

Cost-Sensitive Ensemble Models for Microcredit Risk Prediction: A Comparative Evaluation

Reliya Indah Okta Limabri¹, Ari Muzakir²

^{1,2} Informatics Department, Postgraduate Program, Bina Darma University, Palembang, Indonesia

² Informatics Department, Faculty of Sciences and Technology, Bina Darma University, Palembang, Indonesia

Received:

July 4, 2026

Revised:

September 5, 2026

Accepted:

September 17, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Ari Muzakir

Email*:

arimuzakir@binadarma.ac.id

DOI:

10.63158/journalisi.v8i5.1902

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. Microcredit risk prediction is challenged by class imbalance and asymmetric misclassification costs. This study evaluates cost-sensitive ensemble models and examines model-selection stability across heterogeneous credit sources. Logistic Regression, Random Forest, XGBoost, and CatBoost were evaluated on one operational microcredit portfolio and three public credit datasets using development-only preprocessing, class-balanced fitting weights, and leakage controls. Model and threshold selection used validation data with a five-to-one false-negative cost ratio and minimum specificity of 0.70. XGBoost was selected for the operational, Credit Risk, and FIN ANA sources, while Logistic Regression was selected for German Credit. On the retrospectively held-out evaluation partitions, ROC-AUC values were 0.895, 0.940, 0.461, and 0.620, with recalls of 0.364, 0.839, 0.083, and 0.483, respectively. FIN ANA showed substantial degradation under broader distribution shift. The study jointly evaluates minority-class recall, asymmetric error costs, and validation-selected operating points, showing that cost-sensitive model selection is source-dependent and ROC-AUC alone is insufficient for practical risk detection. The operational analysis represents retrospective adverse-status screening rather than prospective default prediction because temporal feature snapshots and future outcome windows were unavailable.

Keywords: cost-sensitive classification, ensemble models, class imbalance, microcredit risk prediction, machine learning

1. INTRODUCTION

Credit screening combines two tasks that should not be conflated: ranking applicants by adverse-outcome risk and choosing an operating point for intervention. Logistic Regression remains a defensible benchmark because its score construction is inspectable, whereas tree ensembles can represent nonlinear interactions among borrower and loan attributes [1]. Comparative studies do not identify one uniformly superior model family, and the margin an ensemble holds over simpler alternatives shifts with the imbalance ratio of the sample [2]. Reported performance therefore varies with sample size, prevalence, feature quality, target definition, and evaluation design.

Class imbalance intensifies the decision problem. A classifier may appear accurate while missing most adverse cases, particularly when defaults are rare. Cost-sensitive strategies intervene at different points of the pipeline. One line of work embeds misclassification cost in scorecard construction, so the scorecard itself carries the asymmetry [3]; another treats competing cost-sensitive ensemble variants as instances of a single framework [4]. Li et al. instead spread a cost-weighted objective across learning stages to catch early defaults [5]. Data-level treatments differ again: generative models synthesize minority borrowers before fitting [6], whereas resampling alters the training sample directly, and its benefit depends on the classifier it is paired with [7]. These interventions are analytically distinct. Balanced fitting weights affect parameter estimation, whereas a false-negative cost ratio governs the operating point. Combining them without a documented sequence can obscure which mechanism produced an apparent improvement and limits reproducibility.

Leakage can be more consequential than algorithm choice. Inflated estimates arise when duplicate borrower snapshots cross partitions, preprocessing uses full-sample statistics, contemporaneous arrears reconstruct the label, or the final test set informs model and threshold choice [8]. Heterogeneous credit sources add a comparability hazard: collectibility status, loan-status default, loan-quality categories, and historical good/bad risk are related constructs, but they are not interchangeable outcomes. Pooling them can turn source identity and structural missingness into shortcuts rather than evidence of a general credit-risk signal.

Recent credit-scoring research has addressed several related challenges. A non-parametric explanation method extends attribution to skewed class distributions [9], and ensemble classifiers have been coupled with post hoc explanation so that individual decisions remain auditable [10]. Other studies target the learning objective itself, pairing feature selection with gradient-harmonizing loss [11], or comparing resampling families across different imbalance ratios [12]. Chen et al. examine how interpretability behaves on imbalanced scoring data [13], and Dastile and Celik carry explanation into deep credit models [14]. An Indonesian comparison ranked Logistic Regression above Random Forest and XGBoost on its loan-approval data [15], while a separate study applied XGBoost with SMOTE to German Credit [16].

A few researchers have focused on individual aspects of imbalanced credit-risk prediction, cost-sensitive learning, resampling, or model interpretability. However, comparative studies do not consistently examine minority-class recall, asymmetric error costs, and the selection of the final operating point within one unified validation procedure. In addition, reported performance across credit datasets is often difficult to compare because the sources differ in target semantics, prevalence, observation units, and data provenance. Therefore, this research intends to provide a controlled comparative evaluation of Logistic Regression and ensemble models under a common cost-sensitive protocol while preserving source-specific target definitions and validation boundaries. The study treats the operational microcredit portfolio as the primary application and uses three public credit datasets as supplementary robustness benchmarks rather than as a pooled population.

The objectives of this research are threefold: (1) to compare Logistic Regression, Random Forest, XGBoost, and CatBoost for identifying adverse microcredit accounts under class imbalance; (2) to examine how asymmetric false-negative costs and validation-selected operating points affect model selection and minority-class detection; and (3) to assess the stability of model rankings and selected operating points across heterogeneous credit sources while identifying distribution-shift and data-governance limitations relevant to practical credit decision-support systems. The operational portfolio is treated specifically as a retrospective adverse-status screening task, whereas the public datasets represent broader credit-risk classification settings with different target semantics and observation structures.

2. METHODS

2.1. Comparative Study Design and Validation Architecture

A quantitative comparative design was used to evaluate Logistic Regression, Random Forest, XGBoost, and CatBoost under a common cost-sensitive validation protocol. One operational microcredit portfolio was treated as the primary application, while Credit Risk, FIN ANA, and German Credit were evaluated as supplementary public benchmarks. Each source was modeled independently because the observation units, target definitions, prevalence, and data provenance were not equivalent. The operational portfolio used current collectibility; Credit Risk mapped *loan_status* = 1 to adverse; FIN ANA mapped non-G loan quality to adverse; and German Credit mapped bad risk to adverse. Cross-source comparisons therefore assess the stability of model ordering and cost-sensitive operating points rather than transportability of a single probability-of-default model across populations.

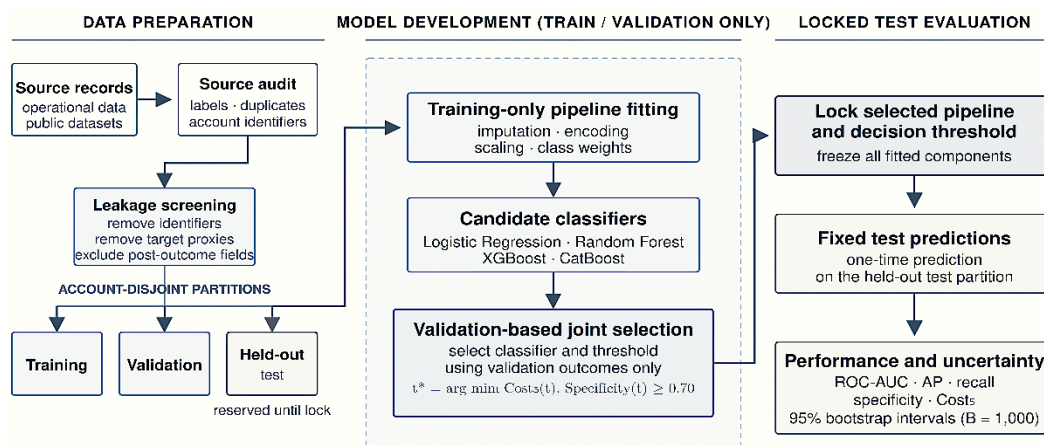


Figure 1. Retrospective Held-Out Evaluation

The validation architecture was designed to separate data auditing, preprocessing, model fitting, threshold selection, and final evaluation. Figure 1 illustrates the overall operational research flow, beginning with source auditing and provenance checking, followed by duplicate detection and target-definition review, leakage screening, source-specific partitioning, development-only preprocessing, class-balanced model fitting, validation-based model and threshold selection, and final test-partition metric calculation. This workflow complements the leakage-aware validation architecture by explicitly mapping the execution sequence used in the empirical analysis.

The final analysis did not use test labels or test metrics for preprocessing, model fitting, model selection, or threshold selection. Earlier exploratory scripts had inspected the evaluation labels, so the evaluation partitions are treated as retrospectively held-out evaluation partitions rather than fully prospectively sequestered confirmatory test sets. The final model and threshold selection procedure did not use these evaluation labels. The remaining invariant leakage controls included preventing duplicate operational accounts from crossing partitions, excluding predictors that reconstructed the target, and fitting all data-dependent preprocessing operations only on development data. Random seed 42 governed stochastic operations, and the model configurations were held fixed for the final selection run.

2.2. Data Sources, Partitions, and Target Semantics

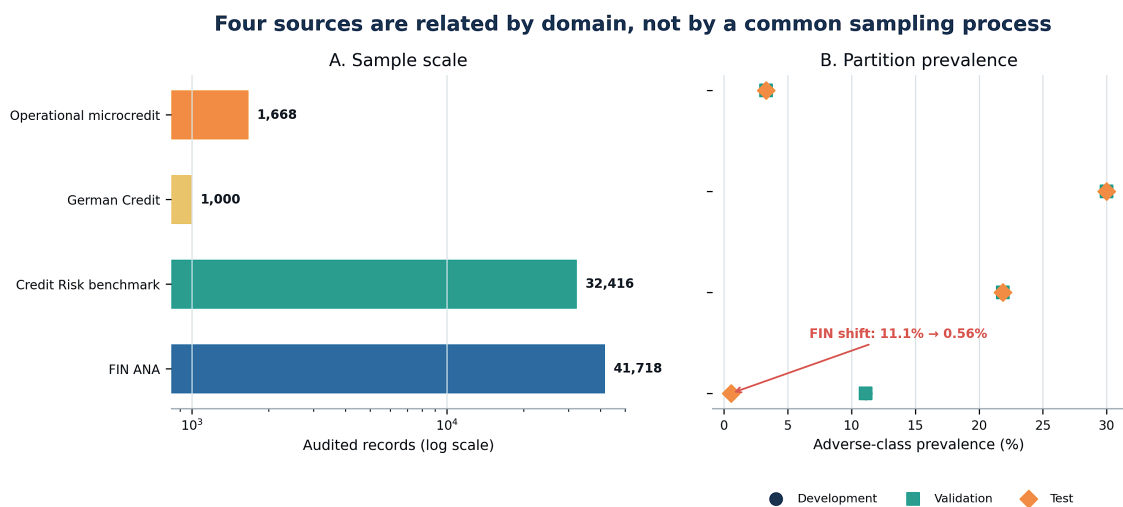
The operational workbook contained a canonical aggregate sheet, *gabungan*, together with overlapping branch/year sheets. A provenance audit identified the same 1,668 accounts across these tabs. Concatenating the sheets would therefore have introduced pseudo-replication. The analysis used only the *gabungan* sheet and removed 50 rows marked as new bookings, resulting in 1,668 labeled accounts with one row per account. Current collectibility values 1 and 2A defined non-adverse status ($n = 1,613$), whereas 2B, 2C, 3, 4, and 5 defined adverse status ($n = 55$; 3.30%). A stratified 60/20/20 split produced 1,000 development, 334 validation, and 334 test accounts, containing 33, 11, and 11 adverse observations, respectively. Thus, the operational analysis represents retrospective adverse-status screening rather than prospective probability-of-default prediction.

After removing 165 exact duplicates, Credit Risk [17] contained 32,416 records with 7,089 defaults (21.87%) and was stratified into 60% development, 20% validation, and 20% test partitions. FIN ANA [18] preserved the uploader-provided separation: 37,408 training-file records were divided 80/20 for development and validation, while all 4,310 supplied test records were retained as the test partition and excluded from the final selection algorithm. The analysis used Kaggle version 1; contrary to its repository description, the downloaded test file contained *QUALITY_OF_LOAN* and 24 adverse labels. File hashes are recorded in the accompanying provenance manifest. Adverse prevalence changed from 11.10% in the training file to 0.56% in the test file. German Credit [19] contained 1,000 records and 300 bad-risk labels, partitioned 60/20/20.

Table 1. Source-specific partitions; adverse observations are shown in parentheses

Source	Audited n	Adverse rate	Development	Validation	Test	Boundary
Operational microcredit	1,668	3.30%	1,000 (33)	334 (11)	334 (11)	Stratified 1/account
Credit Risk	32,416	21.87%	19,449 (4,253)	6,483 (1418)	6,484 (1418)	Stratified
FIN ANA	41,718	10.01%	29,926 (3323)	7,482 (831)	4,310 (24)	Supplied test
German Credit	1,000	30.00%	600 (180)	200 (60)	200 (60)	Stratified

Observation units and target semantics were not harmonized across sources. Operational observations represented current account snapshots, Credit Risk observations represented loan records with native status, FIN ANA observations represented loan-quality categories, and German Credit observations represented historical good/bad applicant risk. Consequently, results are interpreted as within-source benchmarks rather than evidence of a common temporal risk horizon or cross-bank portability. Figure 2 makes the sampling differences explicit. The sources range from 1,000 to 41,718 records, and adverse rates differ by more than an order of magnitude. FIN ANA's prevalence discontinuity and repository-schema inconsistency indicate a population, period, sampling, or labeling change. Its supplied test is therefore a distribution-shift stress case, not an exchangeable random holdout.

**Figure 2.** Sample scale and adverse-class prevalence across partitions

2.3. Predictor Definition and Leakage Gate

Predictors were defined separately within each source. Operational models used log-transformed credit limit, outstanding balance, and instalment; tenor and interest rate; product; balance-adequacy category; economic sector; and purpose code. Credit Risk used the supplied borrower, income, employment, housing, loan, affordability, grade, and bureau-history variables. FIN ANA used investment, current balance, instalment, marital status, gender, client type, and repayment mode. German Credit used age, job category, housing, saving and checking accounts, credit amount, duration, sex, and purpose.

Leakage exclusions were determined from field meaning rather than correlation magnitude. Operational account numbers were retained only for grouping, while names, addresses, contact fields, current days past due, current and previous collectibility, first-payment default, downgrade, arrears components, and the target were excluded. FIN ANA account number, loan quality, due payment, and compensation-charged fields were also removed. Table 2 records these predictor restrictions and leakage exclusions.

These restrictions reduce direct target reconstruction but do not transform the operational analysis into a prospective origination model. Balance adequacy and exposure are contemporaneous portfolio measurements, and the operational workbook does not document an index date followed by a fixed future outcome window. The operational task is therefore described as leakage-reduced adverse-status screening. A prospective prediction study would require dated feature snapshots and outcomes observed after the prediction time.

Table 2. Predictor groups and leakage exclusions

Source	Predictor groups	Excluded from modeling
Operational microcredit	Exposure, balance, instalment, tenor, rate, product, adequacy, sector, purpose	Identifiers; DPD; current/previous collectibility; arrears; FPD; downgrade
Credit Risk	Borrower, income, employment, housing, loan, grade, affordability, bureau history	Target; exact duplicates before splitting
FIN ANA	Investment, current balance, instalment, marital status, gender, client type, repayment mode	Account number; loan quality; due payment; compensation charged

Source	Predictor groups	Excluded from modeling
German Credit	Age, job, housing, savings, checking, amount, duration, sex, purpose	Target; supplied row index

2.4. Preprocessing and Cost-Sensitive Model Selection

Numeric strings were coerced to numbers, invalid values became missing, and right-skewed operational monetary variables were log-transformed. A preprocessing pipeline was fitted on each development partition: numeric variables received median imputation, missingness indicators, and standardization; categorical variables received mode imputation and one-hot encoding. The fitted pipeline was applied unchanged to validation and test records.

Logistic Regression served as the baseline, while Random Forest, XGBoost, and CatBoost were evaluated as ensemble candidates under the same cost-sensitive protocol. Logistic Regression used L2 regularization, $C=1.0$, the liblinear solver, and 2,000 maximum iterations. Random Forest used 250 trees, depth 18, minimum leaf size four, and square-root feature sampling [1]. XGBoost used 250 trees, depth five, learning rate 0.05, row and column subsampling of 0.8, and minimum child weight five. CatBoost used 250 iterations, depth six, learning rate 0.05, and L2 regularization of five. These configurations were held fixed in the final analysis.

Class handling used a single weighting mechanism during model fitting. For class cc , the class-balanced weight was defined as Equation 1.

$$w_c = \frac{N}{2n_c} \quad (1)$$

where N is the number of development observations and n_c is the number of observations belonging to class c . No additional *class_weight* or *scale_pos_weight* multiplier was applied. For XGBoost, the weighted binary logistic objective can be expressed as Equation 2.

$$\mathcal{L}_{XGB} = - \sum_{i=1}^N w_{y_i} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \Omega(f) \quad (2)$$

where $y_i \in \{0,1\}$ is the observed class, p_i is the predicted probability of the adverse class, w_{y_i} is the class-balanced fitting weight, and $\Omega(f)$ represents the regularization term applied to the boosted-tree model. For CatBoost, the corresponding weighted binary log-loss can be written as Equation 3.

$$\mathcal{L}_{CB} = - \sum_{i=1}^N w_{y_i} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (3)$$

where the class-balanced weight w_{y_i} is incorporated into the contribution of each training observation. The equations describe the weighted fitting objective used for the binary classification setting; the internal optimization procedures of XGBoost and CatBoost remain implementation-specific.

Each fitted model generated validation probabilities. A grid of 197 candidate thresholds from 0.01 to 0.99 in increments of 0.005 was evaluated. For a false-negative cost multiplier λ , the normalized cost was defined as Equation 4.

$$C_\lambda = \frac{\lambda FN + FP}{\lambda P + N} \quad (4)$$

where FN denotes false negatives, FP denotes false positives, P denotes the number of adverse observations, and N denotes the number of non-adverse observations.

The primary analysis used $\lambda=5$, meaning that a false negative was assigned five times the decision cost of a false positive within the normalized scenario. Threshold candidates were eligible only when validation specificity was at least 0.70. The classifier and threshold with the minimum eligible validation cost were selected. Equal-cost thresholds were resolved by higher F2 and then recall; model ties were resolved by higher specificity and then recall. The adverse class was encoded as 1, and the probability column was checked to ensure that the adverse class occurred exactly once. Because normalized Cost_λ is calculated from classification errors within each source, its absolute value is prevalence-sensitive and should primarily be used to compare operating rules within the same source rather than directly across datasets with different adverse-class prevalence.

The five-to-one ratio is a reproducible sensitivity scenario rather than an estimate of actual rupiah loss, expected credit loss, or regulatory capital. In practical financial deployment, the false-negative multiplier λ should instead be estimated from institution-specific loss exposure. A conceptual basis is the expected incremental loss from failing to identify an adverse account relative to the operational cost of reviewing or incorrectly flagging a non-adverse account. Expected credit loss can be informed by exposure at default, probability of default, loss given default, and expected recovery, while false-positive cost can incorporate credit-officer review time, customer friction, delayed processing, and other intervention costs. Thus, an institution could estimate λ from the ratio of expected false-negative loss to expected false-positive intervention cost and then evaluate a range of plausible values rather than adopting a fixed 5:1 ratio as a universal policy.

This distinction is important because the selected threshold is a decision-policy parameter, whereas class-balanced fitting weights influence model estimation. Keeping these mechanisms separate makes the cost-sensitive procedure reproducible and prevents the fitting weights from being interpreted as monetary loss parameters. Within each source, the final classifier and threshold were selected using only validation data. Evaluation-set rankings were not used to replace the validation-selected model. For FIN ANA, G was mapped to class 0 and B, DF, or SS to class 1 in both files.

Table 3. Classifier configurations used in the final analysis

Classifier	Configuration
Logistic Regression	L2; C=1.0; liblinear; max_iter=2,000
Random Forest	250 trees; depth=18; min leaf=4; max_features=sqrt
XGBoost	250 trees; depth=5; learning rate=0.05; row/column sample=0.8
CatBoost	250 iterations; depth=6; learning rate=0.05; L2=5

2.5. Evaluation, Interpretation, and Responsible Use

Retrospectively held-out evaluation reporting included ROC-AUC, average precision (AP), accuracy, adverse-class recall, precision, specificity, F1-score, confusion counts, normalized cost, Brier score, and log loss. ROC-AUC measures ranking discrimination, AP emphasizes retrieval of the minority class, threshold-dependent metrics describe the

selected operating point, and Brier score and log loss provide probability-scale diagnostics. Because class-balanced fitting alters the fitted score scale and no independent calibration sample was available, the model outputs were not interpreted as calibrated default probabilities. The Brier score used in the probability-scale diagnostics was calculated as in Equation 5.

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2 \quad (5)$$

where p_i is the predicted probability of the adverse class and $y_i \in \{0,1\}$ is the observed outcome. Lower Brier scores indicate smaller squared probability errors, although interpretation remains sensitive to class prevalence and calibration quality.

For each fixed selected model and threshold, 95% percentile intervals were estimated from 1,000 evaluation-sample bootstrap resample. Operational resampling used the account as the sampling unit, whereas public datasets used records and assumed independence because reusable borrower identifiers were unavailable. These intervals quantify uncertainty within the observed test samples but do not incorporate uncertainty from preprocessing, model selection, threshold selection, or the single partitioning scheme.

Importance was represented by absolute coefficients after preprocessing for Logistic Regression and native tree importance for the ensemble models. These measures describe model associations rather than causal effects. Sensitivity analysis reselected the model and threshold on validation data for false-negative cost ratios of 1, 2, 5, 10, and 20 and validation specificity floors of 0.60, 0.70, 0.80, and 0.90. The specificity floor governed selection and was neither enforced nor guaranteed on the test partitions.

The primary predictions were also rescored across alternative cost ratios. Threshold 0.50, all-non-adverse, and all-adverse rules served as transparent reference baselines; the all-adverse rule was ineligible under every positive specificity constraint. Computations were performed locally using Python 3.12.13, pandas 3.0.5, NumPy 2.4.6, scikit-learn 1.7.2, XGBoost 3.1.2, and CatBoost 1.2.8. Direct identifiers and raw operational records were excluded from the reproducibility package; file hashes, code, split counts, aggregate outputs, and

public-source metadata were retained. The resulting models are research artifacts and do not authorize approval, rejection, pricing, or collection actions. Operational use would require temporal validation, independent calibration, lawful subgroup assessment, privacy review, and model governance [20].

2.6. Ethics and Data Governance

This study used secondary retrospective administrative records and public datasets, with no participant contact or intervention. Direct identifiers in the operational workbook were used only to identify duplicate accounts and enforce account-disjoint partitioning. They were excluded from model inputs, reports, and shared artifacts, and only aggregate, non-identifying results are reported.

3. RESULTS AND DISCUSSION

3.1. Comparative Model Discrimination Across Credit Sources

Table 1 establishes the data-partition context for interpreting the test results. The test sets contain 11 adverse operational accounts, 24 adverse FIN ANA records, 60 bad-risk German Credit records, and 1,418 Credit Risk defaults. Consequently, performance estimates for the operational microcredit and FIN ANA sources are subject to substantially greater sampling uncertainty than those for Credit Risk.

Across the sixteen model-source combinations, test-partition ROC-AUC rankings differed by source, and no classifier ranked first across all four comparisons. Figure 3 summarizes the test-partition ROC-AUC values for all model-source combinations. XGBoost achieved the highest ROC-AUC for the operational microcredit portfolio (0.895) and the Credit Risk benchmark (0.940). Random Forest produced the highest ROC-AUC on German Credit (0.667), although Logistic Regression was selected by the validation procedure. For FIN ANA, Logistic Regression achieved a ROC-AUC of 0.771, whereas the validation-selected XGBoost obtained only 0.461 on the supplied test partition. Logistic Regression was not substituted after observing this test result because doing so would violate the predefined validation-based selection procedure and contaminate the final evaluation.

The differences across sources indicate that model superiority is conditional rather than universal. The relatively strong XGBoost performance on Credit Risk is supported by a

sufficiently large adverse class, whereas the operational and FIN ANA sources contain sparse adverse test observations. In addition, the public datasets differ in target semantics, feature definitions, sampling procedures, and population composition. Therefore, the ROC-AUC values should be interpreted as within-source comparative evidence rather than evidence that one algorithm has a universally higher probability of generalizing to other credit populations.

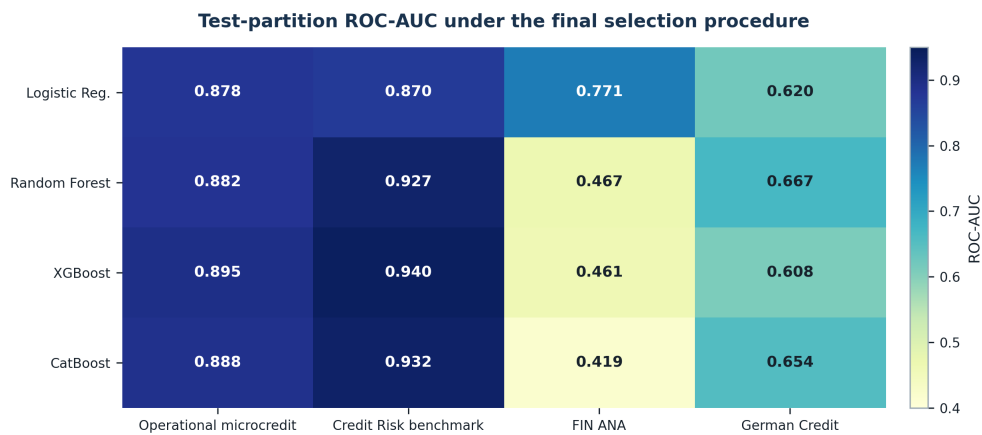


Figure 3. Test-partition ROC-AUC for all model–source combinations

3.2. Cost-Sensitive Model Selection and Test-Partition Performance

Validation selected XGBoost for the operational microcredit, Credit Risk, and FIN ANA sources, and Logistic Regression for German Credit. Table 4 reports the corresponding test-partition results using fixed models and validation-selected thresholds. The threshold was selected exclusively from the validation partition under the five-to-one false-negative cost scenario and the minimum specificity constraint of 0.70.

Table 4. Fixed-model, fixed-threshold test metrics; 95% bootstrap intervals are in parentheses

Source	Model / threshold	ROC-AUC	AP	Recall	Prec.	Spec.	Cost5	Brier
Operational microcredit	XGBoost (t=0.380)	0.895	0.148	0.364	0.129	0.916	0.186	0.062
		(0.853–0.936)	(0.080–0.289)	(0.077–0.688)	(0.027–0.257)	(0.886–0.946)	(0.111–0.269)	(0.043–0.081)
Credit Risk	XGBoost (t=0.405)	0.940	0.894	0.839	0.684	0.892	0.261	0.075
		(0.933–0.948)	(0.882–0.906)	(0.819–0.857)	(0.663–0.704)	(0.883–0.900)	(0.239–0.284)	(0.071–0.078)

Source	Model / threshold	ROC-AUC	AP	Recall	Prec.	Spec.	Cost5	Brier
FIN ANA	XGBoost (t=0.555)	0.461	0.005	0.083	0.003	0.829	0.196	0.184
		(0.377–	(0.003–	(0.000–	(0.000–	(0.817–	(0.181–	(0.179–
		0.555)	0.008)	0.214)	0.007)	0.840)	0.210)	0.187)
German Credit	Logistic Reg. (t=0.550)	0.620	0.486	0.483	0.433	0.729	0.965	0.236
		(0.528–	(0.365–	(0.355–	(0.313–	(0.652–	(0.735–	(0.215–
		0.708)	0.606)	0.607)	0.542)	0.797)	1.205)	0.260)

For the operational microcredit portfolio, XGBoost achieved a ROC-AUC of 0.895 (95% CI: 0.853–0.936), indicating strong ranking discrimination. However, recall was only 0.364 (95% CI: 0.077–0.688) at the selected threshold of 0.380. The model detected 4 of 11 adverse accounts, missed 7 adverse accounts, and flagged 27 non-adverse accounts. This resulted in 31 flagged accounts in total, including 4 adverse and 27 non-adverse accounts. Thus, approximately 7.75 flagged accounts required review for each adverse account detected, illustrating the trade-off between adverse-case detection and credit-officer workload. The apparent contrast between high ROC-AUC and modest adverse-class recall demonstrates why ranking discrimination alone is insufficient for an asymmetric credit-risk decision problem. ROC-AUC evaluates ranking across thresholds, whereas the selected operating point determines which accounts are actually flagged for review.

The operational result also has a direct managerial implication. Under a five-to-one false-negative cost assumption, the model did not simply maximize the number of flagged accounts. Instead, it operated at a threshold constrained by validation specificity, resulting in 27 non-adverse accounts being flagged in order to identify 4 adverse accounts. This represents a practical review trade-off: reducing missed adverse accounts requires additional credit-officer workload and may increase customer friction, while setting a higher threshold would reduce unnecessary reviews but increase the number of adverse accounts that remain undetected. Because the operational test set contains only 11 adverse accounts, this trade-off should not be interpreted as a stable production operating point.

Credit Risk provided the strongest internal benchmark. XGBoost achieved a ROC-AUC of 0.940 (95% CI: 0.933–0.948), AP of 0.894, recall of 0.839 (95% CI: 0.819–0.857), and specificity of 0.892 at a threshold of 0.405. The model detected 1,189 of 1,418 defaults,

with 549 false positives and 229 false negatives. Validation recall was 0.847 compared with 0.839 on the test partition, indicating close within-source consistency. This consistency should be interpreted only as internal split stability and not as external validation. FIN ANA exhibited the largest validation-to-test degradation. Validation selected XGBoost at a threshold of 0.555, where recall was 0.490 and specificity was 0.826. On the supplied test partition, ROC-AUC decreased to 0.461 (95% CI: 0.377–0.555), AP to 0.005, and recall to 0.083 (95% CI: 0.000–0.214). Only 2 of 24 adverse records were detected, while 735 non-adverse records were flagged.

Figure 4 summarizes the validation-to-test changes in recall and specificity at the fixed, validation-selected thresholds. The operational recall decreased from 0.636 during validation to 0.364 on the test partition, while Credit Risk changed only from 0.847 to 0.839. FIN ANA showed the largest validation-to-evaluation degradation. Validation selected XGBoost at threshold 0.555, with recall 0.490 and specificity 0.826, whereas the supplied evaluation partition achieved ROC-AUC 0.461 and recall 0.083. The adverse-class prevalence also decreased from 11.10% in the training data to 0.56% in the evaluation partition, which can affect threshold behavior, calibration, and realized cost. However, prevalence shift alone does not explain the ROC-AUC below 0.50, which indicates that the score ranking itself did not transfer to the evaluation population. This suggests broader distribution differences involving feature distributions, sampling, coding, label semantics, or missingness, and therefore threshold adjustment alone is unlikely to restore performance.

The FIN ANA degradation is technically consistent with the severe change in class prevalence between development and test data. The adverse prevalence changed from approximately 11.10% in the training data to only 0.56% in the supplied test data. A threshold selected on the development distribution therefore operates under a substantially different prior probability when applied to the supplied test population. In addition, the balanced fitting weights used during development modify the training objective and consequently the scale of the resulting scores. The fixed threshold of 0.555 was selected under the development distribution and was not recalibrated using the supplied test labels. Under the extreme prevalence shift, the same score threshold can therefore produce a very different relationship between predicted scores and actual adverse outcomes.

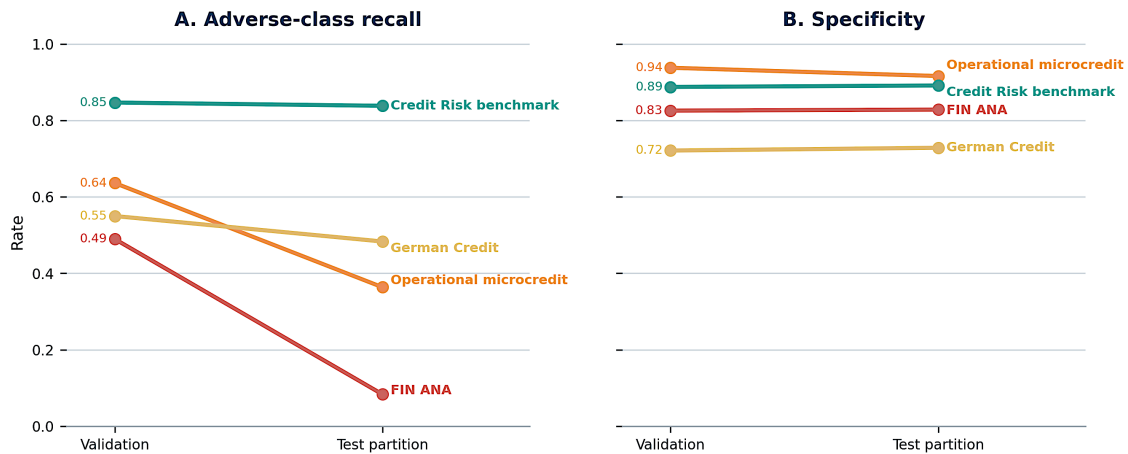


Figure 4. Validation-to-test changes in recall and specificity at fixed thresholds

The resulting ROC-AUC below 0.50 further indicates that the problem is not merely an unfavorable threshold: the ranking itself failed to transfer to the supplied FIN ANA test population. Positive-class and probability-column checks excluded a simple label inversion as the explanation. The result should therefore be interpreted as evidence of severe supplied-test distribution failure rather than as evidence that threshold adjustment alone would restore model performance

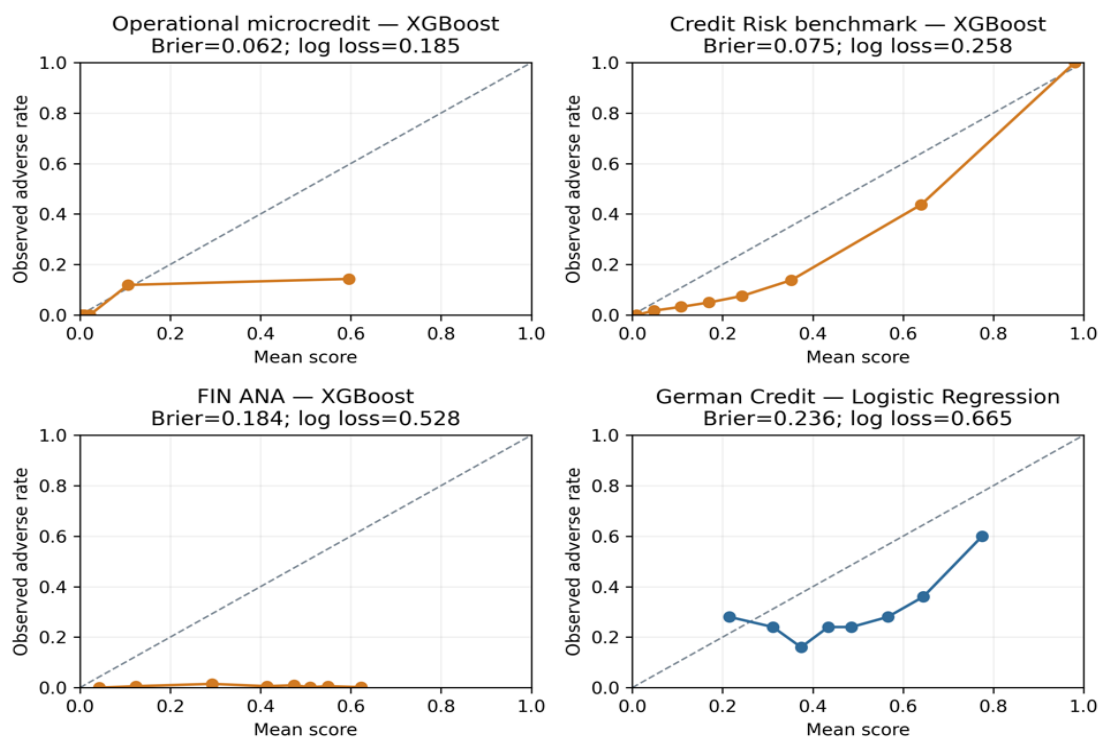


Figure 5. Exploratory probability-scale diagnostics on test partitions

Figure 5 provides exploratory probability-scale diagnostics for the four selected models on their respective test partitions. The accuracy of 0.824 for FIN ANA should therefore not be interpreted as evidence of useful classification performance. The supplied test partition contains 4,286 non-adverse records, so a large proportion of correct non-adverse predictions can coexist with very poor adverse-class detection. This result reinforces the need to evaluate recall, precision, AP, and discrimination together when adverse cases are rare.

German Credit remained exploratory. Validation-selected Logistic Regression achieved ROC-AUC 0.620 (95% CI: 0.528–0.708), recall 0.483 (95% CI: 0.355–0.607), specificity 0.729, and F1-score 0.457. Although Random Forest obtained the highest test ROC-AUC among the four models on this source, Logistic Regression remained the selected model because model selection was performed using validation data rather than test results. With only 200 test records and 60 adverse observations, the observed model ordering should not be interpreted as a definitive leaderboard.

3.3. Misclassification Costs and Error Structure

At the primary five-to-one false-negative cost scenario, the normalized test costs were 0.186 for operational XGBoost, 0.261 for Credit Risk XGBoost, 0.196 for FIN ANA XGBoost, and 0.965 for German Credit Logistic Regression. Relative to the threshold-0.50 baseline, the corresponding cost differences were 0.000 for operational microcredit, -0.014 for Credit Risk, -0.172 for FIN ANA, and -0.010 for German Credit. The all-non-adverse baseline produced costs of 0.165, 1.094, 0.028, and 1.500, respectively. Thus, the selected operating rule improved on the all-non-adverse baseline only for Credit Risk and German Credit. The all-adverse rule was ineligible because its specificity was zero. Notably, the all-non-adverse baseline produced lower realized Cost_s than the selected rule for the operational and FIN ANA evaluations. This does not invalidate the modeling procedure; rather, it shows that a validation-selected cost-sensitive rule does not guarantee lower realized cost when the evaluation population differs from the development distribution.

These normalized costs represent weighted classification counts rather than monetary losses, expected credit losses, or portfolio profitability. The five-to-one ratio should therefore be interpreted as a reproducible decision scenario rather than an empirically estimated financial loss ratio. In an operational credit environment, the false-negative

multiplier should instead be estimated from institution-specific economic consequences. The potential false-negative cost may incorporate exposure at default, probability of default, loss given default or recovery rate, and the effectiveness of subsequent intervention. The false-positive cost may include credit-officer review time, processing delays, customer friction, and the cost of incorrectly escalating a non-adverse account. Consequently, a production threshold should be determined from an institution-specific cost model rather than directly adopting the five-to-one ratio used in this benchmark. Figure 6 presents the confusion matrices at the fixed, validation-selected thresholds. The matrices show the actual numbers of true negatives, false positives, false negatives, and true positives underlying the reported recall, specificity, and precision values.

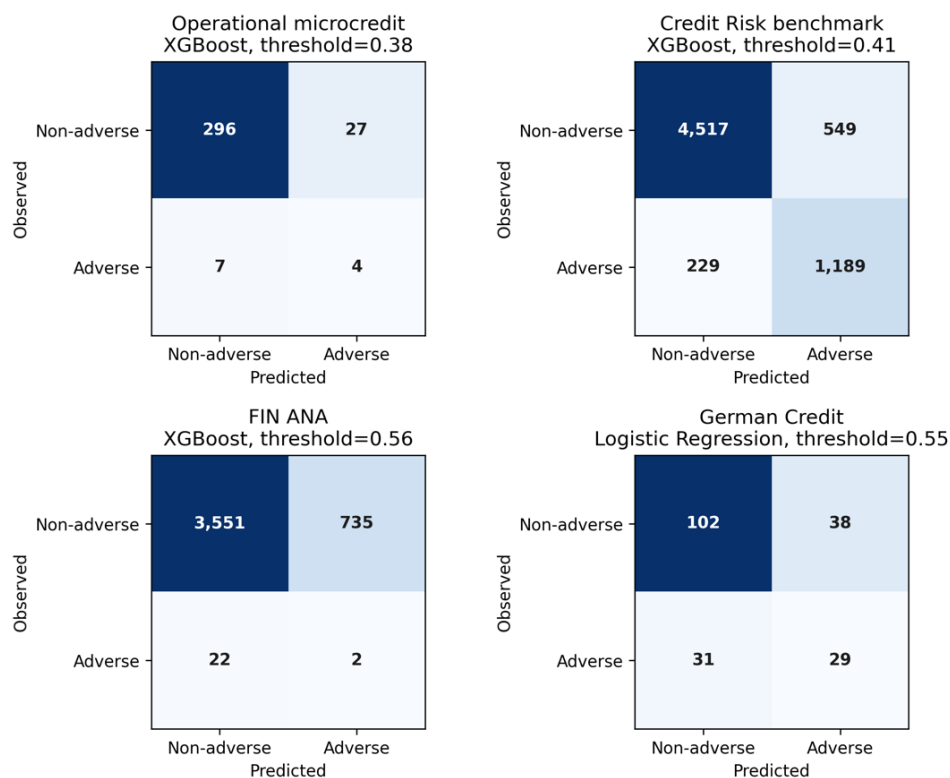


Figure 6. Confusion matrices at fixed, validation-selected thresholds

Scenario re-selection further demonstrates that model choice depends on the assumed cost structure. At a validation specificity floor of 0.70, model selection changed from CatBoost at cost ratios 1–2 to XGBoost at ratios 5–10 and CatBoost at ratio 20 for operational microcredit. For Credit Risk, CatBoost was selected at ratios 1–2 and XGBoost at ratios 5–20. For FIN ANA, Random Forest was selected at ratios 1–2, while XGBoost was

selected only at ratio 5. For German Credit, CatBoost was selected at ratio 1 and Logistic Regression at ratios 2–20.

These changes show that cost-sensitive model selection is not simply a property of the algorithm. It is jointly determined by the classifier, class distribution, validation performance, threshold constraint, and assumed relative cost of errors. For the fixed primary operational predictions, normalized costs increased from 0.102 to 0.123, 0.186, 0.290, and 0.500 as the false-negative multiplier increased from 1 to 2, 5, 10, and 20. The pattern confirms that the selected threshold and model cannot be separated from the decision policy used to evaluate them.

For management, this finding is important because increasing the penalty for missed adverse borrowers does not automatically imply that the model itself becomes better. Instead, it changes the preferred operating point and potentially increases the number of accounts requiring human review. Therefore, the final threshold should jointly consider default prevention benefits, review capacity, intervention effectiveness, and customer impact. Figure 7 illustrates how the normalized test cost of the fixed primary predictions changes under alternative false-negative cost ratios. The increasing cost values under larger false-negative multipliers show why the economic interpretation of the error ratio must be established before selecting an operational threshold.

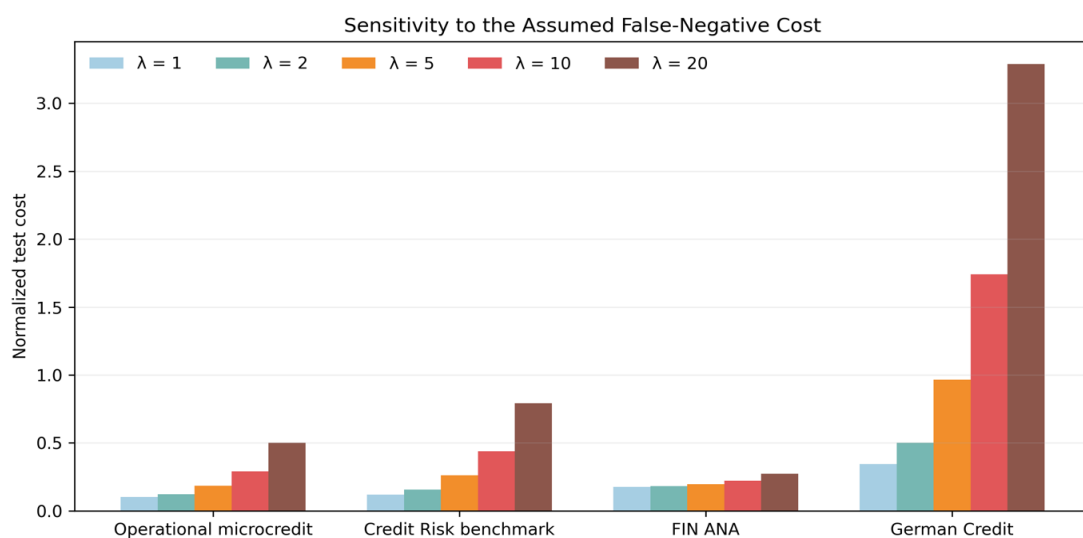


Figure 7. Normalized test cost for fixed primary predictions under alternative false-negative ratios

3.4. Model Interpretation

Figure 5 provides exploratory probability-scale diagnostics on the retrospectively held-out evaluation partitions. The Brier scores were 0.062 for operational microcredit, 0.075 for Credit Risk, 0.184 for FIN ANA, and 0.236 for German Credit. Among the four selected models, only Credit Risk produced a Brier score that was descriptively lower than the constant-score reference based on the observed test prevalence. The reliability bins for the operational microcredit and FIN ANA sources should nevertheless be interpreted cautiously because their adverse test observations are sparse. Moreover, the use of class-balanced fitting weights changes the learning objective, meaning that the resulting scores should not automatically be interpreted as calibrated probabilities of default. Independent calibration would therefore be required before these scores could be used as probability estimates in financial decision-making.

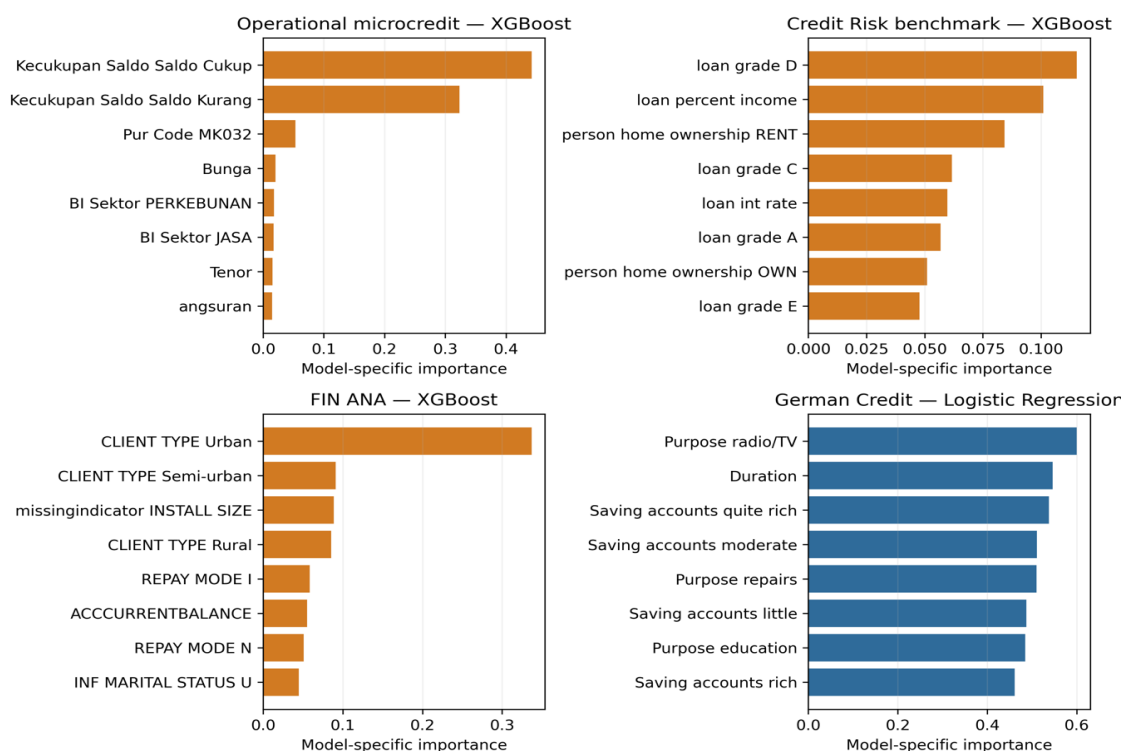


Figure 8. Predictor importance for validation-selected classifiers

The effect of class imbalance is also evident in the confusion matrices. As illustrated in Figure 6, Credit Risk provides substantially more observations across the four confusion-matrix cells, resulting in more stable evidence for the selected operating point. In contrast, the operational microcredit and FIN ANA test partitions contain only 11 and 24

adverse observations, respectively. Consequently, one additional correctly detected adverse case changes recall by approximately 9.09 percentage points for operational microcredit and 4.17 percentage points for FIN ANA. This sensitivity to individual observations explains why apparently small changes in the number of detected adverse cases can produce large changes in recall and precision, and further illustrates why overall accuracy alone is insufficient for evaluating these imbalanced classification problems.

The source-specific nature of the selected models is further reflected in the predictor importance patterns presented in Figure 8. For operational microcredit, the XGBoost model emphasized balance-adequacy categories, purpose/product indicators, interest rate, sector, and tenor. For Credit Risk, the most influential predictors included loan grade, loan-to-income share, housing status, and interest rate. In FIN ANA, XGBoost relied heavily on client type and missing installment information. Given the poor supplied-test ranking of FIN ANA, these importance patterns should be interpreted primarily as characteristics of the development data rather than as durable risk factors. For German Credit, Logistic Regression emphasized purpose, duration, and savings categories.

Taken together, these source-specific patterns indicate that predictor importance should not be interpreted as causal evidence or as a universal hierarchy of credit-risk factors. Changes in population composition, missingness patterns, category frequencies, and data-collection practices can alter the relative importance of predictors. Therefore, practical use of model interpretation in a credit decision-support system should be accompanied by population monitoring, independent calibration, and periodic re-evaluation of feature stability.

3.5. Implications and Limitations for Microcredit Risk Prediction

The comparative evaluation shows that model performance is source-dependent. Credit Risk provided the strongest internal benchmark, while the operational portfolio achieved high discrimination but limited adverse-case detection at the selected threshold. FIN ANA showed substantial degradation on the supplied evaluation partition. These findings indicate that no model is universally superior and that model selection and operating points should be interpreted within the source context [2].

For the operational portfolio, XGBoost achieved ROC-AUC 0.895 but detected only 4 of 11 adverse accounts, while 31 accounts were flagged for review. This corresponds to approximately 7.75 flagged accounts for each adverse account detected. Because temporal feature snapshots and future outcome windows were unavailable, the result represents retrospective adverse-status screening rather than prospective default prediction.

FIN ANA showed a different pattern, with adverse prevalence decreasing from 11.10% in training data to 0.56% in the evaluation partition. Although this shift can affect threshold behavior and realized cost, it alone does not explain the ROC-AUC of 0.461. The result indicates broader distribution differences that should be investigated before further threshold optimization. Its feature importance should therefore be interpreted as development-sample associations rather than stable risk factors [13]. Figure 9 summarizes the appropriate next research actions: prospective temporal validation for the operational portfolio, external or temporal validation for Credit Risk, data and label auditing for FIN ANA, and repeated validation for German Credit.

Operational microcredit	Credit Risk benchmark	FIN ANA	German Credit
High observed discrimination; low fixed-threshold recall	Internal benchmark result	Poor supplied-test performance	Exploratory only
ROC-AUC 0.895 Recall 0.364	ROC-AUC 0.940 Recall 0.839	ROC-AUC 0.461 Recall 0.083	ROC-AUC 0.620 Recall 0.483
Next research step: prospective temporal cohort and review-capacity analysis	Next research step: retain as benchmark; do not infer bank-level portability	Next research step: audit sampling, label process, and feature coding	Next research step: repeat nested validation; avoid definitive model-ranking claims

Figure 9. Evidence-to-action synthesis by source

3.6. Threats to Validity and Methodological Limitations

Several limitations should be considered. First, the evaluation partitions were retrospectively held out because their labels had been inspected by earlier exploratory scripts. Although these labels were not used in the final model or threshold selection, this condition may affect internal and statistical conclusion validity. Second, the operational dataset represents a single institutional portfolio without a temporal feature snapshot and future outcome window. Therefore, prospective predictive validity has not been established. In addition, the operational and FIN ANA evaluation partitions contain

only 11 and 24 adverse observations, respectively, limiting the precision of minority-class estimates. The single split also does not establish model-ranking stability under repeated partitions.

Third, the balanced fitting weights and five-to-one cost ratio are methodological scenarios rather than institution-specific loss estimates, and the resulting scores should not be interpreted as calibrated default probabilities. Fair-lending performance was also not established because reliable subgroup analysis was not possible with the available data. Finally, the public datasets have different target definitions, observation structures, and prevalence, limiting direct portability across sources. Operational confidentiality also limits full reproduction. Therefore, the models should be regarded as research benchmarks rather than autonomous credit decision systems. Future implementation requires temporal validation, independent calibration, institution-specific cost estimation, subgroup evaluation, and formal model governance.

4. CONCLUSION

This study compared Logistic Regression, Random Forest, XGBoost, and CatBoost under a cost-sensitive protocol across one operational microcredit portfolio and three supplementary credit datasets. No model was consistently superior across all sources. XGBoost was selected for the operational, Credit Risk, and FIN ANA sources, while Logistic Regression was selected for German Credit. Credit Risk provided the strongest internal benchmark, with XGBoost achieving ROC-AUC 0.940 and recall 0.839, whereas the operational portfolio achieved ROC-AUC 0.895 but detected only 4 of 11 adverse accounts at the selected operating point. For the operational portfolio, these results represent retrospective adverse-status screening rather than demonstrated prospective default prediction. FIN ANA showed substantial degradation under broader distribution shift, with ROC-AUC 0.461 and recall 0.083. Overall, the findings confirm that ROC-AUC alone is insufficient for cost-sensitive adverse-risk screening and that model selection, thresholds, and error costs should be evaluated within each source context. The reported evaluation remains retrospective and should not be interpreted as prospective predictive validation. Future research should focus on untouched temporal cohorts, independent calibration, institution-specific cost estimation, lawful subgroup evaluation, repeated validation, and threshold reassessment when portfolio conditions materially change.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Science, and Technology through the Research and Community Service Program, Fiscal Year 2026 (Grant Numbers 197 /LL2/AL.04.03/PL-BARU/2026 and 006/SPK/DRPM/Univ-BD/IV/2026). The authors also gratefully acknowledge the support provided by Universitas Bina Darma.

REFERENCES

- [1] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *Eur. J. Oper. Res.*, vol. 297, no. 3, pp. 1178–1192, 2022, doi: 10.1016/j.ejor.2021.06.053.
- [2] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: Adaption of different imbalance ratios," *Expert Syst. Appl.*, vol. 98, pp. 105–117, 2018, doi: 10.1016/j.eswa.2018.01.012.
- [3] J. Martin, S. Taheri, and M. Abdollahian, "Optimizing ensemble learning to reduce misclassification costs in credit risk scorecards," *Mathematics*, vol. 12, no. 6, pp. 855, 2024, doi: 10.3390/math12060855.
- [4] G. Petrides and W. Verbeke, "Cost-sensitive ensemble learning: A unifying framework," *Data Min. Knowl. Discov.*, vol. 36, pp. 1–28, 2022, doi: 10.1007/s10618-021-00790-4.
- [5] Z. Li, J. Zhang, X. Yao, and G. Kou, "How to identify early defaults in online lending: A cost-sensitive multi-layer learning framework," *Knowledge-Based Syst.*, vol. 221, pp. 106963, 2021, doi: 10.1016/j.knosys.2021.106963.
- [6] Y. Kang, L. Chen, N. Jia, W. Wei, J. Deng, and H. Qian, "A CWGAN-GP-based multi-task learning model for consumer credit scoring," *Expert Syst. Appl.*, vol. 206, pp. 117650, 2022, doi: 10.1016/j.eswa.2022.117650.
- [7] J. Xiao, Y. Wang, J. Chen, L. Xie, and J. Huang, "Impact of resampling methods and classification models on the imbalanced credit scoring problems," *Inf. Sci.*, vol. 569, pp. 508–526, 2021, doi: 10.1016/j.ins.2021.05.029.
- [8] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, pp. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

- [9] S. Han and H. Jung, "NATE: Non-parametric approach for explainable credit scoring on imbalanced class," *PLoS One*, vol. 19, no. 12, pp. e0316454, 2024, doi: 10.1371/journal.pone.0316454.
- [10] I. Aruleba and Y. Sun, "Effective credit risk prediction using ensemble classifiers with model explanation," *IEEE Access*, vol. 12, pp. 115015–115025, 2024, doi: 10.1109/ACCESS.2024.3445308.
- [11] Y. Xia, S. Jiang, L. Meng, and X. Ju, "XGBoost-B-GHM: An ensemble model with feature selection and GHM loss function optimization for credit scoring," *Systems*, vol. 12, no. 7, pp. 254, 2024, doi: 10.3390/systems12070254.
- [12] Z. Zhao, T. Cui, S. Ding, J. Li, and A. G. Bellotti, "Resampling techniques study on class imbalance problem in credit risk prediction," *Mathematics*, vol. 12, no. 5, pp. 701, 2024, doi: 10.3390/math12050701.
- [13] Y. Chen, R. Calabrese, and B. Martin-Barragan, "Interpretable machine learning for imbalanced credit scoring datasets," *Eur. J. Oper. Res.*, vol. 312, no. 1, pp. 357–372, 2024, doi: 10.1016/j.ejor.2023.06.036.
- [14] X. Dastile and T. Celik, "Making deep learning-based predictions for credit scoring explainable," *IEEE Access*, vol. 9, pp. 50426–50440, 2021, doi: 10.1109/ACCESS.2021.3068854.
- [15] H. Putra and R. Rumini, "Comparative study of Logistic Regression, Random Forest, and XGBoost for bank loan approval classification," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2822–2835, 2025, doi: 10.30871/jaic.v9i5.10862.
- [16] A. A. Saputra, B. N. Sari, and C. Rozikin, "Penerapan algoritma Extreme Gradient Boosting (XGBoost) untuk analisis risiko kredit," *J. Ilm. Wahana Pendidik*, vol. 10, no. 7, pp. 27–36, 2024, doi: 10.5281/zenodo.10960080.
- [17] L. Tse, "Credit Risk Dataset," *Kaggle*, Jul. 27, 2020.
- [18] M. Tahmid, "Credit Risk Dataset," *Kaggle*, Jul. 27, 2024.
- [19] H. Hofmann, "Statlog (German Credit Data)," *UCI Machine Learning Repository*, 1994.
- [20] K. R. Jammalamadaka and S. Itapu, "Responsible AI in automated credit scoring systems," *AI Ethics*, vol. 3, no. 2, pp. 485–495, 2023, doi: 10.1007/s43681-022-00175-3.