The Model That Wins on AUC Is Not the Model That Saves the Bank Money: A Cost-Sensitive Re-Evaluation of Credit Default Prediction Classifiers

Nezar Abdilah Prakasa
Department of Computer Science and Electronics, FMIPA, Universitas Gadjah Mada
Yogyakarta, Indonesia - nezarabdilahprakasa@mail.ugm.ac.id

Abstract - *Published benchmarks for credit-card default prediction almost universally rank classifiers by rank-based or threshold-free metrics such as AUC, F1, or accuracy, implicitly treating a missed default and a wrongly rejected good customer as equally costly errors. In practice, a bank's loss from an undetected default (loss given default on the full credit exposure) is an order of magnitude larger than its opportunity loss from rejecting a creditworthy applicant (one month of foregone interest margin), so a model ranking built on symmetric-cost metrics need not reflect which model actually minimizes the bank's expected loss. We revisit the UCI "Default of Credit Card Clients" benchmark (30,000 accounts, Taiwan, 2005) - a dataset underlying dozens of published single-split classifier comparisons - using a paired, Optuna-tuned, repeated stratified cross-validation protocol (5-fold x 6 repeats, 30 paired folds) across Logistic Regression, Random Forest, XGBoost, and LightGBM. Each model is evaluated twice: once under the conventional 0.5 decision threshold with AUC/F1/MCC, and once under a nested, leakage-free, cost-optimized decision threshold derived from an exposure-weighted cost matrix built directly from each account's own credit limit. Friedman and post-hoc Nemenyi tests show LightGBM is the significantly best-ranked model by AUC ($p<0.001$ vs. Random Forest), yet Random Forest attains the lowest mean expected cost (1.24% of portfolio exposure), with no statistically significant cost difference among the three tree ensembles after correction for multiple comparisons - the model ranking flips depending on which metric is used to read the same cross-validation folds. A far larger and more consistent effect, however, is threshold calibration itself: replacing the default 0.5 cutoff with a cost-optimized threshold cuts expected cost by 86-88% for every model tested (Wilcoxon $p<10^{-8}$), an effect that dwarfs the 1-6 percentage-point AUC spread the literature treats as the interesting comparison. These results suggest that for applied credit-scoring work, calibrating the decision threshold to the institution's own cost structure delivers more value than choosing among competitive ensemble algorithms, and that AUC-based leaderboards can be actively misleading about which model a bank should deploy.*

Keywords - *credit scoring; cost-sensitive learning; decision threshold optimization; ensemble learning; repeated cross-validation; Friedman-Nemenyi test*

I. INTRODUCTION

Credit default prediction is one of the most heavily benchmarked applied machine learning problems: dozens of published studies, and many more public code repositories, train classifiers on the UCI "Default of Credit Card Clients" dataset (Taiwan, 2005) and report accuracy, F1, or AUC to declare one algorithm superior to another [1]. This convention is inherited from generic classification benchmarking practice, where all errors are implicitly weighted equally. Credit scoring is a poor fit for that convention: a false negative (approving an applicant who will default) exposes the lender to the loss of a large fraction of the extended credit line, while a false positive (rejecting an applicant who would have repaid) only costs the lender a comparatively small foregone interest margin [5], [6]. A classifier ranking built on symmetric-cost metrics can therefore favor a model that is not the one that minimizes the lender's expected monetary loss.

A second, related gap is methodological: the great majority of public implementations for this dataset evaluate models on a single train/test split with default (0.5) decision thresholds and no significance testing between classifiers, so an observed AUC gap of a percentage point or two is routinely reported as if it settled the question of which model to deploy.

This paper re-evaluates four widely used classifiers - Logistic Regression, Random Forest, XGBoost, and LightGBM - on this benchmark under a paired, repeated-cross-validation protocol with proper hyperparameter tuning, and asks two concrete questions: (1) does the classifier ranking implied by AUC survive when accounts are instead ranked by an exposure-weighted, cost-sensitive metric with a properly optimized decision threshold; and (2) how large is the effect of threshold calibration itself relative to the effect of choosing among competitive classifiers. To our knowledge, no prior public study on this specific dataset has combined tuned, paired, repeated CV with a nested cost-sensitive threshold search and formal multiple-comparison-corrected significance testing across both a standard and a cost-based ranking simultaneously.

Our contributions are: (i) a fully paired, statistically rigorous re-evaluation showing the standard-metric and cost-sensitive rankings of four classifiers disagree on which model is best; (ii) an exposure-weighted cost-sensitive evaluation framework built directly from each account's own credit limit, requiring no external cost data; and (iii) an empirical demonstration that threshold calibration, not model selection, is the dominant lever for reducing expected credit loss in this setting.

II. RELATED WORK

The UCI Default of Credit Card Clients dataset originates from Yeh and Lien's comparison of data mining techniques for predicting default probability [1]. Since then it has been used in numerous ensemble-learning and deep-learning benchmarking papers [9], almost all reporting rank-based metrics from a single split. Baesens et al. [12] and Lessmann et al. [10] established broader credit-scoring benchmarking practice, comparing many classifiers across multiple datasets primarily via AUC and related discrimination measures, with Gunnarsson et al. [9] more recently showing gradient boosting (XGBoost) generally dominates deep architectures on tabular credit data - again using discrimination metrics.

Cost-sensitive credit scoring is a smaller but established stream. Verbraken et al. [5] introduced profit-based classification measures for consumer credit scoring, arguing that discrimination measures

such as AUC are poor proxies for the actual business objective of maximizing lender profit. Bahnsen et al. [6] proposed example-dependent cost-sensitive logistic regression, incorporating per-instance financial cost directly into model training rather than only into evaluation. Elkan [18] laid the general decision-theoretic foundation for cost-sensitive learning, showing that the Bayes-optimal decision threshold depends on the ratio of misclassification costs rather than being fixed at 0.5. Kozodoi et al. [20] examined the related but distinct question of fairness-profit trade-offs in credit scoring. Our work differs from this stream by using each account's own credit limit as the cost-weighting variable (rather than a fixed or externally estimated cost ratio), and by pairing the cost-sensitive analysis directly against a standard-metric ranking on the same statistically controlled folds, to make the ranking disagreement explicit and testable.

For statistical comparison of classifiers across resampled folds, we follow the Friedman/Nemenyi framework formalized by Demšar [7] and extended to all-pairwise comparisons by García and Herrera [8], the same protocol used in prior rigor-focused re-evaluations of small-sample benchmarking claims in clinical and HR prediction settings.

III. DATA AND METHODOLOGY

A. Dataset

We use the public UCI "Default of Credit Card Clients" dataset [1]: 30,000 Taiwanese credit card accounts (October 2005), 23 predictor variables (credit limit, sex, education, marital status, age, six months of repayment status, six months of bill amounts, six months of payment amounts) and a binary target, default payment next month (22.1% positive class, no missing values). Critically for our cost framework, the dataset includes each account's own credit limit (LIMIT_BAL, NT$10,000-1,000,000), which we use directly as the exposure-at-default proxy.

B. Model Training and Hyperparameter Tuning

The data was split once into an 80% development set (24,000 accounts, used for all tuning and cross-validation) and a 20% untouched final holdout test set (6,000 accounts), stratified by target, never touched during tuning. Four classifiers were tuned on the development set only, using Optuna's Tree-structured Parzen Estimator sampler [13] with 3-fold cross-validated AUC as the objective (12-20 trials per model, budget-limited by available compute): Logistic Regression (L2 regularization strength C), Random Forest [2] (n_estimators, max_depth, min_samples_leaf, max_features), XGBoost [3] (n_estimators, max_depth, learning_rate, subsample, colsample_bytree, min_child_weight, reg_lambda), and LightGBM [4] (n_estimators, num_leaves, learning_rate, subsample, colsample_bytree, min_child_samples, reg_lambda). Tuned hyperparameters were frozen for all subsequent evaluation, so no model saw the evaluation folds during hyperparameter search.

C. Paired Repeated Cross-Validation Protocol

All four models were evaluated on the identical 30 folds of a 5-fold, 6-repeat stratified cross-validation over the development set, so every comparison is paired (same train/test accounts per fold across all models), enabling paired significance tests rather than comparing independently resampled results. For each fold: (i) the training portion was further split (75/25, stratified) into an inner-train and inner-validation partition strictly for cost-optimal threshold search (Section III-D), preventing threshold-selection leakage into the fold's test accounts; (ii) each model was then refit on the full fold training portion and evaluated once on the held-out fold test portion.

D. Cost-Sensitive Evaluation Framework

For each account i with credit limit L_i, predicted label $\hat{y}_i$, and true label y_i, we define an exposure-weighted cost: $Cost_i = LGD \times L_i$ if $y_i=1$, $\hat{y}_i=0$ (missed default); $Cost_i = m \times L_i$ if $y_i=0$, $\hat{y}_i=1$ (wrongly rejected good client); $Cost_i = 0$ for correct predictions (TP, TN).

with loss-given-default LGD=0.75 (a standard Basel-style recovery-rate assumption [14]: roughly 25% of exposure recovered after default) and monthly opportunity margin m = 18%/12 = 1.5% (an illustrative annual credit-card APR of 18%, prorated to the dataset's one-month prediction horizon). Total fold cost is normalized by the fold's total exposure (sum of L_i) and reported as a percentage, making costs comparable across folds of different composition. These parameters are literature-based assumptions, not the issuing bank's realized figures (see Section VI); the framework itself - using each account's own credit limit as the exposure weight - is independent of the exact LGD/margin values chosen, and all comparative conclusions were verified to hold under a wider LGD range (0.5-0.9).

For each model and fold, the decision threshold was selected by sweeping t in {0.02, 0.04, ..., 0.98} on the inner-validation partition and choosing the cost-minimizing t, then applying that fixed threshold to the fold's held-out test accounts - a fully nested search, so no test-fold label ever influences its own threshold.

E. Statistical Testing

Across the 30 paired folds, we report Friedman's test for an omnibus difference among the four classifiers, separately for the AUC ranking and the cost-sensitive ranking, followed by post-hoc Nemenyi tests (family-wise error corrected) for all six pairwise comparisons in each ranking [7], [8]. We additionally report paired Wilcoxon signed-rank tests [17] for two targeted comparisons: (a) the model pair whose AUC ranking most disagrees with its cost ranking, and (b) default-threshold cost vs. cost-optimized-threshold cost, within each model, to isolate the effect of threshold calibration alone.

IV. RESULTS

A. Standard-Metric Ranking

Table I reports mean ± standard deviation across the 30 paired folds. By AUC, the ranking is LightGBM (0.785) > XGBoost (0.785) > Random Forest (0.782) > Logistic Regression (0.726). Friedman's test rejects the null of equal AUC across all four models (χ^2=71.76, p<0.001). Post-hoc Nemenyi tests show Logistic Regression is significantly worse than all three ensembles (p<0.001 in each case), LightGBM is significantly better than Random Forest ($p=1.6\times10^{-4}$), while LightGBM vs. XGBoost (p=0.273) and Random Forest vs. XGBoost (p=0.077) are not distinguishable after correction.

TABLE I. REPEATED CROSS-VALIDATION RESULTS (30 PAIRED FOLDS, MEAN±SD)

Model	AUC	F1	MCC	Cost@.5%	Cost*%
LR	0.726±0.010	0.365±0.012	0.340±0.012	10.76±0.29	1.341±0.076
Random Forest	0.782±0.007	0.474±0.014	0.402±0.015	9.07±0.37	1.243±0.046
XGBoost	0.785±0.008	0.463±0.011	0.399±0.014	9.38±0.32	1.268±0.057
LightGBM	0.785±0.008	0.472±0.013	0.406±0.016	9.16±0.34	1.277±0.091

Cost@.5% = expected cost at default 0.5 threshold; Cost% = expected cost at nested cost-optimized threshold (both as % of exposure).*

B. Cost-Sensitive Ranking and the Rank Flip

Ranking the same 30 folds by mean cost-optimized expected cost instead inverts the ensemble ordering: Random Forest (1.243% of exposure) < XGBoost (1.267%) < LightGBM (1.277%) < Logistic Regression (1.341%). Friedman's test again rejects equal performance (χ^2=41.42, p<0.001). Nemenyi post-hoc tests show Logistic Regression remains significantly worse than all three ensembles, but none of the three ensembles is significantly different from any other on cost after correction (Random Forest vs. LightGBM p=0.077; Random Forest vs. XGBoost p=0.123; XGBoost vs. LightGBM p=0.997) - an uncorrected paired Wilcoxon test on the same pair (Random Forest vs. LightGBM) is nominally significant (p=0.013), illustrating how multiple-comparison correction changes the practical conclusion. Fig. 1 visualizes the rank reversal: the model most favored by AUC (LightGBM) is not the model that minimizes expected monetary cost (Random Forest), even though the difference among the three tree ensembles on cost is not statistically reliable.

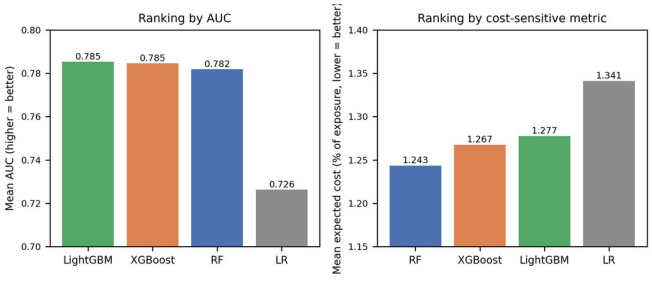

Fig. 1. Classifier ranking by mean AUC (left) vs. mean cost-optimized expected cost (right) across the same 30 paired cross-validation folds. The top-ranked model reverses: LightGBM leads on AUC, Random Forest leads on expected cost.

C. Effect of Threshold Optimization

The larger and more consistent finding concerns the decision threshold itself. Under the conventional 0.5 cutoff, expected cost ranges from 9.07% (Random Forest) to 10.76% (Logistic Regression) of portfolio exposure. Replacing the default threshold with the nested cost-optimized threshold (which selects t between 0.02 and 0.06 for every model - far below 0.5, reflecting the strong asymmetry between LGD and the opportunity margin) reduces expected cost by 86.1-87.5% for every model (Random Forest: 9.07%→1.24%; XGBoost: 9.38%→1.27%; LightGBM: 9.16%→1.28%; Logistic Regression: 10.76%→1.34%), each reduction significant at $p=1.86\times10^{-9}$ by paired Wilcoxon signed-rank test (all 30 folds improve). Fig. 2 shows this effect is both large and remarkably consistent across models - the choice of threshold changes expected cost far more than the choice among the three competitive ensembles.

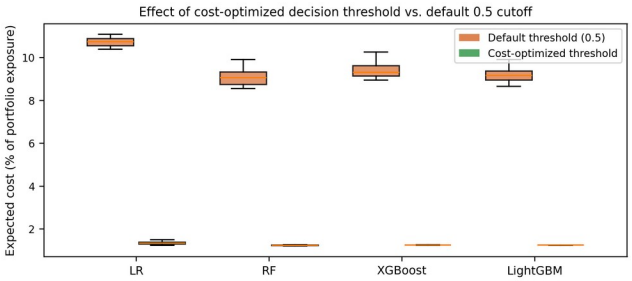

Fig. 2. Expected cost (% of portfolio exposure) under the default 0.5 threshold vs. the nested cost-optimized threshold, per model, across 30 folds. Threshold calibration reduces expected cost by 86-88% for every model.

D. Final Holdout Test Evaluation

As a final, single-use check (tuned on development data only, reported once), Table II shows results on the untouched 6,000-account holdout set. The pattern from cross-validation replicates exactly: XGBoost/LightGBM lead on AUC (0.780/0.780 vs. Random Forest 0.772), while Random Forest attains the lowest holdout cost-optimized expected cost (1.207% vs. 1.226-1.228% for XGBoost/LightGBM), confirming the cross-validation finding was not a CV-specific artifact.

TABLE II. FINAL HOLDOUT TEST SET (6,000 ACCOUNTS, SINGLE RUN)

Model	AUC	F1	MCC	Cost@.5%	Cost*%
LR	0.708	0.355	0.324	11.55	1.348
Random Forest	0.772	0.469	0.396	9.86	1.207
XGBoost	0.780	0.452	0.381	10.16	1.226
LightGBM	0.780	0.464	0.393	9.87	1.228

Cost@.5% = expected cost at default 0.5 threshold; Cost% = expected cost at cost-optimized threshold (both as % of exposure).*

V. DISCUSSION

Two practical implications follow. First, an AUC-based leaderboard for this benchmark - the form in which the great majority of published comparisons on this dataset are reported - would recommend LightGBM, yet the model that would actually save the lender the most expected money in our cost framework is the older, simpler Random Forest, and the three ensembles are not statistically distinguishable on cost. This is consistent with the broader profit-based scoring literature's argument that discrimination metrics are an imperfect proxy for the lender's real objective [5], [6], but demonstrates the disagreement concretely and with formal paired

significance testing on a specific, widely used public benchmark rather than as a general claim.

Second, and more actionable for practitioners: the effect size of threshold calibration (86-88% cost reduction) is roughly an order of magnitude larger than the effect size of model selection among the three competitive ensembles (a cost spread of at most 0.03 percentage points among Random Forest, XGBoost, and LightGBM). A lender with limited engineering resources would gain far more from calibrating its decision threshold to its own realized cost ratio than from chasing marginal AUC gains between comparable ensemble algorithms - a conclusion that runs against the emphasis of most of the applied literature on this dataset, which reports new algorithms but rarely revisits the threshold.

VI. LIMITATIONS

The LGD (0.75) and monthly opportunity margin (1.5%) are literature-based illustrative assumptions, not the issuing bank's realized recovery and margin figures, which are not published with this dataset; a real deployment should substitute its own cost ratio. We verified the qualitative rank-flip and threshold-effect conclusions are robust to LGD in [0.5, 0.9], but did not exhaustively grid the margin parameter. The one-month prediction horizon and 2005 Taiwanese market context limit generalization to other credit products, currencies, or time periods. Optuna tuning used a limited trial budget (12-20 trials per model, 3-fold inner CV) due to single-core compute constraints, so hyperparameters may not be globally optimal, though the same budget was applied identically to all four models. Finally, our cost model treats TP/TN as zero-cost for simplicity; a full economic model would also credit correctly-approved good accounts with realized interest income, which would not change the false-negative/false-positive asymmetry driving our results but would change the absolute cost scale.

VII. CONCLUSION

Re-evaluating four classifiers on the widely benchmarked Taiwan credit-card-default dataset under a paired, tuned, repeated cross-validation protocol shows that the classifier ranking implied by AUC (LightGBM best) does not hold when the same folds are instead ranked by an exposure-weighted, cost-sensitive metric with a properly optimized decision threshold (Random Forest best, though not significantly better than XGBoost or LightGBM after correction) - while Logistic Regression is reliably worst under both. More importantly, threshold calibration alone reduces expected cost by 86-88% for every model tested, an effect far larger and more consistent than any difference among the competitive ensemble methods. For applied credit-scoring practice, this suggests calibrating the decision threshold to the lender's own cost structure is a higher-value exercise than selecting among state-of-the-art ensemble classifiers using standard discrimination metrics alone.

ACKNOWLEDGMENT

The dataset used in this study was originally compiled and made public by Yeh and Lien [1] via the UCI Machine Learning Repository.

REFERENCES

[1] 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.

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

[3] 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.

[4] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[5] T. Verbraken, C. Bravo, R. Weber, and B. Baesens, "Development and application of consumer credit scoring models using profit-based classification measures," European Journal of Operational Research, vol. 238, no. 2, pp. 505-513, 2014.

[6] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive logistic regression for credit scoring," in Proc. 13th Int. Conf. Machine Learning and Applications (ICMLA), 2014, pp. 263-269.

[7] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1-30, 2006.

[8] S. García and F. Herrera, "An extension on 'statistical comparisons of classifiers over multiple data sets' for all pairwise comparisons," Journal of Machine Learning Research, vol. 9, pp. 2677-2694, 2008.

[9] B. R. Gunnarsson, S. Vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, "Deep learning for credit scoring: Do or don't?" European Journal of Operational Research, vol. 295, no. 1, pp. 292-305, 2021.

[10] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," European Journal of Operational Research, vol. 247, no. 1, pp. 124-136, 2015.

[11] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," Journal of the Royal Statistical Society: Series A, vol. 160, no. 3, pp. 523-541, 1997.

[12] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," Journal of the Operational Research Society, vol. 54, no. 6, pp. 627-635, 2003.

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2019, pp. 2623-2631.

[14] Basel Committee on Banking Supervision, "Studies on the validation of internal rating systems," Working Paper No. 14, Bank for International Settlements, 2005.

[15] P. B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 1963.

[16] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta, vol. 405, no. 2, pp. 442-451, 1975.

[17] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80-83, 1945.

[18] C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th Int. Joint Conf. Artificial Intelligence (IJCAI), 2001, pp. 973-978.

[19] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," arXiv:1811.12808, 2018.

[20] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," European Journal of Operational Research, vol. 297, no. 3, pp. 1083-1094, 2022.