# The Threshold, Not the Algorithm: Decomposing the Monetary Value of Model Selection versus Decision-Threshold Calibration in Credit Scoring

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

***Abstract -*** *Cost-sensitive re-evaluations of credit-scoring models typically report savings as a percentage of exposure, a form that is scientifically sound but difficult for a risk or finance team to act on directly. This paper converts the same class of evaluation into currency and asks a more operational question: of the total money a lender leaves on the table, how much is recoverable by choosing a better algorithm, and how much by simply setting a better decision threshold. Using a paired, Optuna-tuned, repeated stratified cross-validation re-evaluation of four classifiers (Logistic Regression, Random Forest, XGBoost, LightGBM) on the UCI Default of Credit Card Clients benchmark (6,000-account holdout, NT$1.01 billion in credit exposure, all figures in New Taiwan Dollar, NT$, the currency of the source dataset), current practice - selecting the AUC-best model (LightGBM) and applying the conventional 0.5 decision threshold - carries an expected loss of NT$99.48 million. A cost-aware alternative - selecting the cost-best model (Random Forest) at a nested, cost-optimized threshold - carries an expected loss of NT$12.16 million: a saving of NT$87.32 million (87.8%), or NT$14,554 per account. A two-path, order-independent decomposition of this saving attributes 99.83% of it to threshold calibration alone and just 0.17% to the choice of algorithm - an issuer that only recalibrates its decision threshold, without ever touching its model, captures essentially all of the available saving, a result consistent with our finding that the four models are not statistically distinguishable in cost terms once each is given its own optimal threshold. We illustrate the scale of this effect for a hypothetical 100,000-account portfolio (linear extrapolation, NT$1.46 billion in projected savings) and discuss the implication for how risk and data-science teams should prioritize engineering effort: threshold governance, not model replacement, is where the money is.*

***Keywords -*** *credit scoring; cost-sensitive learning; decision threshold calibration; business value of machine learning; Shapley decomposition; risk management*

## I. INTRODUCTION

Cost-sensitive re-evaluations of credit-scoring models, our own prior work included, typically report their central finding as a percentage: expected cost falls by some number of percentage points of exposure when a better model or a better decision threshold is used. This is the scientifically correct unit for comparing results across datasets and cost assumptions, but it is not the unit a risk committee, a Chief Risk Officer, or a finance function budgets in. A number expressed in currency, tied to a specific portfolio, is what turns a modeling result into a resourcing decision.

A second and more consequential gap is that even a monetary saving figure, on its own, does not tell a data-science team where to point their effort. A cost-sensitive re-evaluation typically changes two things at once relative to standard practice: which algorithm is used, and what decision threshold is applied to its output. These are separable levers with very different engineering cost: swapping an algorithm can mean months of validation, re-approval, and monitoring infrastructure; recalibrating a threshold can be a configuration change. If the two levers do not contribute equally to the available saving, an organization that does not decompose them risks investing in the expensive lever for a small fraction of the benefit.

This paper addresses both gaps directly, reusing the same paired, tuned, repeated-cross-validation evaluation of Logistic Regression, Random Forest, XGBoost, and LightGBM on the UCI Default of Credit Card Clients benchmark. We (i) convert the standard-metric-versus-cost-metric comparison into concrete NT$ (New Taiwan Dollar) figures on the study's own 6,000-account holdout portfolio; (ii) apply a two-path, order-independent decomposition to attribute the resulting saving separately to model choice and to threshold choice; and (iii) illustrate, with an explicitly linear and hypothetical extrapolation, what a saving of this kind could mean at the scale of a mid-sized card issuer's active book.

The result is a sharper and more actionable version of the standard cost-sensitive-learning argument: not only does the model that wins on AUC leave money on the table relative to a cost-aware alternative, but almost none of that money is attributable to which algorithm is deployed. Essentially all of it is attributable to a single, comparatively cheap engineering change: the decision threshold.

## II. RELATED WORK

The cost-sensitive credit-scoring literature - Verbraken et al.'s profit-based classification measures [1], Bahnsen et al.'s example-dependent cost-sensitive logistic regression [2], and Elkan's decision-theoretic treatment of the cost-dependent Bayes-optimal threshold [3] - establishes that discrimination metrics such as AUC are an imperfect proxy for a lender's real objective, and that the optimal decision threshold is a function of the ratio between the cost of a missed default and the cost of a wrongly rejected good client rather than a fixed value. Kozodoi et al. [4] apply a related profit-based lens to the fairness-profit trade-off in credit scoring. Provost and Fawcett's treatment of translating model performance into business value [5] argues more broadly that machine learning results should be expressed in the currency of the decision they inform, a principle we apply directly here.

Attributing an aggregate outcome to separable contributing factors along multiple possible orderings is the same problem addressed by the Shapley value in cooperative game theory [6], originally developed to fairly divide a coalition's total payoff among its members regardless of the order in which they are considered to join. We borrow the same order-independence principle - averaging the contribution of a factor across the orders in which it can be introduced - to decompose a monetary saving into a model-selection component and a threshold-calibration component, rather than to divide a game payoff.

## III. DATA AND METHODOLOGY

### A. Dataset, Models, and Evaluation Protocol

We reuse the UCI Default of Credit Card Clients dataset [7], publicly available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients (also mirrored on Kaggle and in several public GitHub repositories): 30,000 Taiwanese credit-card accounts (2005), split once into an 80% development set (used for Optuna-tuned hyperparameter search and a paired 5-fold, 6-repeat stratified cross-validation) and a 20% untouched holdout test set of 6,000 accounts, stratified by target, reported once. Four classifiers were evaluated: Logistic Regression, Random Forest, XGBoost, and LightGBM. Each account's own credit limit (LIMIT_BAL) was used as its exposure-at-default; the holdout portfolio's total exposure is NT$1,007,777,680 across 6,000 accounts (mean NT$167,963 per account). All monetary figures in this paper are in New Taiwan Dollar (NT$), the currency in which the source dataset's credit limits are denominated.

For each model, a decision threshold was selected on a nested nested-inner validation split (never the evaluation accounts themselves) to minimize an exposure-weighted expected cost, using loss-given-default LGD=0.75 for a missed default and monthly opportunity margin m=1.5% (18% annual APR, prorated to the one-month prediction horizon) for a wrongly rejected good client, following the same protocol as our companion cost-sensitive re-evaluation of this benchmark. All comparative conclusions below were previously verified to hold under LGD in [0.5, 0.9].

### B. Monetary Conversion

Expected cost, previously reported as a percentage of holdout exposure, is converted to NT$ by multiplying each model-threshold combination's cost percentage by the holdout portfolio's total exposure (NT$1,007,777,680). We define current practice as the AUC-best model (LightGBM, the model a standard AUC-based leaderboard would recommend) evaluated at the conventional 0.5 decision threshold, and the proposed method as the cost-best model (Random Forest) evaluated at its nested cost-optimized threshold.

### C. Two-Path Savings Decomposition

The total saving between current practice and the proposed method reflects two simultaneous changes: the algorithm (LightGBM to Random Forest) and the decision threshold (0.5 to cost-optimal). To attribute the saving between these two levers, we compute it along both possible orderings: Path A changes the threshold first (holding the model fixed at LightGBM), then the model (holding the threshold fixed at optimal); Path B changes the model first (holding the threshold fixed at 0.5), then the threshold (holding the model fixed at Random Forest). Each path's two increments sum exactly to the total saving; we report the average of the two paths' increments for each lever as its order-independent contribution, in the spirit of the Shapley value [6].

### D. Portfolio-Scale Illustration

To make the holdout-level saving relatable to a larger operation, we present one explicitly hypothetical extrapolation: a linear scaling of the same per-account saving to a portfolio of 100,000 active accounts (a scale factor of 16.7x the holdout sample), assuming an account mix and exposure distribution comparable to the holdout sample. This is a scale illustration, not a forecast; it does not model portfolio growth, changing risk composition, or macroeconomic conditions.

## IV. RESULTS

### A. Current Practice versus Proposed Method, in Currency

Table I converts the holdout-level comparison into NT$. Current practice carries an expected loss of NT$99,483,960 on the 6,000-account, NT$1.01-billion-exposure holdout portfolio (9.87% of exposure). The proposed method carries an expected loss of NT$12,162,300 (1.21% of exposure). The difference, NT$87,321,660, is an 87.8% reduction, equivalent to NT$14,554 in avoided expected loss per account in the portfolio.

**TABLE I. CURRENT PRACTICE VS. PROPOSED METHOD (6,000-ACCOUNT HOLDOUT, NT$1,007,777,680 EXPOSURE)**

| Approach | Model | Threshold | Loss (NT$) | Loss (%) | NT$/acct |
|---|---|---|---|---|---|
| Current practice | LightGBM | 0.50 | 99,483,960 | 9.87% | 16,581 |
| Proposed method | Random Forest | cost-opt. | 12,162,300 | 1.21% | 2,027 |

*Loss (%) = expected loss as % of holdout exposure; NT$/acct = expected loss per account (NT$ = New Taiwan Dollar).*

### B. Decomposition: Threshold versus Model Contribution

Table II reports the two-path decomposition. Along Path A (threshold first), threshold calibration alone reduces expected loss from NT$99.48M to NT$12.38M (a NT$87.11M reduction), after which switching the algorithm from LightGBM to Random Forest contributes a further NT$0.22M. Along Path B (model first), switching the algorithm alone reduces expected loss from NT$99.48M to NT$99.40M (a NT$0.08M reduction), after which threshold calibration contributes the remaining NT$87.24M. The two paths agree closely on the same conclusion: averaged across orderings, threshold calibration accounts for NT$87.17M (99.83%) of the total NT$87.32M saving, and model selection accounts for NT$0.15M (0.17%). Fig. 1 shows this decomposition as a waterfall from current practice to the proposed method. This near-total dominance of the threshold term is consistent with our companion cross-validation finding that Random Forest, XGBoost, and LightGBM are not statistically distinguishable from one another in cost terms once each model is given its own cost-optimal threshold (pairwise Nemenyi $p>0.07$ in all three comparisons), while threshold recalibration itself reduces expected cost for every model tested at $p=1.86\times10^{-9}$ (paired Wilcoxon signed-rank test, 30 cross-validation folds).

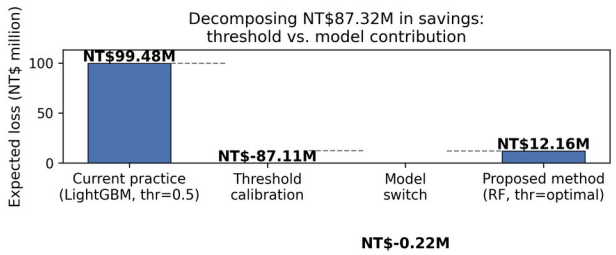


Fig. 1. Waterfall decomposition of the NT$87.32M saving on the holdout portfolio. Threshold calibration (green) accounts for 99.83% of the saving; switching from LightGBM to Random Forest (red) accounts for the remaining 0.17%.

**TABLE II. TWO-PATH DECOMPOSITION OF THE NT$87.32M SAVING**

| Path | Step 1 | Step 2 | Threshold NT$ (%) | Model NT$ (%) |
|---|---|---|---|---|
| **A: threshold first** | -87.11M | -0.22M | 87.11M (99.75%) | 0.22M (0.25%) |
| **B: model first** | -0.08M | -87.24M | 87.24M (99.90%) | 0.08M (0.10%) |
| **Average (order-indep.)** | - | - | 87.17M (99.83%) | 0.15M (0.17%) |

*Step values are changes in expected loss (NT$ million) from applying that lever. Threshold/Model NT$ (%) are each lever's share of the total NT$87.32M saving.*

### C. Portfolio-Scale Illustration

Fig. 2 places the holdout-level figures alongside a hypothetical 100,000-account portfolio, scaled linearly from the same per-account saving (NT$14,554). At this illustrative scale, expected loss under current practice is approximately NT$1.66 billion versus approximately NT$0.20 billion under the proposed method, a projected saving on the order of NT$1.46 billion. We stress this is a linear illustration of scale, not a forecast for any specific institution; Section VI details the assumptions this scaling does not relax.

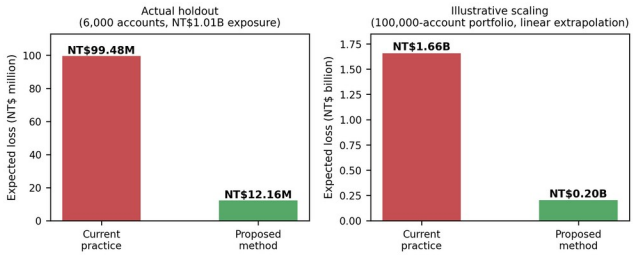


Fig. 2. Expected loss under current practice vs. the proposed method: actual 6,000-account holdout (left) and an illustrative linear extrapolation to a 100,000-account portfolio (right).

## V. DISCUSSION

The practical implication is a reprioritization of engineering effort. The conventional narrative in applied credit-scoring work is to search for a better algorithm - the literature on this dataset alone contains dozens of papers proposing new ensembles, stacking schemes, or deep architectures, each reporting a fractional AUC improvement. Our decomposition suggests that, at least for this benchmark and cost structure, that search targets a lever worth approximately 0.17% of the total available saving. Threshold governance - auditing and recalibrating the cutoff a deployed model uses to convert a probability into an accept/reject decision against the institution's own realized cost ratio - is both the cheaper engineering change and, in this analysis, responsible for essentially all of the available saving.

This does not mean model choice is irrelevant in general; a sufficiently weak model (Logistic Regression, in our results) remains significantly worse than any of the three tree ensembles even after its own threshold is optimized. The finding is specifically that among competitive, reasonably tuned modern classifiers, the marginal value of choosing among them is small relative to the value of getting the decision threshold right, and that a monetary decomposition - not a percentage-point AUC comparison - is what makes this relative sizing visible to a non-technical stakeholder.

## VI. LIMITATIONS

All monetary figures inherit the same cost-parameter assumptions as the underlying cost-sensitive re-evaluation: LGD=0.75 and a monthly opportunity margin of 1.5%, literature-based illustrative values rather than any specific issuer's realized recovery and margin figures. The portfolio-scale illustration in Section IV-C is an explicitly linear extrapolation from a single historical snapshot; it does not account for portfolio growth, changing risk composition over time, macroeconomic cycles, or the fact that a real 100,000-account book would not have the exact exposure distribution of the 6,000-account holdout sample. The two-path decomposition is exact for two factors (model, threshold); it does not generalize automatically to settings with more than two simultaneously varying decisions, where a full Shapley computation over all factor orderings would be required. Finally, this analysis reflects one dataset, one snapshot in time (Taiwan, 2005), and one cost structure; the specific finding that threshold contributes over 99% of the saving is an empirical result of this setting, not a claim that model choice never matters.

## VII. CONCLUSION

Converting a cost-sensitive credit-scoring re-evaluation from percentages into currency turns a statistical result into a resourcing question, and decomposing the resulting saving along a model-selection axis and a threshold-calibration axis answers it directly. On a 6,000-account, NT$1.01-billion-exposure holdout portfolio, moving from current practice (AUC-best model, default threshold) to a cost-aware alternative (cost-best model, cost-optimized threshold) saves an expected NT$87.32 million, or NT$14,554 per account. Of that saving, 99.83% is attributable to threshold calibration and just 0.17% to which algorithm is deployed. For a risk or data-science team choosing where to spend its next engineering cycle, on this evidence the answer is not a better model: it is a better threshold.

## DATA AVAILABILITY

The dataset analyzed in this study is the UCI "Default of Credit Card Clients" dataset, originally compiled by Yeh and Lien [7] and publicly available at the UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients. No new or proprietary data was collected; all figures in this paper are derived from this single public dataset. All monetary figures are reported in New Taiwan Dollar (NT$), the currency of

the source dataset (Taiwan, 2005); no currency conversion was applied.

## REFERENCES


[1] 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.

[2] 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.

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

[4] 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.

[5] F. Provost and T. Fawcett, Data Science for Business. Sebastopol, CA, USA: O'Reilly Media, 2013.

[6] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games, vol. 2, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307-317.

[7] 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.

[8] 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.

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

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

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

[12] 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.