



# Lead Scoring Case Study 2023

Created By

Vandana Parashar  
Pragnya Paramita  
Reena



# Problem Statement

- X Education offers online courses tailored for professionals in various industries. The company leverages multiple online platforms to promote these courses, including websites and popular search engines like Google.
- When individuals visit the company's website, they can explore available courses, submit course inquiry forms, or engage with instructional content, such as videos. These individuals are categorized as potential leads upon submitting a form with their contact details, such as email addresses or phone numbers. Additionally, X Education also acquires leads through referrals from previous customers.
- Subsequently, the company's sales team initiates a series of outreach activities, which may involve phone calls, email communications, and more. Some of the leads transition into paying customers throughout this engagement process, while the majority do not. X Education typically achieves a lead conversion rate of approximately 30%.



# Goals

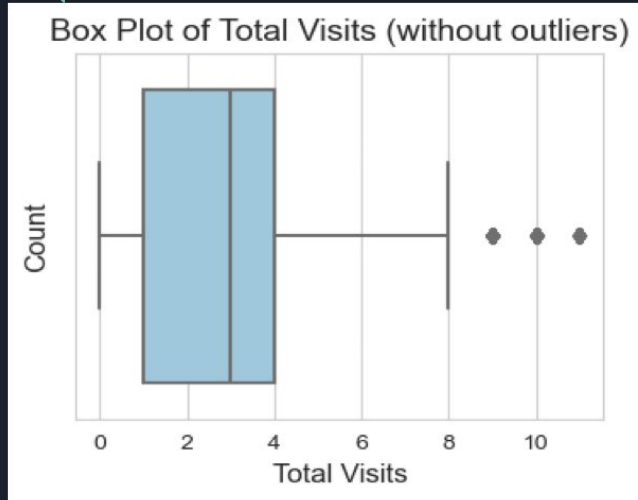
- X Education is seeking assistance in identifying the most prospective leads with the highest potential to become paying customers.
- To achieve this, the company requires the development of a lead scoring model, which assigns a score to each lead based on their likelihood of converting into paying customers. Leads with higher scores are expected to have a greater chance of conversion, while those with lower scores are anticipated to have a lower likelihood of converting.
- The CEO, in particular, has set a target lead conversion rate of approximately 80% as a desirable goal for the company to work towards.



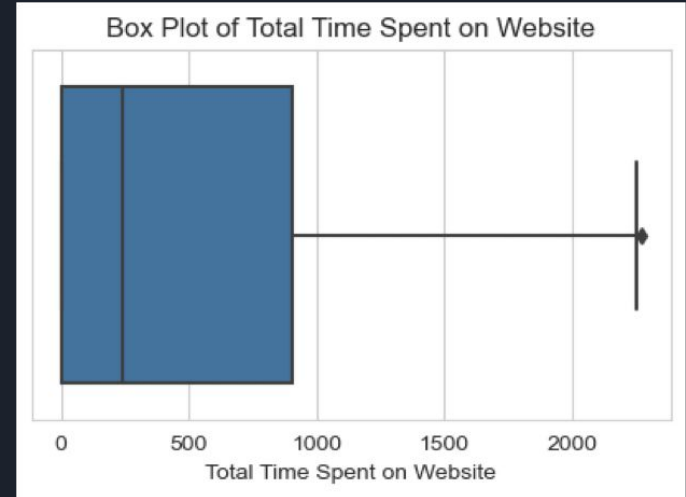
# Problem Solving Methodology

- Source Data for analysis
- Data Pre-processing: Data cleaning, Data Manipulation
- EDA: Univariate Data Analysis, Bivariate Data Analysis
- Feature Scaling
- Model Building: Logistic regression Model
- Model Training
- Model Evaluation
- Model Performance
- Predictions
- Conclusion

# Exploratory Data Analysis

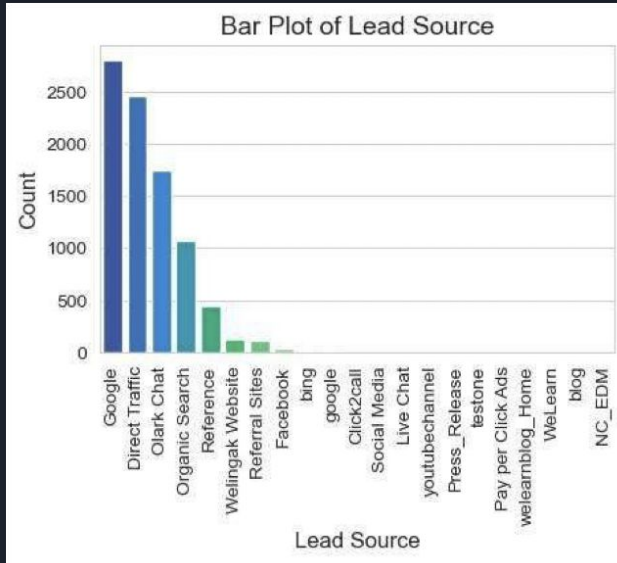


The box spans roughly 1 to 4, signifying that most Total Visit values are concentrated within this interval. The median, positioned at approximately 3, indicates that 50% of Total Visit values lie below this value, while the other 50% are above it.

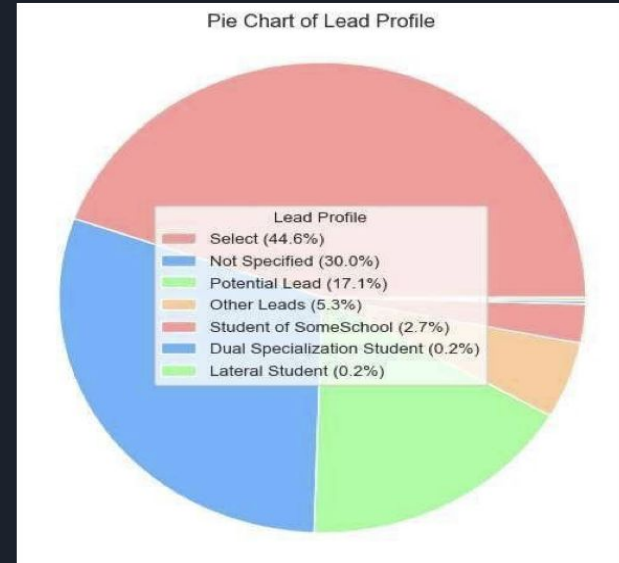


The box plot implies that most of the data points for 'Total Time Spent on Website' fall within the range of roughly 0 to 800, and the median value is approximately 250.

# Lead Source & Lead Profile



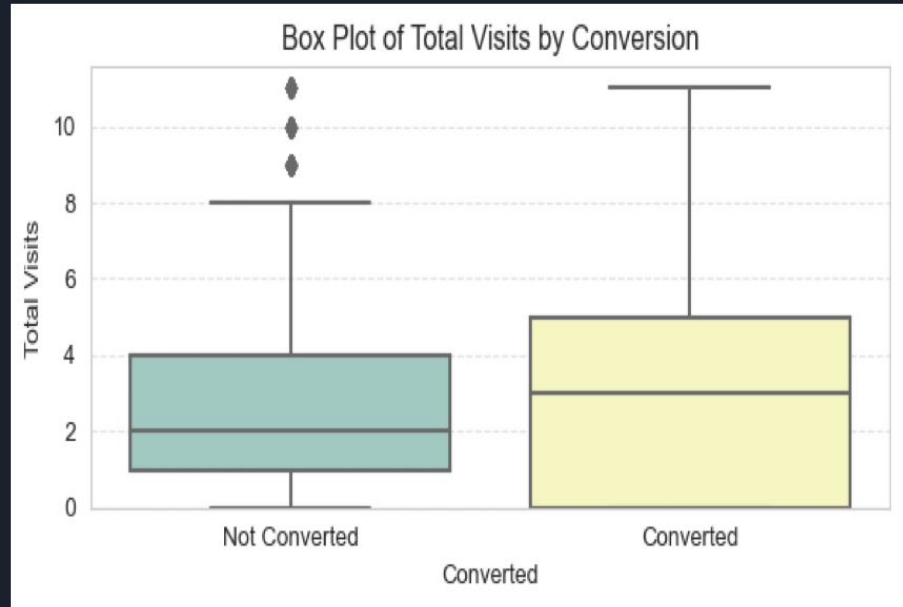
The predominant sources of traffic appear to be Google and direct visits.



Most leads are categorized with a "Select" label for their profile, indicating incomplete information.

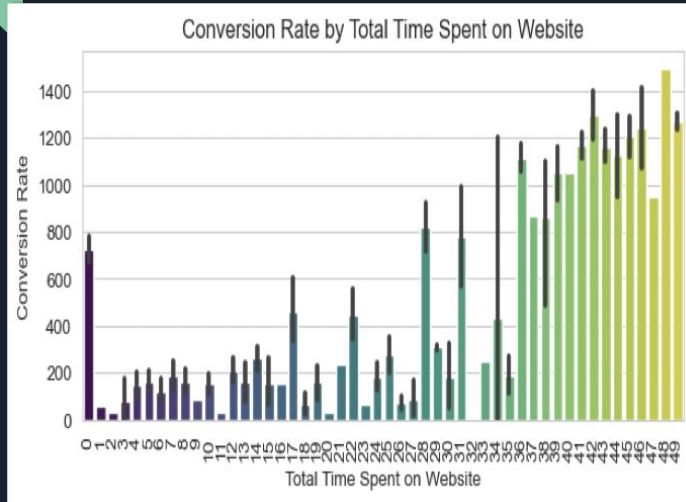
# Box Plot Conversion

- The median value for the 'Converted' group is slightly higher than that of the 'Not Converted' group, implying that, on average, converted leads tend to have slightly more 'Total Visits' compared to non-converted leads.
- The vertical span of the box illustrates where the bulk of the data points are concentrated. The box spans roughly from 0 to 4.5 in the 'Converted' group, while in the 'Not Converted' group, it extends from around 1 to 4.
- This indicates that the range of 'Total Visits' values for converted leads is slightly broader than for non-converted leads.



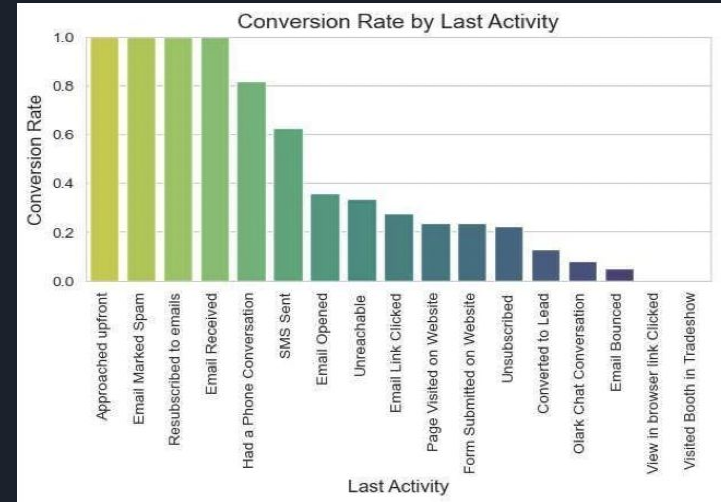
# Conversion Rate Graph

## Total Time



It is noticeable that the conversion rate experiences a notable increase among users who spend more time on the website.

## Last Activity

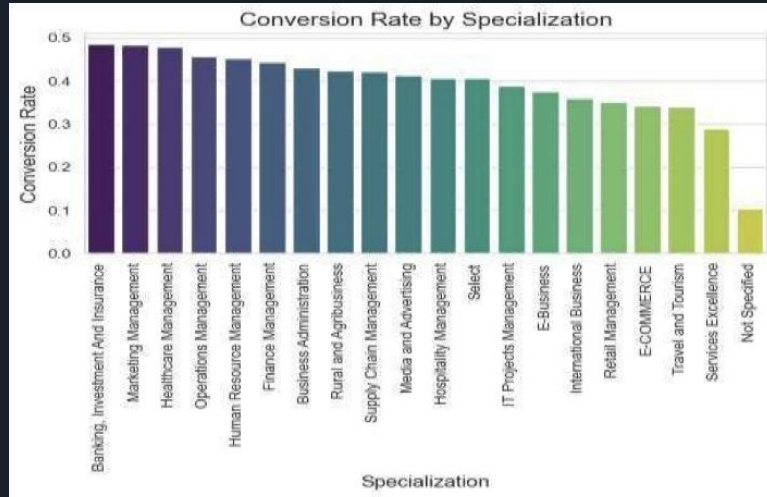


The increased conversion rate for actions like "Approached up front," "Email Marked Spam," "Resubscribe to emails," and "Emails Received" indicates that these activities boost the likelihood of conversion, making leads who engage in them more likely to become customers.



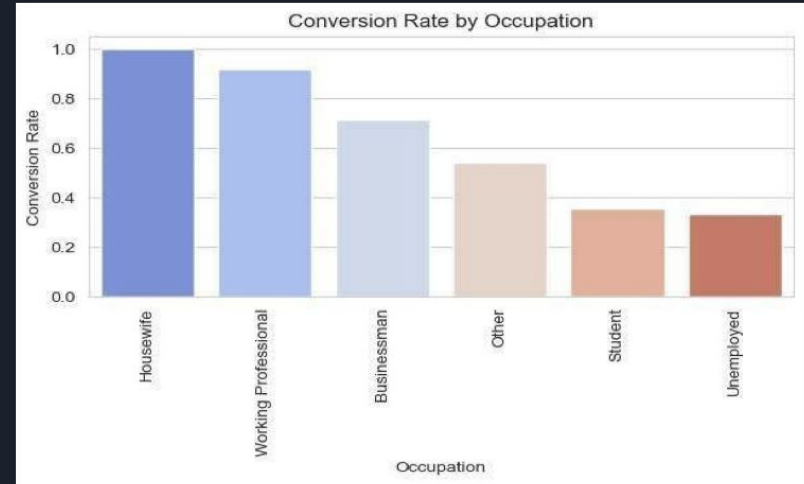
# Conversion Rate Graph

## Specialisation



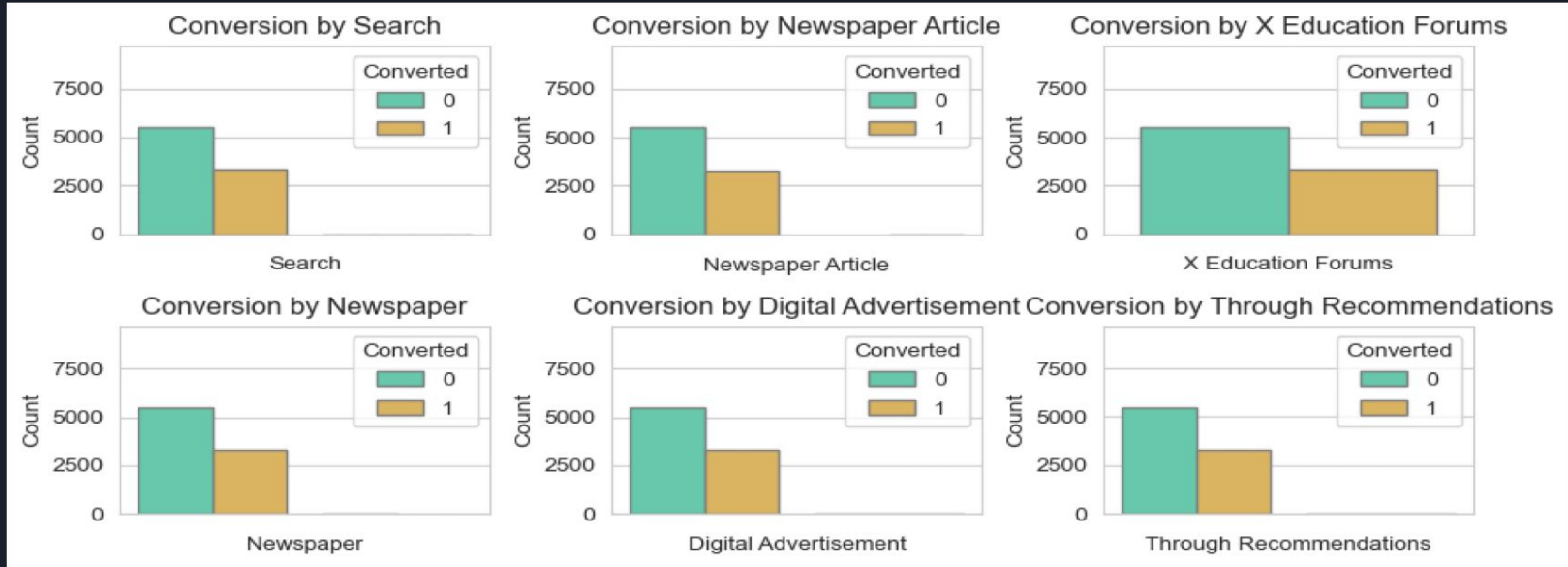
Most specializations boast a conversion rate exceeding 30%, with a few even achieving an impressive 50% conversion rate.

## Occupation



Homemakers and working professionals exhibit notably high conversion rates.

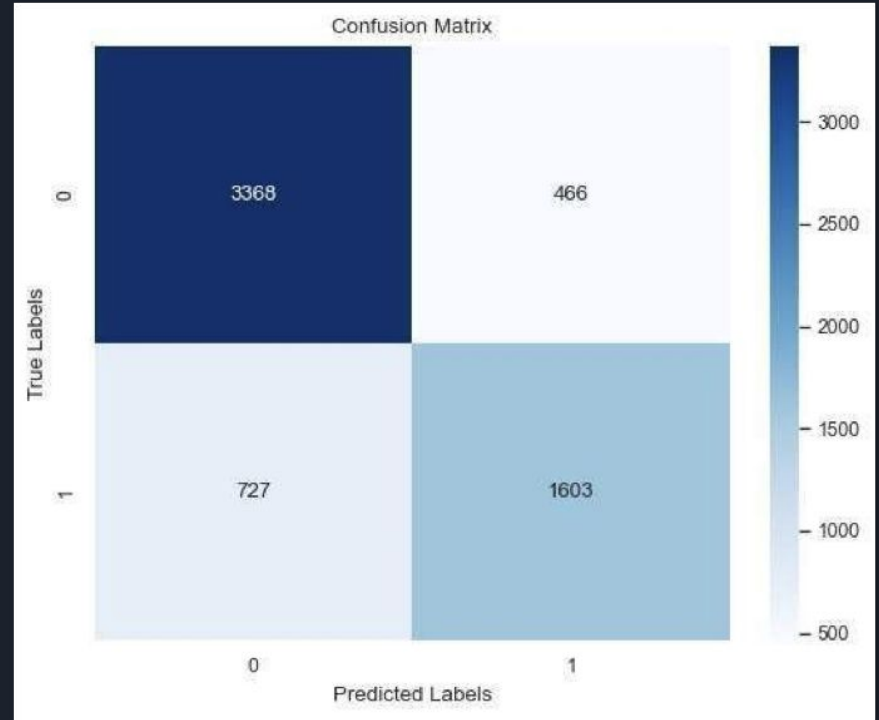
# Conversion Rates- Others



The conversion rates from various portals are relatively consistent or quite similar.

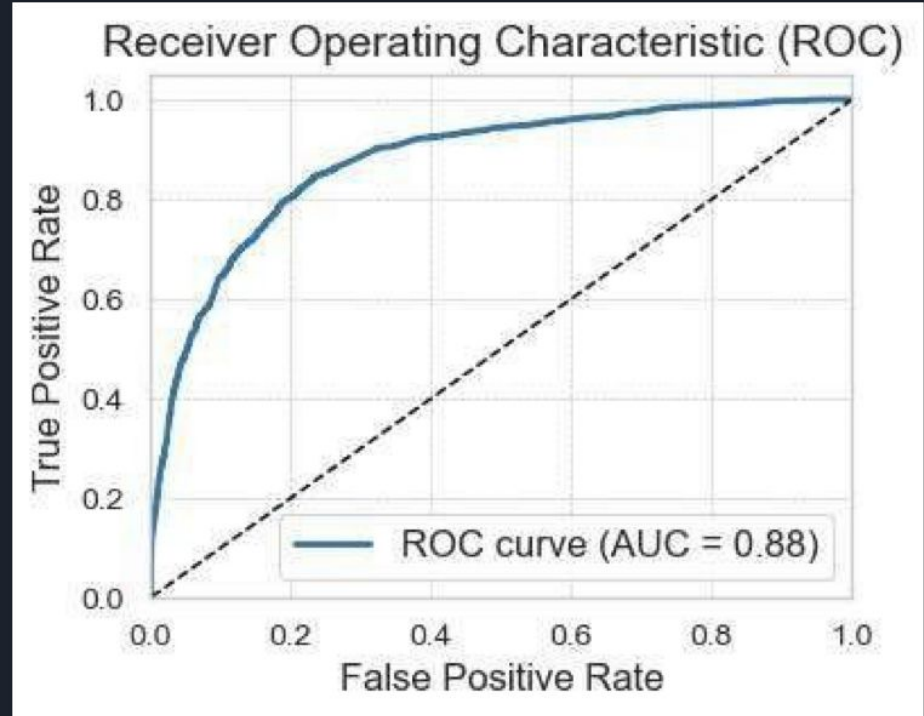
# Model Evaluation

- The confusion matrix furnishes insights into the performance of a binary classification model. In this instance, the model accurately predicted 3368 cases of the positive class (actual conversions) and 1603 instances of the hostile class (actual non-conversions).
- Nonetheless, it also made 466 false optimistic predictions (predicted as conversions but not converted) and 727 wrong pessimistic predictions (predicted as non-conversions but converted).

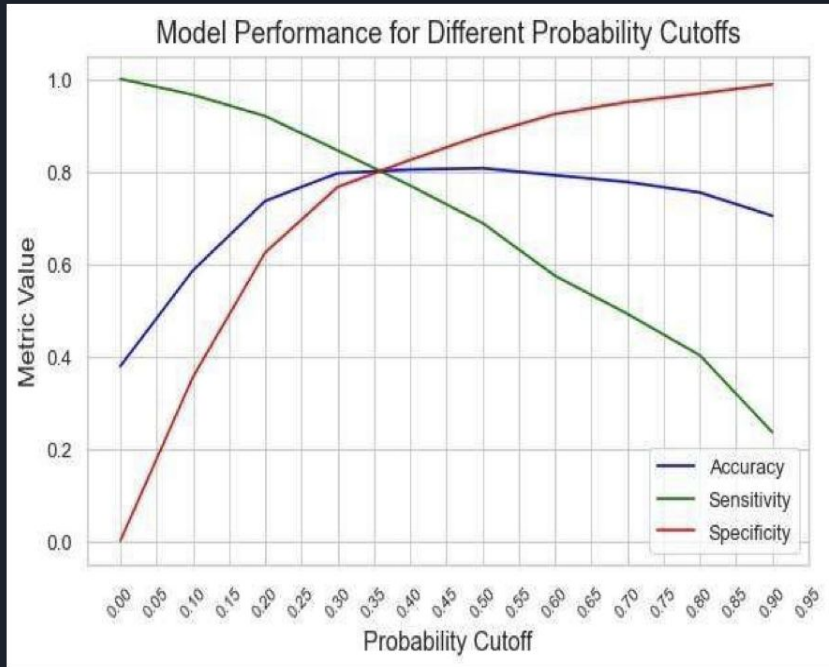


# ROC: Receiver Operating Characteristics

- A ROC AUC (Receiver Operating Characteristic Area Under the Curve) of 0.88 signifies the model's discriminatory solid capability when distinguishing between positive and negative instances.
- A higher ROC AUC value, which falls from 0 to 1, indicates a more excellent model proficiency in accurately classifying instances. In this context, an ROC AUC of 0.88 implies that the model predicts the conversion flag.



# Model Performance

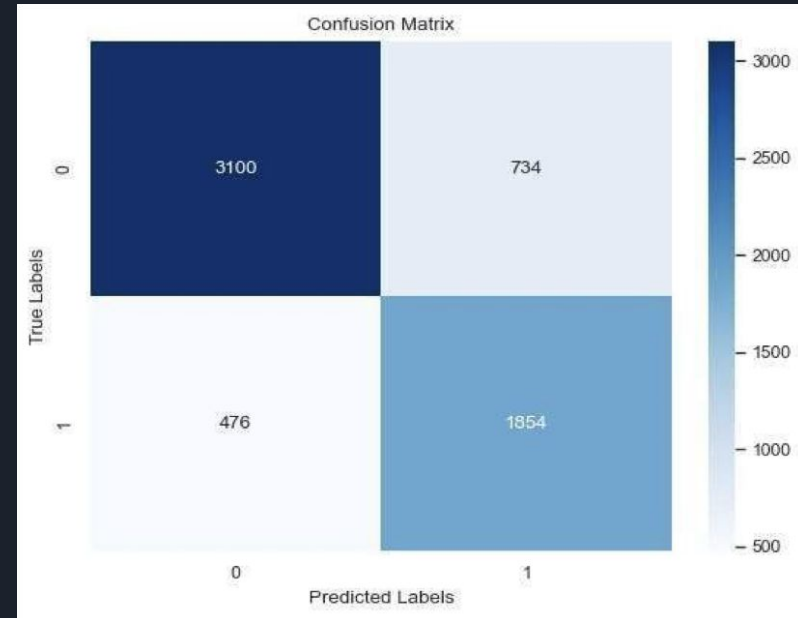


- The optimal cutoff point is situated around 0.37. The model strikes a harmonious balance between accuracy, sensitivity, and specificity at this particular threshold.
- This equilibrium implies that the model can proficiently classify positive and negative instances without displaying a bias towards either category.

# Binary Classification Model

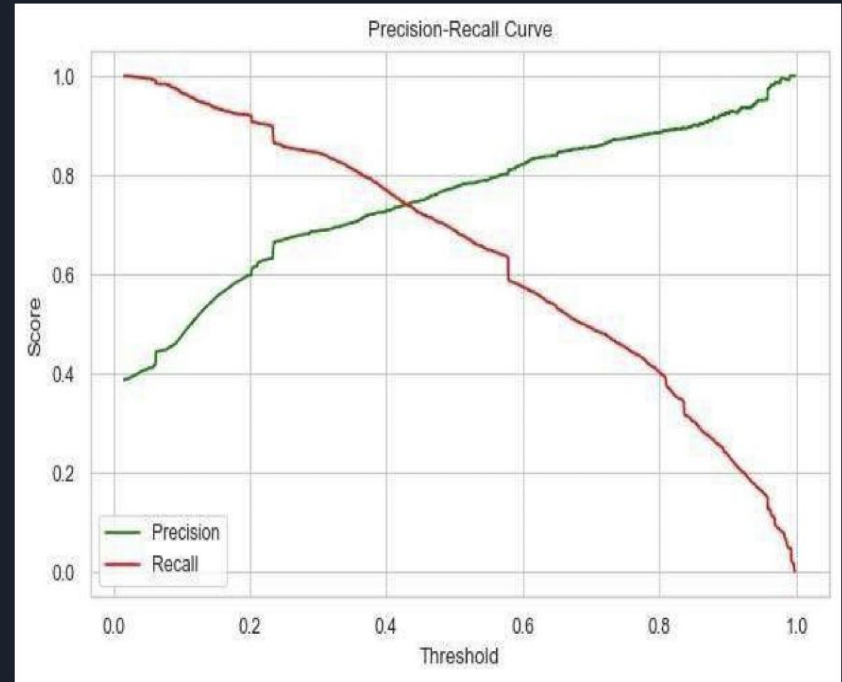
The confusion matrix provided is a 2x2 matrix used to assess the performance of a binary classification model. Each element in the matrix has a specific interpretation:

- True Negative (TN): 3100 - This represents the number of genuinely negative observations correctly predicted as unfavorable by the model.
- False Positive (FP): 734 - This signifies the number of negative observations that have been incorrectly predicted as positive by the model.
- False Negative (FN): 476 - This indicates the number of positive observations incorrectly predicted as unfavorable by the model.
- True Positive (TP): 1854 - This denotes the number of genuinely positive observations that have been correctly predicted as positive by the model.



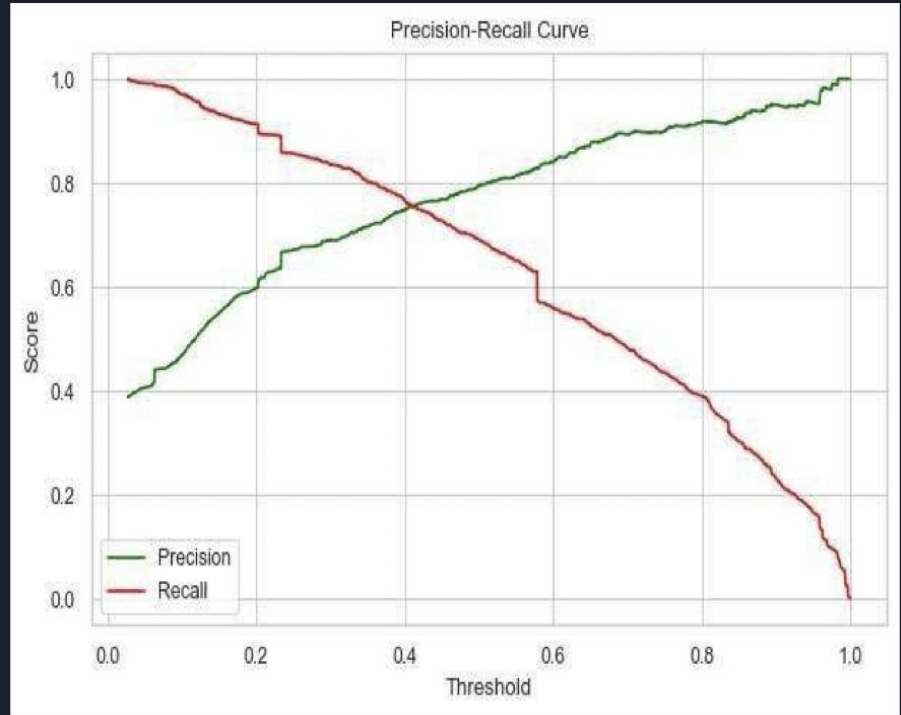
# Precision and Recall View

- When we use a threshold score of 0.4, the precision and recall values are notably high. This implies that the model effectively classifies a substantial proportion of positive instances (high recall) while keeping false optimistic predictions relatively low.
- The point where the precision and recall curves intersect, occurring at a score of 0.7 and a threshold of 0.4, indicates a well-balanced performance by the model. It signifies that the model adeptly identifies positive instances while minimizing false positives.



# Predictions

- The point where the precision and recall curves intersect is around (0.8, 0.4). This signifies that by setting a threshold of 0.8, we can attain a precision of 0.8 and a recall of 0.4.
- This intersection point serves as a threshold value where precision and recall are reasonably balanced in the model's performance.







# Conclusion

Based on the analysis and findings from the logistic regression model, here are some valuable recommendations for the customer:

1. Lead Scoring: Implement a scoring system (0-100) to prioritize leads for effective targeting.
2. Conversion Factors: Focus on key factors influencing conversions identified by the model.
3. Personalized Communication: Tailor messages based on lead data for better engagement.
4. Website Improvement: Enhance user experience to increase engagement and conversion chances.
5. Analyze Lead Quality: Target high-quality leads and regularly assess lead quality metrics.
6. Referral Programs: Encourage satisfied customers to refer others for broader reach.
7. Model Monitoring: Continuously monitor and update the model for accuracy.

8.

These recommendations aim to optimize lead conversion strategies.



# Thank You

---

---