

Capacity-Constrained Credit Monitoring: Explainable and Drift-Adaptive Prediction of MSME Repayment Stress

D Ajay Kumar

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

Abstract— Small and medium enterprises (MSMEs) often exhibit volatile cash flows, sparse formal records, and repayment signals that change with seasonality and macroeconomic conditions. Credit monitoring teams also operate under a finite review budget, so a useful early-warning system must rank cases by expected stress while accounting for the number of cases officers can investigate. This paper specifies a capacity-constrained framework that estimates a calibrated probability of repayment stress over a fixed horizon, explains each score using local feature attributions, monitors population and performance drift, and selects up to a monthly review capacity using expected avoidable loss. The proposed pipeline uses a temporal gradient-boosted model with monotonic constraints where domain direction is defensible, rolling-origin validation, delayed-label aware calibration, a drift trigger with approval-gated updates, and a budgeted top-K policy. A worked example with five constructed borrower profiles illustrates how risk, exposure, explanation, and review capacity combine; it is not an empirical evaluation on bank records. The paper therefore provides an implementation and evaluation protocol, metric definitions, governance controls, and a reproducibility checklist rather than claiming measured portfolio gains. The design targets operational decision support and retains human review for adverse actions.

Index Terms— Credit risk, explainable machine learning, concept drift, early warning, MSME lending, capacity-constrained ranking, probability calibration.

I. INTRODUCTION

MSME lending decisions are made in an environment where financial statements may arrive infrequently while transaction, repayment, and account activity signals change continuously. A monitoring model must identify emerging stress early enough for a relationship manager to verify the situation and offer an appropriate intervention. A binary default label alone is insufficient: it ignores time-to-event, risk severity, exposure, and the cost of investigating a false alarm. The practical decision is therefore a ranked allocation of scarce review capacity.

Credit scoring research has compared statistical and machine-learning methods across diverse datasets [1]–[4]. Flexible ensembles can capture nonlinear interactions, but financial institutions need stable probabilities, defensible reasons, and controls over data and model changes [5], [6]. MSME monitoring adds two constraints: repayment outcomes are delayed, and a risk model that was calibrated in one economic period may become unreliable after a shift in borrower mix or operating conditions [7], [8].

This paper proposes a decision-support design with four linked components: (1) horizon-specific repayment-stress prediction; (2) local explanations and monotonic guardrails; (3) drift monitoring that separates a signal from permission to retrain; and (4) a capacity-constrained review queue. Its central objective is not to maximize a threshold-free score in isolation. It is to prioritize cases with the greatest expected benefit from human attention, subject to a monthly review budget.

The contribution is a specification and worked example rather than a claim of deployment performance. No confidential bank data were supplied. Accordingly, all borrower records in Section IV are constructed solely to make the selection rule transparent. The equations and experiment plan define what must be implemented and measured before an empirical claim can be made.

II. RELATED WORK

A. Credit scoring and small-business risk

Traditional credit scorecards rely on interpretable statistical relationships and remain useful baselines. Comparative work has documented that model ranking depends on dataset and evaluation design, with ensemble methods often competitive but not uniformly superior [1], [2]. Business lending research further notes that transaction-level data can support more timely risk estimation, while the source and temporal granularity of features affect validity [3], [4]. Small-business banking work demonstrates how constrained boosting and feature attribution can support reviewable risk scores in operational contexts [5]. These findings motivate comparison against logistic regression and a nonadaptive tree ensemble instead of assuming that a complex model is always preferable.

B. Explainability and decision governance

Local Interpretable Model-agnostic Explanations approximate a predictor near one observation [9], while SHapley Additive exPlanations assign additive contributions relative to a background distribution [10]. For monitoring, the explanation

should answer why a specific account was ranked and whether the reason is actionable, not merely provide a global importance plot. Attribution is not causal evidence: a high feature contribution does not establish that changing the feature will reduce default risk. Credit scoring guidance recommends model governance, oversight of unintended consequences, and clear ownership through development and use [11]. Therefore, explanations are paired with reason codes, stability checks, and officer verification.

C. Drift, temporal validation, and calibration

Concept drift describes a change in the relationship between inputs and outcomes; data may also shift in distribution without an immediate change in the conditional relationship [7]. Dynamic credit-risk work has adapted drift-handling concepts to credit scoring [8], and local competence methods have studied adaptive scoring under changing populations [12]. Online change detectors such as ADWIN maintain adaptive windows to test whether a monitored sequence has changed [13]. In credit portfolios, labels mature late, so monitoring input distributions and probability calibration must be distinguished from confirming performance drift after labels arrive. Calibration is critical because business rules consume probability estimates, not only class ranks [14].

D. Gap addressed

Prior work commonly treats prediction quality, explanation, drift, and operational workload as separate questions. The proposed framework connects them in one evaluation loop: a risk estimate is calibrated on temporally prior data; explanation accompanies queue selection; drift signals prompt controlled review; and the queue is constrained by available officer capacity. This integration is a testable design proposition, not a claim that these components guarantee better outcomes.

III. METHODOLOGY

A. Prediction target and observation unit

Let i denote an active MSME facility observed at month t . The target $Y_{i,t+h}$ equals one when a pre-registered stress event occurs in the next h months. An operational definition may combine 30+ days past due, a restructuring flag, or a verified covenant breach; the exact event and horizon must be fixed before analysis and aligned with the lender's policy. Multiple events should be retained as separate labels in sensitivity analysis. The prediction cutoff must precede each label window to prevent outcome leakage.

Candidate inputs include repayment roll rates, days past due, utilization, balance volatility, inflow trend, returned payments, overdue supplier obligations where lawfully available, sector, loan vintage, and exposure. Features are computed as-of t . Protected or sensitive attributes should be excluded from production scoring unless a lawful, approved fairness audit specifically requires their controlled use. Missingness indicators and reporting delays are retained because absence of information may itself reflect operational processes.

B. Horizon-specific score and calibration

A gradient-boosted decision tree is the primary candidate because it captures nonlinear effects on tabular data and supports monotonic constraints for relationships with clear domain direction. The model output is transformed into an estimated event probability using a calibrator fitted only on an earlier temporal validation window. Logistic regression and random forest are benchmark models. A risk estimate is denoted $p_{i,t}$; for each borrower the score is updated monthly only after all features pass the point-in-time validation checks.

$$p_{i,t} = \text{Cal}_{t-1} [f\theta (x_{i,t})] \quad (1)$$

where $f\theta$ is the fitted model and Cal_{t-1} is an isotonic or logistic calibration map trained on matured labels preceding t . The calibrator is refit less frequently than the score is computed, avoiding unstable month-to-month probability shifts. Brier score, log loss, calibration slope, and reliability diagrams are reported alongside AUROC and area under the precision-recall curve (AUPRC).

C. Explainable risk factors

For a tree ensemble, local SHAP values provide a decomposition of the model output relative to a fixed, documented background sample. In probability space, this attribution need not sum exactly unless the explanation method is configured for the chosen output scale; log-odds explanations are therefore preferred for additive consistency.

$$g(p_{i,t}) = \phi_0 + \sum_{j=1}^M \phi_{ij} \quad (2)$$

Here g is the logit link, ϕ_0 is the reference prediction and ϕ_{ij} is the contribution of feature j for borrower i . The interface exposes the top positive and negative contributors, the underlying observed value, the comparison reference, timestamp, and data quality flag. A “reason” is not an instruction to change the feature; it is a prompt for verification and a borrower conversation. Stability is checked by comparing reason-code distributions across adjacent periods and by perturbing plausible input values.

D. Capacity-constrained prioritization

Let q_i be the estimated probability that review of borrower i will enable an effective, policy-compliant intervention, and let L_i be avoidable loss if stress is addressed. Let c_i denote review effort and B the total officer-hours available in a period. The selection problem is a 0–1 knapsack objective:

$$\max \sum_i z_i (p_i L_i q_i - \lambda c_i), \quad \text{s.t. } \sum_i z_i c_i \leq B, z_i \in \{0,1\} \quad (3)$$

where $z_i=1$ means the case is assigned for review and λ prices operational effort. If all cases have equal review effort, selecting the top K expected benefit values is a special case. If q_i or L_i is not estimated from valid historical interventions, use a transparent risk-times-exposure proxy and state this limitation; do not present it as a causal estimate of avoided loss. The capacity rule should also reserve a small, predeclared share for random audits and underserved segments to detect blind spots, without changing their risk probability.

E. Drift monitoring and adaptation

The monitoring layer tracks (i) input drift, using population stability or distributional tests on key variables; (ii) score drift; (iii) label-based performance after outcomes mature; and (iv) explanation drift in reason-code distributions. An ADWIN-style detector may monitor a bounded loss sequence, such as absolute calibration error, once outcomes are available [13]. A signal triggers an investigation ticket, not automatic retraining. The model owner checks data pipelines, policy changes, seasonality, label delay, and subgroup performance. Retraining requires a minimum sample, temporal holdout validation, calibration, fairness review, sign-off, and rollback artifact.

$$D_t = 1\{|\hat{\mu}_{\text{recent}} - \hat{\mu}_{\text{reference}}| > \varepsilon_t\} \quad (4)$$

D_t is a simplified drift alarm for a monitored statistic; ε_t is determined using a documented sequential test or control limit. The production trigger should account for multiple testing and baseline seasonal variation. A confirmed deterioration is defined in advance, for example as a sustained increase in expected loss or calibration error on matured outcomes, rather than a single alert.

G. Data contract and leakage controls

Each feature must have a named source system, event time, availability time, refresh cadence, and fallback rule. A row may enter a training set only if its feature values were available by the scoring cutoff, not merely posted later with an earlier transaction date. The target window begins after that cutoff, and any account with unresolved outcome maturity is censored or excluded according to a documented rule. Restructuring, write-off, and delinquency should be timestamped separately so an event caused by an earlier internal action is not inadvertently used as a predictor for the same episode. Duplicate accounts and facilities are reconciled to a stable borrower identifier before splitting.

Table III gives an example feature contract. The exact set depends on lawful availability, data quality, and the lender's product design. The purpose is to make the observation-time logic explicit and to prevent a seemingly strong model from relying on post-event information.

Example variables	Time and quality rule
DPD, missed instalments, cure count	As known at month-end; exclude future posting
Inflow trend, balance volatility	Rolling window ending before score date
Utilization, outstanding, maturity	Snapshot aligned to score timestamp
Return item, contact history	Record event time optimizability on
Sector, vintage, geography	Version taxonomy; monitor mix shifts

TABLE III. EXAMPLE POINT-IN-TIME FEATURE CONTRACT

H. Decision thresholds and borrower treatment

A risk probability should not be interpreted as a credit decision. Monitoring thresholds determine the order of human review, while the reviewer verifies current circumstances and follows existing policy. Where risk bands are used, each band should map to a permitted action such as routine check-in, document verification, or specialist review; it must not trigger automatic

denial, repricing, or collection escalation. Thresholds should be set against the available capacity and the relative cost of missed stress, then approved by the credit policy owner. When capacity is exceeded, the queue should state the reason a case was deferred and when it will be rescored.

The review interface should distinguish model uncertainty from high risk. For example, a borrower outside the training distribution may receive a wide uncertainty interval even when its point estimate is moderate. Such cases can be routed to a manual review bucket with a controlled capacity allowance. This is preferable to treating a point estimate as precise when the data are sparse. Uncertainty can be approximated through calibrated ensembles or bootstrap variation, but the method must be validated and displayed separately from the stress probability.

F. Evaluation design

Use rolling-origin evaluation: train on months 1...T, calibrate on T+1...T+m, and test on the next nonoverlapping block; then advance the origin. Split by borrower so that repeated observations do not leak across folds. Include periods with different sector and macroeconomic conditions where available. Compare logistic regression, random forest, gradient boosting without adaptation, and the proposed drift-aware queue. Evaluate both model and operations: AUROC, AUPRC, Brier score, recall at K, precision at K, expected review cost, workload utilization, lead time, and the share of stress events detected before delinquency. Report confidence intervals by bootstrap over borrower identifiers and subgroup results where lawful and adequately powered.

IV. IMPLEMENTATION AND RESULTS

A. Reference implementation

A practical implementation consists of a point-in-time feature store, monthly batch scoring, a versioned calibration service, explanation generation, and an auditable queue interface. Each score record stores borrower key, as-of timestamp, model and calibrator versions, predicted horizon risk, expected exposure, top reason codes, data freshness, drift status, and queue decision. Training uses temporal folds and class-weighted loss restricted to training folds and never applied to test periods. The same feature transformations and missing-value rules are used in training and production. The scoring service produces a candidate queue; a deterministic optimizer chooses assignments under the staffing limit. Reviewers record the contact outcome, reason for escalation, and intervention. These action records are essential to estimate q_i and avoid interpreting post-review outcomes as unbiased labels: reviewed accounts receive different treatment from unreviewed accounts. Until an appropriate causal or policy evaluation is performed, operational impact is reported as queue efficiency and early identification, not "loss prevented."

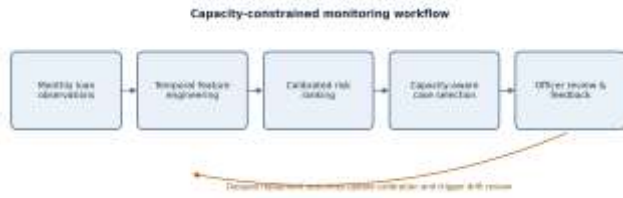


Fig. 1. Proposed monthly monitoring and feedback workflow.

The monthly scoring job first validates source freshness and schema versions, then assembles borrower-month features. A point-in-time check blocks fields whose availability timestamp is later than the scoring cutoff. The model and calibrator are loaded from immutable registry versions; the score is generated only for active facilities with a mature observation window. After scoring, a queue service combines the probability, exposure, review effort estimate, and any approved feasibility weight. The output is stored before the reviewer interface is updated, so later changes can be traced to a model version or an authorized override.

A second job receives repayment outcomes after a maturity delay. It links each event to the original score timestamp, calculates the agreed performance measures, and updates drift dashboards. Data drift is reported immediately; outcome drift is reported only when enough labels have matured. If the number of labels is small, the dashboard shows uncertainty intervals and suppresses unstable subgroup rates. A model release cannot be triggered solely by a dashboard alarm; the model-risk process records the investigation, decision, and rollback option.

The architecture is intentionally batch-oriented. Monthly scoring is often sufficient for portfolio monitoring and is easier to audit than continuous online updates. A lender with daily transactions can still compute recent aggregates daily while issuing a governed monthly queue. The scoring cadence should be selected from repayment frequency, contact capacity, feed latency, and the expected lead time needed for useful borrower support. A faster refresh has little value if the review team cannot act on additional alerts.

B. Worked capacity example

Table I uses five constructed accounts, illustrative probabilities, hypothetical exposure amounts, and equal 30-minute review effort. It demonstrates the risk-times-exposure proxy under a two-case capacity. No bank observations were used, and these rows are not model predictions or measured results. Actual implementations should replace the values with validated out-of-time estimates and policy-approved loss weights.

Stress p	Exposure (₹ lakh)	$p \times \text{exposure}$ (₹ lakh)
0.32	12	3.84
0.21	18	3.78
0.44	6	2.64
0.12	20	2.40
0.27	7	1.89

TABLE I. CONSTRUCTED EXAMPLE OF CAPACITY-BASED PRIORITY

The example shows why highest probability alone can differ from highest expected exposure at risk: Case C has the largest probability but Case B has a slightly larger risk-exposure proxy. In an operational system, the ranking may change after review effort and intervention feasibility are incorporated. The example is illustrative arithmetic only and supports no claim of predictive accuracy.

For the constructed queue, the two selected cases have a combined risk-exposure proxy of ₹7.62 lakh ($3.84 + 3.78$), while using probability alone would select Cases C and A, with a combined exposure proxy of ₹6.48 lakh ($2.64 + 3.84$). This arithmetic illustrates the possible effect of including exposure under a fixed capacity; it does not estimate losses or demonstrate that review prevents them. If case-specific review time differed, the knapsack rule could select a different set. If an intervention feasibility estimate were introduced, it would need to be learned from documented interventions or treated as a transparent policy weight.

C. Baseline and ablation comparisons

A full experiment should isolate the value of each proposed component. The baseline sequence begins with logistic regression using the same features and temporal splits, then a static random forest, then gradient boosting. The next comparison adds probability calibration, followed by explanation monitoring, drift alarms with approval-gated updates, and capacity-aware ranking. For each addition, report both the change in predictive measures and the operational consequence at the same review budget. This ablation design can reveal, for example, whether recalibration improves Brier score while leaving recall@K unchanged, or whether drift adaptation helps only in a subset of test windows.

The main comparison should be paired by test period: each candidate is evaluated on the same borrower-month cohort, with the same label definition and review capacity. Confidence intervals should resample borrowers rather than individual rows because observations from one enterprise are correlated across months. Report the number of stress events and non-events in each test window, and avoid a single aggregate score when time-specific behavior differs. A challenger should be retained only when it offers a useful, stable advantage over the simpler benchmark, not merely a favorable average from one period.

D. Interpretation of the demonstration

The constructed result supports three limited observations. First, an operational objective that includes exposure can produce a different queue from a probability-only sort. Second, the review capacity is a binding policy input: expanding K changes which cases receive attention and changes the measured precision and recall. Third, the queue result depends on assumptions about avoidable loss and review feasibility. These statements follow directly from the arithmetic and optimization definition. They are not empirical claims about MSME populations and should not be generalized to a bank without out-of-time data.

In the proposed real-world evaluation, a useful result would therefore include a table for each test window showing event

prevalence, AUROC, AUPRC, Brier score, recall@K, precision@K, workload hours, and lead time. A separate table should show drift alarms, how many were investigated, how many represented a data defect or policy change, and whether a governed update improved subsequent calibration. Publishing the negative or null results is essential: a system that does not beat the current process under realistic capacity is not an operational improvement.

C. Empirical result reporting plan

Because access to a governed MSME loan dataset is required, this paper does not report fabricated AUROC, recall, calibration, or drift-recovery values. Table II defines the minimum results that should be filled after a proper experiment. Report separate values for each temporal test window and give borrower-level confidence intervals; avoid averaging away a crisis or seasonal regime.

	Measure to report	Required comparison
	AUROC; AUPRC; recall@K	LR, RF, static boosting, proposed
	Brier; log loss; calibration slope	Before and after reviewed update
	Precision@K; cases/hour; utilization	Fixed monthly B and effort-aware queue
	Median months before 30+ DPD	Model queue versus policy baseline
	Detection delay; post-drift loss	Static versus alarm-reviewed update
	Reason stability; subgroup gaps	Across months and optimal group

TABLE II. REQUIRED OUT-OF-TIME REPORTING MATRIX

D. Operational controls and limitations

Key failure modes include stale transaction feeds, changed definitions of delinquency, selective labels, macroeconomic shocks, and officer capacity changing faster than model governance. A flagged account may be “high risk” because of an erroneous feed or policy change. The interface must permit a reviewer to inspect source values and record a correction. Fairness review should assess error and calibration differences across permissible groups and segments, with minimum sample rules and privacy controls. Protected attributes must not be used to prioritize individual borrowers unless an approved legal basis and governance process explicitly allow it.

The framework’s main limitations are data dependence, label delay, intervention confounding, and reliance on proxy loss when causal intervention effects are unavailable. SHAP explanations are model attributions, not causal reasons. Drift detectors can alarm on benign seasonality and miss changes in unmonitored variables. A production claim requires prospective shadow testing, security review, user acceptance, model-risk approval, and an evaluation of whether interventions improve borrower and lender outcomes.

E. Sensitivity and stress testing

Before a live pilot, analysts should test the ranking under plausible changes to review capacity, stress prevalence, and loss severity. Capacity sensitivity evaluates at least three budgets (for example, 5%, 10%, and 20% of the eligible monthly population) and reports precision, recall, workload, and the

number of high-risk accounts left unreviewed. Severity sensitivity compares a uniform exposure proxy with policy-approved segment-level loss weights. These are decision sensitivities, not predictive model performance. The chosen operating point should be selected by the credit policy owner after reviewing the cost of missed stress and the burden of unnecessary contact.

Temporal stress tests should include a sharp sector-level contraction, a change in repayment policy, and delayed or missing transaction feeds. Such tests can be carried out by replaying historical windows or perturbing a held-out test set, but the perturbation assumptions must be stated. A robust design should degrade gracefully: when data freshness falls below a threshold, it should surface a data-quality warning and route the account for manual review or defer the score according to policy. It should not silently impute a benign value that lowers risk.

F. Selection bias and intervention outcomes

Monitoring produces a selective-label problem because officers contact a subset of accounts and the contact can change repayment behavior. A naive comparison of reviewed and unreviewed accounts will therefore mix model selection, borrower condition, and treatment effects. During a shadow phase, scores can be generated without influencing actions to estimate ranking and calibration under current policy. For estimating intervention effects, a later controlled pilot may randomize eligible cases within approved risk bands, or use a carefully justified quasi-experimental design, with safeguards for vulnerable borrowers and urgent cases. Assignment probability, contact attempts, intervention type, and subsequent outcomes must be logged. This separation prevents an operational success claim from being inferred merely because flagged borrowers received more attention.

G. Reproducibility and governance record

Every evaluation release should include a data dictionary, event-label definition, feature cutoff, cohort inclusion rules, temporal split dates, code commit, model parameters, random seed, calibration method, and queue capacity assumptions. The report should document missingness, class prevalence, sector composition, and label maturity in every period. A model card should specify intended use, prohibited use, known limitations, escalation rules, and the person accountable for approval. Explanations, alerts, and overrides should be retained under the institution’s access and retention policy, with sensitive borrower data protected through role-based access and audit logs.

These controls make the proposed method falsifiable. If drift alerts do not precede performance deterioration, if calibrated probabilities remain unreliable, or if fixed-capacity recall fails to exceed the incumbent policy across out-of-time windows, the model should not be advanced to production. A technically more complex predictor is justified only when it improves decision utility under the same review budget and remains understandable to the institution’s users.

V. CONCLUSION

This paper presented a capacity-constrained framework for explainable, drift-adaptive prediction of MSME repayment stress. It combines a horizon-specific calibrated risk estimate, local explanations, monitored changes in data and performance, and a review queue limited by officer capacity. The proposed evaluation separates predictive discrimination from calibration and operational utility. A worked example illustrates how probability and exposure can prioritize review under a fixed budget; it is explicitly constructed and is not empirical evidence.

The next research step is a temporal evaluation on a governed MSME portfolio with point-in-time features, matured outcomes, stable stress definitions, and documented review capacity. Only after the model is compared with simple baselines across multiple time windows, audited for explanation and subgroup stability, and assessed prospectively should predictive or loss-reduction benefits be claimed. The design treats adaptation as a controlled model-governance process and leaves final borrower decisions with authorized staff.

REFERENCES

- [1] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: A review," *J. Roy. Stat. Soc.: Series A*, vol. 160, no. 3, pp. 523–541, 1997, doi: 10.1111/1467-985X.00078.
- [2] 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," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015, doi: 10.1016/j.ejor.2015.05.030.
- [3] L. C. Thomas, "A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers," *Int. J. Forecasting*, vol. 16, no. 2, pp. 149–172, 2000, doi: 10.1016/S0169-2070(00)00034-0.
- [4] M. G. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, 2012, doi: 10.1016/j.eswa.2011.09.033.
- [5] J. Bracke, A. Datta, C. Jung, and S. Z. Shah, "Using small business banking data for explainable credit risk scoring," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 8, pp. 13996–14001, 2020, doi: 10.1609/aaai.v34i08.7055.
- [6] A. Bussmann, A. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," *Comput. Econ.*, vol. 57, pp. 203–216, 2021, doi: 10.1007/s10614-020-10042-0.
- [7] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, Art. no. 44, 2014, doi: 10.1145/2523813.
- [8] M. R. Sousa, J. Gama, and E. Brandão, "A new dynamic modeling framework for credit risk assessment," *Expert Syst. Appl.*, vol. 45, pp. 341–351, 2016, doi: 10.1016/j.eswa.2015.09.055.
- [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] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [11] World Bank, "Credit Scoring Approaches Guidelines," Washington, DC, USA, 2020. [Online]. Available: <https://documents.worldbank.org/en/publication/documents-reports/documentdetail/935891585869698451>
- [12] D. Nikolaidis and M. Doumpos, "Credit scoring with drift adaptation using local regions of competence," *Oper. Res. Forum*, vol. 3, Art. no. 67, 2022, doi: 10.1007/s43069-022-00177-1.
- [13] 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.
- [14] T. Silva Filho et al., "Classifier calibration: A survey on how to assess and improve predicted class probabilities," *Mach. Learn.*, vol. 112, pp. 3211–3260, 2023, doi: 10.1007/s10994-023-06336-7.
- [15] G. Babaei, P. Giudici, and E. Raffinetti, "Explainable FinTech lending," *J. Econ. Bus.*, vol. 125–126, Art. no. 106126, 2023, doi: 10.1016/j.jeconbus.2023.106126.