INTERNATIONAL JOURNAL OF ACTA INFORMATICA
VOLUME (2026)

DRIFT-RESILIENT EARLY-WARNING MODELS FOR SME LOAN DISTRESS: A CROSS-INSTITUTION VALIDATION FRAMEWORK

Zeynep Grigoryan[1*], Elif Ani [2]
[1] Yerevan State University, ARMENIA
[2] Baku State University, AZERBAIJAN

ABSTRACT

Early-warning models for small and medium-sized enterprise (SME) loan distress are usually evaluated within a single lending institution, where apparently strong discrimination may reflect local portfolio composition, underwriting policy, and data-definition choices rather than risk relationships that transfer elsewhere. This article develops a multi-institution validation framework for drift-resilient SME early-warning models, building on the multi-institution architecture that Rashid and Sattiraju propose for SME loan distress and transaction-fraud pilots across U.S. community financial institutions, and applies it to public data. The framework separates two questions that single-site evaluation conflates, temporal stability within an institution and transfer to an institution the model has never seen, and specifies harmonized outcomes, leakage-safe partitions within each site, staged cross-site transfer tests, unsupervised drift detection with D3 and OCDD, and site-level reporting. It is implemented on 375,261 SBA-guaranteed small-business loans, with eight industry-sector portfolios standing in for institutions because the public file identifies neither lender nor approval date. Transferred models ranked borrowers as well as site-specific ones (leave-one-site-out AUC 0.68 to 0.79), but calibration did not transfer: predicted charge-off rates were off by up to 8.4 percentage points, and between-site heterogeneity was high ($I^2 = 0.96$). Local recalibration on about 1,000 labelled loans cut the calibration error by roughly 72 percent, whereas 100 labels made it worse. At their default thresholds, D3 and OCDD rarely flagged the cross-site shifts, and no detector statistic tracked the performance cost of transfer. The results support site-level reporting and labelled local recalibration over label-free drift alarms. Because sector portfolios are not institutions and no approval dates were available, the evaluation does not test temporal stability or true institutional transfer.

KEYWORDS: SME Lending; Loan Distress; Early-Warning Models; Multi-Institution Validation; Concept Drift; Calibration; Temporal Validation; Model Governance

INTRODUCTION

Small and medium-sized enterprises can move from apparent stability into financial distress well before conventional accounting indicators, such as a missed covenant test or a formally reported delinquency, fully reveal that deterioration. An early-warning model attempts to close that gap by identifying risk earlier, from signals such as payment behavior, cash-flow volatility, account activity, and other indicators a lender routinely observes in the ordinary course of

servicing a loan. The appeal of such a model is operational as much as statistical: an institution that can flag a distressed relationship months before a missed payment has a genuine opportunity to intervene, whereas a model that only confirms distress after it has already become visible in the account record offers little beyond what a credit officer would have noticed regardless.

The difficulty is that a model trained and validated at a single financial institution is, almost by construction, trained on that institution's own portfolio composition, underwriting policy, data definitions, collections practices, and prevailing local economic conditions. A model that discriminates well on a held-out sample drawn from the same institution has demonstrated, at most, that it fits patterns present in that one lending environment; it has not demonstrated that those patterns reflect a genuine, transferable relationship between early indicators and eventual distress, as opposed to an artifact of how that particular institution originates, services, and labels its loans. The two possibilities are observationally similar from inside a single-site evaluation and only become distinguishable once a model's predictions are confronted with a second institution's data.

Rashid and Sattiraju[1] address exactly this problem by proposing a multi-institution validation architecture for drift-resilient early-warning models that spans both SME loan distress and transaction fraud, piloted across a sample of U.S. community financial institutions. Their contribution emphasizes the need to distinguish genuine generalization from institution-specific historical fit, and it is grounded in unsupervised drift-detection methods that do not require immediate access to confirmed outcome labels, a practical necessity given how slowly SME distress labels typically mature relative to the pace at which an institution's portfolio and operating environment can change. This article develops that theme into a fuller methodological framework specifically for validating SME distress models across both institutions and time, and it is written without assuming that the architecture Rashid and Sattiraju propose has already demonstrated universal effectiveness; their pilot evidence motivates the problem this article addresses, but the validation design developed here is offered as a methodological proposal in its own right, to be tested rather than taken as already confirmed.

RESEARCH OBJECTIVES

This article pursues four linked objectives. First, it distinguishes temporal validation, whether a model continues to perform acceptably as time passes within the institution that trained it, from institutional validation, whether the relationships the model has learned transfer to an institution it has never seen; the two are frequently collapsed into a single pooled evaluation in practice, with the result that neither question is actually answered cleanly. Second, it identifies the principal sources of cross-site distribution shift that a validation design must anticipate, including differences in portfolio composition, underwriting standards, outcome definitions, and data availability. Third, it proposes a leakage-safe evaluation design that respects chronological ordering within each participating institution while still supporting a meaningful cross-site comparison. Fourth, it defines a set of metrics for early-warning usefulness that go beyond ordinary classification performance to capture lead time, calibration, and cross-site heterogeneity. Running alongside these four objectives is a fifth, necessarily practical concern: the governance and data-sharing arrangements that determine whether cross-institution

evidence, once produced, can actually be interpreted, audited, and trusted by the institutions that contributed it.

LITERATURE REVIEW

SME DISTRESS AS AN EARLY-WARNING PROBLEM

SME financial distress is not a single, cleanly defined event in the way that, say, a bond default with a contractual trigger might be. Depending on the study and the institution, a model may target missed payments, covenant breaches, a formal restructuring, delinquency past a given number of days, or a defined default outcome, and these candidate outcomes differ meaningfully in both severity and timing relative to the underlying deterioration they are meant to signal. A credible early-warning study must therefore specify, before any model is fit, the prediction horizon over which an alert is meant to anticipate the outcome, the point in time at which the predictors used to generate that alert actually become available to the institution, and the precise definition of the event itself. Without this specification, performance estimates reported by two institutions, or even by two studies at the same institution conducted a few years apart, may simply not be comparable, however similar the headline accuracy or recall figures appear.

WHY INSTITUTION-LEVEL VALIDATION MATTERS

Financial institutions differ from one another along dimensions that bear directly on whether a learned risk relationship can be expected to transfer: portfolio composition, geography, sector exposure, the specific lending products offered, underwriting standards, collections practices, and the structure and quality of the underlying data systems. A model trained at a large, geographically diversified bank may rely, often implicitly, on predictor relationships that simply do not hold, or do not hold with the same strength, at a community institution with a different customer relationship model, a narrower sectoral footprint, and more limited data infrastructure. The multi-institution proposal advanced by Rashid and Sattiraju[1] makes this generalization problem explicit rather than leaving it as an unstated assumption behind a single-site study, and it is precisely this explicitness that the present framework seeks to operationalize into a repeatable validation protocol: institution-level validation should separate within-institution temporal stability, the question of whether a model trained at a given site remains useful there over time, from cross-institution transfer, the logically distinct question of whether that same model is useful anywhere else at all.

DRIFT AND ADAPTATION

Gama *et al.*[2] describe concept drift as a change over time in the relationship between predictors and outcomes, distinct from a mere change in the marginal distribution of the predictors themselves, and this distinction carries directly into the SME lending context. Drift in SME distress risk may follow macroeconomic conditions, interest-rate changes, supply-chain disruptions, sector-specific policy shifts, or a gradual change in the composition of an institution's own SME book as it grows or repositions its lending strategy. A shift in the distribution of a given predictor, such as average account balance or transaction frequency, does not necessarily imply any change in the conditional risk of distress given that predictor; conversely, conditional risk can change even while the marginal distributions of the predictors

an institution monitors remain superficially stable, which is precisely why monitoring the input distribution alone is an incomplete strategy. A monitoring layer able to flag such a change as early as possible, and without waiting for a large volume of confirmed distress outcomes to accumulate, benefits from an unsupervised detection method that compares feature or score distributions directly; Zuberi *et al.*[3] demonstrate the practical value of exactly this kind of leakage-safe, chronologically evaluated drift-detection procedure in a closely related financial-fraud setting, and the evaluation protocol developed later in this article implements the same chronological-splitting discipline for the SME distress case. Adaptive-windowing methods, in which the effective size of the monitoring window is automatically reduced once a change-detection statistic crosses a calibrated threshold, offer one concrete mechanism for operationalizing this kind of continuous drift sensing without committing to a fixed lookback period in advance[4].

IMBALANCED OUTCOMES AND EARLY-WARNING METRICS

Severe distress events are typically rare relative to the full population of performing SME loans, so an aggregate accuracy figure can conceal a large share of missed cases simply by favoring the majority class. Precision and recall should therefore be reported at an operationally meaningful decision threshold rather than left implicit inside a single summary number, and the area under the precision-recall curve is generally more informative than the area under the receiver operating characteristic curve once the event of interest is this rare, because a large population of non-distressed accounts can make a model's false-positive rate look reassuringly small in relative terms even when the absolute number of false positives it implies is operationally unmanageable[5]. Fawcett's foundational treatment of ROC analysis is itself explicit that the curve's apparent robustness to class skew can be misleading in exactly this way, which is one reason a precision-recall view is preferred as a complement rather than a substitute[6]; Hand's critique of the area under the ROC curve as a single operating-point-free summary adds a further caution, since that statistic implicitly averages performance over a range of relative misclassification costs that may bear no resemblance to the cost structure an SME lender actually faces[7]. The Matthews correlation coefficient offers a useful complementary summary precisely because it incorporates all four confusion-matrix cells into one balanced figure that is considerably less distorted by severe class imbalance than accuracy or even F1[8]. Resampling techniques such as the Synthetic Minority Over-sampling Technique are a standard response to this imbalance during model training, generating synthetic minority-class examples by interpolating between confirmed distress cases and their nearest neighbors in feature space[9], though, as with fraud detection, a resampling strategy calibrated once at training time can itself become a source of staleness if the distribution of genuine distress cases drifts without the resampling step being revisited alongside the model it feeds. Beyond ordinary classification quality, an early-warning model also needs a lead-time measure: identifying distress several months before the event may be considerably more useful to an institution than identifying it only shortly before the outcome is confirmed, even though both would count identically as a true positive under a conventional confusion matrix.

PROPOSED MULTI-INSTITUTION VALIDATION ARCHITECTURE

The architecture proposed here uses institution-specific data preparation conducted under a shared analysis protocol, a harmonized outcome definition, temporal holdouts computed independently within each participating institution, a staged sequence of cross-institution transfer tests, and a central analysis plan that governs how results from different sites are combined and reported. Data is not required to be physically pooled: where regulation or privacy constraints prevent that, aggregate metric exchange or a federated evaluation procedure may support a valid cross-site comparison, but only on the condition that metric definitions, outcome definitions, and data-quality checks are harmonized across sites before any comparison is attempted; without that harmonization, an apparent difference in performance between two institutions may simply reflect a difference in how each institution defines or records the outcome being predicted.

OUTCOME AND HORIZON HARMONIZATION

Before any cross-institution model comparison is attempted, participating institutions should agree explicitly on the target event, the event date, the observation window over which predictors are computed, the prediction horizon, applicable censoring rules, and the label-maturity period required before an outcome is treated as confirmed. Differences in how delinquency, restructuring, or default are recorded operationally can produce what look like genuine model-performance differences across sites but are, in fact, entirely attributable to differences in outcome definition rather than to any difference in predictive validity. A shared data dictionary, documenting the meaning, units, availability, and missingness pattern of every feature used in the comparison, is therefore not a bureaucratic formality but a methodological precondition for any claim about cross-site transfer to be interpretable at all.

WITHIN-SITE TEMPORAL VALIDATION

Each participating institution should partition its own observations chronologically into training, validation, and test windows, written here as $D_{train,s} \prec D_{validation,s} \prec D_{test,s}$ for each institution s, where the ordering symbol denotes strict chronological precedence. All feature engineering, imputation, scaling, feature selection, and decision-threshold tuning must use only data that would genuinely have been available before the boundary of the test window, mirroring the leakage-safe chronological discipline that Zuberi *et al.* demonstrate for drift evaluation in the closely related fraud-detection setting[3]. Rolling-origin evaluation, in which the training and test windows are advanced repeatedly through the available history rather than fixed at a single split, should be used whenever an institution's history is long enough to support it, because a single chronological split answers only whether the model generalizes forward across one particular boundary, while a rolling design tests whether that forward generalization holds up repeatedly across different points in time.

CROSS-SITE EVALUATION

At least three distinct evaluation questions should be kept separate rather than collapsed into a single number: pooled training combined with pooled testing, which estimates performance under a shared mixture of institutions and is the easiest to compute but the least informative about transfer; training at one institution and testing at another, which directly tests

transferability in its strictest form; and local adaptation, in which a limited amount of data from the target institution is used to calibrate or lightly adjust a model trained elsewhere, which tests whether a practical, low-cost adaptation step can bridge genuine site differences without a full retraining cycle. These three designs answer genuinely different questions, and results from them should never be merged into a single average without separately reporting site-level variability, since an encouraging pooled figure can be entirely consistent with the model performing poorly at one or more individual institutions.

DRIFT DETECTION AND DIAGNOSIS

Drift monitoring within this architecture compares feature distributions, model-score distributions, and mature-label performance across both time and institutions, rather than relying on any single one of these views in isolation. D3 and OCDD are identified in Rashid and Sattiraju's source architecture as relevant unsupervised drift-detection approaches for exactly this purpose[1], and their practical value in the SME-distress setting should be assessed against detection delay, false-alarm rate, and downstream performance impact, since the exact performance profile reported in the original fraud-and-distress pilot context need not transfer unchanged to every deployment. A detector's label efficiency, its ability to flag a meaningful change without requiring a large volume of already-confirmed distress outcomes, is particularly important here because institutions frequently cannot obtain confirmed distress labels quickly, a constraint that closely parallels the verification-latency problem documented in credit-card fraud detection, where confirmed outcome labels for a given transaction routinely arrive only after a meaningful delay and the timing of that delay becomes part of the learning problem rather than an incidental detail[10]. Explanation-based diagnostic evidence can usefully complement distributional drift indicators at this stage: comparing the ranked importance that a SHAP-based attribution method assigns to each predictor before and after a monitoring alert gives an analyst concrete, auditable evidence about which input is associated with an observed change in model behavior, rather than leaving the analyst with only an aggregate metric that has moved without any visibility into why[11], although such an explanation shift should, as in the fraud-detection literature, be treated as corroborating evidence to weigh alongside distributional and outcome-based signals rather than as sufficient justification for a model change on its own.

EVALUATION MEASURES

DISCRIMINATION AND CLASSIFICATION

Precision, defined as TP/(TP+FP), and recall, defined as TP/(TP+FN), together with the area under the precision-recall curve, the area under the receiver operating characteristic curve, and the Matthews correlation coefficient, provide complementary rather than redundant views of classification quality, and an evaluation that reports only one of them risks a blind spot the others would have caught. The decision threshold used to compute precision and recall should be selected on validation data according to an explicit, stated operational goal rather than a default value carried over from a different setting. In an early-warning context, the cost of a false alarm, typically unnecessary relationship-manager time and a possible unwarranted tightening of credit terms, must be weighed against the cost of a missed distress case, typically

a late intervention and realized loss exposure, and this trade-off should be documented explicitly in the evaluation report rather than hidden inside a single blended metric. Fawcett's broader caution about what the ROC curve can and cannot reveal under class imbalance[6], together with Hand's argument that the area under that curve implicitly averages over a range of misclassification-cost assumptions that may not match an institution's actual priorities[7], both counsel against treating any single discrimination statistic as sufficient on its own.

LEAD TIME AND INTERVENTION VALUE

Early-warning usefulness should be measured in part by lead time, defined as the interval between an alert's first valid occurrence and the actual event date. This measure should be computed only over correctly identified events, since a lead-time figure averaged across both true and false alerts would conflate a model's timing behavior with its discrimination quality, and it should be reported as a distribution rather than reduced to a single mean, because the practical value of an early-warning system depends heavily on how much of its lead time is concentrated usefully far in advance as opposed to clustered just before the event. A model that generates alerts too close to the event date may deliver little genuine intervention value even when its classification metrics look acceptable in isolation, while a model that alerts very early may trade away some precision and require a staged human-review process rather than an automatic response.

CROSS-SITE HETEROGENEITY

Each participating site's metrics, confidence intervals, event prevalence, data completeness, and portfolio composition should be reported individually rather than only as a pooled summary, because a pooled estimate can conceal materially poor performance at a single, often smaller, institution. Between-site variability should itself be treated as a substantive empirical result that the framework is specifically designed to surface, not as statistical noise to be averaged away. Where only a small number of institutions participate in a given evaluation, inferential claims about the entire sector or about community financial institutions generally should remain appropriately cautious, since a handful of sites cannot support a confident generalization to the full population of similarly sized lenders. The empirical evaluation below gives a concrete example: a pooled AUC of 0.741 summarises site values ranging from 0.681 to 0.792.

CALIBRATION

Risk scores should be calibrated whenever they are interpreted as probabilities or used directly to set an intervention threshold, and calibration quality can differ substantially across institutions because baseline distress prevalence and underwriting practice both differ from site to site; a single global threshold tuned at one institution may therefore be poorly suited to another even when the underlying ranking of relative risk transfers reasonably well. Beyond simple scaling methods, a Dirichlet calibration procedure of the kind Kull and colleagues propose offers a principled way to recover well-calibrated probability estimates when the relationship between raw model output and true outcome probability is more complex than a single scalar adjustment can capture[12], which is a realistic possibility once a model is applied outside the institution where it was originally fit. Any site-specific calibration procedure should use a designated calibration sample, held separate from the sample used to fit the underlying

model, and should be evaluated on a later, untouched period rather than on the same window used to perform the calibration itself.

GOVERNANCE, PRIVACY, AND IMPLEMENTATION

DATA-SHARING MODELS

Cross-institution validation can be organized around centralized analysis, secure data enclaves, or federated computation, and the right choice depends on legal constraints, the sensitivity of the underlying data, each institution's technical capacity, and the degree of reproducibility the evaluation needs to support. Institutions should not assume that a federated computation approach automatically removes privacy risk merely because raw records never leave each site; metadata, intermediate model updates, and even aggregate outputs can still reveal sensitive information under some combinations of model structure and query pattern. Privacy review and access-control design should therefore be built into the study design from the outset rather than addressed only after a data-sharing architecture has already been selected on other grounds.

GOVERNANCE RESPONSIBILITIES

Participating institutions should agree explicitly on model ownership, the authority responsible for approving a model change, incident-reporting obligations, the permitted uses of any shared model or shared evidence, audit-trail requirements, and the procedure by which a site may withdraw from an ongoing evaluation. Model documentation should record data versions, feature definitions, outcome rules, and the precise scope of whatever validation has actually been performed, so that a reader of that documentation can tell exactly what has and has not been tested. A successful cross-site validation result, however strong, does not remove the need for continued local monitoring after deployment, since the governance burden this framework describes does not end once a model has cleared an initial cross-institution evaluation.

OPERATIONAL RESPONSE

A monitoring alert should trigger a defined workflow rather than an ad hoc response: verify the integrity of the underlying data, inspect the specific segments the alert appears to implicate, compare mature-label performance against the alert's implied severity, review the relevant business context, and determine whether the appropriate response is intensified monitoring, recalibration, retraining, or escalation to human review. This structured response is especially important given that the models under discussion support decisions that can materially affect an SME's access to credit; model outputs in this setting should support, rather than automatically replace, lending decisions for which the institution itself remains accountable.

EMPIRICAL EVALUATION ON PUBLIC SBA LOAN DATA

DATA AND ADAPTATIONS

The evaluation uses the public U.S. Small Business Administration loan-level file prepared by Li, Mickel and Taylor[13], which records the outcome of loans guaranteed under the SBA 7(a) programme between 1987 and 2014. A cleaned public extract of 375,261 loans from the eight largest industry sectors was used. The outcome is charge-off (distress) versus paid in full. The public extract does not identify the lender or the approval date, so two adaptations were necessary. First, each sector portfolio (retail trade, professional services, construction,

manufacturing, other services, wholesale trade, health care, and accommodation and food) was treated as a site; these portfolios differ in size, charge-off rate (from 10.5 to 25.0 percent) and loan mix, which is the kind of heterogeneity institutions present, but they are not institutions. Second, partitions within each site were random (60/20/20, stratified) rather than chronological, so this evaluation tests cross-site transfer but not temporal stability.

The early-warning model is a LightGBM classifier[14] using information available at origination: new-business status, the log of the gross amount approved, the SBA-guaranteed share, whether the loan is a revolving line, participation in the LowDoc programme, and whether the amount is a round multiple of USD 10,000. The loan term was deliberately excluded. With term included, within-site and transferred AUCs rose to between 0.96 and 0.98, implausibly high for an origination-time model, which suggests that the recorded term in this extract carries information from after origination. Reporting that sensitivity rather than the inflated figures follows the leakage discipline of Zuberi *et al.*[3].

WITHIN-SITE AND CROSS-SITE PERFORMANCE

Table 1 reports, for each site, discrimination of a model trained on that site and of a model trained on the other seven sites (leave-one-site-out, LOSO), together with LOSO calibration on the site's held-out loans.

Table 1. Site-level results on held-out loans: within-site versus leave-one-site-out (LOSO) transfer.

Sector portfolio (site)	Loans	Charge-off rate	AUC within site	AUC transferred (LOSO)	Mean predicted (LOSO)	Calib. intercept (LOSO)	Calib. slope (LOSO)
Retail trade	81,319	23.7%	0.738	0.739	20.8%	0.13	0.95
Professional services	49,548	20.0%	0.740	0.746	23.7%	-0.07	1.19
Construction	49,511	24.6%	0.722	0.717	21.9%	0.20	1.03
Manufacturing	45,782	16.1%	0.747	0.755	17.2%	-0.20	0.91
Other services	45,001	21.0%	0.733	0.738	20.9%	0.00	0.99
Wholesale trade	35,021	20.3%	0.791	0.792	18.8%	0.40	1.25
Health care	34,718	10.5%	0.759	0.770	18.9%	-0.60	1.17
Accommodation & food	34,361	25.0%	0.675	0.681	22.3%	-0.08	0.77

Two results stand out. First, transfer cost almost nothing in ranking: LOSO AUC ranged from 0.681 to 0.792, against 0.675 to 0.791 within site, and was at least as high in most sites because the transferred models were trained on more loans. Second, calibration transferred poorly. Within-site calibration intercepts lay between -0.17 and -0.03; LOSO intercepts ranged from -0.60 to 0.40. In health care, where the charge-off rate is 10.5 percent, the transferred model predicted 18.9 percent. A ranking-based report would call these models transferable, but a

lender acting on their probabilities, for example to set provisions or watch-list capacity, would be systematically misled at several sites.

Site-level reporting matters as much as the framework claims. The fixed-effect pooled LOSO AUC was 0.741, but Cochran's Q was 176.8 on 7 degrees of freedom and I^2 was 0.96, so almost all of the variation in AUC across sites was real rather than sampling noise. Accommodation and food services had the weakest discrimination (AUC 0.681), a result invisible in the pooled figure.

LOCAL RECALIBRATION AND LABEL EFFICIENCY

Each held-out site was given n of its own labelled loans (n = 100, 250, 1,000 or 5,000; two random draws each) and four strategies were compared: the transferred model unchanged; the transferred model with logistic recalibration on the local labels; a model trained on the local labels only; and a model retrained on the pooled data with the local labels up-weighted. Figure 1 shows the averages across sites.

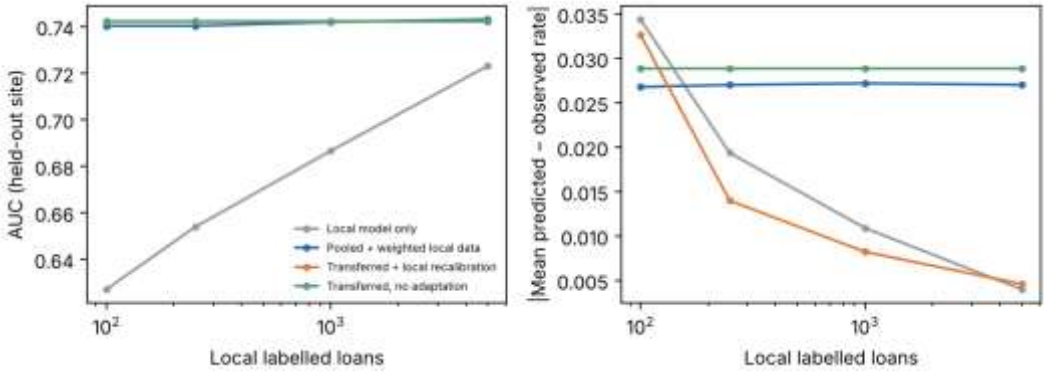


Figure 1. Discrimination (left) and absolute calibration-in-the-large error (right) on held-out site loans by number of local labels and adaptation strategy, averaged over eight sites [authors' analysis].

Recalibration was the efficient choice once a few hundred labels were available. With 1,000 local labels, it reduced the mean absolute gap between predicted and observed charge-off rates from 0.029 to 0.008 while leaving AUC unchanged (0.742); with 5,000 labels the gap was 0.005. With only 100 labels, recalibration made calibration worse (0.033), because the fitted intercept and slope were too noisy. A model trained on local data alone needed 5,000 labels to reach an AUC of 0.723, still below the transferred model, and up-weighting local data in a pooled retrain did not correct calibration (gap 0.027 at 1,000 labels). The practical rule this suggests is to transfer the ranking, recalibrate locally once roughly a thousand outcomes have matured, and not to recalibrate on a few dozen early outcomes.

UNSUPERVISED DRIFT DETECTORS

The D3 detector trains a discriminative classifier to separate a reference window from a new window and signals drift when the cross-validated AUC reaches 0.70[15]; it was run with a logistic and a tree-based discriminator. OCDD fits a one-class support vector machine to the reference window and signals drift when more than 30 percent of the new window falls outside

it[16]. Both were compared with two common baselines: maximum feature PSI of 0.25 or more, and a Bonferroni-corrected Kolmogorov-Smirnov test. Each detector compared a reference sample from one site with a new sample either from the same site (no shift) or from another site (shift), at window sizes of 100 to 1,000 loans. A shift was classed as material when deploying the reference site's model on the new site raised the absolute calibration error or lowered AUC by more than 0.02, which held for 36 of the 56 site pairs. Table 2 reports the results at 250 and 1,000 loans, and Figure 2 shows the full range.

Table 2. Unsupervised drift detectors on cross-site shift (default thresholds; 112 shifted and 32 unshifted comparisons per window size).

Window (loans)	Detector	False-alarm rate (no shift)	Detection rate (material shift)	Flag rate (immaterial shift)	Spearman: statistic vs Brier excess
250	D3 (logistic)	0.00	0.01	0.10	0.12
250	D3 (tree)	0.00	0.03	0.07	0.16
250	KS (Bonferroni)	0.00	0.68	0.90	0.07
250	Max-feature PSI	0.00	0.28	0.50	-0.02
250	OCDD	0.00	0.00	0.07	-0.04
1000	D3 (logistic)	0.00	0.00	0.07	0.02
1000	D3 (tree)	0.00	0.08	0.15	0.05
1000	KS (Bonferroni)	0.00	0.94	0.95	0.01
1000	Max-feature PSI	0.00	0.12	0.25	-0.10
1000	OCDD	0.00	0.00	0.00	0.04

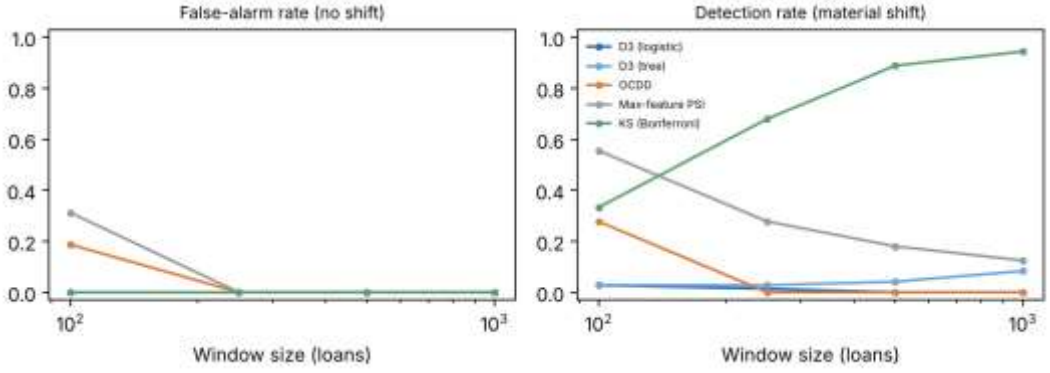


Figure 2. False-alarm rate under no shift (left) and detection rate under material shift (right) by window size and detector [authors' analysis].

Neither detector named in the source architecture was useful at its default threshold for this kind of shift. At 1,000 loans, D3 flagged 0.00 (logistic) and 0.08 (tree) of material shifts and

OCDD flagged 0.00, although neither raised false alarms. The KS test flagged almost everything, material or not (0.94 against 0.95), because with large windows any difference becomes significant. Most importantly, no detector statistic was meaningfully correlated with the actual excess Brier score from transfer (|Spearman| at most 0.19). The sector portfolios differ mostly in their charge-off rate, a prior shift that moves outcomes more than inputs, and input-based detectors cannot see that. This supports the framework's insistence on diagnosis with labelled evidence, and it shows that D3 and OCDD thresholds must be re-tuned for each portfolio rather than carried over from the fraud and distress pilots in which they were originally assessed.

ANALYTICAL SYNTHESIS AND PROPOSED HYPOTHESES

The framework developed above treats cross-site generalization as a sequence of increasingly demanding tests rather than as a single pass-or-fail evaluation. A model that performs well on a random split drawn from within one institution has passed only the least demanding test in that sequence, since a random split does not even establish that the model generalizes forward in time at its own site, let alone to a different one. A chronological holdout within a single site raises the bar by testing temporal generalization specifically. A leave-one-institution-out test raises it further by examining whether the learned relationship transfers to an environment the model has never encountered at all. Local calibration, finally, tests a narrower but practically important question: whether a small, targeted adaptation step, requiring only a limited amount of target-institution data, can recover acceptable performance without committing to a full retraining cycle. Table 3 summarizes the question each of these designs actually answers and the specific limitation attached to it, since no single design in this sequence is sufficient on its own to support a general deployment claim.

Table 3. Validation designs for SME distress early-warning models.

Validation design	Question answered	Key limitation
Random split within one site	Can the model fit this site's sampled data?	May conceal temporal dependence entirely
Chronological holdout within site	Does performance transfer forward in time at this site?	Does not test institutional transfer
Train at one site, test at another	Does the model transfer across institutions?	Sensitive to differences in outcome and feature definitions
Leave-one-institution-out validation	Does performance generalize to a site never seen at all?	Requires enough participating institutions to be informative
Local calibration, then later test	Can limited local adaptation improve transfer?	Needs a strict separation between the calibration sample and the test sample

TESTABLE PROPOSITIONS

Three propositions follow from this framework. The first, P1, is that cross-institution evaluation will reveal materially larger performance dispersion than within-site testing alone. The SBA evaluation supports P1 for calibration but not for discrimination: transferred and within-site AUCs were almost identical, whereas calibration intercepts spread from -0.60 to 0.40 under transfer against -0.17 to -0.03 within site. The second, P2, is that harmonizing outcome and horizon definitions will reduce the apparent between-site variation that is attributable to definitional inconsistency. P2 could not be tested, because every site in the public file already shares one outcome definition (charge-off versus paid in full); the dispersion observed is therefore genuine rather than definitional. The third, P3, is that limited local calibration will improve cross-site transfer at lower cost than full retraining. The evaluation supports P3 with a qualification: recalibration on 1,000 local outcomes cut calibration error by about 72 percent without loss of ranking and outperformed local retraining, but with 100 outcomes it was harmful. None of these results speaks to temporal stability, which requires dated data.

DISCUSSION

The central methodological implication of this framework is that multi-institution validation should never be treated as a single pooled test. It requires an explicit separation of the temporal axis, whether a model remains useful as conditions evolve within the institution that trained it, from the institutional axis, whether the relationships that model has learned transfer to an environment it has never seen. A model can perform well on average across several institutions while failing badly at one of them, or it can transfer well across institutions at a single point in time while nonetheless degrading as conditions evolve within each of them; both failure modes matter for any serious deployment claim, and neither is visible from a single pooled metric computed once.

Rashid and Sattiraju's source architecture[1] motivates multi-institution validation as a direct response to the risk that single-site evaluation confuses local statistical fit with genuine, transferable generalization, and the present framework operationalizes that underlying concern through harmonized outcome definitions, site-specific chronological splits, a staged sequence of transfer tests, explicit label-maturity management, and mandatory site-level reporting rather than pooled-only reporting. The evaluation on SBA loans implements the source architecture's components (multi-site validation, transfer tests, local recalibration, and the D3 and OCDD detectors) on public data. Its central limitation is that sector portfolios stand in for institutions and no approval dates were available, so the framework's temporal arm and its claims about genuine institutional transfer still await dated, lender-identified data.

CONCLUSION

SME distress early-warning models should be evaluated across both time and institutions, because these two dimensions test logically distinct properties that neither substitutes for the other. A chronological holdout within a single institution assesses whether a model remains useful as conditions evolve at that institution; a cross-site transfer test assesses whether the relationships that model has learned generalize to a different institutional environment the model has never encountered. A model that performs well on only one of these two tests has

not yet demonstrated that it is safe to deploy broadly, and a pooled evaluation that collapses both questions into a single number risks concealing exactly the failure modes this framework is designed to surface. This article proposes a validation architecture that combines harmonized outcome and horizon definitions, leakage-safe within-site temporal evaluation, a staged sequence of cross-site transfer tests, unsupervised drift monitoring and explanation-based diagnosis, lead-time and calibration reporting disaggregated by institution, and an explicit governance and data-sharing structure. Applied to 375,261 public SBA loans across eight sector portfolios, it showed that ranking transferred while calibration did not, that a pooled AUC concealed high between-site heterogeneity, that about 1,000 local outcomes were enough for recalibration, and that default-threshold D3 and OCDD alarms did not track the performance cost of shift. It has not yet been validated across U.S. community financial institutions or over time. Future studies should test the framework's predictive and operational value using transparent, site-level reporting and temporally later outcomes, following the chronological-evaluation discipline set out above[3], so that the three propositions advanced in this article can be assessed under conditions that approximate the generalization problem a real multi-institution deployment actually faces.

REFERENCES

[1] Rashid, S. U., & Sattiraju, S. R. (2025). Multi-Institution Validation of Drift-Resilient Early-Warning Models for SME Loan Distress and Transaction Fraud: Evidence from US Community Financial Institution Pilots. *International Journal of Modern Computing*, *8*(1), 40-54.

[2] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44.

[3] Zuberi, N., Rahman, M. M., Rashid, S. U., Ahsan, M. N., Rahman, M. D., & Mollah, M. H. O. R. (2026). A robust machine learning framework for detecting temporal drift in financial fraud prevention. *Scientific Reports*.

[4] Bifet, A., & Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In Proceedings of the 2007 SIAM International Conference on Data Mining (pp. 443-448).

[5] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432.

[6] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

[7] Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 77, 103-123.

[8] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, Article 6.

[9] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[10] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784-3797.

[11] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765-4774). Curran Associates, Inc.

[12] Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc.

[13] Li, M., Mickel, A., & Taylor, S. (2018). "Should this loan be approved or denied?": A large dataset with class assignment guidelines. *Journal of Statistics Education*, 26(1), 55-66.

[14] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146-3154).

[15] Gözüaçık, Ö., Büyükçakır, A., Bonab, H., & Can, F. (2019). Unsupervised concept drift detection with a discriminative classifier. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2365-2368).

[16] Gözüaçık, Ö., & Can, F. (2021). Concept learning using one-class classifiers for implicit drift detection in evolving data streams. *Artificial Intelligence Review*, 54, 3725-3747.