Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

A Model-Risk-Friendly Probability of Default Workflow: Calibration, Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset

Jiaying Jin[1], *Tina Huang*[2], *Sam Lu*[3]
[1]*Applied Analytics, Columbia University, NY, USA*
[2]*Computer Engineering, Columbia University, NY, USA*
[3]*Computer Science, Columbia University, NY, USA*
Jiaying@columbia.edu

DOI: 10.69987/JACS.2024.40606

Keywords

Probability of default; credit risk; model risk management; calibration; conformal prediction; bootstrap ensembles; SHAP.

Abstract

Probability of default (PD) models are central to retail credit risk, but governance frameworks require more than a single score. This paper develops an auditable PD workflow that combines predictive performance, probability calibration, uncertainty quantification, and explanation. Using the UCI Default of Credit Card Clients dataset (30,000 observations, 23 explanatory variables), we evaluate logistic regression, random forest, XGBoost, and LightGBM under a strict 60/10/10/20 train/probability-calibration/conformal-calibration/test split. Ranking performance is assessed with AUC, PR-AUC, and KS; probability quality with Brier score, log loss, and expected calibration error (ECE). Raw XGBoost gives the strongest overall balance of discrimination and proper-score probability quality on the held-out test set (AUC=0.7796, PR-AUC=0.5526, Brier=0.1351, log loss=0.4301), while sigmoid-calibrated logistic regression achieves the lowest ECE (0.0068). Split conformal prediction for the final model achieves empirical coverage 0.9017 and 0.9462 at 90% and 95% targets, with average set sizes 1.222 and 1.469. Bootstrap sensitivity bands have average widths 0.0852 (90%) and 0.0934 (95%), and the decile-level observed default rate falls within the mean interval bounds in every decile. SHAP analysis identifies recent repayment status and credit limit as the dominant global drivers. The analysis adopts a strict validation logic in which proper scoring rules are primary for probability selection, while ECE and reliability plots are diagnostic complements. Overall, the results show that a PD workflow can be made calibration-aware, uncertainty-aware, and governance-ready without departing from standard supervised-learning tools.

1. Introduction

Probability of default (PD) estimation is a core input to retail credit risk management. PD enters expected loss, pricing, portfolio monitoring, provisioning, and, in many settings, regulatory capital calculations [1]. Because a PD is used as a probability and not merely as a rank score, model evaluation must go beyond classification accuracy.

Model risk management frameworks make this explicit. SR 11-7 frames model risk as the risk of adverse consequences from incorrect or misused model outputs and emphasizes robust development, effective validation, and sound governance [2]. For PD models, that means checking not only whether risky borrowers are ranked correctly, but also whether the reported PDs are numerically reliable, whether uncertainty is characterized, and whether the model can be explained to validators and business users.

Discrimination and calibration answer different questions. Discrimination asks whether the model separates defaulters from non-defaulters; AUC is a standard threshold-free summary and KS remains common in credit reporting [14]. Calibration asks whether a predicted PD of, say, 0.20 corresponds to an

empirical default rate near 0.20. Proper scoring rules such as Brier score and log loss are important because PD errors propagate directly into expected loss and decision policies [12],[13]. ECE is useful as a reliability-gap diagnostic, but it should supplement rather than replace proper scoring rules [12].

Uncertainty quantification is a second gap between typical benchmark studies and governance needs. A single PD point estimate suppresses uncertainty arising from limited sample size, model instability, or boundary cases. Two complementary tools are attractive in practice. Bootstrap ensembles quantify how sensitive PD estimates are to data perturbation [18],[19], while conformal prediction produces classification sets with finite-sample marginal coverage under exchangeability [15]-[17]. In credit operations, singleton conformal sets support automated action and ambiguous {0,1} sets flag cases for manual review.

Interpretability is equally important. Credit risk teams need to understand which variables drive the model, whether the drivers are economically plausible, and how a particular applicant's score can be decomposed. SHAP addresses this by attributing model outputs to input features in a way that supports both local and global interpretation [20],[21]. For tree ensembles, TreeSHAP is computationally efficient and well suited to production-scale tabular models.

This paper develops an end-to-end PD workflow on the UCI Default of Credit Card Clients dataset [3],[4]. The contribution is not a new learning algorithm. Instead, the contribution is a stricter, model-risk-oriented evaluation design: disjoint training, calibration, and uncertainty splits; explicit reporting of both ranking and probability quality; calibrated uncertainty artifacts; and governance-focused explanation.

2. Method

2.1 Workflow overview

The workflow has four stages. First, a base model is fit on the training split. Second, post-hoc calibration is fit on a disjoint probability-calibration split. Third, uncertainty artifacts are computed using the selected final model, namely split conformal prediction sets and bootstrap sensitivity bands. Fourth, TreeSHAP explanations are produced for the final tree ensemble. Figure 1 summarizes the pipeline.

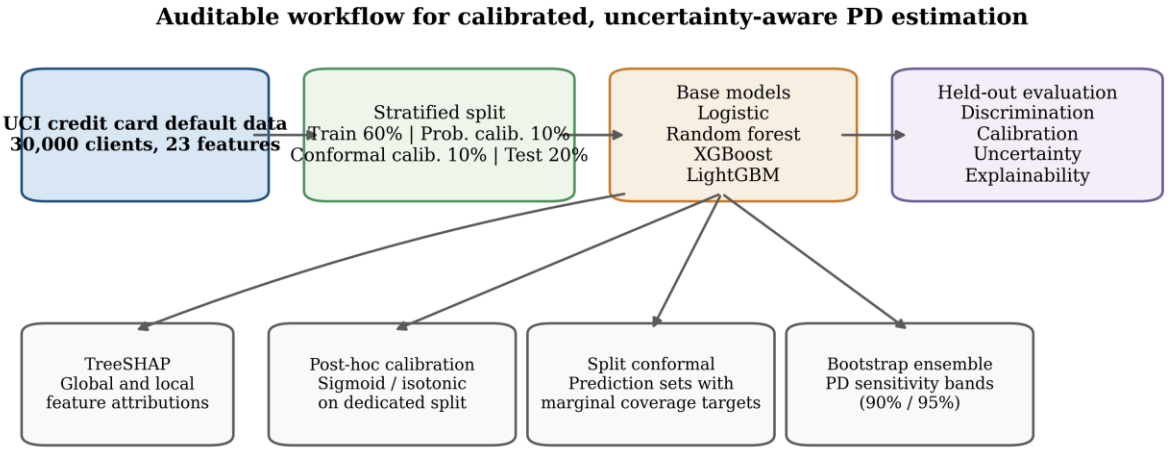

Figure 1. End-to-end workflow for calibrated, uncertainty-aware PD estimation.

Table 1. Dataset summary.

Dataset	Samples	Features	Missing values	Default rate
UCI Default of Credit Card Clients	30000	23	0	0.2212

Table 2. Stratified split statistics.

Split	N	Default rate
Train	18000	0.2212
Prob. calibration	3000	0.2213
Conformal calibration	3000	0.2210
Test	6000	0.2212

Table 3. Fixed model hyperparameters.

Model family	Key hyperparameters
Logistic regression	C=1.0, penalty=L2, solver=saga, max_iter=4000
Random forest	n_estimators=200, min_samples_leaf=50
XGBoost	n_estimators=200, max_depth=4, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8, tree_method=hist
LightGBM	n_estimators=300, num_leaves=31, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8

2.2 Data, split protocol, and model families

The dataset contains 30,000 credit-card clients and 23 explanatory variables after excluding the record identifier [3],[4]. Inputs include credit limit, age, sex, education, marital status, six months of repayment status, six months of bill amounts, and six months of payment amounts. The target is default payment in the following month.

We use a fixed stratified 60/10/10/20 split: 18,000 training cases, 3,000 probability-calibration cases, 3,000 conformal-calibration cases, and 6,000 test cases. Stratification preserves the portfolio default rate across splits and prevents avoidable prevalence-shift artifacts. The disjoint calibration splits are deliberate: each post-estimation step is evaluated on data not used for fitting the step itself, which is consistent with model-risk control objectives [2].

We compare four standard model families: L2-regularized logistic regression [5], random forest [6], XGBoost [7], and LightGBM [8]. The comparison is intentionally conservative. The goal is not to exhaustively tune every algorithm, but to contrast a transparent baseline with well-established nonlinear alternatives under a reproducible protocol.

Preprocessing is handled inside model-specific pipelines fit on the training split only [9]. Transformations are applied conservatively: scaling is used where needed for numerical stability and coefficient interpretation, while tree-based models operate on the original feature scale. The emphasis is leakage control and reproducible transformation, not a blanket claim that one universal encoding strategy is optimal for every model class.

2.3 Calibration, uncertainty quantification, and explanation

Probabilities are examined in raw form and after two monotone post-hoc maps: Platt sigmoid calibration [10] and isotonic regression [11]. Both are fit on the dedicated 3,000-case probability-calibration split. Sigmoid calibration is parsimonious and preserves ranking; isotonic regression is more flexible but can overfit local structure and may alter ranking metrics because it introduces ties [10]-[12].

We report AUC, PR-AUC, and KS for discrimination; Brier score and log loss for probability quality; and ECE with 10 bins as a diagnostic summary of reliability [12]-[14]. Proper scoring rules are treated as primary for selecting a deployable probability model, while ECE and reliability graphics are used as complements. Accuracy and F1 at a 0.5 threshold are informative but secondary because a fixed 0.5 cutoff is not a calibrated business policy in an imbalanced credit portfolio.

For uncertainty quantification we take raw XGBoost as the final model family because it has the strongest held-out discrimination and the best overall proper-score performance in Table 4. Split conformal prediction uses the nonconformity score $s(x,y)=1-p_y(x)$ on the disjoint conformal-calibration split. For a target miscoverage alpha, the prediction set for a new case includes class y

whenever the score is below the conformal quantile threshold [15]-[17].

We also fit 10 bootstrap replicas of the final model family and compute empirical 90% and 95% PD bands from the ensemble distribution. These are reported as resampling-based sensitivity bands rather than formal pointwise confidence intervals. With 10 replicas, the tail quantiles are coarse, but the resulting intervals are still informative for segment-level reasonableness checks.

Explainability is assessed with TreeSHAP [20],[21]. We use aggregated mean absolute SHAP values to rank global drivers and retain the model-specific output decomposition for local case review. All tables and figures in the paper are regenerated directly from the reported empirical outputs so that the manuscript is internally consistent and audit-ready.

3. Results and discussion

3.1 Base-model comparison

Table 4 and Figure 2 show that raw XGBoost dominates the held-out comparison on the main threshold-free metrics: AUC=0.7796, PR-AUC=0.5526, KS=0.4352, Brier=0.1351, and log loss=0.4301. LightGBM and random forest are close on discrimination, while logistic regression remains competitive given its simpler functional form.

These gaps are substantively plausible for the dataset. Recent repayment-status variables and recent payment behavior create nonlinear interactions that tree ensembles can exploit more aggressively than the linear baseline. At the same time, the performance differences are not extreme, which is consistent with the fact that the UCI benchmark is a structured tabular problem rather than a setting where linear models are uniformly inadequate.

Because threshold metrics depend on business cutoffs, the analysis focuses on ranking and proper scores rather than on very small differences in classification accuracy at a fixed 0.5 threshold.

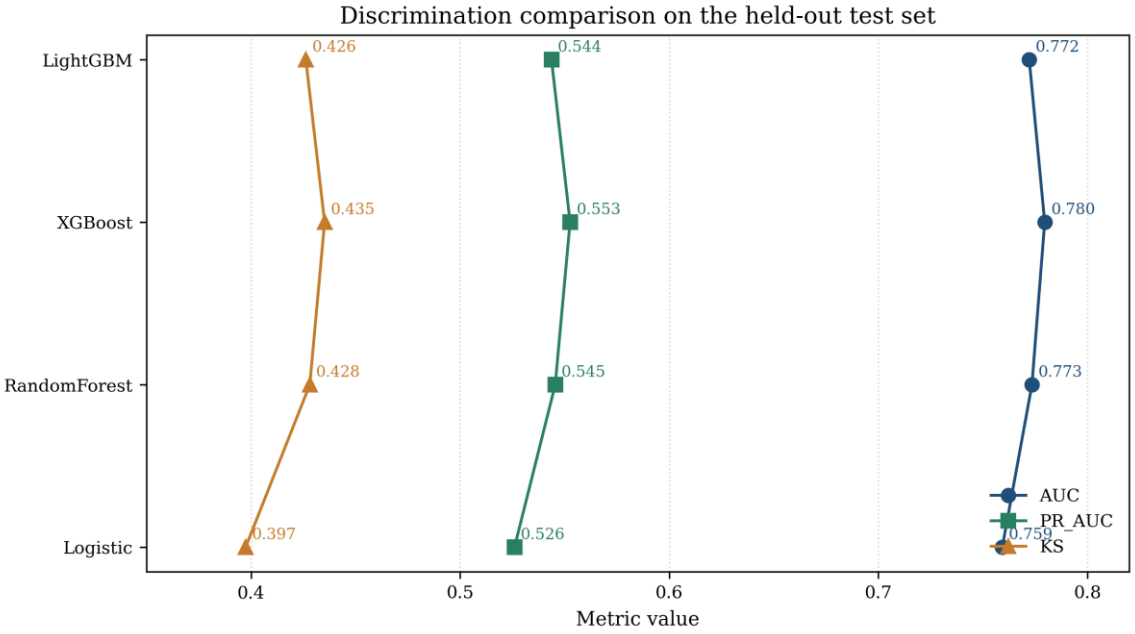

Figure 2. Comparison of uncalibrated discrimination metrics on the held-out test set.

Table 4. Held-out performance metrics for uncalibrated models.

Model	AUC	PR-AUC	KS	Brier	Log loss	ECE (10 bins)
Logistic	0.7593	0.5260	0.3974	0.1387	0.4421	0.0121
RandomForest	0.7735	0.5455	0.4281	0.1373	0.4361	0.0221
XGBoost	0.7796	0.5526	0.4352	0.1351	0.4301	0.0150

Model	AUC	PR-AUC	KS	Brier	Log loss	ECE (10 bins)
LightGBM	0.7722	0.5437	0.4262	0.1370	0.4372	0.0235

Note: Lower is better for Brier, log loss, and ECE. Raw XGBoost is selected as the final model family for uncertainty quantification and SHAP because it gives the strongest overall held-out combination of discrimination and proper-score probability quality.

3.2 Effect of post-hoc calibration

Figure 3 and Table 5 show that calibration effects are model-specific. Logistic regression benefits clearly from sigmoid calibration: ECE falls from 0.0121 to 0.0068 and both Brier and log loss improve slightly. Random forest also improves modestly under sigmoid calibration on proper scores, while isotonic reduces ECE further but without a corresponding Brier or log-loss gain.

For XGBoost, raw probabilities are already strongest; both sigmoid and isotonic worsen proper scores, and sigmoid materially worsens ECE. LightGBM is mixed: sigmoid slightly improves log loss and ECE, raw retains the lowest Brier score, and isotonic lowers ECE at the cost of a large log-loss deterioration. The main conclusion is therefore not that one calibrator always wins, but that calibration must be validated empirically and interpreted with the right metric priority.

Because ECE depends on binning, the final interpretation does not rely on ECE alone. A model family may show a lower ECE yet worse proper scores, which is not a persuasive reason to prefer it for PD reporting. This is especially clear for the random forest and LightGBM variants.

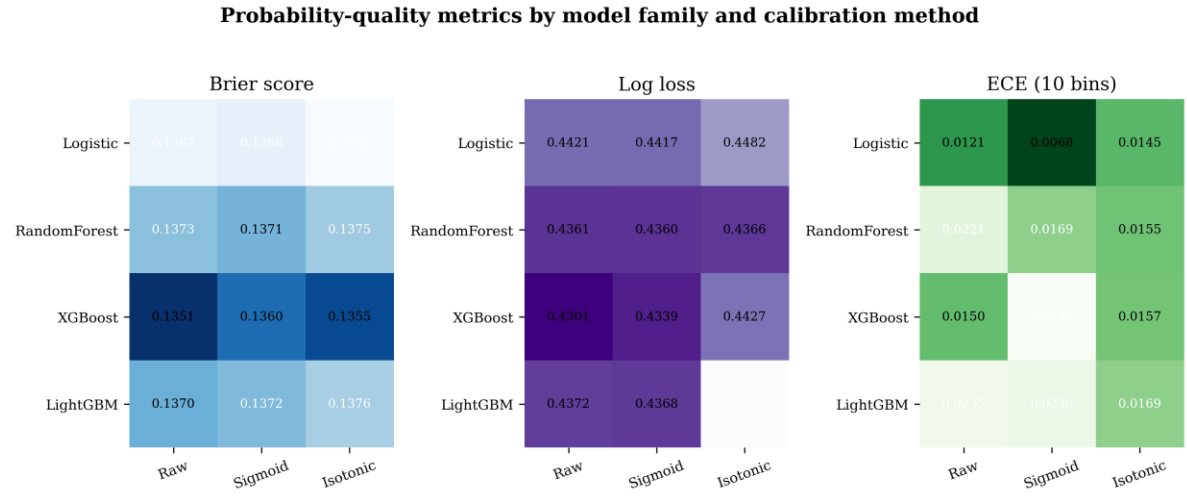

Figure 3. Brier score, log loss, and ECE by model family and calibration method. Darker shading indicates a lower value within each metric.

Table 5. Calibration comparison across model families.

Model	Variant	AUC	Brier	Log loss	ECE (10 bins)
Logistic	Raw	0.7593	0.1387	0.4421	0.0121
RandomForest	Raw	0.7735	0.1373	0.4361	0.0221
XGBoost	Raw	0.7796	0.1351	0.4301	0.0150
LightGBM	Raw	0.7722	0.1370	0.4372	0.0235
Logistic	Sigmoid	0.7593	0.1386	0.4417	0.0068
RandomForest	Sigmoid	0.7735	0.1371	0.4360	0.0169
XGBoost	Sigmoid	0.7796	0.1360	0.4339	0.0245
LightGBM	Sigmoid	0.7722	0.1372	0.4368	0.0230

Model	Variant	AUC	Brier	Log loss	ECE (10 bins)
Logistic	Isotonic	0.7555	0.1389	0.4482	0.0145
RandomForest	Isotonic	0.7725	0.1375	0.4366	0.0155
XGBoost	Isotonic	0.7782	0.1355	0.4427	0.0157
LightGBM	Isotonic	0.7699	0.1376	0.4663	0.0169

Note: Sigmoid preserves ranking; isotonic can alter AUC slightly because it introduces probability ties. Proper scores are the primary basis for probability-model selection; ECE is used as a diagnostic complement.

3.3 Decile behavior of the selected final model

Figure 4 and Table 6 translate the final-model probabilities into a portfolio view. Observed default rate increases monotonically from 4.5% in the lowest-risk decile to 67.8% in the highest-risk decile, while mean predicted PD rises from 4.4% to 72.0%. This monotonic profile is operationally valuable because it supports risk-tier design, manual review rules, and collections prioritization.

The decile comparison also reveals where remaining calibration error lives. The model slightly underpredicts in deciles 4, 5, 6, and 9, and slightly overpredicts in the top decile. None of these gaps reverse the ranking, but they are exactly the kind of segment-level deviations that a validation team would monitor after deployment.

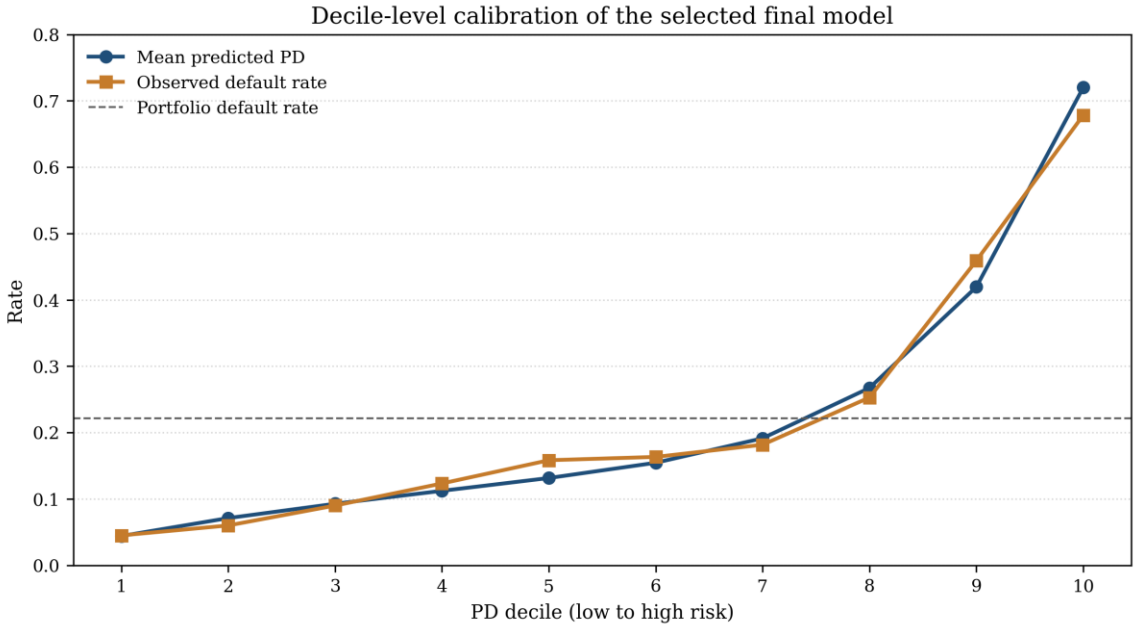

Figure 4. Mean predicted PD and observed default rate by test-set decile for the selected final model.

Table 6. PD decile analysis for the selected final model.

Decile	**Count**	**Mean predicted PD**	**Observed default rate**	**Lift vs. overall**
1	600	0.0441	0.0450	0.2034
2	600	0.0711	0.0600	0.2712
3	600	0.0927	0.0900	0.4069
4	600	0.1122	0.1233	0.5576

Decile	Count	Mean predicted PD	Observed default rate	Lift vs. overall
5	600	0.1315	0.1583	0.7158
6	600	0.1545	0.1633	0.7384
7	600	0.1914	0.1817	0.8213
8	601	0.2673	0.2529	1.1434
9	599	0.4195	0.4591	2.0755
10	600	0.7200	0.6783	3.0666

Note: Lift is the observed default rate in the decile divided by the overall test default rate.

3.4 Uncertainty quantification

Table 7 and Figure 5 show that split conformal prediction delivers the expected coverage-versus-informativeness trade-off. At the 90% target, empirical coverage is 0.9017 and 77.8% of test cases receive singleton predictions. At the 95% target, coverage rises to 0.9462 but singleton predictions fall to 53.1%, so nearly half of cases are explicitly marked ambiguous.

This is a useful governance control. A lender can define a simple policy in which singleton conformal outputs proceed through automated treatment and ambiguous {0,1} cases are routed to manual review or higher documentation requirements. Because the guarantee is marginal and exchangeability-based, this policy is statistically interpretable but should still be accompanied by drift monitoring.

Bootstrap sensitivity bands give a complementary view. Table 8 and Figure 6 show average widths of 0.0852 for 90% bands and 0.0934 for 95% bands, with interval width increasing in higher-risk deciles. Appendix Table A1 shows that the observed decile default rate lies within the mean bootstrap bounds in every decile. This is a coarse but operationally interpretable segment-level stability check.

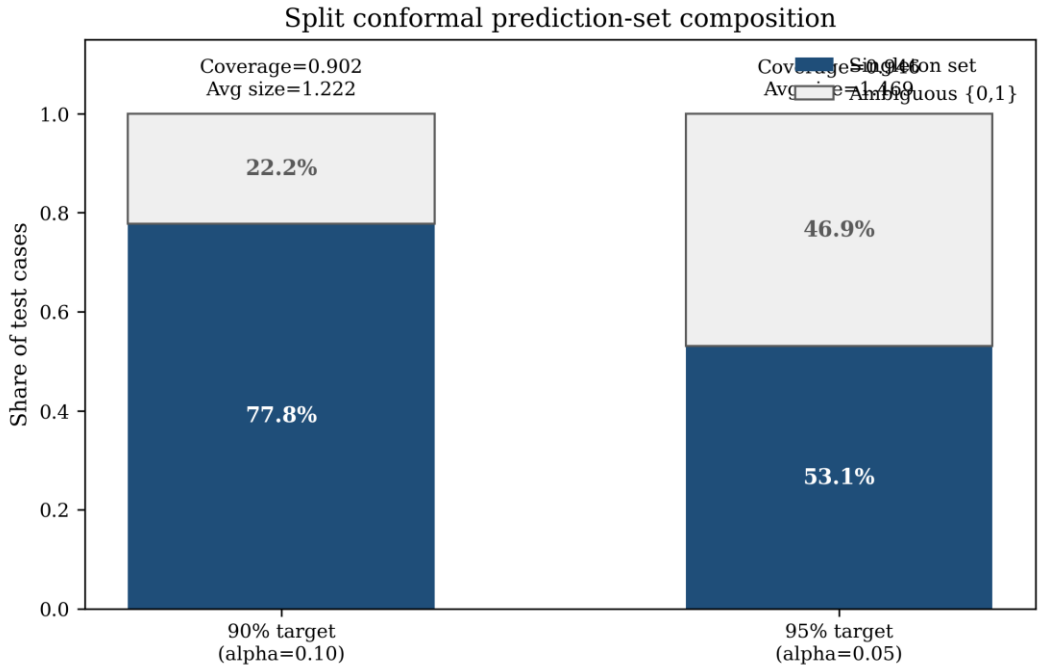

Figure 5. Composition of split conformal prediction sets for 90% and 95% coverage targets.

Table 7. Split conformal prediction-set summary on the test set.

Model	alpha	q_hat	Empirical coverage	Avg. set size	Frac. singleton	Frac. ambiguous {0,1}
XGBoost+raw	0.1000	0.7465	0.9017	1.2220	0.7780	0.2220
XGBoost+raw	0.0500	0.8515	0.9462	1.4687	0.5313	0.4687

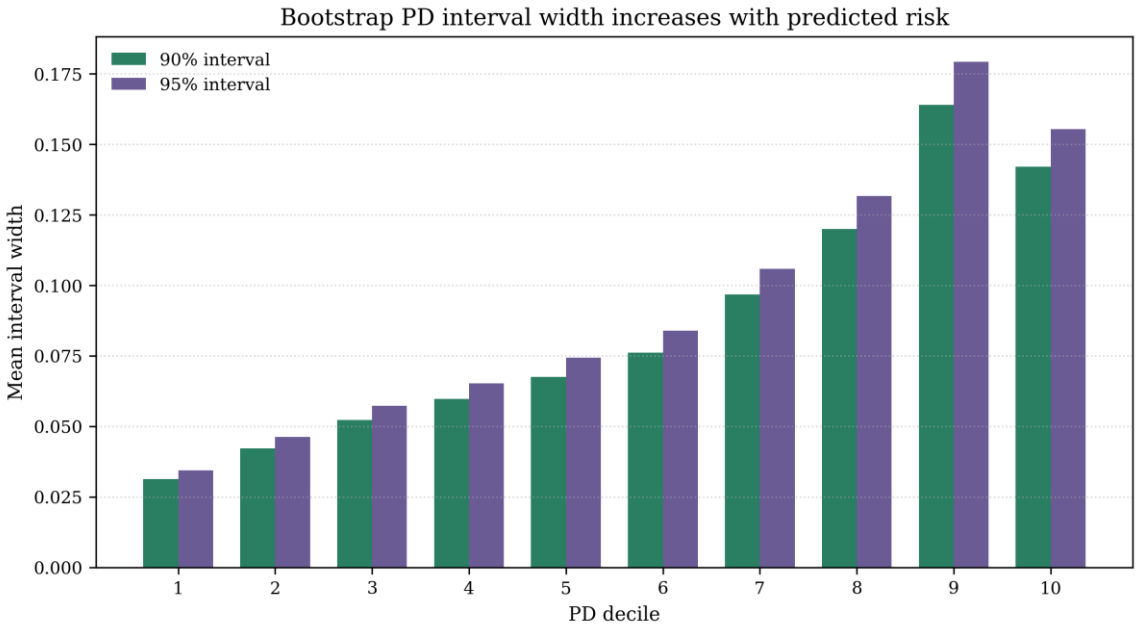

Figure 6. Mean bootstrap PD interval width by decile for 90% and 95% bands.

Table 8. Bootstrap interval summary for the selected final model family.

Model family	Members	Interval	Avg. width	Bin-level coverage
XGBoost-bootstrap	10	90%	0.0852	1.0000
XGBoost-bootstrap	10	95%	0.0934	1.0000

3.5 Explainability and governance interpretation

Figure 7 and Table 9 show that recent repayment status dominates the model. PAY_0 is by far the largest global driver, followed by LIMIT_BAL, PAY_AMT2, PAY_2, BILL_AMT1, and PAY_AMT1. This ranking is economically reasonable: recent delinquency indicators capture immediate deterioration in payment behavior, credit limit proxies borrowing capacity, and bill/payment amounts reflect current balance pressure and repayment effort.

SHAP is most useful here as a governance lens rather than as a claim of causal structure. The contributions help validators test data quality, feature stability, and policy acceptability, and they give case reviewers a coherent explanation of why a borrower received a high or low PD. Combined with the decile and uncertainty

diagnostics, the explanation layer makes the final model substantially easier to defend.

Taken together, the results support a two-layer recommendation. If the primary objective is the strongest overall held-out performance with no evidence that post-hoc calibration improves probabilities, raw XGBoost is the natural final model. If the primary objective is an especially transparent benchmark with the lowest reliability gap, sigmoid-calibrated logistic regression remains a strong challenger. This is the kind of trade-off a model-risk committee should document explicitly.

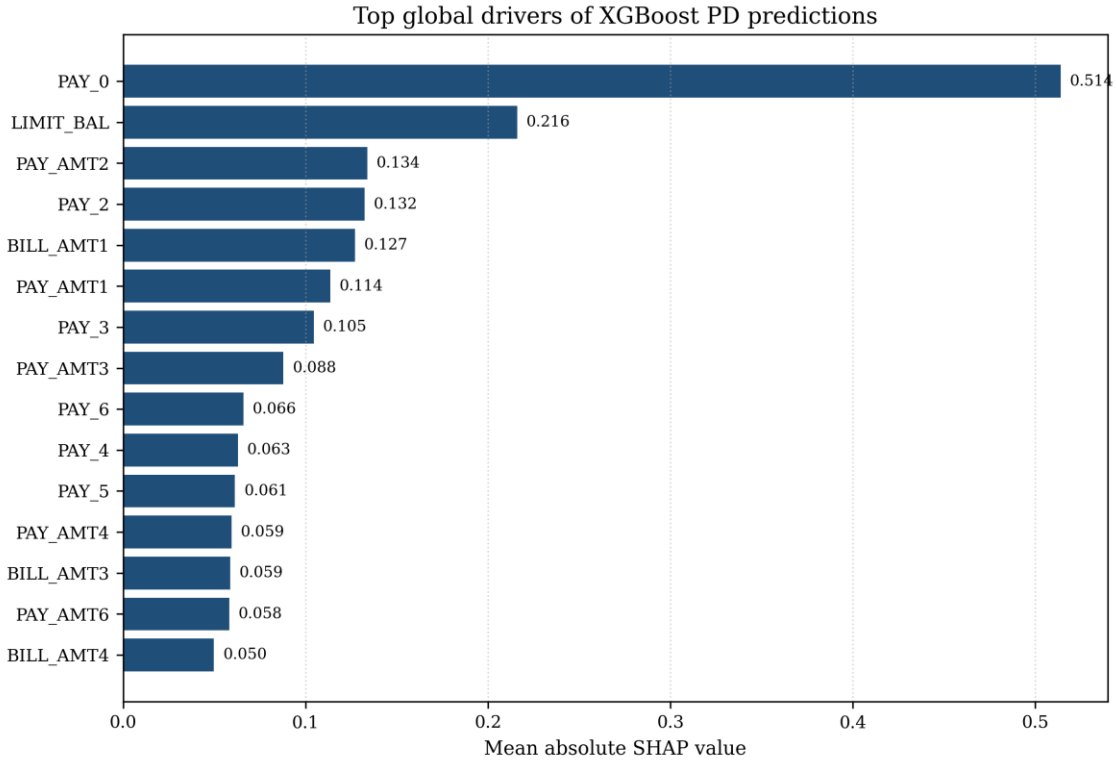

Figure 7. Top global SHAP drivers for the selected XGBoost PD model.

Table 9. Top global drivers by aggregated mean absolute SHAP value.

Feature	**Mean absolute SHAP**
PAY_0	0.5140
LIMIT_BAL	0.2161
PAY_AMT2	0.1339
PAY_2	0.1324
BILL_AMT1	0.1270
PAY_AMT1	0.1136
PAY_3	0.1046
PAY_AMT3	0.0879
PAY_6	0.0660
PAY_4	0.0630
PAY_5	0.0613
PAY_AMT4	0.0594

Feature	Mean absolute SHAP
BILL_AMT3	0.0587
PAY_AMT6	0.0581
BILL_AMT4	0.0498

4. Limitations

The study uses a single public dataset from one institution and one historical period, so the reported metrics should not be read as direct expectations for another lender's portfolio. A production implementation would require institution-specific development data, out-of-time validation, and macro-stress analysis.

Conformal guarantees are marginal and rely on exchangeability between calibration and deployment data. Real portfolios drift. The ambiguous-set rate and the decile calibration profile should therefore be monitored over time and recalibrated when data drift becomes material.

The bootstrap analysis is intentionally lightweight. Ten replicas are sufficient for a segment-level sensitivity demonstration, but not for high-precision tail-quantile estimation. In production, a larger ensemble and repeated-time validation would be preferable.

Finally, the dataset includes demographic variables such as sex and marital status. This paper does not perform a full fairness or adverse-impact analysis. Any real deployment would require policy review, legal sign-off, and subgroup monitoring in addition to the performance checks reported here.

5. Conclusion

This paper presented a model-risk-friendly PD workflow that integrates ranking evaluation, probability validation, uncertainty quantification, and explanation into one auditable protocol. On the UCI credit card default benchmark, raw XGBoost provided the strongest overall combination of discrimination and proper-score probability quality, while sigmoid-calibrated logistic regression achieved the lowest ECE.

The interpretation of the results is deliberately stricter than a pure benchmark comparison. Proper scoring rules are treated as the primary basis for probability-model selection, ECE is used as a diagnostic rather than a sole selector, and uncertainty artifacts are interpreted in operational rather than purely academic terms.

Overall, the study shows that standard tabular models can support governance-ready PD estimation when they are evaluated with disjoint calibration splits, uncertainty diagnostics, and explanation outputs. The resulting template is practical for validation reporting and can be extended with out-of-time testing, fairness review, and institution-specific policy thresholds before deployment.

References

[1] Basel Committee on Banking Supervision, Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Bank for International Settlements, 2006.

[2] Daren Zheng, Boning Zhang, and Julie Geibel, “VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification”, JACS, vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[3] I.-C. Yeh and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications, vol. 36, no. 2, pp. 2473-2480, 2009.

[4] I.-C. Yeh, Default of Credit Card Clients [Dataset]. UCI Machine Learning Repository. doi:10.24432/C55S3H.

[5] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ: Wiley, 2013.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[8] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, 2017, pp. 3146-3154.

[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[10] Xinzhuo Sun, Yifei Lu, and Jing Chen, “Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-

Aware Retrieval-Augmented Generation, and Update/Forgetting", JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[11] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Machine Learning, 2005, pp. 625-632.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning, 2017, pp. 1321-1330.

[13] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1-3, 1950.

[14] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29-36, 1982.

[15] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York: Springer, 2005.

[16] G. Shafer and V. Vovk, "A tutorial on conformal prediction," Journal of Machine Learning Research, vol. 9, pp. 371-421, 2008.

[17] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," arXiv:2107.07511, 2021.

[18] T. G. Dietterich, "Ensemble methods in machine learning," in Multiple Classifier Systems. Berlin: Springer, 2000, pp. 1-15.

[19] Yuanzheng Chen, Yitian Zhang, and Matt Sherman, "Going Concern and Bankruptcy Prediction under Extreme Class Imbalance: Cost-Sensitive Learning, Resampling, and Focal Loss with Explainable Financial-Ratio Portraits", JACS, vol. 4, no. 4, pp. 80–96, Apr. 2024, doi: 10.69987/JACS.2024.40407.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765-4774.

[21] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, pp. 56-67, 2020.

Appendix A. Detailed bootstrap coverage diagnostic

Appendix Table A1 provides the decile-level bootstrap interval diagnostic used to support Table 8 and Figure 6. The observed decile default rate falls inside the mean 90% and 95% interval bounds in every decile.

Appendix Table A1. Bootstrap PD interval coverage by decile.

Decile	N	Mean PD	Obs. rate	Lo90	Hi90	Cov90	Lo95	Hi95	Cov95
1	600	0.0476	0.0500	0.0335	0.0649	1.0000	0.0325	0.0670	1.0000
2	600	0.0754	0.0667	0.0560	0.0982	1.0000	0.0546	0.1009	1.0000
3	600	0.0957	0.0833	0.0720	0.1243	1.0000	0.0701	0.1275	1.0000
4	600	0.1154	0.1217	0.0878	0.1475	1.0000	0.0858	0.1511	1.0000
5	600	0.1358	0.1617	0.1044	0.1720	1.0000	0.1018	0.1762	1.0000
6	600	0.1571	0.1600	0.1209	0.1971	1.0000	0.1177	0.2016	1.0000
7	600	0.1913	0.1700	0.1455	0.2423	1.0000	0.1418	0.2476	1.0000
8	600	0.2630	0.2667	0.2073	0.3272	1.0000	0.2029	0.3345	1.0000
9	600	0.4092	0.4483	0.3303	0.4942	1.0000	0.3236	0.5028	1.0000
10	600	0.7140	0.6833	0.6388	0.7807	1.0000	0.6307	0.7861	1.0000

Note: Cov90 and Cov95 indicate whether the observed decile default rate lies within the corresponding mean interval bounds