



LEAD SCORE CASE STUDY



PROBLEM STATEMENT

X Education, a provider of online courses aimed at industry professionals, is facing a critical challenge in its operations. Despite acquiring a significant number of leads daily, their lead conversion rate remains suboptimal. For example, out of 100 leads generated each day, only 30 are successfully converted into paying customers. This inefficiency is hindering their growth and overall profitability.

To tackle this issue, X Education aims to identify the most promising leads—referred to as "Hot Leads"—to help their sales team focus on high-potential customers with a greater likelihood of conversion. This approach is expected to significantly improve the company's lead conversion rates and operational efficiency.



BUSINESS OBJECTIVE

X Education's primary objective is to develop and implement a predictive model to identify the most promising leads. This model will streamline the sales process and be adaptable for long-term deployment, thereby ensuring sustained improvements in lead conversion rates.

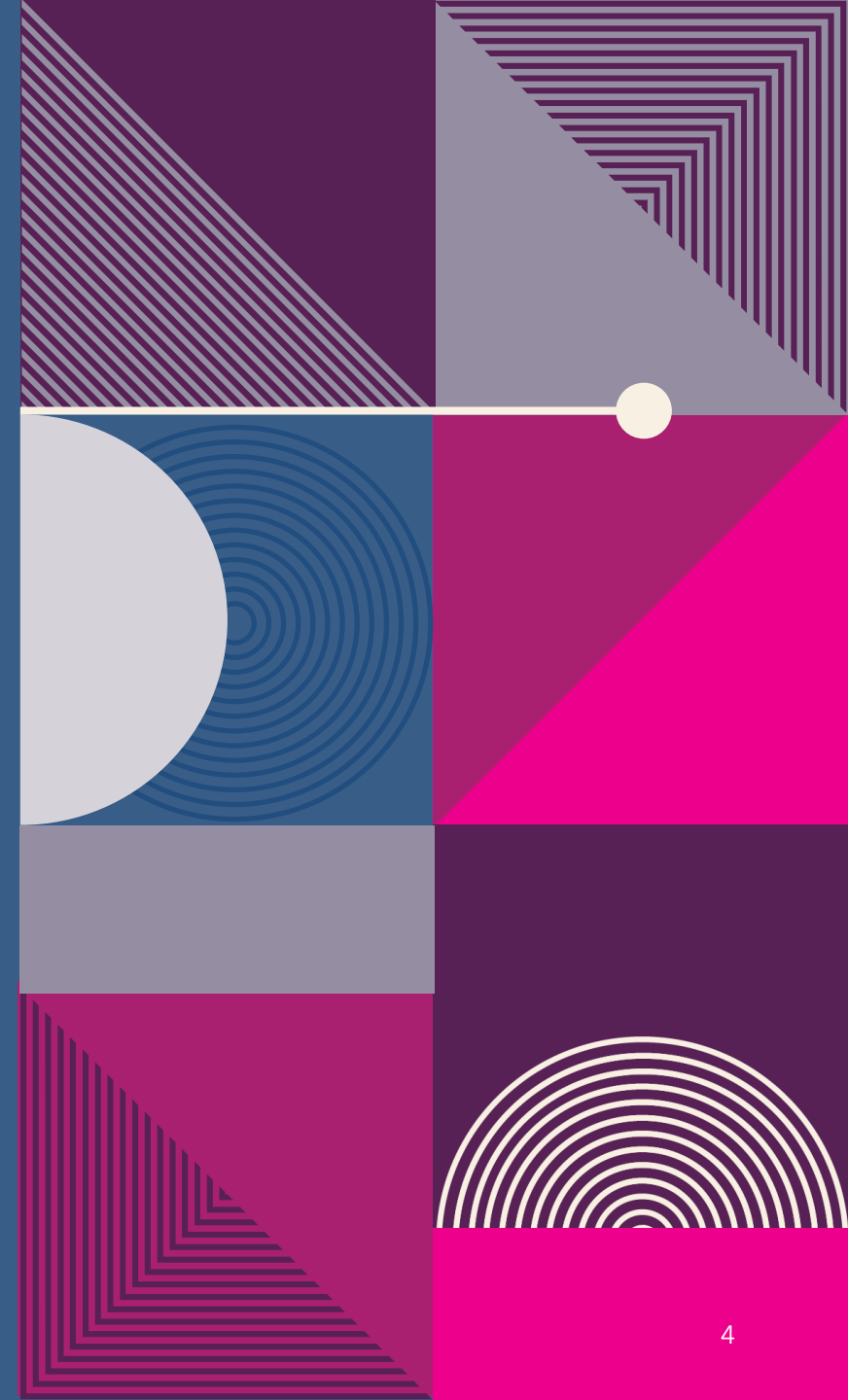
SOLUTION METHODOLOGY

1. Data Cleaning and Manipulation

- Duplicate Handling: Identified and removed duplicate entries to ensure data accuracy.
- Missing Values: Managed missing and NA values through appropriate imputation or by discarding columns with excessive missing data.
- Irrelevant Features: Eliminated columns with single unique values or minimal relevance, such as "Magazine" and "Receive More Updates About Our Courses."
- Variance Analysis: Dropped features with insufficient variance, including "Do Not Call" and "Newspaper Article."
- High Missing Data Columns: Removed columns with more than 35% missing values, such as "How did you hear about X Education" and "Lead Profile."

2. Exploratory Data Analysis (EDA)

- Univariate Analysis: Conducted assessments of distributions and value counts to understand variable behavior.
- Bivariate Analysis: Evaluated correlations and patterns between variables to identify key relation



3. Data Transformation

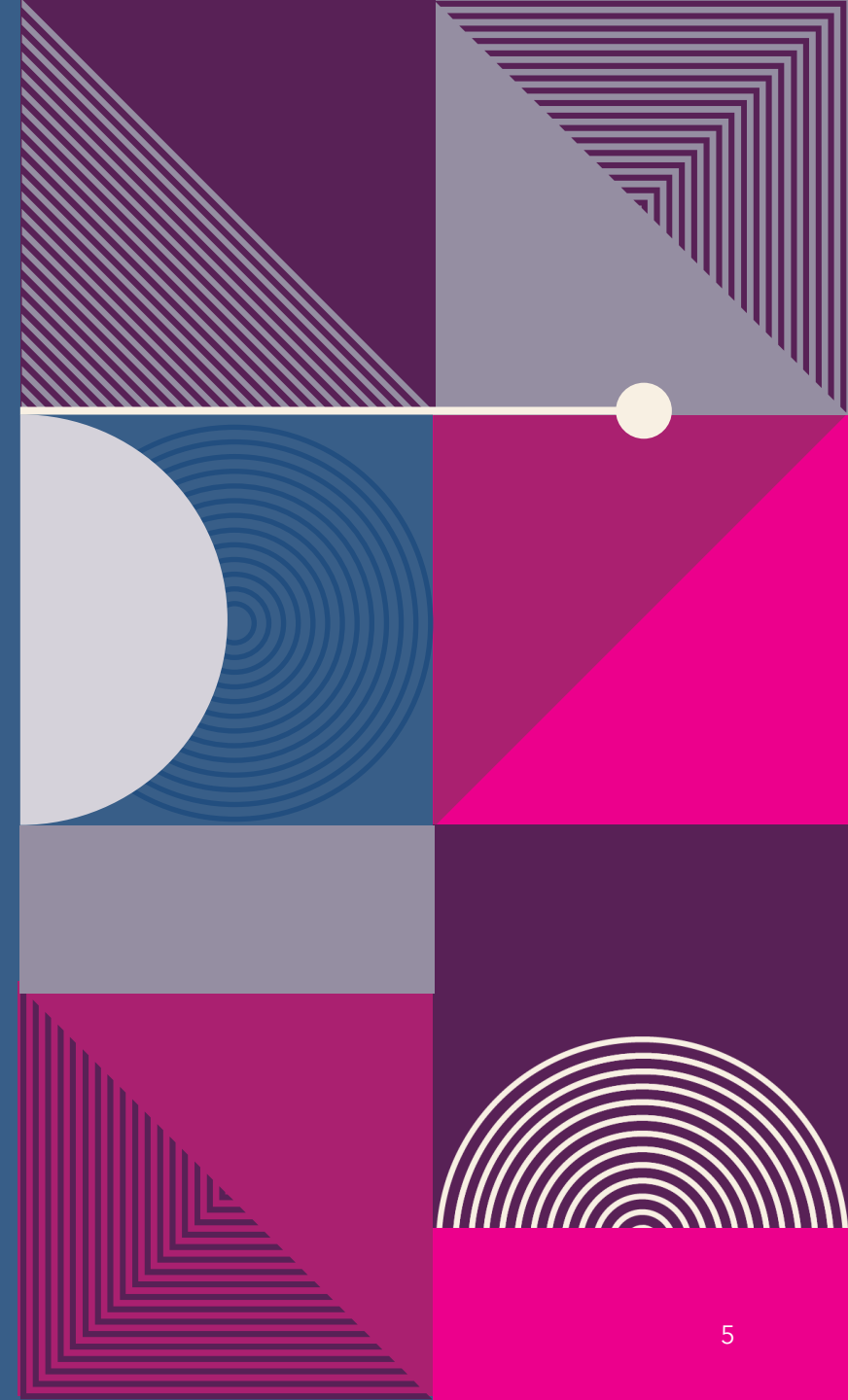
- Numerical Variables: Normalized to ensure uniform scaling.
- Categorical Variables: Created dummy variables for object-type features to enable their use in the model.
- Final Dataset: After cleaning and transformation, the dataset contained 8,792 rows and 43 columns for analysis.

4. Model Development

- Train-Test Split: Divided the dataset into training and testing sets in a 70:30 ratio.
- Feature Selection: Applied Recursive Feature Elimination (RFE) to identify the 15 most relevant variables.
- Model Refinement: Excluded features with p-values > 0.05 and VIF values > 5 to maintain statistical significance and minimize multicollinearity.
- Classification Technique: Built and optimized a logistic regression model.

5. Model Validation and Performance


- Accuracy: Achieved an overall accuracy of 82% on the test dataset.
- Evaluation: Validated the model using the ROC curve to confirm predictive performance.





KEY INSIGHTS

The model identified several key factors that strongly influence lead conversion rates:

- 1.Total Time Spent on Website: Higher engagement correlates with higher likelihood of conversion.
 - 2.Total Number of Visits: Frequent visits indicate stronger interest.
 - 3.Lead Source: Leads from Google, direct traffic, organic search, and Welingak's website displayed higher conversion probabilities.
 - 4.Last Activity: Interactions like SMS and Olark chat conversations were significant predictors.
 - 5.Lead Origin: Leads originating from "Lead Add Format" showed greater promise.
 - 6.Current Occupation: Working professionals were more likely to convert than others.
- 

DATA MANIPULATION

- Dataset Overview:

- Total Number of Rows: 37
- Total Number of Columns: 9240

•Single Value Features: Features with only a single unique value, which do not contribute to analysis, were dropped. These include:

- "Magazine"
- "Receive More Updates About Our Courses"
- "Update me on Supply Chain Content"
- "Get updates on DM Content"
- "I agree to pay the amount through cheque"

•Redundant Identifiers: Columns like "Prospect ID" and "Lead Number" were removed as they do not provide meaningful insights for the analysis.

•Low-Variance Features: Features that lack sufficient variance across observations were excluded. Examples include:

- "Do Not Call"
- "What matters most to you in choosing course"
- "Search"
- "Newspaper Article"
- "X Education Forums"
- "Newspaper"
- "Digital Advertisement"

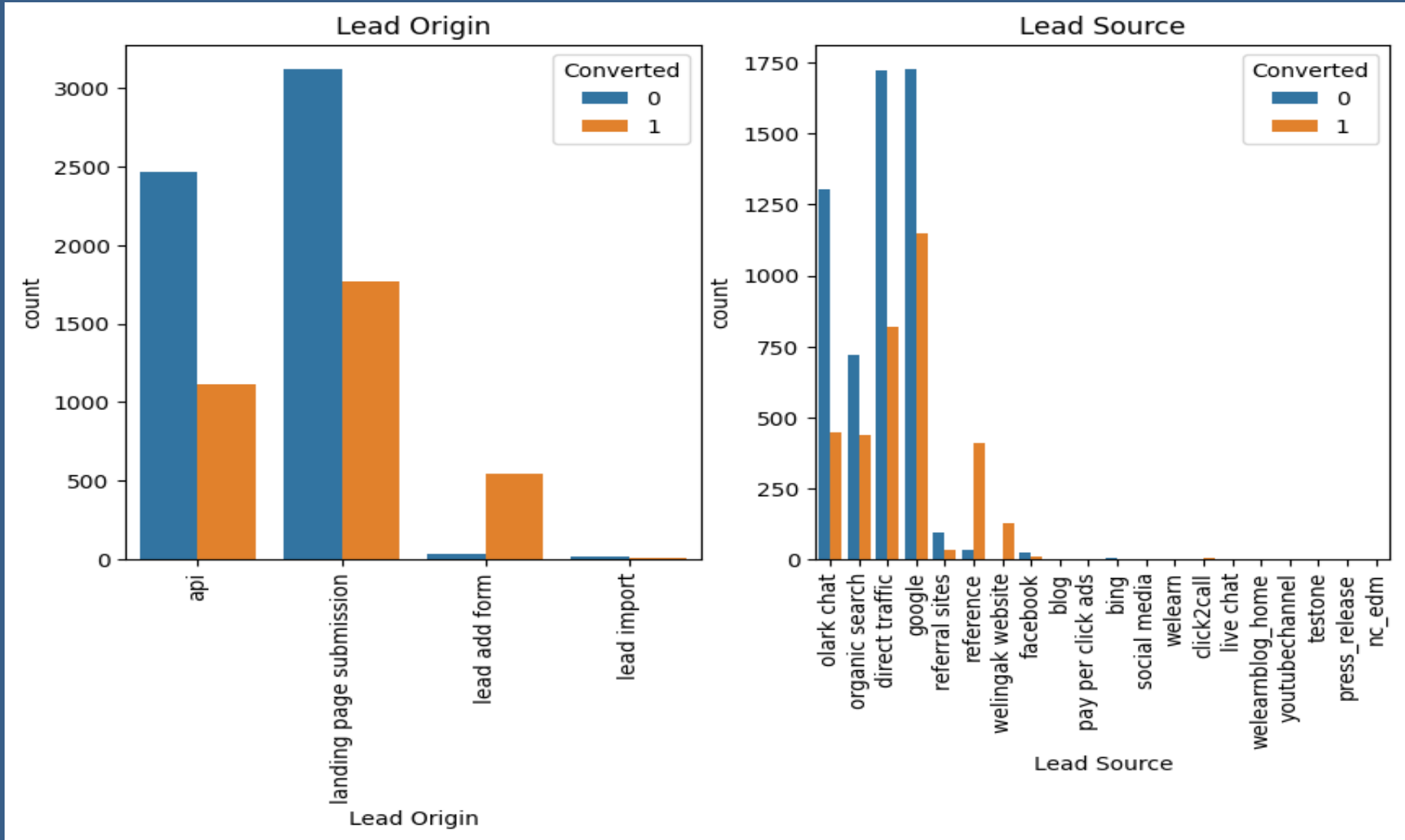
•High Missing Value Features: Columns with more than 35% missing values were dropped, as imputing them would compromise data quality. These features include:

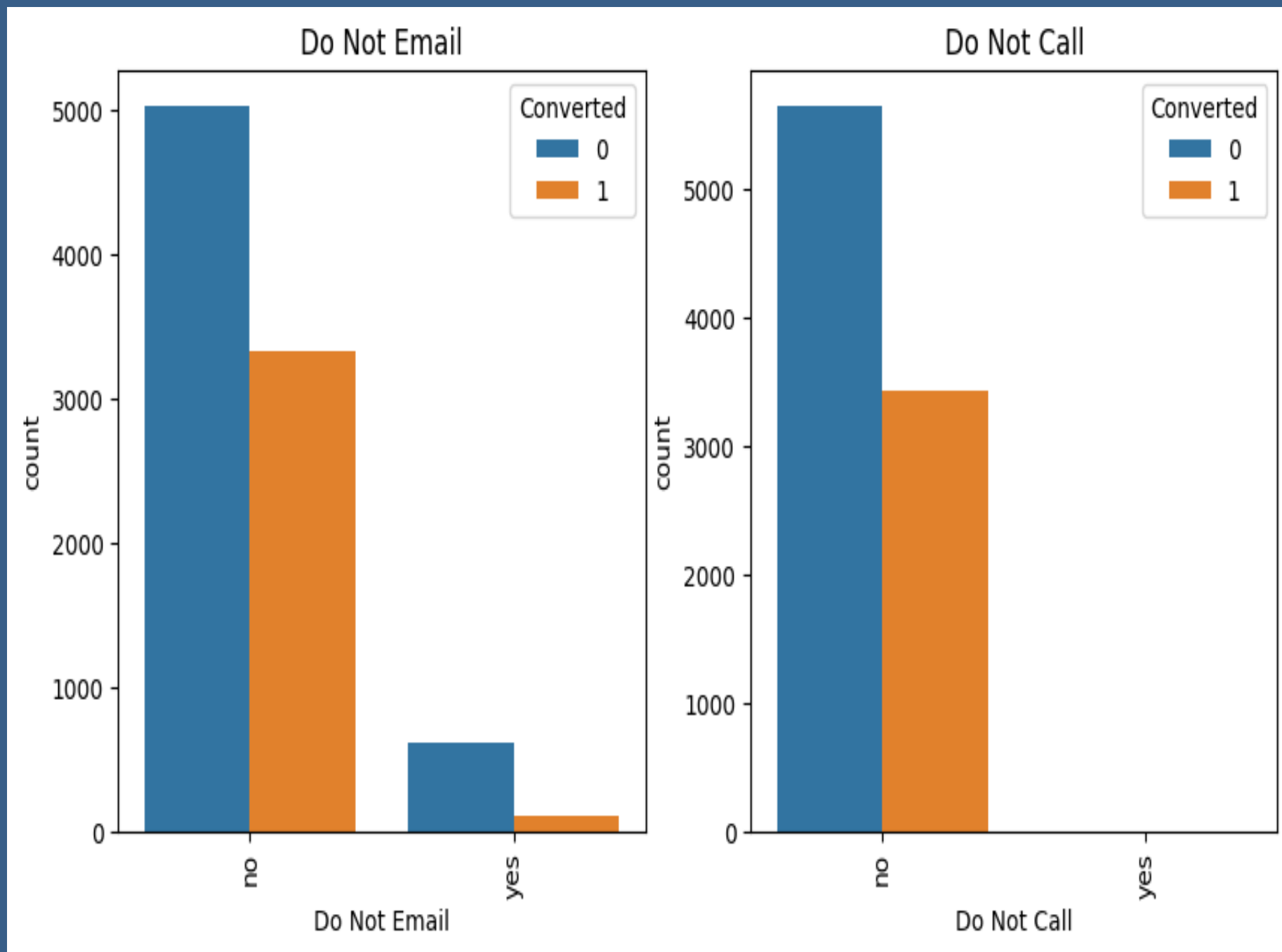
- "How did you hear about X Education"
- "Lead Profile"

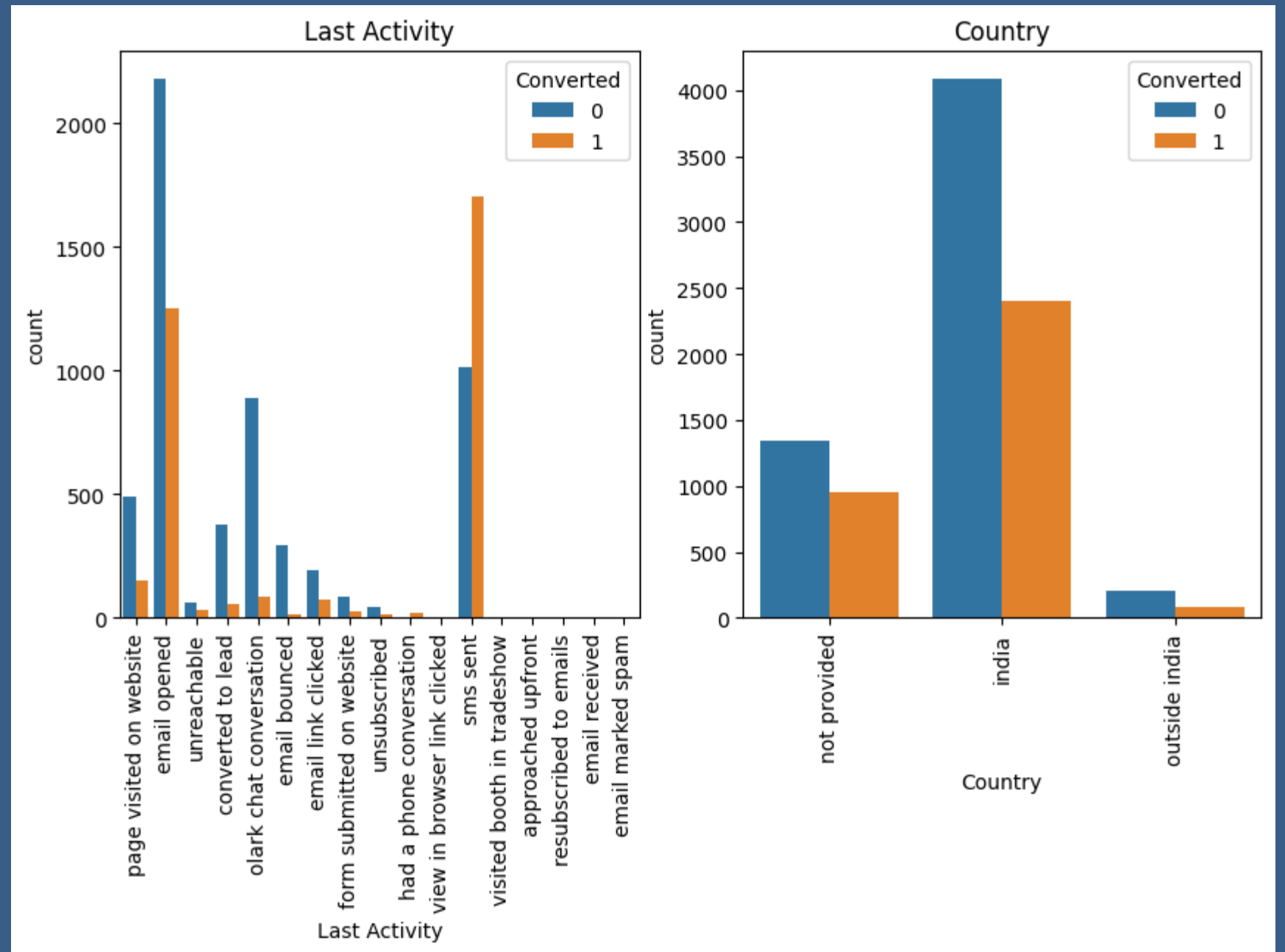
EDA



CATEGORICAL VARIABLE RELATION







DATA CONVERSION

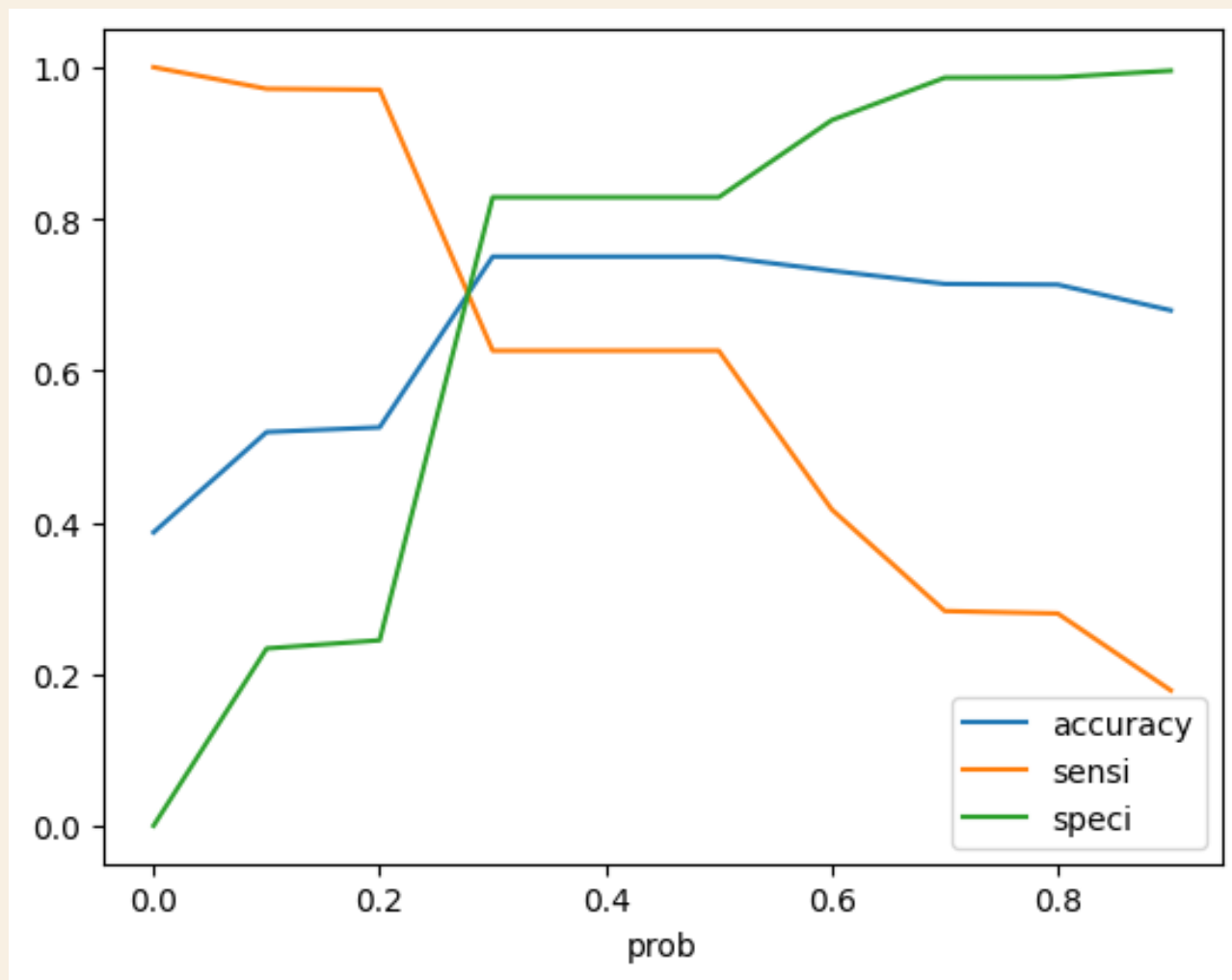
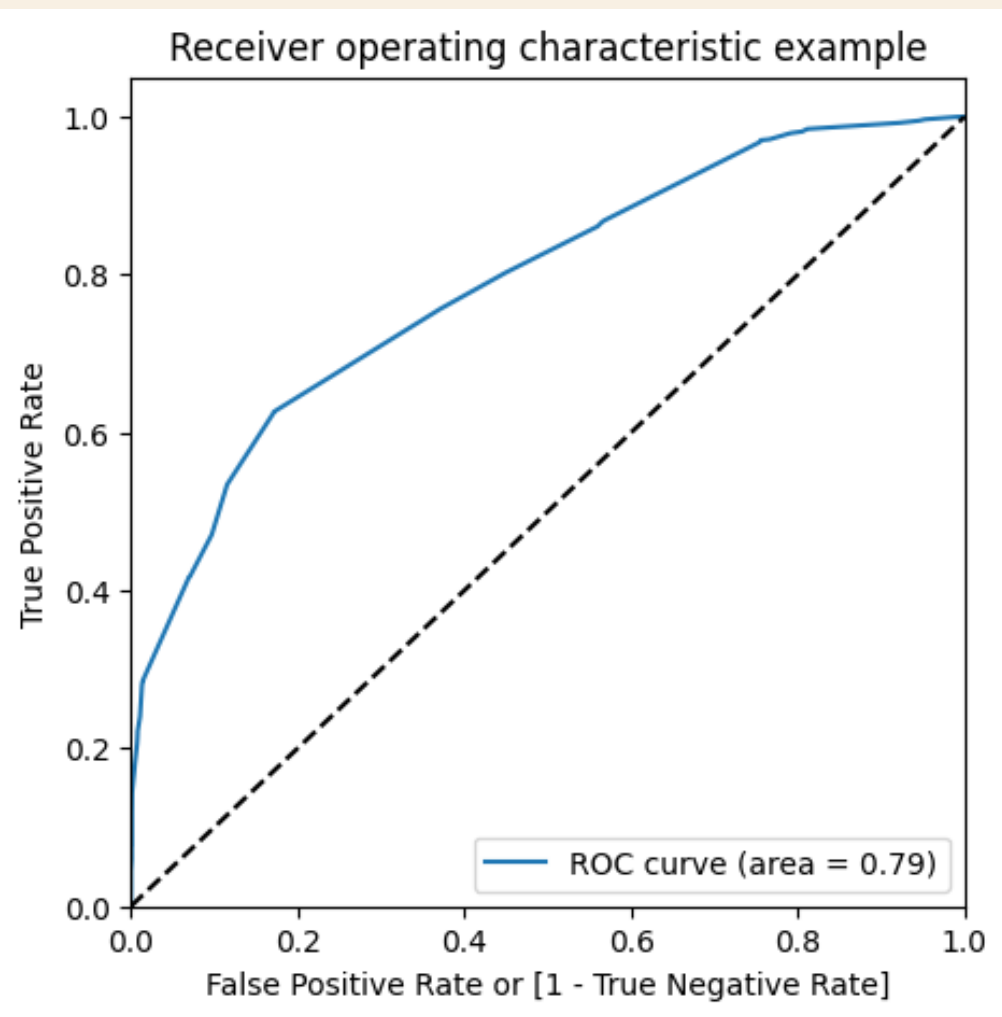
- Numerical Variables: Normalized to ensure uniform scaling.
- Categorical Variables: Created dummy variables for object-type features to enable their use in the model.
- Final Dataset: After cleaning and transformation, the dataset contained 8,792 rows and 43 columns for analysis.



MODEL BUILDING

- Train-Test Split: Divided the dataset into training and testing sets in a 70:30 ratio.
- Feature Selection: Applied Recursive Feature Elimination (RFE) to identify the 15 most relevant variables.
- Model Refinement: Excluded features with p-values > 0.05 and VIF values > 5 to maintain statistical significance and minimize multicollinearity.
- Classification Technique: Built and optimized a logistic regression model.

ROC CURVE



CONCLUSION



- 1.Total Time Spent on Website: Higher engagement correlates with higher likelihood of conversion.
- 2.Total Number of Visits: Frequent visits indicate stronger interest.
- 3.Lead Source: Leads from Google, direct traffic, organic search, and Welingak's website displayed higher conversion probabilities.
- 4.Last Activity: Interactions like SMS and Olark chat conversations were significant predictors.
- 5.Lead Origin: Leads originating from "Lead Add Format" showed greater promise.
- 6.Current Occupation: Working professionals were more likely to convert than others.