Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediction Time Information Governance for Explainable Machine Learning in Nonperforming Asset Risk Prioritization

Mithilesh Kumar Dubey
VIT-AP University
Syed Tahir Abbas Hasani
abbas.23bce8807@vitapstudent.ac.in

VIT-AP University

Research Article

Keywords: non-performing assets, credit risk, explainable machine learning, prediction-time governance, data leakage, risk prioritization, decision support, logistic regression

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11056341/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Prediction Time Information Governance for Explainable Machine Learning in Nonperforming Asset Risk Prioritization

Mithilesh Kumar Dubey
School of Computer Science and Engineering, VIT-AP University, Amaravati, India
mithileshkumar.dubey@vitap.ac.in

Syed Tahir Abbas Hasani*
School of Computer Science and Engineering, VIT-AP University, Amaravati, India
abbas.23bce8807@vitapstudent.ac.in

Abstract

Credit-risk models can exhibit strong retrospective discrimination while relying on variables that would not be available when monitoring or intervention decisions must actually be made. This paper presents the Adaptive Modular Risk Prioritization Framework (AMRPF), an explainable machine-learning decision-support methodology for non-performing asset (NPA) monitoring that treats prediction-time information availability as part of the predictive problem. AMRPF integrates feature governance, leakage-aware validation, interpretable probability estimation, portfolio ranking, case-level explanation, and cost-sensitive intervention analysis. Evaluation used an industry-derived dataset containing 131 observations, including 49 adverse and 82 non-adverse cases. A prediction-time audit reduced 39 candidate predictors to 21 strict pre-event variables. Logistic regression was retained as the primary interpretable model. Repeated stratified five-fold cross-validation with ten repetitions yielded mean ROC-AUC of 0.9294 and mean PR-AUC of 0.9143, with an empirical 95% ROC-AUC interval of 0.8389–0.9921. Restricting the feature space from full information to the strict pre-event configuration reduced mean ROC-AUC by only 0.0257, while a 1,000-permutation test produced an empirical p-value of 0.000999. In a separate 27-case holdout analysis, ROC-AUC was 0.9176 and PR-AUC was 0.9335; at threshold 0.40, precision and recall were both 0.90. Feature-removal, calibration, cost-sensitive, and chronological analyses provided complementary robustness checks. The results support AMRPF as a prediction-time-governed analytical prototype, while the small single-source dataset requires independent external and prospective validation before operational deployment.

Keywords: non-performing assets; credit risk; explainable machine learning; prediction-time governance; data leakage; risk prioritization; decision support; logistic regression

1 Introduction

Non-performing assets (NPAs) and other distressed-credit exposures create financial and operational pressure for lending institutions because deterioration can increase provisioning requirements, recovery costs, servicing workload, and loss exposure. In practice, risk monitoring is therefore not only a classification problem. Institutions must determine which accounts deserve attention first, whether the information used for scoring would genuinely have existed at the intended intervention point, and whether the resulting priorities can be explained to human reviewers.

*Corresponding author

Statistical credit-scoring methods have long been used to distinguish higher-risk and lower-risk borrowers [1]. Machine-learning research has broadened the range of candidate models and shown that predictive performance can depend strongly on the dataset, preprocessing choices, class balance, and experimental protocol [3]. Recent credit-risk studies have continued this direction through ensemble learning, AI-based scoring, and model comparison [11–13]. Explainability has simultaneously become increasingly important in financial decision support because high predictive performance alone does not establish that a model is auditable or operationally defensible [7, 8].

A central difficulty is that retrospective financial datasets often contain variables populated after deterioration, delinquency escalation, servicing action, legal action, or recovery activity. Such variables may be highly predictive of an adverse outcome while being unavailable at the time an early-warning decision would need to be made. Their inclusion can create prediction-time leakage and produce evaluation results that are optimistic relative to the intended use case [10]. Ordinary train-test separation does not by itself resolve this problem if the feature set already contains information that belongs to the future of the decision point.

A second difficulty is evaluation under limited data. A single favorable split can produce unstable estimates, particularly when the number of adverse cases is modest. Multiple forms of evidence are therefore useful: repeated resampling, uncertainty reporting, permutation testing, feature sensitivity, calibration assessment, explicit holdout analysis, and temporal separation. These procedures do not create new independent observations, but they can test whether a reported result is fragile to different analytical challenges.

A third difficulty is the gap between prediction and action. Monitoring and recovery teams typically need more than a probability. They need an ordered account queue, interpretable reasons for prioritization, explicit operating thresholds, and a way to alter intervention policy when missing an adverse account is more costly than reviewing an ultimately non-adverse account.

This paper addresses these issues through the Adaptive Modular Risk Prioritization Framework (AMRPF). AMRPF is a modular analytical architecture rather than an autonomous artificial-intelligence system. Its purpose is to organize six functions—prediction-time governance, leakage-aware validation, interpretable prediction, account prioritization, case-level explanation, and cost-sensitive intervention—around the information state that should exist when a risk decision is made.

The empirical study uses an industry-derived research extract supplied by a private-sector collections organization managing recovery activities across multiple financial-institution portfolios. The dataset contains 131 observations, including 49 adverse and 82 non-adverse cases. From 39 candidate predictors, a prediction-time audit retained 21 variables judged suitable for a strict pre-event information set. Because the sample is limited and originates through one source organization, the study does not claim population-level generalization or production readiness.

The study is organized around three research questions:

RQ1:
Does substantial predictive discrimination remain when the feature space is restricted to variables judged available before the adverse event?

RQ2:
Is the resulting discrimination internally robust across repeated resampling, randomized-label testing, feature-removal analysis, calibration diagnostics, and temporal separation?

RQ3:
Can governed predictive probabilities be translated into interpretable portfolio priorities and operational threshold policies under asymmetric error costs?

The principal contributions are:

- a prediction-time feature-governance stage that separates strict pre-event candidates from target-proximal, event-related, and identifier-like variables;
- a leakage-aware evaluation strategy in which the strict pre-event information set is treated as the principal scientific configuration rather than as a secondary sensitivity check;
- an interpretable decision-support workflow that transforms probabilities into continuous risk scores, portfolio ranks, case-level explanations, and intervention decisions;
- a chain of complementary robustness analyses—repeated cross-validation, information-set sensitivity, permutation testing, feature removal, calibration, holdout evaluation, and chronological testing—rather than reliance on one split or one metric; and
- explicit threshold and cost-sensitive analyses linking predictive performance with the operational consequences of false-negative and false-positive errors.

1.1 Methodological Contribution and Scope

The contribution of AMRPF is methodological rather than algorithmic. The framework does not claim a new classifier, explanation algorithm, ranking method, or cost function in isolation. Instead, it treats the prediction-time information state as a first-class part of model design and connects that governed information state to explainable portfolio prioritization.

AMRPF integrates six analytical functions:

1. **Prediction-time governance**, in which candidate predictors are classified according to expected availability relative to the adverse event;
2. **Leakage-aware validation**, in which the principal evaluation uses only the strict pre-event information set;
3. **Interpretable prediction**, in which adverse-risk probabilities are estimated using a transparent primary model;
4. **Account prioritization**, in which probabilities are transformed into scores, ranks, and operational groupings;
5. **Case-level explanation**, in which influential contributions to an individual prediction are surfaced for human review; and
6. **Cost-sensitive intervention**, in which operating thresholds are evaluated under alternative false-negative and false-positive cost assumptions.

These functions are supported by repeated cross-validation, permutation testing, information-set sensitivity analysis, feature-removal experiments, calibration diagnostics, holdout evaluation, and chronological testing. The integrated methodology can be summarized as

$$\begin{aligned} \text{AMRPF} = {} & \text{Prediction-Time Governance} \\ & + \text{Leakage-Aware Validation} \\ & + \text{Interpretable Prediction} \\ & + \text{Account Prioritization} \\ & + \text{Case-Level Explanation} \\ & + \text{Cost-Sensitive Intervention}. \end{aligned} \tag{1}$$

The integration matters because validity in one dimension does not guarantee validity in another. An interpretable model can still rely on post-event information; a leakage-aware probability model can still fail to support resource allocation; and a ranked portfolio can still be difficult to justify if reviewers cannot understand why particular accounts were prioritized. AMRPF therefore reframes the analytical objective from isolated classification toward governed, explainable, and intervention-oriented risk prioritization.

2 Related Work

2.1 Credit-Risk Prediction and Machine Learning

Statistical classification and scoring have historically formed the foundation of credit-risk prediction. Hand and Henley [1] reviewed classical credit-scoring methods and emphasized the importance of data quality, model specification, predictor characteristics, and validation design. Lessmann et al. [3] later benchmarked a broad range of credit-scoring classifiers and showed that apparent model superiority depends materially on the dataset and experimental protocol.

Recent studies continue to investigate whether more flexible machine-learning systems improve financial risk assessment. Hade and Elhia [11] compared multiple machine-learning approaches for mortgage default prediction on a real-world confidential dataset. Vakrani et al. [12] compared AI-driven credit-scoring models with traditional statistical techniques, while Soni and Jethava [13] focused on ensemble selection for imbalanced credit-risk data. These studies illustrate the continuing importance of model comparison, class imbalance, and predictive performance in credit-risk research.

AMRPF is complementary to this literature. Its primary question is not which classifier produces the largest retrospective metric, but whether useful predictive discrimination remains after the available information is constrained toward what should exist at the intended decision point.

2.2 Explainable Credit-Risk Modeling

Financial decision-support systems benefit from exposing the factors associated with a prediction rather than returning only a class label or probability. Lundberg and Lee [4] introduced SHAP as a general additive framework for explaining machine-learning predictions. Bussmann et al. [7] applied explainable machine learning to credit-risk management and demonstrated the value of feature-level attribution for borrower-level estimates. Chang et al. [8] similarly incorporated transparency into credit-worthiness prediction.

Recent AI-based credit-scoring work has also paired model performance with explainability [12]. AMRPF shares the objective of transparent financial prediction but uses explanation inside a broader workflow that also governs prediction-time information, ranks accounts, evaluates intervention thresholds, and tests sensitivity to potentially event-proximal variables.

2.3 Temporal Information and Prediction-Time Leakage

Credit risk is inherently temporal. A variable can be associated with an adverse outcome yet still be unsuitable for early warning if it is populated only after deterioration has begun. Bellotti and Crook [2] incorporated temporal information into credit scoring through survival analysis, while Bone-Winkel and Reichenbach [9] combined explainable machine-learning survival analysis with credit-risk assessment and risk stratification.

Information leakage is a related but distinct problem. Kaufman et al. [10] formalized leakage in data mining and described how information unavailable at prediction time can contaminate

model development and evaluation. In financial data, this concern is particularly important because repayment status, servicing activity, collections actions, and legal or recovery fields may be populated downstream of the event that a model is supposed to anticipate.

AMRPF therefore treats prediction-time feature governance as an explicit analytical stage rather than assuming that ordinary random train-test separation is sufficient. Candidate variables are categorized by their expected relationship to the prediction point, and the strict pre-event information set is used for the principal evaluation.

2.4 From Prediction to Operational Prioritization

Predictive accuracy and operational usefulness are related but different objectives. A classifier may estimate adverse risk, but a monitoring or recovery team still needs to decide which accounts should be reviewed first, how many accounts can be acted upon, and how operating thresholds should change when missed adverse accounts are substantially more costly than unnecessary reviews.

AMRPF extends conventional classification by translating model probabilities into continuous risk scores, portfolio ranks, case-level explanations, and intervention decisions. This creates an explicit bridge between predictive modelling and limited-resource portfolio management while retaining human review authority.

2.5 Research Gap and Methodological Positioning

Existing literature provides strong foundations for statistical scoring, machine-learning classification, explainability, temporal modelling, class-imbalance handling, and leakage prevention. The gap addressed here is therefore not the absence of these techniques individually. It concerns how they are organized around the complete prediction-time decision problem faced by an NPA-monitoring process.

Table 1: Methodological Positioning Relative to Representative Credit-Risk Literature

Study	Primary Emphasis	Relationship to AMRPF
Hand and Henley [1]	Statistical credit scoring	Establishes the statistical scoring foundation; AMRPF extends the task toward prediction-time governance and portfolio prioritization.
Lessmann et al. [3]	Classifier benchmarking	Demonstrates the importance of evaluation protocol; AMRPF additionally constrains the information state and connects prediction to intervention.
Bussmann et al. [7]	Explainable credit-risk modelling	Provides a foundation for interpretable borrower-level prediction; AMRPF embeds explanation in a governed prioritization workflow.
Chang et al. [8]	Explainable creditworthiness prediction	Reinforces transparent financial modelling; AMRPF additionally evaluates pre-event information restriction and cost-sensitive intervention.
Hade and Elhia [11]	Mortgage default prediction with multiple ML models	Focuses on model performance and class handling; AMRPF instead treats prediction-time information availability as a core design constraint.

Study	Primary Emphasis	Relationship to AMRPF
Vakrani et al. [12]	AI-driven versus traditional credit scoring	Motivates comparison between predictive power and interpretability; AMRPF prioritizes a transparent primary model within a governance-first workflow.
Soni and Jethava [13]	Ensemble selection for imbalanced credit risk	Advances algorithmic classification under imbalance; AMRPF focuses on information governance, explanation, ranking, and intervention policy.
Kaufman et al. [10]	Information-leakage detection and avoidance	Provides the leakage-control foundation operationalized in AMRPF through feature auditing and full-versus-strict information-set comparison.
AMRPF	Prediction-time-governed NPA prioritization	Integrates governance, leakage-aware validation, interpretable prediction, ranking, explanation, and cost-sensitive decision support.

A practically defensible workflow should answer three linked questions. First, what information would actually have been available while intervention was still possible? Second, does the apparent predictive signal survive when event-proximal variables are removed and the model is challenged through multiple robustness checks? Third, can the resulting probabilities be converted into interpretable portfolio priorities under explicit workload and error-cost trade-offs?

AMRPF combines these requirements within one analytical architecture. This positioning deliberately limits the novelty claim: the contribution lies in the prediction-time-centered integration and empirical stress testing of the workflow, not in relabeling established algorithms as new methods.

3 Problem Formulation

Let account i be described by the predictor vector

$$\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]. \tag{2}$$

where p represents predictors considered available at the defined prediction point. The target variable is

$$y_i \in \{0, 1\}. \tag{3}$$

where $y_i = 1$ represents an adverse case and $y_i = 0$ represents a non-adverse case. The objective is to estimate

$$P(y_i = 1 \mid \mathbf{x}_i). \tag{4}$$

and use the resulting probability to prioritize accounts for monitoring or intervention. For logistic regression,

$$\hat{p}_i = \frac{1}{1 + \exp[-(\beta_0 + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})]}. \tag{5}$$

The probability is converted into a continuous score

$$R_i = 100\hat{p}_i. \tag{6}$$

This score is subsequently used for ranking and tier assignment. The prioritization stage does not alter the underlying model probability.

4 AMRPF Framework

4.1 Overall Architecture

AMRPF consists of six specialized analytical modules:

1. Data Governance Module;
2. Prediction Module;
3. Risk Prioritization Module;
4. Explainability Module;
5. Decision-Support Module; and
6. Validation and Monitoring Module.

The primary analytical sequence is

$$\text{Data} \rightarrow \text{Governance} \rightarrow \text{Prediction} \rightarrow \text{Prioritization} \rightarrow \text{Explanation} \rightarrow \text{Decision Support}. \tag{7}$$

Validation and monitoring operate across the workflow rather than being restricted to the final stage.

4.2 Data Governance Module

The Data Governance Module evaluates target integrity, missingness, numerical validity, variable types, identifiers, and predictor availability. Its central responsibility is the prediction-time feature audit. Candidate variables are assessed according to whether they are reasonably expected to exist at the intended decision point or whether they may instead be target-proximal, event-related, or identifier-like.

4.3 Prediction Module

The Prediction Module estimates adverse-outcome probability using the governed feature set. Logistic regression is used as the primary model because interpretability is a core requirement of AMRPF. Alternative algorithms are explored for comparison, but maximum predictive complexity is not treated as the sole model-selection criterion.

4.4 Risk Prioritization Module

The Risk Prioritization Module maps predicted probabilities to continuous 0–100 scores and orders accounts from highest to lowest predicted risk. Operational groupings can then be layered on top of the continuous ranking according to institutional policy. The ranking itself does not alter the underlying model probability.

4.5 Explainability Module

For logistic regression, the contribution of transformed feature j to case i is expressed as

$$C_{ij} = z_{ij}\beta_j. \tag{8}$$

where z_{ij} is the transformed feature value and β_j is the corresponding fitted coefficient. Positive values increase the modeled log-odds of the adverse class, while negative values decrease them. The strongest positive contributions are surfaced as case-level "WHY RISKY?" explanations. These contributions should be interpreted as predictive associations rather than causal effects.

4.6 Decision-Support Module

For each account, the Decision-Support Module returns the estimated adverse probability, continuous risk score, portfolio priority rank, major predictive contributors, and an intervention category derived from the selected operating policy. These outputs are intended to assist human decision-making and do not represent autonomous lending or recovery decisions.

4.7 Validation and Monitoring Module

The Validation and Monitoring Module examines discrimination, probability error, calibration, statistical significance, information-set sensitivity, feature-removal stability, threshold behavior, cost sensitivity, holdout risk concentration, and temporal robustness. The objective is not to convert repeated analysis into artificial evidence of a larger sample, but to test whether the principal conclusions survive multiple distinct methodological challenges.

5 Data and Methodology

5.1 Industry-Derived Dataset

The study uses a research extract supplied by a private-sector collections organization managing recovery activities across several financial-institution portfolios.

The dataset contains 131 observations and 40 raw columns, consisting of 39 candidate predictors and one binary target variable.

Of the 131 observations, 49 correspond to adverse cases and 82 correspond to non-adverse cases, producing an adverse-case rate of approximately 37.4%.

The observations reflect heterogeneous real-world collection activity. However, because the research extract was supplied through one collections organization and institution-level provenance was not used for stratified evaluation, the dataset is treated as single-source data throughout this study.

No synthetic observations were introduced.

5.2 Target Definition

The binary adverse target was defined using observed account outcomes:

$$Y = \begin{cases} 1, & \text{REPOSSESSED or WRITEOFF,} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Variables used directly to construct the target were removed from the predictor space prior to model fitting. The target should be interpreted as an adverse recovery outcome within

an NPA/collections setting; it is not intended to redefine the regulatory criteria by which an exposure is formally classified as an NPA.

5.3 Prediction-Time Feature Audit

The 39 candidate predictors were categorized according to their relationship with the prediction point used in the study.

The audit identified 21 pre-event candidate predictors, 16 variables judged target-proximal or event-related, and two identifier-like predictors.

The resulting 21-variable configuration is referred to as the strict pre-event information set. These variables were retained because their operational roles suggested that they could reasonably exist before the defined adverse event.

Variables that appeared to arise downstream of delinquency escalation, deterioration, recovery activity, or target construction were excluded.

The designation pre-event is an analytical classification based on the available dataset and operational interpretation. It does not imply that production timestamps were independently verified for every variable.

A real deployment would therefore require formal field-level data lineage and timestamp verification to confirm that each retained predictor is populated before operational scoring occurs.

The strict information set is intended to provide a more conservative evaluation than unrestricted full-information modeling and to quantify how strongly predictive performance depends on potentially event-proximal variables.

5.4 Preprocessing

Numerical and categorical predictors were processed through separate preprocessing paths.

Missing numerical values were imputed where required. Categorical variables were imputed and one-hot encoded.

All preprocessing transformations were fitted exclusively on the corresponding training partition inside each validation procedure. This prevents statistics or categorical information from the validation observations from entering model preparation.

5.5 Primary Predictive Model

Logistic regression was selected as the primary predictive model because AMRPF emphasizes transparency and case-level explanation.

The resulting probability is interpreted as a conditional predictive risk estimate based on the available information rather than as a causal probability of default.

5.6 Evaluation Metrics

Model evaluation uses ROC-AUC, PR-AUC, balanced accuracy, precision, recall, F1 score, and Brier score.

ROC-AUC evaluates ranking discrimination across classification thresholds [6], while PR-AUC provides complementary information about the precision-recall relationship [5]. Probability behavior is assessed separately using Brier score and calibration diagnostics.

6 Experimental Design

The experiments were organized to answer the three research questions. RQ1 is addressed primarily through repeated stratified cross-validation and the full-versus-strict information-set comparison. RQ2 is assessed through permutation testing, feature-removal sensitivity, calibration analysis, and exploratory chronological testing. RQ3 is evaluated through explicit threshold analysis, asymmetric error-cost scenarios, and the additional 27-case holdout analysis.

6.1 Repeated Stratified Cross-Validation

The strict pre-event model was evaluated using repeated stratified five-fold cross-validation with ten repetitions.

This procedure produced 50 fold-level evaluations.

Stratification preserved the relative proportions of adverse and non-adverse cases in each fold.

6.2 Permutation Significance Test

A permutation experiment was used to determine whether the observed predictive discrimination could plausibly occur under randomly assigned target labels.

A total of 1,000 target permutations were evaluated. The empirical p-value was calculated as

$$p_{\text{emp}} = \frac{1 + b}{1 + B}. \tag{10}$$

where B = 1000 and b denotes the number of permutation statistics at least as large as the observed statistic.

6.3 Information-Set Sensitivity

The full-information configuration was compared with the strict pre-event configuration.

The purpose of this experiment was to quantify the reduction in discrimination after variables judged potentially target-proximal were removed.

6.4 Feature-Removal Sensitivity

Selected financial predictors were removed individually or in groups from the strict feature set.

This experiment examined whether predictive performance depended excessively on any one variable or predictor family.

6.5 Probability Calibration

Probability behavior was evaluated using several internal procedures.

Repeated cross-validation produced an average Brier score across validation folds. A separate strict-model calibration analysis reported Brier score, calibration intercept, calibration slope, and expected calibration error.

An additional out-of-fold calibration procedure produced a separate probability-error diagnostic.

Because these procedures are not numerically identical, their results are reported separately.

6.6 Threshold Analysis

Several probability thresholds were evaluated to measure the relationship among intervention volume, precision, recall, F1 score, and adverse-case capture.

6.7 Cost-Sensitive Analysis

Operational scenarios were evaluated in which a false negative was assigned a cost five, ten, or twenty times greater than a false positive.

This analysis was used to determine how the recommended intervention threshold changes under asymmetric error costs.

6.8 Held-Out Decision-Support Evaluation

An additional 27-case holdout partition was used to examine predictive performance, confusion-matrix behavior, portfolio ranking, operational grouping, and adverse-case concentration. Because all observations originate from the same 131-case source dataset and are reused in other resampling analyses, this partition is an internal holdout analysis rather than independent external validation.

6.9 Temporal Validation

A chronological 80/20 experiment was conducted using available origination dates.

The earlier 104 observations were used for model training, while the later 27 observations formed the test partition.

Because the chronological test set is small, the experiment is treated as exploratory.

7 Results

7.1 Primary Repeated-Cross-Validation Performance

Table 2: Strict Pre-Event Repeated-CV Performance

Metric	Result
Mean ROC-AUC	0.9294
ROC-AUC SD	0.0387
ROC-AUC empirical 95% interval	0.8389–0.9921
Mean PR-AUC	0.9143
Mean Brier Score	0.1043
Validation folds	50

The repeated-validation results indicate strong discrimination within the evaluated dataset while also showing meaningful variability across resamples.

Because the dataset is limited, the empirical interval is reported in addition to the average value rather than presenting only a single point estimate.

Table 3: Information-Set Sensitivity

Information Set	ROC-AUC	PR-AUC
Full information	0.9551	0.9597
Strict pre-event	0.9294	0.9143

7.2 Information-Set Sensitivity

The full-information configuration achieved mean ROC-AUC of 0.9551 and mean PR-AUC of 0.9597.

The strict pre-event configuration achieved mean ROC-AUC of 0.9294 and mean PR-AUC of 0.9143.

The difference in ROC-AUC was

$$\Delta\text{AUC} = 0.9551 - 0.9294 = 0.0257. \tag{11}$$

Therefore, excluding variables judged potentially target-proximal caused a relatively modest reduction in discrimination.

This result does not establish that the broader dataset is completely free of leakage. It instead shows that strong predictive signal remains after moving to the more conservative information set.

7.3 Permutation Significance

Table 4: Permutation Significance Test

Statistic	Value
Observed ROC-AUC	0.934690
Permutation Mean	0.503159
Permutation SD	0.070781
Permutations	1000
Empirical p-value	0.000999

The observed strict-model ROC-AUC in the permutation experiment was 0.934690. Across 1,000 randomized target assignments, the mean permutation ROC-AUC was 0.503159 with a standard deviation of 0.070781.

The resulting empirical p-value indicates that the observed discrimination is substantially stronger than that produced under the randomized-label null condition.

The value 0.934690 belongs specifically to the permutation experiment and should not be confused with the repeated-CV mean ROC-AUC of 0.9294.

7.4 Feature-Removal Sensitivity

The lowest ROC-AUC among the evaluated removal configurations was 0.9196 after excluding tenure-related variables.

The comparatively small performance changes indicate that the model's discrimination is not attributable exclusively to any single tested financial variable.

Table 5: Feature-Removal Sensitivity

Configuration	ROC-AUC	PR-AUC
All strict features	0.9294	0.9143
Without loan amount	0.9291	0.9145
Without disbursed amount	0.9299	0.9152
Without market value	0.9261	0.9112
Without ROI	0.9256	0.8978
Without tenure variables	0.9196	0.8953

7.5 Strict-Model Operating and Calibration Diagnostics

A separate strict-model validation procedure produced ROC-AUC of 0.9258, PR-AUC of 0.9037, balanced accuracy of 0.8432, precision of 0.7500, recall of 0.8571, F1 score of 0.8000, Brier score of 0.1069, calibration intercept of 0.0470, calibration slope of 1.2140, and expected calibration error of 0.0507.

An additional out-of-fold calibration procedure produced a Brier score of 0.1127 and a mean absolute calibration error of approximately 0.0923.

These values arise from separate internal procedures and are therefore reported independently from the repeated-CV mean Brier score of 0.1043.

7.6 Threshold Sensitivity

Table 6: Intervention Trade-Off

Threshold	Cases	Precision	Recall	F1
0.20	71	0.634	0.918	0.750
0.30	61	0.705	0.878	0.782
0.40	56	0.750	0.857	0.800
0.50	46	0.804	0.755	0.779
0.60	43	0.837	0.735	0.783
0.70	36	0.889	0.653	0.753
0.80	23	1.000	0.469	0.639

At a threshold of 0.40, 56 of the 131 observations were selected for intervention, corresponding to approximately 42.7% of the evaluated dataset.

Precision was 0.750 and recall was 0.857 at this operating point.

Lower thresholds increased adverse-case capture but also expanded the intervention workload, while higher thresholds improved precision at the cost of additional missed adverse cases.

7.7 Exploratory Broader-Information Model Comparison

This broader-information comparison was performed before the strict pre-event configuration was established.

The experiment is included only to document algorithmic robustness and is not used as the primary estimate of AMRPF's prospective performance because it relies on a broader information configuration.

Table 7: Exploratory Broader-Information Model Comparison

Model	CV ROC-AUC	Test ROC-AUC
Logistic Regression	0.9670	0.9176
Random Forest	0.9471	0.9235
Extra Trees	0.9586	0.9588
Hist. Gradient Boosting	0.9526	0.9588

The ROC-AUC values from this experiment should therefore not be directly compared with the strict-model mean ROC-AUC of 0.9294.

Logistic regression was retained as the primary AMRPF model because interpretability is required for the case-level explanation layer. The strict pre-event configuration was retained as the principal scientific evaluation because it is more defensible with respect to information availability.

8 Operational Prioritization and Explainability

8.1 Held-Out Predictive Performance

The additional holdout evaluation contained 27 observations, including 17 non-adverse and 10 adverse cases. It is reported as a distinct partitioning analysis, not as external validation.

At the selected threshold of 0.40, logistic regression achieved an accuracy of 0.9259, balanced accuracy of 0.9206, precision of 0.9000, recall of 0.9000, F1 score of 0.9000, ROC-AUC of 0.9176, and PR-AUC of 0.9335.

The corresponding confusion matrix contained 16 true negatives, one false positive, one false negative, and nine true positives.

8.2 Descriptive Risk-Tier Prioritization

Table 8: Held-Out Risk-Tier Evaluation

Risk Tier	Cases	Mean Prob.	Adverse Rate
Critical	8	0.9316	1.0000
High	2	0.4213	0.5000
Moderate	2	0.3490	0.0000
Low	15	0.0533	0.0667

The Critical and High tiers contained 10 of the 27 held-out accounts, corresponding to 37.0% of the test portfolio.

These two tiers captured nine of the ten observed adverse cases, producing a 90% high-risk capture rate.

The Critical tier had a mean predicted probability of 0.9316, whereas the Low tier had a mean probability of 0.0533.

Because only 27 observations are included, these tier-level adverse rates should be interpreted as descriptive evidence of risk concentration rather than stable population estimates. The tier labels are operational groupings rather than universal model parameters; the reproducible intervention analysis is therefore anchored primarily to the explicit probability thresholds reported

separately.

8.3 Case-Level "WHY RISKY?" Explanations

For each prioritized account, AMRPF identifies the strongest positive model contributions associated with predicted adverse risk.

Within the strict pre-event model, prominent predictive associations included variables related to money required for normalization, EMI amount, balance tenure, market value, selected geographic attributes, product-related characteristics, and selected origination and branch variables.

The explanation mechanism therefore supplements a simple statement such as "this account is high risk" with information about which observed characteristics contributed most strongly to the model score.

The explanation layer does not alter the underlying probability or prediction.

9 Cost-Sensitive Decision Analysis

The cost of failing to detect an adverse account may be substantially greater than the cost of reviewing an account that later proves non-adverse.

AMRPF therefore evaluates operating thresholds under several false-negative cost assumptions.

Table 9: Cost-Sensitive Intervention Analysis

Scenario	Threshold	FN	FP	Capture
FN cost = 5× FP	0.09	1	39	97.96%
FN cost = 10× FP	0.06	0	48	100.00%
FN cost = 20× FP	0.06	0	48	100.00%

Under the five-to-one false-negative cost setting, the selected threshold captured 97.96% of observed adverse cases.

Under ten-to-one and twenty-to-one cost ratios, all observed adverse cases were captured, although the number of false-positive interventions increased substantially.

These results show that intervention policy need not rely on one fixed threshold. The operating point can instead be adapted according to the institutional consequences of missed adverse accounts.

10 Temporal Robustness

A chronological 80/20 evaluation was conducted using the available origination dates.

The model was trained on 104 earlier observations and tested on 27 later observations.

The chronological experiment achieved ROC-AUC of 1.0000, PR-AUC of 1.0000, balanced accuracy of 0.9545, F1 score of 0.9524, recall of 0.9091, and precision of 1.0000. These unusually strong values require careful interpretation because the temporal test partition contains only 27 observations. With such a small test set, a limited number of cases can materially change the evaluation metrics.

The experiment therefore demonstrates that temporal separation is feasible within the dataset, but it does not establish robust out-of-time generalization.

11 Consolidated Validation Evidence

Because the dataset is limited, no single performance metric is sufficient to establish robustness. AMRPF was therefore evaluated through several complementary procedures, each addressing a different potential failure mode.

Repeated cross-validation measures sensitivity to resampling. Information-set restriction tests dependence on event-proximal predictors. Permutation testing compares the observed discrimination with a randomized-label baseline. Feature-removal analysis evaluates dependence on individual predictors. Calibration analysis examines probability behavior instead of ranking alone. Threshold and cost-sensitive analyses study operational trade-offs, while held-out and chronological evaluations provide additional evidence under separate partitioning strategies.

Importantly, all of these experiments reuse the same underlying set of 131 observations. They therefore strengthen the assessment of internal robustness but do not increase the amount of independent data.

The combined value of these procedures lies in testing whether the principal result survives several different methodological challenges rather than treating repeated analysis as additional observations.

External validity remains an empirical question that requires genuinely independent data.

12 Discussion

12.1 RQ1: Does Predictive Signal Survive Prediction-Time Restriction?

The central empirical result is that useful discrimination remains after the problem is reformulated around prediction-time information governance. The strict pre-event configuration achieved mean repeated-CV ROC-AUC of 0.9294 and mean PR-AUC of 0.9143, compared with ROC-AUC of 0.9551 for the broader information configuration. The resulting ROC-AUC reduction of 0.0257 indicates that the model does not depend entirely on variables judged potentially event-proximal.

This comparison is more important to the methodological argument than maximizing retrospective performance with every available field. AMRPF intentionally accepts a modest loss in discrimination in exchange for a more defensible approximation of the information state that could exist while intervention is still meaningful.

The result should not be interpreted as proof that the strict set is completely leakage-free. The pre-event designation was based on the available data documentation and operational interpretation rather than a field-by-field production timestamp audit. The experiment instead quantifies how much predictive signal remains after applying a conservative information restriction.

12.2 RQ2: How Robust Is the Result Internally?

Several analyses challenge alternative explanations for the observed discrimination. The 1,000-permutation experiment produced an empirical p-value of 0.000999, indicating that the observed ranking performance is not compatible with the randomized-label baseline used in that test. Feature-removal experiments produced ROC-AUC values no lower than 0.9196 among the evaluated configurations, suggesting that the result is not attributable exclusively to one tested financial predictor or predictor family. Calibration diagnostics provide a probability-oriented complement to ranking metrics, while the chronological experiment tests whether temporal separation is feasible within the available data.

These procedures provide convergent evidence of internal robustness, but their evidential scope must remain clear. They repeatedly interrogate the same underlying set of 131 observations

and therefore do not substitute for external validation. The unusually strong chronological ROC-AUC and PR-AUC of 1.0000 are particularly sensitive to the 27-case test size and should be treated as exploratory.

12.3 RQ3: Can the Predictions Support Operational Prioritization?

AMRPF converts estimated probabilities into a continuous 0–100 score and an ordered portfolio queue. At the explicit operating threshold of 0.40, the additional 27-case holdout analysis achieved precision of 0.90 and recall of 0.90, with one false positive and one false negative. The threshold-sensitivity analysis further shows how intervention workload changes as the operating point moves from 0.20 to 0.80.

The cost-sensitive analysis makes this trade-off explicit. When a false negative was treated as five times as costly as a false positive, the selected threshold was 0.09 and 97.96% of observed adverse cases were captured. Under ten-to-one and twenty-to-one cost ratios, the selected threshold decreased to 0.06 and all observed adverse cases were captured, at the cost of substantially more false-positive interventions. The value of this analysis is not the claim that these particular ratios are universally correct, but that the operating policy can be made explicit and adapted to institutional consequences.

The case-level explanation layer complements the ranking by identifying the strongest positive logistic-regression contributions associated with each score. This provides a transparent reason trace without implying that the identified variables are causal determinants of deterioration.

12.4 Relation to Recent Credit-Risk AI Research

Recent credit-risk research has emphasized stronger classifiers, ensemble selection, class-imbalance handling, and AI-versus-statistical model comparisons [11–13]. AMRPF addresses a different but complementary problem: how to structure a risk-prioritization workflow when the validity of the information available at prediction time is itself uncertain. This distinction also explains the choice of logistic regression as the primary model.

Exploratory ensemble models achieved competitive or higher discrimination in the broader-information comparison, but the principal AMRPF analysis prioritizes transparent case-level contributions and a conservative information state over maximum model complexity. The framework does not argue that logistic regression is universally preferable; it argues that model choice should be aligned with the requirements of the decision-support workflow.

12.5 Practical Interpretation

AMRPF should be understood as a human-in-the-loop analytical prototype. It can support the ordering and review of accounts, but it should not autonomously determine credit eligibility, legal action, repossession, or recovery treatment. Before operational deployment, an institution would need independent validation, formal field-level data lineage, monitoring for distribution shift, external calibration, subgroup and fairness assessment where relevant, governance of threshold changes, and prospective evaluation of whether AMRPF-guided interventions actually improve recovery outcomes.

13 Limitations

The first limitation is sample size. Only 131 observations were available, including 49 adverse cases. Repeated cross-validation, permutation testing, feature-removal analysis, and other

robustness procedures reduce dependence on one favorable split but cannot substitute for additional independent observations.

Second, although the source organization manages activities across several financial-institution portfolios, the research dataset was obtained through one private-sector collections organization. The study does not perform institution-stratified validation or establish independence across originating institutions. Differences in borrower populations, lending policies, product types, geographic distributions, servicing processes, recovery practices, and data-generation mechanisms may therefore affect performance elsewhere.

Third, the prediction-time classification of variables depends on the operational definition of the scoring point and available dataset documentation. The strict pre-event set is an analytical classification rather than the result of a complete field-level production timestamp audit. A production implementation would require formal data-lineage and timestamp verification for every retained predictor.

Fourth, the additional 27-case holdout and the chronological 27-case test are small internal evaluations. They originate from the same underlying source dataset and should not be interpreted as independent external validation. The chronological ROC-AUC and PR-AUC of 1.0000 are especially vulnerable to small-sample variation.

Fifth, logistic-regression coefficients and case-level feature contributions describe predictive associations, not causal effects. The framework therefore cannot establish that changing an influential variable would change the adverse outcome.

Sixth, internal calibration diagnostics do not establish external probability calibration. Model probabilities should not be interpreted as population-level default probabilities without validation on larger independent samples from the intended deployment environment.

Seventh, the intervention policies are retrospective analytical scenarios. No prospective controlled study has yet tested whether acting on AMRPF priorities improves recovery outcomes or produces unintended operational consequences.

Eighth, the current dataset is not sufficient for reliable subgroup fairness auditing. No claim is therefore made that the resulting priorities are fair across demographic or legally protected groups. Any deployment in a consequential financial setting should include appropriate fairness, bias, and governance assessments based on variables that can be evaluated lawfully and responsibly.

Finally, the literature comparison is representative rather than systematic and is used to position the methodological contribution, not to claim an exhaustive review of credit-risk modelling research.

14 Conclusion

This study presented AMRPF, a prediction-time-governed framework for explainable non-performing asset risk prioritization and decision support. Its central contribution is not a new classifier. It is a methodology that treats information availability at the intended decision point as part of model design and then links the governed prediction process to interpretation, ranking, and intervention policy.

Using an industry-derived dataset of 131 observations, a feature audit reduced 39 candidate predictors to a strict pre-event set of 21 variables. The primary logistic-regression model achieved mean repeated-CV ROC-AUC of 0.9294 and mean PR-AUC of 0.9143. Restricting the model from the full information set to the strict pre-event configuration reduced mean ROC-AUC by only 0.0257. A 1,000-permutation test produced an empirical p-value of 0.000999, and feature-removal experiments produced ROC-AUC values no lower than 0.9196 among the evaluated configurations.

Operational analyses showed how governed probabilities can be translated into explicit

intervention choices. In the additional 27-case holdout analysis, ROC-AUC was 0.9176 and PR-AUC was 0.9335; at threshold 0.40, precision and recall were both 0.90. Cost-sensitive scenarios demonstrated how lower thresholds increase adverse-case capture when false negatives are treated as more consequential than false positives.

Taken together, the experiments provide encouraging internal evidence that prediction-time governance can preserve useful predictive signal while supporting transparent portfolio prioritization. They do not establish generalization to another institution, portfolio, region, or future period. AMRPF should therefore be treated as a validated analytical prototype requiring larger independent datasets, explicit field-level temporal provenance, external calibration, fairness assessment, and prospective intervention studies before operational deployment.

Acknowledgements

The authors acknowledge the industry source that provided the research extract used in this study and the academic and institutional support received during the development and evaluation of the AMRPF framework.

Funding

The authors declare that no funds, grants, or other external financial support were received for this study.

Data Availability

The datasets analysed during the current study are not publicly available because the account-level research extract was supplied by a private-sector collections organization and is subject to confidentiality and proprietary-data restrictions. Requests for access may be directed to the corresponding author on reasonable request. Any data sharing is subject to prior written authorization from the data provider and applicable confidentiality requirements and therefore cannot be guaranteed.

Author Contributions

Syed Tahir Abbas Hasani contributed to conceptualization, methodology, software implementation, analysis, validation, and preparation of the original manuscript. Mithilesh Kumar Dubey contributed to supervision, methodological review, and manuscript review and editing. Both authors reviewed and approved the final manuscript.

Ethics Approval

Not Applicable. This study involved retrospective analysis of an industry-supplied operational research extract. No human participants were recruited, surveyed, contacted, or involved in a questionnaire, feedback process, or intervention. The manuscript reports only aggregate analytical results and does not disclose direct personal identifiers.

Consent to Participate

Not Applicable.

Consent for Publication

Not Applicable.

Competing Interests

The authors declare no competing financial or non-financial interests related to this work.

AI Use Declaration

OpenAI ChatGPT (GPT-5.6 Sol, accessed September 2026) was used to assist with language editing, manuscript organization, and journal-oriented restructuring of author-generated text. It was not used to generate or alter the study data, perform the reported statistical analyses, or create empirical results. The authors reviewed and verified the final manuscript and take full responsibility for its content.

References

[1] 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. https://doi.org/10.1111/j.1467-985X.1997.00078.x

[2] T. Bellotti and J. Crook, "Credit scoring with macroeconomic variables using survival analysis," Journal of the Operational Research Society, vol. 60, no. 12, pp. 1699–1707, 2009. https://doi.org/10.1057/jors.2008.130

[3] 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. https://doi.org/10.1016/j.ejor.2015.05.030

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[5] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 233–240. https://doi.org/10.1145/1143844.1143874

[6] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, 1997. https://doi.org/10.1016/S0031-3203(96)00142-2

[7] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," Computational Economics, vol. 57, pp. 203–216, 2021. https://doi.org/10.1007/s10614-020-10042-0

[8] V. Chang, Q. A. Xu, S. H. Akinloye, V. Benson, K. Hall, et al., "Prediction of bank credit worthiness through credit risk analysis: An explainable machine learning study," Annals of Operations Research, vol. 354, pp. 247–271, 2025. https://doi.org/10.1007/s10479-024-06134-x

[9] G. F. Bone-Winkel and F. Reichenbach, "Improving credit risk assessment in P2P lending with explainable machine learning survival analysis," Digital Finance, vol. 6, pp. 501–542, 2024. https://doi.org/10.1007/s42521-024-00114-3

[10] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 556–563. https://doi.org/10.1145/2020408.2020496

[11] A. Hade and M. Elhia, "Predicting mortgage credit defaults in Morocco using machine learning approaches," Discover Artificial Intelligence, vol. 5, article 113, 2025. https://doi.org/10.1007/s44163-025-00303-y

[12] D. S. Vakrani, P. S. Padhye, S. K. Gupta, J. K. Roy, V. S. Nerlekar, and N. Parashar, "Evaluating AI-driven credit scoring models versus traditional statistical techniques," Discover Artificial Intelligence, vol. 6, article 72, 2026. https://doi.org/10.1007/s44163-025-00772-1

[13] U. Soni and G. Jethava, "Enhancing credit risk prediction via voting classifier and meta-dynamic ensemble selection on imbalanced data," Discover Artificial Intelligence, vol. 6, article 138, 2026. https://doi.org/10.1007/s44163-026-00849-5