

A Budget-Constrained, Drift-Aware Framework for Early Warning of MSME Repayment Stress

D Ajay Kumar

Independent Researcher, Telangana, India

Abstract—Micro, small, and medium enterprise (MSME) lenders must identify emerging repayment stress early while working with limited field-investigation capacity and borrower populations whose behavior changes over time. This paper proposes a budgeted early-warning framework that converts calibrated short-horizon stress probabilities into a capacity-constrained review queue. The framework combines a transparent logistic baseline and a gradient-boosted risk model, time-ordered validation, population and performance drift monitoring, delayed-label adaptation, probability calibration, and an uncertainty-based abstention rule. A lender-defined review budget B is represented as a fraction of active accounts; accounts are ranked by expected avoidable loss per review cost, with uncertainty used to route ambiguous cases for manual assessment. The target is an actionable stress event within a stated horizon (for example, transition from a performing state to SMA-1/SMA-2 or a separately defined 30+ days-past-due event), not a regulatory classification or an automatic lending decision. The paper specifies data requirements, leakage controls, baselines, metrics, ablation tests, and operational safeguards. Because no borrower-level dataset or executed experiment was supplied, numerical performance is not asserted; Section IV provides a reproducible results protocol and a reporting template. The contribution is a testable design for evaluating whether drift-aware ranking and uncertainty estimation improve the yield and reliability of a fixed-size MSME early-warning queue.

Index Terms—MSME credit risk, early-warning system, concept drift, risk ranking, uncertainty estimation, selective prediction, review-budget allocation.

I. INTRODUCTION

MSMEs contribute to employment, local supply chains, and regional production, yet smaller firms often have shorter credit histories, concentrated cash flows, and limited audited information. A conventional annual or quarterly score can therefore miss a rapid deterioration in collections, account conduct, or debt-service capacity. Early warning is operationally valuable only when a lender can review the flagged cases and take proportionate action. In practice, relationship managers, credit officers, and field teams have finite time; a model that labels most borrowers as “high risk” cannot serve as an actionable queue.

The Indian regulatory setting already recognizes incipient stress. The Reserve Bank of India (RBI) describes special-mention categories and, for the MSME revival framework, specifies mechanisms for identifying stress before non-performing-asset classification [1], [2]. This paper treats the model as a decision-support layer that may prioritize a human review. It does not replace RBI classifications, lender policy, or borrower-specific judgement. RBI’s Financial Stability Report reported an improvement in the scheduled commercial bank MSME gross NPA ratio from 4.5% in March 2024 to 3.6% in March 2025, alongside moderation in SMA-2 [3]. Portfolio improvement does not remove the need to detect local, borrower-level deterioration.

Credit scoring research shows that complex classifiers can improve discrimination in some settings, while also emphasizing that performance depends on dataset and evaluation design [4], [5]. A further difficulty is temporal instability: a model trained on past borrowers may become

miscalibrated when interest rates, supply chains, payment practices, products, or borrower composition change. Concept-drift methods address evolving data distributions, but they are rarely evaluated together with a fixed review budget, delayed repayment labels, and calibrated uncertainty in an MSME early-warning workflow [6], [7].

We propose a unified, testable framework around three operational questions: (1) among only the accounts a lender can review, how many future stress events are captured? (2) do predicted risks remain calibrated across time and borrower segments? and (3) when evidence is weak or shifted, can the system abstain and route a case for human review? The intended outcome is prioritized outreach or diagnosis, not an automatic decline, repricing, or collection action.

A. Contributions

The paper makes four design contributions. First, it defines a budgeted early-warning objective using top-B recall, precision, and expected loss captured rather than relying only on whole-population AUC. Second, it separates risk ranking from uncertainty: the model estimates event risk, while a calibration and abstention layer indicates when a score should not be acted on without additional review. Third, it specifies a delayed-label drift loop that monitors input and outcome changes and allows controlled recalibration or retraining. Fourth, it provides a time-ordered evaluation protocol, baselines, ablations, and reporting requirements that can be applied to an authorized lender dataset.

B. Scope and target definition

For each borrower-account observation at month t , let $Y(t,h)=1$ indicate a pre-agreed stress event within h months. One

defensible target is the first transition from a performing account to a lender-defined early-delinquency event (for example, 30+ days past due), while regulatory SMA status may be included as a distinct secondary target. Definitions, cure rules, restructuring, write-offs, and competing exits must be fixed before modeling. A stress event is not synonymous with default, fraud, or insolvency.

II. RELATED WORK

A. Credit-risk prediction for smaller firms

Benchmark comparisons of credit-scoring classifiers show that model rankings depend on the dataset, sample design, and metric; no single model is uniformly superior [4]. For bankruptcy prediction, machine-learning approaches have been compared with statistical baselines across multiple datasets [5]. Recent work directly studies small-business credit risk and finds that advanced methods can be useful when lenders have limited information, but also stresses the need to interpret prediction drivers [8]. These studies motivate strong but controlled baselines: logistic regression for transparency, gradient boosting for nonlinear tabular effects, and a lender's existing score when available.

B. Temporal change and adaptation

Concept drift denotes change in the relationship between features and outcomes, while covariate shift changes the feature distribution. Surveys distinguish detection, adaptation, and recurring-drift settings [6]. Adaptive-window methods such as ADWIN provide a general mechanism for detecting changes in a monitored stream [7]. Financial-distress work has also considered dynamic prediction under concept drift [9]. However, a warning system must account for outcome-label delay: a risk estimate emitted today may not be validated until the observation horizon closes. Consequently, immediate feature-distribution alarms and later outcome-performance alarms should be reported separately.

C. Calibration, selective prediction, and review capacity

Probability calibration matters when scores guide decisions with asymmetric costs; calibration methods such as beta calibration can adjust classifier outputs using a held-out temporal calibration set [10]. Selective classification permits a predictor to abstain on uncertain examples. Pugnana and Ruggieri formulate this trade-off for ranking metrics such as AUC [11]. For a lender, however, the relevant question is not merely global discrimination: the number of accounts receiving manual attention is capped. This work combines top-budget retrieval with calibrated probabilities and an uncertainty-based review route, and evaluates coverage, precision, recall, calibration, and workload together.

D. Research gap

Prior work supplies components—credit scoring, drift detection, calibration, and selective prediction—but does not itself establish that the proposed combination improves MSME outcomes. The gap addressed here is a unified experimental design that binds prediction quality to a declared review

capacity, tests chronological generalization, and makes uncertainty and drift part of the operational output. The framework is a research proposal; its superiority remains an empirical question.

III. METHODOLOGY

A. Data unit, labels, and point-in-time features

The unit of analysis is a borrower-account-month snapshot. A suitable study requires de-identified monthly snapshots over enough calendar time to include multiple economic and lending conditions, with a unique borrower key for grouped splitting. Candidate point-in-time features include days past due and payment transitions; utilization and limit changes; repayment regularity; returned payments; cash-flow or statement summaries where lawful; account age; exposure and repayment burden; sector, geography, enterprise-size class, and product; and prior restructuring or relief events. Features must be available as of the prediction date. Post-event flags, later statements, and retrospectively updated values are excluded to prevent leakage.

The primary horizon h should be chosen with credit operations (e.g., 3 or 6 months) and frozen before analysis. Overlapping snapshots require borrower-grouped confidence intervals or non-overlapping landmark windows. Censored observations are excluded or handled with a pre-specified survival method. Missingness indicators are retained where absence itself may carry process information; imputation parameters are fitted on training folds only. Personally identifying attributes are excluded from predictors and protected in a governed audit dataset.

B. Predictive and uncertainty layers

Let $x(i,t)$ denote the point-in-time feature vector for account i and let $f\theta(x)$ produce a raw risk score. The baseline is regularized logistic regression; the nonlinear candidate is gradient-boosted trees. Both are trained with identical time splits and feature availability. Class weighting may be selected within training data, but the reported probabilities are recalibrated on a later, disjoint calibration window. Beta calibration is a candidate map $g\phi(f\theta(x))$ yielding $p(i,t)=g\phi(f\theta(x))$ [10]. Calibration is evaluated without resampling the test period.

Uncertainty is estimated using a small ensemble of models fit on temporally blocked bootstrap samples. Let $\bar{p}$ be the ensemble mean and u the dispersion of member probabilities. A separate shift score $d(x)$ can combine population stability or a distance-to-training measure. The abstention score $q=\lambda u+(1-\lambda)d$ is standardized on the calibration window. Accounts above a pre-registered uncertainty threshold are marked “review—low confidence”; they are not silently assigned a trustworthy rank. As a sensitivity analysis, compare ensemble dispersion with a conformal or selective-prediction strategy [11], [12].

C. Drift-aware monitoring

Two monitoring channels are maintained. An unlabeled channel tracks feature and score distributions monthly using population

stability index, Jensen–Shannon divergence, and segment-specific missingness. A labeled channel, activated only when outcomes mature, tracks calibration error, Brier score, log loss, and top-B recall on the latest completed cohort. An ADWIN-style detector may be applied to a bounded stream of loss or calibration residuals [7]. Thresholds are set using training/validation history and an explicit false-alarm policy; they are not tuned on final test outcomes.

A detected shift triggers a graduated response: (i) verify data pipeline and coding; (ii) inspect affected segments and explanations; (iii) increase manual review for low-confidence cases; (iv) recalibrate if ranking remains stable but probabilities drift; or (v) retrain after sufficient mature labels and independent validation. Each model version, alarm, override, and decision is logged. Retraining is not automatic upon a single population-shift alarm, since benign seasonality may alter feature distributions without changing event risk.

D. Budgeted ranking and decision rule

Let $c(i)$ be the estimated review cost and $L(i)$ the avoidable loss if an actionable stress event is missed. For a fixed human-review budget K , define priority $s(i)=p(i)L(i)/c(i)$. If reliable account-level loss estimates are unavailable, rank by calibrated $p(i)$ and report that this is a proxy. Select the top K eligible accounts, subject to operational exclusions and a protected uncertainty route. Report results for budgets $B \in \{1\%, 2\%, 5\%, 10\%\}$ of active accounts or the lender's actual capacity. To avoid double-counting, uncertain accounts consume the same review budget unless the lender has a separate specialist queue.

E. Experimental design and measures

Use expanding-window validation: train on earlier periods, calibrate on the next period, and test on the subsequent period; roll this sequence forward. The final period remains untouched until design choices are frozen. Baselines are (1) current policy score, if available; (2) logistic regression; (3) static gradient boosting; (4) recency-weighted boosting without drift alarms; and (5) the full framework. Include ablations removing calibration, uncertainty routing, or drift response. Handle imbalance through training-only weighting; do not use random oversampling across time.

For a test cohort, let N be the eligible accounts, P the observed stress events, K the reviewed accounts, S_K the selected top- K set, $y_i \in \{0,1\}$ the event label, and p_i the calibrated predicted probability. Let $TP_K = \sum_{i \in S_K} y_i$ and prevalence $\pi = P/N$. The primary budgeted ranking measures are:

$$\text{Precision@K} = TP_K / K \quad (1)$$

$$\text{Recall@K} = TP_K / P \quad (2)$$

$$\text{Lift@K} = \text{Precision@K} / \pi, \quad \text{where } \pi = P / N \quad (3)$$

Probability quality is assessed on all eligible test accounts, not only the reviewed queue. The Brier score and binary log loss are:

$$\text{Brier} = (1/N) \sum_{i=1}^N (p_i - y_i)^2 \quad (4)$$

$$\text{LogLoss} = -(1/N) \sum_{i=1}^N [y_i \ln(p_i) + (1-y_i) \ln(1-p_i)] \quad (5)$$

For expected calibration error, divide predictions into M bins. In bin b , n_b is the account count, $\text{acc}(b)$ the observed event fraction, and $\text{conf}(b)$ the mean predicted probability:

$$\text{ECE} = \sum_{b=1}^M (n_b/N) |\text{acc}(b) - \text{conf}(b)| \quad (6)$$

If validated realized loss L_i is available for each event, loss capture at budget K is the share of total realized event loss among the reviewed accounts. It measures ranking of observed loss and does not by itself estimate the causal benefit of intervention:

$$\text{LossCapture@K} = [\sum_{i \in S_K} y_i L_i] / [\sum_{i=1}^N y_i L_i] \quad (7)$$

Calibration slope and intercept are estimated by logistic regression of outcomes on the logit of predicted probabilities; ideal values are intercept $\alpha=0$ and slope $\beta=1$:

$$\text{logit Pr}(Y_i=1 | p_i) = \alpha + \beta \text{logit}(p_i) \quad (8)$$

Secondary measures include PR-AUC, ROC-AUC, abstention rate and model coverage, false alarms per month, and average workload. Confidence intervals are clustered by borrower and resampled by test month; paired temporal folds are compared. Report the positive-event count and metric denominator for every test window. Do not claim improvement from a single split.

F. Governance and safeguards

The system is a prioritization aid. A trained officer checks underlying facts and contacts borrowers through existing policy before corrective action. No adverse credit action is made solely from the score. Validate performance by enterprise size, geography, sector, gender of proprietor where lawful and appropriate, and other permitted groups; investigate disparities and missing-data effects. Restrict access, retain only necessary features, document purpose and retention, and obtain lender legal, compliance, and model-risk approval before deployment. Data minimization and applicable privacy, banking, and fair-lending requirements remain controlling.

G. Reproducible training and queue construction

The training pipeline follows a strict temporal order. First, an eligible snapshot date is selected and the outcome window is checked for complete follow-up. Second, all preprocessing is fitted using the training interval only. Third, candidate models are selected using an inner rolling validation loop. Fourth, the selected model is refit on the permitted training interval, calibrated on a chronologically later calibration interval, and frozen before testing. Fifth, each account receives a probability, uncertainty value, drift flag, and reason codes. Finally, eligible accounts are ordered by priority and the top K are selected. The order and tie-breaking rule are deterministic, with ties resolved using a stable account hash rather than an outcome-related field.

To prevent train-test contamination, duplicate monthly observations for a borrower must not be treated as independent random examples. A borrower may appear in multiple historical training months, but test months must occur strictly after training and calibration dates. Where the deployment question concerns unseen firms, a separate borrower-disjoint sensitivity test is required. Where it concerns monitoring existing customers, chronological borrower overlap is appropriate,

provided features are point-in-time correct. These two estimands answer different operational questions and should not be conflated.

H. Feature governance and explanation

The initial feature set should favor stable, interpretable aggregates over opaque high-frequency signals. For transaction or cash-flow data, candidate summaries include rolling inflow/outflow trend, inflow volatility, concentration of top payers, and the ratio of scheduled debt service to observed operating inflow. These variables require lender validation and must not be derived from post-cutoff transactions. Categorical sector or region values should be grouped according to a documented taxonomy, and rare categories should be pooled using training data only. Every feature needs a name, definition, source system, frequency, availability timestamp, missingness rate, owner, and permitted use.

For each alert, the interface should expose the event horizon, estimated probability, uncertainty band or flag, main contributing factors, recent change from the account's own history, and recommended next review step. SHAP-style feature attributions may summarize model behavior, but they do not establish causality and should be checked for stability across time. A reason code should map to an observable record or ratio so that an officer can verify it with the borrower. Model explanations must not be presented as a definitive account of why a firm is distressed.

I. Operational interpretation of the budget

A review budget may be stated as a daily or monthly capacity rather than a fixed percentage. In that case, K is the number of accounts that the operations team can investigate within the decision window. If review costs vary substantially, a knapsack formulation can select accounts maximizing expected value subject to total review hours H : maximize the sum of $p(i)L(i)z(i)$, subject to the sum of $c(i)z(i) \leq H$ and $z(i) \in \{0,1\}$. This optimization should only be used when loss and intervention-effect estimates are defensible. Otherwise, a transparent top- K ranking is preferable and easier to audit.

The queue should preserve a small exploration component only if approved by the lender's model-risk process, so that low-ranked accounts can be sampled for outcome and fairness monitoring. That sample is excluded from the operational top- K outcome comparison or separately accounted for. Officer overrides should record the pre-review score, override direction, reason, evidence considered, and eventual outcome. These records support future validation but should not be used as labels without careful treatment of selection bias: reviewed borrowers receive attention precisely because the model ranked them highly.

J. Metric definitions

The notation in (1)–(8) is used consistently for every test window and review budget. For selective prediction, let U be the set routed to abstention and M the number of tested months. Let FP_t denote false-positive alerts in month t , and let c_i be

the recorded review cost for selected account i . Report the abstention and monthly alert burden as:

$$\text{AbstentionRate} = |U| / N; \quad \text{ModelCoverage} = 1 - |U| / N \quad (9)$$

$$\text{EventRate}_U = (\sum_{i \in U} y_i) / |U| \quad (10)$$

$$\text{FalseAlarmsPerMonth} = (\sum_{t=1}^M FP_t) / M \quad (11)$$

$$\text{MeanReviewCost} = (\sum_{i \in S_k} c_i) / M \quad (12)$$

PR-AUC summarizes the precision–recall curve over thresholds; it must be reported with event prevalence because its baseline depends on prevalence:

$$\text{PR-AUC} = \int_0^1 \text{Precision}(r) dr \quad (13)$$

For expected calibration error, disclose the number and construction of bins and provide a reliability plot; ECE changes with binning. Calibration slope and intercept are estimated from logit outcomes versus logit predictions as in (8); the ideal values are $\alpha=0$ and $\beta=1$. Confidence intervals are computed by resampling test months and borrowers within months, preserving temporal clustering as far as the sample allows. Report event counts and denominators for every segment. These metrics are diagnostics; thresholds and model choices must not be retuned on the final test set.

K. Deployment gates

Before a pilot, the lender should define the minimum useful recall or lift at capacity, the maximum acceptable calibration error, the review response time, the policy for uncertainty cases, and the rollback trigger. Shadow testing should compare model queues with existing processes without changing borrower treatment. A limited pilot can then assess alert precision, time-to-contact, officer workload, override reasons, false-alert burden, and whether interventions lead to appropriate resolution. A model that ranks well but creates unmanageable workload, unstable alerts, or unexplained segment disparities fails the deployment objective.

L. Target construction and observation bias

The event definition controls the scientific meaning of every reported metric. The primary label should be derived from servicing records using an auditable rule and a frozen observation cutoff. If the lender chooses 30+ days past due as an outcome, the label should identify the first qualifying event and distinguish it from technical reversals or data corrections. SMA status, restructuring, missed covenant, and NPA status can be evaluated as separate outcomes rather than collapsed into an ambiguous “stress” label. The analysis should also state whether accounts that close, refinance, are sold, or enter restructuring before horizon h are censored, treated as events, or treated as competing risks.

Historical review and intervention create selective labels. Accounts previously contacted by officers may have had stress prevented, while unreviewed accounts may lack equally detailed records. A model trained naively on such data can learn past review policy as if it were underlying borrower risk. The study should therefore document historical policy changes and review coverage, stratify results by prior intervention, and, where feasible, use a prospective shadow period or randomized audit sample to observe outcomes across risk strata. Any

inverse-propensity or causal correction should be reported as a sensitivity analysis with its assumptions, not as a substitute for better data collection.

M. Training objective and imbalance

The primary objective is ranking future events rather than maximizing overall accuracy. For the logistic baseline, optimize regularized binary cross-entropy on training data. For boosted trees, use a proper probabilistic loss or a ranking objective followed by independent calibration. Class weights are estimated from training prevalence and fixed before each evaluation window. Synthetic minority oversampling is not applied to the test data and is avoided for temporal sequences unless a leakage-safe design is justified. Threshold selection is performed only on training/validation periods and is subordinate to the operational top-K policy.

When stress events are rare, accuracy can remain high even if nearly all events are missed. The paper therefore reports prevalence next to PR-AUC and top-K retrieval. In addition to pooled metrics, performance is tabulated by calendar quarter and major borrower segments. A model with high pooled recall but a complete failure in one sector should not be described as robust. The number of positive outcomes in each cell is always shown; small cells receive intervals and are not overinterpreted.

N. Choosing and validating drift signals

Feature drift is not necessarily predictive drift. A rise in sales volatility, for example, may be seasonal and harmless, while a stable feature distribution can coexist with a changed relationship between cash-flow decline and delinquency. For this reason, the unlabeled monitor is an early diagnostic and the mature-label monitor is the performance test. Each signal has a reference window, update frequency, minimum sample size, severity threshold, and named owner. Reference windows are refreshed only under documented governance to avoid gradually redefining abnormality as normal.

A drift study should use historically held-out periods containing naturally occurring changes where possible. If these are insufficient, controlled perturbations can test pipeline sensitivity (for example, changes in missingness or scale), but simulated perturbations must be reported separately from observed market-period results. Report detection delay, false-alarm rate, missed material degradation, and the time needed for a model update to restore calibration. These measures make drift detection comparable as an operational mechanism rather than a binary feature checkbox.

O. Uncertainty and manual routing

Ensemble disagreement is a practical proxy for epistemic uncertainty, not a guarantee of correctness. A model may be confidently wrong under a new regime, so uncertainty is interpreted together with drift and data-quality checks. The system should distinguish at least three states: routine ranked alert, high predicted risk with usable confidence, and low-confidence or out-of-support observation. The third state calls

for information verification or specialist review; it should not automatically be treated as either safe or distressed.

The abstention threshold is selected on a calibration period to meet a pre-declared workload and risk-coverage target. Risk-coverage curves plot error or event capture against the fraction of cases assigned a model-led ranking. Results should be given for multiple thresholds and compared with random abstention. Where the institution needs a single prioritized queue, low-confidence cases can be placed in a reserved review band and their capacity cost counted explicitly. This prevents uncertainty routing from appearing to improve predictive performance merely by removing difficult cases from the denominator.

P. Reporting reproducibility

For replication, the study record should include data cut dates, eligibility criteria, event prevalence, horizon, feature-generation code version, missingness treatment, model hyperparameters, random seeds, calibration method, drift thresholds, review budgets, and metric definitions. The borrower-level data need not be published; a data dictionary, synthetic schema, and executable evaluation code can still be shared where policy permits. Model artifacts and code should be versioned so that a reviewer can reproduce each rolling split and each reported queue.

The manuscript should disclose funding, lender participation, conflicts of interest, and whether authors had access to raw servicing data. Results should distinguish model performance from intervention impact: a high-risk list can identify accounts likely to deteriorate, but only a controlled operational study can show whether outreach improves repayment or business continuity. The evaluation should also track whether false alerts cause unnecessary borrower burden, and whether missed alerts delay access to restructuring or support.

Q. End-to-end alert procedure

The monthly scoring cycle is specified as follows. Input: eligible active accounts $A(t)$, fitted model f_θ , calibration map g_ϕ , uncertainty threshold τ_u , drift policy D , and review capacity K . (1) Validate the monthly extract against schema, freshness, duplicate, and missingness checks; quarantine invalid records. (2) Build point-in-time features using only information timestamped on or before t . (3) Compute calibrated event probabilities $p(i,t)$, ensemble dispersion $u(i,t)$, and feature-shift score $d(i,t)$. (4) Apply the drift policy: if population shift is material, preserve the prior model version, flag affected accounts, and route low-confidence observations for verification; do not silently replace the model. (5) Compute priority $s(i)=p(i,t)L(i)/c(i)$, or $p(i,t)$ when validated L and c are unavailable. (6) Create the top-K queue subject to eligibility and review rules, counting uncertainty referrals against the stated capacity. (7) Attach the event horizon, uncertainty state, reason codes, and source timestamps to each alert. (8) Store the scored cohort, model version, selected queue, exclusions, and review outcomes in an immutable audit log. (9) When labels mature, compute monthly and subgroup metrics and evaluate whether an approved recalibration or retraining gate has been reached.

The procedure deliberately separates a warning from an intervention. A high score causes a timely review under the lender’s existing process; the officer confirms data, asks about relevant business conditions, and documents a proportionate next step. A low score does not certify that the borrower is safe. If the lender learns that the account data are stale or incomplete, the case is marked for data repair and is not presented as a confident model assessment. Queue-level monitoring should report the fraction of records excluded, delayed, manually overridden, or routed as uncertain, as these operational outcomes affect the effective review capacity.

IV. RESULTS

A. Evidence status

No anonymized MSME account-month dataset, feature dictionary, label definition, review-cost schedule, or executed model outputs were supplied for this manuscript. Therefore, there are no empirical performance results to report. In particular, no AUC, recall, precision, calibration, workload reduction, or loss-savings figure is claimed. A made-up numerical comparison would not be scientifically valid. The material below defines the result tables and tests that should be completed when authorized data become available.

B. Required reporting table

Evaluation item	Required reporting
Temporal holdout	Dates; borrower and event counts; prevalence
Budgeted retrieval	Precision@K, recall@K, and lift@K for each K
Probability quality	Brier score, log loss, ECE, calibration slope
Uncertainty route	Abstention rate and event rate among abstained cases
Drift response	Alarm dates, false alarms, recovery after recalibration
Subgroup analysis	Metrics and uncertainty intervals by segment

C. Hypotheses and acceptance criteria

H1: At the same K, the full model captures more future stress events than the current-policy and static-model baselines on multiple chronological test windows. H2: recalibration reduces Brier score and calibration error after observed prevalence or covariate change without materially reducing ranking quality. H3: uncertainty routing concentrates a higher event rate among abstained cases than the overall portfolio, while keeping the additional review load within budget. H4: the drift-response policy reduces the duration of material calibration degradation compared with a static model. Each hypothesis is evaluated with paired, month-clustered intervals and a pre-specified minimum operational effect chosen with the lender. Statistical significance alone is insufficient.

A result table should include point estimates and 95% intervals, sample counts, number of stress events, test months, and the exact denominator used for each metric. For instance, recall@K is the share of all observed events captured among the K reviewed accounts; precision@K is the event proportion among those K. PR-AUC is prevalence-sensitive and should be read with the event rate. Expected loss captured requires validated

loss and intervention-effect assumptions; absent these, it must not be presented as monetary benefit.

D. Robustness and error analysis

The final evaluation should vary the stress horizon, budget, calibration window, uncertainty threshold, and drift threshold. Examine false negatives by subsector and event lead time, and false positives that arise from seasonal working-capital needs or one-off payment delays. Conduct a temporal permutation check and a label-leakage audit. Compare predicted-risk distributions before and after each shift alarm, and assess whether officer overrides add useful information. If results are unstable across months, the correct conclusion is that the model is not yet reliable for operational use.

V. CONCLUSION

This paper presents a budgeted early-warning framework for prioritizing MSME credit accounts under changing borrower and economic conditions. Its design joins calibrated short-horizon risk estimation, uncertainty-based abstention, drift monitoring, delayed-label adaptation, and top-budget evaluation. The framework is intended to help lenders direct limited human review toward accounts that may need timely diagnosis, while preserving officer judgement and existing regulatory processes.

The central research claim is deliberately testable: drift-aware ranking and uncertainty estimation may improve event capture and probability reliability at a fixed review budget, but this must be demonstrated on chronological, lender-authorized data. As no such data or model runs were supplied, this manuscript reports a protocol rather than empirical findings. Future work should evaluate the proposed system across lenders, products, and economic periods; estimate intervention effects; and assess borrower outcomes such as successful rectification, restructuring, and sustainable credit access. Only after these tests and governance review should operational deployment be considered.

REFERENCES

[1] Reserve Bank of India, “Framework for Revival and Rehabilitation of Micro, Small and Medium Enterprises (MSMEs),” FIDD.MSME & NFS.BC.No.21/06.02.31/2015-16, Mar. 17, 2016.

[2] Reserve Bank of India, “Frequently Asked Questions: Micro, Small and Medium Enterprises,” sections on revival and rehabilitation and identification of incipient stress. [Online]. Available: <https://www.rbi.org.in/commonperson/English/Scripts/FAQs.aspx?Id=966>. Accessed: Oct. 16, 2026.

[3] Reserve Bank of India, Financial Stability Report, Jun. 2025, ch. I, section on MSME credit and asset quality.

[4] 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.

[5] F. Barboza, H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction,” Expert Syst. Appl., vol. 83, pp. 405–417, 2017, doi: 10.1016/j.eswa.2017.04.006.

[6] 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, pp. 1–37, 2014, doi: 10.1145/2523813.

[7] A. Bifet and R. Gavaldà, “Learning from time-changing data with adaptive windowing,” in Proc. SIAM Int. Conf. Data Mining, 2007, pp. 443–448.

[8] A. Bitetto, P. Cerchietto, S. Filomeni, A. Tanda, and B. Tarantino, “Can we trust machine learning to predict the credit risk of small businesses?” Rev.

Quant. Finance Account., vol. 63, pp. 925–954, 2024, doi: 10.1007/s11156-024-01278-0.

- [9] J. Sun and H. Li, “Dynamic financial distress prediction using instance selection for the disposal of concept drift,” *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4272–4280, 2011.
- [10] M. Kull, T. Silva Filho, and P. Flach, “Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers,” in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, PMLR, vol. 54, 2017, pp. 623–631.
- [11] A. Pugnana and S. Ruggieri, “AUC-based selective classification,” in *Proc. 26th Int. Conf. Artif. Intell. Stat. (AISTATS)*, PMLR, vol. 206, 2023, pp. 2494–2514.
- [12] Y. Romano, E. Patterson, and E. C. Candès, “Conformalized quantile regression,” in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.