Gaussian Naive Bayes for Personal Loan Classification

Sarthak Sablania

April 2025

1 Problem Statement

The objective is to identify customers who are likely to accept a personal loan. By doing so, the bank can selectively target these individuals in marketing campaigns, increasing conversion while reducing cost.

2 Dataset Overview

The dataset contains 11 input features and 1 target variable (`Personal Loan`).

Feature Types

- **Continuous Features (5/11):**
  - `Age`
  - `Experience`
  - `Income`
  - `CCAvg`
  - `Mortgage`
- **Binary Categorical Features (4/11):**
  - `Securities Account`
  - `CD Account`
  - `Online`
  - `CreditCard`
- **Ordinal Features (2/11):**
  - `Family`
  - `Education`

Target

`Personal Loan: 0 or 1`

Analysis

I performed exploratory analysis to understand feature distributions, class imbalance, and feature correlations.

- **Data Size:** 5000 Rows
- **Feature Independence and Correlations:** Since `Experience` is highly correlated with `Age`, I removed the `Experience` column and checked the performance of the classifier, but it was the same with and without the column.

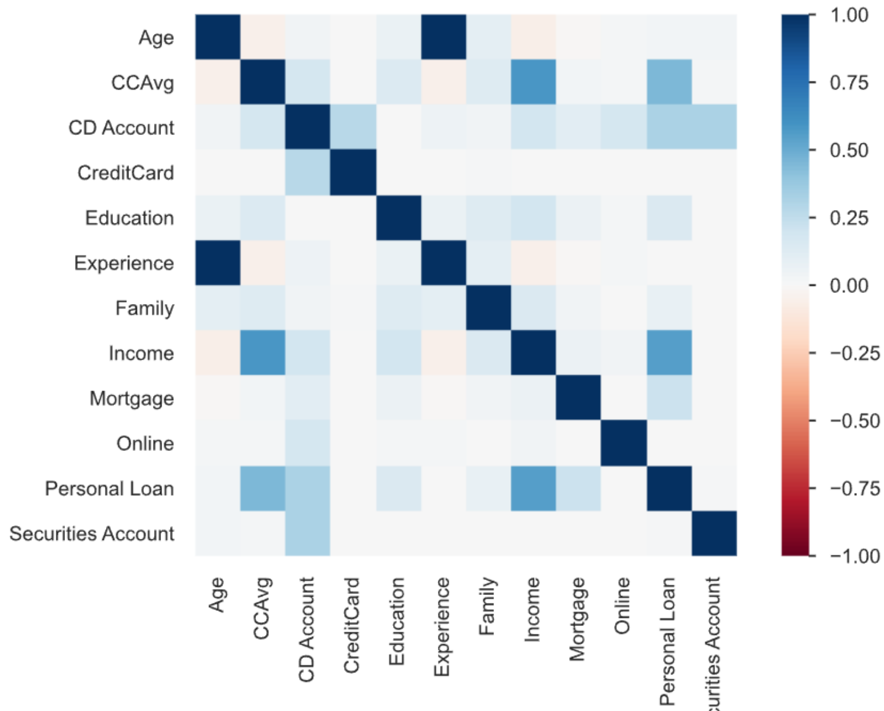

Figure 1: Correlation Matrix of Features and Target. Only `Experience` and `Age` are highly correlated, so I removed `Experience` column.

- **Class Imbalance:** 9.6% of the data is labelled as 1, rest labelled as 0.

3 Results and Threshold Tuning for Business Context

Motivation

The bank prefers **high recall** (capturing as many potential loan-takers as possible) over precision. Missing a potential customer is more costly than targeting an uninterested one.

Strategy

I lowered the classification threshold from the default of 0.5 to 0.02 to increase recall. Therefore, we classify a person as a potential personal loan customer if their predicted probability exceeds 0.02 instead of 0.5.

Results at Different Thresholds

Threshold = 0.02

Metric	Value
Recall (Class 1)	**0.94**
Precision (Class 1)	0.43
F1 Score	0.59
Accuracy	86%
Customers Targeted	**23%**

Threshold = 0.05

Metric	Value
Recall (Class 1)	0.89
Precision (Class 1)	0.45
F1 Score	0.60
Accuracy	88%
Customers Targeted	20%

Threshold = 0.16

Metric	Value
Recall (Class 1)	0.80
Precision (Class 1)	0.50
F1 Score	0.62
Accuracy	90%
Customers Targeted	17%

Threshold = 0.50

Metric	Value
Recall (Class 1)	0.61
Precision (Class 1)	0.50
F1 Score	0.55
Accuracy	89%
Customers Targeted	13%

Threshold = 0.93

Metric	Value
Recall (Class 1)	0.46
Precision (Class 1)	0.59
F1 Score	0.52
Accuracy	91%
Customers Targeted	8%

- As threshold decreases, **recall increases** but **precision decreases**.
- At threshold ≈ 0.02, recall reaches ∼0.94, while only ∼23% of the population is targeted.
- At threshold ≈ 0.93, recall ≈ 0.46, while only ∼8% of the population is targeted.

If we choose the threshold of 0.02, we successfully identify 94% of actual positive cases while only targeting 23% of the population, making it a highly focused and cost-effective strategy.
Alternatively, a higher threshold (e.g., 0.93) could be chosen to precisely target the most probable potential customers, depending on business goals.

Results from Various Implementations

Results from all three types of implementations — from scratch, using `scikit-learn`, and AI-generated — were consistent.

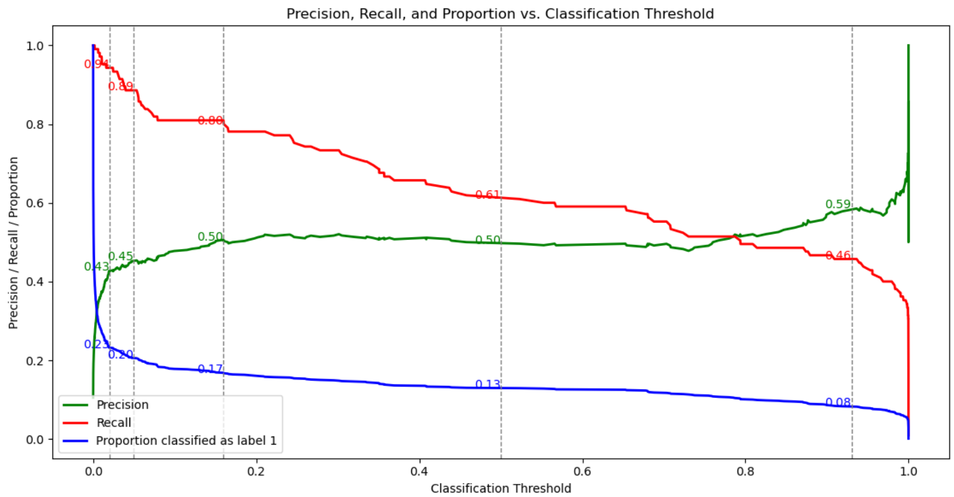

Figure 2: Precision, Recall, and Proportion of Label 1 Predictions vs. Threshold

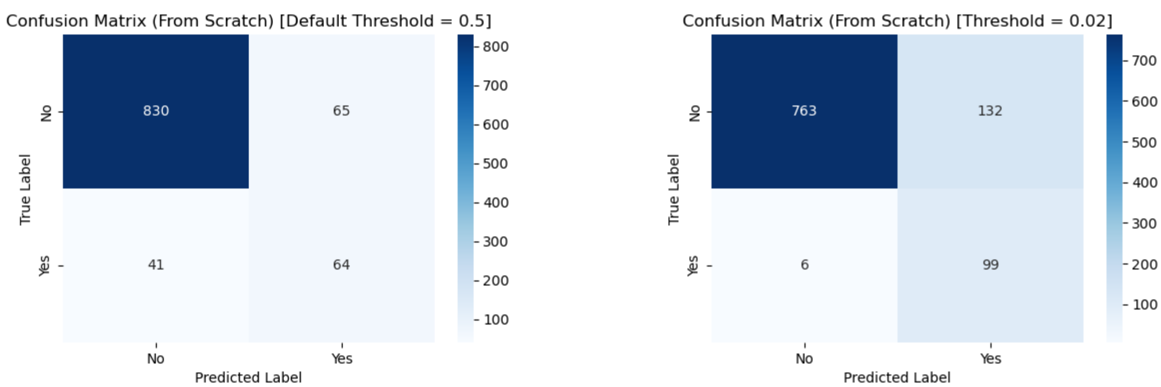

Figure 3: Confusion Matrices for From-Scratch Implementation: (a) Threshold = 0.5 and (b) Threshold = 0.02. We correctly identify 99 out of 105 of the potential PL customers if we choose 0.02 as threshold.

4 Conclusion

In this project, we demonstrated the effectiveness of Gaussian Naive Bayes (GNB) for personal loan classification. Despite the presence of categorical and non-Gaussian features, GNB delivered competitive results, underscoring its robustness and computational simplicity.

Through threshold tuning, we aligned model behavior with business objectives — particularly the emphasis on recall. By lowering the classification threshold, we were able to identify 94% of potential loan customers while only targeting 23% of the total population. This strategy significantly enhances campaign efficiency by minimizing missed opportunities without overwhelming marketing resources.

The consistency of results across different implementations — from-scratch, scikit-learn, and AI-assisted — validates the reliability of our modeling approach. Overall, this analysis provides a strong foundation for a targeted marketing strategy that balances business impact

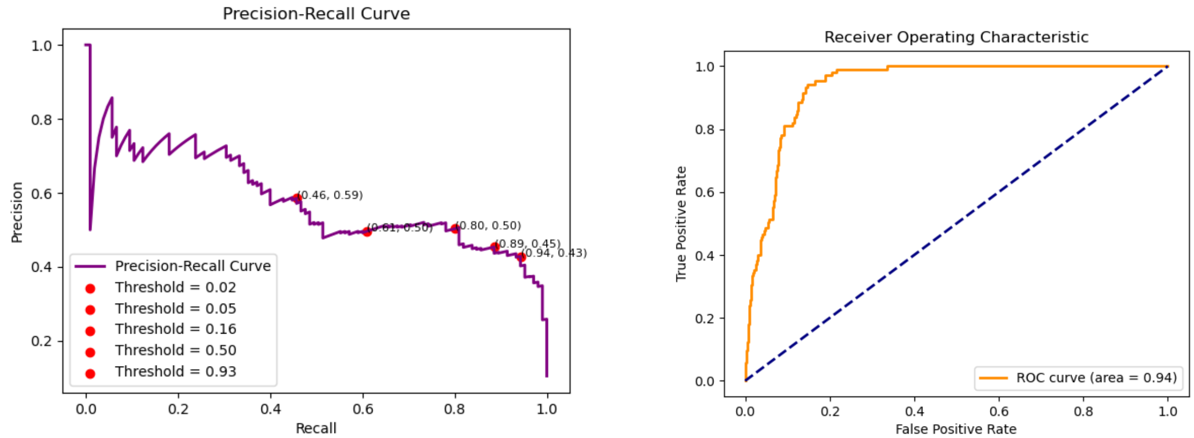

Figure 4: Model Evaluation Plots: (a) Precision-Recall Curve and (b) ROC Curve

with operational efficiency.

GitHub Repository

The code and analysis for this project are available at:
`github.com/sablania-dev/Personal-Loan--Naive-Bayes`