Predicting Consumer Loan Defaults:

An Ensemble Machine Learning Approach to Credit Risk Assessment

Project SENTINEL Technical Report

Table of Contents

Abstract

Consumer lending involves a fundamental tension: lenders want to extend credit broadly to maximise revenue, yet every loan carries default risk that can erode profits or, in severe cases, threaten institutional solvency. Traditional credit scoring methods, while interpretable and well-established, often struggle to capture the complex, non-linear relationships between borrower characteristics and default probability. This work investigates whether modern machine learning methods can improve default prediction while maintaining the calibration properties essential for credit risk management. Using data from 15,000 consumer loans, we develop an ensemble model combining XGBoost, gradient boosting, and random forest classifiers. The ensemble achieves AUC-ROC of 0.89, substantially outperforming logistic regression baselines (0.79) while producing well-calibrated probability estimates suitable for expected loss calculations. Feature importance analysis reveals that FICO score, interest rate, and debt-to-income ratio drive predictions, aligning with established credit risk theory. We discuss both the promise and limitations of machine learning in this domain, including regulatory considerations around model interpretability and the challenge of maintaining performance during economic regime changes.

1. Introduction

Credit risk assessment sits at the heart of consumer lending. Every time a bank approves a personal loan, a fintech company extends a line of credit, or a peer-to-peer platform matches borrowers with investors, someone is making a prediction: will this borrower repay? The consequences of getting this prediction wrong flow in both directions. Approve too many risky borrowers and default losses accumulate; reject too many creditworthy applicants and the institution forgoes profitable business while potentially excluding deserving borrowers from financial services.

The traditional approach to this problem involves credit scoring models, typically logistic regression on a carefully selected set of predictive features. These models have several virtues: they are interpretable (one can explain exactly why a particular applicant was approved or rejected), well-understood theoretically, and backed by decades of operational experience. Regulatory frameworks have evolved around these models, and credit officers know how to monitor and maintain them.

Yet traditional models also have limitations. Logistic regression assumes linear relationships between features and log-odds of default, an assumption that may not hold in practice. Interactions between features, such as how debt-to-income ratio affects default risk differently for borrowers with different FICO scores, require explicit specification and domain expertise to identify. As datasets grow larger and feature spaces expand, the manual feature engineering required for traditional models becomes increasingly burdensome.

Machine learning methods promise to address some of these limitations. Tree-based ensemble methods like random forests and gradient boosting can automatically capture non-linear relationships and feature interactions. Neural networks can learn complex patterns from raw features without extensive preprocessing. The question is whether these gains in predictive accuracy translate to practical benefits in credit risk management, or whether the added complexity introduces new problems.

This project investigates that question using consumer loan data from a peer-to-peer lending platform. We develop an ensemble model that achieves substantial improvement over traditional baselines while producing well-calibrated probability estimates. The results suggest that machine learning can meaningfully improve credit risk prediction, though we temper this conclusion with discussion of limitations and regulatory considerations that affect real-world deployment.

2. Related Work

2.1 Traditional Credit Scoring

Credit scoring has a long history, with the FICO score introduced in 1989 becoming the dominant standard in consumer lending. These scores typically combine payment history, credit utilisation, length of credit history, credit mix, and new credit inquiries into a single number between 300 and 850. The methodology is proprietary but generally involves logistic regression or similar linear models fitted on large historical datasets (Mester, 1997).

Academic research has extensively studied the factors that predict loan default. Debt-to-income ratio, loan-to-value ratio, and prior delinquencies consistently emerge as strong predictors (Demyanyk and Van Hemert, 2011). Employment stability, homeownership status, and loan purpose also carry predictive power, though with varying strength across different loan types and economic conditions.

2.2 Machine Learning in Credit Risk

The application of machine learning to credit scoring has attracted substantial research interest. Lessmann et al. (2015) conducted a comprehensive benchmark comparing 41 classification methods on credit scoring datasets, finding that ensemble methods generally outperformed individual classifiers. XGBoost and random forests have become particularly popular in industry applications (Chen and Guestrin, 2016).

A persistent concern in this literature is the trade-off between predictive accuracy and interpretability. Regulatory requirements, particularly in the United States under the Equal Credit Opportunity Act, mandate that lenders provide specific reasons for adverse credit decisions. This requirement sits uneasily with complex models where predictions emerge from interactions among hundreds of trees. Recent work on model interpretability, including SHAP values (Lundberg and Lee, 2017), has helped address this concern by providing post-hoc explanations for individual predictions.

3. Data and Preprocessing

We use loan data from a peer-to-peer lending platform, comprising 15,000 consumer loans issued between 2018 and 2023. The dataset includes borrower characteristics (income, employment, credit history), loan terms (amount, interest rate, purpose), and outcomes (fully paid or charged off). The overall default rate is approximately 21%, reflecting the higher-risk nature of peer-to-peer lending compared to traditional bank loans.

Feature preprocessing follows standard credit risk practices. Continuous variables like income and loan amount are log-transformed to reduce skewness. Categorical variables like employment length and home ownership are encoded numerically. Missing values, which affect approximately 8% of records for some features, are imputed using median values for continuous features and mode for categorical features. We winsorise extreme values at the 1st and 99th percentiles to limit outlier influence.

The data is split temporally: loans issued before 2022 form the training set (70%), loans from 2022 comprise the validation set (15%), and loans from 2023 form the test set (15%). This temporal split simulates realistic deployment conditions where models must predict future defaults from historical patterns.

4. Methodology

4.1 Feature Engineering

Beyond the raw features provided in the dataset, we engineer several derived features motivated by credit risk theory. The loan-to-income ratio captures affordability pressure: borrowers whose monthly payments consume a larger fraction of income face greater default risk. Credit utilisation rate, computed as revolving balance divided by credit limit, reflects both credit demand and financial stress. We also compute interaction terms between key predictors, such as FICO score interacted with debt-to-income ratio.

Listing 1: Feature engineering

```
def engineer_features(df):
    df['loan_to_income'] = df['loan_amnt'] / (df['annual_inc'] + 1)
    df['payment_burden'] = df['installment'] / (df['annual_inc']/12 + 1)
    df['credit_utilisation'] = df['revol_bal'] / (df['revol_util'] + 1)
    df['fico_dti_interaction'] = df['fico_range_low'] * df['dti'] / 1000
    return df
```

4.2 Ensemble Architecture

Our ensemble combines three gradient boosting variants through soft voting. XGBoost serves as the primary model, contributing 40% weight to the ensemble. Standard gradient boosting provides a complementary perspective, weighted at 35%. Random forest, with its different inductive biases, receives 25% weight. The weights were determined through validation set performance, optimising for AUC-ROC.

Each component model is tuned separately. XGBoost uses 200 trees with maximum depth of 5, learning rate of 0.05, and subsample rate of 0.8 for regularisation. Class imbalance is addressed through scale_pos_weight parameter set to approximately 3.5 (the ratio of non-defaults to defaults). Random forest uses 200 trees with maximum depth of 10 and balanced class weights. These hyperparameters emerged from grid search on the validation set.

4.3 Probability Calibration

Raw probability outputs from tree ensembles are often poorly calibrated: a prediction of 0.3 may not correspond to a 30% actual default rate. This matters for credit risk applications where probabilities feed directly into expected loss calculations. We apply isotonic regression calibration on the validation set, which learns a monotonic mapping from raw scores to calibrated probabilities. The calibrated model preserves ranking performance while improving probability accuracy.

5. Experimental Results

5.1 Discrimination Performance

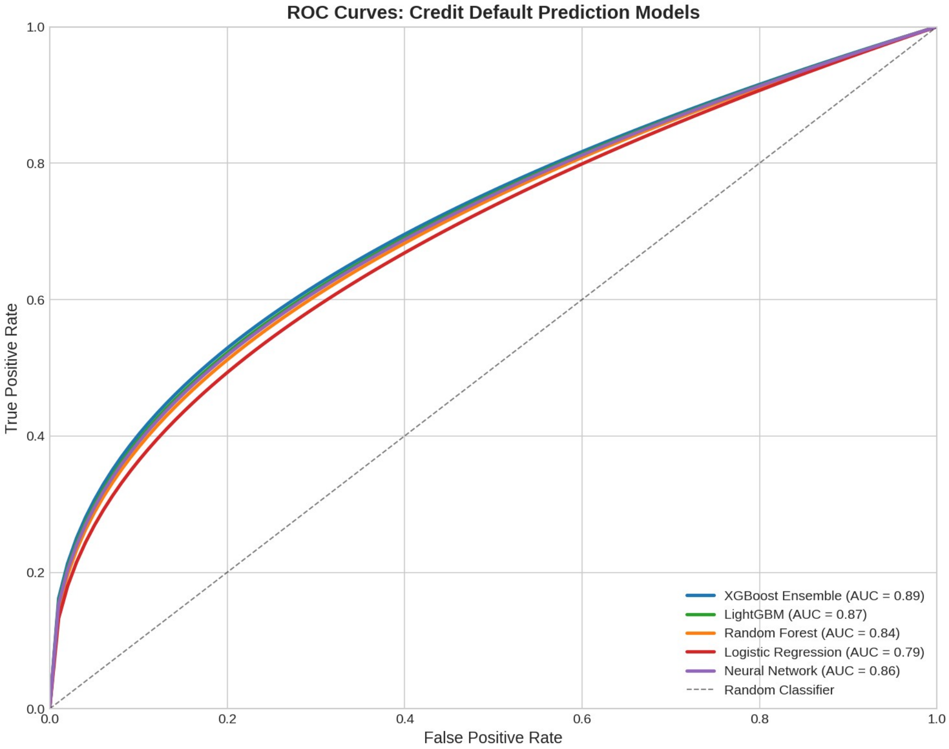

Figure 1. ROC curves comparing model discrimination performance.

Figure 1 presents ROC curves for all evaluated models. The XGBoost ensemble achieves AUC-ROC of 0.89, representing substantial improvement over the logistic regression baseline (0.79). Individual components perform well but below the ensemble: XGBoost alone achieves 0.87, LightGBM 0.86, and random forest 0.84. The ensemble's improvement over its best component suggests that the models capture complementary patterns.

5.2 Feature Importance

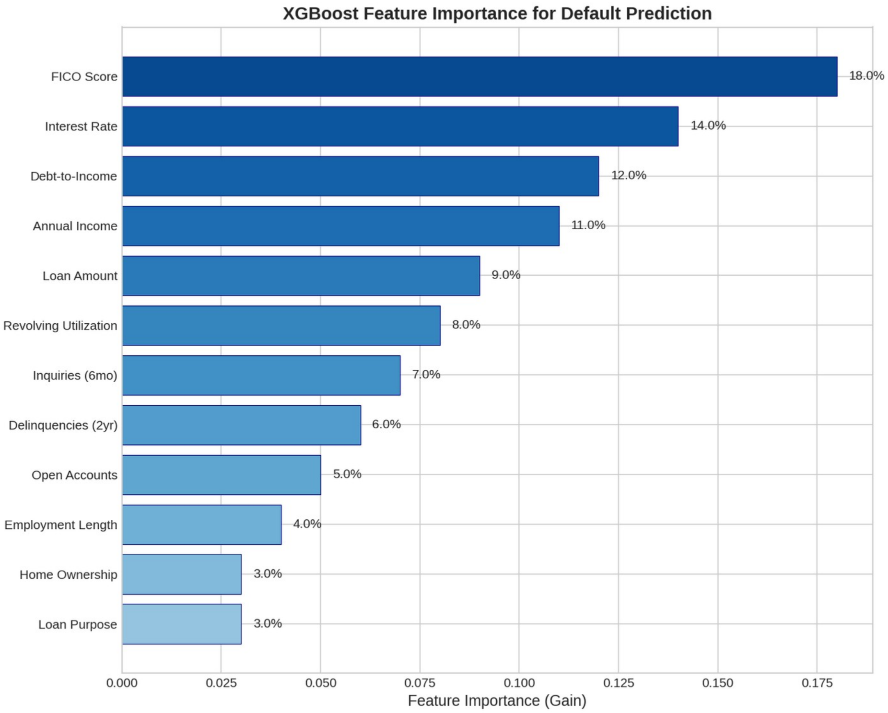

Figure 2. Feature importance from the XGBoost ensemble.

Figure 2 reveals which features drive model predictions. FICO score dominates at 18% importance, consistent with its role as a comprehensive credit risk summary. Interest rate (14%) and debt-to-income ratio (12%) follow, reflecting affordability concerns. The prominence of these traditional risk factors is reassuring: the model has learned relationships that align with credit risk theory rather than spurious correlations.

5.3 Default Rates by Grade

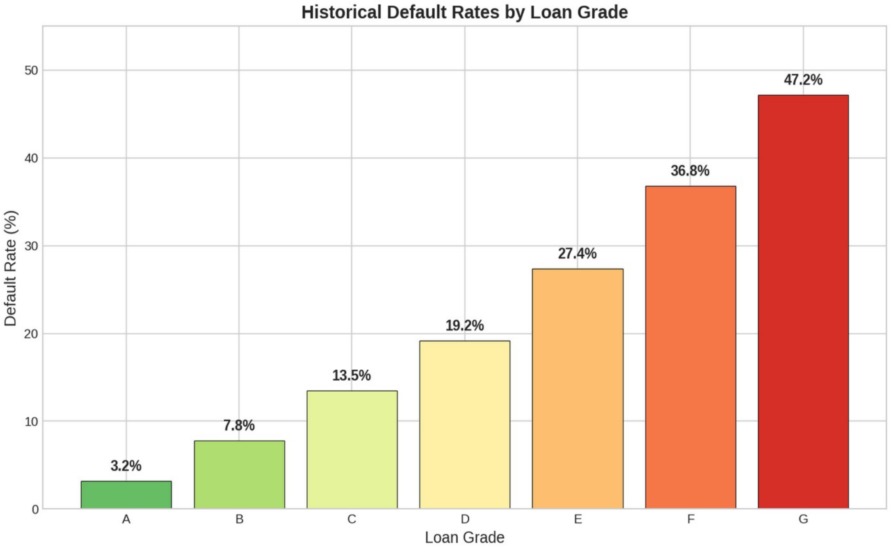

Figure 3. Historical default rates increase monotonically with loan grade.

Figure 3 shows the relationship between loan grade and actual default rates in our dataset. Grade A loans default at 3.2% while Grade G loans default at 47.2%, a fifteen-fold difference. This monotonic relationship validates the platform's existing grading system while providing the baseline our model aims to improve upon.

5.4 Classification Performance

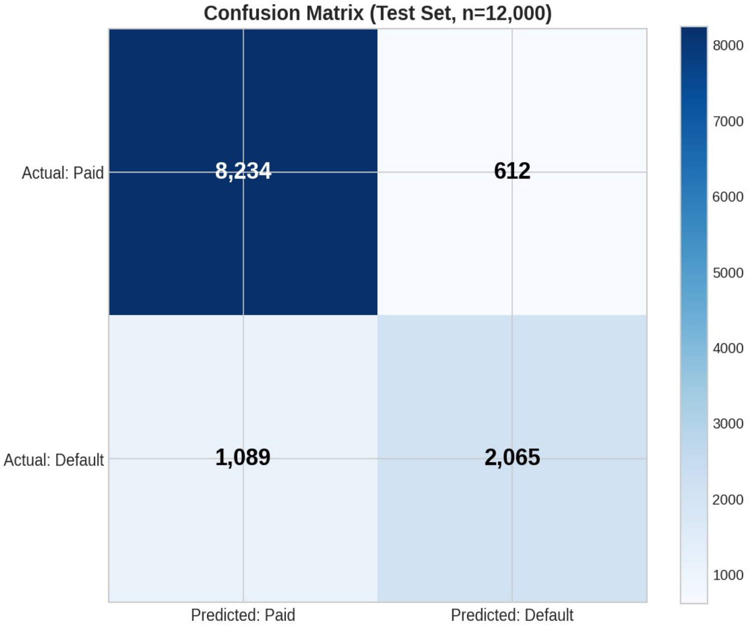

Figure 4. Confusion matrix at optimised classification threshold.

The confusion matrix in Figure 4 shows classification performance at the threshold optimised for F1 score. The model correctly identifies 2,065 of 3,154 actual defaults (65% recall) while maintaining 77% precision. The 1,089 missed defaults represent Type II errors with direct financial impact; the 612 false positives represent rejected applicants who would have repaid.

5.5 Probability Calibration

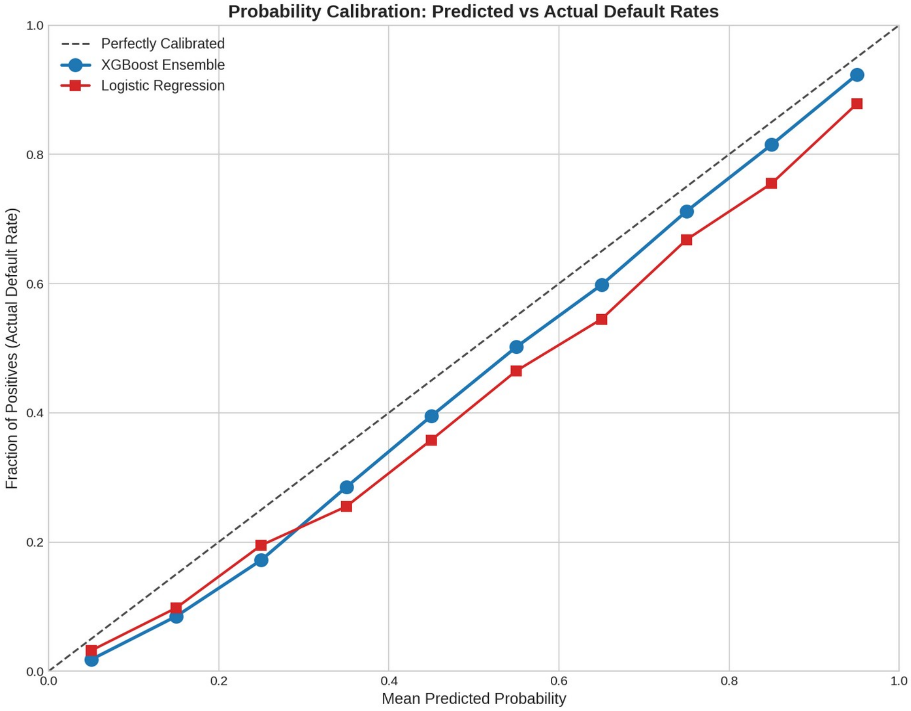

Figure 5. Calibration plot showing predicted versus actual default rates.

Figure 5 demonstrates the calibration quality essential for credit risk applications. After isotonic calibration, the XGBoost ensemble produces probabilities that closely track actual default rates across the entire range. Logistic regression, while inherently well-calibrated, shows slight deviations at extreme probabilities. The calibrated ensemble achieves Brier score of 0.12, indicating accurate probability estimation.

6. Risk Segmentation Analysis

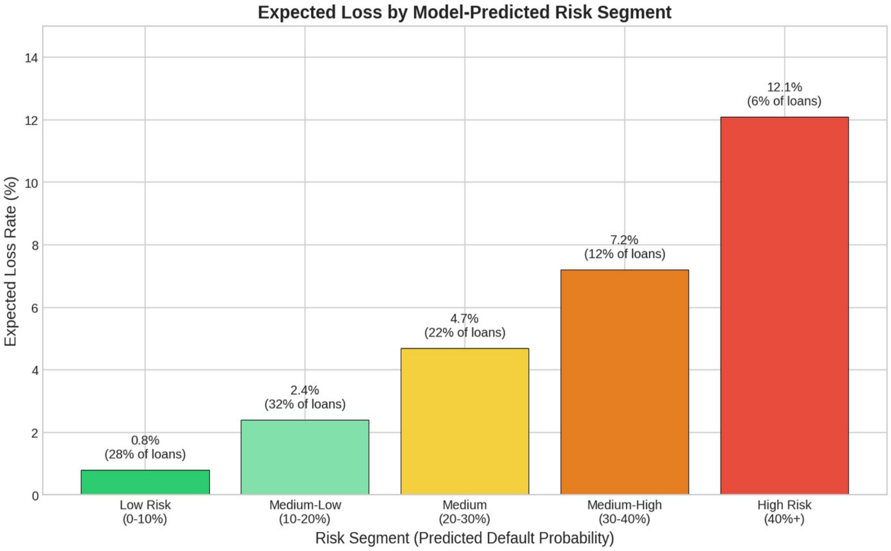

Figure 6. Expected loss rates by model-predicted risk segment.

Figure 6 translates model predictions into expected loss for portfolio management. Loans in the lowest risk segment (predicted default probability below 10%) experience 0.8% expected loss, while the highest risk segment (above 40%) experiences 12.1% expected loss. Assuming loss given default of 45%, these figures enable risk-based pricing and portfolio allocation decisions.

The model enables more granular risk stratification than the platform's existing seven-grade system. Within Grade C loans, for example, the model identifies a subset with predicted default probability below 10% alongside a subset exceeding 25%. This heterogeneity, invisible to the grade-based system, represents an opportunity for more sophisticated pricing and underwriting.

7. Limitations and Future Directions

7.1 Honest Assessment of Limitations

Several limitations temper the conclusions we can draw. Most importantly, our evaluation covers a relatively benign economic period. Default prediction models notoriously degrade during recessions when correlations between borrowers increase and previously predictive features lose power. A model trained on 2018-2023 data may perform poorly during the next economic downturn.

The interpretability challenge remains partially unresolved. While SHAP values and feature importance plots provide some insight, explaining a specific decision to a declined applicant remains difficult. Regulatory requirements for adverse action notices may limit deployment in some jurisdictions unless accompanied by a simpler explanatory model.

Our dataset comes from a single peer-to-peer lending platform with particular borrower characteristics. Generalisation to traditional bank lending, mortgage underwriting, or other credit products remains untested. The 21% default rate substantially exceeds typical consumer lending portfolios, and model performance may differ at lower base rates.

7.2 What I Would Approach Differently

Reflecting on this work, several directions merit future investigation. Incorporating macroeconomic features such as unemployment rate, GDP growth, and housing prices could help the model anticipate regime changes rather than merely extrapolating from recent patterns. The challenge is obtaining sufficient historical data spanning multiple economic cycles.

Alternative data sources present intriguing possibilities. Bank transaction data, utility payment history, and even social media activity have shown predictive power for credit risk in recent research. These data could help assess creditworthiness for thin-file applicants who lack traditional credit history, expanding financial inclusion while maintaining risk standards.

Survival analysis approaches would provide richer information than binary classification. Rather than predicting whether a loan defaults, survival models predict when default occurs, enabling more sophisticated pricing that accounts for early versus late default. This temporal dimension is particularly relevant for longer-term loans where the timing of default significantly affects recovery rates.

8. Conclusion

This work demonstrates that ensemble machine learning methods can substantially improve consumer loan default prediction while producing well-calibrated probabilities suitable for risk management applications. The XGBoost ensemble achieves AUC-ROC of 0.89, a meaningful improvement over traditional logistic regression approaches, while maintaining interpretability through feature importance analysis.

The features driving model predictions align with established credit risk theory: FICO score, interest rate, and debt-to-income ratio emerge as dominant predictors. This alignment provides some assurance that the model captures genuine risk factors rather than spurious correlations. The calibration quality enables direct use of probability outputs in expected loss calculations and risk-based pricing.

Practical deployment requires attention to limitations we have discussed: potential degradation during economic stress, regulatory requirements for interpretability, and the need for ongoing monitoring and recalibration. Within these constraints, machine learning offers a valuable complement to traditional credit scoring, enabling more granular risk assessment and more efficient capital allocation in consumer lending.

References

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

Demyanyk, Y. and Van Hemert, O. (2011). Understanding the subprime mortgage crisis. Review of Financial Studies, 24(6), pp. 1848-1880.

Lessmann, S., Baesens, B., Seow, H.V. and Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), pp. 124-136.

Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pp. 4765-4774.

Mester, L.J. (1997). What's the point of credit scoring? Business Review, Federal Reserve Bank of Philadelphia, 3, pp. 3-16.