Mathematics of Loan Default Prediction and Risk Analytics

Project Documentation

August 2025

Abstract

This document is a rigorous mathematical exploration of the algorithms, feature engineering, and profit optimization used in our loan default risk modeling platform. It is designed for graduate students and practitioners with backgrounds in statistics, machine learning, and financial risk.

1 Overview

The objective is to predict the default event $Y \in \{0, 1\}$ on a given loan application, maximize business profit, and segment customers for optimal portfolio management. Core phases include: data preprocessing, feature engineering, business cost modeling, supervised ML training, and post-hoc explainability.

2 Data Preparation

Let the feature vector for each applicant be $\mathbf{x} = [x_1, x_2, \ldots, x_n]$. Dataset: $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. Missing data handled by impute(x_j) and outlier filtering by Z-score: $z_j = \frac{x_j - \mu_j}{\sigma_j}$.

3 Feature Engineering

3.1 Debt-to-Income Ratio (DTI)

$$\text{DTI}_i = \frac{D_i^{\text{monthly}}}{I_i^{\text{monthly}}}$$

where D^{monthly} is total monthly debt, I^{monthly} is monthly income.

3.2 Debt Service Coverage Ratio (DSCR)

$$\text{DSCR}_i = \frac{I_i^{\text{monthly}}}{D_i^{\text{monthly}} + P_i^{\text{proposed}}}$$

where P^{proposed} is the proposed loan payment.

3.3 Payment Shock

$$\text{Shock}_i = \text{BackEnd}_i - \text{CurrentBurden}_i$$

3.4 Credit Risk Tier

By quantized cutoffs in CreditScore.

3.5 Opportunity Index

A composite: $O_i = 0.4S_i + 0.35C_i + 0.25Q_i$, where S_i = stability score, C_i = capacity score, Q_i = quality score.

4 Business Cost Optimization

Define profit as:

$$\text{NetProfit} = \sum_{i=1}^{N} [\text{TP}_i \cdot r_{TP} - \text{FP}_i \cdot c_{FP} - \text{FN}_i \cdot c_{FN}]$$

with r_{TP}, c_{FP}, c_{FN} reference rates:

$$\text{TP}_i = I[y_i = 0 \wedge \hat{y}_i = 0],\ \text{FP}_i = I[y_i = 0 \wedge \hat{y}_i = 1],\ \text{FN}_i = I[y_i = 1 \wedge \hat{y}_i = 0]$$

Threshold t^* maximizing expected profit:

$$t^* = \arg\max_{t \in [0,1]} \text{NetProfit}(t)$$

5 Machine Learning Model

5.1 Gradient Boosted Trees (LightGBM)

Fitted by minimizing loss $L(\mathbf{y}, \hat{\mathbf{y}})$, weighted for business cost:

$$L = \sum_{i=1}^{N} w_i \cdot \ell(y_i, \hat{y}_i)$$

where w_i reflects business cost (from above). $\ell()$ usually is logistic loss.

5.2 Interpretability

Permutation importance: assess decrease in accuracy by randomly shuffling x_j across D. Decision tree extraction: $f_{tree}(\mathbf{x})$ yields rules for boundary splits.

6 Bias and Fairness

For protected groups G (e.g., Education), disparate impact ratio:

$$DI = \frac{\min_g R_g}{\max_g R_g},\ R_g = \text{Rejection Rate in Group } g$$

Model passes the 80% rule if $DI \geq 0.8$.

7 Portfolio Segmentation

Multi-dimensional scoring assigns each customer to a segment $S_i \in \{\text{Premier}, \text{Prime+}, \ldots, \text{HighTouch}\}$ according to cut-offs in score composites.

8 Model Documentation and Auditing

Model Card summarises:

- Model class, feature input
- Business objective and metrics
- Fairness and bias test results
- Usage guidelines and limitations

Audit-ready for regulators and compliance teams.

9 Conclusion

This pipeline combines best practices in statistical modeling, ML, business cost optimisation, fairness, and financial risk engineering. The mathematical framework is modular and extensible: practitioners can swap models, adjust cost parameters, or add fairness constraints.