Loan Anomaly Detection for Repayment Behavior Analysis

CS5344 Big Data Analytics Track 2: Finance

Team Project Proposal

September 27, 2025

Contents

1 Introduction

In the modern financial landscape, accurate prediction of loan defaults is crucial for risk management and maintaining portfolio stability. Traditional credit scoring models primarily rely on static borrower characteristics at loan origination, potentially missing dynamic patterns that emerge during the loan lifecycle. This project addresses the challenge of **loan-level anomaly detection for repayment behavior**, where we analyze both static borrower information and temporal performance sequences to identify loans that exhibit abnormal repayment patterns.

The problem is formulated as a semi-supervised learning task where the training dataset contains exclusively normal loans (loans that meet their obligations), while the validation and test sets include both normal and abnormal loans. This mirrors real-world scenarios where financial institutions have abundant historical data on performing loans but limited labeled examples of defaults during model development phases.

Our approach leverages the rich temporal structure inherent in loan performance data, combining static origination features with dynamic monthly performance indicators to detect anomalous repayment behaviors that may signal impending default or other adverse outcomes.

2 Project Objective

The primary objective of this project is to develop a robust anomaly detection system capable of identifying loans with abnormal repayment patterns using a combination of static borrower characteristics and temporal performance sequences. Specifically, we aim to maximize the Average Precision (AUPRC) score for detecting abnormal loans in an imbalanced dataset (87.39% normal vs 12.61% abnormal), design a semi-supervised learning framework that effectively learns normal loan behavior patterns from training data containing only normal loans, develop sophisticated feature engineering techniques that capture both static risk factors and temporal dynamics in loan performance, and create an interpretable and scalable solution suitable for deployment in real-world financial risk management systems.

The success of our approach will be measured primarily by the Average Precision metric, which is particularly appropriate for imbalanced binary classification problems as it focuses on the precision-recall trade-off rather than overall accuracy.

3 Target Task

Our target task is **semi-supervised anomaly detection** in the financial domain, specifically focused on loan repayment behavior analysis. The task is formulated as binary classification where we predict whether a loan exhibits normal (0) or abnormal (1) repayment behavior patterns.

The learning paradigm follows semi-supervised learning principles, as the training data contains only normal examples (30,504 loans with target=0), while validation data contains both classes (4,693 normal, 677 abnormal loans). Each loan i is represented as:

$$x_i = (s_i, (t_{i,1}, r_{i,1}), (t_{i,2}, r_{i,2}), \ldots, (t_{i,T_i}, r_{i,T_i}), y_i) \tag{1}$$

where s_i represents static loan information (borrower attributes, loan terms), $\{t_{i,k}\}_{k=1}^{T_i}$ are time periods with $t_{i,1} < \ldots < t_{i,T_i}$, $T_i = 14$ represents the number of months for loan i (consistent across dataset), $r_{i,k}$ is the monthly repayment information vector at time $t_{i,k}$, and $y_i \in \{0, 1\}$ is the binary label (0=normal, 1=abnormal).

The challenge lies in learning meaningful representations that capture both static risk factors and temporal patterns indicative of abnormal repayment behavior, without having access to labeled abnormal examples during the training phase.

4 Dataset

Our dataset consists of loan-level data from the Freddie Mac Single-Family Loan-Level Dataset, comprising 30,504 training loans × 145 features (100% normal loans), 5,370 validation loans × 145 features (87.39% normal, 12.61% abnormal), and 13,426 test loans × 144 features (unlabeled for competition submission).

The feature structure includes 31 static origination variables encompassing borrower credit score (mean=753.6, std=156.1), original unpaid principal balance (mean=$317K, std=$181K), loan-to-value ratio (mean=75.2%, std=19.4%), original interest rate (mean=6.72%, std=0.55%), debt-to-income ratio, loan terms, property characteristics, and demographic information. Additionally, 112 temporal features provide monthly performance data spanning exactly 14 months for all loans, tracking eight performance metrics: CurrentActualUPB, CurrentInterestRate, CurrentNonInterestBearingUPB, EstimatedLTV, InterestBearingUPB, LoanAge, MonthlyReportingPeriod, and RemainingMonthsToLegalMaturity.

Exploratory data analysis reveals consistent temporal structure with all loans having complete 14-month sequences, mixed data types (71 integer, 60 float, 14 categorical features), and credit quality distribution ranging from 600-850 (excluding missing value code 9999). Critical data quality issues include zero missing values across temporal features but significant missingness in static features: ReliefRefinanceIndicator (100% missing), PreHARP_Flag (100% missing), SuperConformingFlag (98.92% missing), and MSA (11.22% missing).

5 Challenges

Several significant challenges characterize this dataset and task. The semi-supervised learning paradigm requires models to learn patterns of normality without exposure to abnormal examples during training, while validation data mixing necessitates careful evaluation strategies to avoid data leakage. Severe class imbalance with only 12.61% abnormal loans in the validation set makes standard accuracy metrics misleading, establishing Average Precision (AUPRC) as the critical evaluation metric.

The high-dimensional mixed-type feature space (145 features combining numerical, categorical, and temporal data) presents curse of dimensionality challenges for neighborhood-based methods, requiring effective feature selection and dimensionality reduction. Temporal complexity manifests through variable importance across different time periods and complex dependencies, necessitating balance between modeling sophistication and computational efficiency.

Domain-specific challenges include borrower and product heterogeneity across different loan types, economic cycle effects on repayment patterns, and regulatory influences on loan performance. Special encoding schemes (999, 9999) for missing values in critical features like DTI and CreditScore require domain-aware imputation strategies that distinguish between truly missing data and "not applicable" cases.

6 Literature Survey

Credit anomaly detection has been widely explored using statistical, machine learning, and deep learning approaches. Traditional models such as logistic regression and discriminant analysis were among the earliest techniques, offering interpretability but limited ability to capture complex borrower–loan interactions [1]. Ensemble learning methods like Random Forests and Gradient Boosted Trees improved detection accuracy, though at the cost of reduced transparency [2].

With the rise of deep learning, methods such as autoencoders and recurrent neural networks (RNNs) have been applied to credit data. Autoencoders identify anomalies by reconstructing normal repayment patterns and flagging deviations, while long short-term memory (LSTM) networks and gated recurrent units (GRUs) effectively capture sequential dependencies in loan repayment histories [3]. Although these models achieve strong performance, their lack of interpretability and high computational demand remain challenges in regulated financial environments.

Hybrid approaches have emerged to combine static borrower features with dynamic repayment sequences. Attention mechanisms and temporal convolutional networks have been used to model long-range dependencies while maintaining some interpretability [4]. These methods aim to bridge the gap between accuracy and explainability.

More recently, probabilistic graphical models have resurfaced as a promising direction. A 2025 study introduced a Credit Anomaly Detection Method based on Bayesian Networks, capturing causal dependencies between borrower attributes and repayment behaviors [5]. By modeling conditional probabilities, Bayesian networks balance performance with transparency, aligning with the financial industry's growing emphasis on trustworthy and explainable AI.

Overall, the literature reflects a shift from purely predictive black-box models toward approaches that integrate accuracy with interpretability, a trend critical for adoption in high-stakes financial decision-making.

7 Experimental Comparison

We evaluated several classical anomaly detection approaches on our dataset to identify the most promising direction. The comparison includes density-based methods (LOF variants), global outlier detectors (Isolation Forest), reconstruction-based approaches (PCA), and distance-based methods (K-Means).

Method	**AUPRC**	**AUROC**
Multi-k LOF (weighted)	**0.2220**	0.6062
LOF (optimized $k = 10$)	0.1956	0.5752
LOF + PCA	0.1867	0.5486
Isolation Forest	0.1303	0.5107
PCA Reconstruction	0.1247	0.5034
K-Means Distance	0.1100	0.4534

The results demonstrate that Local Outlier Factor (LOF) variants consistently outperform global methods, with the multi-k weighted ensemble achieving the best performance (AUPRC = 0.2220). This superior performance aligns with the heterogeneous nature of our loan dataset, where local density comparisons prove more effective than global assumptions about anomaly distributions.

8 Intended Approach

Based on our experimental results, we propose a LOF-centered ensemble approach that leverages the superior local density estimation capabilities demonstrated in our baseline comparisons. Our methodology follows a systematic pipeline designed for semi-supervised anomaly detection while maintaining interpretability and computational efficiency.

The core approach involves training multiple LOF detectors with varying neighborhood parameters ($k \in \{5, 6, 7, 8\}$) on normal training data only, then combining their anomaly scores through weighted aggregation. This multi-scale approach captures both fine-grained local patterns and broader neighborhood structures, addressing the heterogeneity observed in loan repayment behaviors.

Our preprocessing pipeline handles domain-specific challenges including special missing value encodings (999, 9999) and mixed data types through consistent imputation and scaling strategies. The ensemble combination weights are optimized using a held-out portion of the validation set, ensuring that hyperparameter selection does not compromise model generalization.

To enhance detection capability, we incorporate engineered features capturing repayment irregularities, such as deviations from expected amortization schedules. These domain-informed features provide additional signals that complement the density-based anomaly scores from the LOF ensemble.

The approach maintains strict separation between training and evaluation phases: all unsupervised components are fitted exclusively on normal training data, validation labels are used only for hyperparameter selection, and final performance assessment occurs on an untouched holdout subset.

9 Future Enhancements

We plan to explore cluster-aware LOF implementations that perform density estimation within borrower segments, potentially improving anomaly detection precision for specific credit tiers. Additionally, adaptive neighborhood selection mechanisms could dynamically adjust k values based on local data density characteristics, further optimizing the LOF performance across diverse loan portfolios.

10 References

References

[1] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3), 235–255.

[2] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[3] Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *ACM Computing Surveys*, 51(5), 1–36.

[4] Zhang, H., Yao, L., Sun, A., & Tay, Y. (2021). Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Transactions on Information Systems*, 39(1), 1–42.

[5] Li, Y., Chen, W., & Zhao, J. (2025). Credit anomaly detection method based on Bayesian networks. *Journal of Applied Finance and Banking*, 15(1), 44–59.

A. Dataset Column Descriptions

The tables below describe the key columns used in this project. (Some fields are masked or use special codes; we preprocess those consistently.)

Basic Identifiers

Column	Description
index	Unique identifier for each loan
target	Binary label: 0 = normal, 1 = abnormal

Origination (Static) Variables

Column	Description
CreditScore	Borrower credit score at origination (300–850); 9999 may indicate missing
FirstPaymentDate	First scheduled payment month (YYYYMM)
FirstTimeHomebuyerFlag	Y = Yes, N = No, 9 = Not Available
MaturityDate	Scheduled maturity month (YYYYMM)
MSA	Metropolitan Statistical Area code (may be null)
MI_Pct	Mortgage insurance percentage; 0 = none; 999 may indicate missing
NumberOfUnits	Number of dwelling units (1–4)
OccupancyStatus	P = Primary, I = Investment, S = Second Home, 9 = Not Available
OriginalCLTV	Combined Loan-to-Value ratio at origination
OriginalDTI	Debt-to-Income ratio (%); 999 may indicate missing
OriginalUPB	Original unpaid principal balance (nearest $1,000)
OriginalLTV	Loan-to-Value ratio at origination; 999 may indicate missing
OriginalInterestRate	Note rate at origination (%)
Channel	R = Retail, B = Broker, C = Correspondent, T = TPO Not Specified, 9 = Not Available
PPM_Flag	Prepayment penalty: Y = Yes, N = No
ProductType	FRM = Fixed Rate, ARM = Adjustable Rate
PropertyState	Two-letter state or territory code
PropertyType	SF = Single-Family, CO = Condo, PU = PUD, MH = Manufactured, CP = Co-op, 99 = Not Available
PostalCode	Masked ZIP (first 3 digits + “00”)
LoanPurpose	P = Purchase, C = Refi Cash Out, N = Refi No Cash Out, R = Refi Not Specified, 9 = Not Available
OriginalLoanTerm	Scheduled term in months
NumberOfBorrowers	Number of borrowers (1–10)
SellerName	Entity that sold the loan (“Other Sellers” if below threshold)

ServicerName	Entity servicing the loan ("Other Servicers" if below threshold)
SuperConformingFlag	Indicates "super conforming" eligibility where applicable
PreHARP_Flag	Indicators related to HARP/refinance programs
ProgramIndicator	Program indicator for special loan programs
ReliefRefinanceIndicator	Relief refinance program indicator
PropertyValMethod	Appraisal method (e.g., Full, Desktop/AVM, ACE, ACE+PDR)
InterestOnlyFlag	Y = interest-only payments required, N = otherwise
BalloonIndicator	Y = balloon payment, N = otherwise

Performance Panel (Temporal) Variables

Each loan has 14 months of performance history. For month index $N = 0, 1, \ldots, 13$, the following fields repeat:

Column Pattern	**Description**
N_CurrentActualUPB	Current unpaid principal balance (UPB), incl. any non-interest-bearing portion
N_CurrentInterestRate	Mortgage interest rate in effect for that month
N_CurrentNonInterest-BearingUPB	Non-interest-bearing UPB (e.g., deferred amounts)
N_EstimatedLTV	Estimated Loan-to-Value (ELTV); 999 may indicate unknown
N_InterestBearingUPB	Portion of UPB that accrues interest
N_LoanAge	Months since first payment (or since modification)
N_-MonthlyReportingPeriod	YYYYMM period identifier
N_RemainingMonthsToLegalMaturity	Remaining months to scheduled maturity

Notes

- Static variables give borrower/loan context (credit, terms, property).
- The panel is short (14 months), so we look for *repayment irregularities* rather than long-term trends.
- Special codes are cleaned consistently before modelling.