Type I and II Errors

Problem Statement

- **Context**
  - **Moringa School** sells online courses to industry professionals.
  - Prospects land on the website, browse courses, fill out forms, complete tests, or watch videos.
  - Leads are generated when visitors provide their contact info (email/phone) or through past referrals.
- **Lead Conversion Process:**
  - Current lead conversion rate is 30% (e.g., for 100 leads, 30 convert to paying students).
- **Problem:**
  - **Lead Conversion Rate** is low despite generating many leads.

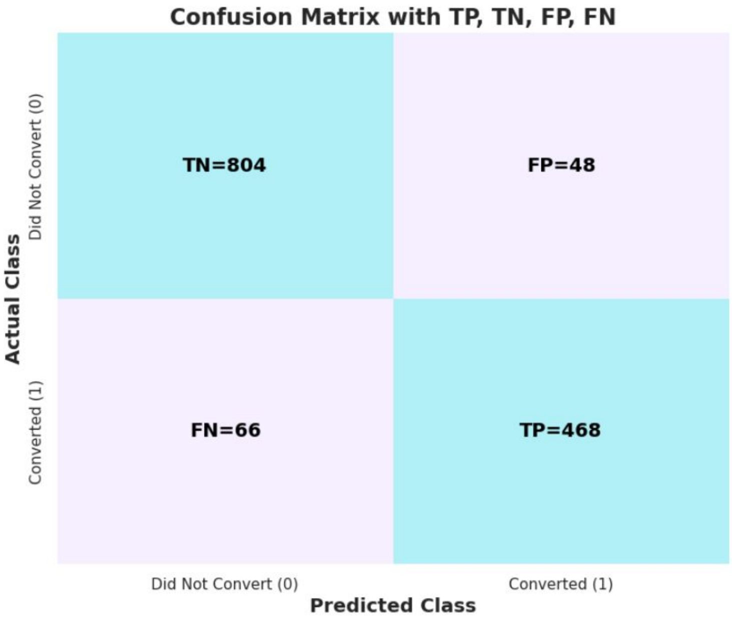
Problem Statement

- **Objective:**
    - Moringa is launching a new targeted marketing campaign
    - **Save cost on targeted marketing** by focusing on leads most likely to convert.
    - Identify **‘Hot Leads’** — the most promising leads to prioritize marketing efforts.
- **Solution:**
    - Build a **lead scoring model** to assign scores to leads.
    - Higher lead score = higher likelihood of conversion.
    - Target lead conversion rate: **80%**.

Target Variable (Class)

- The target variable for this analysis is the **Class**, which indicates whether a lead has converted to a paying student or not.
    - **Class 0:**
        - Represents leads that **did not convert** into paying students.
        - These leads show no engagement or interest to the point of payment, i.e., they did not follow through with the course registration or purchase.
    - **Class 1:**
        - Represents leads that **converted** into paying students.
        - These are the leads that successfully enrolled and made the purchase, turning into paying customers.
- The goal is to classify leads as either **Class 0 (Did not convert)** or **Class 1 (Converted)** based on the model, and to evaluate the classification performance using **Type I** and **Type II errors**.

Target Variable (Class)

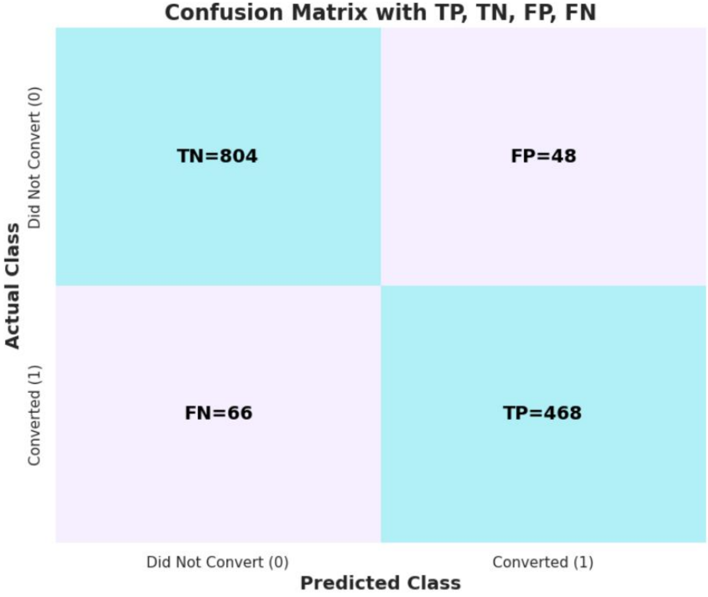


Confusion Matrix with TP, TN, FP, FN

Actual Class \ Predicted Class	Did Not Convert (0)	Converted (1)
Did Not Convert (0)	TN=804	FP=48
Converted (1)	FN=66	TP=468

- **True Negatives (TN) = 804:** These are the leads that did not convert (Class 0) and were correctly predicted as so.
- **False Positives (FP) = 48:** These are the leads that did not convert (Class 0) but were incorrectly predicted as converted (Class 1).
- **False Negatives (FN) = 66:** These are the leads that converted (Class 1) but were incorrectly predicted as did not convert
- **True Positives (TP) = 468:** These are the leads that converted (Class 1) and were correctly predicted as converted (Class 1).

Type I and Type II Errors:



Confusion Matrix with TP, TN, FP, FN

Actual Class \ Predicted Class	Did Not Convert (0)	Converted (1)
Did Not Convert (0)	TN=804	FP=48
Converted (1)	FN=66	TP=468

- **Type I Error (False Positive) - 48**
  - The model incorrectly predicts a lead will convert when it actually does not. In this case, 48 leads were predicted to convert but did not, leading to marketing or sales efforts being wasted on leads that weren’t likely to convert.
- **Type II Error (False Negative) - 66**
  - The model incorrectly predicts a lead will not convert when it actually does. Here, 66 leads were predicted not to convert but actually did, meaning missed opportunities to follow up with potentially valuable leads

Model Evaluation

Summary:

- The confusion matrix gives a direct view of how well the model is classifying leads.
- **804 TN** and **468 TP** indicate that the model is performing well on predicting leads that didn’t convert and those that did, respectively.
- **48 FP** and **66 FN** indicate room for improvement, particularly in reducing the Type I and Type II errors by adjusting the threshold or improving the model further

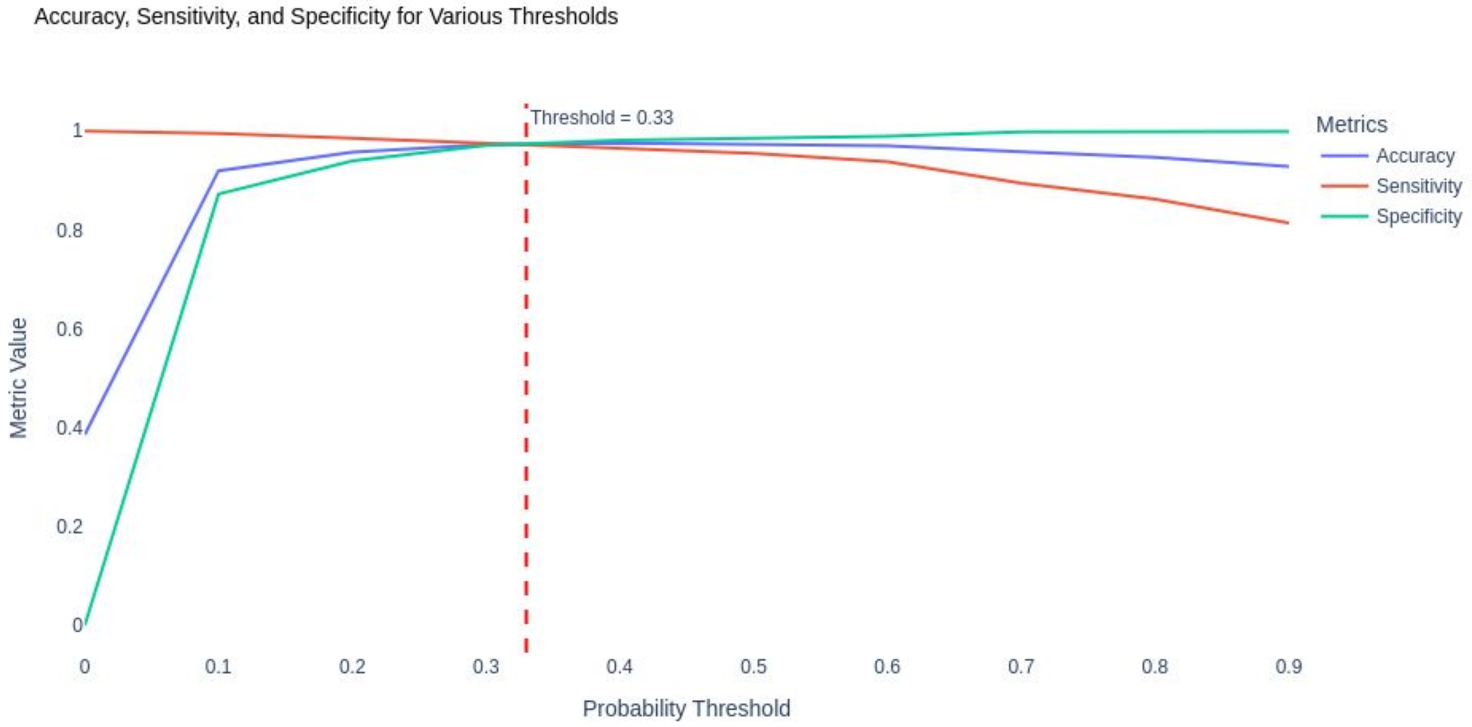
Model Improvement

- **Objective:**
  - **Save on marketing and sales costs** by focusing on the most promising leads (those with a higher probability of conversion).
- **Optimizing for Type I Error (False Positive):**
  - Type I errors involve predicting a lead will convert when they actually will not, **we want to minimize false positives**.
  - Reducing false positives ensures we don’t waste resources (calls, emails, ads) on leads that are unlikely to convert.
- **Why Focus on Type I Error?**
  - **Minimizing Type I errors** helps save costs by avoiding unnecessary efforts on leads that don’t convert, making the sales process more efficient.
  - This optimization strategy aligns with the goal of focusing marketing efforts on leads with the highest likelihood of conversion, thus improving the return on investment (ROI).

Model Improvement : Thresholding for Classification

- **Create Different Cutoff Values**
  - We apply various thresholds (from 0 to 0.9) to classify leads.
  - For each threshold value, we decide if a lead is classified as **converted** or **not converted** based on its predicted probability.
- Threshold range from from 0.0 to 0.9 for more granular control.
- **Metrics Calculated for Each Threshold:**
  - **Accuracy:** The percentage of correct predictions.
  - **Sensitivity (Recall):** The proportion of actual positive leads (converted) correctly predicted.
  - **Specificity:** The proportion of actual negative leads (did not convert) correctly predicted.

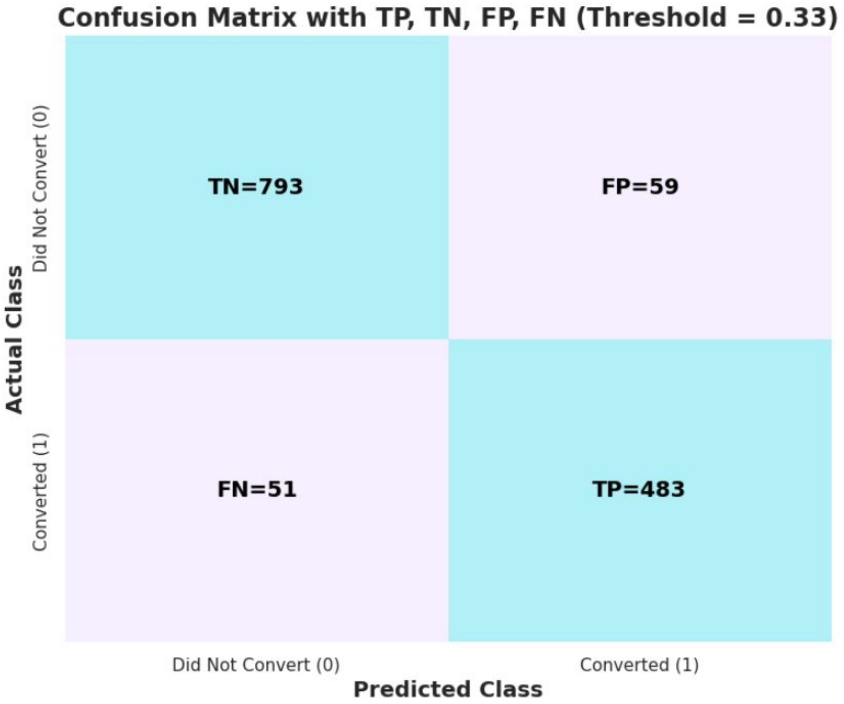
New Threshold for Classification



Finding the Optimal Threshold

- **Best Threshold:**
  - The optimal threshold will likely balance **high sensitivity** (minimizing Type II error) and **high specificity** (minimizing Type I error).
- **Threshold Adjustment Impact:**
  - Adjusting the threshold helps reduce false positives (Type I error) and false negatives (Type II error).
  - This leads to more targeted marketing, saving costs by focusing on high-probability leads.
- **Next Steps:**
  - Apply the selected threshold to make final predictions on the test set.
  - Evaluate the impact on the lead conversion rate and overall performance.

Adjusted Confusion Matrix



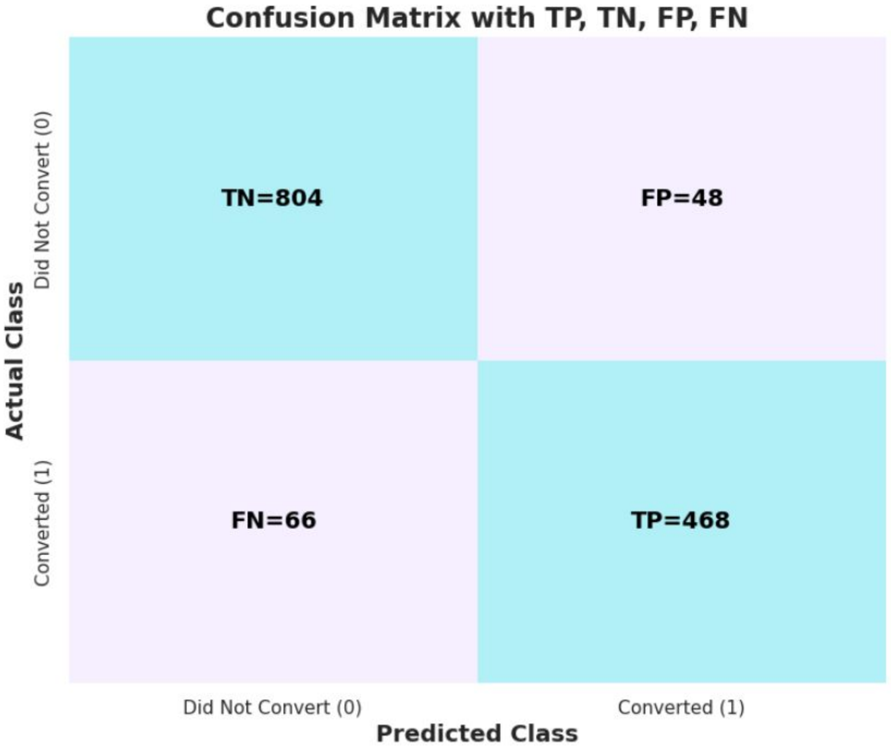
Confusion Matrix with TP, TN, FP, FN (Threshold = 0.33)

Actual Class \ Predicted Class	Did Not Convert (0)	Converted (1)
Did Not Convert (0)	TN=793	FP=59
Converted (1)	FN=51	TP=483

- **Confusion Matrix (Adjusted Threshold = 0.33)**
  - **True Negatives (TN)**: 793 — Correctly identified leads that did not convert.
  - **False Positives (FP)**: 59 — Incorrectly identified leads as converts.
  - **False Negatives (FN)**: 51 — Incorrectly identified leads that converted.
  - **True Positives (TP)**: 483 — Correctly identified leads that converted.

Comparing the Confusion Matrices

Before Adjustment

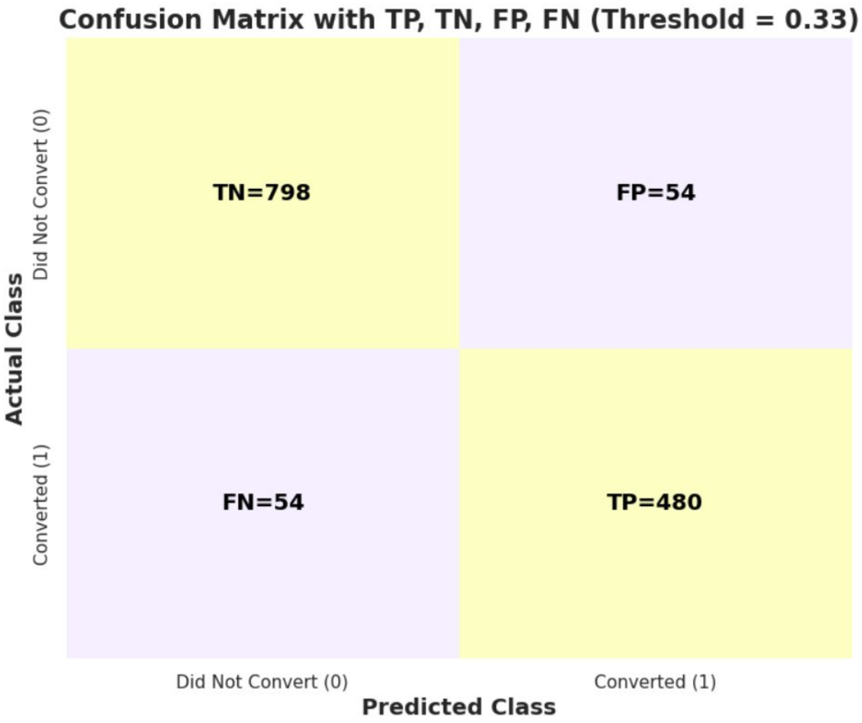


Confusion Matrix with TP, TN, FP, FN

	Did Not Convert (0)	Converted (1)
Did Not Convert (0)	TN=804	FP=48
Converted (1)	FN=66	TP=468

Actual Class
Predicted Class

After Adjustment



Confusion Matrix with TP, TN, FP, FN (Threshold = 0.33)

	Did Not Convert (0)	Converted (1)
Did Not Convert (0)	TN=798	FP=54
Converted (1)	FN=54	TP=480

Actual Class
Predicted Class

Model Improvement Analysis: Did the Threshold Adjustment Improve the Model?

- **True Negatives (TN)**: Slight decrease (804 → 793). This is minor and suggests the model is still performing well in identifying non-converting leads. Acceptable since the goal was to increase conversions
- **False Positives (FP)**: Increased (48 → 59). Adjusting the threshold led to more non-converting leads being incorrectly classified as converted. Typical trade-off when trying to increase True Positives.
- **False Negatives (FN)**: Decreased (66 → 51). Fewer converting leads are missed, which is a **positive improvement** (capturing more potential leads that could convert.
- **Model improved in identifying True Positives (TP)**: Increased (468 → 483). The model correctly identifies more converting leads with the adjusted threshold.