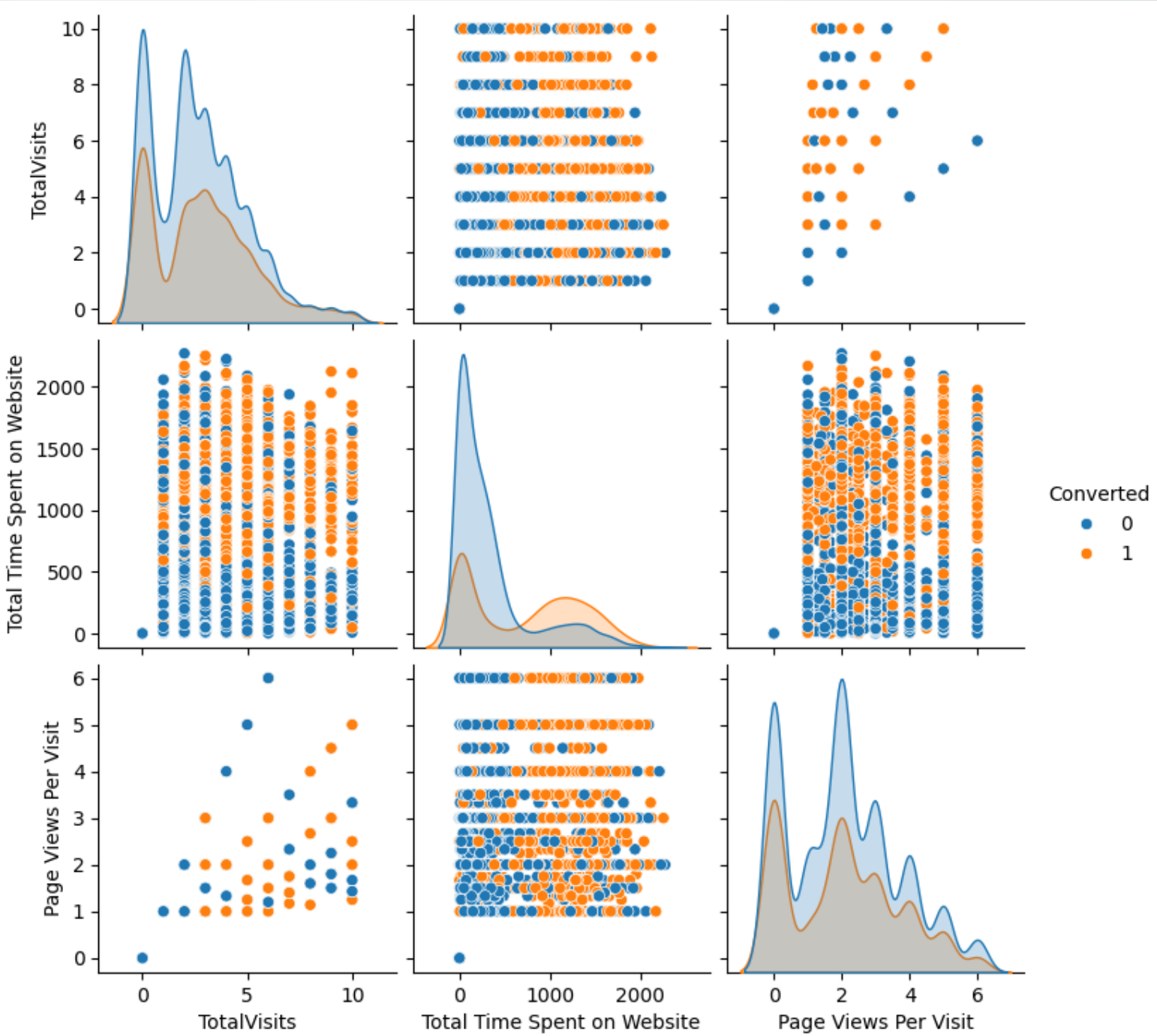
Data Preprocessing

- **Handling "Select" Values:** Converted all instances of the string "Select" to NaN (Not a Number) for consistent missing value handling.
- **Identifying and Removing Single-Value Columns:** Identified columns with only one unique value
- **Missing Value Analysis:** Checked the number of missing values in each remaining column
- **Handling High Missing Value Columns:** Removed features with a missing value percentage greater than or equal to 30%.
- **Columns with < 30% Missing Values:** Addressed missing values in columns with less than 30% missing data, as these were retained
- **Imputation Strategy: Continuous Variables (TotalVisits, Page Views Per Visit):** Imputed missing values in these continuous variables using the *median* due to the presence of outliers. This is a robust approach as the median is less sensitive to extreme values than the mean.
- **Categorical Variables:** Imputed missing values in categorical columns using the *mode* (most frequent value) for each respective column. Specifically

EDA

From this pair plot 'Total Time Spent on Website' variable seems more important

Leads who spent more time on the website were much more likely to convert (**notice the concentration of orange points (Converted = 1) towards the higher end of the x-axis**).



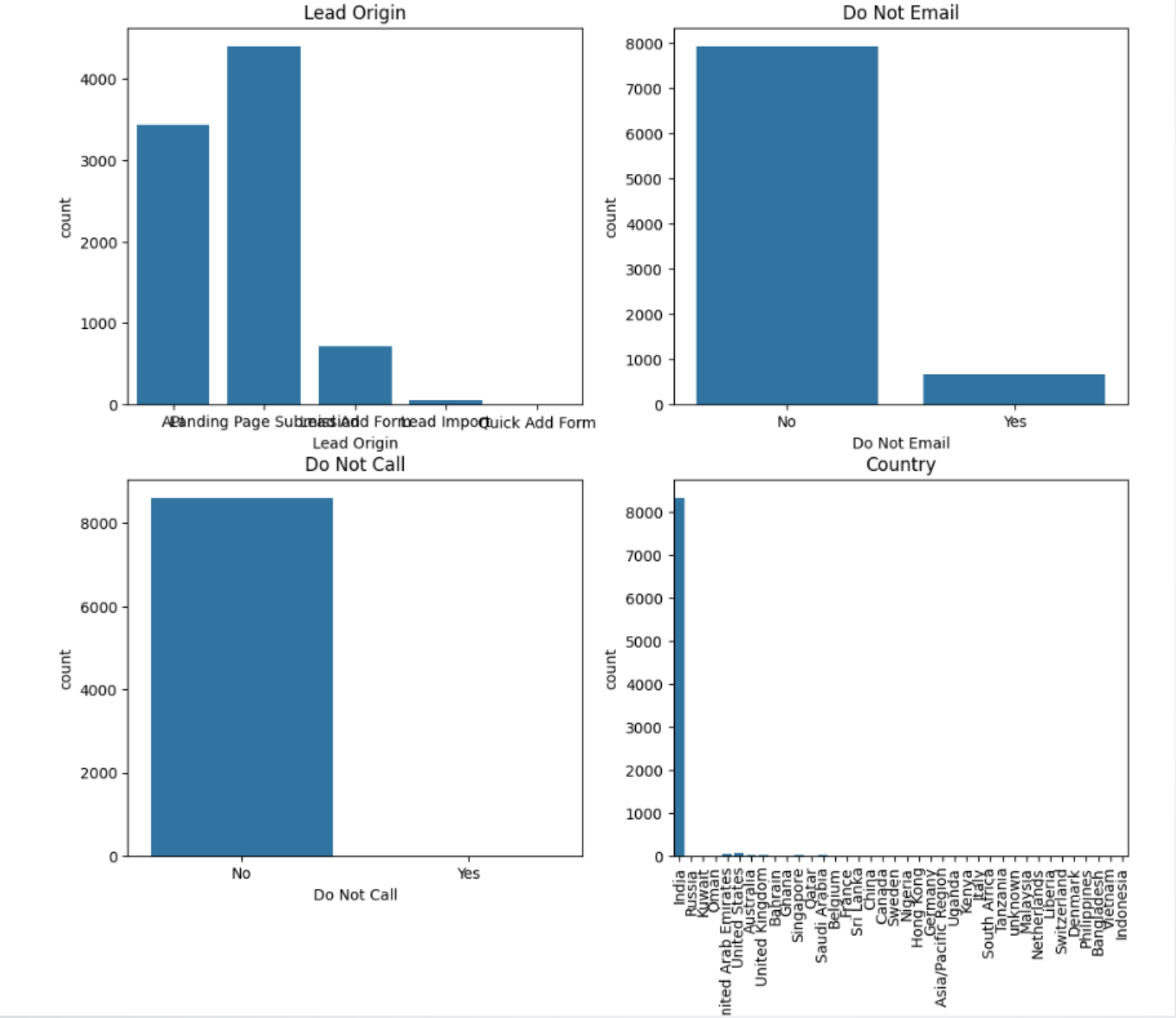
EDA

Key Takeaways for X-Education (Based on the Graph in previous slide)

- **Focus on Time Spent:** The most important factor appears to be "Total Time Spent on Website." X-Education should focus on strategies that encourage leads to spend more time on their site. This could involve improving content, user experience, and site navigation.
- **Further Investigation of Total Visits:** While "Total Visits" doesn't show a strong relationship on its own, it might be worth exploring further. Perhaps leads who visit more frequently *and* spend a significant amount of time are the most likely to convert. Feature engineering (creating new features from existing ones) could be helpful here.
- **Page Views Per Visit - Less Important (Initially):** "Page Views Per Visit" seems less influential at first glance. However, it's essential to keep this variable in the model initially and assess its importance through statistical measures during model building. It might contribute in combination with other variables.

Univariate Analysis 1

EDA

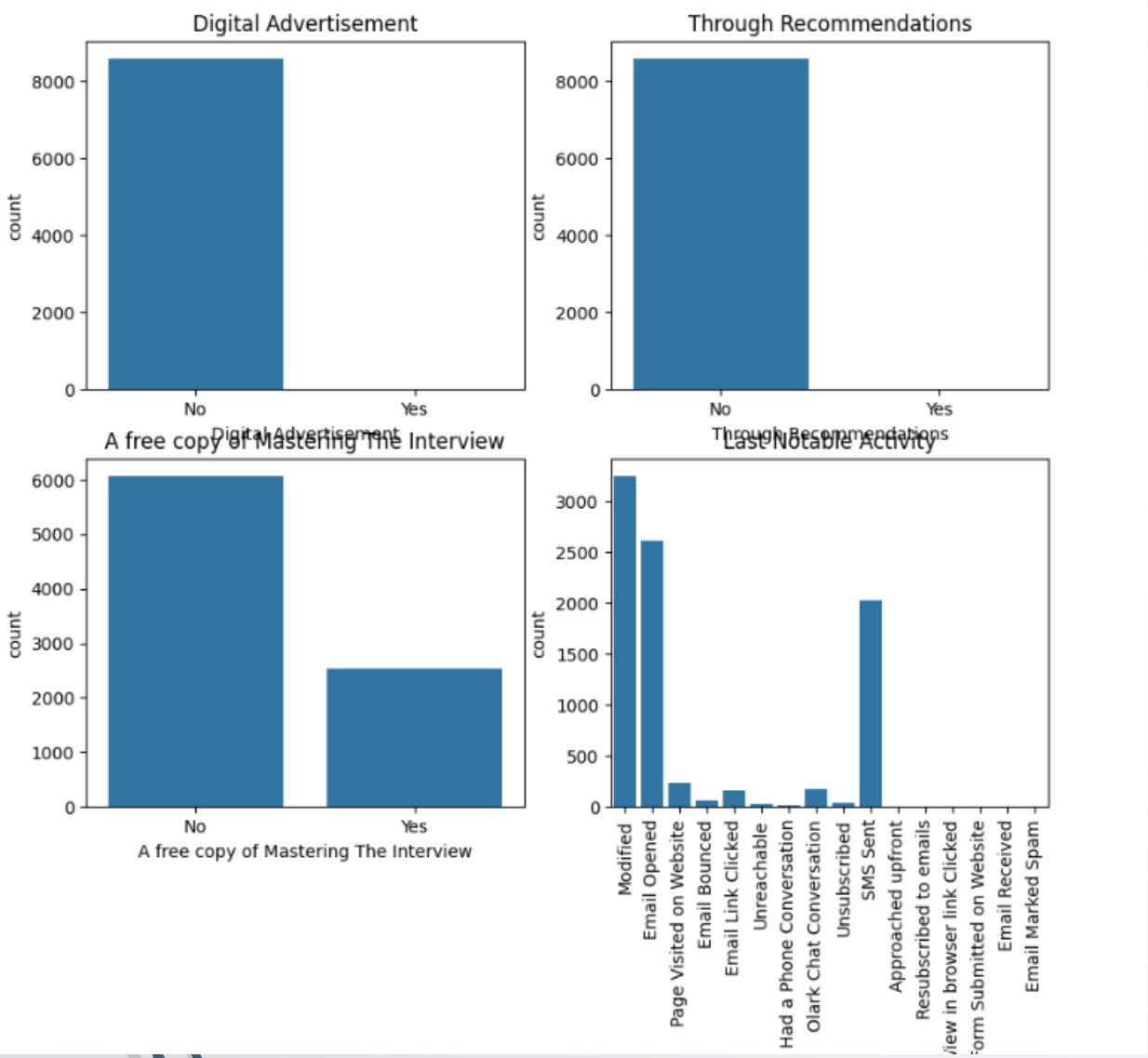


Key Takeaways

- **Lead Origin** is depicted in the first chart, showing the count of leads originating from different sources. The most frequent source is "Landing Page Submission", followed by "API"
- The second chart, **Do Not Email**, presents a binary distribution. The vast majority of leads are not flagged with a "Do Not Email" status, while a small fraction are.
- The third chart, **Do Not Call**, also shows a binary distribution. Similar to "Do Not Email", the overwhelming majority of leads are not flagged with a "Do Not Call" status.

Univariate Analysis 2

EDA

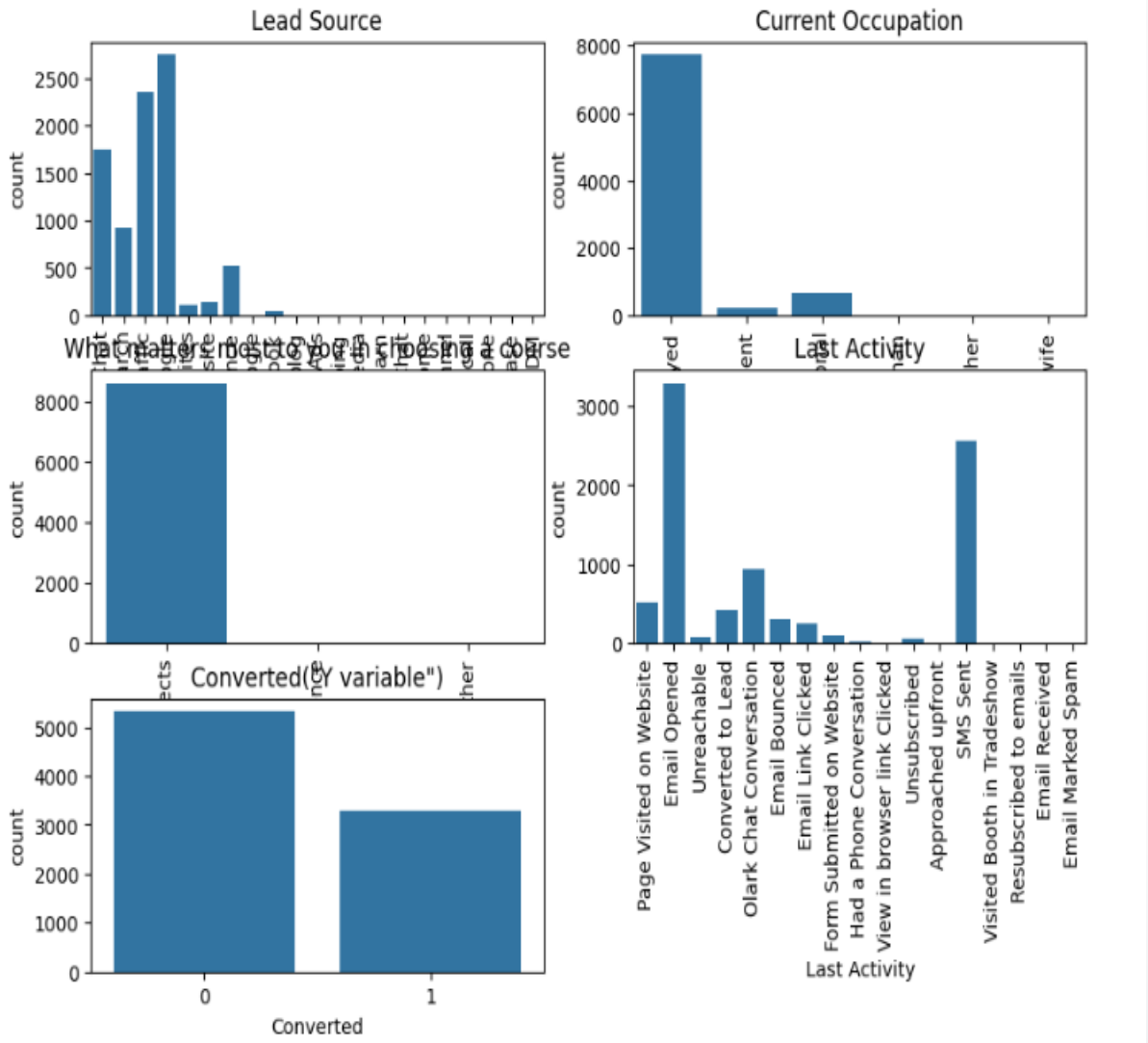


Key Takeaways - Category variables

- These charts depict the counts of different categories related to leads or customers
- Digital advertisements are not a significant source of leads compared to other channels (as we'll see in the other charts).
- Recommendations, like digital ads, are not a primary driver of leads in this dataset.
- Offering the free book appears to be a more effective lead generation tactic than digital ads or recommendations, as it has a noticeably higher number of associated leads.
- Lead Notable Activity -Email engagement and website visits are common lead behaviors.

Univariate Analysis 3

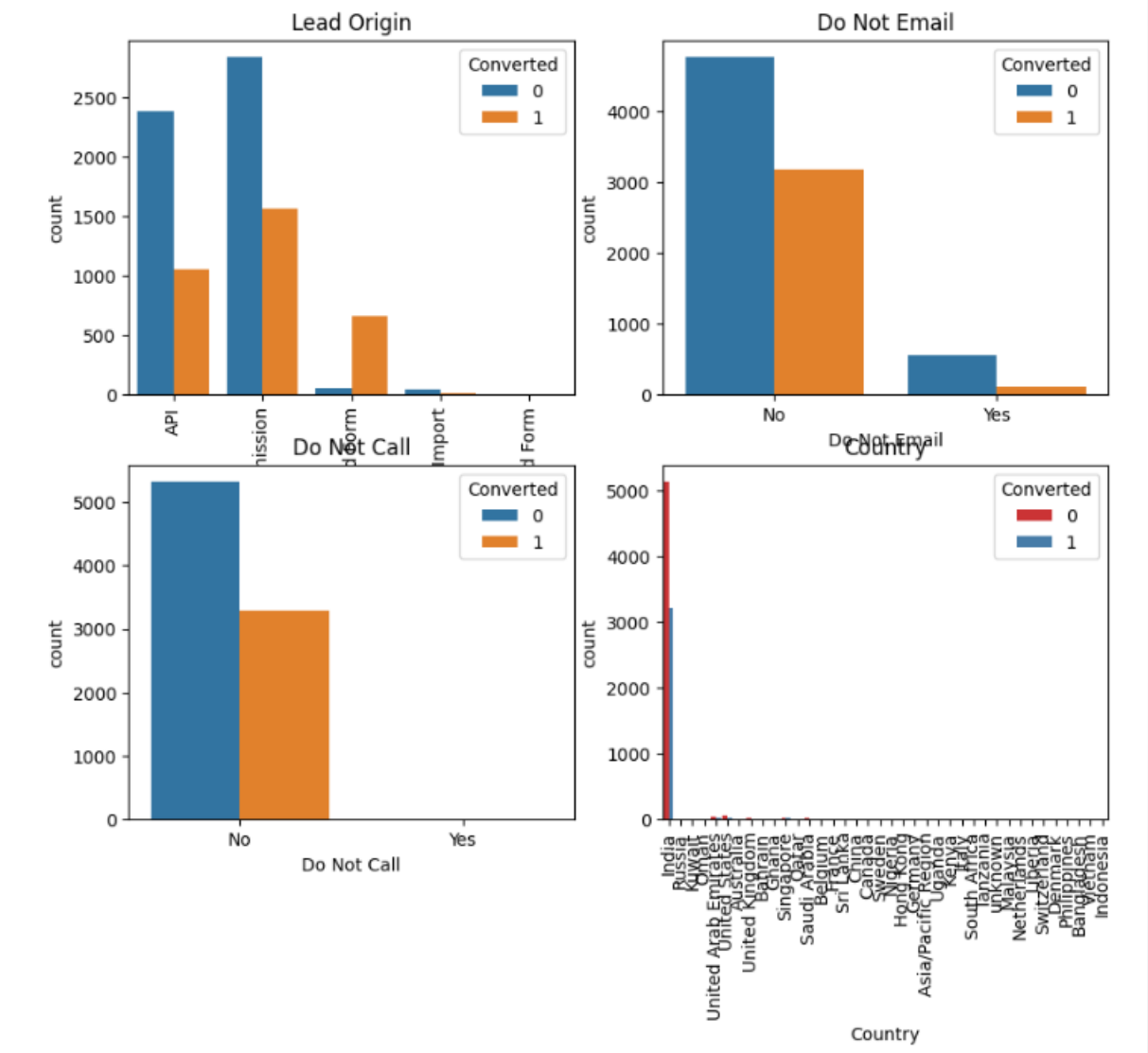
EDA



- Overall Analysis and Insights: **Category variables**
- **Lead Source:** This chart shows the distribution of leads based on their origin. "What is your current occupation" dominates, followed by "Select," with other sources having significantly lower counts.
- **Current Occupation:**The vast majority are categorized as "Unemployed," followed by "Student," with other occupations having much lower representation.
- **Last Activity:** "Email Opened" is the most frequent activity, followed by "Page Visited on Website," with other activities like "Unreachable" and "Had a Phone Conversation" having lower counts.
- **Converted (by variable):** The majority of leads are not converted ("0"), while a smaller portion is converted ("1").

Univariate Analysis 4

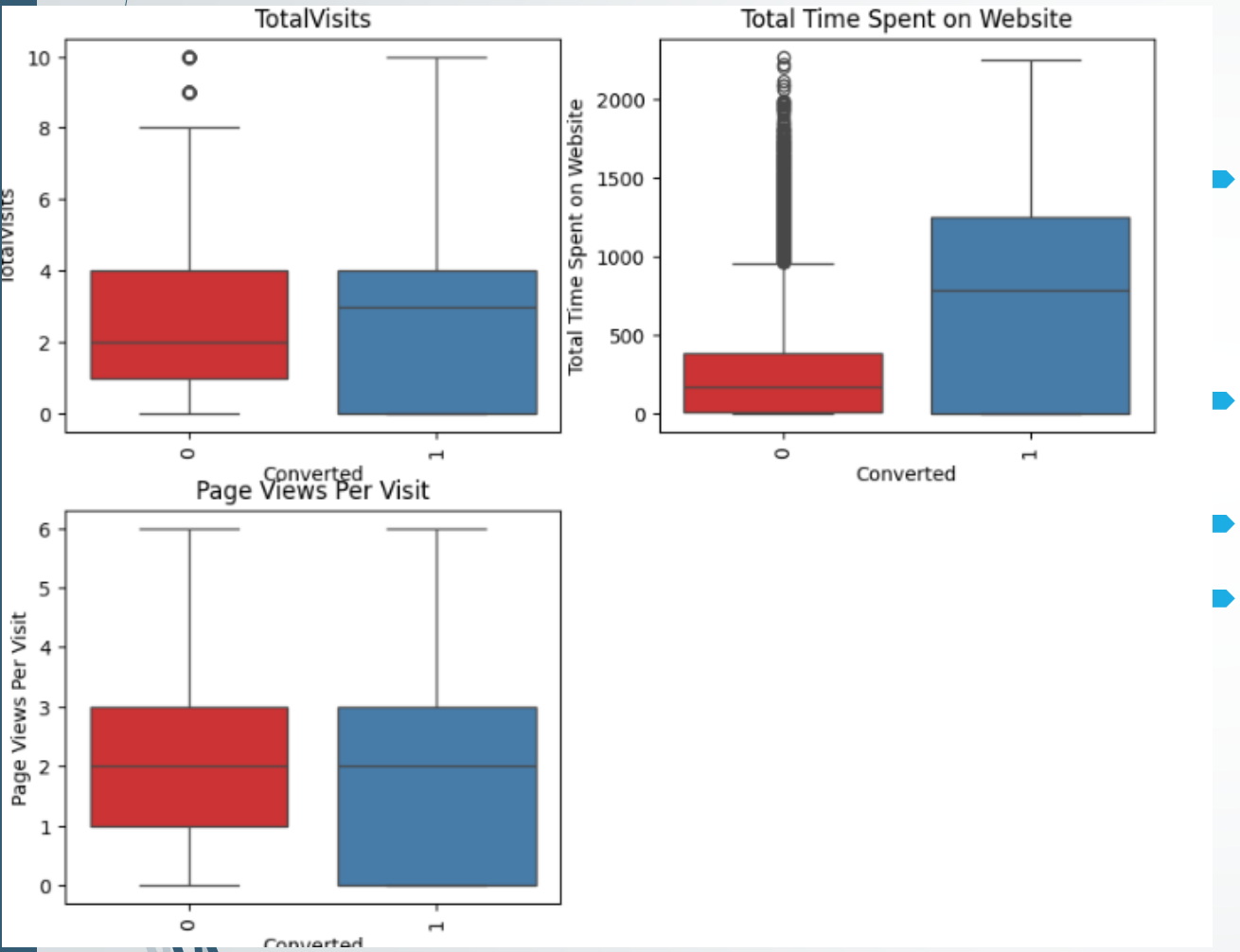
EDA



- Key Recommendation -
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- Lead Source- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

Univariate Analysis 5

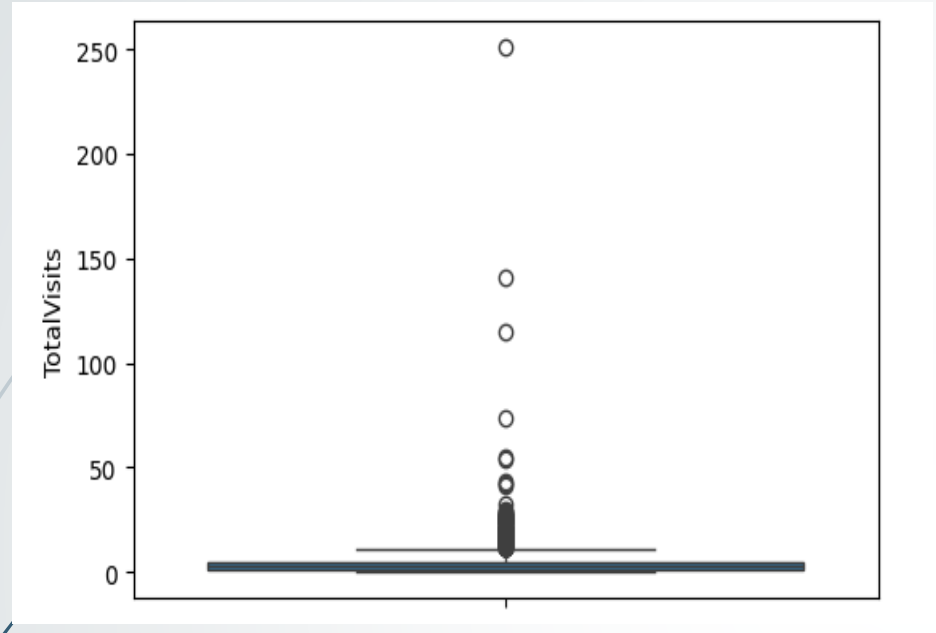
EDA



- **Individual Box Plot Insights and Interpretations**

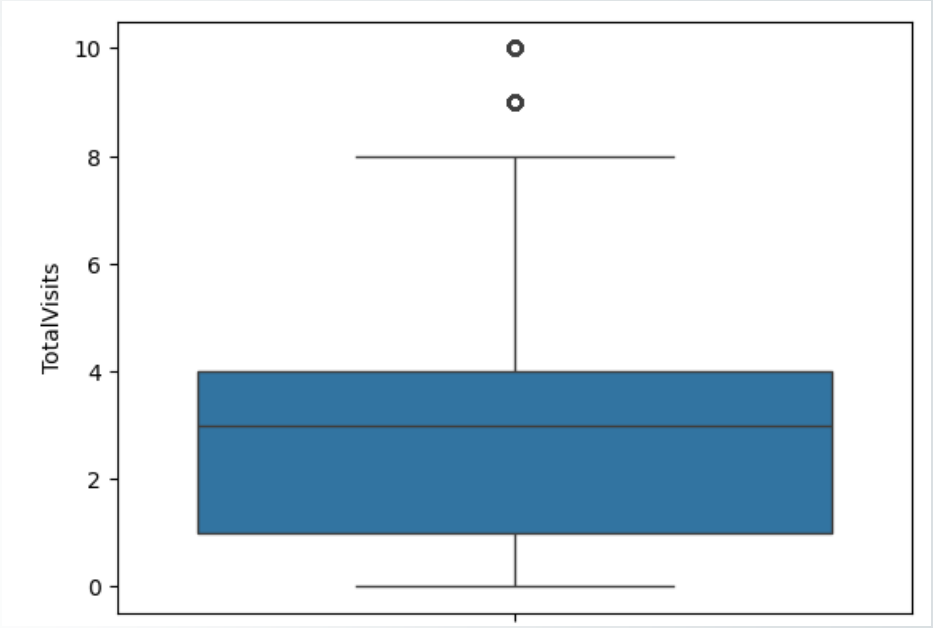
- **Total Visits:** The median number of visits is slightly higher for converted users (Converted = 1) compared to non-converted users (Converted = 0).
- **Total Time Spent on Website:** There are several high outliers in the converted group, indicating some users spend exceptionally long periods on the site.
- **Page Views Per Visit:** There's one high outlier in the non-converted group.
- Time Spent as a Strong Indicator
- **Potential Outliers:** The presence of outliers, especially in the "Total Time Spent" for converted users, warrants further investigation.

Outliers Detection & Treatment



Before Treatment -outliers detected

- Here we can see several data points far above the upper whisker, indicating unusually high visit counts



After Treatment -outliers Handled

- The extreme outliers have been removed. The plot now shows a more compact distribution with no extremely high values plotted individually.

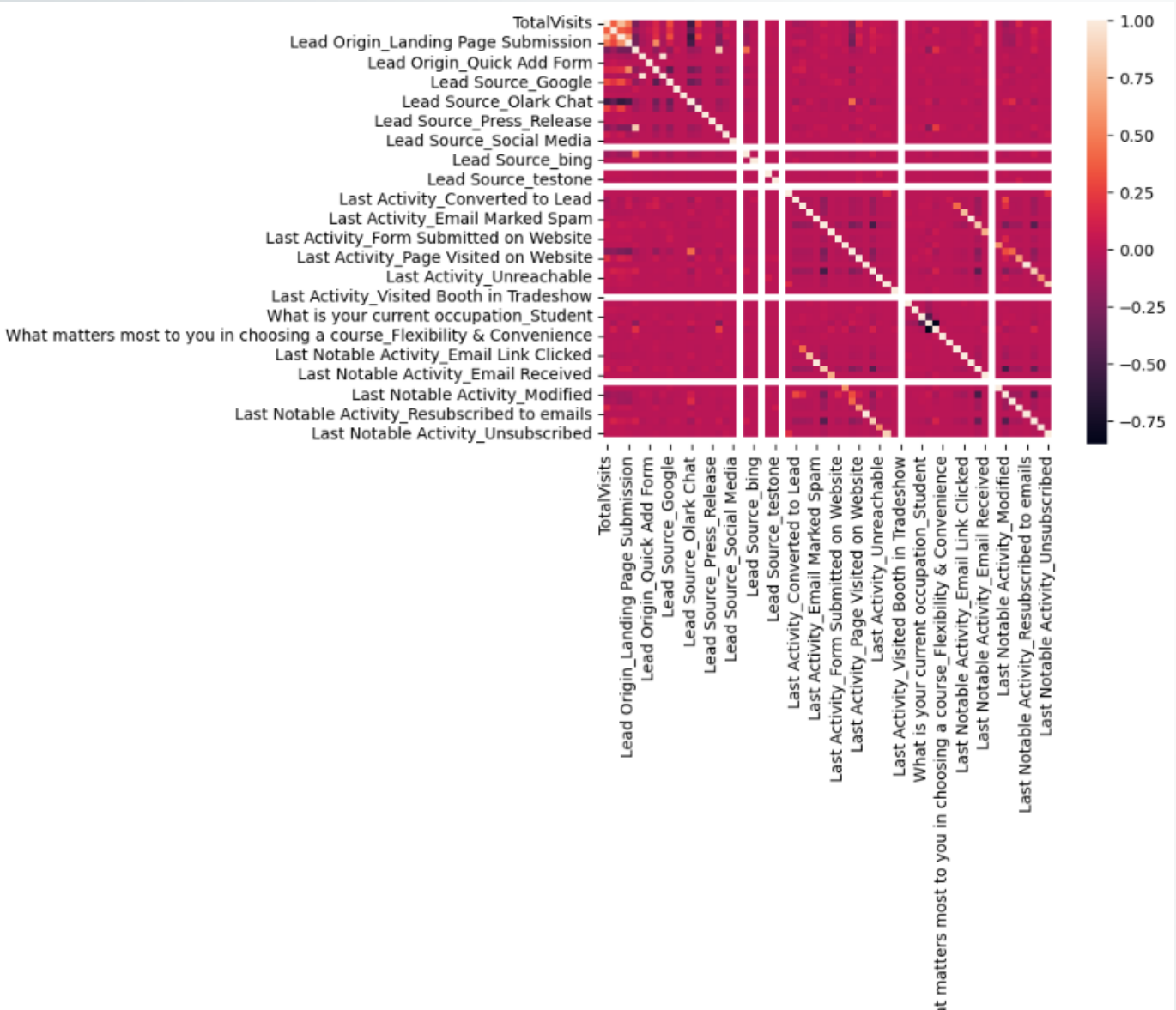
Impact of Outlier Treatment: Comparing the two box plots clearly demonstrates the impact of outlier treatment. The second plot provides a much clearer picture of the typical distribution of "TotalVisits" without the distortion caused by the extreme values.

Feature Engineering (Dummy Variables creation):

Benefits of Dummy Variables:
Enables the use of categorical data in numerical models.
Prevents the model from misinterpreting categorical values as ordinal.

Used RFE (Recursive Feature Elimination)

Correlation among variables



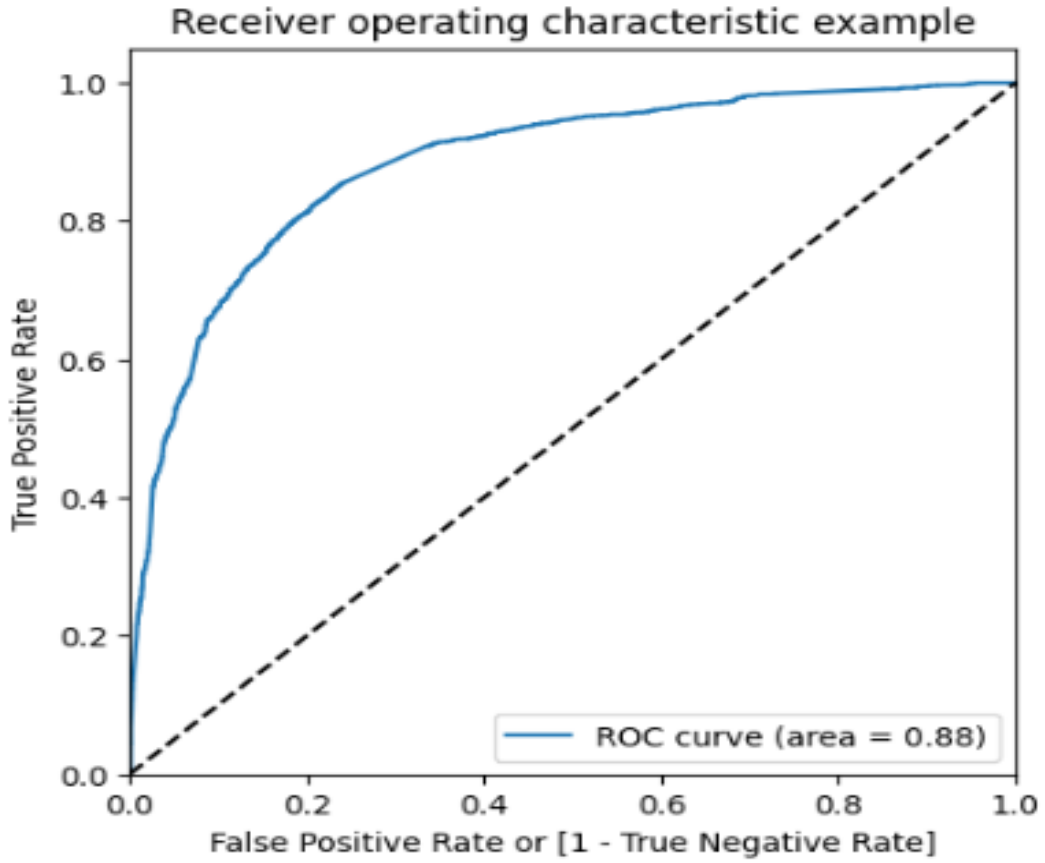
Insights and Observations:

Weak to Moderate Correlations:

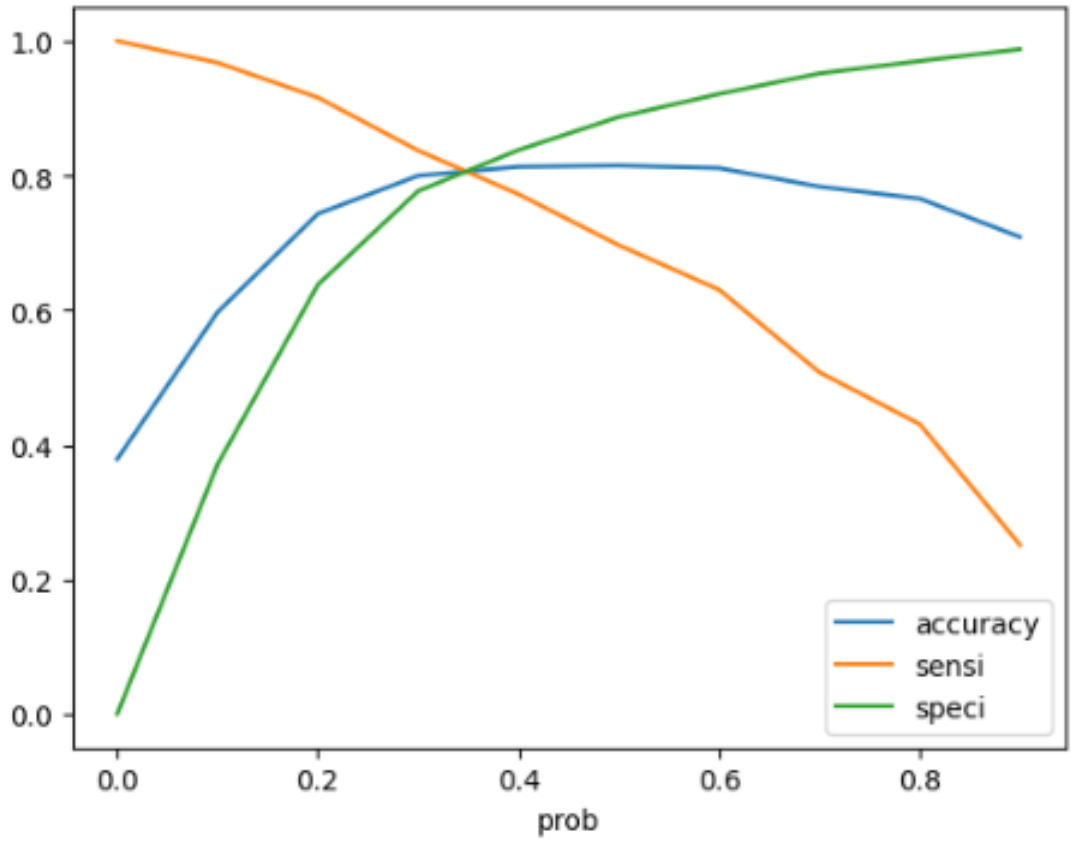
- Most of the correlations appear to be weak to moderate.
- There are no extremely dark red or blue cells, suggesting that no single variable is an overwhelmingly strong predictor of another.

Model Building Stage

- Logistic Regression Model Process:
  - Split data into training (70%) and testing (30%)
  - Used Logistic Regression with optimal threshold
  - Performed cross-validation



The area under ROC curv is 0.88, which is very good value

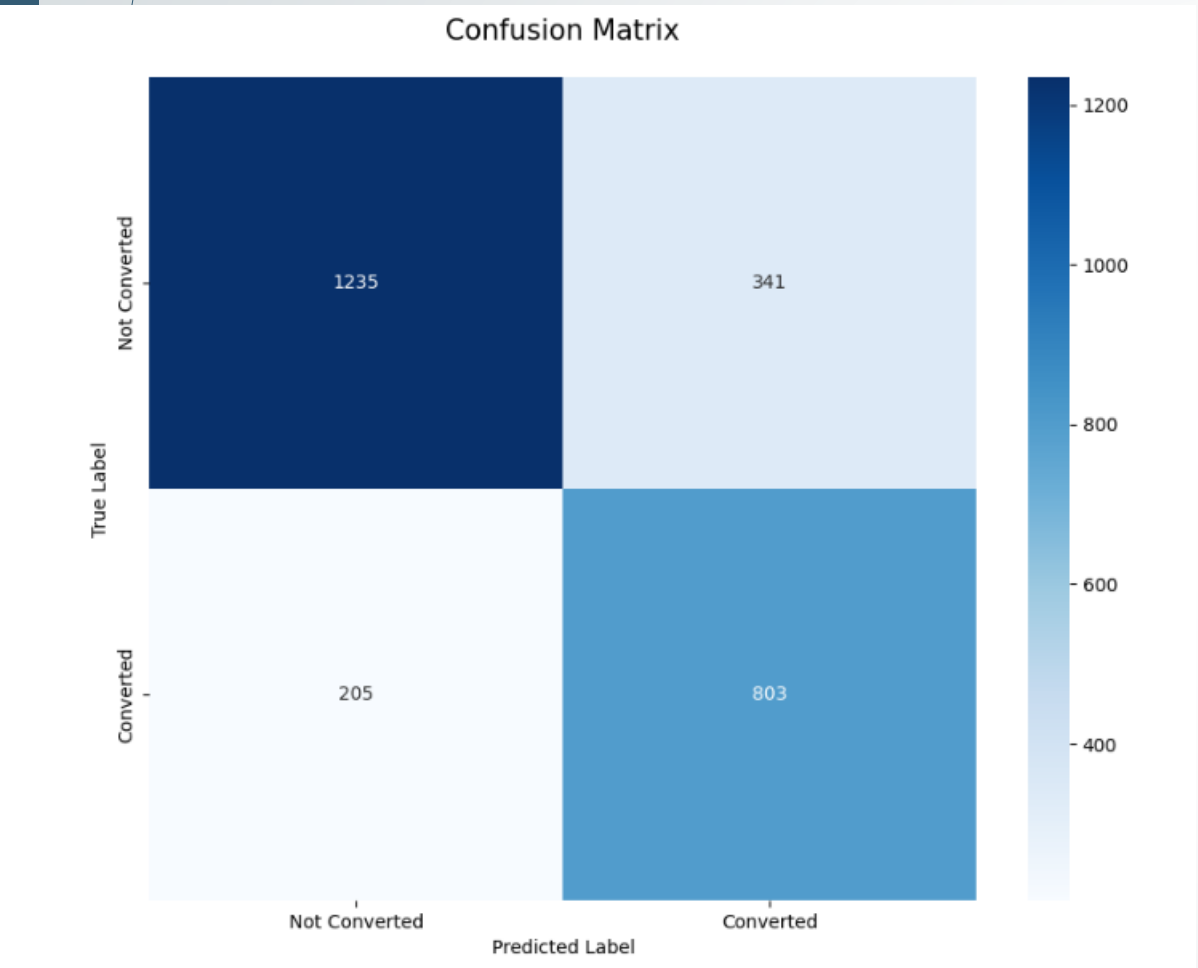


From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

Model Evaluation

- **Precision-recall tradeoff** is a crucial tool for evaluating the performance of classification models, especially when dealing with imbalanced datasets. Here's a breakdown of the insights –
- **Trade-off:** As you increase recall (find more actual positives), precision tends to decrease (more false positives among the identified cases) and vice-versa.
- **Green Line Superiority:** The green line clearly outperforms the red line. It indicates a model that is better at balancing precision and recall. For most recall values, the green line has a higher precision than the red line.
- **Effectiveness at Different Thresholds:** By examining specific points on the curve, you can understand how the model behaves at different classification thresholds.

Model Evaluation



- Key Insights and Metrics - Analysing the Matrix

- **TN = 1235:** The model correctly predicted 1235 cases as "Not Converted" when they were actually "Not Converted.“
- **FP = 341:** The model incorrectly predicted 341 cases as "Converted" when they were actually "Not Converted.“
- **FN = 205:** The model incorrectly predicted 205 cases as "Not Converted" when they were actually "Converted.“
- **TP = 803:** The model correctly predicted 803 cases as ""Converted" when they were actually "Converted.

- "**Accuracy:** Calculated as: (TP + TN) / (TP + TN + FP + FN)
- In this case: (803 + 1235) / (803 + 1235 + 341 + 205) = 2038 / 2584 ≈ 0.79 or **79%**
- **Recall (for "Converted" class):** Calculated as: TP / (TP + FN)
- In this case: 803 / (803 + 205) = 803 / 1008 ≈ 0.80 or **80%**

Recommendation

- Business Actions:
  - Focus on leads with high website engagement
  - Prioritize leads from high-converting sources
  - Monitor and follow up based on last activity
  - Implement lead scoring system for real-time prioritization
  - Regular model monitoring and updates

Conclusion

- It was found that the variables that mattered the most in the potential buyers are (In descending order)
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Last Activity_Email Bounced
- Last Activity_Olark Chat Conversation
- Last Notable Activity_SMS Sent
- Last Notable Activity_Unreachable
- Lead Source_Olark Chat
- Last Activity_Converted to Lead
- Last Activity_Page Visited on Website
- TotalVisits

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses

THANK YOU