MM.DD.YY

Lead Scoring Case Study

Data-Driven Insights and Model-Based Strategies
Submitted By –Divvy Pratap Singh, Yajat S, B. Sravan Kumar

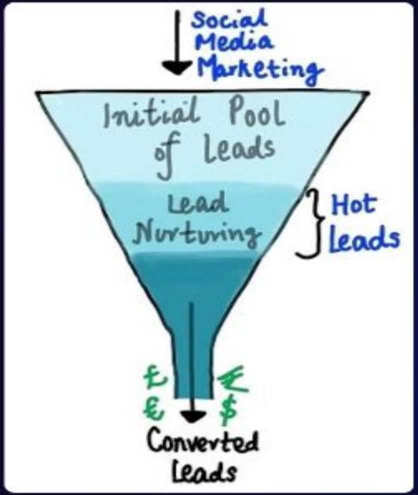
Lorem Ipsum

Business Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Although the company successfully generates numerous leads through various marketing channels, a significant portion of these leads do not convert into paying customers. As the sales team spends considerable time contacting leads that may not be viable, there is an urgent need to enhance the efficiency of the lead conversion process.



Specifically, the company faces the challenge of identifying high-potential leads, referred to as ‘Hot Leads,’ which are more likely to convert into sales. The current funnel demonstrates a large number of potential leads at the top, but only a small fraction successfully progresses to becoming customers at the bottom of the funnel.

Business Objective

1. Improve Lead Conversion Rate

The foremost objective is to increase the lead conversion rate from **30% to 80%** by effectively identifying and prioritizing **‘Hot Leads.’**This enhancement should lead to an **increase in revenue** generated from course sales.

3. Data-Driven Decision Making

Establish a **data-driven approach** to lead management by utilizing **historical data and predictive modeling**.This will allow the company to make **informed decisions** regarding marketing strategies and customer engagement.

2. Optimize Sales Efforts

By developing a **scoring system** for leads, the company aims to streamline the sales process.Sales personnel will invest their **time and resources in leads** that are statistically more likely to convert, hence reducing the time spent on **less promising leads** and improving overall productivity.

Scalability and Future Adaptability

Build a **flexible logistic regression model** that can be adjusted to accommodate evolving **company requirements and market conditions**.The model should be robust enough to **reassess lead scoring criteria** as new data and metrics become available.

Comprehensive Reporting

Deliver a **well-structured report and presentation** that summarizes the **methodology employed, results obtained, and actionable insights** derived from the analysis.

This documentation should facilitate **stakeholder understanding and buy-in** for the new **lead prioritization** approach.

1. Data Collection & Preprocessing

- Thoroughly analyzed the dataset to understand its structure and content.
- Cleaned the provided leads dataset, addressing any null values represented by ‘Select’ in categorical variables.
- This step is crucial for ensuring the data’s integrity.

2. Exploratory Data Analysis (EDA)

- Conducted an EDA to understand the relationships between various attributes (Lead Source, Total Time Spent, Last Activity, etc.) and the lead conversion outcome.
- Used visualizations to interpret these relationships.

3. Model Development

- Identified the most relevant features using Recursive Feature Elimination (RFE).
- Built a logistic regression model utilizing the cleaned dataset to predict lead conversion probabilities.
- The model assigns a score to each lead, indicating its likelihood of converting into a paying customer.

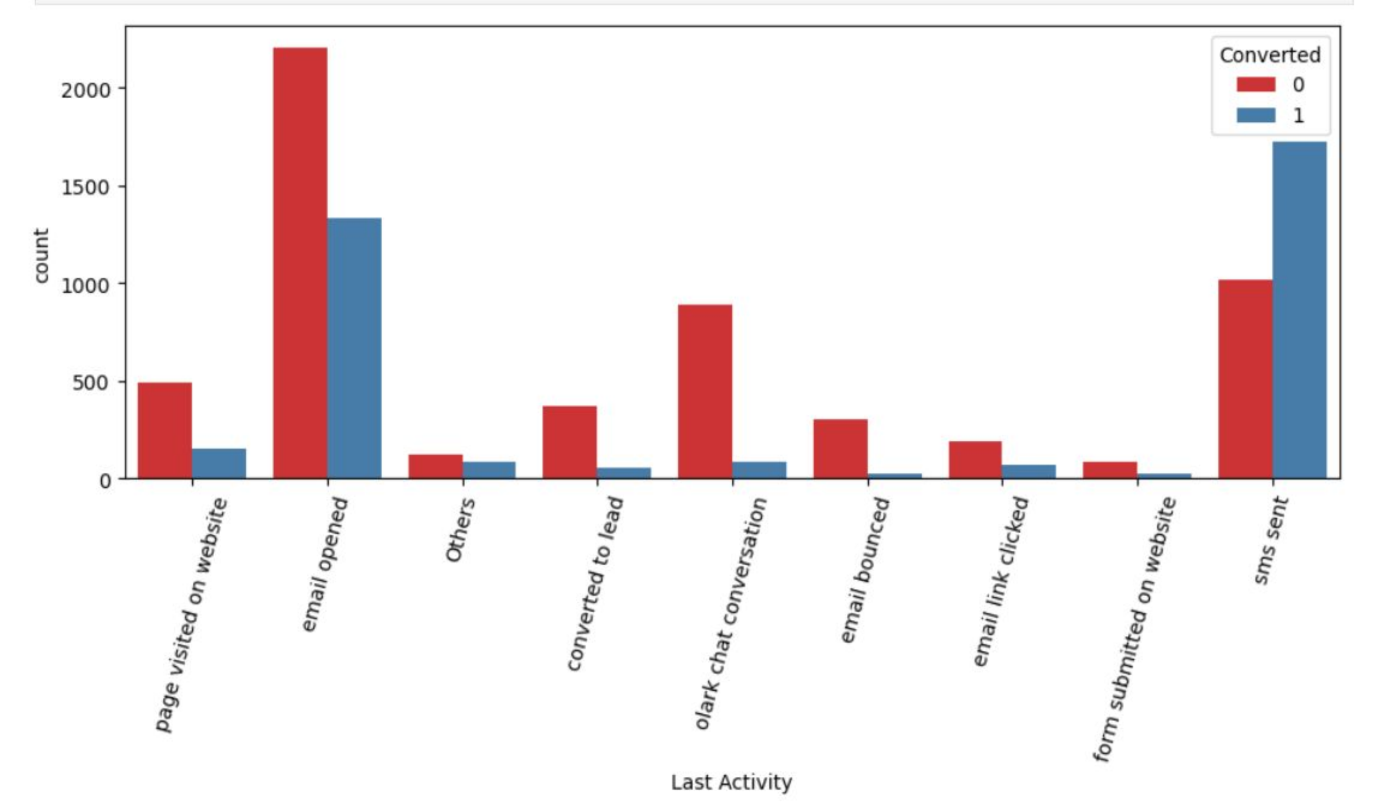
4. Model Evaluation

- Assessed model performance using accuracy, sensitivity, specificity, and the ROC-AUC curve.
- These metrics helped determine the model's effectiveness in distinguishing between hot and cold leads.

5. Recommendations and Reporting

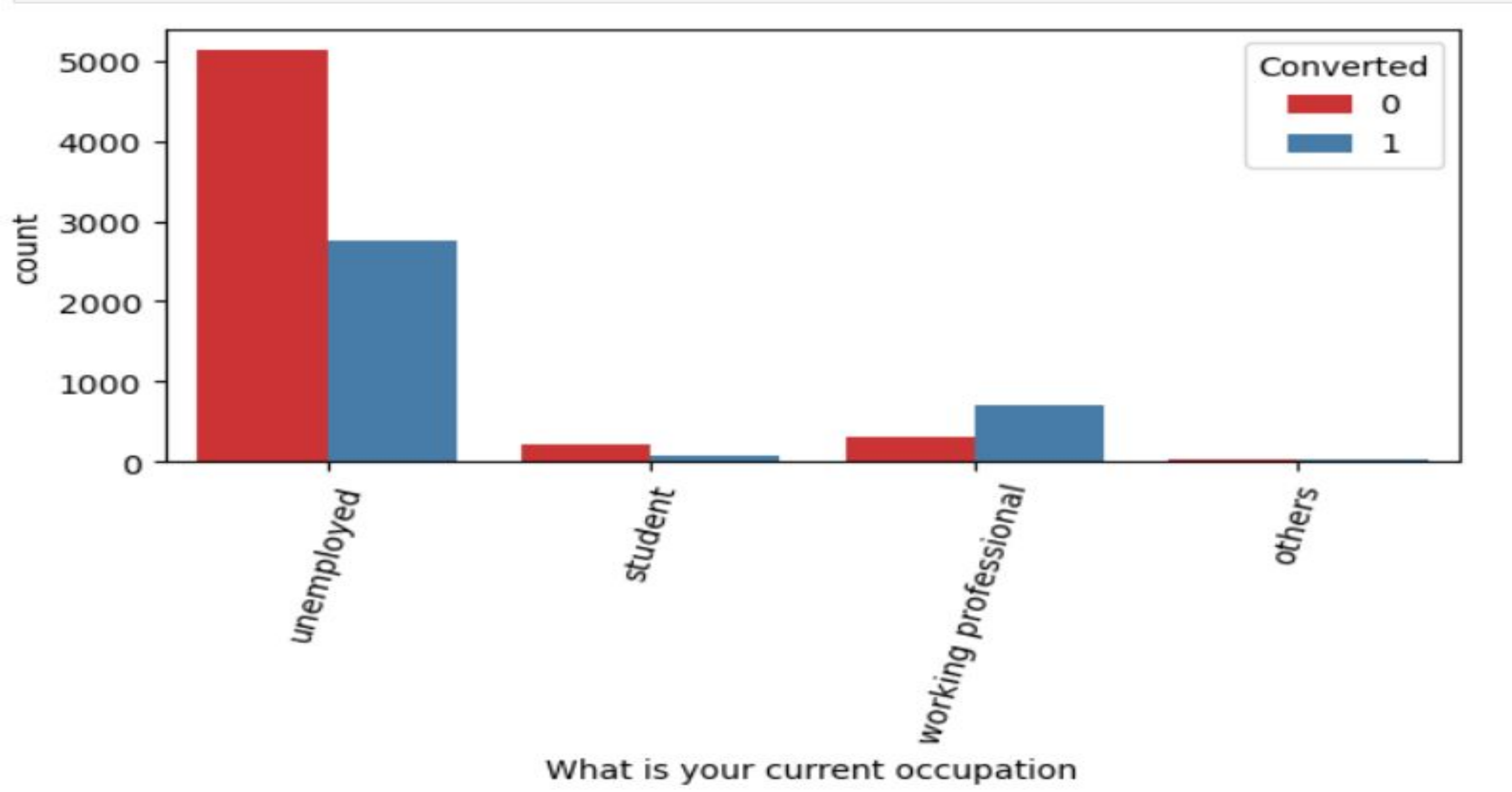
- Prepared a detailed report summarizing the analysis process, results, and recommendations for the sales team.
- Included visualizations to aid in understanding key insights and findings.

EXPLORING DATA ANALYSIS



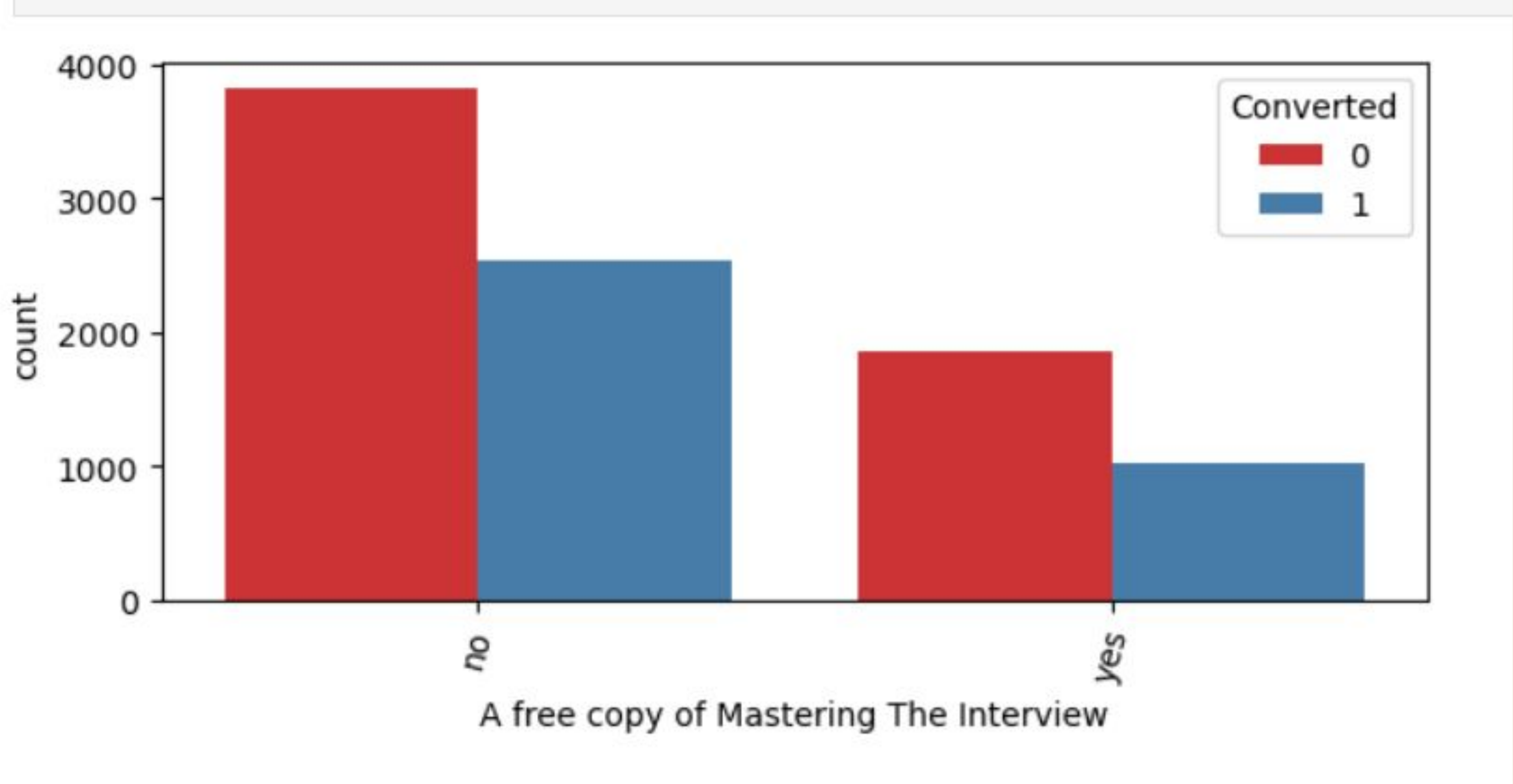
Key Insights:

- Most leads had "Email Opened" as their last activity.
- "SMS Sent" has a significantly higher conversion rate compared to other activities.
- Activities like "Olark Chat Conversation" and "Page Visited on Website" show fewer conversions.



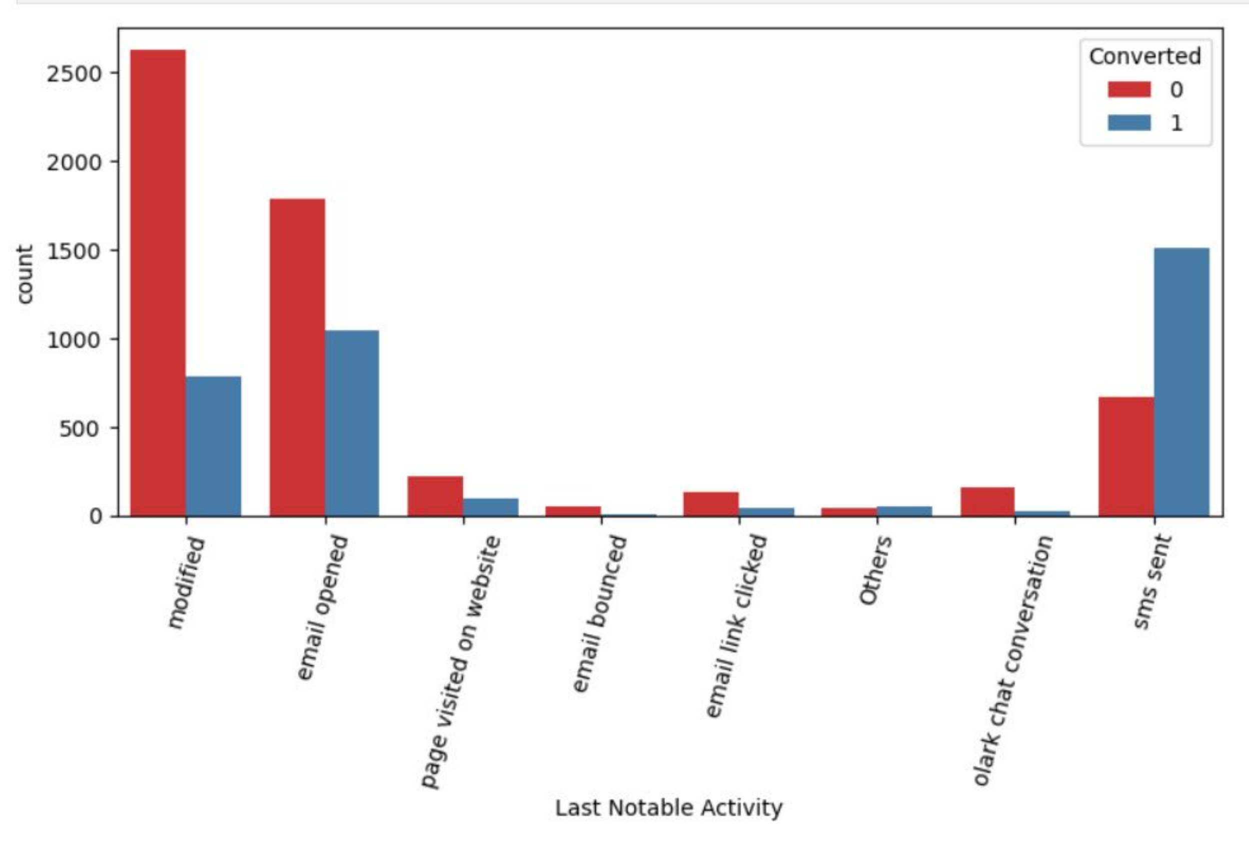
Key Insights:

- Majority of leads are **Unemployed**, but they have lower conversion rates.
- **Working Professionals** have a higher conversion rate.



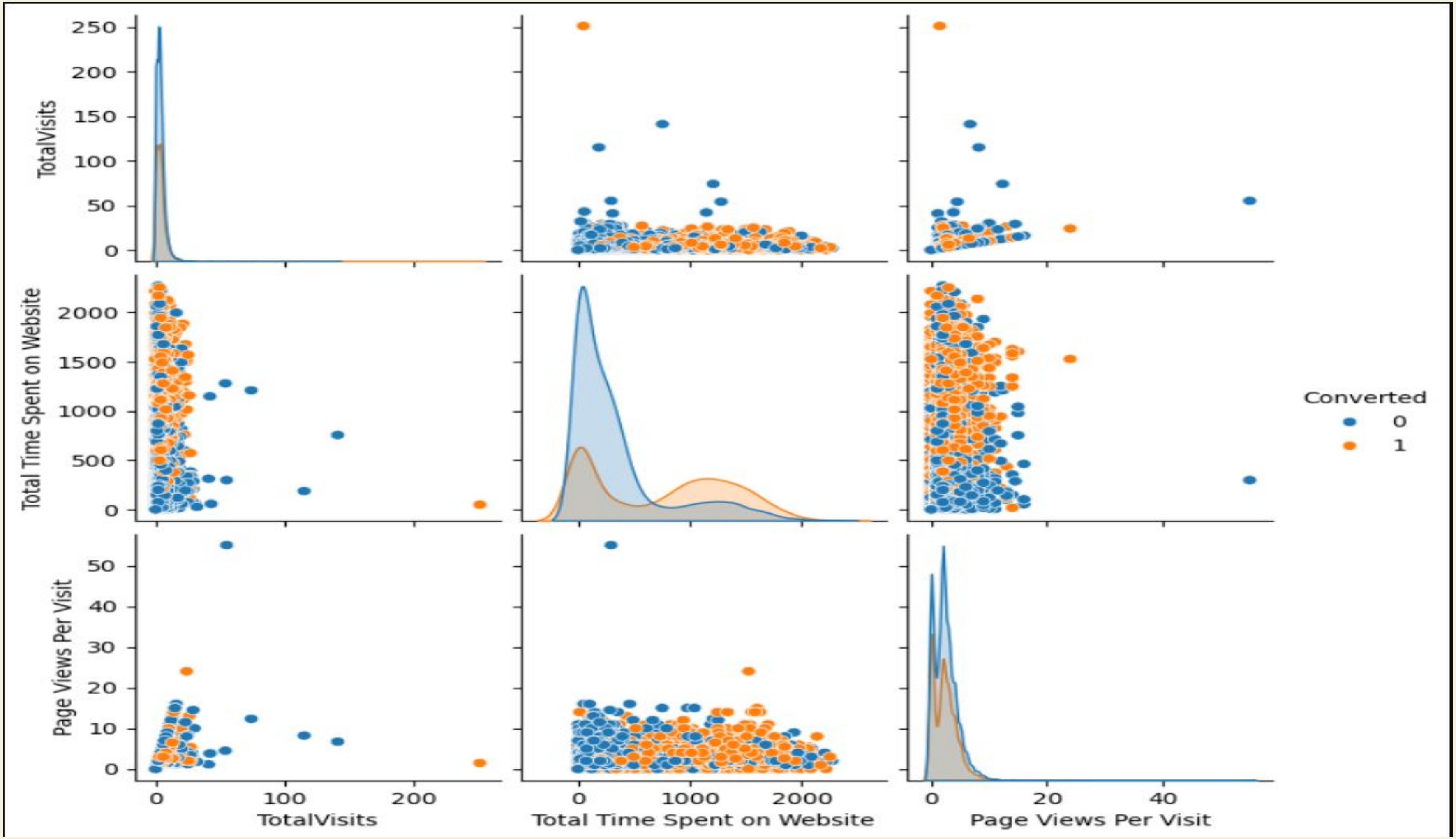
Key Insights:

- Leads who received a **free copy** had a lower conversion rate.
- People who did **not** receive the free copy converted at a higher rate.



Key Insights:

- **Modified leads** and **email opened** dominate last notable activities.
- "SMS Sent" remains a strong driver of conversions.


TotalVisits
Total Time Spent on Website
Page Views Per Visit
0
50
100
150
200
250
500
1000
1500
2000
10
20
30
40
Converted
0
1

Higher Total Visits & Time Spent Correlate with Conversions

- Leads who converted (orange points) are often found in the **upper-right region**, indicating that users who visit the website frequently and spend more time have a higher chance of conversion.

Time Spent on the Website is a Stronger Predictor than Page Views

- Even if leads do not navigate many pages per visit, those who spend **more time** on the site have a **higher probability of conversion**.
- This suggests that engagement depth (time spent) matters more than breadth (number of pages viewed).

Most Leads Have Low Page Views Per Visit

- The majority of leads, regardless of conversion status, view **only a few pages per visit**.
- This implies that users prefer concise, well-structured content rather than navigating through multiple pages.

Conversion Occurs Even with Fewer Pages Viewed

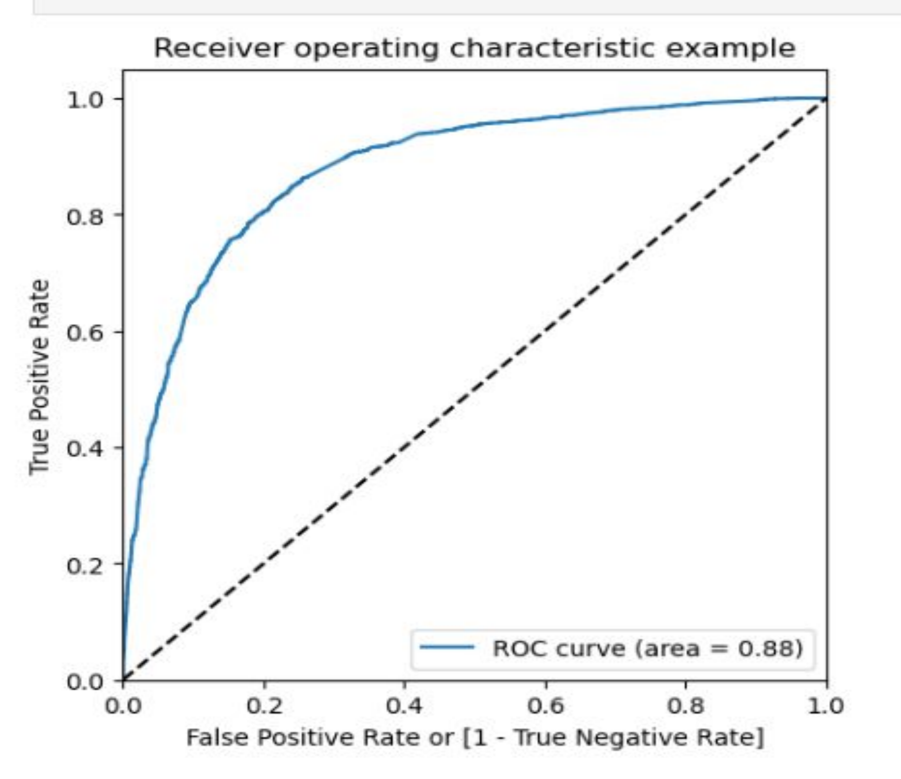
- Leads who convert **may not explore too many pages**, but they engage deeply with important content.
- This emphasizes the need to optimize key pages with relevant, compelling information.

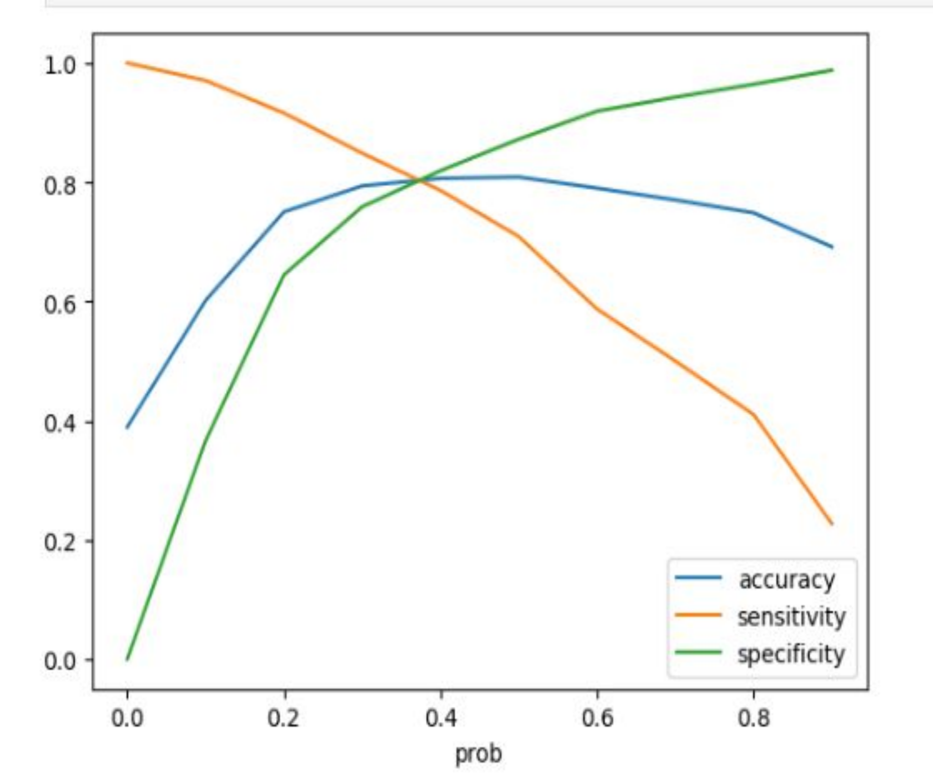
A Small Segment of Leads Spends Extensive Time and Still Doesn’t Convert

- Some leads spend a lot of time on the website without converting, suggesting potential **hesitation** or **barriers** (e.g., unclear CTAs, pricing concerns, or lack of trust).

Optimization Strategies for Better Lead Conversion

- Focus on **enhancing engagement** on high-traffic pages.
- Ensure that key landing pages **hold user attention** and drive action.
- Reduce distractions and **simplify the conversion process** for users spending a long time without converting.


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.88)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensitivity
specificity
prob
0.0
0.2
0.4
0.6
0.8
1.0

Model Evaluation Key Insights

The model achieves an overall accuracy of around **79-81%**, indicating strong performance with slight overfitting.

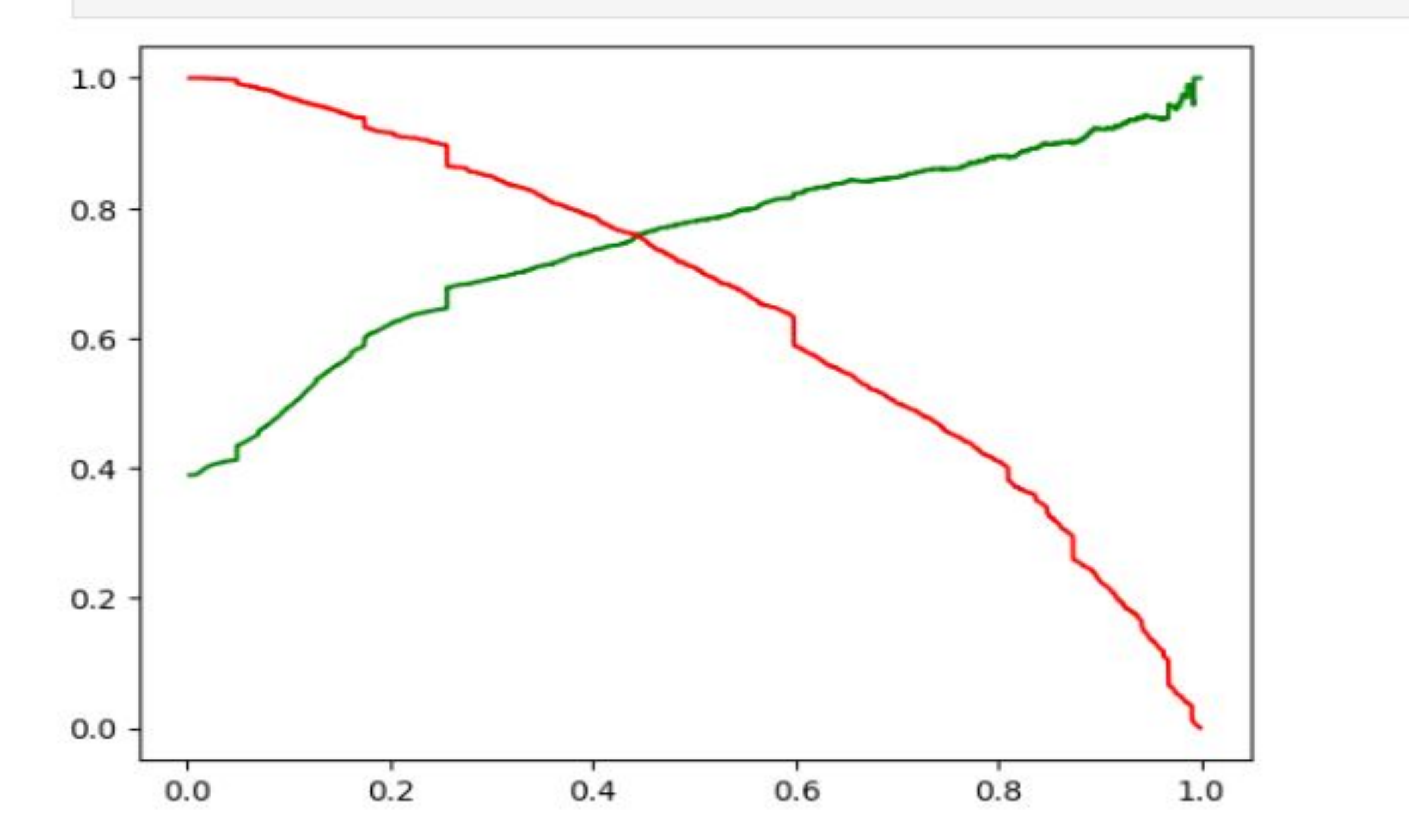
Sensitivity (recall) is lower than specificity, meaning it correctly classifies more non-converters than converters.

Initially, the model had lower sensitivity, but after **threshold tuning (0.36)**, it improved while maintaining balanced accuracy and specificity.

Precision is **~78%**, meaning when the model predicts a conversion, it's correct most of the time, but recall remains at **~71%**, indicating some missed conversions.

The **PRC curve (0.44)** confirms that the model is well-calibrated, and after tuning, both **test sensitivity (~81%) and specificity (~80%)** are balanced.

Future improvements should focus on boosting recall without sacrificing precision, possibly through better feature engineering or alternative models.


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

The Precision-Recall curve helps determine the optimal threshold, and **0.44** was chosen for a balanced precision (~75%) and recall (~76%).

The **final test accuracy is 79.47%**, with a slight drop from training accuracy, indicating a well-generalized model.

The confusion matrix shows **1,435 true negatives, 768 true positives, 294 false positives, and 275 false negatives**, maintaining a good balance between precision and recall.

Precision (72%) suggests that when the model predicts a conversion, it is correct **72% of the time**, while **recall (73%)** means **73% of actual conversions were correctly identified**.

Comparing the ROC and PR curve-based thresholds (**0.36 vs. 0.44**), the **0.44 threshold maintains higher precision while keeping recall balanced**.

Overall, the model performs well, and future improvements could focus on **further increasing recall without reducing precision**, possibly through advanced feature engineering or ensemble learning.

Thank
You