LEAD SCORING CASE STUDY

By: Ruchika Raju
Sanskar Gupta
Muhammad Rushad Arab

PROBLEM STATEMENT

- X Education, an online course provider, is struggling with a low lead conversion rate of around 30%. The company generates leads through marketing efforts and referrals, but only a small fraction converts into paying customers.

- To optimize sales efforts and improve efficiency, X Education wants to identify high-potential leads (Hot Leads) who are more likely to convert. The goal is to develop a Lead Scoring Model using logistic regression to assign a lead score between 0 and 100.

- The target is to increase the lead conversion rate to 80% by enabling the sales team to focus on the most promising leads.

OBJECTIVES

1. Develop a Lead Scoring Model
   Use logistic regression to assign a lead score between 0 and 100.

2. Improve Lead Conversion Rate
   Identify and prioritize Hot Leads to help the sales team focus on high-potential customers.

3. Handle Data Challenges
   Process missing values, deal with categorical levels like 'Select', and optimize feature selection.

4. Ensure Model Adaptability
   Build a scalable model that can adjust to future business requirements and data changes.

5. Provide Data-Driven Recommendations
   Analyze insights and propose actionable steps to optimize lead conversion strategies.

APPROACH

1. Data Understanding & Initial Exploration

- The dataset was loaded and examined to understand its structure.
- Checked for missing values and handled them based on logical assumptions or statistical measures.
- Dropped unnecessary columns such as "Prospect ID" and "Lead Number" since they were non-informative for prediction.

2. Handling Missing Values

- Columns with very high missing percentages were removed.
- Other missing values were imputed using statistical methods or domain-based assumptions.

3. Handling Categorical Variables

- Categorical variables were converted into numerical form using one-hot encoding.
- Categorical levels like 'Select' (which acted as missing data) were properly handled.

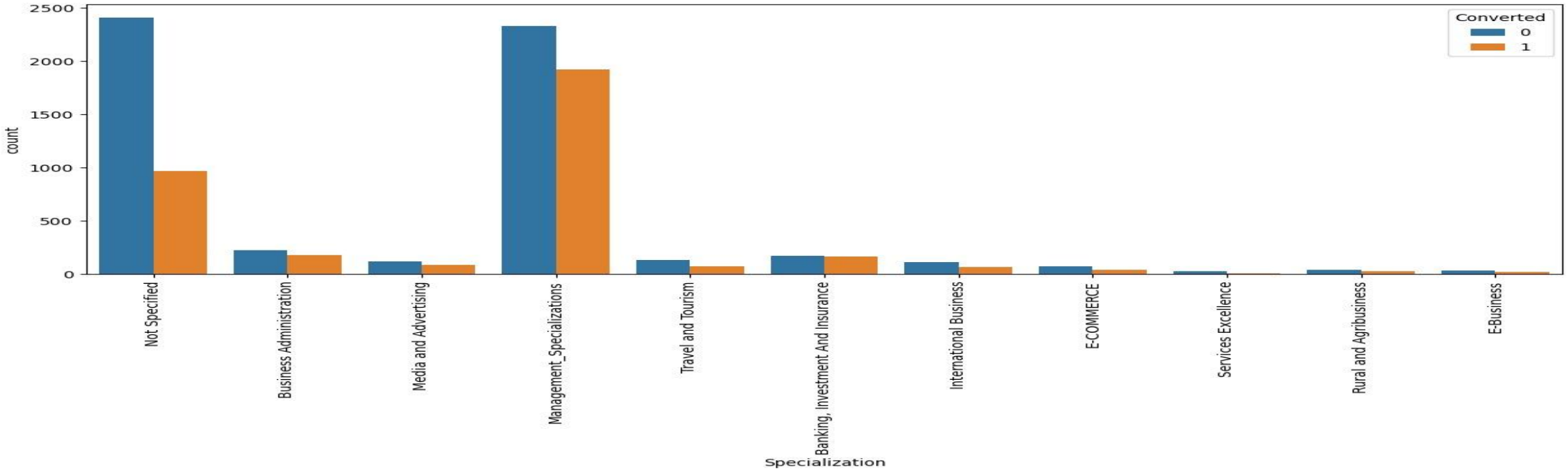
4. Visualizing Data

- Distributions of features were analyzed to understand how different variables affect lead conversion.
- Correlation analysis was performed to find relationships between numerical features and the target variable ("Converted").
- Important categorical features like Lead Source, Last Activity, and Lead Profile were examined to identify trends in conversions.

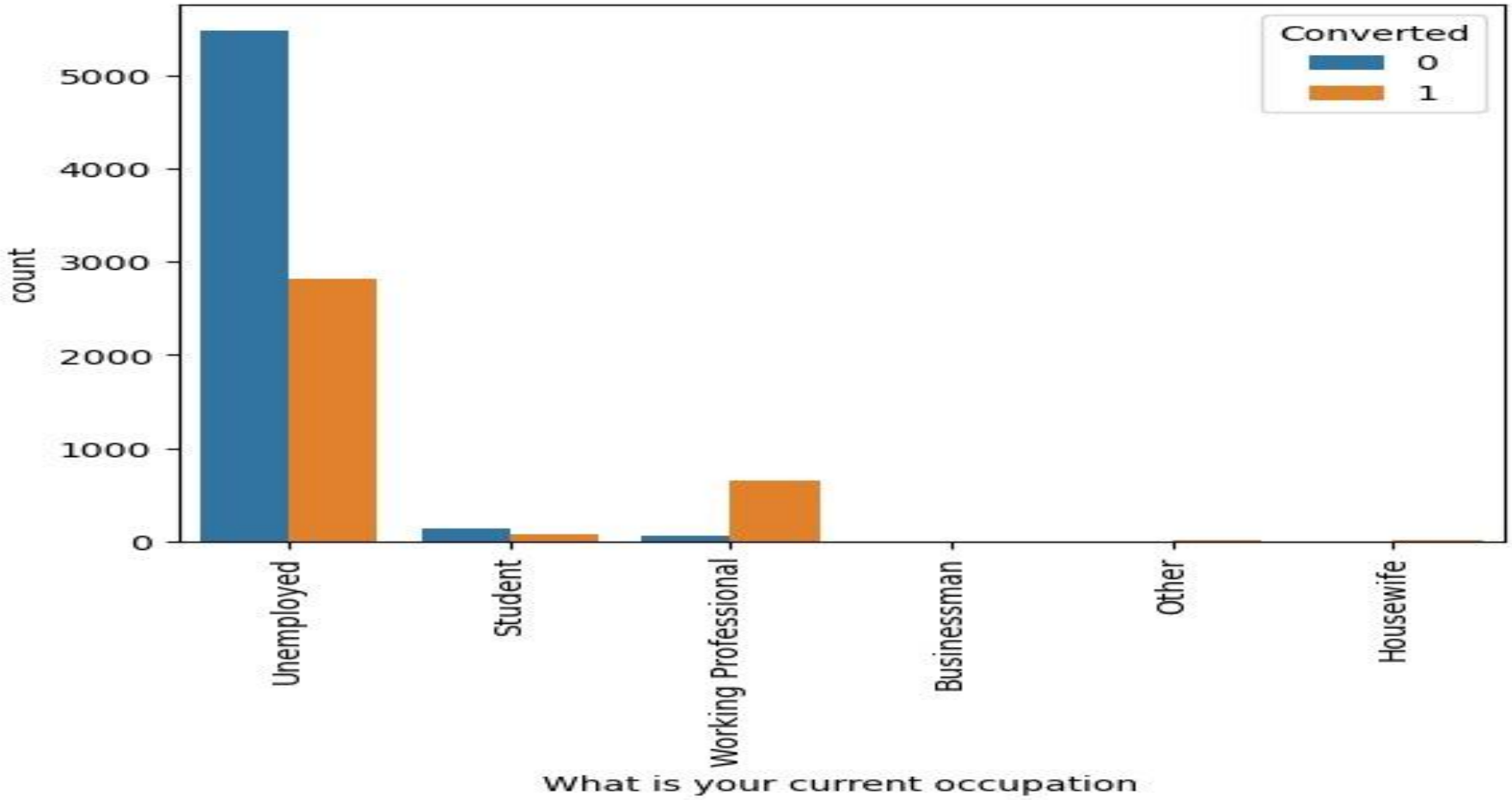
5. Identifying Key Features

- Features that showed a significant impact on lead conversion were prioritized.
- Insights were gathered on attributes like Total Time Spent on Website, Number of Visits, and Engagement Through Emails.

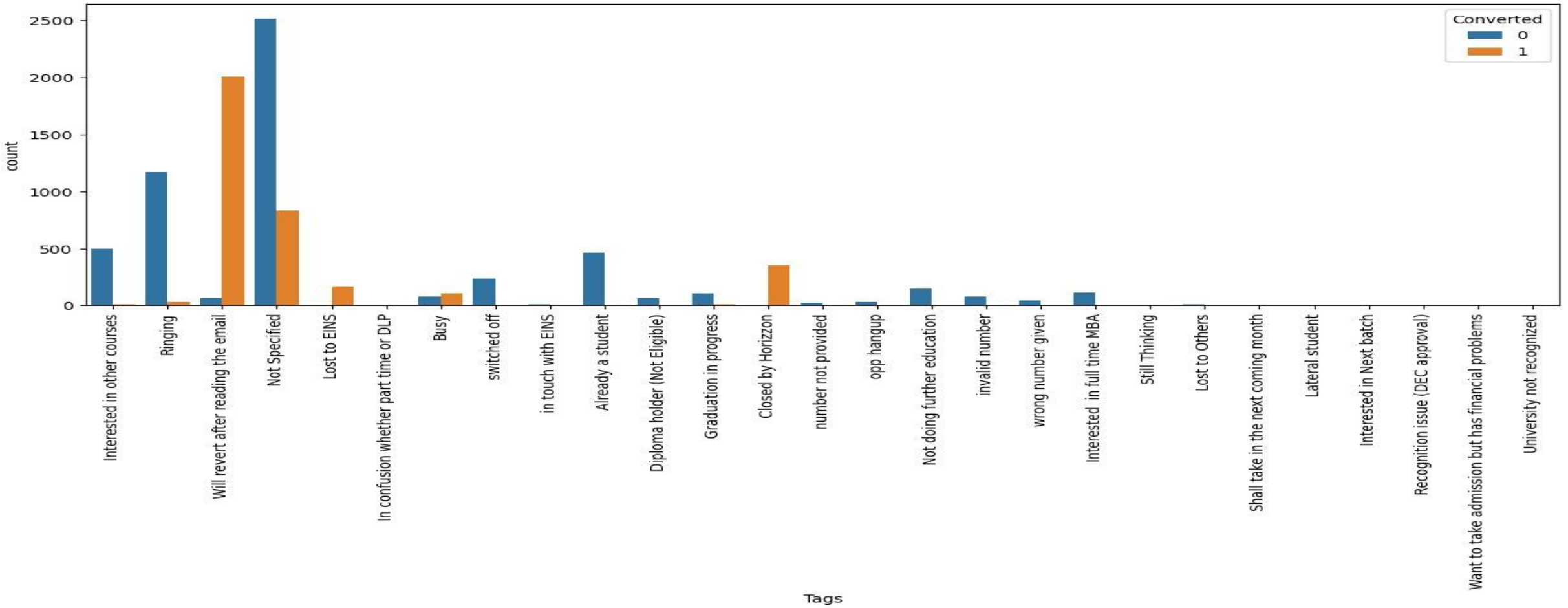
EXPLORATORY DATA ANALYSIS (EDA)



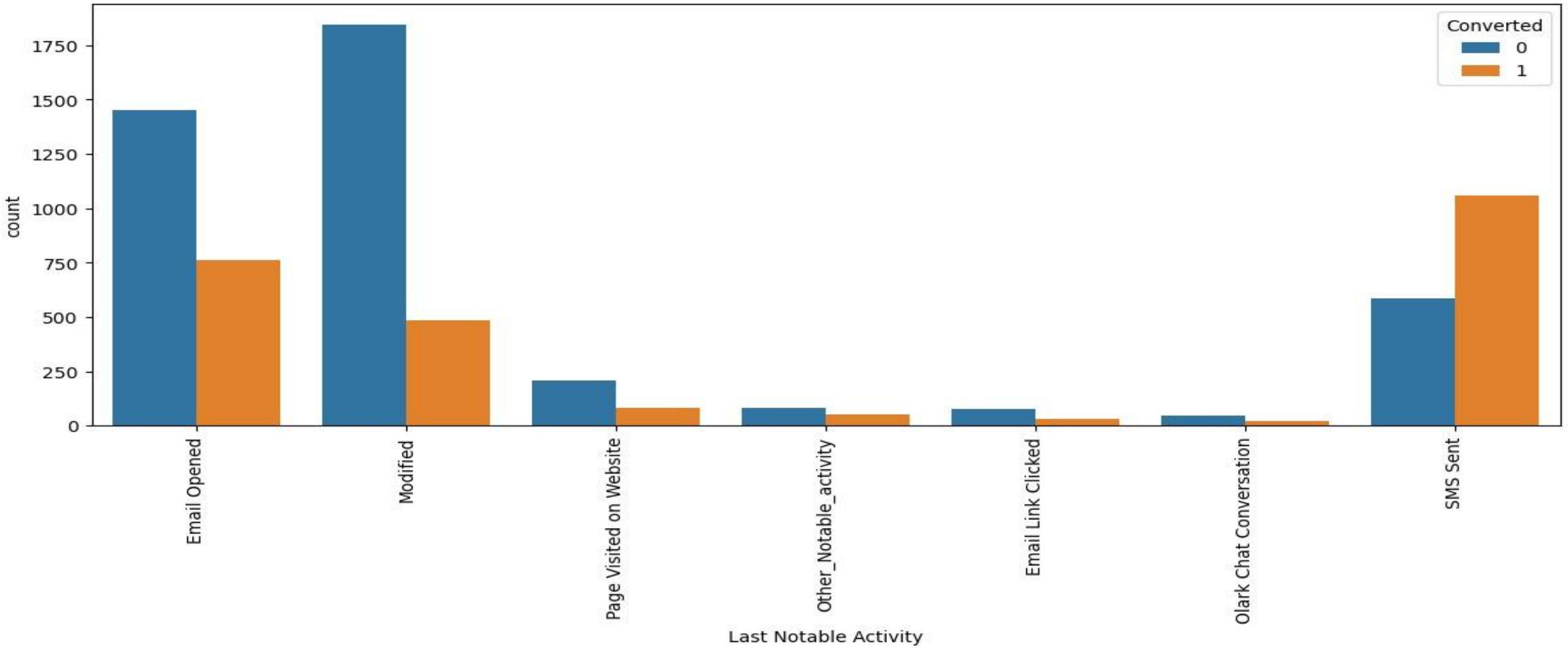
- Management specialisation has the highest number of leads compared to other specialisations
- The conversion rate in management specialisation is not a lot but a good amount of leads are converted in that field



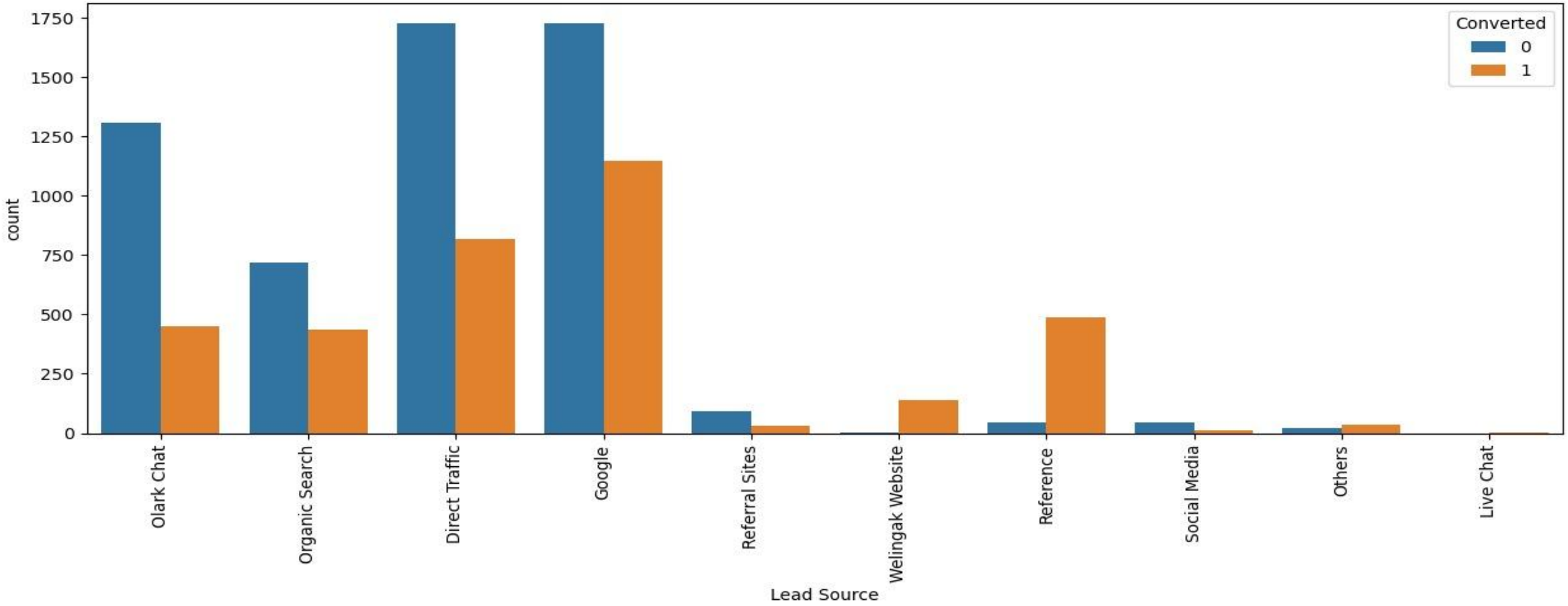
- Unemployed people have the maximum number of leads but the conversion rate is not up to the mark
- Working professionals have the highest amount of leads that are converted



- The tags that are not specified have the highest number of leads
- The tags “Will revert after reading the email” have the highest number of lead conversion rate as they are interested after reading the email.
- Closed by Horizzon and Lost to EINS also has a higher conversion rate compared to other tags



- Email opened and Modified leads have the highest number of leads
- SMS sent sets the bar for the highest number of lead conversion



- Most leads are generated through Google and Direct Traffic.
- The highest conversion rates are from the Welingak website.
- API and Landing Page Submission bring in the highest number of leads and conversions.

Correlation between numeric variables

- Total Time Spent on Website is strongly correlated with conversion. Leads who spend more time on the website have a higher likelihood of converting.

- Higher Total Visits also indicate a higher chance of conversion, but it is not as strong a predictor as time spent.

- Engagement through SMS and email increases conversion rates significantly.

Recommendations

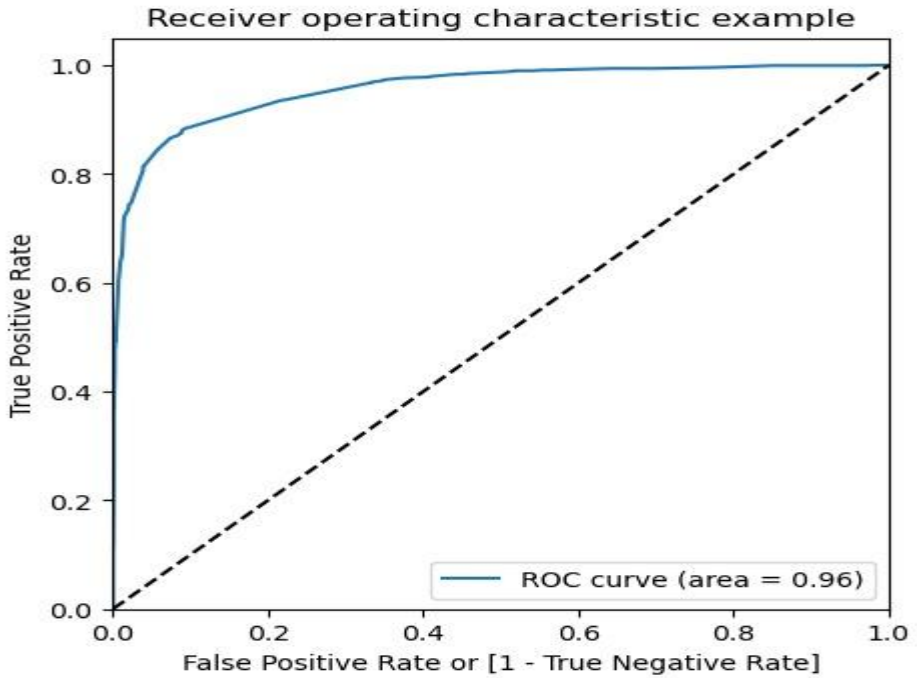
Optimization Strategies for Better Lead Conversion

- Prioritize high-conversion sources (Google, Direct Traffic, Welingak website).
- Increase engagement via SMS and email to improve conversion rates.
- Optimize high-traffic landing pages to improve user retention and lead conversion.
- Simplify the conversion process for leads spending extensive time on the website but not converting.

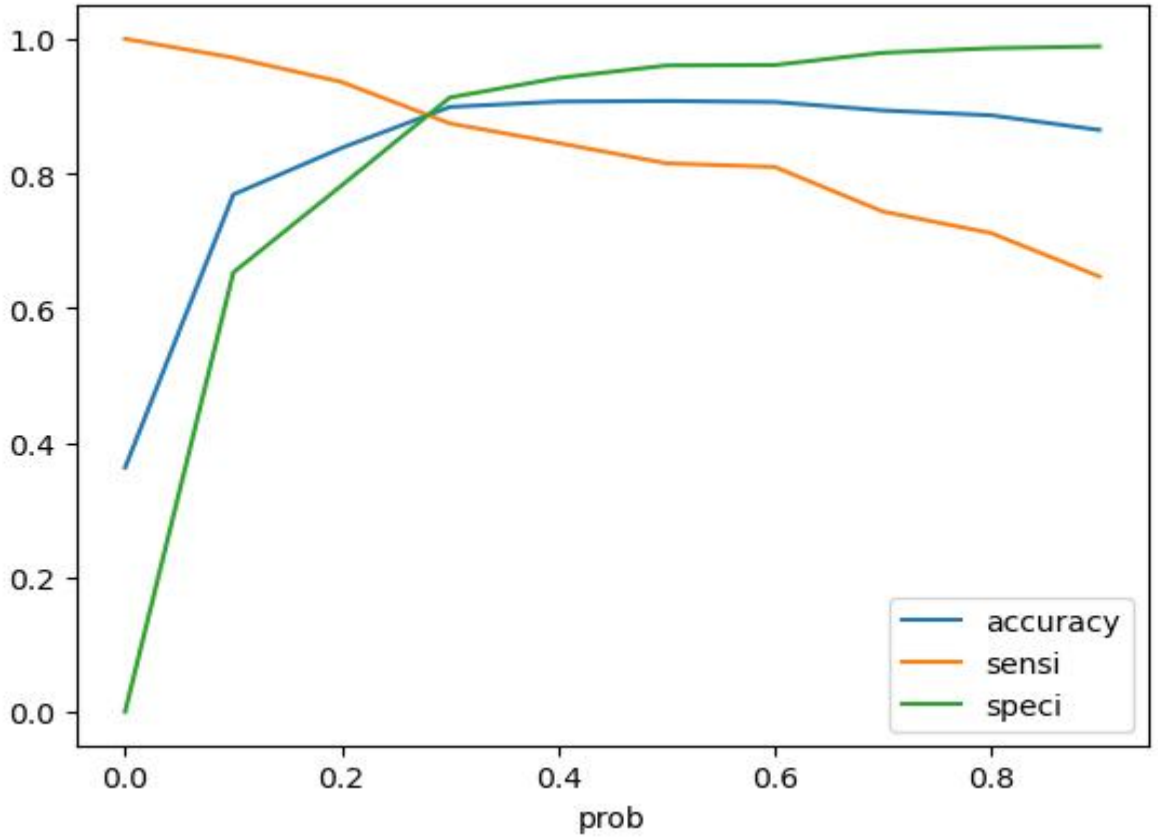
Improving the Lead Scoring Model

- Consider alternative models (e.g., ensemble learning) to further improve recall.
- Fine-tune feature selection to enhance predictive power.
- Implement automated workflows to target high-potential leads.

MODEL EVALUATION INSIGHTS



- The AUC (Area Under the Curve) = 0.96, which indicates an excellent classification performance (since an AUC of 1 is perfect, and 0.5 is random guessing).
- Model is highly capable of distinguishing between converted and non-converted leads.



Need to choose a probability threshold where sensitivity and specificity are well-balanced (in this case, around 0.3 to 0.4).

Train-Test Performance

Training Data:

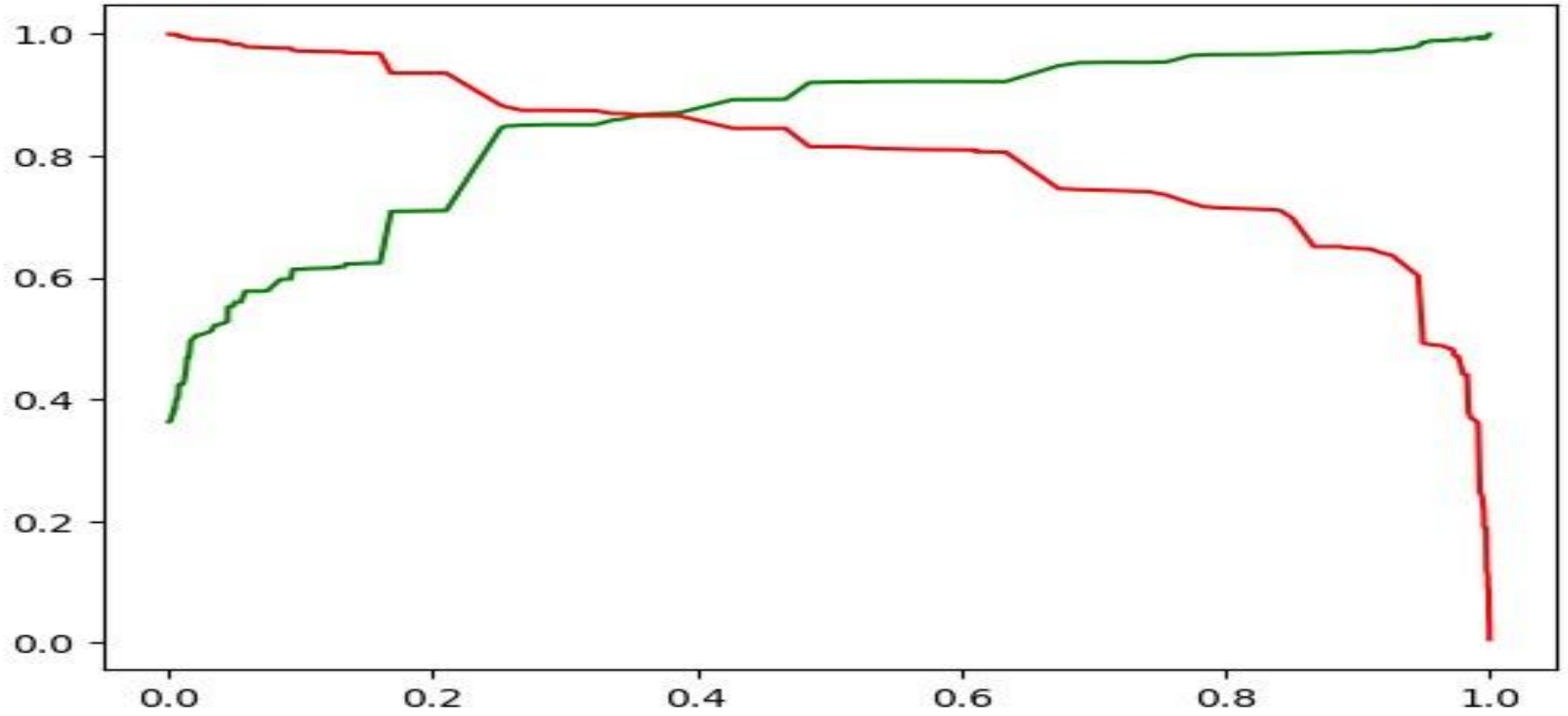
- Sensitivity (Recall) = 87.43% → The model correctly identifies 87.43% of actual conversions.
- Specificity = 91.28% → The model correctly identifies 91.28% of non-converting leads.

Test Data:

- Sensitivity (Recall) = 91.00% → A slightly improved recall, meaning the model captures more conversions.
- Specificity = 90.69% → The ability to correctly classify non-converters remains strong.

Precision, Recall & Model Balance

- Precision (~78%) → When the model predicts a conversion, it is correct 78% of the time.
- Recall (~87-91%) → The model successfully identifies a high percentage of actual conversions.



ROC-AUC Score (~0.96) → The model has high discrimination ability between converting and non-converting leads.

PR Curve (Optimal threshold = 0.44) → Precision (~75%) and recall (~76%) are balanced at this threshold.

(True Positives (TP): Correctly predicted conversions.

True Negatives (TN): Correctly predicted non-conversions.

False Positives (FP): Leads predicted as converting but did not convert.

False Negatives (FN): Leads that converted but were not predicted to do so.)

- The confusion matrix has 2703 True Negatives, 258 False Positives, 212 False Negatives, 1475 True Positives which exhibits good conversion rate.

- The model is well-calibrated and maintains high accuracy (~79-81%) across training and test datasets.

- There is minimal overfitting, as the test performance is comparable to training performance.

- Threshold tuning (0.44) optimally balances precision and recall to focus on high-quality leads.

- Future improvements can focus on boosting recall further while maintaining precision, possibly through feature engineering or ensemble methods.