



A Validation-Weighted Explainable Ensemble for Consumer Credit Default Prediction: Balancing Predictive Performance and Regulatory Interpretability

Kelin Qian¹, Zhanpeng Wang¹ and Chunjiang Li^{2*}

¹School of Finance and Economics, Xizang University, Lhasa 850000, China

²School of Economics and Management, Hunan University of Technology, Zhuzhou 412007, China

* Corresponding author: li13650320074@163.com

Abstract. The growth of digital and online consumer lending has intensified the need for accurate yet transparent default-prediction models. Black-box machine-learning models often achieve strong discrimination but are difficult to deploy under financial-regulatory regimes that demand explainability. This paper proposes an explainable ensemble-learning framework that couples five heterogeneous base learners—logistic regression, random forest, extremely randomised trees, gradient boosting and histogram-based gradient boosting—through a convex meta-combination whose weights are selected to maximise the area under the ROC curve on a held-out validation set, and integrates a model-agnostic Shapley-value (SHAP) explanation layer. Synthetic minority over-sampling (SMOTE) mitigates class imbalance, and the conclusions are shown to be robust to this choice. On the public ‘default of credit card clients’ benchmark (30,000 records, 22.1% default rate), the proposed ensemble attains the highest AUC (0.776) and Kolmogorov–Smirnov statistic (0.419); the margin over the strongest single learner is small but statistically significant (DeLong $p < 0.05$) and stable across five repeated runs (mean AUC 0.781 ± 0.007), and the validation-based weighting effectively reduces to a two-learner combination of the bagging models. The SHAP analysis identifies the recent repayment-status history and the credit limit as the dominant default drivers and yields per-applicant explanations for adverse-action reporting, and a preliminary subgroup analysis shows only a small fairness gap across applicant sex. The results indicate that a carefully designed ensemble can improve ranking performance without sacrificing the transparency required for compliant credit decision-making.

Keywords: Consumer Credit Risk, Credit Default Prediction, Ensemble Learning, Explainable AI, SHAP, FinTech Governance

1 Introduction

The digitalisation of financial services has reshaped how credit is originated, priced and monitored. Internet-finance platforms—including online consumer lending, peer-to-

peer marketplaces and embedded buy-now-pay-later products—now extend credit to millions of borrowers who may lack a conventional credit history^[1]. While this broadening of access supports financial inclusion, it simultaneously exposes lenders to elevated default risk and the financial system to new channels of contagion. Reliable estimation of a borrower’s default probability is therefore a central problem of modern financial risk management.

Traditional credit scoring relied on linear models such as logistic regression, valued for their interpretability. Internet-finance data, however, are high-dimensional, noisy and rich in non-linear interactions among repayment behaviour, utilisation and demographics; machine-learning methods, and tree-based ensembles in particular, capture this structure and improve discrimination over linear baselines^[2]. The gain comes at a cost: the resulting models are opaque, and a credit decision that cannot be explained is hard to reconcile with consumer-protection law and prudential supervision. This tension between accuracy and transparency is the defining methodological challenge for credit modelling in the digital economy, and supervisory guidance on artificial intelligence in finance increasingly emphasises interpretability, robustness and fairness. Post-hoc feature-attribution methods grounded in cooperative game theory^[3] reconcile the two objectives: a high-capacity model performs prediction while a principled attribution layer renders each decision auditable. Building on this idea, we develop an explainable ensemble-learning framework for credit-default prediction that combines heterogeneous base learners through a validation-weighted meta-combination, addresses class imbalance through synthetic over-sampling, and equips the predictor with a Shapley-value explanation layer yielding both global rankings and per-applicant explanations.

Against this backdrop, the contributions of this work are threefold. **(1) An explainable ensemble framework for consumer-credit risk.** We design an end-to-end framework unifying accurate risk prediction, decision-level interpretability and regulatory adaptation in a single pipeline; internet-finance lending is the motivating application, while the empirical study uses a public credit benchmark as a representative proxy. **(2) A validation-AUC weighted ensemble.** The learners are combined with weights chosen to maximise validation-set AUC; repeated experiments and a DeLong test show the combination improves on the strongest single learner by a small but statistically significant margin, and an ablation shows the weighting prunes the ensemble to its two strongest, decorrelated members. **(3) Coupling SHAP explanations with financial-regulatory concepts.** We map the model’s global and local Shapley attributions onto responsible-finance practices—model risk management, adverse-action reporting and fairness auditing—as a conceptual bridge the present experiments illustrate.

2 Related Work

Credit scoring is a long-standing application of statistical classification. Benchmarking studies establish that ensemble methods—gradient-boosted trees and random forests^[4,5]—are typically the strongest individual performers on tabular credit data^[2,6], while logistic regression remains a competitive interpretable baseline. Because default events

are rare, credit datasets are intrinsically imbalanced; the synthetic minority over-sampling technique (SMOTE) ^[7] and cost-sensitive or class-weighting remedies improve recall and class-balanced metrics such as the G-mean.

As models have grown more complex, explainability has become a priority. Model-agnostic attribution methods ^[3] estimate each feature's contribution to a prediction; Shapley-value attributions, which uniquely satisfy a set of game-theoretic axioms, have become the de facto standard for explaining credit-risk models ^[8,9], while a parallel strand argues for inherently interpretable models in high-stakes decisions ^[10]. We adopt the now-common position that a high-capacity predictor paired with a faithful attribution layer delivers both strong performance and decision-level transparency.

3 Proposed Framework

3.1 Problem Formulation

Let $D = \{(x_i, y_i)\}$, $i = 1, \dots, N$, denote a credit dataset in which $x_i \in \mathbb{R}^d$ is the feature vector of applicant i and $y_i \in \{0, 1\}$ indicates default (1) or non-default (0). The goal is a scoring function $f: \mathbb{R}^d \rightarrow [0, 1]$ estimating the default probability while admitting a faithful per-prediction explanation. The framework comprises preprocessing, imbalance correction, a heterogeneous base-learner layer, a weighted meta-combination and a Shapley-value explanation layer.

3.2 Data Preprocessing and Feature Engineering

Raw records are cleaned by resolving out-of-range categorical codes and consolidating sparse education and marital-status categories; the monthly repayment-status indicators are treated as ordinal risk signals. Features are standardised using training-partition statistics only, avoiding leakage into the validation or test partitions.

3.3 Class-Imbalance Handling

Because defaulters constitute a small minority, the synthetic minority over-sampling technique (SMOTE) ^[7] is applied to the training partition. For each minority instance x , one of its k nearest minority neighbours $\hat{x}$ is selected and a synthetic example is generated as

$$x_{\text{new}} = x + \delta \cdot (\hat{x} - x), \quad \delta \sim U(0, 1) \quad (1)$$

with $k = 5$. Over-sampling is restricted to the training data, so the balanced training set improves minority detection without distorting the held-out assessment.

Because several predictors are categorical or ordinal codes that ordinary SMOTE may interpolate into non-meaningful values, we treat it as only one of several imbalance remedies; Section 5.2 shows that ranking performance is essentially unchanged under class weighting and no resampling, so the conclusions do not hinge on SMOTE's treatment of the discrete features.

3.4 Heterogeneous Base Learners

Ensemble diversity drives generalisation, so the base layer mixes five learners with different inductive biases: logistic regression (a smooth linear surface); a random forest^[4] (bagged de-correlated deep trees); extremely randomised trees^[11], which randomise split thresholds for a complementary low-variance learner; gradient boosting (shallow trees fitted sequentially to residuals); and a fast, regularised histogram-based gradient-boosting variant representative of the gradient-boosting family underpinning the most widely used credit-scoring models, of the kind popularised by LightGBM^[5], so the comparison covers this family without depending on any single third-party implementation.

3.5 Weighted Ensemble Integration

Each base learner is trained on the rebalanced training set and produces a default probability $p_m(x)$. The ensemble forms a convex combination of these probabilities,

$$f(x) = \sum_m w_m \cdot p_m(x), \quad w_m \geq 0, \quad \sum_m w_m = 1 \quad (2)$$

The non-negative weights, summing to one, are chosen by grid search over the probability simplex to maximise validation-set AUC. Selecting them on a validation set drawn from the real class distribution, rather than on the over-sampled training set, prevents over-fitting the resampled distribution. The procedure concentrates the weight on the two bagging learners (random forest and extra trees), assigning negligible weight to the others—an informative finding about which learners carry complementary signal.

3.6 Shapley-Value Explanation Layer

To explain predictions, the framework computes Shapley values, the unique attribution scheme satisfying local accuracy, missingness and consistency. The contribution of feature j to a prediction is its average marginal contribution over all feature orderings,

$$\varphi_j = \sum |S|!(d-|S|-1)!/d! \cdot [f(S \cup \{j\}) - f(S)] \quad (3)$$

where the sum runs over feature subsets S . As the exact computation is exponential, the values are estimated by Monte-Carlo permutation sampling^[3] against a background set—the model-agnostic estimator underlying modern SHAP^[3]. Averaging $|\varphi_j|$ over applicants gives a global ranking; the signed values give an additive local explanation.

4 Experimental Setup

4.1 Dataset

Experiments use the public ‘default of credit card clients’ dataset^[12], a widely adopted consumer-credit benchmark of 30,000 records. Each record contains 23 predictors: the credit limit, demographics (sex, education, marital status, age), and six months each of

repayment-status, bill-statement and previous-payment amounts. The binary target indicates default in the next billing cycle. The overall default rate is 22.1%. The data are partitioned into 70% training and 30% test sets by stratified sampling, with one fifth of the training partition held out as a validation set for ensemble-weight selection (final split $\approx$ 56% training, 14% validation, 30% test). SMOTE is applied only to the training partition. We use this benchmark as a reproducible proxy for consumer-credit default; validation on real internet-finance platform data is left to future work.

4.2 Evaluation Metrics

As accuracy is misleading under imbalance, we report a panel of metrics: the area under the ROC curve (AUC); precision, recall and their harmonic mean (F1) for minority detection; the geometric mean of sensitivity and specificity (G-mean) for class balance; and the Kolmogorov–Smirnov statistic $KS = \max_i (TPR(t) - FPR(t))$, a standard credit-scoring measure of score separability.

4.3 Implementation

All models are implemented in Python with scikit-learn and tuned on the training data (random forest: 300 trees, depth 14; extra trees: 400 trees, depth 16; gradient boosting: 250 stages, depth 3, learning rate 0.05; histogram gradient boosting: 400 iterations, depth 6; logistic regression with L2 penalty). Ensemble weights, selected by grid search to maximise validation-set AUC, are 0.40 for random forest and 0.60 for extra trees. Shapley values are estimated by Monte-Carlo permutation sampling (60 permutations per applicant, background of 80 instances). A fixed random seed ensures reproducibility, and the complete pipeline is released as a single script.

5 Results and Discussion

5.1 Predictive Performance

Table 1 reports test-set performance. The proposed ensemble attains the highest AUC (0.776) and KS statistic (0.419) of all evaluated models, edging extra trees (0.774) and random forest (0.771) and clearly beating the linear baseline (0.718), while essentially matching extra trees on the class-balanced metrics (F1 0.536, G-mean 0.689). The single learners show a clear accuracy–recall trade-off—histogram gradient boosting has the highest accuracy (0.801) but lowest recall (0.435), logistic regression the highest recall (0.644) at the cost of precision—whereas the ensemble balances the two. We stress that the margin over the best single learner is small (about 0.003 AUC); its value lies in being statistically significant and stable rather than large.

Table 1. Test-set performance of the five base learners and the proposed weighted ensemble (last row, in bold). The ensemble attains the highest AUC and KS statistic.

Model	AUC	Acc.	Prec.	Rec.	F1	Gmean	KS
Logistic Regression	0.7184	0.6641	0.3566	0.6444	0.4591	0.6569	0.3655
Random Forest	0.7705	0.7840	0.5109	0.5550	0.5320	0.6865	0.4173
Extra Trees	0.7735	0.7854	0.5138	0.5605	0.5362	0.6900	0.4161
Gradient Boosting	0.7651	0.7963	0.5439	0.4917	0.5165	0.6589	0.4032
Hist. Gradient Boosting	0.7564	0.8012	0.5660	0.4350	0.4919	0.6275	0.3880
Weighted Ensemble (ours)	0.7761	0.7869	0.5170	0.5560	0.5358	0.6885	0.4187

To confirm that these gains are not an artefact of a single split, the entire pipeline was repeated over five random seeds. The ensemble attained the highest mean AUC, 0.781 ± 0.007 , ahead of extra trees (0.779 ± 0.007) and random forest (0.776 ± 0.008). The difference between the ensemble and the best single learner was assessed with the DeLong test for two correlated ROC curves: the ensemble significantly outperformed extra trees ($p = 0.047$) and random forest ($p < 0.001$). The improvement over the strongest single model is therefore modest in magnitude but statistically significant and stable across repetitions, and is obtained without any test-set tuning because the combination weights are fixed on the validation set.

5.2 Imbalance Handling and Robustness

At the 0.5 threshold the ensemble identifies 1,107 of 1,991 true defaulters (recall 0.556) while limiting false positives to 1,034 of 7,009 non-defaulters. The training-set over-sampling drives this substantial recall relative to the accuracy-oriented boosting learners. In practice the threshold would be tuned to the institution’s relative cost of false acceptances and rejections, and the favourable KS statistic (0.419) indicates ample separability for cost-sensitive selection.

To check that the results do not depend on how class imbalance is handled, we compared the ensemble under SMOTE, class weighting and no resampling. Ranking is robust: the test AUC is 0.776 with SMOTE and 0.780 with both alternatives. SMOTE and class weighting both restore recall (0.55 and 0.59), whereas without rebalancing recall collapses to 0.35 at an essentially unchanged AUC. Class weighting, which synthesises no discrete values, matches or slightly exceeds SMOTE, so the conclusions do not rest on SMOTE’s interpolation of the categorical or ordinal variables.

5.3 Ensemble Ablation

Because the validation-weighted ensemble assigns zero weight to three of the five learners, we examine what it contributes by comparing combination strategies. Naive equal averaging of all five learners (AUC 0.771, F1 0.530) is in fact worse than the best

single learner (extra trees, 0.774 / 0.536), because the weaker members drag the average down. The proposed ensemble (0.776 / 0.536) is numerically identical to a two-learner random-forest / extra-trees combination, since the validation procedure places all weight on those two bagging models; it is therefore best read as an automatic, data-driven selection of the two strongest decorrelated learners rather than a many-model fusion. Relative to the common explainable-credit baseline of a single gradient-boosting model paired with SHAP, our framework retains that interpretability while improving discrimination by a small but statistically significant margin (DeLong $p = 0.047$).

5.4 Global Interpretability

Fig. 1 ranks features by mean absolute Shapley value. The most recent repayment status (PAY_0) dominates—over twice the impact of any other feature—followed by the earlier repayment indicators (PAY_2–PAY_6) in decreasing order of recency and then the credit limit, with bill/payment amounts and demographics far less important. This accords with the credit-scoring literature, in which recent delinquency and exposure are the strongest predictors [6], confirming that the ensemble has learned economically sensible structure.

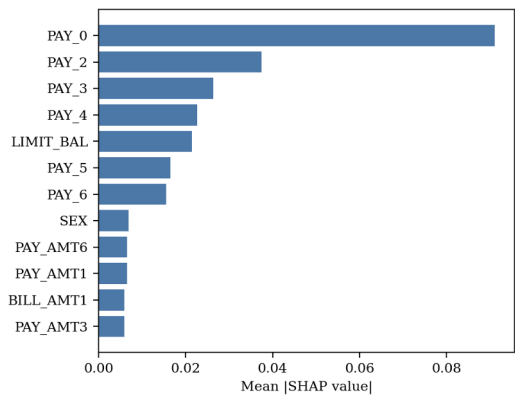


Fig. 1. Global feature importance measured by mean absolute SHAP value. The most recent repayment status (PAY_0) is the dominant driver of predicted default probability.

5.5 Local Interpretability: A Case Study

Beyond global rankings, the framework explains individual decisions. For a high-risk applicant whose predicted default probability (0.94) far exceeds the background reference (0.51), the Shapley decomposition attributes the gap almost entirely to the repayment-status history—the delinquent most-recent status (PAY_0, about +0.14) reinforced by earlier cycles. As the contributions sum exactly to this gap, the explanation directly supplies the principal reasons for a decline required by adverse-action notices and lets an auditor confirm the decision rests on legitimate factors.

5.6 Regulatory Implications

These implications are conceptual and illustrative rather than validated by dedicated deployment experiments. The framework maps onto four pillars of responsible-finance governance. For model risk management, the global SHAP ranking makes the model's logic inspectable and lets validators monitor for drift or reliance on prohibited proxies. For AI governance, every score carries an auditable attribution, keeping a human reviewer able to interrogate or override automated decisions. For responsible lending, local explanations let an institution verify case by case that a decision rests on legitimate repayment-related factors rather than sensitive attributes. And for explainable credit decisions^[13], the additive Shapley contributions give the principal reasons for a decline in a directly reportable, contestable form.

5.7 Fairness Discussion

As a preliminary, single-attribute check illustrating how the explanation layer supports fairness analysis, we examine the framework across the sex attribute. Disaggregating the test set, the ensemble achieves an AUC of 0.765 for men ($n = 3,563$, default rate 23.9%) and 0.783 for women ($n = 5,437$, default rate 21.0%), an AUC gap of 0.018. At the 0.5 threshold, the false-positive rate is 0.171 for men versus 0.133 for women (gap 0.038) and the overall error rate is 0.234 versus 0.200 (gap 0.034); the differences are modest and partly track the higher observed default rate among men. This evidence suggests the sex gap is not pronounced, but a more rigorous audit—covering additional protected attributes, calibration and group-fairness criteria such as equal opportunity—is needed before deployment, for which the per-group transparency of the explanation layer provides a natural foundation.

5.8 Limitations

Several limitations qualify the findings. The experiments use a single public benchmark, so validation on multiple real-world internet-lending portfolios would strengthen external validity. The Monte-Carlo Shapley estimator assumes feature independence, which can bias attributions under strong feature correlation. The evaluation focuses on discrimination, class balance and group-wise separation, leaving probability calibration and a comprehensive fairness audit for future work. The methodological novelty is incremental: the contribution lies in the integrated, validation-weighted and explainable design rather than a new algorithm.

6 Conclusion

This paper proposed a validation-weighted explainable ensemble for consumer-credit default prediction: five heterogeneous learners combined with weights chosen on a held-out validation set, SMOTE for class imbalance, and a SHAP explanation layer.

On the public benchmark it attained the highest AUC and Kolmogorov–Smirnov statistic among all models while matching the strongest single learner on class-balanced metrics. We report this advantage cautiously: it is small (about 0.003 AUC) and, as an ablation shows, the weighting effectively reduces the ensemble to a random-forest / extra-trees combination, so the contribution is an integrated, interpretable design rather than a large accuracy gain. The SHAP explanations were economically sensible and usable for adverse-action reporting and model validation, illustrating how predictive performance and regulatory interpretability can be reconciled. Future work will validate on real internet-finance platform data, add correlation-aware attribution and cost-sensitive thresholds, and conduct a comprehensive fairness audit.

Acknowledgments.

This study was not supported by any external funding. The authors thank the anonymous reviewers for their constructive comments.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used generative AI tools for language editing and manuscript organization, and carefully reviewed and verified all AI-assisted content, taking full responsibility for the final manuscript. No AI tool was used for data collection, data analysis, result generation, or scientific decision-making.

References

1. Tian, X., Tian, Z., Khatib, S.F.A., Wang, Y.: Machine learning in internet financial risk management: a systematic literature review. *PLOS ONE* 19(4), e0300195 (2024). <https://doi.org/10.1371/journal.pone.0300195>.
2. Mestiri, S.: Credit scoring using machine learning and deep learning-based models. *Data Science in Finance and Economics* 4(2), 236-248 (2024). <https://doi.org/10.3934/DSFE.2024009>.
3. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* 30, pp. 4765-4774 (2017).
4. Breiman, L.: Random forests. *Machine Learning* 45(1), 5-32 (2001). <https://doi.org/10.1023/A:1010933404324>.
5. Ke, G., Meng, Q., Finley, T., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems* 30, pp. 3146-3154 (2017).

6. Chen, Y., Calabrese, R., Martin-Barragan, B.: Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research* 312(1), 357-372 (2024). <https://doi.org/10.1016/j.ejor.2023.06.036>.
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321-357 (2002). <https://doi.org/10.1613/jair.953>.
8. Babaei, G., Giudici, P., Raffinetti, E.: Explainable FinTech lending. *Journal of Economics and Business* 125-126, 106126 (2023). <https://doi.org/10.1016/j.jeconbus.2023.106126>.
9. Nallakaruppan, M.K., Chaturvedi, H., Grover, V., et al.: Credit risk assessment and financial decision support using explainable artificial intelligence. *Risks* 12(10), 164 (2024). <https://doi.org/10.3390/risks12100164>.
10. Tu, J., Wu, Z.: Inherently interpretable machine learning for credit scoring: optimal classification tree with hyperplane splits. *European Journal of Operational Research* 322(2), 647-664 (2025). <https://doi.org/10.1016/j.ejor.2024.10.046>.
11. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. *Machine Learning* 63(1), 3-42 (2006). <https://doi.org/10.1007/s10994-006-6226-1>.
12. Yeh, I.C., Lien, C.H.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36(2), 2473-2480 (2009). <https://doi.org/10.1016/j.eswa.2007.12.020>.
13. Weber, P., Carl, K.V., Hinz, O.: Applications of explainable artificial intelligence in finance: a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly* 74(2), 867-907 (2024). <https://doi.org/10.1007/s11301-023-00320-0>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

