Robust Credit-Stress Prediction Under Borrower Behavioural Drift: An Explainable Framework for MSME Lending

D Ajay Kumar

Independent Researcher, Khammam, Telangana, India,507209.
ajayupgc@gmail.com

Abstract—MSME borrowers operate in changing markets where payment cycles, cash conversion, input costs, and account behaviour can evolve after loan origination. A risk model trained on historical patterns may therefore mis-rank emerging stress even when its input fields remain unchanged. This paper proposes an explainable robust credit-stress framework that distinguishes population shift from deterioration in predictive performance, adapts only after matured-label validation, and monitors whether explanations remain stable across time. The design combines point-in-time behavioral features, a regularized probability model, drift diagnosis using complementary feature and performance signals, a recency-weighted candidate update, calibration checks, explanation stability tests, and controlled deployment with rollback. A synthetic temporal benchmark containing abrupt and gradual shifts compares a frozen logistic model, a rolling update, and a robust rolling model with bounded feature values and stronger regularization. Across six held-out periods, the robust model achieved AUROC 0.828, average precision 0.667, Brier score 0.181, and event recall 0.374 at a fixed 15% alert capacity; the frozen model achieved 0.776, 0.590, 0.223, and 0.336, respectively. These outputs illustrate a reproducible evaluation under known simulated shifts and are not evidence from an actual lender or MSME portfolio. The paper contributes a practical validation and governance design for future institution-specific testing.

Index Terms—MSME lending, credit stress, borrower behaviour, concept drift, robust prediction, explainable AI, temporal validation.

I. INTRODUCTION

Micro, small, and medium enterprises (MSMEs) often exhibit uneven cash flows, concentrated customers, seasonal revenues, and delayed receivables. A facility that is current today may show early deterioration through a sequence of signals: lower operating inflows, higher drawing power utilization, missed instalments, larger payment gaps, and volatile balances. Credit monitoring models seek to detect such transitions early enough for a human officer to verify the account and consider an appropriate response. Credit scoring literature demonstrates the value of statistical and machine-learning methods, but performance depends on the population, outcome, and validation design [1], [2].

Borrower behaviour is not stationary. A change in economic conditions can alter feature distributions (covariate shift), the prevalence of stress (prior shift), or the relationship between features and the event (concept drift). These mechanisms are often grouped together under "drift," although each has a different operational implication [3]–[5]. A shift in average utilization may simply reflect a changed portfolio mix; it does not necessarily mean that the model is less accurate. Conversely, the stress mechanism may evolve while aggregate feature histograms remain stable. A robust monitoring system must examine both incoming data and delayed predictive outcomes.

Adaptation alone is not sufficient. Frequent retraining can chase short-lived noise, overfit recent labels, or produce score changes that reviewers cannot explain. In lending, model outputs are part of a governed risk process, so performance, calibration, interpretability, and auditability must be considered together [6], [7]. Explanation methods provide model attribution rather than causal proof, and their background distributions or feature interactions can change as borrower populations evolve [8]–[10].

This study proposes a robust credit-stress prediction framework for MSME loan monitoring. It uses a two-signal drift diagnosis: input-distribution change raises a monitoring flag, while deterioration in matured-label performance provides evidence of predictive degradation. Candidate updates use recency-weighted observations and robust preprocessing, but are deployed only if they pass chronological validation, probability-calibration checks, explanation-stability analysis, and subgroup review. The central premise is that adaptation should be evidence-gated and reversible.

The paper makes four contributions: (1) a clear separation of feature shift from performance degradation; (2) a robust update and model-promotion procedure for delayed labels; (3) an explanation-stability measure alongside predictive metrics; and (4) a synthetic experiment with explicit shift conditions, designed to demonstrate the protocol without fabricating institutional findings.

II. RELATED WORK

A. Credit Risk Modelling

Comparative studies benchmark classification algorithms for credit scoring and show that predictive rankings vary with the dataset and metric [1], [11]. Reviews of credit scoring further identify recurring issues in outcome definition, class imbalance, and validation design [12]. In behavioral monitoring, the problem differs from origination scoring: repeated account observations arrive over time, outcomes mature after a horizon, and intervention decisions may affect later labels. Therefore, random train-test splits may allow future-like information into training or place adjacent observations from the same borrower in both partitions.

Doi:10.48047/ijiee.2024.14.4.19

Recent credit-risk modelling research emphasizes that complex learners can improve nonlinear pattern capture, but deployment requires reliable probability estimates and audit-ready interpretation [6], [7]. Calibration methods such as Platt scaling and isotonic regression can correct score reliability on a validation sample, but calibration itself may drift and needs temporal reassessment [13]. The proposed framework treats calibration as a monitored model component rather than a one-time post-processing step.

B. Concept Drift and Adaptation

Concept-drift research distinguishes abrupt, gradual, recurring, and incremental changes and surveys detection and adaptation strategies [3]. Adaptive windowing methods update the retained sample as evidence of change accumulates [4]. Performance-aware detectors use prediction errors or losses to determine when model quality has changed, while feature-only detectors can provide earlier but weaker evidence [5]. In behavioral credit scoring, local competence approaches have been proposed to adapt models to evolving borrower populations [14].

These methods motivate the present design but also expose limitations. A detector may produce false alarms under small samples, and a statistical signal does not state whether performance changed enough to justify model replacement. Delayed credit outcomes make rapid error-based detection impossible for long horizons. This paper therefore combines immediate feature monitoring with delayed performance validation and applies guardrails at the candidate-promotion stage.

C. Explainability and Robustness

SHAP provides additive attributions based on Shapley values, while LIME approximates local model behaviour with a simpler surrogate [8], [9]. These approaches help communicate which observed variables influenced a score, but explanation stability under changing populations remains a separate property from accuracy. A model may preserve AUROC while changing the dominant risk drivers, or retain similar explanations while becoming poorly calibrated. The framework consequently monitors both local explanations for reviewed accounts and global driver rankings across temporal windows.

Credit-risk explainability research argues that model transparency and auditability are necessary for institutional use [7]. However, explanations should not be interpreted as counterfactual advice or causal evidence. Our use case is narrower: expose influential evidence, compare driver patterns over time, and flag large changes for human investigation. The system must preserve the explanation method, reference sample, feature version, and model version used for each score.

III. METHODOLOGY

A. Prediction Target and Point-in-Time Features

For borrower i observed at monthly cut-off t, let $x_{i,t}$ be a feature vector and $y_{i,t}^{H}$ a binary event that equals one if a lender-defined deterioration state occurs during the next H months. The event may combine pre-specified arrears, restructuring, covenant, or watch-list conditions. Credit and model-risk owners define the event before experimentation. A feature is valid only if it was available at t; records created after the cut-off are excluded even if later backfilled.

Candidate features include instalment lateness, payment-gap statistics, revolving-limit utilization, operating-account inflow trend, balance volatility, sales or turnover proxies, returned payments, leverage, sector, geography, facility age, and macroeconomic series. Ratios are winsorized or transformed under rules fitted on the training window. Missingness is represented explicitly when it is informative and is not automatically treated as low risk. Borrower-linked facilities are grouped during validation to reduce leakage from near-duplicate records.

B. Drift Diagnosis

The feature monitor compares current-window distributions with a stable reference. For feature j, the population stability index (PSI) can summarize changes in prespecified bins. Jensen–Shannon divergence or a two-sample statistic may be preferred for sparse or continuous features. The threshold is selected using historical stability and false-alarm analysis, not adopted as a universal rule. Feature warnings are tagged by cohort, data source, and feature family to help distinguish actual economic change from pipeline breakage.

$$\mathrm{PSI}_j = \Sigma_b\,(P_j,b - Q_j,b)\,\ln[(P_j,b + \varepsilon)/(Q_j,b + \varepsilon)] \quad (1)$$

Predictive degradation is assessed only when labels mature. A rolling loss L_t (e.g., Brier score or log loss), event recall at the operating review rate, and calibration slope are compared with the incumbent baseline. A drift alarm is raised when a control rule is exceeded over a prespecified number of consecutive evaluation windows, with minimum sample size and confidence bounds. This avoids allowing a single noisy month to trigger a production change.

$$D_t = 1\{\, L_t - \mathrm{Lref} > \delta L \text{ and } n_t \geq \mathrm{nmin}\,\};\quad C_t = 1\{\, \mathrm{PSI}_t > \delta X\,\} \quad (2)$$

C_t flags covariate change; D_t flags evidence of degraded predictive performance. They are interpreted jointly: C_t=1,D_t=0 means investigate and monitor; C_t=1,D_t=1 permits candidate adaptation; C_t=0,D_t=1 prompts review for label, pipeline, or relationship shift that feature statistics did not reveal. D_t is delayed by the outcome horizon, an inherent limitation rather than a system defect.

C. Robust Estimation and Update Gate

A regularized logistic regression is retained as an interpretable incumbent. A tree ensemble may be evaluated as a challenger. The candidate model uses an exponentially decayed sample weight so recent matured observations receive greater weight while older periods still constrain the fit. To limit sensitivity to outliers, continuous features are clipped at quantiles estimated only from the candidate training period; robust scaling and missingness flags are applied consistently. The decay factor and clipping bounds are fixed on rolling validation and not selected using the final holdout.

$$w_s = \exp[-\kappa(t - s)];\quad \hat{\theta}_t = \arg\min\theta\ \Sigma_s\, w_s\, \ell(y_s, f\theta(x_s)) + \lambda||\theta||_2{}^2 \quad (3)$$

The candidate is calibrated on a separate, chronologically later window of matured labels. It is promoted only when it meets all predefined conditions: acceptable discrimination and calibration; no material deterioration in event capture at the

Doi:10.48047/ijiee.2024.14.4.19

intended review capacity; stable or explainably changed driver patterns; no unresolved data-quality alarms; and satisfactory performance across approved portfolio segments. If the candidate fails any guardrail, the incumbent remains in service and the failure is recorded. A rollback artifact is retained for every release.

D. Explanation Stability and Risk Reasons

For model version v and time window t, global importance can be summarized by mean absolute attribution across accounts. Stability between successive windows is calculated as Spearman rank correlation over the shared feature set. A low correlation does not prove the model is wrong; it triggers inspection of data availability, portfolio mix, model version, and potential structural change. Local explanations are stored with the input values, attribution baseline, score, and timestamp. Missing data and unusual input values are displayed as explicit review reasons.

$$ES_t = \rho S(\mathrm{rank}(|\bar{\varphi}_t|), \mathrm{rank}(|\bar{\varphi}_{t-1}|)); \quad \bar{\varphi}_{j,t} = \mathrm{mean}_i\, |\varphi_{i,j,t}| \quad (4)$$

E. Validation Design and Measures

Validation follows a forward-chaining scheme. At each scoring date, training includes only labels whose full H-month outcome window is complete. Hyperparameters and thresholds are tuned on earlier rolling folds; one final chronological block is untouched until the design is frozen. Repeated observations are clustered by borrower for uncertainty estimation. A model comparison reports AUROC, average precision, Brier score, calibration slope, event recall and precision at a fixed workload, and explanation stability across adjacent test windows. Metrics are reported overall and for policy-approved segments with sample counts and intervals.

This evaluation separates statistical quality from operational robustness. AUROC measures ranking across thresholds; average precision is informative for minority events; Brier score measures probability error; recall at capacity measures how many future stress cases enter the actual review queue. Explanation stability checks whether the model's stated evidence persists. No single metric is sufficient, and a higher AUROC does not imply that the worklist is more useful or that intervention prevents loss.

F. Robustness Checks and Subgroup Evaluation

Robustness analysis should test whether the model relies on brittle features or on a stable combination of repayment and operating signals. Perturbations are applied within plausible ranges and never used to alter the true target: small changes to balance volatility, late-payment counts, missingness flags, or sector proxies should not produce implausibly large swings in risk. For each account, score sensitivity can be summarized as the maximum absolute change across a predefined set of plausible perturbations. The sensitivity distribution is reported by risk band and portfolio segment; extreme cases are reviewed for data errors or discontinuous decision boundaries.

$$RS_i = \max_k |\hat{p}(x_i + \Delta_k) - \hat{p}(x_i)|, \quad \Delta_k \in P \quad (5)$$

A temporal holdout should also be stratified by facility type, sector, region, borrower age, and exposure band where data and policy permit. Group metrics include event prevalence, Brier score, calibration slope, recall at the common workload, and confidence intervals. Small cohorts can produce unstable point estimates, so minimum sample rules are defined and suppressed values are not interpreted as evidence of parity. Segment comparisons are diagnostic; they do not by themselves establish fairness or legal compliance. A separate governance review determines which attributes are allowed, how proxy variables are treated, and what remedial action is appropriate.

Counterfactual or "what-if" explanations should be constrained to actionable and plausible changes. For example, an analyst may ask whether updated verified turnover or a corrected repayment record changes a score. The system should not imply that a borrower can improve a model score by changing an unverified proxy or by taking on additional debt. Explanations remain descriptive of model behavior, and any communication to the borrower follows the lender's established policies and applicable rules.

G. Model Promotion and Monitoring Governance

The framework uses a shadow-to-challenger release sequence. The first stage computes candidate predictions without changing the analyst worklist. Credit and model-risk teams compare candidate and incumbent scores, calibration, explanation changes, and proposed queue composition over multiple matured-label windows. The second stage permits a controlled pilot for a limited segment or business unit while the incumbent remains available. Only after the pilot meets predefined acceptance conditions is the candidate promoted to the primary monitoring flow. This staged approach reduces the risk that a statistical improvement on a historical replay becomes an immediate workflow disruption.

A model card for each version records the target event and horizon, included populations, excluded cases, training and calibration windows, drift signals, update cadence, feature treatment, intended use, known limits, approval owner, and rollback plan. Change records include the reason for retraining, whether the change was driven by feature shift or performance degradation, the validation evidence, and any accepted trade-offs. Model versions and feature definitions are immutable for replay. If a data source changes its semantics, the feature is versioned and the validation protocol restarts rather than silently reusing an old definition.

Human review is itself a selective measurement process: outcomes are more likely to be recorded for borrowers who were alerted or escalated. To reduce biased learning, the study distinguishes verified outcomes from analyst disposition, tracks which cases were reviewed, and evaluates label coverage across risk bands. When legally and operationally permissible, a small audit sample outside the alert queue can be reviewed to estimate missed-event rates. These labels are collected under an approved sampling plan and do not alter customer treatment merely to improve the model.

H. Probability Calibration and Update Uncertainty

A high-ranking score is not necessarily a reliable probability. The calibration window is separated from model fitting and is drawn from the most recent fully matured labels available before the test period. Reliability diagrams compare observed event frequency with predicted probability within prespecified score bins. The Brier score is monitored as an aggregate proper

Doi:10.48047/ijiee.2024.14.4.19

scoring rule, while calibration slope and intercept are used to detect systematic under- or over-confidence. If a cohort has too few events for stable bin estimates, the system reports uncertainty rather than an apparently precise reliability curve.

$$ECE = \Sigma_\beta\ (n_\beta/N)\ |\ acc(y_\beta) - conf(p_\beta)\ |\quad (6)$$

Calibration updates are versioned and tested independently of the underlying classifier. A recalibration is not assumed to repair ranking degradation. If both ranking and calibration have deteriorated, a model candidate may be evaluated; if ranking remains adequate but probabilities shift, a constrained recalibration may be safer. This distinction helps avoid retraining the entire system when the problem lies in the probability scale. Calibration, threshold selection, and alert capacity are frozen for each temporal test window so that evaluation is not optimized on the observed outcomes.

I. Reliability Under Delayed and Selective Labels

Credit-stress labels are delayed by the horizon H and may be censored when loans are refinanced, transferred, or closed. The study protocol specifies whether such accounts are excluded, treated as competing events, or followed through successor facilities. An outcome-maturity table reports the number of eligible observations by scoring month and segment. If a label definition changes during the study, earlier and later outcomes are not mixed without an explicit mapping and sensitivity analysis.

Alert review also influences what becomes known. When analysts inspect high-scoring borrowers more frequently, actionable issues may be documented sooner in the selected group, creating differential observation. The monitoring team should maintain an audit sample or use a documented outcome-ascertainment process where permitted. In offline evaluation, metrics are calculated against outcomes collected independently of the model's alert disposition whenever possible. If only reviewed cases have reliable labels, the report describes this limitation and avoids treating unreviewed borrowers as confirmed negatives.

IV. IMPLEMENTATION AND RESULTS

A. Synthetic Benchmark

No institution-provided MSME data were available. Accordingly, we generated a deterministic panel of 18,000 synthetic borrower-period records over 15 periods using seven standardized behavioral features. Period 7 introduces abrupt shifts in several feature distributions and changes the coefficients linking borrower behaviour to future stress. From period 10, a gradual shift changes cash-flow and balance-volatility patterns. A binary event is sampled from a logistic data-generating process. This known construction allows controlled measurement of adaptation but cannot represent real-world borrower heterogeneity, censored observations, intervention effects, or lending policy.

Periods 0–5 train the frozen baseline. Periods 9–14 are held out for pooled evaluation. For each test period, the rolling models use only observations whose two-period outcome horizon has matured. The frozen comparator is class-weighted logistic regression trained once. The rolling comparator refits logistic regression on a recent matured window. The robust comparator uses the same rolling window with 1st–99th percentile clipping fitted on training data and stronger L2 regularization. All models use the same feature set and 15% alert capacity; the synthetic event prevalence and random seed are fixed.

TABLE I SYNTHETIC EXPERIMENT DESIGN

Item	Specification
Panel	15 periods × 1,200 synthetic records
Features	Seven behavioural and portfolio indicators
Shift pattern	Abrupt at period 7; gradual from period 10
Outcome delay	Two periods before labels are eligible
Test block	Periods 9–14, forward in time
Capacity	Highest-scored 15% of cases per pooled test set

B. Predictive Results

Table II summarizes pooled performance. The robust rolling model achieved AUROC 0.828 and average precision 0.667, compared with 0.776 and 0.590 for the frozen model. Brier score was 0.181 for the robust model and 0.223 for the frozen model. Under the fixed 15% alert budget, recall was 37.4% for robust adaptation and 33.6% for the frozen baseline. These differences are generated by the specified feature and concept shifts. They show that the protocol can compare methods under a known non-stationary process; they do not establish that winsorization, stronger regularization, or rolling retraining will improve any real lender's performance.

TABLE II POOLED SYNTHETIC TEST PERFORMANCE

Metric	Frozen	Rolling	Robust rolling
AUROC ↑	0.776	0.828	0.828
Average precision ↑	0.590	0.665	0.667
Brier score ↓	0.223	0.181	0.181
Recall at 15% ↑	0.336	0.374	0.374
Precision at 15%	0.654	0.728	0.727

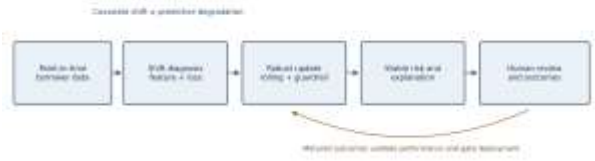


Fig. 1. Drift-aware modelling loop. Feature shift raises a warning; matured predictive evidence gates candidate promotion.

C. Operational Interpretation

The synthetic test event prevalence was 29.2%. At 15% capacity, the queue captures only a fraction of all future positive cases, emphasizing the need to report the capacity-recall curve rather than one threshold result. The robust candidate's alert precision was 72.7%. In an actual portfolio, a queue should also measure exposure captured, review hours, time to first contact, completed reviews, and subsequent resolution. These metrics distinguish a better risk ranking from a more effective credit intervention.

Doi:10.48047/ijiee.2024.14.4.19

Implementation should preserve a reproducible snapshot and record feature lineage, label definition, model version, calibration version, score, explanation, and queue position. A reviewer interface should show the evidence date for each driver and a clear reason for an alert. Drift monitoring must distinguish source outages from borrower change; an unavailable cash-flow feed must never silently reduce risk. Candidate updates should run in shadow mode before promotion, with an incumbent fallback and a documented rollback decision.

D. Shift-Specific Interpretation and Ablation Plan

Abrupt and gradual drift create different operational demands. A sudden shift in payment behavior may require rapid diagnosis of an external shock, a source-system change, or a policy event. A gradual shift may remain below a single-period threshold while accumulating into meaningful degradation. The proposed two-signal monitor addresses this by tracking feature distributions frequently and evaluating rolling matured-label loss across overlapping windows. A reference window is periodically reviewed so that an old baseline does not make a permanent economic regime appear anomalous. Reference updates are approved changes and are not performed automatically.

An ablation study should compare the full robust pipeline with (i) no feature-shift monitor, (ii) no performance-aware gate, (iii) no recency weighting, (iv) no robust clipping, (v) no temporal calibration, and (vi) no explanation-stability check. For each ablation, training windows, validation periods, class weights, and alert budget remain fixed. The outcomes should include model-wide ranking, probability reliability, event capture at capacity, update frequency, false drift alarms, score-change magnitude, explanation-rank stability, and analyst hours. A component is useful only if it improves the intended outcome without causing unacceptable instability or workload.

A stress test can vary the outcome delay, drift magnitude, prevalence, and missingness mechanism. Delays of one, two, or more periods reveal how much predictive deterioration can be detected before the target has matured. Higher missingness in a single sector tests whether the model confuses reporting gaps with distress. A return to an earlier regime tests whether the rolling model can recover from recurring concepts. These experiments should be pre-registered in the evaluation plan to avoid selecting scenarios that happen to favor a proposed method.

E. Reading the Results Against the Research Question

The proposed research question is whether a robust update can preserve early-warning performance when borrower behavior changes, while maintaining interpretable and governable model behavior. In the synthetic benchmark, both rolling models improve on the frozen model because recent observations contain information about the changed regime. The robust variant is close to the ordinary rolling model: its AUROC is marginally lower, average precision marginally higher, Brier score marginally higher, and event recall nearly identical. This mixed comparison is useful. It shows that robust clipping and stronger regularization are not guaranteed to improve every objective even when the overall adaptation is beneficial relative to a stale model.

A deployment decision should therefore depend on the institution's acceptable trade-offs and confidence intervals, not on rounding differences in a single simulation. If probability accuracy is central to expected-loss estimation, the Brier and calibration results may carry more weight. If scarce analysts need to capture as many deteriorating accounts as possible, recall at the fixed workload may be primary. If an update substantially changes driver explanations or increases score volatility, model-risk owners may require additional review even when ranking improves. The appropriate acceptance rule is agreed in advance and tied to the model's intended use.

A complete empirical paper would repeat the evaluation over multiple chronological cut-offs and economic conditions, report paired borrower-clustered confidence intervals, and compare the proposed update with an incumbent scorecard and at least one strong nonlinear challenger. It would publish the queue capacity, event prevalence, label delay, feature availability, and model-update schedule so that others can assess whether reported gains are operationally plausible. The present synthetic results provide an example of the table structure and analysis needed for that future study.

D. Limitations

The benchmark uses synthetic data with known coefficients, complete features, one binary event, and controlled shifts. Its event prevalence is generated rather than estimated and may differ substantially from lending portfolios. The analysis does not establish fairness, causal impact of alerts, calibration under severe label delay, or robustness to strategic borrower behaviour. It also does not compare local models, boosted trees, survival methods, or all detector families. A bank-level study requires authorized point-in-time loan histories spanning multiple business cycles, a documented stress label, leakage audits, subgroup checks, and prospective silent evaluation before operational use.

V. CONCLUSION

This paper presented an explainable framework for robust MSME credit-stress prediction under borrower behavioural drift. It separates input-distribution warnings from matured evidence of predictive degradation, applies recency-aware robust updates, monitors calibration and explanation stability, and promotes a candidate only after chronological validation and governance checks. The synthetic benchmark shows how to compare a frozen model with rolling and robust updates under controlled abrupt and gradual shifts. Its numerical findings are illustrative and do not constitute empirical evidence from a bank. Future work should validate the framework with an authorized MSME portfolio, measure alert utility in a silent pilot, and assess whether human review outcomes improve borrower support and risk management.

REFERENCES

[1] 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, doi: 10.1016/j.ejor.2015.05.030.

[2] World Bank, Credit Scoring Approaches Guidelines. Washington, DC, USA: World Bank, 2020.

[3] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Computing Surveys, vol. 46, no. 4, Art. no. 44, 2014, doi: 10.1145/2523813.

Doi:10.48047/ijiee.2024.14.4.19

[4] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in Proc. SIAM Int. Conf. Data Mining, 2007, pp. 443–448, doi: 10.1137/1.9781611972771.42.

[5] F. Bayram, B. S. Ahmed, and A. Kassler, "From concept drift to model degradation: An overview on performance-aware drift detectors," Knowledge-Based Systems, vol. 245, Art. no. 108632, 2022, doi: 10.1016/j.knosys.2022.108632.

[6] V. Chang, Q. A. Xu, S. H. Akinloye, V. Benson, and K. Hall, "Prediction of bank credit worthiness through credit risk analysis: An explainable machine learning study," Annals of Operations Research, 2025, doi: 10.1007/s10479-024-06134-x.

[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, doi: 10.1007/s10614-020-10042-0.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[10] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. 23rd Int. Conf. Machine Learning, 2006, pp. 233–240, doi: 10.1145/1143844.1143874.

[11] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," Journal of the Operational Research Society, vol. 54, no. 6, pp. 627–635, 2003, doi: 10.1057/palgrave.jors.2601545.

[12] F. Louzada, A. Ara, and G. B. Fernandes, "Classification methods applied to credit scoring: A systematic review and overall comparison," Surveys in Operations Research and Management Science, vol. 21, no. 2, pp. 117–134, 2016, doi: 10.1016/j.sorms.2016.10.001.

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

[14] D. Nikolaidis and M. Doumpos, "Credit scoring with drift adaptation using local regions of competence," Operations Research Forum, vol. 3, no. 4, pp. 1–28, 2022, doi: 10.1007/s43069-022-00177-1.

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

Doi:10.48047/ijiee.2024.14.4.19