Data Science in Finance and Economics

https://www.aimspress.com/journal/dsfe

DSFE, 6(4): 622–643.
DOI:10.3934/DSFE.2026022
Received: 13 May 2026
Revised: 12 September 2026
Accepted: 13 October 2026
Published: 09 November 2026

Research article

Machine-learning-based prediction and interpretable analysis of survey-based consumer credit default risk: Evidence from self-reported late payment

Yangting Huang[1], Yifan Liao[1], Ying Li[2], Zihuan Yue[2] and Keyue Yan[1,*]

[1] School of Data Science and Artificial Intelligence, Guangdong University of Finance, Guangzhou 510521, China

[2] School of International and Continuing Education, Beijing Institute of Technology (Zhuhai), Zhuhai 519088, China

* **Correspondence:** Email: conneryan@foxmail.com.

Abstract: This study develops an explainable machine-learning framework for survey-based consumer-credit late-payment risk identification. Using public microdata from Mexico's 2021 Encuesta Nacional de Inclusión Financiera (ENIF), the outcome is defined as self-reported late payment on at least one formal credit product since July 2020. The analytical cohort contains 4549 respondents with valid repayment responses, including 1216 late-payment reports (26.73%). Outcome components, contemporaneous repayment information, formal-credit-holding information, COVID-related late-payment items, routing variables, and survey-design identifiers are excluded from the predictors. Fourteen statistical-learning, bagging, boosting, neural-network, and ensemble candidates are compared under one leakage-controlled protocol: Each training fold selects 25 raw predictors, and the observed class distribution is retained without class resampling. Gradient boosting is selected using repeated cross-validation on training data only, with a mean out-of-fold area under the receiver operating characteristic curve (AUC) of 0.695. On the independent test set, its AUC, Kolmogorov–Smirnov (KS) statistic, and Brier score are 0.706, 0.312, and 0.174, respectively. Permutation importance and Shapley additive explanations (SHAP) analyses identify financial resilience, prior credit constraints, and informal borrowing as important correlates of reported late payment. The findings concern survey-based risk identification and should not be interpreted as an administrative-data probability-of-default score or an automated lending rule.

Keywords: consumer credit; default-risk identification; late payment; machine learning; explainable artificial intelligence; SHAP

JEL Codes: C45, C52, C53, C55, G21, G32

1. Introduction

In recent years, the expansion of internet finance, platform-based economies, and digital consumption scenarios has accelerated the growth of consumer credit. Consumer credit can ease short-term household liquidity constraints, support consumption upgrading, and improve financial inclusion. At the same time, borrower default risk has continued to accumulate. Against the background of income volatility, rising household expenditure pressure, and cross-platform borrowing, consumer credit risk has become more heterogeneous and difficult to identify. Evidence from comparative machine-learning research and credit-risk applications indicates that flexible learners can strengthen nonlinear identification and risk ranking when they are evaluated under a controlled protocol (Fernández-Delgado et al., 2014; Shreya & Pathak, 2025). Accordingly, accurately identifying high-risk borrowers while maintaining risk controllability, approval efficiency, and resource allocation efficiency has become an important issue in financial risk management.

In existing credit risk management practice, traditional credit scoring methods mainly rely on statistical models such as logistic regression and scorecards. These models have transparent structures, clear coefficient directions, and readily interpretable outputs, making them suitable for highly regulated financial contexts. However, they usually impose strong linear assumptions. When applied to high-dimensional survey variables, complex behavioral characteristics, and variable interactions, they may fail to capture latent risk information sufficiently (Fernández-Delgado et al., 2014; Shreya & Pathak, 2025). With the development of data science and artificial intelligence, random forests, support vector machines, extremely randomized trees, neural networks, and ensemble learning have been increasingly introduced into credit risk research and have shown stronger fitting capacity and classification performance in default prediction tasks (Mapfumo & Shongwe, 2026; Pedregosa et al., 2011).

In empirical research and business applications, consumer finance risk modeling is constrained not only by model choice but also by feature engineering, imbalanced class distributions, data-leakage prevention, and model interpretability. Raw survey fields often contain weight variables, low-variance variables, identifiers, and noisy variables weakly associated with the target. If used directly without standardized screening, models may overfit, generalize poorly, and become less applicable in real risk management. Existing studies have examined feature importance, class-imbalance handling, explainable machine learning, and transparency in credit risk management (Saarela & Jauhiainen, 2021; Chawla et al., 2002; Han & Jung, 2024; Bussmann et al., 2020, 2021; Wang et al., 2025). However, the integrated design of the full pipeline from feature selection to model comparison and interpretability analysis still requires further refinement. Accordingly, this study develops a standardized, rigorous, and reproducible machine-learning framework for consumer-credit late-payment risk identification.

The contribution of this study is an auditable survey-data protocol for identifying consumer-credit late-payment risk. Predictor governance separates outcome components, contemporaneous repayment fields, routing variables, weights, and design identifiers from the candidate predictors. In the comparison of fourteen candidate models, the feature budget, sample split, feature-selection rule, and absence of class resampling are held constant; the candidates include transparent statistical baselines, tree ensembles, boosting methods, a neural network, and a stacking ensemble. Explanation is conducted through agreement among global importance measures, Shapley additive explanations (SHAP) effect distributions, and representative local SHAP cases. All displayed variables use the plain-language meanings of the questionnaire items rather than opaque field codes.

The empirical design also distinguishes discrimination from an operating decision. A risk model may rank respondents effectively, although different institutions choose different alert thresholds according to intervention capacity and the relative costs of missed payment difficulty and unnecessary outreach. The study therefore reports threshold-free discrimination and probability-quality measures alongside threshold-dependent classification measures. This distinction is particularly important for the present outcome: Late-payment reports are less frequent than observations without reported repayment difficulty, so predicting the majority class alone can yield a high overall accuracy, whereas a useful alert model should identify a meaningful proportion of late-payment respondents. The operating threshold is selected from training-only out-of-fold predictions, and the independent test set is never used for threshold optimization.

2. Literature review

Research on consumer credit default prediction and credit risk identification can generally be divided into three strands: traditional credit risk models, machine-learning-based credit risk models, and model interpretability. Traditional statistical models, including logistic regression and scorecards, have long been used in banking and consumer finance risk management. Their main advantages are stability, transparency, and interpretability, but their fitting capacity is relatively limited in complex, nonlinear settings. By contrast, machine learning methods such as random forests, support vector machines, neural networks, and ensemble learning are better suited to high-dimensional features, complex interactions, and class-imbalance problems and can improve predictive performance when model comparison is conducted consistently (Fernández-Delgado et al., 2014; Shreya & Pathak, 2025).

Across the credit risk modeling pipeline, data preprocessing and feature engineering are critical prerequisites for model reliability. Existing studies commonly use missing-value treatment, outlier identification, redundant-variable removal, variance filtering, and correlation analysis. Some studies further use statistical indicators such as the information value (IV) and the Gini coefficient for preliminary variable screening to reduce dimensional redundancy and noise. These approaches can simplify data structure and improve training efficiency (Saarela & Jauhiainen, 2021), but preprocessing may be fragmented, survey design variables may be retained inconsistently, full-sample selection can cause leakage, and purely statistical filters may not align with the prediction target (Pedregosa et al., 2011).

Existing research also suggests that the evidential strength of credit-risk studies depends not only on the algorithms considered but also on whether the information boundary is clearly defined. Administrative credit records can typically observe contractual status, repayment timing, and account balances, whereas household surveys are better suited to describing financial resilience, financial access, and informal coping practices. The outcome definition, sample scope, and permissible operational inferences therefore differ between the two data sources. Directly equating self-reported late payment in a survey with an account-level bank probability of default or retaining overlapping information between outcome components and predictors may overstate the practical meaning of a model. For survey data, a more appropriate research question is whether risk signals associated with late payment can be identified within a clearly defined eligible cohort, with the scope of inference stated explicitly.

Class imbalance can substantially affect a model's ability to identify late-payment observations in consumer-credit prediction. Common responses include random over- or undersampling, synthetic methods such as the synthetic minority oversampling technique (SMOTE) and adaptive synthetic

sampling (ADASYN), and algorithm-level approaches such as class weighting or cost-sensitive learning (Chawla et al., 2002; Han & Jung, 2024). These methods change the effective training distribution: Oversampling may amplify idiosyncratic observations, undersampling may discard information, and synthetic observations can be difficult to interpret for sparse or noisy survey responses. The present experiment therefore preserves the observed class distribution for every candidate model so that performance differences can be attributed to differences among learners under a common protocol.

In comparative designs, previous work often reports the best result obtained by each algorithm under different numbers of variables, data partitions, or imbalance treatments. Although this practice can illustrate model potential, it is difficult to determine whether a performance difference is attributable to the learner itself or to feature availability, sampling strategy, or threshold choice. To improve comparability, candidate models should share the same training/test split, information-governance rule, and fold-internal feature-selection procedure; model selection, threshold determination, and final testing should also be separated into distinct information stages. This approach shifts algorithm comparison from a display of individually optimized settings to an empirical judgment under controlled conditions.

In terms of model selection and algorithmic development, consumer credit risk prediction has evolved from linear models to nonlinear models, from single models to ensembles, and from conventional machine learning to deep learning. Linear models such as logistic regression and scorecards are highly interpretable but limited in capturing complex nonlinear relationships and higher-order interactions (Mapfumo & Shongwe, 2026). Tree-based and ensemble models, including random forests, Extra Trees, and gradient boosting trees, have clear advantages in predictive accuracy, robustness, and feature-importance output and are therefore common choices in current research and practice. Recent machine-learning research has also introduced deep neural representations, graph-based learning, and automated model-search designs (Schmitt, 2024; Kipf & Welling, 2017). However, complex models also face practical constraints, including larger parameter scales, higher training costs, greater difficulty of interpretation, and higher deployment thresholds. Their performance gains therefore do not necessarily translate into business usability (Monje et al., 2026).

Explainable machine learning has become an important research direction in financial risk management and an essential approach for addressing the black-box problem of high-accuracy models under regulatory compliance requirements (Bussmann et al., 2020, 2021; Lipton, 2016). Existing interpretability research distinguishes inherently interpretable models from post-hoc explanation methods (Rudin, 2019; Doshi-Velez & Kim, 2017). Logistic regression and shallow decision trees are transparent but may have limited predictive capacity; SHAP provides a coherent additive-attribution framework for global importance, effect direction, and local attribution (Lundberg & Lee, 2017; Gramegna & Giudici, 2021). More broadly, explainable artificial intelligence (XAI) research emphasizes that explanation methods need explicit scope, audience, and evaluation criteria (Arrieta et al., 2020); local explanations such as local interpretable model-agnostic explanations (LIME) make individual predictions inspectable but do not by themselves establish global reliability (Ribeiro et al., 2016). Explanation should therefore be interpreted jointly with feature governance, model selection, and validation (Guidotti et al., 2018).

Interpretability should not be equated with assigning causal meaning to a variable ranking. Permutation importance, model-based importance, and SHAP address different questions: The first examines whether predictive quality deteriorates after a variable association is disrupted, whereas the latter two describe the fitted model's usage pattern and an individual prediction's contribution relative to

the baseline. When survey items measure related aspects of financial capability or household circumstances, correlated variables may share importance. A more robust approach is to assess agreement across multiple measures and to summarize the findings as recurrent risk-related themes rather than as independent causal effects of single questionnaire items.

This study addresses the above limitations through explicit outcome construction, leakage control, and one controlled benchmark rule. The outcome is interpreted as evidence of self-reported late-payment risk in a household survey and is clearly distinguished from an operational probability-of-default model constructed from administrative repayment records.

3. Research design

This study uses the 2021 Encuesta Nacional de Inclusión Financiera (ENIF) public microdata for Mexico. The prediction target is not formal-credit experience; it is self-reported late payment on at least one formal credit product since July 2020. The outcome is constructed from valid repayment responses for formal credit products in Q6.4; its component items, related follow-up questions, and fields directly concerning contemporaneous repayment are excluded from the predictors. Respondents with at least one valid repayment response form the eligible analytical cohort (4549 respondents); 1216 respondents report late payment, and 3333 report no late payment. This is a survey-defined repayment-risk cohort rather than a representative portfolio of bank customers. The data are split in a stratified manner into an 80% training set and a 20% independent test set. The training data are used for fold-internal feature selection, model estimation, and threshold determination; the test set is reserved for final performance validation and explanation after the model has been locked.

ENIF 2021 was fielded in Mexico from June 28 to August 13, 2021 using a probabilistic, stratified, multistage household-sampling design. The outcome question asks, for each formal credit product reported in Q6.2: "Since July 2020 to the present, have you been late in paying for your [reported product]?" The substantive response options are yes and no; nonsubstantive responses are not used to define eligibility. The indicator equals one when at least one of the nine Q6.4 product items records a yes response and equals zero when at least one valid item is observed, and none records late payment.

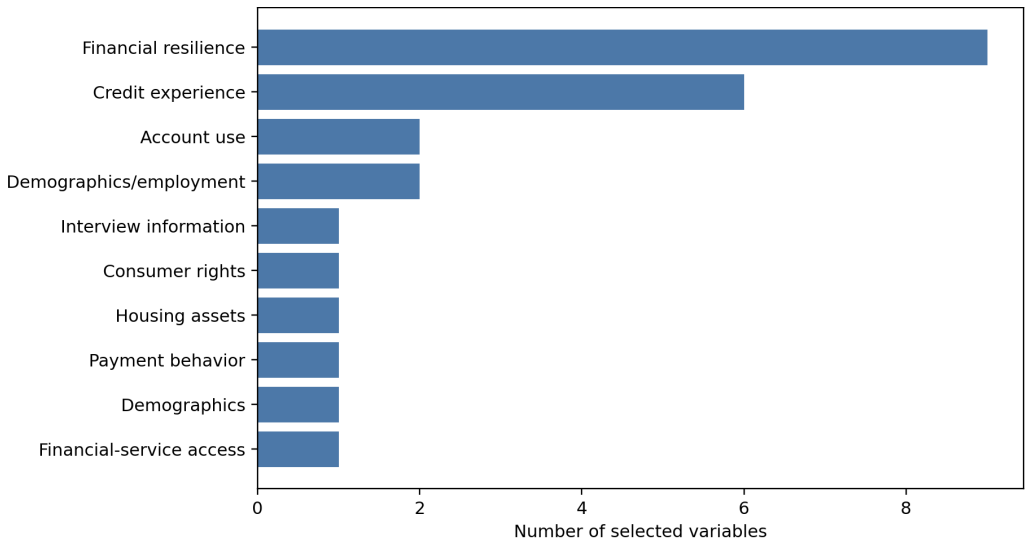


Figure 1. Questionnaire domains retained for predictor governance.

Figure 1 organizes the candidate survey fields into demographic characteristics, household circumstances, financial resilience, formal-financial access, credit experience, borrowing practices, and financial knowledge. The labels describe the respondent-facing meaning of a question rather than an internal questionnaire code. This organization is substantive as well as technical: Repayment outcomes and repayment-adjacent questions are separated from predictors, and the remaining domains provide potential antecedents of payment difficulty. Categorical questions are represented through encoded response indicators during estimation; the figures report the underlying question meaning once rather than repeating opaque dummy-variable labels.

Table 1. Substantive meaning of retained predictor domains.

Domain	Examples of respondent-facing predictor meanings
Financial resilience	Whether monthly income covers expenses, money left at month end, and months that savings could cover expenses
Credit access and experience	Prior formal-credit denial, perceived ability to repay, and reasons for a credit decision
Financial-service use	Use of formal products and access to retail or financial-service outlets
Household and personal circumstances	Age, marital status, employment-related coverage, and confidence in obtaining desired goods
Borrowing practices	Use of informal borrowing from friends or relatives

Table 1 gives the substantive interpretation behind the predictor-governance scheme. Financial-resilience items describe the household's immediate capacity to absorb income or expenditure shocks; credit-access items capture past restrictions on formal liquidity; financial-service and household-circumstance items provide the institutional and demographic context in which repayment difficulty may arise; and informal borrowing records a potential coping channel. These domains are not asserted to be causes of late payment, nor are all listed items necessarily retained in every fold. Rather, the table specifies the permissible pool from which the fold-internal selection procedure draws predictors, thereby making the feature boundary auditable and the graphical labels interpretable.

Predictors are governed before model fitting to avoid outcome leakage. The repayment items used to construct the outcome, their frequency follow-ups, formal-credit holding items, directly contemporaneous repayment information, COVID-related late-payment items, survey-routing variables, weights, and geographic design identifiers are excluded. Within each training fold, variables with excessive missingness, no variation, or near-zero variation are screened out. Extra Trees then ranks the remaining raw variables, and the top 25 are retained. The same fold-specific feature-selection procedure is applied to every candidate model.

The value $K = 25$ is a prespecified common parsimony budget, not a claim that 25 is a universally optimal number of predictors. It gives every candidate the same information budget while limiting the dimension of a survey-based model. Because ranking is recomputed within each training fold, the exact set of 25 may differ across folds; Figure 3 therefore reports selection frequencies rather than incorrectly presenting one fold-specific list as universal. After development is closed, one final 25-predictor set is refitted on the complete training partition only for the single locked test evaluation. Alternative feature budgets are a worthwhile robustness exercise but were not selected using the independent test set and are left for future work.

The analytical cohort is deliberately not described as a representative customer portfolio. ENIF is a nationwide Mexican financial-inclusion survey, whereas the present cohort is conditional on valid responses to the formal-credit repayment module. Restricting the survey to this eligible group is necessary to define the outcome, but it changes the estimand: The findings characterize associations with reported late payment among eligible respondents, not the population probability of default (PD) of all borrowers or all Mexican households.

This distinction is important for both interpretation and reproducibility. A household survey records respondents' answers in a defined reference period and cannot provide the contractual performance history, account balances, delinquency days, collection actions, or approval policy that would be available to a lending institution. The outcome should consequently be read as a reported late-payment indicator, not as a bank's regulatory probability-of-default estimate. At the same time, the survey perspective has value because it makes financial resilience, formal access, and informal coping practices observable in one harmonized questionnaire. The analysis therefore asks whether these respondent-reported circumstances can rank late-payment risk within the eligible cohort while deliberately avoiding claims about automated approval, pricing, or adverse action.

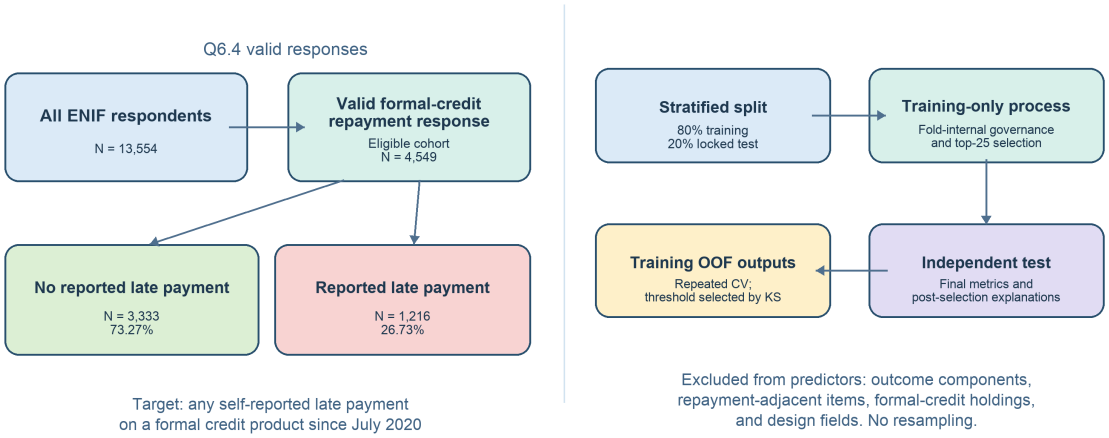


Figure 2. Analytical-cohort construction and leakage-controlled evaluation workflow.

Figure 2 separates two distinct design decisions. The left panel shows that the analytical cohort is formed only by respondents with a valid response to the formal-credit repayment question: 1216 of the 4549 eligible respondents report late payment, and 3333 do not. The right panel records the order of evaluation safeguards: A stratified split is made before model development, predictor governance and top-25 selection occur inside training folds, training-only out-of-fold (OOF) predictions support repeated validation and threshold selection, and the locked test set is used only after the primary model is fixed. Thus, eligibility for the outcome is not itself treated as a predictor, and no test-set information enters feature selection or threshold determination.

The controlled comparison includes logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), random forest, Extra Trees, support vector machine with radial basis function (RBF-SVM), bagging trees, AdaBoost, gradient boosting, HistGradientBoosting, XGBoost, light gradient-boosting machine (LightGBM), multilayer perceptron (MLP), and stacking. All models use the same stratified split and the same fold-internal feature-selection procedure, with no oversampling, undersampling, or synthetic-sample generation. Let $\hat{p}_i = \Pr(Y_i = 1 \mid X_i)$ denote the predicted probability

of reported late payment. The classification indicator is $\hat{y}_i = \mathbb{I}(\hat{p}_i \geq t)$, where t is selected from training-only OOF predictions by maximizing the Kolmogorov–Smirnov (KS) statistic. After the training-only selection stage is closed, the same Extra Trees ranking is refitted on the complete training partition to obtain one final 25-predictor set. Each candidate is then refitted on that partition using this fixed final set, and the locked test partition is scored once for descriptive final validation; these test results do not determine model, feature, or threshold choice.

For reproducibility, the primary gradient boosting specification uses 300 estimators, learning rate 0.025, maximum tree depth 2, minimum leaf size 12, and subsampling rate 0.80; QDA uses a regularization parameter of 0.15. The stacking candidate combines logistic regression, gradient boosting, and Extra Trees base learners through a logistic-regression metalearner, with three-fold internal cross-validation and probability outputs as the meta-level inputs. It is evaluated under exactly the same split, fold-internal selector, 25-predictor budget, and no-resampling rule as the other candidates; it is not given a separate feature set or privileged tuning process.

To evaluate model performance comprehensively and objectively, this study constructs a multidimensional metric system consisting of Accuracy, Precision, Recall, F1-score, area under the receiver operating characteristic curve (AUC), the KS statistic, and the Brier score. The calculation formulas are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR}, \tag{5}$$

$$\text{KS} = \max_t \left|\text{TPR}(t) - \text{FPR}(t)\right|, \tag{6}$$

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_i - y_i)^2 . \tag{7}$$

These metrics evaluate model performance from the perspectives of overall classification accuracy, precision of late-payment alerts, coverage of late-payment observations, balance between Precision and Recall, risk-ranking ability, and probability-prediction quality. The Brier score is a proper scoring rule that reflects both calibration and resolution; calibration is additionally examined with the calibration curve. They are aligned with the dual requirements of prudence and practical usability in financial risk management.

The metrics have distinct roles in the reporting design. AUC and KS evaluate ranking separation without fixing one operating point, and the Brier score evaluates probability-prediction quality rather than calibration alone. Accuracy, Precision, Recall, and F1 depend on the threshold and are therefore reported for the training-selected operating threshold. The complete seven-metric panel is given for the selected model on the independent test set; the repeated cross-validation (CV) model-selection table focuses on AUC, Brier score, and KS because those are the criteria used before the test set is accessed.

No single metric is treated as a sufficient measure of usefulness. In a sample with fewer observations in the late-payment class than in the nonevent class, accuracy at an arbitrary threshold can appear satisfactory even when a model identifies few late-payment observations. AUC summarizes the ordering of predicted risks across all thresholds, KS captures the maximum separation of the two cumulative score distributions, and the Brier score penalizes inaccurate probability assignments. Precision, recall, and F1 make the consequences of the selected operating point visible. Reporting this panel prevents a presentation in which a favorable accuracy or AUC is detached from coverage of the late-payment class or from the quality of the estimated probabilities.

The explanation framework uses permutation importance on the independent test set, model-based feature importance, and SHAP values for the selected gradient-boosting model. Permutation importance assesses the loss in predictive quality after a variable is disrupted, and SHAP provides global rankings, signed effect distributions, and additive local attributions. These explanations are conducted only after the primary model, predictor-governance rule, and operating threshold have been fixed; they play no role in model selection, feature selection, or threshold determination. Agreement among the global rankings and the stability of selected features are reported, so the interpretation is evidence about robust correlates within this survey rather than a visual description alone.

4. Empirical results

4.1. Comparison of multiple experimental results

The empirical comparison is designed to isolate algorithmic differences rather than to compare different feature counts or class-resampling settings. All 14 candidates receive the same training/test split and the same fold-internal 25-variable selection budget. A prespecified, two-stage, training-only rule is used: The initial five-fold OOF comparison screens all 14 candidates, and the two leading candidates, gradient boosting and QDA, then undergo 10 repeated five-fold cross-validation runs. Gradient boosting has a mean repeated-OOF AUC of 0.695 (2.5–97.5% empirical interval: 0.692–0.698) and a mean Brier score of 0.177. QDA has a mean AUC of 0.690 (0.685–0.695) and a mean Brier score of 0.194. Gradient boosting achieves a higher AUC in each repeated-CV run and is selected as the primary model before final test evaluation.

The repeated-validation stage is a confirmation exercise for the two candidates identified by the initial all-model screen; it is not a complete repeated-CV ranking of all 14 candidates. The observed AUC difference between gradient boosting and QDA is modest (0.005), and the manuscript therefore does not interpret it as evidence of a universally or statistically significant superior algorithm. Gradient boosting is retained as the primary model under the prespecified development rule because it has the higher mean repeated-OOF AUC, the lower mean Brier score, and a higher AUC in each of the repeated runs. A full repeated-CV comparison of all candidates, accompanied by formal paired inference, would be a useful extension but is outside the locked development protocol used for the present independent test evaluation.

Table 2 is the formal basis for selecting the primary model. Gradient boosting exceeds QDA by 0.005 in mean OOF AUC (0.695 versus 0.690), has a lower mean Brier score (0.177 versus 0.194), and has a slightly higher mean KS statistic (0.289 versus 0.282). The difference is modest and is therefore not interpreted as a universal dominance claim. Its relevance is that the advantage is obtained under the same split, the same fold-internal 25-predictor budget, and the same no-resampling rule; the decision consequently reflects repeated discrimination and probability-quality evidence rather than a single favorable run.

Table 2. Repeated-CV comparison of the leading candidates under a common protocol.

Model	Mean OOF AUC	Mean Brier	Mean KS	Selection status
Gradient boosting	0.695	0.177	0.289	Selected
QDA	0.690	0.194	0.282	Comparator

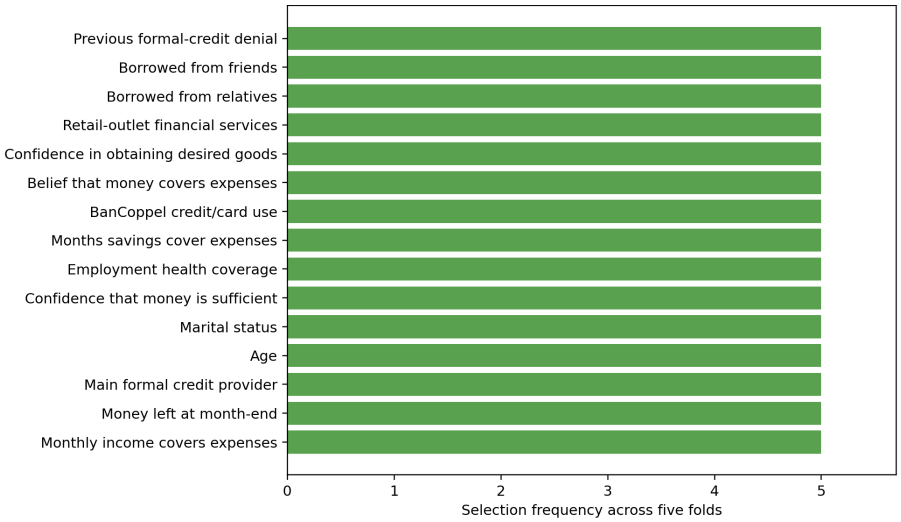


Figure 3. Fold-level stability of the common 25-predictor selection budget.

Figure 3 shows the variables that were selected most consistently across the initial five training folds; it is not a complete list of all 25 variables retained in each fold. Prior formal-credit denial, borrowing from friends or relatives, access to retail financial services, and indicators of financial sufficiency recur at the top of the plot, which is consistent with the subsequent household-finance interpretation. At the same time, selection frequency is a stability diagnostic rather than a causal ranking: Related survey questions may share information, and a feature can be useful without being selected in every fold. The common 25-predictor budget remains fixed for every candidate model, so the comparison does not rank models using their best result under different experimental settings.

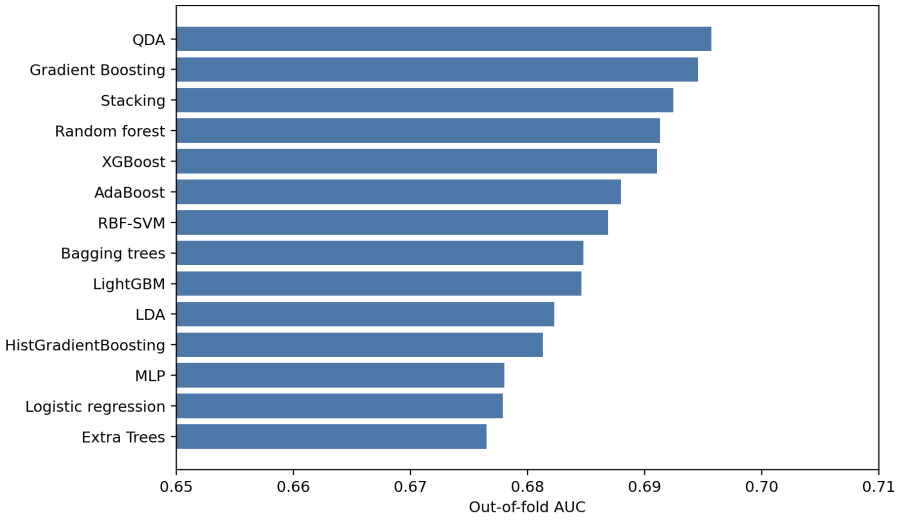


Figure 4. Cross-validated comparison of candidate models under the common protocol.

Figure 4 is the initial five-fold OOF screening comparison under the common protocol. QDA is marginally higher than gradient boosting in this single screening partition; it is therefore not used as the final selection rule. The plot is useful for showing that the shortlisted methods are close competitors rather than that one method dominates every possible training partition. The repeated-CV evidence in Table 2, obtained before test access, shows that gradient boosting has the higher mean AUC and lower mean Brier score across the repeated folds. Model selection consequently follows the repeated-CV summary rather than a one-run maximum score.

4.2. Main model selection and results

Gradient boosting is the primary model under the repeated-CV selection rule. On the independent test set, it achieves Accuracy = 0.675, Precision = 0.424, Recall = 0.609, F1-score = 0.500, AUC = 0.706, KS = 0.312, and Brier score = 0.174. The operating threshold, 0.290, is determined only from training out-of-fold predictions. Accuracy is interpreted jointly with Recall and KS: Raising the threshold to 0.50 increases the accuracy mechanically but reduces the recall from 0.578 to 0.151 in the training OOF analysis.

The comparison also provides an independent-holdout consistency check. The mean training out-of-fold AUC and Brier score for gradient boosting are 0.695 and 0.177, whereas the corresponding independent-test values are 0.706 and 0.174. The test results are therefore not materially weaker than the repeated-CV evidence. This agreement does not prove an absence of overfitting, but it is consistent with the view that the selected model does not rely solely on one favorable training split.

The magnitude of these results should be interpreted in relation to the task. AUC values near 0.70 indicate useful but imperfect discrimination in a self-reported survey setting, where repayment behavior is affected by unobserved contractual terms, household shocks, and measurement limitations. The modest gap between OOF and independent-test performance is more informative than an isolated high score because it indicates that model selection did not rely on a split-specific result. The selected threshold is likewise a transparent operating choice rather than a universal cutoff. Organizations using a related diagnostic would need to choose a threshold according to the cost of missed payment difficulty, the burden of outreach, and the availability of human review; none of these institutional choices is identified by the present survey alone.

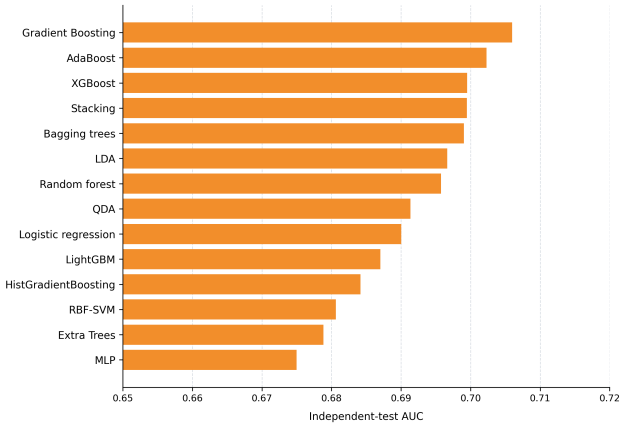


Figure 5. Independent-test AUC comparison of all candidate models under the common protocol.

Figure 5 presents the independent-test AUC of all candidate models under the common protocol. Gradient boosting has the highest displayed test AUC, but this panel is descriptive rather than a second selection exercise: It is reported after the repeated-CV rule has already fixed the primary model. The main value of the figure is therefore transparency. It lets readers assess whether the training-only decision remains plausible on the locked holdout while preserving the principle that a one-time test result must not feed back into model, feature, or threshold selection.

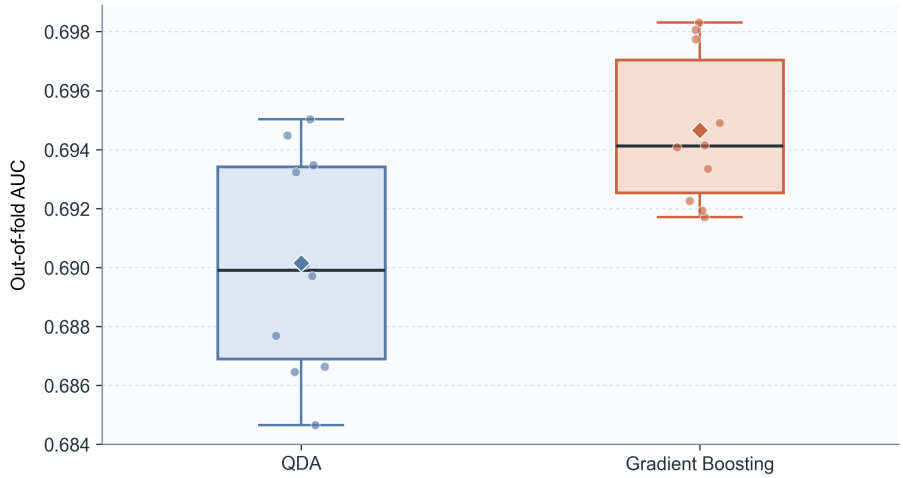


Figure 6. Repeated cross-validation AUC distributions for QDA and gradient boosting.

Figure 6 shows the out-of-fold AUC distributions of QDA and gradient boosting across 10 repeated five-fold cross-validation runs. Boxes represent interquartile ranges, horizontal lines represent medians, diamonds represent means, and points represent the mean out-of-fold AUC from each repeated run. The mean AUC of gradient boosting is 0.695, compared with 0.690 for QDA; its distribution is shifted upward while remaining concentrated. This matters because the model decision is not based only on the difference between two averages: The repeated runs show that the result is not created by one unusually favourable fold allocation. The figure should nevertheless be read as internal stability evidence under one survey split, not as proof that the same ranking will hold across periods, countries, institutions, or administrative loan portfolios.

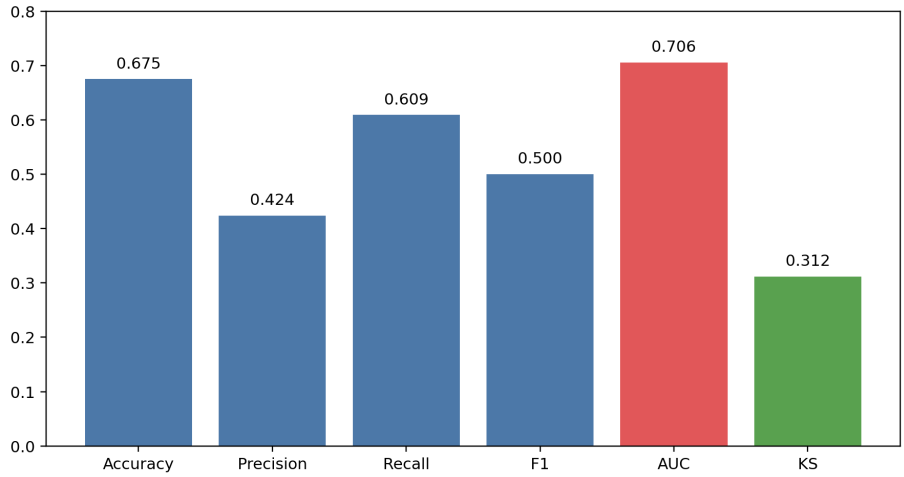


Figure 7. Independent-test metrics for gradient boosting.

Figure 7 reports threshold-dependent and threshold-free metrics jointly. Accuracy (0.675) is not treated as the sole indication of performance because the non-late-payment class is more frequent in the analytical cohort. At the training-selected threshold, Recall is 0.609, meaning that the model identifies a material share of respondents reporting late payment; Precision is 0.424, so many alerts are necessarily false positives in this survey setting; and the F1-score of 0.500 summarizes that trade-off. The relatively stronger AUC (0.706) and KS (0.312) show that threshold-free ranking and separation should be distinguished from the performance of a particular alert cutoff.

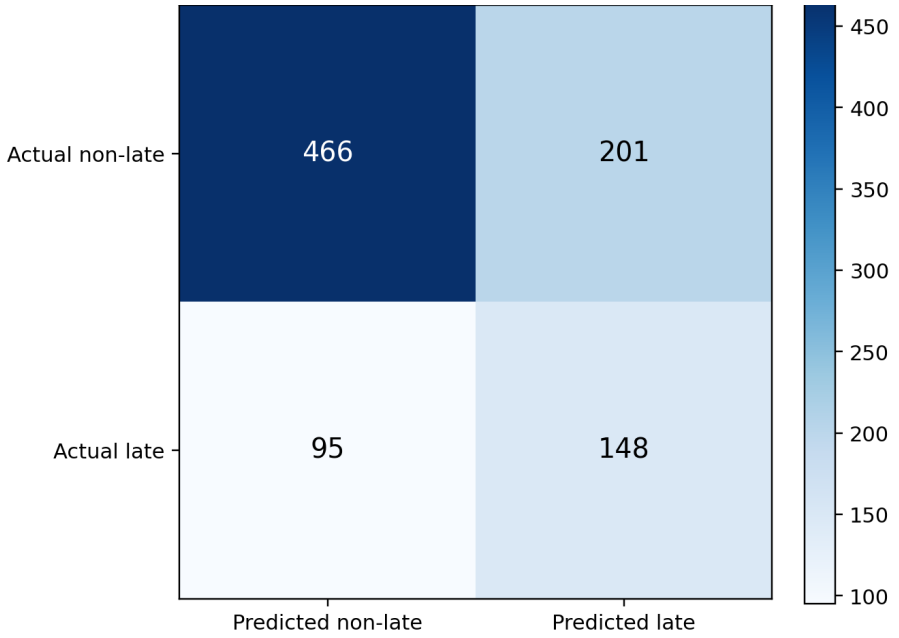


Figure 8. Independent-test confusion matrix at the training-selected threshold.

Figure 8 makes the associated false-positive and false-negative trade-off visible at the threshold chosen from training-only OOF predictions. Of the 243 late-payment reports in the test subset, 148 are identified, and 95 are missed; among the 667 non-late-payment reports, 201 are alerted. These counts explain why recall and precision must be discussed together. A lower threshold could capture more reported late payment at the cost of more outreach to nonevents, whereas a higher threshold would reduce alerts but miss more potential payment difficulty. The present threshold is a transparent training-derived operating point, not a recommendation for automated approval, denial, or pricing.

4.3. Overall Discussion of the Experimental Results

A strict data-leakage prevention procedure is used during model training. Cross-validation is established within the training set; variable screening and ranking are completed separately inside each training fold; out-of-fold predictions determine the threshold; and the locked test set is not accessed until the primary model is fixed. It is not used for feature selection, hyperparameter tuning, threshold selection, or model choice. The test set is used only for final performance reporting and locked-post hoc explanation and does not feed back into any development decision.

This procedure evaluates generalization to observations from the same survey that were not used for training; it is not temporal, geographic, or institutional external validation. Because some variables in consumer credit survey data may be strongly associated with the target variable, completing feature

screening and threshold optimization on the full sample before model evaluation would easily overestimate model performance. Fold-internal feature selection and OOF threshold optimization are therefore used to reduce this risk.

The results show that survey-based late-payment identification depends on target definition, predictor governance, model selection, and transparent validation. Gradient boosting is selected by average repeated-CV discrimination and probability quality under one common protocol; it is not selected from a collection of incomparable feature counts or class-resampling settings.

The empirical section therefore contributes not only by reporting model performance but also by demonstrating how a leakage-controlled design can be applied to survey-based consumer-finance data. Its contribution lies not merely in selecting a relatively well-performing model but in providing a standardized, reproducible, and interpretable pathway for survey-based late-payment risk identification.

5. Core variables and explainability analysis

5.1. Identification of key risk variables and global feature importance based on tree models

Global importance is assessed for the selected gradient-boosting model using independent-test permutation importance and model-based importance. Across these measures, the recurring domains are financial resilience and financial exclusion, reflected in whether monthly income covers expenses, prior credit constraints, informal borrowing, insurance and retirement-savings participation, and access to formal financial services. These patterns are associations in the eligible survey cohort and are not causal claims or applicant-level lending rules.

The substantive interpretation is consistent with a household-finance mechanism. Respondents reporting that income does not comfortably cover expenses or that savings cover few months of expenses may face tighter liquidity buffers when an income or expenditure shock occurs. Previous formal-credit denial and credit-bureau-related denial can indicate restricted access to conventional liquidity, whereas borrowing from friends or relatives may signal reliance on informal coping channels. Insurance and retirement-savings participation are interpreted as markers of financial planning and formal-system engagement, not as direct causes of repayment. In a nonlinear model, these dimensions can matter jointly: For example, low perceived financial sufficiency may convey a different risk signal when it occurs alongside limited savings and informal borrowing than when it appears alone.

Correlation among survey responses is expected because several questions measure related aspects of financial capability. The interpretation therefore does not treat a single rank as a unique structural effect. The agreement of permutation, model-based, and SHAP rankings, together with fold-level selection frequencies, is used as a robustness check. Where related variables carry overlapping information, their importance may be shared; the relevant conclusion concerns the broader financial-resilience domain rather than an assertion that one questionnaire item is independently causal.

The figures should therefore be interpreted as a connected explanatory record. Permutation importance asks whether predictive performance deteriorates when the observed association of a variable is broken. Model-based importance records how frequently and effectively a variable is used by the fitted ensemble. SHAP values show the direction and magnitude of individual contributions relative to the model baseline. Their convergence supports the conclusion that liquidity pressure, prior credit constraints, and informal borrowing are recurrent risk-related domains in this cohort. Divergence would not necessarily invalidate the model, but it would warn against a strong substantive interpretation of one

graphical ranking. This triangulation is particularly useful for survey features, whose categories may overlap conceptually and whose associations need not be stable across countries, periods, or credit products.

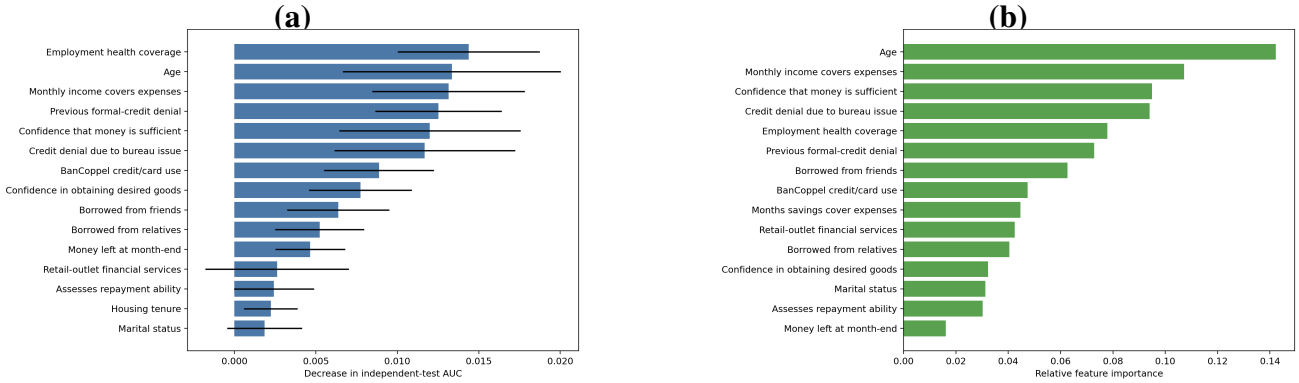


Figure 9. Global importance diagnostics for gradient boosting: (a) independent-test permutation importance and (b) model-based feature importance.

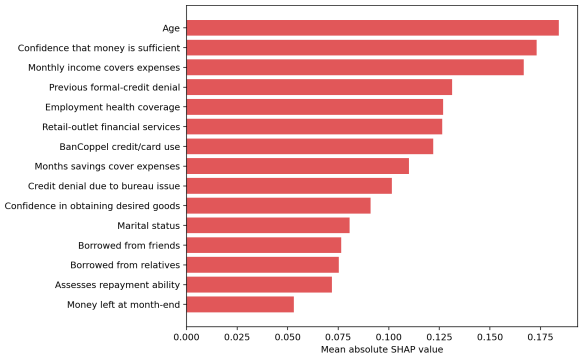


Figure 10. Global SHAP importance ranking for gradient boosting.

Figures 9 and 10 present complementary global views: perturbation-based importance, fitted-model importance, and SHAP importance. The permutation panel asks whether predictive performance declines when a feature's observed association is disrupted, whereas the fitted-model panel records the role used by the ensemble and the SHAP panel reports average contribution magnitude. The recurring prominence of financial sufficiency, prior formal-credit constraints, and informal borrowing across these distinct calculations strengthens the interpretation at the domain level. It does not turn any one item into a causal determinant: Correlated questionnaire items can exchange or share importance, and the rankings remain conditional on this cohort, this outcome definition, and the locked model.

5.2. Global and local explanation based on SHAP

The SHAP analysis makes the model's associations inspectable at two levels. The global summaries show the relative magnitude and direction of effects across respondents. Domain aggregation translates the individual questionnaire items into meaningful financial-resilience, financial-access, borrowing, household, and demographic dimensions. The agreement plot and the feature-selection frequency plot check whether the explanation rests on a single importance measure or a single training/validation split. This distinction is important for practice: The survey variables can support screening, outreach, or financial-capability interventions in a study population, but they cannot substitute for contractual repayment histories in a production credit-decision system.

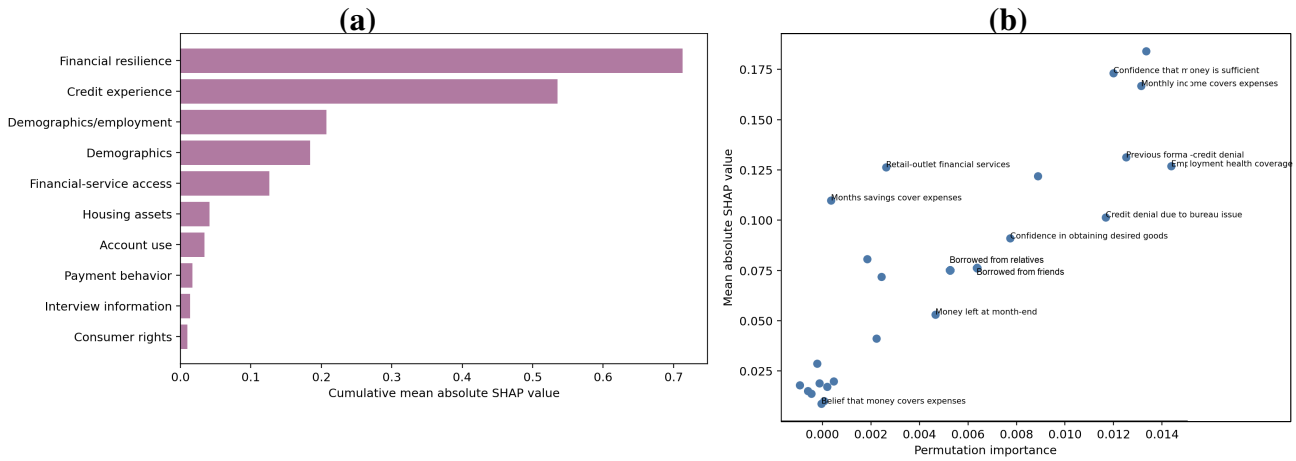


Figure 11. Cross-checks for global explanation: (a) Questionnaire-domain contribution summary and (b) Agreement of global importance measures.

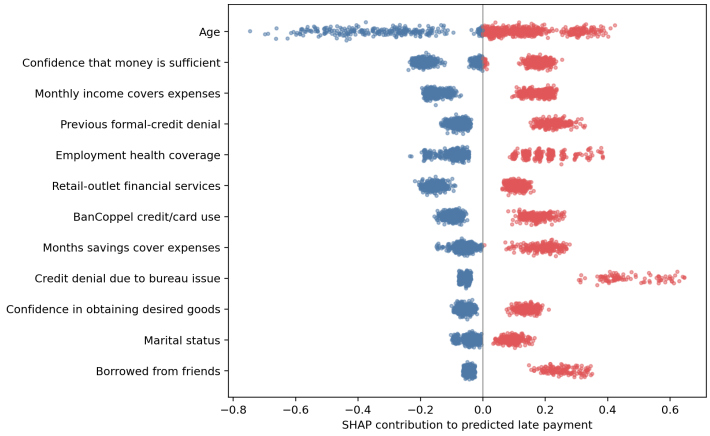


Figure 12. SHAP value distributions for the leading questionnaire variables.

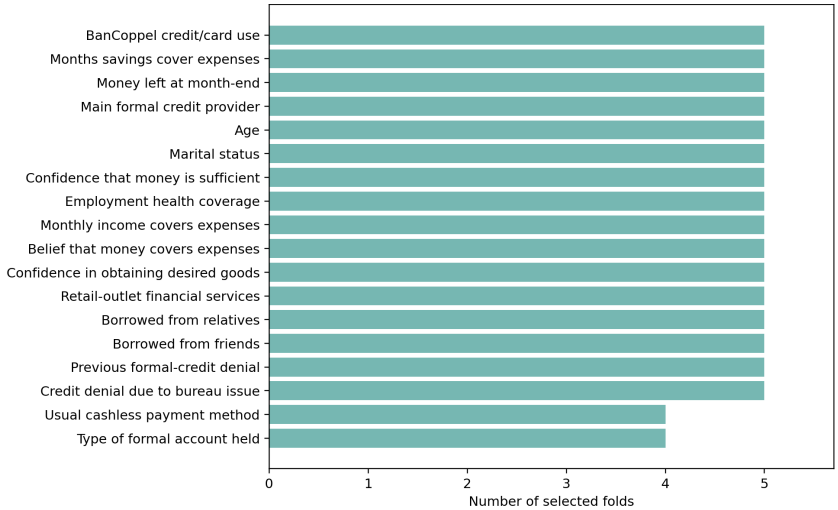


Figure 13. Frequency with which leading predictors are selected across folds.

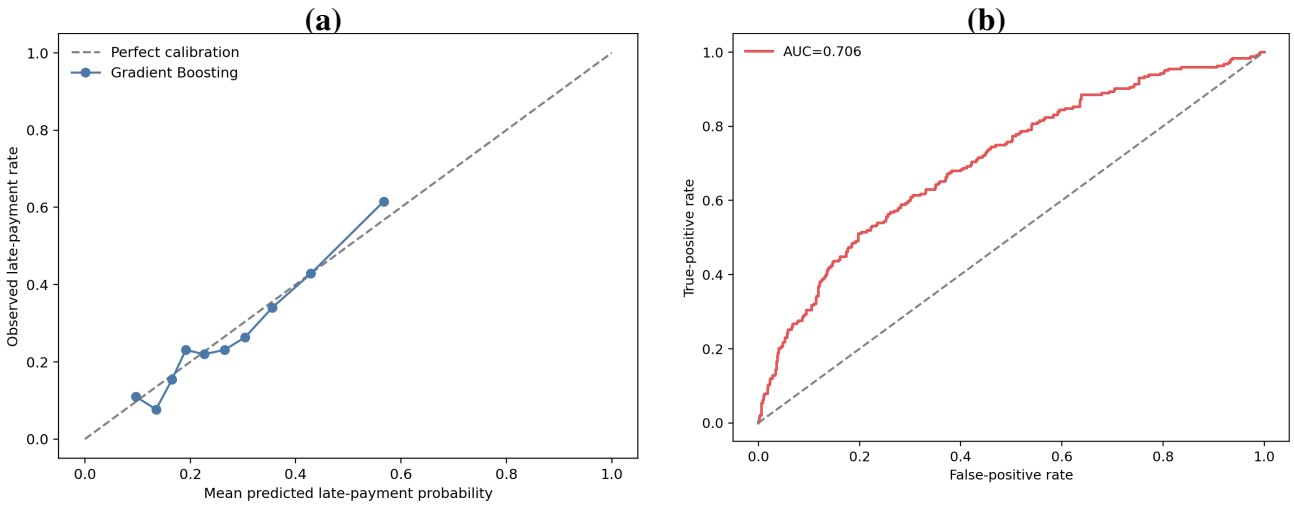


Figure 14. Independent-test probability diagnostics: (a) Calibration curve and (b) ROC curve.

Figures 11–14 connect domain-level explanation, agreement across importance measures, feature-selection stability, calibration, and discrimination. Figure 11(a) aggregates item-level SHAP evidence and shows that financial resilience and credit experience account for the largest cumulative contribution, and Figure 11(b) indicates that the principal items remain visible when two global importance measures are compared. Figure 12 adds direction and heterogeneity: The same response can move predictions differently across respondents because the gradient-boosting model captures interactions. Figure 13 then checks that the explanatory record is not driven by a single fold. Finally, the paired panels in Figure 14 distinguish two properties that should not be conflated: The ROC curve summarizes discrimination across thresholds (AUC = 0.706), whereas the calibration curve compares predicted and observed event frequencies in probability bins. Together these diagnostics support a moderate, internally validated survey-risk signal; they do not establish transportability or a production probability-of-default calibration.

5.3. Local explanation of representative samples based on SHAP

Representative independent-test observations are shown with additive SHAP decompositions. Each plot displays the predicted late-payment probability and the questionnaire responses that move the prediction upward or downward from the model baseline. The case illustrations do not establish causality and are not intended to justify automated credit decisions; their purpose is to make a model output traceable to interpretable survey responses.

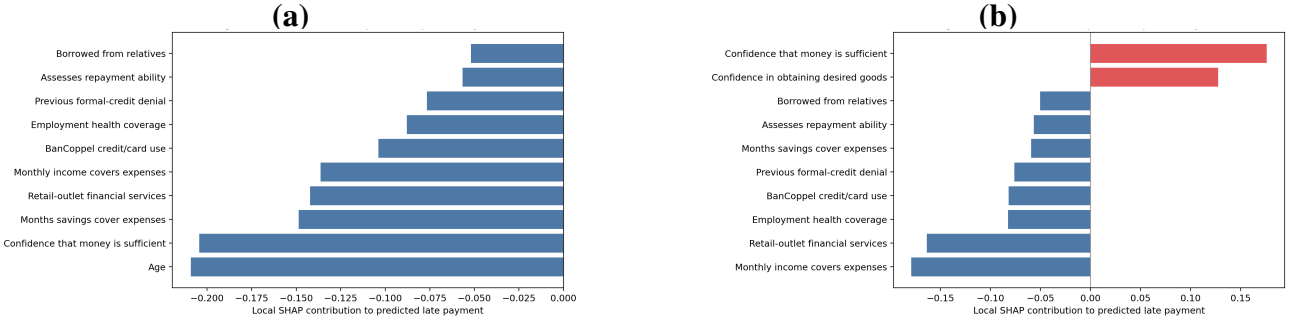


Figure 15. Local SHAP explanations for two representative lower-risk observations: (a) Predicted probability = 0.064 and (b) Predicted probability = 0.164.

Figure 15 places two lower-risk cases side by side so that their shared and distinct protective contributions can be inspected without treating either respondent as a fixed risk type. In panel (a), the displayed contributions are predominantly negative relative to the model baseline, producing a predicted probability of 0.064. Panel (b) remains below the training-selected threshold but has a higher predicted probability (0.164), illustrating that lower predicted risk is a continuum rather than a single profile. The figures make the additive accounting of the fitted model inspectable; they do not imply that the displayed responses caused a favorable repayment outcome.

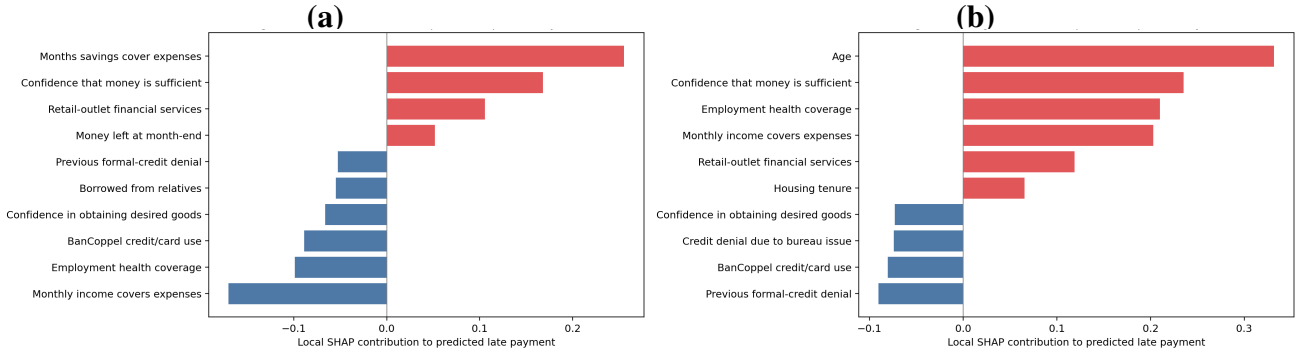


Figure 16. Local SHAP explanations near the operating threshold: (a) Predicted probability = 0.244 and (b) Predicted probability = 0.354.

Figure 16 shows two cases on opposite sides of the training-selected threshold of 0.290. The first case remains below the alert threshold, whereas the second is above it; the contrast is driven by the combined contribution of several questionnaire responses rather than by a single decisive field. This comparison is especially useful for interpreting threshold-dependent classification: Two respondents with broadly similar reported circumstances can receive different alert statuses when their aggregate model contributions fall on different sides of the operating point. The figure does not validate a lending decision or establish false-positive/false-negative status because the purpose of the local display is to explain the model score, not to make an individual-level causal or policy judgment.

6. Conclusion

Using ENIF 2021 survey data, this study examines self-reported late-payment risk on formal credit products and the interpretability of the corresponding prediction model. The analysis begins with an explicitly defined eligible repayment cohort and removes outcomes, outcome-adjacent fields, routing variables, weights, and design identifiers from the predictors. A common fold-internal 25-variable protocol then benchmarks fourteen candidate models without class resampling.

6.1. Research conclusions and academic contributions

Gradient boosting is selected from repeated cross-validation because it has the highest average AUC (0.695) and the lowest average Brier score (0.177) among the leading candidates that entered repeated validation under the controlled rule. On the one-time independent test set, its AUC is 0.706, its KS is 0.312, its Brier score is 0.174, its Accuracy is 0.675, and its Recall is 0.609 at the training-selected threshold. These values characterize a moderate survey-risk signal validated on an independent holdout; they do not represent a bank portfolio's production PD performance.

The explanatory results have substantive and diagnostic roles. Financial resilience, prior credit constraints, informal borrowing, insurance and retirement-savings participation, and formal financial access repeatedly appear in the global rankings. SHAP distributions and local decompositions identify how response patterns move a particular model prediction relative to its baseline. The cross-checks among importance measures and fold-level feature-selection frequency strengthen the evidence that these are not merely artifacts of one visualization.

The selected gradient-boosting specification also offers a pragmatic balance between discrimination, calibration, and explanation. It is not presented as a replacement for a bank scorecard or an administrative PD model. Rather, it is a reproducible survey-risk framework whose outputs can be audited through a fixed feature-governance rule, a fixed comparison rule, and respondent-level SHAP decompositions.

Methodologically, the main contribution lies in keeping several decisions that are often conflated separate and visible. Outcome construction determines what event the model is permitted to learn; predictor governance prevents it from learning that event indirectly through follow-up or contemporaneous repayment fields; cross-validation and fold-internal selection determine the information available during training; and threshold choice determines the operational balance between coverage and false alerts. Treating these as distinct stages improves the credibility of a comparison between algorithms. It also permits subsequent studies to retain the protocol while substituting an institution-specific outcome, a different survey instrument, or a new time period. The framework is therefore intended as a transparent empirical template rather than as a claim that one algorithm is universally optimal for every consumer-credit context.

6.2. Practical implications

These findings indicate that reported late-payment risk is associated with household liquidity, prior access to formal credit, financial participation, and informal support channels. For financial institutions, the practical implication is not to replace administrative credit records with survey answers. A more appropriate use is as a complementary diagnostic framework in customer research, financial-inclusion programs, or outreach studies in which a bank or regulator collects comparable survey information from a defined sample. The model can identify clusters of vulnerability that merit financial counselling, product redesign, or closer human review; it should not independently approve, decline, or price credit.

For model-risk governance, the protocol is transferable even where the questionnaire is not. An institution can define an operational target from its own repayment records, apply the same separation between outcome components and predictors, conduct feature selection within cross-validation folds, fix a common benchmark rule, and retain test data for one final evaluation. Such a workflow makes it easier to distinguish genuine algorithmic improvement from changes caused by feature availability, sample composition, or threshold choice.

6.3. Future research

Future research may integrate unstructured and high-dimensional alternative data, such as consumption behavior, transaction records, platform credit data, social relationships, and location behavior, with traditional survey data. This would help build a more complete risk-factor system, improve identification of long-tail customer groups such as thin-file borrowers, new urban residents, and low-income borrowers and extend research from static cross-sectional modeling to dynamic,

full-cycle risk management. This study is based on one cross-sectional survey and its independent holdout; future work can use data from different periods, regions, or institutions to test the stability of the model relationships and operating threshold. In expanded feature spaces, model-agnostic diagnostics can help distinguish stable predictive signals from importance changes induced by correlated inputs (Fisher et al., 2019). Moreover, existing models may generate differential bias due to factors such as income, education, and region (Szepannek & Lübke, 2021). Future studies can embed algorithmic fairness constraints into the model objective function and construct inclusive finance risk management frameworks that jointly consider predictive accuracy (Misheva et al., 2021), decision fairness, and model explainability. This approach can help mitigate potential credit discrimination and improve the inclusiveness and compliance of financial services (Nwafor et al., 2024). In parallel, future research may examine how large language models can complement established explanation methods such as SHAP and LIME in credit-risk workflows while preserving auditable and human-reviewable outputs (Golec & AlabdulJalil, 2025; Rudin, 2019). Finally, as global regulation of financial AI becomes increasingly stringent, interpretability is likely to become a mandatory compliance requirement. Future research may examine standardized XAI metrics, explanation stability measures, and explanation robustness tests (Guidotti et al., 2018) and further integrate explainable models with regulatory technology. This would provide technical support for compliant deployment, auditing, and regulatory reporting of intelligent risk management models.

Author contributions

Yangting Huang: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – original draft, Writing – review and editing; Yifan Liao: Software, Data curation, Validation, Writing – review and editing; Ying Li: Investigation, Validation, Writing – review and editing; Zihuan Yue: Formal analysis, Visualization, Writing – review and editing; Keyue Yan: Conceptualization, Methodology, Supervision, Writing – review and editing.

Use of AI tools declaration

The authors declare that DeepSeek, an Artificial Intelligence (AI) tool, was used to improve the readability and language of this manuscript. The authors take full responsibility for the content, interpretation, and conclusions of this work.

Acknowledgments

This research was supported by the Guangdong University of Finance Research Start-Up Fund (0000-KC2026014001126) and the 2024 Guangdong Province University Youth Innovation Talent Project (2024WQNCX123).

Conflict of interest

All authors declare no conflicts of interest.

References

Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform Fusion* 58: 82–115. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103.

Bussmann N, Giudici P, Marinelli D, et al. (2020) Explainable AI in fintech risk management. *Front Artif Intell* 3: 26. https://doi.org/10.3389/frai.2020.00026

Bussmann N, Giudici P, Marinelli D, et al. (2021) Explainable machine learning in credit risk management. *Comput Econ* 57: 203–216. https://doi.org/10.1007/s10614-020-10042-0

Chawla NV, Bowyer KW, Hall LO, et al. (2002) SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* 16: 321–357. https://doi.org/10.1613/jair.953

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608. https://arxiv.org/abs/1702.08608

Fernández-Delgado M, Cernadas E, Barro S, et al. (2014) Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res* 15: 3133–3181. Available from: https://jmlr.org/papers/v15/delgado14a.html

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *arXiv preprint* arXiv:1801.01489. https://arxiv.org/abs/1801.01489

Golec M, AlabdulJalil M (2025) Interpretable LLMs for credit risk: A systematic review and taxonomy. *arXiv preprint* arXiv:2506.04290. https://arxiv.org/abs/2506.04290

Gramegna A, Giudici P (2021) SHAP and LIME: An evaluation of discriminative power in credit risk. *Front Artif Intell* 4: 752558. https://doi.org/10.3389/frai.2021.752558

Guidotti R, Monreale A, Ruggieri S, et al. (2018) A survey of methods for explaining black box models. *arXiv preprint* arXiv:1802.01933. https://arxiv.org/abs/1802.01933

Han S, Jung H (2024) NATE: Non-pArameTric approach for Explainable credit scoring on imbalanced class. *PLOS ONE* 19: e0316454. https://doi.org/10.1371/journal.pone.0316454

Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. *arXiv preprint* arXiv:1609.02907. https://arxiv.org/abs/1609.02907

Lipton ZC (2016) The mythos of model interpretability. *arXiv preprint* arXiv:1606.03490. https://arxiv.org/abs/1606.03490

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv Neural Inform Process Syst* 30: 4765–4774.

Mapfumo I, Shongwe T (2026) Performance evaluation of machine learning and deep learning models for credit risk prediction. *J Risk Financ Manage* 19: 210. https://doi.org/10.3390/jrfm19030210

Misheva BH, Osterrieder J, Hirsa A, et al. (2021) Explainable AI in credit risk management. *arXiv preprint* arXiv:2103.00949. https://arxiv.org/abs/2103.00949

Monje L, Carrasco RA, Sánchez-Montañés M (2026) Machine learning XAI for early loan default prediction. *Comput Econ* 67: 4033–4062. https://doi.org/10.1007/s10614-025-10962-9

Nwafor CN, Nwafor O, Brahma S (2024) Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach. *Sci Rep* 14: 25174. https://doi.org/10.1038/s41598-024-75026-8

Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12: 2825–2830.

Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?: Explaining the predictions of any classifier. *arXiv preprint* arXiv:1602.04938. https://arxiv.org/abs/1602.04938

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *arXiv preprint* arXiv:1811.10154. https://arxiv.org/abs/1811.10154

Saarela M, Jauhiainen S (2021) Comparison of feature importance measures as explanations for classification models. *Sn Appl Sci* 3: 272. https://doi.org/10.1007/s42452-021-04148-9

Schmitt M (2024) Explainable automated machine learning for credit decisions: Enhancing human-artificial intelligence collaboration in financial engineering. *arXiv preprint* arXiv:2402.03806. https://doi.org/10.48550/arXiv.2402.03806

Shreya, Pathak H (2025) Explainable artificial intelligence credit risk assessment using machine learning. *arXiv preprint* arXiv:2506.19383. https://doi.org/10.48550/arXiv.2506.19383

Szepannek G, Lübke K (2021) Facing the challenges of developing fair risk scoring models. *Front Artif Intell* 4: 681915. https://doi.org/10.3389/frai.2021.681915

Wang H, Huang J, Sua LS, et al. (2025) Enhancing credit risk prediction: A multi-stage ensemble pipeline. *arXiv preprint* arXiv:2509.22381. https://doi.org/10.48550/arXiv.2509.22381



AIMS Press

© 2026 Huang et al., licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)