LEAD SCORING CASE STUDY

LOGISTIC REGRESSION

SUBMITTED BY : **Rajat Sachan**
Arun Challa
Sonmoy Jana

Contents

- Problem Statement
- Problem Approach
- Exploratory Data Analysis
- Model Evaluation
- Observations
- Conclusion

PROBLEM STATEMENT

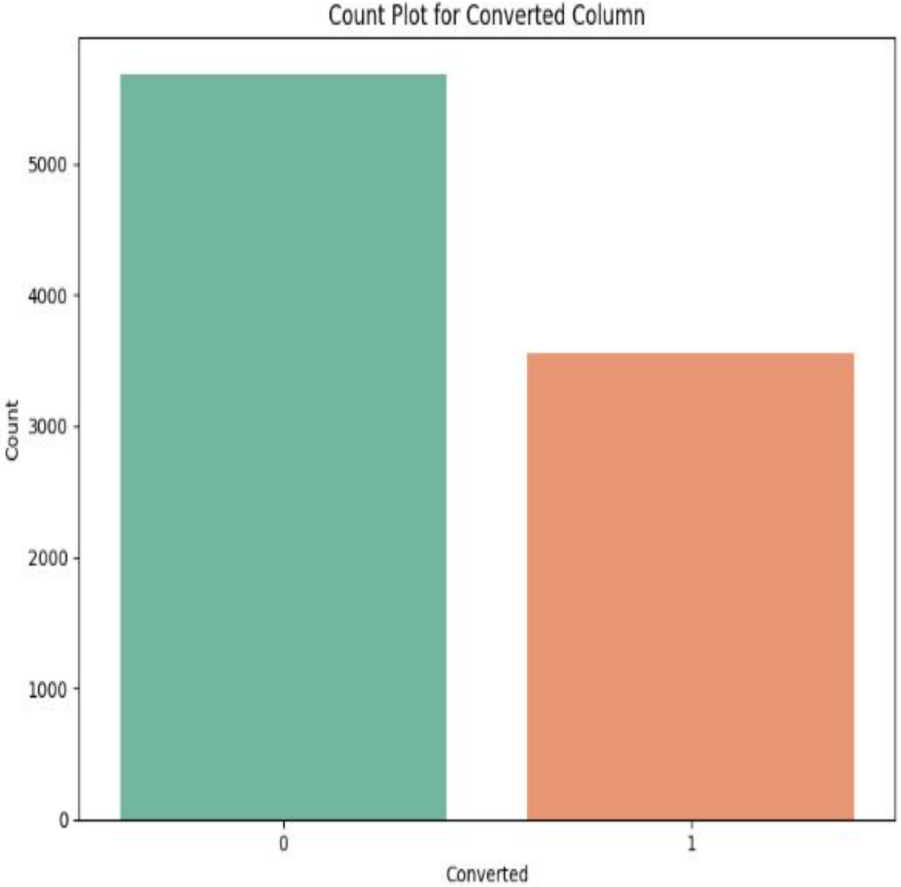
X Education struggles to convert website visitors into paying customers, with a mere 30% conversion rate. They seek your help building a model that scores leads based on their conversion potential, allowing the sales team to prioritize high-value leads and achieve the CEO's ambitious target of 80% conversion. This model will not only identify promising leads but also guide future lead nurturing strategies, optimizing the journey from curious visitor to loyal customer.

Problem Approach

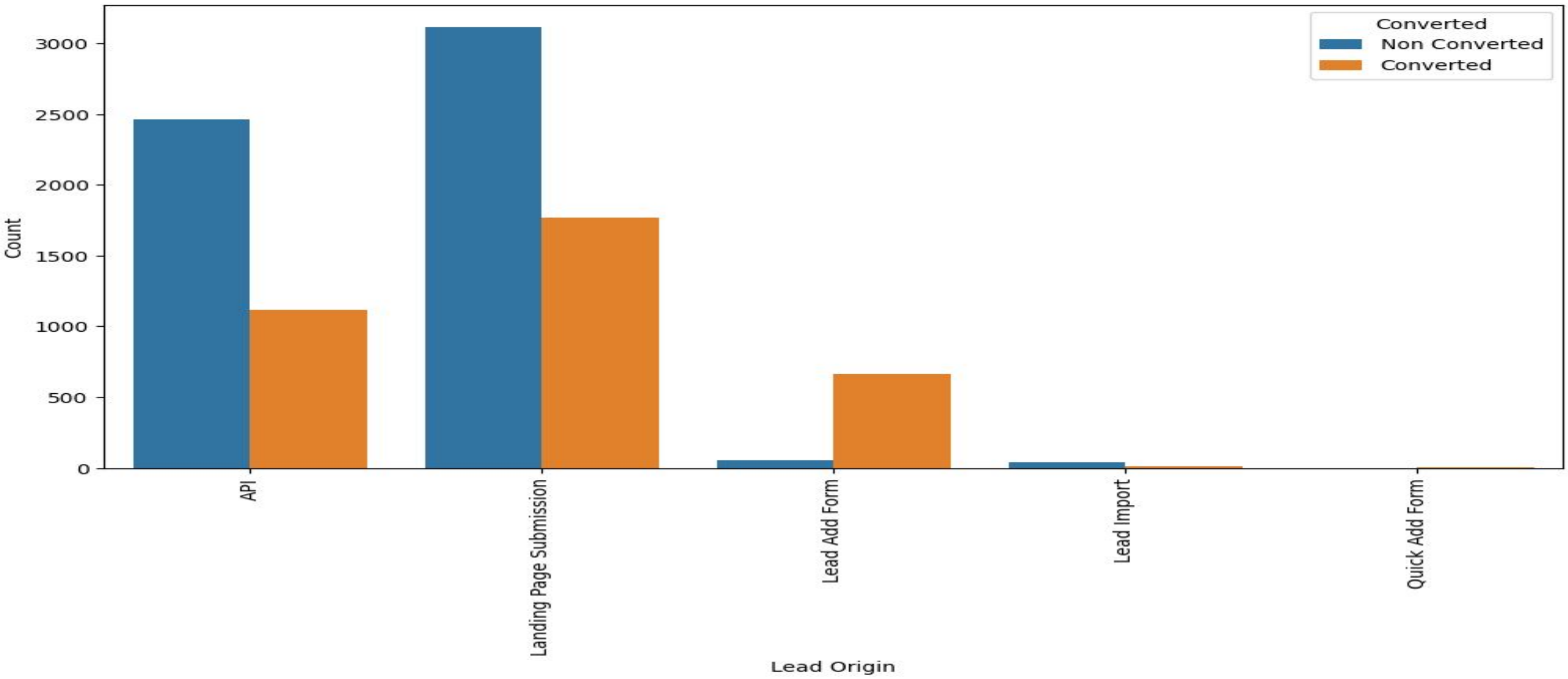
- Exploratory data Analysis
- Cleaning of data
- Scaling of data
- Modelling of data
- Evaluation using various metrics
- Change of threshold
- Applying the technique to test set

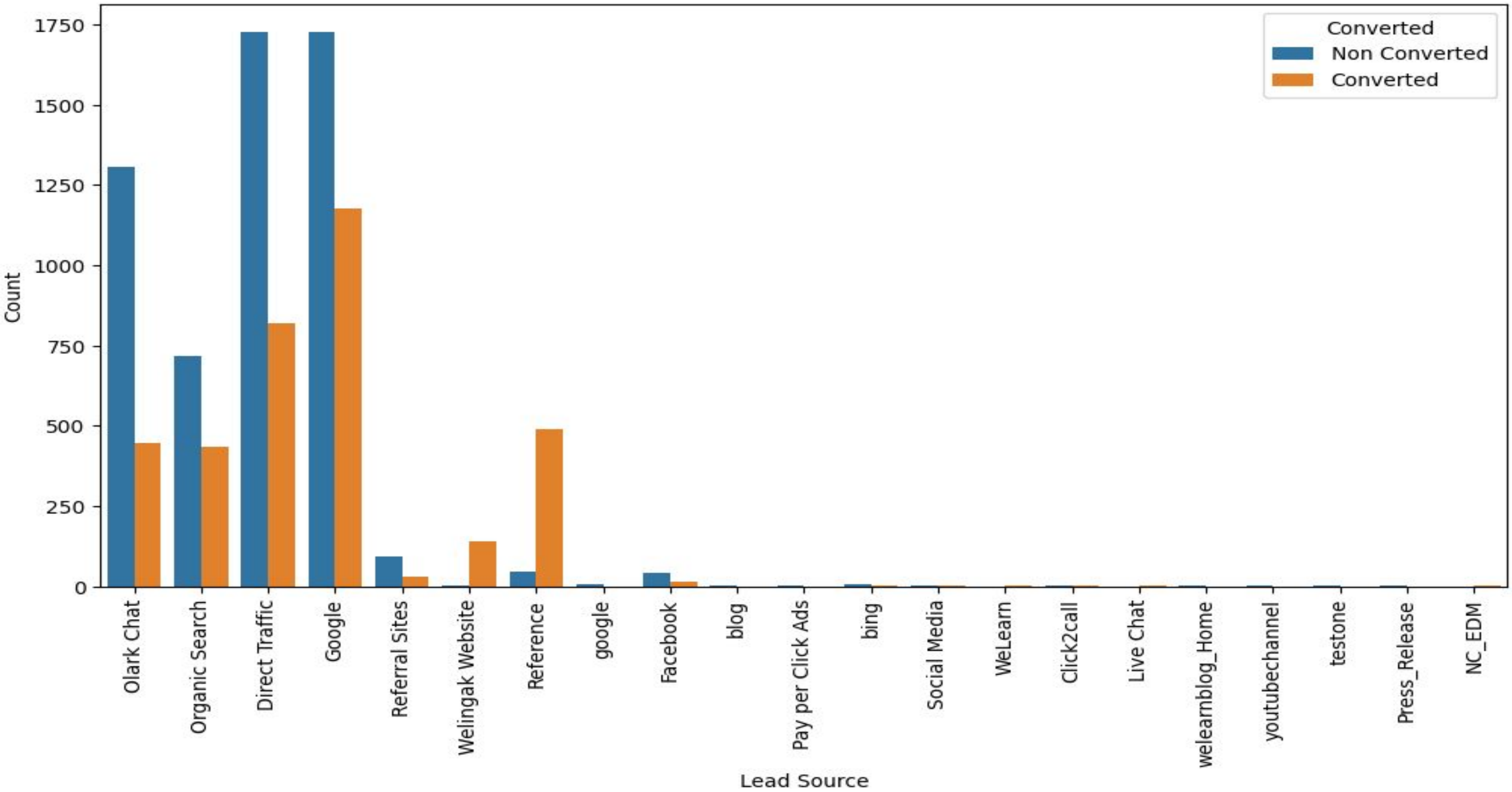
Exploratory Data Analysis

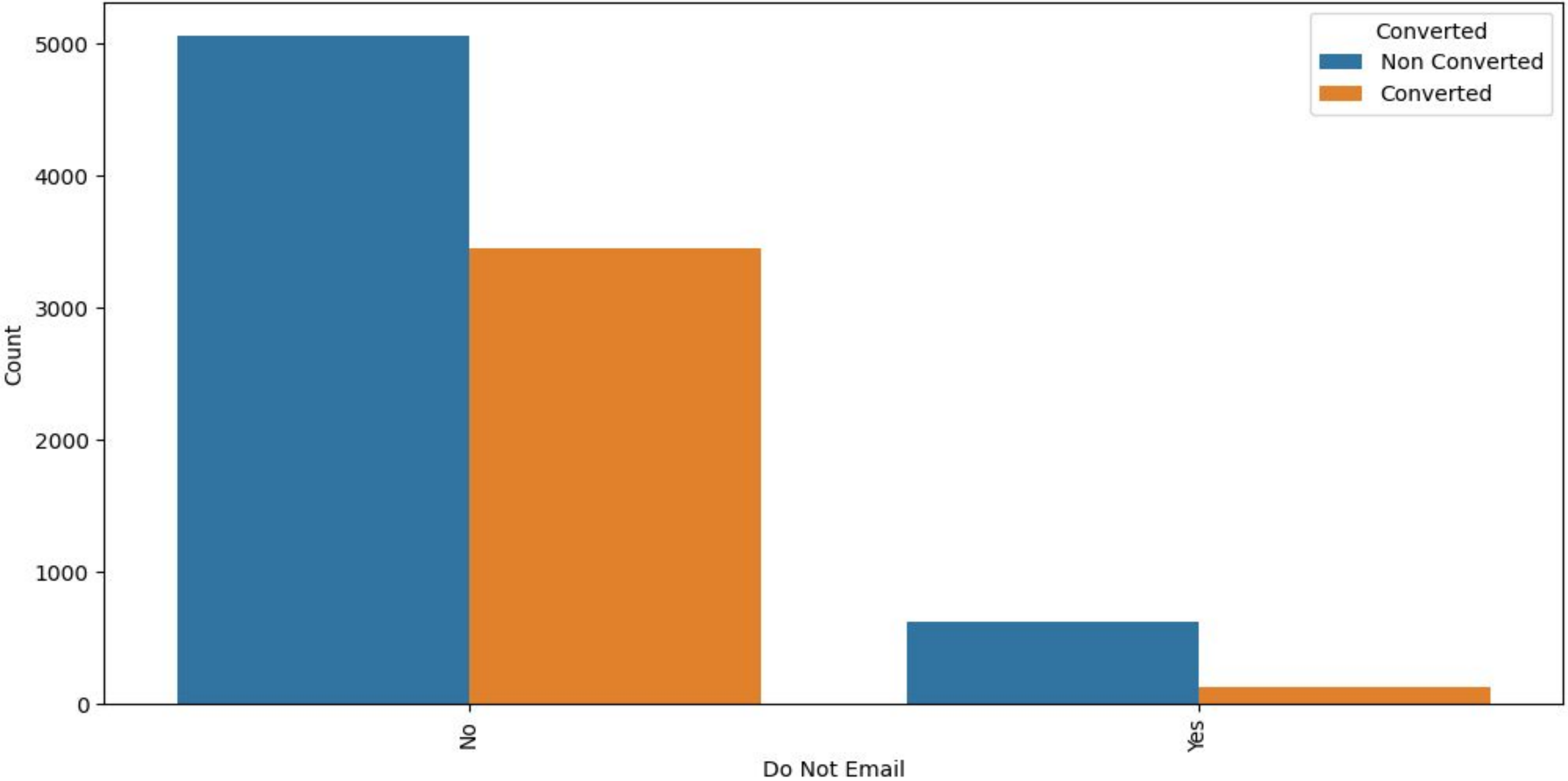
We have 9240 unique customer entries and aim to identify those with the highest likelihood of conversion. The decision criteria involve categorizing potential leads based on their Leads Score, representing the probability of conversion. Notably, approximately 37% of leads are converted, while 73% are not.

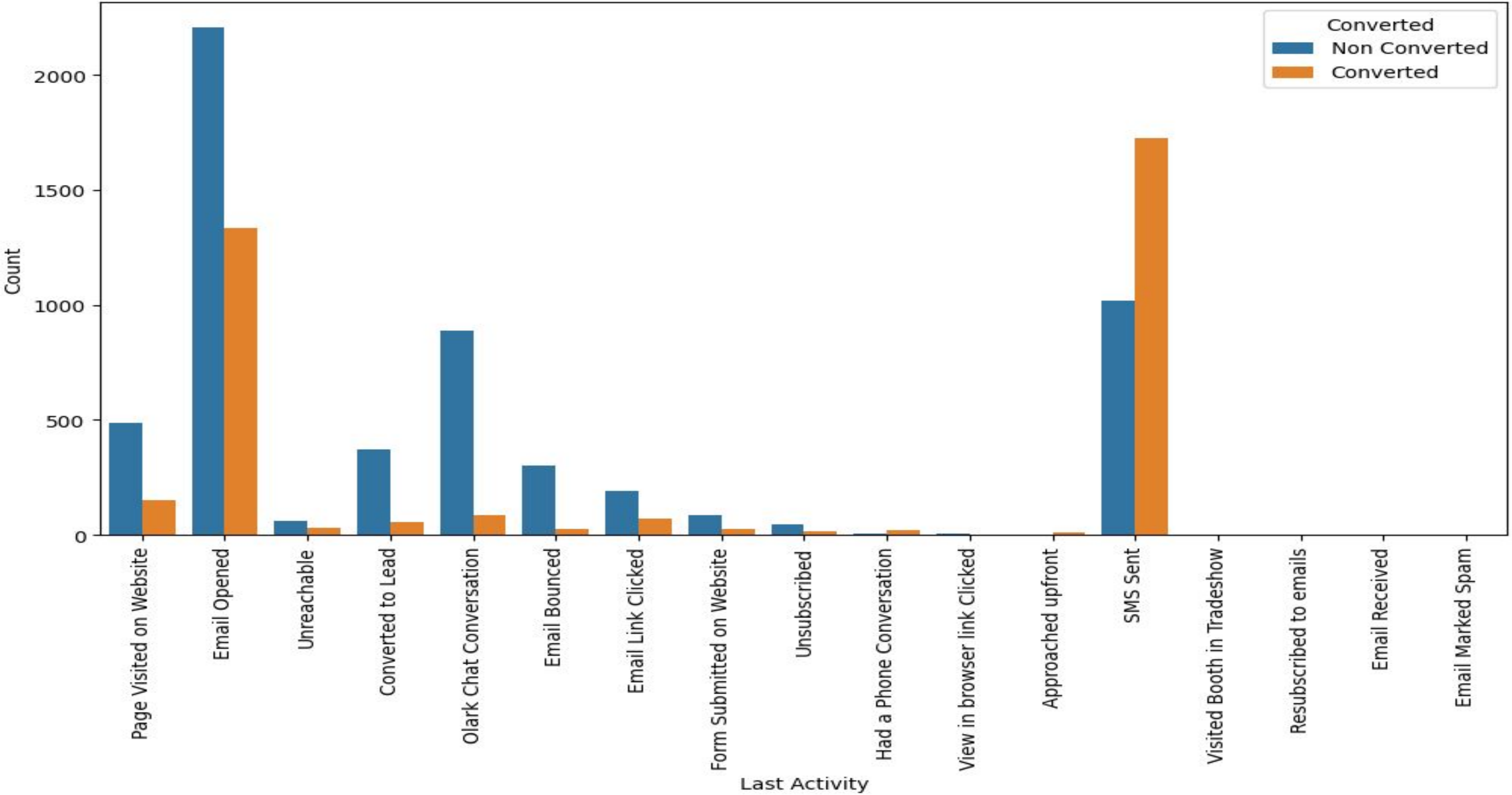


Data Visualization of Categorical Columns

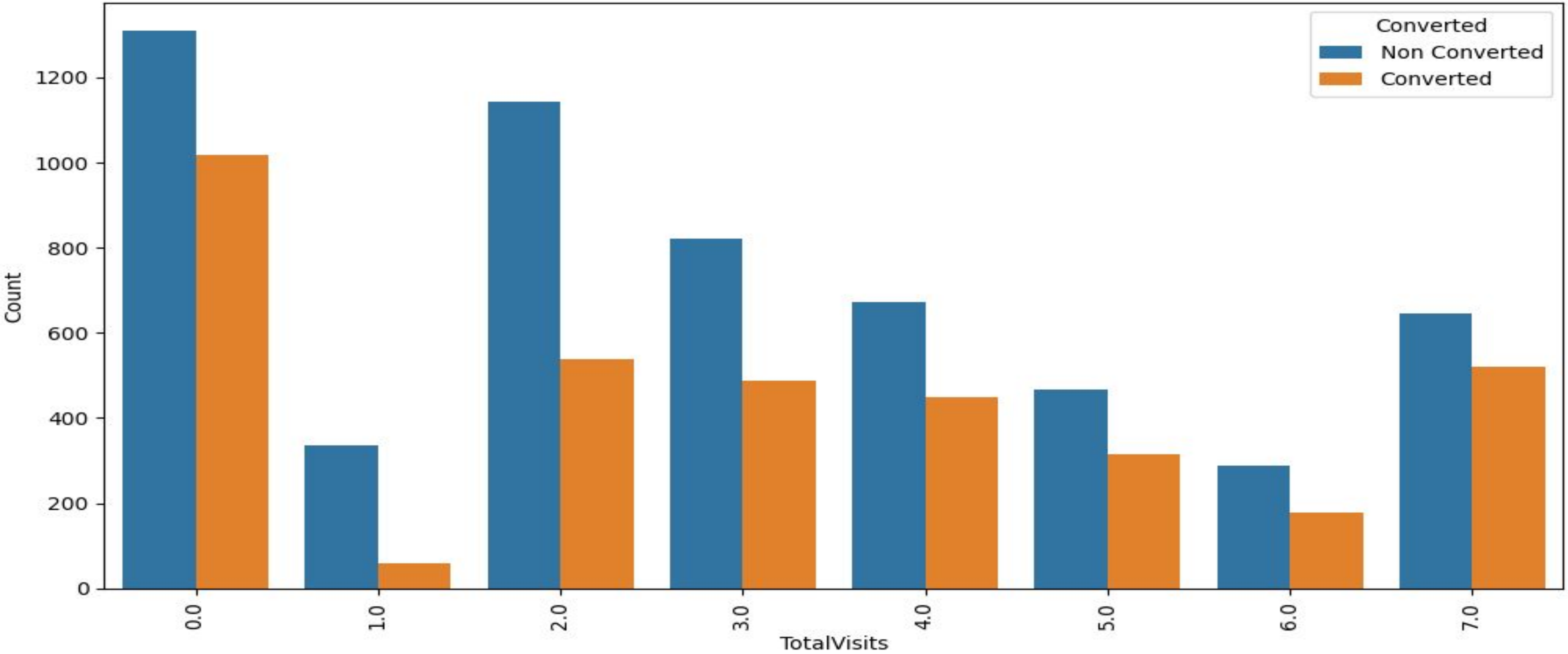


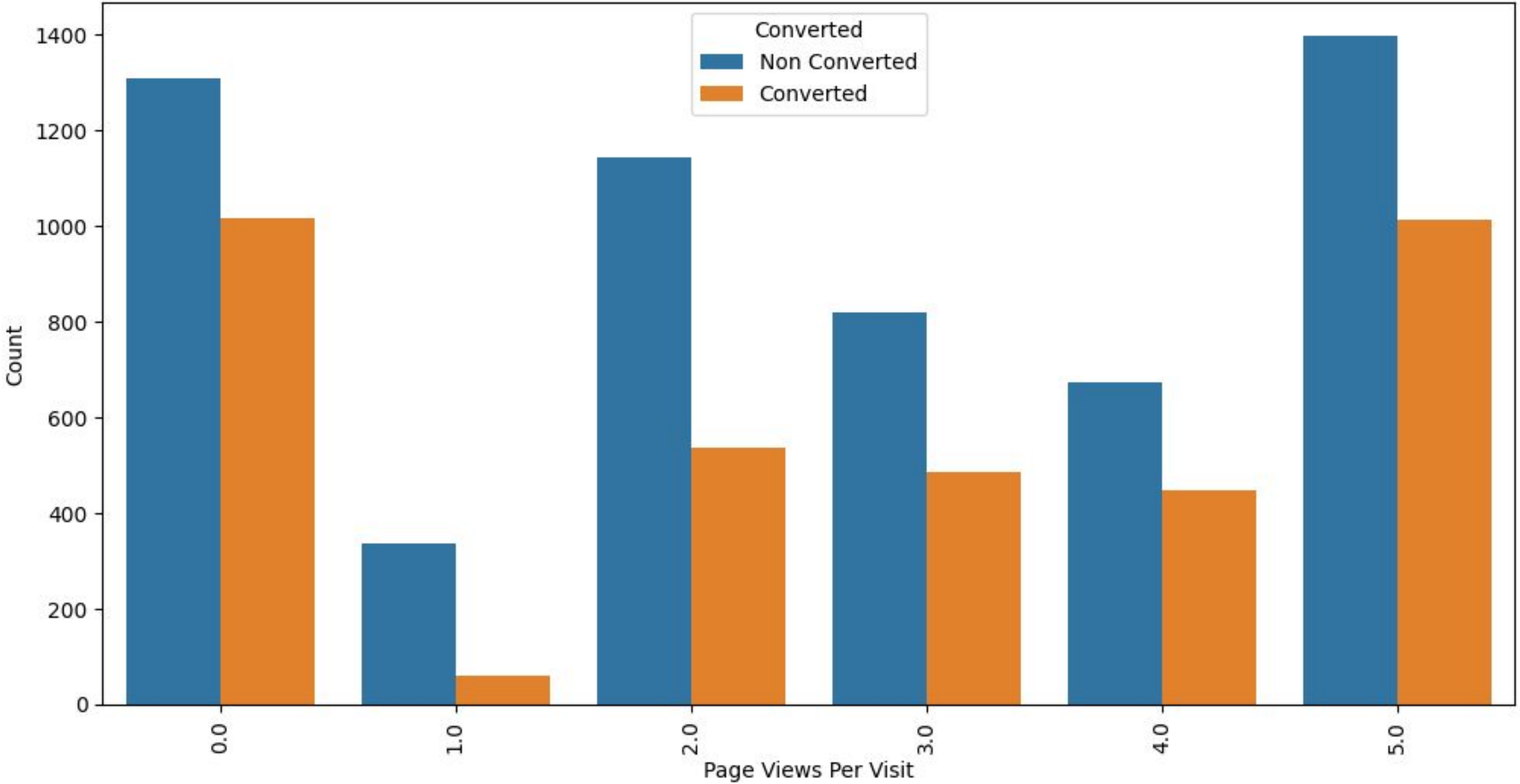

Converted
Non Converted
Converted
1750
1500
1250
1000
750
500
250
0
Count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


Converted
Non Converted
Converted
5000
4000
3000
2000
1000
0
Count
No
Yes
Do Not Email


Converted
Non Converted
Converted
Count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

Data Visualization of Numerical Columns




Converted
Non Converted
Converted
Count
1400
1200
1000
800
600
400
200
0
0.0
1.0
2.0
3.0
4.0
5.0
Page Views Per Visit

Model Evaluation

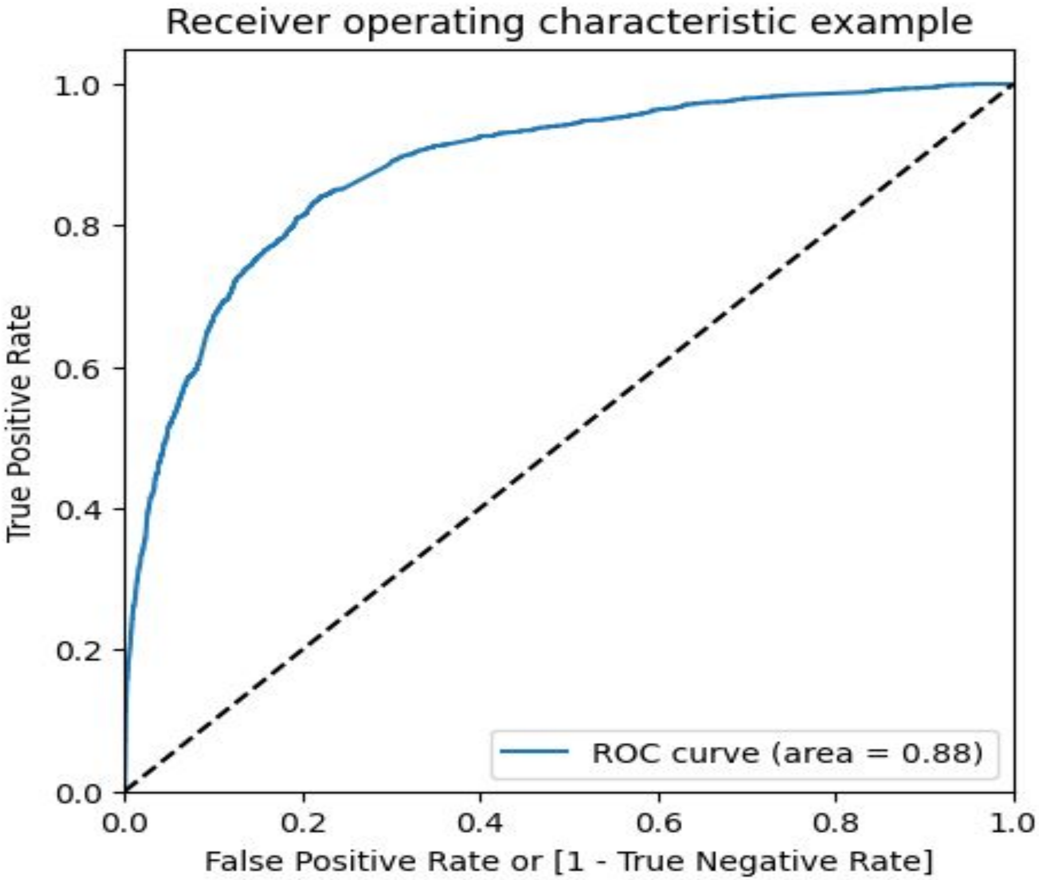
The model underwent rigorous evaluation to assess its performance and ensure its suitability for target lead prediction.

Key metrics including accuracy, sensitivity, specificity, ROC curve analysis, and precision/recall tradeoff were examined. The model achieved an accuracy of 80.8% on the test set, exceeding the target prediction of 80%. Notably, it demonstrated both high sensitivity (82%) and specificity (79%), indicating reliable capture of both converted and unconverted leads.

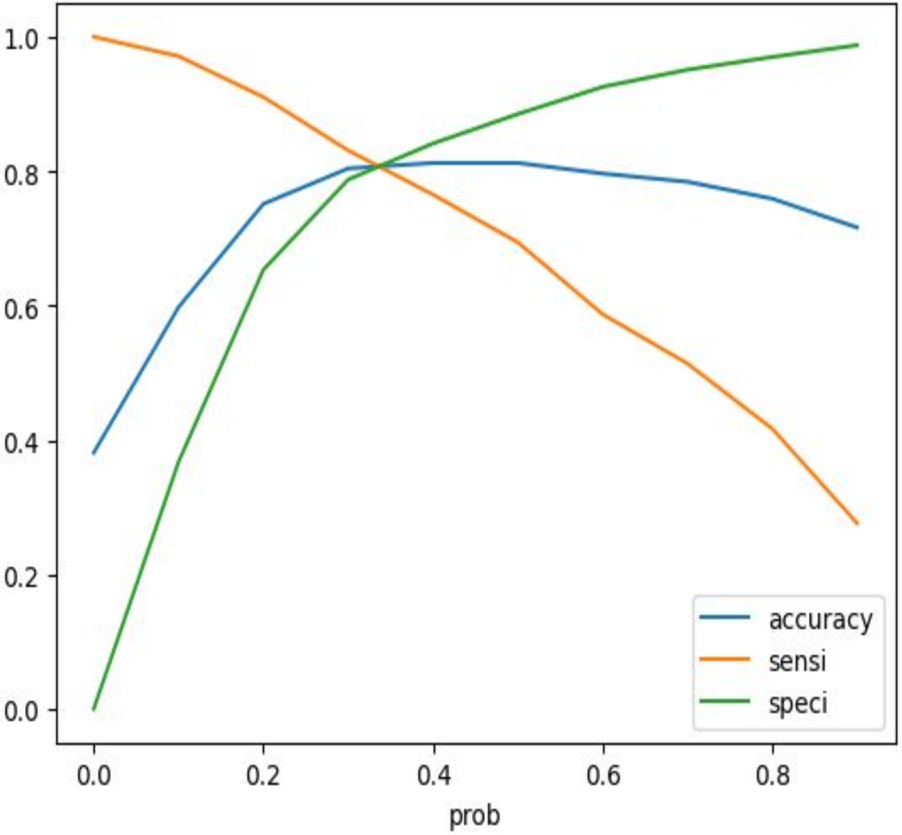
Furthermore, feature selection via RFE identified the top 15 most influential features, ensuring model interpretability and focusing on critical factors. Additionally, analysis of the ROC curve and precision/recall metrics optimized the cutoff probability threshold to balance lead prediction accuracy with desired sensitivity and specificity levels.

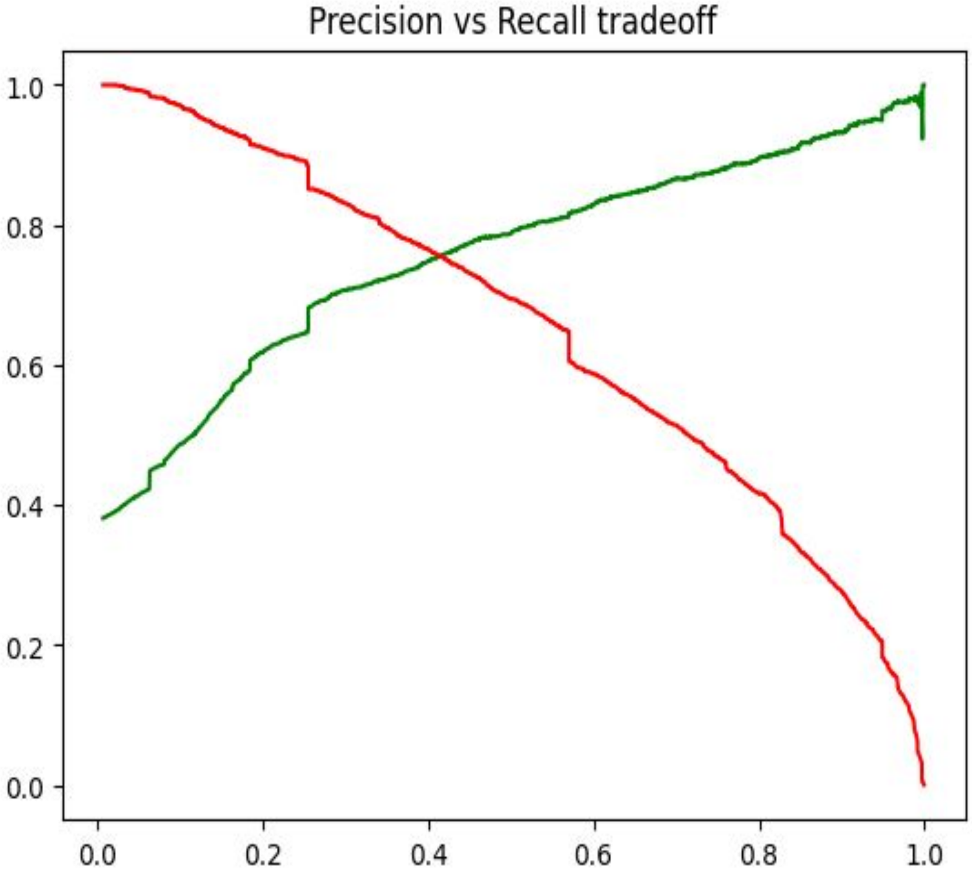
ROC Curve

The ROC curve boasts an impressive area under the curve (AUC) of 0.88, indicating excellent model discrimination between positive and negative cases.



Precision and Recall





Observations

Train Data:

- Accuracy : 80%
- Sensitivity : 83%
- Specificity : 78%

Test Data:

- Accuracy : 80%
- Sensitivity : 82%
- Specificity : 79%

Final Features list:

- Total Time Spent on Website
- Do Not Email
- What is your current occupation
- Lead Source_Reference
- Lead Source_Olark Chat
- Lead Origin_Landing Page

Conclusion

- ❖ Accuracy, Sensitivity, and Specificity: All metrics on the test set closely matched those of the training set, hovering around 80% and 82%, indicating robust generalizability.
- ❖ Optimal Cutoff Point: Analyzing Sensitivity and Specificity led to an optimized probability threshold, further boosting predicted conversion rates to 80% (train) and 79% (test).
- ❖ Feature Selection: Top influential features were identified, shedding light on key drivers of lead conversion. The top three include total website time, lead origin from Lead Add Form, and having a recent phone conversation.
- ❖ Business Adaptability: The model exhibits flexibility to adjust to evolving company needs and campaign strategies.

These findings suggest a reliable and insightful tool for optimizing lead conversion efforts. Focusing on leads from API and Landing page submissions, targeting website engagement, and prioritizing specific lead origins and activities can further enhance lead nurturing strategies.