THE COMPUTERTECH
(An International Peer Review Journal)
VOLUME 12; ISSUE 1 (JAN-JUNE); (2026)
WEBSITE: THE COMPUTERTECH

Cross-Institution Generalizability of Financial Risk Models: Validation, Reliability, and Operational Governance

[1*]Thandiwe Aya [1], Tendai Phiri [2]
[1] Middle East Technical University, TURKEY
[2] Tbilisi State University, GEORGIA

Abstract
Financial risk models can perform well within the institution that trained them yet fail when transferred to another, because institutions differ in customer composition, product mix, data definitions, label-confirmation processes, and operating policy, and a model's apparent accuracy can rest on regularities specific to the environment it was built in rather than on any property that travels across environments. This article develops a cross-institution validation framework for drift-resilient early-warning models used in SME loan distress and transaction fraud, building on the multi-institution validation architecture that Rashid and Sattiraju propose and on the broader concept-drift literature's treatment of non-stationary learning problems. The framework separates four questions that are often collapsed into a single reported accuracy figure: whether a model is stable over time within the institution that trained it; whether its ranking quality transfers to an institution it never trained on; whether a small amount of local, target-site data can correct its probability calibration without retraining; and whether a statistically transferable model is actually deployable given an institution's workflow, privacy constraints, and investigation capacity. It specifies evaluation designs for each question, including leave-one-institution-out and pairwise transfer testing, and proposes criteria, beyond detection speed alone, for judging unsupervised drift detectors: false-alert burden, computational cost, robustness to changing sample sizes, and demonstrated association with subsequent, outcome-confirmed performance deterioration. The framework is then applied to 375,261 public SBA-guaranteed small-business loans, with eight industry-sector portfolios standing in for institutions because the public file identifies neither lender nor approval date. Ranking transferred almost intact (leave-one-site-out AUC 0.68 to 0.79), but calibration did not: across 56 site pairs the median absolute error in the predicted charge-off rate was 0.032 and the worst was 0.117. About 1,000 local outcomes were enough for recalibration. Unsupervised detectors at default thresholds missed most shifts; re-tuned to a 5 percent false-alarm rate, D3 detected almost all of them, but flagged harmful and harmless shifts alike, so detection did not identify which transfers needed action. Because sector portfolios are not institutions and the data carry no dates, these results test cross-segment transfer but not temporal stability. Its central argument is that cross-institution generalizability is not a binary property a model either has or lacks, but a bounded set of claims that must be supported by explicit, site-level evidence across time, institutions, and operational conditions before transfer is assumed.

Keywords: Cross-Institution Validation; Financial Risk Models; SME Distress; Transaction Fraud; Concept Drift; Generalizability; Calibration; Model Governance

Introduction

A machine-learning model used inside a financial institution must generalize not only forward in time, to observations that have not yet occurred, but also sideways, to institutional settings that differ from the one in which the model was developed. The second kind of generalization is routinely underexamined relative to the first. Cross-institution validation matters most acutely for community financial institutions, whose portfolios, data infrastructures, underwriting practices, and customer populations commonly differ, sometimes substantially, from the larger institutions or pooled datasets on which a candidate model may originally have been built and validated. A model that performs well at one institution can be exploiting local regularities, specific to that institution's customer base, product mix, or historical period, that simply do not hold at another institution, so that a single reported accuracy figure says far less about deployability elsewhere than its precision suggests.

Rashid and Sattiraju[1] propose a multi-institution validation architecture for drift-resilient early-warning models covering both SME loan distress and transaction fraud, fielding it across a sample of U.S. community financial institution pilots. Their study is valuable precisely because it surfaces, from practice rather than from abstract argument, how difficult it is to distinguish genuine generalization from institution-specific historical fit, and how operationally costly unsupervised drift detection becomes once it must run across independently operating pilot sites with their own data infrastructures and governance constraints. This article takes their architecture as a starting point and develops a more fully specified critical framework for evaluating cross-institution generalizability, reliability, and governance: one that separates evidence of statistical transfer from evidence of operational suitability, and that argues explicitly against letting a single pooled average substitute for site-level performance analysis.

Four questions organize the remainder of this article. First, what does cross-institution generalizability actually mean for a financial risk model, as distinct from the more familiar notion of temporal generalization within a single institution? Second, how should temporal validation, which asks whether a model remains accurate over time at the institution that trained it, be kept conceptually and empirically separate from institutional validation, which asks whether the same model transfers to a different institution altogether? Third, how should unsupervised drift detectors, an attractive option precisely where confirmed labels are scarce or delayed, be assessed in terms of both their statistical detection properties and their operational cost once deployed? Fourth, what governance controls, spanning accountability, privacy, and deployment gating, must be in place before a model is allowed to transfer into a new institution's production environment at all? The sections that follow address each of these in turn, building toward a framework that treats generalizability as a profile of bounded evidence rather than a single pass-fail determination, and then apply the transfer, recalibration and detector-evaluation layers of that framework to public SBA loan data.

Generalization, Drift, and Prior Validation Evidence

Generalization Across Environments and the Limits of Random Holdouts

Conventional machine-learning evaluation practice often assumes, implicitly, that training and test samples are drawn from sufficiently similar distributions that a random holdout from the same pool is an adequate test of generalization. That assumption is routinely violated by institution-specific

processes in financial risk modeling: differences can arise in how a feature is defined and recorded, in customer composition and product mix, in the prevalence of the event being predicted, in investigation and escalation policy, and in the timing with which a label is confirmed as final rather than provisional. A random holdout drawn from a single institution's own historical data cannot, by construction, detect any of these cross-institution differences, because every observation in both the training and the holdout set was generated by the same institutional process. Generalization claims for a financial risk model should therefore be evaluated explicitly across environments, including but not limited to different institutions, rather than inferred from a single within-institution random split, however large that split happens to be.

This is not merely a methodological nicety. A model whose reported performance rests entirely on a within-institution holdout carries an implicit, usually unstated, generalization claim that the broader financial-risk-modeling literature has increasingly scrutinized: that the regularities the model has learned are properties of the underlying fraud or distress process rather than artifacts of one institution's particular data-generating environment. Separating these two possibilities requires a validation design that varies the environment deliberately, which is precisely what the framework developed in the remainder of this article is intended to provide.

Temporal Drift Versus Institution Shift

Temporal drift refers to change across time within a single environment, while institution shift refers to a difference between organizations at a given point in time; the two axes can interact, but they are not interchangeable, and a validation design that only tests one cannot be read as evidence about the other. A model may be stable over time at Institution A, passing every rolling-origin temporal test an evaluator could reasonably construct, yet perform poorly from the moment it is deployed at Institution B, because the institutional difference, not any temporal change, is what degrades it. Equally plausibly, a model may transfer acceptably across institutions at a given period, only to deteriorate later as the environment at the receiving institution itself evolves, which is a temporal-drift problem layered on top of an initially successful transfer. The concept-drift literature's treatment of non-stationary learning, distinguishing virtual drift in the input distribution from real drift in the underlying input-label relationship[2], provides a vocabulary for the first axis; it does not, by itself, address the second. A companion methodological contribution demonstrating leakage-safe, chronologically evaluated drift detection on a benchmark credit-card dataset[3] is similarly a contribution to the temporal axis specifically, and its chronological-splitting discipline, valuable as it is, does not substitute for an explicit cross-site transfer test of the kind this article argues is a separate and equally necessary requirement.

Treating these two axes as a single undifferentiated notion of "generalization" obscures exactly the information a deploying institution most needs: whether a candidate model's weaknesses, if any, are a temporal problem that monitoring and periodic refresh can address, an institutional problem that no amount of monitoring at the original site can reveal, or some combination of the two that only a validation design testing both axes explicitly can disentangle.

Unsupervised Drift Detection and Its Costs

Unsupervised drift detectors, which flag a possible distributional change without requiring confirmed outcome labels, are attractive wherever labels are expensive, delayed, or both, which describes the label-confirmation processes of most financial risk-prediction settings reasonably

well. Rashid and Sattiraju's source architecture identifies two such detectors, D3 and OCDD, as relevant candidates and discusses their differing label-efficiency profiles in detail[1]; that discussion is a genuine contribution, but a detector's practical value should not be judged by detection speed or label efficiency alone. A detector should also be evaluated for its false-alarm rate, its rate of missed genuine changes, its computational cost at the scale a production monitoring system actually runs at, its robustness to the kind of changing sample sizes a growing or shrinking institution naturally produces, and, most importantly, whether its alerts are actually associated with subsequent, outcome-confirmed performance deterioration once mature labels eventually become available.

An adaptive-windowing approach, in which the effective size of the monitoring window automatically shrinks when a statistical test detects a significant change between its older and newer portions[4], illustrates one way to operationalize continuous unsupervised monitoring without committing in advance to a fixed lookback period, and it is a useful complement to the fixed-window detectors that Rashid and Sattiraju's architecture emphasizes. Whichever specific detector or windowing scheme an institution adopts, the evaluation standard proposed in this article is the same: a statistical alert is evidence to be weighed against the detector's own false-alarm and operational-cost profile, not a self-sufficient trigger for institutional or model-level action.

Statistical Versus Operational Generalizability

A model that is statistically transferable, in the sense that its ranking quality holds up under a rigorous cross-site test, may nonetheless be operationally unsuitable for deployment at the receiving institution. It may produce more alerts than the institution's investigation capacity can absorb, arrive with probability estimates that are poorly calibrated to the receiving institution's true event prevalence, depend on a feature that the receiving institution's data governance rules do not permit it to use, or simply fail to integrate cleanly into the institution's existing case-management workflow. Interpretability is itself part of this operational picture rather than a separate concern: an explanation method such as SHAP, which attributes an individual prediction to the input features that produced it[5], can materially affect whether reviewers and regulators are willing to act on a transferred model's outputs, independent of that model's measured statistical transfer performance. Generalizability, properly understood, must therefore include predictive metrics, calibration quality, alert burden, latency, interpretability, and governance feasibility together, and cross-site statistical validation, however carefully it is conducted, should be treated as a necessary but not a sufficient condition for deployment at a new institution.

Cross-Institution Validation Framework

The framework proposed in this article contains five evaluation layers, each of which answers a different question about a candidate model and should, for that reason, be reported separately rather than collapsed into a single composite score: data and outcome harmonization, within-site temporal validation, cross-site transfer testing, local adaptation and recalibration, and operational and governance review. No layer substitutes for another; a model that performs acceptably at one layer still requires separate evidence at each of the remaining layers before an institution can responsibly treat it as validated for cross-institution use.

Data and Outcome Harmonization

Before any cross-institution comparison is meaningful, participating institutions should document their feature definitions, measurement units, patterns of missingness, observation dates, label-maturity windows, and the precise rules that define a target event. For SME distress specifically, default, delinquency, restructuring, and covenant breach are related but distinct events, and treating them as interchangeable without an explicit, documented mapping will produce apparent cross-institution performance differences that are in fact differences in what is being measured rather than differences in model quality. For transaction fraud, confirmation policies and chargeback windows directly affect observed label prevalence; an institution with a longer confirmation window will, all else equal, appear to have a different fraud rate than one with a shorter window, purely as an artifact of label timing. Because the two domains this framework addresses, SME distress and transaction fraud, both involve rare target events, resampling methods developed for severe class imbalance[6] are frequently used upstream of model training at the harmonization stage itself, and any such resampling choice should be documented alongside the outcome definitions it was applied to, since a resampling strategy calibrated at one institution's event prevalence can itself become a source of apparent cross-institution difference if it is not revisited when the target institution's prevalence differs materially.

Differences in label process are easy to mistake for differences in model performance, and the harmonization layer exists specifically to prevent that mistake. An institution evaluating a transferred model should ask, before attributing any observed performance gap to the model itself, whether the two institutions being compared actually share a common definition of the event the model is trying to predict, and should treat a documented mapping between definitions as a prerequisite for any subsequent comparison rather than as an optional appendix to it.

Within-Site Temporal Validation

Each participating site should preserve chronological order in its own evaluation, using separate training, validation, and test windows drawn in that temporal sequence, formally $D_{train,s} \prec D_{validation,s} \prec D_{test,s}$ for each site s, so that no information from a later period leaks, directly or through tuning, into an earlier period's model fitting. Any learned preprocessing, feature selection, resampling, or threshold tuning should be confined to data genuinely available before the test period begins, and rolling-origin evaluation, repeating this chronological split at successive cutoffs, should be used wherever a site's history is long enough to support it. This establishes, independently for each site, a baseline measure of temporal generalization: whether the model remains accurate over time within that one institution, before any question of cross-site transfer is even raised.

This within-site baseline is a precondition for everything that follows, not a redundant check. A model that cannot sustain accuracy over time even within the single institution that trained it has failed a strictly easier test than cross-site transfer, and an institution should establish this baseline before investing in a cross-site transfer evaluation whose results would be uninterpretable without it.

Cross-Site Transfer Designs

Several distinct transfer designs are useful, and they estimate materially different deployment scenarios, which is why they should be reported separately rather than averaged together into one transfer score. In leave-one-institution-out evaluation, a model is trained on a pooled set of institutions and tested on an institution held out entirely from training, estimating how a model trained centrally would perform if deployed, unmodified, at a new site. In pairwise transfer, training occurs at one specific institution and testing at another specific institution, estimating a narrower, site-pair-specific transfer claim rather than a general one. In local calibration, a small, explicitly designated sample from the target institution is used only to adjust output probabilities or decision thresholds, with evaluation then conducted on a later, untouched period at that same target institution, estimating how much a light-touch local adjustment, short of full retraining, can recover.

These three designs are not substitutes for one another, and a report that only presents the most favorable of the three without disclosing which design produced it misrepresents what has actually been established. An institution reading a transfer result should always be told which of these three designs generated it, since a leave-one-institution-out result, a pairwise result, and a locally calibrated result license meaningfully different deployment decisions.

Local Adaptation and Recalibration

Local calibration can correct a difference in event prevalence or in the overall scale of a model's output probabilities without altering the model's underlying ranking of which transactions or borrowers are riskier than others; it cannot, by the same token, necessarily repair a genuinely changed relationship between input features and the outcome, which is the harder case recalibration is not designed to address. A calibration method such as Dirichlet calibration, which extends temperature-scaling-style approaches to produce well-calibrated class probabilities without distorting the underlying ranking[7], is one concrete technique an institution can apply at this layer, but applying any calibration technique still requires documenting how much local data was used and its temporal position relative to the evaluation period. If the same local sample used to fit a calibration adjustment is also used to tune a decision threshold, a separate, later test period is required to avoid an optimistically biased evaluation, since otherwise the calibration and the threshold have, in effect, been tuned and graded on the same data.

The broader point this layer is meant to enforce is that local adaptation is a graded, documented intervention with its own evidentiary requirements, not a one-line footnote that a model was "recalibrated for the new site" without further specification of how, on how much data, or evaluated against what subsequent period.

Site-Level Reporting

Performance should be reported separately for each participating institution, together with that institution's event prevalence, sample size, missingness profile, and confidence intervals around its reported metrics; pooled, cross-institution metrics should be secondary to this site-level breakdown and should always be accompanied by an explicit measure of between-site heterogeneity rather than presented as a single number that implies uniform performance. A model that performs very well at several large institutions but poorly at one small institution should not be described as uniformly generalizable merely because its sample-weighted pooled average looks strong; the small

institution's poor performance is precisely the information a pooled average is structurally prone to concealing, and it is often the information that matters most to the institution considering whether to adopt the model.

Evaluating Drift Detectors

Detection Quality

Drift detectors intended to support cross-institution monitoring should be assessed on detection delay, the rate of false-positive alerts, the rate of missed genuine changes, computational cost at production scale, and robustness to the changing sample sizes that naturally occur as an institution's transaction or loan volume grows or shrinks over time. Where a detector is unsupervised, as most of the detectors relevant to this framework are, the reference event used to judge whether a given alert was a true or a false alert must itself be defined with care, because a statistical distributional change is not automatically the same thing as a harmful change in the model's actual predictive performance; a detector can be perfectly accurate at detecting distributional movement while that movement turns out to be operationally inconsequential. Outcome-based evaluation, conducted once mature labels eventually become available, should be used specifically to determine whether a detector's alerts are associated with subsequent performance deterioration, which is the test that actually matters for deciding whether to act on an alert rather than merely log it.

An adaptive-windowing detector of the kind discussed above[4] ties its effective monitoring window directly to detected change, which can reduce the need to commit in advance to a fixed lookback period, but it does not remove the obligation to validate, separately and empirically, that its window-shrinking events are in fact associated with the kind of outcome deterioration this evaluation layer is designed to catch.

Label Efficiency and Operational Cost

Rashid and Sattiraju's source architecture emphasizes label-efficiency differences between its candidate unsupervised detectors[1], and label efficiency genuinely matters in settings where institutions have a limited supply of confirmed outcomes, which describes most SME-distress and transaction-fraud settings given realistic label-confirmation delays. A realistic treatment of credit-card fraud detection makes a closely related point from a practitioner's vantage point: verification latency, the delay between a transaction occurring and its outcome being confirmed, is itself a first-order feature of the learning and monitoring problem rather than a secondary data-engineering inconvenience that can be abstracted away[8]. The number of labels a detector requires to calibrate is therefore only one dimension of its true operational cost; an institution evaluating a detector should also weigh how frequently it generates alerts, the staff burden of reviewing each alert, and whether acting on its alerts measurably improves the timing of intervention relative to not monitoring at all.

A detector that is cheap in confirmed labels but expensive in investigator time has not actually reduced an institution's operational burden; it has merely relocated that burden from the labeling process to the alert-review process, and a complete cost accounting must capture both.

Decision Thresholds for Alerts

Alert thresholds should be calibrated against an institution's own historical variability and its specific operational priorities, rather than imported wholesale from another institution's published benchmark or from a generic statistical convention. A highly sensitive detector threshold will generate a correspondingly high rate of false alarms, consuming investigation capacity on alerts that further diagnosis will not confirm as genuine; an overly conservative threshold will, by the same logic, miss consequential shifts until they have already caused material harm. A threshold should be selected through a validation process conducted on one period and then assessed on a later, independent period, and no fixed population-stability-index, Kolmogorov-Smirnov, or divergence threshold should be presented as universally valid across institutions and features, because the natural variability of a given feature or score differs across institutions for reasons unrelated to the presence or absence of genuine drift.

Metrics and Analytical Decision Framework

Predictive Metrics

Classification metrics reported under this framework should include precision, recall, the area under the precision-recall curve, the Matthews correlation coefficient, and calibration quality wherever calibration is relevant to the deployment decision. Precision-recall analysis is particularly informative for the rare-outcome settings this framework addresses, since it focuses directly on the operational question of how many flagged cases are genuinely positive and how much of the genuine positive population is actually captured[9], while the Matthews correlation coefficient offers a complementary, single-figure summary that incorporates all four confusion-matrix counts and is markedly less distorted by severe class imbalance than accuracy or even F1[10]. The receiver operating characteristic curve remains informative for comparing ranking quality independent of a specific threshold[11], but a coherent critique of AUC as a universal summary measure argues that it implicitly uses a different misclassification-cost distribution for different classifiers being compared, which can make AUC-based comparisons misleading precisely in the cross-institution setting this framework addresses, where the relevant cost structure can itself differ from one institution to the next[12]. For early-warning models specifically, lead time, the interval between an alert and the point at which the alert would need to be confirmed to still be useful, and the proportion of alerts that actually arrive early enough to permit meaningful intervention, should also be reported alongside these more conventional classification metrics.

$$\mathrm{MCC} = (\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}) \,/\, \sqrt{[(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})]}$$

Generalizability Profile

A generalizability profile, as proposed here, summarizes within-site temporal performance, cross-site transfer performance under each of the designs discussed above, performance after local calibration, and the variation in each of these across participating institutions, rather than reducing all of this to a single headline number. The profile should report both the average outcome across sites and the worst-performing site, or a suitably conservative lower quantile across sites, specifically because an average, by construction, is the statistic most likely to conceal an operationally important single-site failure. Table 1 summarizes the four evaluation layers this framework treats as load-bearing, the central question each answers, the evidence each requires,

and the specific risk of omitting it; no single layer in the table substitutes for another, and a generalizability claim that rests on evidence from only one or two of the four layers should be reported as exactly that, a partial claim, rather than as general evidence of cross-institution validity.

Table 1. Evaluation layers in the proposed cross-institution generalizability profile.

Evaluation layer	**Central question**	**Required evidence**	**Risk if the layer is skipped**
Within-site temporal validation	Is the model stable over time at the institution that trained it?	Rolling-origin precision, recall, PR-AUC, and MCC computed on strictly later, previously unseen windows	A model that looks stable only because it was never tested on genuinely later data
Cross-site transfer (leave-one-institution-out / pairwise)	Does the model's ranking quality transfer to an institution it never trained on?	Site-level predictive metrics and calibration, computed and reported separately for each held-out institution	A pooled average that conceals one or more sites where the model actually fails
Local calibration	Can a small amount of target-site data correct probability scale without full retraining?	Calibration error before and after adjustment, evaluated on a later, untouched period at the target site	A calibration step that is itself overfit to the same data used to tune the decision threshold
Operational / governance review	Is the statistically validated model actually deployable under this institution's workflow and constraints?	Alert volume, investigation burden, data-sharing constraints, and rollback readiness	A model that is statistically sound but operationally unusable once it reaches production

Cost and Utility

Where an institution has access to reliable cost estimates, an expected-loss figure can combine the cost of false negatives, false positives, review effort, and delayed intervention into a single decision-relevant quantity, and such a figure should always be computed separately per site rather than pooled, for the same reason site-level predictive metrics should not be pooled. Where credible cost estimates are not available, which is common, a sensitivity analysis showing how the resulting decision changes across a plausible range of cost assumptions is preferable to a single point estimate that conceals how much the conclusion actually depends on an assumption the institution cannot verify. A study operating under this framework should take particular care not to present assumed costs, however reasonable they appear, as though they were observed financial benefits actually realized in deployment; the two are categorically different kinds of claims; and the resampling and imbalance-correction choices discussed earlier in the Data and Outcome Harmonization section[6]

can themselves shift the effective cost trade-off a model is optimized against, which is a further reason to document them explicitly rather than treat them as incidental preprocessing detail.

Empirical Application to Public SBA Loan Data

Data and Design

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].

Generalizability Profile

Table 2 gives the site-level generalizability profile that the Generalizability Profile section calls for: discrimination within each site and under leave-one-site-out (LOSO) transfer, with calibration of the transferred model on each site's held-out loans.

Table 2. Generalizability profile 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

Sector portfolio (site)	Loans	Charge-off rate	AUC within site	AUC transferred (LOSO)	Mean predicted (LOSO)	Calib. intercept (LOSO)	Calib. slope (LOSO)
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

The profile separates two properties that a pooled figure merges. Ranking quality transferred: averaged across sites, LOSO AUC was 0.741, and the worst site, accommodation & food, still reached 0.681 under transfer. Calibration did not transfer: the transferred model predicted a charge-off rate of 18.9 percent for health care, where the observed rate was 10.5 percent. Pairwise transfer, from each single site to each other site, showed the same pattern more sharply. Across the 56 ordered pairs, the median AUC loss relative to the target site's own model was 0.007 (maximum 0.022), but the median absolute calibration error was 0.032, reaching 0.117 when the health-care model was applied to accommodation and food services. Heterogeneity across sites was high (I^2 = 0.96), so reporting the worst site alongside the mean, as the framework requires, changes the conclusion a reader would draw.

Local Recalibration

For each held-out site, the transferred model was adapted using n local labelled loans (100, 250, 1,000 or 5,000) by logistic recalibration, by training a local model only, or by retraining on pooled data with the local loans up-weighted. Figure 1 averages the results over sites.

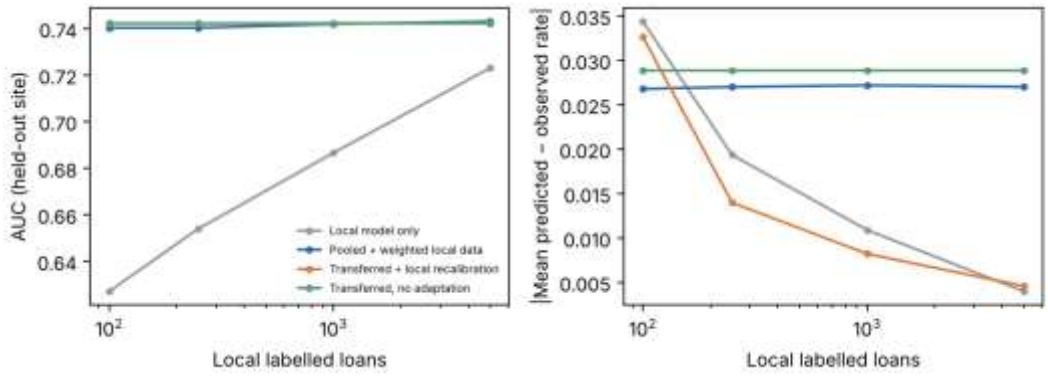


Figure 1. Discrimination (left) and absolute calibration-in-the-large error (right) on held-out site loans by number of local labels and adaptation strategy [authors' analysis].

Recalibration corrected calibration without retraining, as the Local Adaptation and Recalibration section anticipates, but only once enough outcomes had accumulated. With 1,000 local labels, the mean calibration gap fell from 0.029 to 0.008 with AUC unchanged; with 100 labels it rose to 0.033. A local-only model was worse on both counts at every sample size tested (AUC 0.687 with 1,000 labels), and up-weighting local loans in a pooled retrain left calibration almost untouched. In label-efficiency terms, a few hundred to a thousand matured outcomes per site is the practical minimum before recalibration should be allowed under a deployment gate.

Evaluating the Drift Detectors

The four criteria proposed in the Evaluating Drift Detectors section were applied to D3[15], OCDD[16], a maximum-feature PSI rule and a Bonferroni-corrected Kolmogorov-Smirnov test. Each detector compared a reference sample from one site with a new sample from the same site (no shift) or another site (shift), using windows of 100 to 1,000 loans. A shift was classed as material when applying the reference site's model to the new site raised the absolute calibration error or lowered AUC by more than 0.02; 36 of the 56 pairs (64 percent) met this definition.

Table 3 compares each detector at its default threshold with the same detector re-tuned so that its false-alarm rate on unshifted comparisons was 5 percent, the procedure the Decision Thresholds for Alerts section recommends. All results use 1,000-loan windows.

Table 3. Drift detectors on cross-site shift at 1,000-loan windows: default versus re-tuned thresholds (re-tuned to a 5% false-alarm rate on unshifted comparisons).

Detector	Default: detects material	Default: flags immaterial	Re-tuned: detects material	Re-tuned: flags immaterial	Re-tuned: share of flags that were material	Time per check (ms)
D3 (logistic)	0.00	0.07	0.97	0.97	0.64	9.6
D3 (tree)	0.08	0.15	1.00	1.00	0.64	358.4
OCDD	0.00	0.00	0.22	0.40	0.50	11.9
Max-feature PSI	0.12	0.25	0.93	1.00	0.63	0.6
KS (Bonferroni)	0.94	0.95	0.99	0.97	0.65	6.2

Three findings follow. First, default thresholds were badly mismatched to this portfolio. D3 at its published AUC threshold of 0.70 and OCDD at a 30 percent outlier share almost never fired, whereas the KS test fired on nearly every comparison. Second, re-tuning fixed sensitivity but not relevance. At a 5 percent false-alarm rate, D3 with a tree discriminator flagged 1.00 of material shifts and 1.00 of immaterial ones, and the share of its flags that were material (0.64) was no better than the base rate of material pairs (0.64). The detectors were detecting that portfolios differed, which they always did, not whether the difference harmed the model. Third, costs varied by three orders of magnitude, from 0.6 ms per check for PSI to 358 ms for tree-based D3, which matters when checks run continuously across many sites. No detector statistic was associated with the actual excess Brier score of transfer (|Spearman| at most 0.19), which fails the fourth criterion: demonstrated association with outcome-confirmed deterioration.

For this kind of shift, a deployment gate based on label-free alarms would therefore either block nothing or block every transfer. A gate based on a few hundred matured local outcomes, used to estimate calibration error directly, would have separated the harmful transfers from the harmless ones.

Governance and Implementation

Model and Data Accountability

Each participating institution should identify, by name or role, a model owner, a data owner, the operational decision-maker who acts on the model's outputs, and the authority responsible for approving any subsequent change to the model or its thresholds. The validation record maintained for a transferred model should include its training data periods, feature definitions, outcome definitions, decision thresholds, drift-detector settings, and known limitations, and any cross-institution transfer result reported from this record should explicitly specify the population and the conditions under which the evidence was obtained, rather than being stated as though it were an unconditional, context-free claim about the model's quality.

Privacy and Cross-Institution Data Sharing

Cross-institution validation, almost by definition, raises data-sharing questions that a purely within-institution evaluation does not, and it may require data minimization, secure computation environments, contractual controls on permitted use, and a documented review of exactly what is shared and why. Institutions participating in a cross-institution validation exercise should document explicitly whether raw transaction or loan-level records ever leave their own environment, and what can be inferred about an individual record from shared summary statistics or from updates to a shared model even when raw records themselves are never transmitted. Privacy-preserving computation techniques can usefully reduce what is exposed during a cross-institution validation exercise, but they do not eliminate the underlying need for access control, contractual governance, and an explicit risk assessment conducted before, not after, data or model updates begin moving between institutions.

Deployment Gates and Rollback

Before a model is deployed at a new institution, it should pass a defined sequence of gates: a data-quality check specific to the receiving institution's own data, a temporal holdout evaluation at that institution if sufficient history exists, a calibration review, a workflow-integration test, and a documented operational approval step with a named approver. The receiving institution should retain an explicit fallback, either a previously validated model or a manual procedure, that it can revert to if the newly deployed model underperforms once live. After deployment, ongoing monitoring should compare the model's actual, observed performance against the generalizability profile established before deployment, and should trigger a formal review whenever a pre-agreed condition, rather than an ad hoc judgment call made after the fact, indicates that the model's performance has materially diverged from what the profile led the institution to expect.

Conclusion

Cross-institution validation is essential whenever a financial risk model is intended to operate beyond the single institution that supplied its training data, which, in practice, is most deployments a vendor or a central modeling team would consider economically worthwhile. Temporal validation and institutional transfer address genuinely different sources of uncertainty, as the Generalization, Drift, and Prior Validation Evidence section argues at length, and a credible evaluation must therefore assess both separately rather than allowing strong performance on one to stand in for the other. Unsupervised drift detection can meaningfully support ongoing monitoring once a model is

deployed across institutions, but its alerts must always be interpreted against outcome-based performance evidence and against the operational consequences of acting, or failing to act, on a given alert, rather than being treated as self-sufficient triggers in their own right. This article has proposed a validation profile that integrates data and outcome harmonization, rolling within-site temporal testing, cross-site transfer under multiple designs, local calibration with its own documented evidentiary requirements, systematic drift-detector evaluation, disaggregated site-level reporting, and explicit governance gates, building directly on, and substantially extending, the multi-institution architecture that Rashid and Sattiraju[1] first proposed. Applied to 375,261 public SBA loans across eight sector portfolios, the profile showed ranking transferring while calibration did not, local recalibration working once about 1,000 outcomes were available, and unsupervised detectors that, even when re-tuned, detected shift rather than harm. These portfolios are not institutions and carry no dates, so the framework should still be tested using diverse, longer-duration data from a genuinely varied set of institutions, with transparent reporting of site-specific outcomes rather than pooled summaries alone. Until that evidence exists, any claim that a given financial risk model "generalizes across institutions" should remain explicitly bounded by the particular institutions, periods, labels, and operational workflows that were actually evaluated, and not extended, by assumption, to institutions, periods, or conditions that were not.

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] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

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

[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.