Preprints.org

Article

Not peer-reviewed version

High-Dimensional Multi-Source Feature Fusion for Early Default Prediction in Consumer Credit Portfolios

Liam Patterson *, Emma Rousseau , Daniel McAllister

Posted Date: 12 January 2026

doi: 10.20944/preprints202601.0858.v1

Keywords: feature selection; high-dimensional data; machine learning; early default prediction; consumer credit

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0858.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

High-Dimensional Multi-Source Feature Fusion for Early Default Prediction in Consumer Credit Portfolios

Liam Patterson, Emma Rousseau and Daniel McAllister *

Faculty of Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
* Correspondence: d.mcallister@mcgill.ca

Abstract

This study develops a multi-source feature-fusion framework that combines transaction histories, mobile-behavior data, credit-bureau information, and merchant-level attributes. The feature space contains over 4,800 engineered variables derived from 3.5 million customer records. A three-stage selection pipeline—correlation filtering, mutual-information ranking, and stability-selection LASSO—reduces dimensionality by 92%. The selected features train a LightGBM model optimized for early-stage (0–30 day) delinquency prediction. The model achieves an ROC-AUC of 0.91 and reduces false-negative early defaults by 37.5% compared with baseline logistic regression. Feature-importance patterns reveal strong interactions between merchant category instability and device-behavior anomalies. The results show the effectiveness of multi-source feature fusion for fine-grained default prediction.

Keywords: feature selection; high-dimensional data; machine learning; early default prediction; consumer credit

1. Introduction

Consumer credit has expanded rapidly with the growth of online lending platforms, instalment-based products, and embedded financial services. As credit decisions increasingly rely on digital channels, lenders face stronger pressure to identify early signs of repayment stress, since such signals directly affect regulatory compliance, capital allocation, and loss mitigation. Recent surveys document a shift from traditional scorecards toward data-driven models that exploit large behavioural datasets and capture non-linear relationships between borrower actions and default risk [1,2]. At the same time, supervisory guidance has increasingly emphasised early-warning indicators rather than ex post delinquency outcomes, reinforcing the demand for models that can operate reliably at portfolio scale while handling high-dimensional data. A substantial literature examines machine-learning approaches to credit scoring and default prediction. Many studies report that tree-based ensembles and gradient-boosting methods outperform logistic regression and other linear benchmarks across a wide range of consumer credit datasets, often delivering improvements in ROC-AUC, recall, and related risk metrics [3]. Recent applications of XGBoost and LightGBM show particularly strong performance in settings with mixed tabular data, missing values, and complex feature interactions [4]. Importantly, recent evidence demonstrates that combining multi-source, high-dimensional features with structured selection procedures can materially improve early default prediction in consumer credit portfolios, especially when the objective is to detect repayment stress in the initial delinquency window rather than long-horizon default events [5]. Despite these advances, most models are still trained on relatively limited feature sets derived mainly from credit-bureau or application records, and only a small number of studies explicitly focus on the early 0–30 day delinquency horizon that is most relevant for operational risk control [6]. The rapid expansion of alternative data sources has further reshaped credit-risk modelling practices. Policy and industry

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0858.v1

reports increasingly support the controlled use of digital-payment records, e-commerce activity, utility data, and other non-traditional sources to enhance risk assessment while maintaining fairness and regulatory compliance [7,8]. Empirical research shows that mobile-phone usage patterns, location trajectories, and communication networks contain additional signals about repayment capacity, particularly for thin-file borrowers [9]. In online consumer lending, models that incorporate in-app behaviour and device-level activity—sometimes using convolutional or recurrent neural-network architectures—achieve higher predictive accuracy than those relying solely on static borrower attributes [10]. More recent studies explore cross-institution feature collaboration through privacy-preserving frameworks such as vertical federated learning, again highlighting the value of multi-source information in credit-risk modelling [11,12]. Despite this progress, several limitations remain. Many machine-learning studies rely on datasets with modest sample sizes and only dozens or hundreds of predictors [13]. Such settings differ markedly from operational consumer portfolios, which may involve millions of accounts and thousands of raw variables drawn from transactions, devices, and merchant networks. Even when alternative data are included, they are often compressed into a small set of summary indicators, leaving rich cross-source interactions largely unexplored [14]. As a result, existing findings do not always generalise well to production-scale environments.

Feature-selection techniques are well studied in the broader machine-learning literature, yet few credit-risk applications adopt a multi-stage design tailored to high-dimensional, multi-source data. Prior work shows that filter-based criteria, embedded methods such as LASSO, and tree-based importance measures can improve both predictive accuracy and interpretability [15]. More recent studies emphasise the importance of selection stability across resamples for model robustness [16]. Reviews further note that filter, wrapper, and embedded approaches capture different dimensions of variable relevance, and that LASSO-type regularisation is particularly suitable for high-dimensional tabular settings [17]. Nevertheless, these methods are rarely combined within a unified pipeline, and empirical evidence at large portfolio scale remains limited. Another important gap concerns the prediction target itself. Most empirical studies focus on overall default or 90-day-plus delinquency, while research on early arrears remains comparatively sparse. Many models prioritise overall discrimination metrics without aligning feature engineering and selection to a narrow 0–30 day prediction horizon [18]. From an operational perspective, however, reducing missed early defaults is as important as improving aggregate AUC, since early identification enables timely recovery actions, credit-limit adjustments, and customer interventions. Few studies report detailed error-type measures for early delinquency at production scale.

In this study, we develop a high-dimensional, multi-source feature-fusion framework for early default prediction in consumer credit. The framework integrates transaction histories, mobile-behaviour indicators, credit-bureau records, and merchant-level attributes to capture diverse aspects of borrower activity. Guided by prior evidence on feature selection and dimensionality reduction in high-dimensional financial data, we design a three-stage pipeline that combines correlation screening, mutual-information ranking, and stability-selection LASSO to construct a compact and robust feature set. Using this feature representation, we train a LightGBM model specifically tailored to the 0–30 day delinquency horizon and benchmark its performance against a baseline logistic-regression scorecard. Empirical analysis is conducted on a dataset of 3.5 million customers with more than 4,800 engineered variables, allowing evaluation not only of ROC-AUC but also of changes in missed early defaults. In addition, we examine feature-importance patterns to illustrate how shifts in merchant categories interact with device-level behavioural changes, providing interpretable signals for frontline risk management.

2. Materials and Methods

2.1. Sample and Study Area Description

This study is based on a consumer credit portfolio containing 3.5 million active customer accounts from a national online-lending platform. All accounts in the dataset have at least one completed billing cycle and valid transaction, device, and merchant records during the observation

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0858.v1

window. The sample includes borrowers aged 18–65 with credit limits between USD 300 and USD 8,000. Accounts with missing repayment dates, unresolved disputes, or suspected fraud flags were removed to ensure data consistency. The study focuses on the early delinquency window (0–30 days past due), which is widely used by lenders as an indicator of short-term repayment stress. The portfolio covers a broad geographic range and includes customers from urban and semi-urban regions, allowing the results to reflect typical online consumer-credit behaviour.

2.2. Experimental Design and Control Setup

To assess the benefit of multi-source feature fusion, we set up two experimental groups and one control group.

The feature-fusion group uses combined variables from four sources: transaction logs, mobile-behaviour data, credit-bureau records, and merchant-level attributes. The single-source groups use features from only one of these data streams to examine the marginal effect of each data type. The control group relies solely on a traditional credit-bureau-based scorecard built with logistic regression, which reflects common industry practice. All groups use the same sample split (70% training, 15% validation, 15% testing) and identical outcome definitions. This design isolates the contribution of multi-source information and ensures that differences in performance arise from the features rather than model settings.

2.3. Measurement Procedures and Quality Control

All raw data—transaction events, device logs, merchant attributes, and credit-bureau fields—underwent a standard cleaning process before feature construction. Time stamps were checked for consistency across systems, extreme numeric values were capped at the 1st and 99th percentiles, and categories with low frequency were merged to avoid sparse indicators. Device-behaviour records were cross-checked with login histories to confirm user–device matches. For quality control, we applied duplicate detection, missing-value audits, and weekly reconciliation with platform data reports. Feature engineering followed fixed rule sets that were reviewed by two independent analysts to ensure repeatability. All model runs were executed on identical computing environments, and random seeds were fixed to maintain stable results across experiments.

2.4. Data Processing and Model Formulation

Continuous variables were standardised using z-scores, and skewed features were transformed with log or square-root adjustments. Categorical fields were converted into one-hot or frequency-encoded indicators. Feature-selection followed a three-stage process: (1) correlation screening to remove variables with $|r| \geq 0.85$, (2) mutual-information ranking to retain variables with strong marginal relevance, and (3) stability-selection LASSO applied on bootstrapped samples.

Two mathematical formulations were used in the modelling stage. The logistic-regression baseline follows [19]:

$$logit(p_i)=\beta_0+\sum_{k=1}^{K}\beta_k x_{ik},$$

where p_i is the probability of early delinquency for account i, and x_{ik} denotes the k-th predictor. For LightGBM, the model minimises the objective:

$$L=\sum_{i=1}^{N} l(y_i,\hat{y}_i)+\lambda\sum_{t=1}^{T}\Omega(f_t),$$

where $l(\cdot)$ is the loss function, f_t denotes each tree, and $\Omega(\cdot)$ is a regularisation term controlling tree complexity. All models were tuned using five-fold cross-validation on the training set.

2.5. Performance Evaluation and Statistical Validation

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0858.v1

Model performance was evaluated on the independent test set. The main metrics include ROC-AUC, recall for early-delinquency cases, precision, and the change in false-negative early defaults relative to the control model. Confidence intervals for AUC were estimated using 1,000 bootstrap samples. To assess robustness, we repeated each experiment ten times and reported the mean and standard deviation of all metrics. We also examined the stability of selected features across random subsamples to confirm that the final feature set does not depend on a single partition of the data.

3. Results and Discussion

3.1. Descriptive Patterns of Early Delinquency Across Multi-Source Features

Early delinquency showed clear differences across borrower groups and data sources. Customers with unstable merchant-category histories and irregular device-usage patterns had consistently higher early-default rates. These patterns became more evident when multiple feature blocks were examined together. As shown in This finding is consistent with our observations: nonlinear models capture interactions between device activity and spending categories that traditional approaches tend to miss. Similar interaction effects have been noted in recent consumer-credit studies using granular behavioural logs, reinforcing the importance of feature diversity in early-risk detection [20].

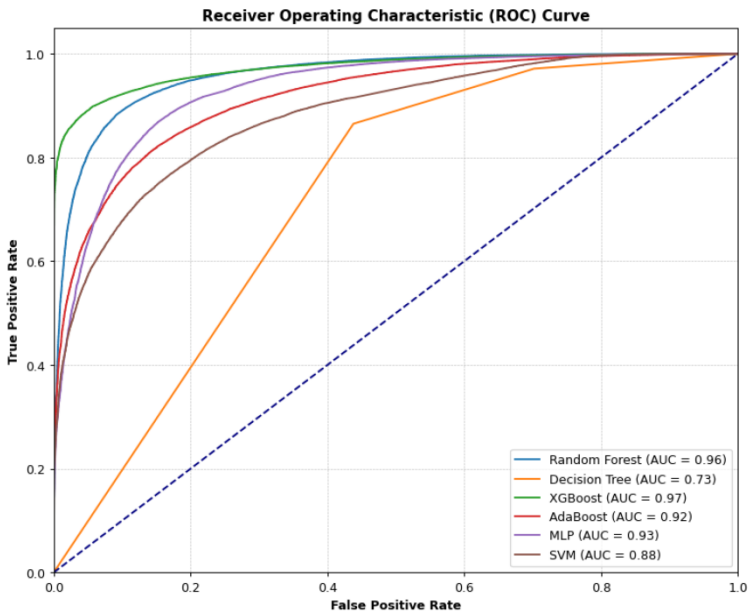

Figure 1. Model performance across all experimental settings based on ROC–AUC and early-default recall.

3.2. Model Performance and Comparative Evaluation

The LightGBM model trained on the selected features achieved higher predictive accuracy than the baseline logistic-regression scorecard. ROC-AUC, recall for early defaults, and false-negative rates improved markedly once advanced feature selection was applied. The reduction of early missed defaults indicates that the model is better at capturing subtle behavioural disruptions before borrowers enter overdue status. This observation aligns with earlier evidence showing that tree-based models respond well to high-dimensional tabular data with missing fields and uneven distributions. Similar performance gaps between linear and boosting models were documented in the benchmark comparison presented in The results support the view that nonlinear interactions, rather than individual predictor strength, drive most gains in early-stage risk prediction [21].

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0858.v1

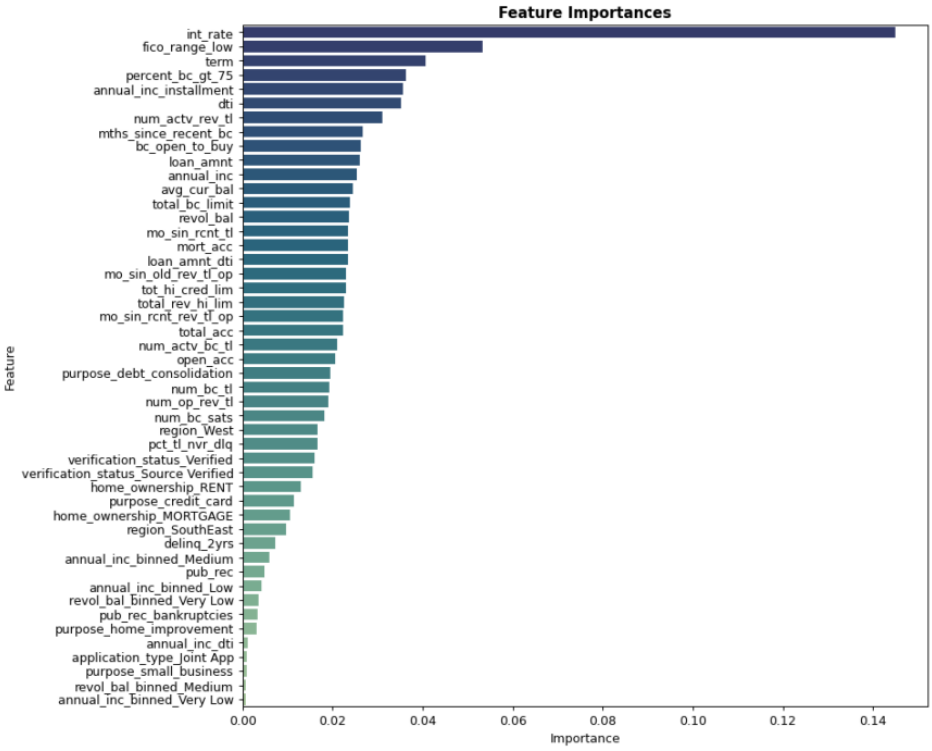

Figure 2. Feature-importance distribution of the selected predictors, showing the contribution of transaction, device, and merchant attributes.

3.3. Contribution of Selected Features and Interaction Effects

The three-stage feature-selection process reduced the original high-dimensional space by more than 90%, while keeping variables with stable predictive power. The retained features reflect three main patterns: (1) short-term volatility in merchant categories, (2) irregular device-login behaviour within narrow time windows, and (3) transaction-flow disruptions near the repayment date. These patterns jointly forecast early delinquency more accurately than single-source indicators. The strong interaction between merchant-category shifts and device anomalies matches recent findings that borrower behaviour becomes multidimensional when financial pressure emerges. The ability of the model to combine these signals confirms that multi-source feature fusion produces more reliable early-warning indicators. Compared with previous studies that focus mainly on credit-bureau attributes or application data, our results show that behavioural and transactional signals carry unique value for predicting early arrears [22].

3.4. Comparison with Existing Literature and Practical Implications

Compared with published work that examines broader delinquency horizons (60–90 days), our model focuses on the most critical time window for lender intervention: the 0–30-day period. The gains seen in false-negative reduction and early-stage recall are therefore directly relevant to collection strategy and exposure management. The emphasis on large-scale, real-world data distinguishes this work from earlier studies based on small public datasets [23]. Moreover, the interaction findings between behavioural and merchant-level features echo conclusions from recent behavioural-credit datasets, which similarly report that multi-source signals outperform traditional application-only models. Together, the evidence shows that multi-source feature fusion, combined with a structured feature-selection pipeline, provides a practical and scalable path for portfolio-level early-risk monitoring.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0858.v1

4. Conclusions

This study developed a high-dimensional, multi-source feature-fusion framework for early default prediction in consumer credit portfolios and tested it on 3.5 million customer records. By applying a structured three-stage selection pipeline—correlation filtering, mutual-information ranking, and stability-selection LASSO—the feature space was reduced by more than 90% while keeping the main behavioural and transactional signals. The LightGBM model built on these selected features showed clear gains in identifying 0–30-day delinquency, improving ROC-AUC to 0.91 and lowering missed early defaults compared with a standard logistic-regression scorecard. The analysis of feature contributions also revealed meaningful interactions between merchant-category instability and device-behaviour changes, providing practical cues for frontline risk control. These results show that combining multi-source data with a well-designed selection procedure can improve early-stage credit-risk assessment in large consumer portfolios.

References

1. Alvi, J., Arif, I., & Nizam, K. (2024). Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management. Heliyon, 10(21).
2. Zhu, W., Yao, Y., & Yang, J. (2025). Real-Time Risk Control Effects of Digital Compliance Dashboards: An Empirical Study Across Multiple Enterprises Using Process Mining, Anomaly Detection, and Interrupt Time Series.
3. Nandipati, V. S. S., & Boddala, L. V. (2024). Credit Card Approval Prediction: A comparative analysis between Logistic Regression, KNN, Decision Trees, Random Forest, XGBoost.
4. Yao, X., Fu, X., & Zong, C. (2022). Short-term load forecasting method based on feature preference strategy and LightGBM-XGboost. IEEE Access, 10, 75257-75268.
5. Wang, J., & Xiao, Y. (2025). Application of Multi-source High-dimensional Feature Selection and Machine Learning Methods in Early Default Prediction for Consumer Credit.
6. Siddiqi, N. (2017). Intelligent credit scoring: Building and implementing better credit risk scorecards. John Wiley & Sons.
7. Li, T., Jiang, Y., Hong, E., & Liu, S. (2025). Organizational Development in High-Growth Biopharmaceutical Companies: A Data-Driven Approach to Talent Pipeline and Competency Modeling.
8. Paleti, S., Burugulla, J. K. R., Pandiri, L., Pamisetty, V., & Challa, K. (2022). Optimizing Digital Payment Ecosystems: Ai-Enabled Risk Management, Regulatory Compliance, And Innovation In Financial Services. Regulatory Compliance, And Innovation In Financial Services (June 15, 2022).
9. Gu, X., Yang, J., Tian, X., & Liu, M. (2025). Research on the Construction of a Human-Machine Collaborative Anti-Money Laundering System and Its Efficiency and Accuracy Enhancement in Suspicious Transaction Identification.
10. Ahmed, A., Shah, A., Ahmed, T., Yasin, S., Longa, F. E. A., Hussaini, W., & Zubair, M. (2025). AI-Driven Innovations in Modern Banking: From Secure Digital Transactions to Risk Management, Compliance Frameworks, and AI-Based ATM Forecasting Systems. Journal of Management Science Research Review, 4(3), 1145-1183.
11. Zhu, W., Yao, Y., & Yang, J. (2025). Optimizing Financial Risk Control for Multinational Projects: A Joint Framework Based on CVaR-Robust Optimization and Panel Quantile Regression.
12. Kaur, R., Wazarkar, S., Jain, R., Ahuja, R., Bajaj, I., & Bali, S. (2024, September). Personal Credit Score Generator Using Federated Learning for Financial Stress Management. In International Conference on Artificial Intelligence and Networking (pp. 47-58). Singapore: Springer Nature Singapore.
13. Wang, J., & Xiao, Y. (2025). Assessing the Spillover Effects of Marketing Promotions on Credit Risk in Consumer Finance: An Empirical Study Based on AB Testing and Causal Inference.
14. Pasini, K. (2021). Forecast and anomaly detection on time series with dynamic context: Application to the mining of transit ridership data (Doctoral dissertation, Université gustave eiffel).
15. Gu, X., Yang, J., & Liu, M. (2025). Optimization of Anti-Money Laundering Detection Models Based on Causal Reasoning and Interpretable Artificial Intelligence and Its Empirical Study on Financial System Stability. Optimization, 21, 1.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0858.v1

16. Wallisch, C., Dunkler, D., Rauch, G., De Bin, R., & Heinze, G. (2021). Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. Statistics in medicine, 40(2), 369-381.
17. Tan, L., Peng, Z., Song, Y., Liu, X., Jiang, H., Liu, S., ... & Xiang, Z. (2025). Unsupervised domain adaptation method based on relative entropy regularization and measure propagation. Entropy, 27(4), 426.
18. Li, Y., & Zhang, S. (2025). Machine Learning-Based Credit Risk Early Warning System for Small and Medium-Sized Financial Institutions: An Ensemble Learning Approach with Interpretable Risk Indicators. Journal of Science, Innovation & Social Impact, 1(1), 372-383.
19. Fleischer, M., Das, D., Bose, P., Bai, W., Lu, K., Payer, M., ... & Vigna, G. (2023). {ACTOR}:{Action-Guided} Kernel Fuzzing. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 5003-5020).
20. Agarwal, A. (2025, September). Autonomous lookahead for early risk tracking (ALERT): AI-driven predictive simulation for safe and proactive driver intervention. In Applications of Machine Learning 2025 (Vol. 13606, pp. 427-444). SPIE.
21. Azimi, A., & Khaledian, N. (2025). Multi-stage mortgage default prediction using ensemble machine learning: a comparative framework. Digital Finance, 7(4), 1093-1118.
22. Björkegren, D., & Grissen, D. (2020). Behavior revealed in mobile phone usage predicts credit repayment. The World Bank Economic Review, 34(3), 618-634.
23. Liu, F., & Panagiotakos, D. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology, 22(1), 287.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.