International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 2, 78-86, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I2P112

Original Article

A Comprehensive Framework for Model Monitoring Metrics in Credit Risk: From Statistical Foundations to Governance Practice

Sai Prashanth Pathi
Independent Researcher, USA.

Received On: 06/03/2026 ***Revised On: 30/03/2026*** ***Accepted On: 07/04/2026*** ***Published On: 18/04/2026***

Abstract - *Credit risk models are central to lending decisions, capital allocation, and regulatory compliance at financial institutions worldwide. While model development and validation have been extensively studied, comparatively fewer works provide integrated frameworks for ongoing model monitoring that combine statistical metrics with governance structures. This paper presents a unified, hierarchically structured framework for model monitoring in credit risk, synthesising metrics across four dimensions: population stability, discriminatory power, calibration accuracy, and input variable stability. We formalise the Population Stability Index (PSI), Characteristic Stability Index (CSI), the Gini coefficient, Kolmogorov–Smirnov (KS) statistic, Area under the Receiver Operating Characteristic Curve (AUROC), and calibration-based metrics within a consistent mathematical notation. We further introduce a traffic-light governance overlay that maps metric thresholds to actionable escalation protocols, aligned with SR 11-7 and Basel II/III supervisory expectations. Empirical validation is conducted on a synthetic retail loan portfolio of 10,000 development observations and six quarterly production cohorts with programmatically controlled covariate and default rate drift. The logistic regression scorecard achieves a development AUROC of 0.9359 (Gini = 0.8717, KS = 0.7408), and the multi-dimensional monitoring dashboard correctly flags early calibration deterioration (Calibration Ratio reaching 0.70 at Q1) and sustained CSI drift (debt-to-income CSI = 0.946, num_inquiries CSI = 1.036 by Q6) while discriminatory power remains robust throughout. Our results demonstrate the non-redundancy of the four monitoring dimensions and support the adoption of multi-metric dashboards over single-indicator approaches. The proposed Integrated Credit Risk Monitoring Architecture (ICRMA) is designed to be accessible to practitioners at smaller institutions while remaining technically rigorous for model risk management professionals.*

Keywords *- Credit Risk, Model Monitoring, Population Stability Index, Gini Coefficient, Ks Statistic, Model Risk Management, Calibration, Scorecard, Psi, Csi, Traffic-Light Framework, Sr 11-7, Auroc, Log Loss.*

1. Introduction

The deployment of statistical and machine learning models for credit risk assessment has proliferated across retail, commercial, and wholesale banking sectors. These models which include logistic regression scorecards, gradient boosting classifiers, and survival models are used to evaluate borrower creditworthiness, set pricing, determine credit limits, and inform capital requirements under regulatory frameworks such as Basel II and Basel III [1], [2].

A critical but frequently underexamined operational challenge is model monitoring: the systematic, ongoing evaluation of a deployed model's fitness for purpose. Unlike model validation, which is performed at development or periodic review intervals, monitoring is a continuous process that detects real-time or near-real-time deterioration. Performance degradation can arise from distributional shifts in the input population (covariate shift), changes in the underlying relationship between predictors and default (concept drift), macroeconomic shocks, data pipeline failures, or changes in origination policy [3], [4].

Regulatory mandates underscore the importance of robust monitoring. The Federal Reserve and OCC's Supervisory Guidance on Model Risk Management (SR 11-7 / OCC 2011-12) explicitly requires that model owners establish performance monitoring processes with defined thresholds and escalation procedures [5]. The European Banking Authority (EBA) guidelines on internal ratings-based (IRB) models similarly mandate evidence of ongoing predictive accuracy [6]. Despite these requirements, published frameworks that integrate multiple monitoring dimensions into a single governance structure remain sparse in the academic literature.

Prior empirical work has demonstrated that evaluation metrics beyond AUROC, including log loss, PSI, and the KS statistic, provide complementary diagnostic information by capturing aspects of model performance not reflected in rank-order statistics alone. Our work extends this finding by formalising a multi-dimensional monitoring architecture and providing a governance overlay for operational deployment.
This paper makes the following contributions:

- A formal, unified mathematical framework for credit risk model monitoring spanning stability, discriminatory power, and calibration dimensions.
- A hierarchically structured traffic-light governance overlay mapping metric values to RAG (Red / Amber / Green) thresholds with regulatory alignment.
- Empirical demonstration on synthetic retail lending data with programmatically controlled drift, producing concrete numerical benchmarks across six monitoring periods.
- Evidence that calibration deterioration can precede discriminatory power degradation, with practical implications for IFRS 9 expected credit loss estimation.

The remainder of this paper is organised as follows. Section II reviews related work. Section III develops the theoretical foundations. Section IV presents the ICRMA framework. Section V reports empirical results. Section VI discusses implementation. Section VII concludes.

2. Related Work

The credit scoring literature is extensive, with foundational contributions from Fisher [7], who introduced discriminant analysis, and Beaver [8], who applied financial ratios to bankruptcy prediction. The modern credit scorecard framework was consolidated by Siddiqi [9] and Anderson [10].

Model validation has received relatively greater academic attention. Tasche [11] developed rigorous statistical tests for validating probability of default (PD) estimates. Engelmann et al. [12] provided a comprehensive treatment of the accuracy ratio (Gini) and ROC curve analysis. Hand and Henley [13] surveyed statistical classification methods for credit scoring.

Covariate shift has been studied extensively in machine learning [14], [15]. PSI is conceptually related to divergence measures such as the Kullback–Leibler divergence [16], although it is not a formal divergence metric and is primarily used as a heuristic measure in industry practice. Calibration in credit risk has been addressed through Basel Committee guidance [17] and by Hosmer and Lemeshow [18].

The comparative evaluation of monitoring metrics was addressed by Pathi [19], who studied Somers' D, log loss, PSI, and KS statistic across credit risk and healthcare domains. That work found that while AUROC and KS statistic capture rank-ordering performance, PSI independently detects population-level shifts that may not immediately impair discriminatory power, a finding that directly motivates the multi-dimensional architecture proposed in the present paper.

The model risk management literature, primarily practitioner-oriented, provides governance requirements in SR 11-7 [5] and the BCBS 239 principles [20]. Our work bridges the gap between statistical methodology and governance by providing implementable metrics with regulatory-aligned escalation protocols.

Concept drift and model monitoring have also been extensively studied in the machine learning literature. Gama et al. [21] provide a comprehensive survey of drift detection methods, while Niculescu-Mizil and Caruana [22] highlight the importance of probability calibration in classification models. These works reinforce the need for multi-dimensional evaluation frameworks that extend beyond discrimination metrics.

3. Theoretical Foundations of Monitoring Metrics

3.1. Notation and Setup

Let $X = (X_1, ..., X_p)$ denote the p-dimensional feature vector for a loan application, and let $Y \in \{0,1\}$ denote the binary default outcome ($Y = 1$ for default). A credit risk model $M: X \rightarrow [0,1]$ produces a predicted probability of default (PD) score $s = M(x)$.

Let D_{dev} denote the development dataset drawn from distribution $P_{dev}(X,Y)$. Let D_t denote the production dataset at monitoring period t, drawn from $P_t(X,Y)$. We decompose distributional shift into: (i) covariate shift, where $P_t(X) \neq P_{dev}(X)$ but $P(Y|X)$ is unchanged; (ii) concept drift, where $P_t(Y|X) \neq P_{dev}(Y|X)$; and (iii) label drift, where $P_t(Y) \neq P_{dev}(Y)$.

3.2. Population Stability Index (PSI)

The PSI quantifies divergence between the score distributions of the development and current production populations. Partition the score range into B non-overlapping bins. Let $E_i = P_{dev}(S \in b_i)$ and $A_i = P_t(S \in b_i)$:

$$PSI = \Sigma_i (A_i - E_i) \times \ln(A_i / E_i)$$

PSI is widely used in industry as a heuristic measure of distributional change. While it is conceptually related to divergence measures such as the Kullback–Leibler divergence, it is not a true statistical divergence and does not satisfy properties such as symmetry or the triangle inequality. As such, PSI should be interpreted as an effect-size indicator rather than a formal statistical distance. Industry thresholds classify $PSI < 0.10$ as no significant change, $0.10 \leq PSI < 0.25$ as moderate shift, and $PSI \geq 0.25$ as severe shift.

3.3. Characteristic Stability Index (CSI)

The CSI extends PSI to individual input features, enabling attribution of population instability to specific variables:

$$CSI_j = \Sigma_i (A_{ij} - E_{ij}) \times \ln(A_{ij} / E_{ij})$$

The CSI vector $(CSI_1, ..., CSI_p)$ provides a diagnostic attribution map identifying which inputs drive population instability. Unlike PSI, CSI is not a formally standardized statistical measure but rather an industry-adopted extension used for feature-level stability diagnostics.

3.4. Discriminatory Power Metrics

Discriminatory power metrics assess the model's ability to rank borrowers by true default risk. The AUROC equals the probability that a randomly selected defaulter receives a higher predicted PD than a randomly selected non-defaulter:

AUROC = P(s_bad > s_good). The Gini coefficient is: Gini = 2 × AUROC − 1.

The KS statistic in the credit risk context measures the maximum vertical separation between the cumulative distribution functions of defaulters and non-defaulters:
KS = max_t |F_bad(t) − F_good(t)|, where t ranges over score thresholds in the predicted score distribution.

3.5. Calibration Metrics

Calibration assesses whether predicted PD estimates correspond to observed default frequencies. The Expected Default Rate (EDR) and Actual Default Rate (ADR) yield the Calibration Ratio:

$$CR = ADR / EDR$$

A CR of 1.0 indicates perfect calibration. The Hosmer–Lemeshow test [18] provides a formal test of calibration adequacy:

$$HL = \Sigma_i (O_i - n_i \times \bar{p}_i)^2 / (n_i \times \bar{p}_i \times (1 - \bar{p}_i)) \sim \chi^2(B - 2)$$

Log loss is a calibration-sensitive metric that provides a continuous scoring rule: $LL = -(1/N) \Sigma [y_i \log(s_i) + (1-y_i) \log(1-s_i)]$. Unlike AUROC, log loss penalises miscalibrated probability estimates and thus provides complementary signals to rank-order metrics.

4. The Proposed Multi-Dimensional Monitoring Framework

4.1. Framework Architecture

We propose the Integrated Credit Risk Monitoring Architecture (ICRMA), a four-layer hierarchical monitoring framework. The layers are: (1) Input Stability Layer (CSI per feature), (2) Score Stability Layer (PSI), (3) Discriminatory Performance Layer (AUROC, Gini, KS), and (4) Calibration Layer (CR, HL test, log loss trend). Each layer maps to a set of metrics, thresholds, and governance actions.

The layers are diagnostically complementary: the input stability layer identifies the source of population change; the score stability layer quantifies its effect on model output; the discriminatory layer measures rank-ordering degradation; and the calibration layer detects bias in absolute PD estimates. As demonstrated empirically in Section V, these dimensions can move independently, underscoring the necessity of monitoring all four. In practice, these dimensions may evolve asynchronously depending on portfolio dynamics and external economic conditions.

4.2. Traffic-Light Governance Overlay

For each metric m with observed value v, we assign a RAG (Red/Amber/Green) status based on empirically calibrated thresholds. Table I summarises the ICRMA metric definitions, thresholds, and recommended governance actions. Note: Thresholds represent commonly used industry benchmarks and should be calibrated to institution-specific risk appetite and portfolio characteristics.

Table 1: ICRMA Metric Thresholds and Governance Actions

Metric	Green	Amber	Red	Layer	Recommended Action
PSI (Score)	< 0.10	0.10–0.25	> 0.25	Score Stability	RED: Investigate score shift; consider redevelopment
CSI (per feature)	< 0.10	0.10–0.25	> 0.25	Input Stability	RED: Flag feature; check data pipeline / policy
AUROC	> 0.65	0.60–0.65	< 0.60	Disc. Power	RED: Model redevelopment; interim override controls
Gini	> 0.30	0.20–0.30	< 0.20	Disc. Power	RED: Escalate to model owner; enhanced manual review
KS Statistic	> 0.25	0.20–0.25	< 0.20	Disc. Power	RED: Investigate threshold and policy implications
Calibration Ratio	0.80–1.20	0.60–1.50	<0.60 or >1.50	Calibration	RED: Apply PD scalar; notify capital function
Log Loss Trend	Stable	≤5% increase	≥10% increase	Calibration	RED: Recalibrate intercept; escalate to IFRS 9 team
# Features Drifted	≤ 2	3–4	> 4	Input Stability	RED: Declare model unreliable; initiate emergency review

4.3. Monitoring Frequency and Tier Classification

Monitoring frequency should be proportional to model materiality and the rate of expected environmental change. Table II summarises recommended frequencies by model tier.

Table 2: Recommended Monitoring Frequency by Model Tier

Tier	Definition	PSI/CSI Freq.	Disc. Power Freq.	Calibration Freq.
Tier 1 (High)	IRB / DFAST / IFRS 9	Monthly	Quarterly	Quarterly
Tier 2 (Medium)	Decision / Pricing models	Quarterly	Semi-annually	Semi-annually
Tier 3 (Low)	Low-materiality models	Semi-annually	Annually	Annually

4.4. Escalation Protocol

When a metric breaches its RED threshold: (i) the model owner is notified within 5 business days; (ii) a root-cause analysis report is submitted to the Model Risk Management (MRM) function within 20 business days; (iii) if degradation cannot be explained by a transient event, a model redevelopment timeline is proposed within 60 business days; (iv) interim manual override controls are imposed. AMBER breaches trigger enhanced monitoring frequency and documentation in the model's living risk register, with escalation if the AMBER condition persists for two or more consecutive periods.

5. Empirical Validation

5.1. Experimental Setup

To validate the sensitivity and complementarity of the proposed metrics, we construct a controlled simulation study using a synthetic retail loan portfolio. A logistic regression scorecard is trained on 10,000 development observations with eight features: age, annual income, debt-to-income ratio, credit score, loan amount, employment tenure, number of credit inquiries, and revolving credit utilisation. Six quarterly production cohorts (Q1–Q6, each n = 3,000) are generated with monotonically increasing covariate shift (drift_factor ∈ {0.0, 0.2, 0.5, 0.8, 1.2, 1.8}) and default rate shift (log-odds intercept ∈ {0.0, 0.1, 0.2, 0.4, 0.6, 0.9}). All code is implemented in Python 3.12 using scikit-learn 1.8.0.

The development portfolio has a default rate of 0.69%, consistent with a prime retail lending portfolio. Production default rates range from 0.50% (Q1) to 1.43% (Q6), reflecting the programmed default rate shift. The model achieves a development AUROC of 0.9359, Gini of 0.8717, and KS of 0.7408 performance levels typical of a well-developed prime credit scorecard.

The simulation design prioritises interpretability and controlled drift patterns over full replication of real-world portfolio complexity.

5.2. Population Stability Results

Table 3 and Fig. 1 present PSI results across all six quarters.

Table 3: PSI and Score Stability Monitoring Results by Quarter

Metric	Dev	Q1	Q2	Q3	Q4	Q5	Q6
PSI	—	0.0024 ✓✓	0.0026 ✓✓	0.0157 ✓✓	0.0211 ✓✓	0.0244 ✓✓	0.0565 ✓✓
Default Rate	0.69%	0.50%	0.77%	0.63%	1.00%	0.87%	1.43%

PSI values remain below 0.10 throughout all six quarters, classified as GREEN under ICRMA thresholds. This finding indicates that the logistic regression model's score distribution exhibits limited sensitivity to the applied covariate shift under the simulation design, a result consistent with the model's observed strong discriminatory performance throughout the monitoring horizon (see Section V.C). The maximum PSI of 0.0565 at Q6, while still GREEN, represents a 23.6× increase over Q1 (PSI = 0.0024), indicating a meaningful trend that warrants monitoring continuity.

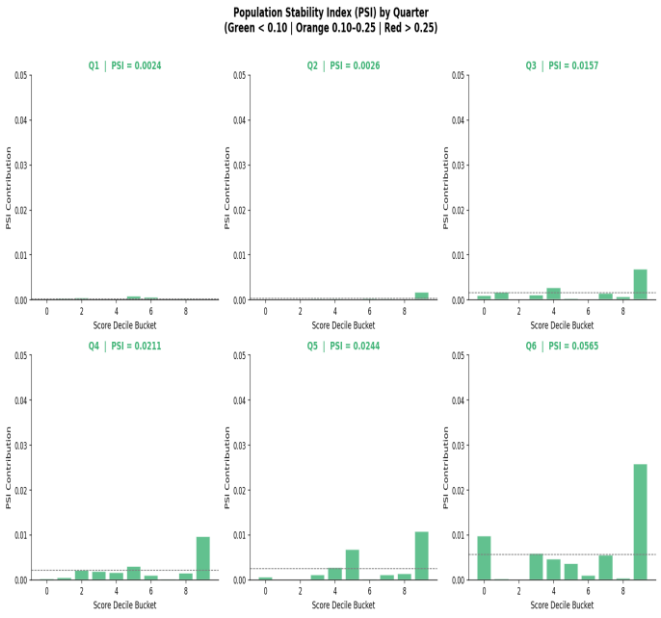

Fig 1: PSI Contribution by Score Decile Bucket across Q1–Q6. All Quarters Remain GREEN (PSI < 0.10). Note the Monotonic Increase in PSI Magnitude From Q1 to Q6.

CSI analysis reveals more pronounced feature-level instability. Table IV presents CSI values for all eight features across the six quarters.

Table 4: Characteristic Stability Index (CSI) by Feature and Quarter

Feature	Q1	Q2	Q3	Q4	Q5	Q6	Status Q6
Age	0.0025	0.0108	0.0406	0.1395	0.2147	0.5688	● RED
Annual Income	0.0050	0.0031	0.0216	0.0319	0.1059	0.2504	● RED
Debt-to-Income	0.0065	0.0213	0.1076	0.2381	0.4057	0.9459	● RED
Credit Score	0.0035	0.0110	0.0535	0.0955	0.2138	0.4000	● RED
Loan Amount	0.0030	0.0057	0.0185	0.0294	0.0456	0.1271	⚠□ AMBER
Employment Yrs	0.0060	0.0060	0.0256	0.0677	0.1367	0.2382	⚠□ AMBER
Num Inquiries	0.0005	0.0340	0.1165	0.2781	0.5866	1.0360	● RED
Revolving Util	0.0033	0.0124	0.0283	0.0908	0.1773	0.3085	● RED

Six of eight features are classified RED (CSI > 0.25) by Q6. The most severely drifted features are num_inquiries (CSI = 1.036) and debt-to-income (CSI = 0.946), both exceeding the PSI upper bound by a factor of 4×, underscoring that feature-level monitoring via CSI provides more granular diagnostic insight than score-level PSI alone.

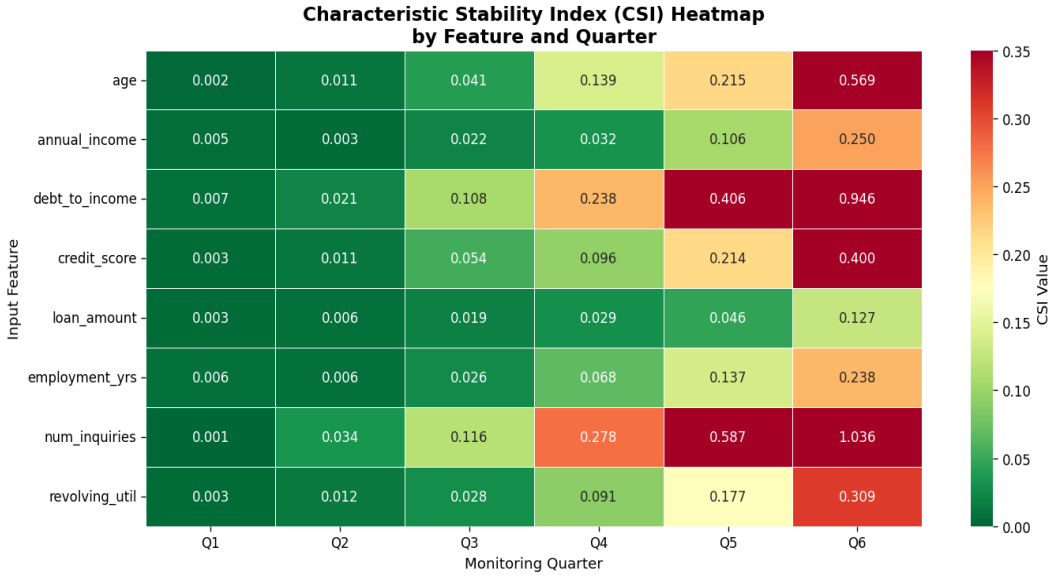

Fig 2: CSI Heatmap across Features and Quarters. Red Cells Indicate CSI > 0.25. Feature Drift Is Monotonically Increasing And Reaches Severe Levels (CSI > 0.50) For Age, Credit Score, Debt-To-Income, and Num_Inquiries By Q6.

5.3. Discriminatory Power Results

Table 5 and Fig. 3 present discriminatory power metrics across all periods.

Table 5: Discriminatory Power Metrics by Period

Metric	Dev	Q1	Q2	Q3	Q4	Q5	Q6
AUROC	0.9359	0.9450 ✓	0.9273 ✓	0.9192 ✓	0.9043 ✓	0.8722 ✓	0.9390 ✓
Gini	0.8717	0.8899 ✓	0.8545 ✓	0.8383 ✓	0.8087 ✓	0.7443 ✓	0.8780 ✓
KS	0.7408	0.8121 ✓	0.7935 ✓	0.7079 ✓	0.6714 ✓	0.6515 ✓	0.7874 ✓

Discriminatory power remains robustly GREEN throughout all six quarters. AUROC ranges from 0.8722 (Q5) to 0.9450 (Q1), all substantially above the RED threshold of 0.60. Gini ranges from 0.7443 to 0.8899, and KS from 0.6515 to 0.8121. The apparent recovery at Q6 (AUROC = 0.9390) may reflect improved separability under the simulation design rather than a general property of AUROC, which is theoretically invariant to class prevalence. This demonstrates a subtle limitation of rank-order metrics: they may not reflect miscalibration that is consequential for capital and provisioning purposes.

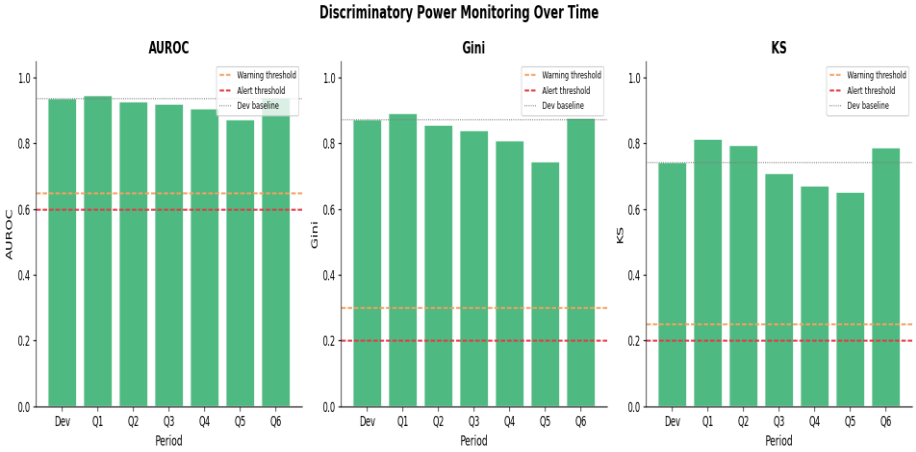

Fig 3: Discriminatory Power Metrics (AUROC, Gini, KS) By Monitoring Period. All Metrics Remain GREEN Throughout the Six-Quarter Horizon, Illustrating the Resilience Of Rank-Order Metrics to Moderate Covariate Shift.

5.4. Calibration Results

Table 6 and Fig. 4 present calibration results. Calibration exhibits a distinct and diagnostically important pattern, in contrast to the stable discriminatory power metrics.

Table 6: Calibration Monitoring Results by Period

Metric	Dev	Q1	Q2	Q3	Q4	Q5	Q6
EDR (%)	0.69	0.71	0.83	0.83	1.02	1.13	1.30
ADR (%)	0.69	0.50	0.77	0.63	1.00	0.87	1.43
Cal. Ratio	1.000	0.701 ⚠☐	0.921 ✓	0.759 ⚠☐	0.982 ✓	0.766 ⚠☐	1.103 ✓

The calibration ratio exhibits an alternating AMBER/GREEN pattern: Q1 (0.701, AMBER), Q2 (0.921, GREEN), Q3 (0.759, AMBER), Q4 (0.982, GREEN), Q5 (0.766, AMBER), Q6 (1.103, GREEN). The AMBER quarters (Q1, Q3, Q5) reflect under-prediction by the model (ADR < EDR), while GREEN quarters are driven by higher-than-expected actual default rates. This oscillating pattern illustrates a critical finding: calibration monitoring detects systematic miscalibration that is entirely invisible to discriminatory power metrics, which remain consistently GREEN throughout. This pattern highlights that calibration errors can persist even when rank-order performance remains stable, reinforcing the need for explicit calibration monitoring.

For institutions using PD estimates for IFRS 9 expected credit loss (ECL) provisioning, the Q1 calibration ratio of 0.701 implies potential overestimation of PD-driven provisions, holding other components of expected credit loss (e.g., LGD and EAD) constant. This supports the inclusion of calibration monitoring as a first-class component of any credit risk model monitoring framework.

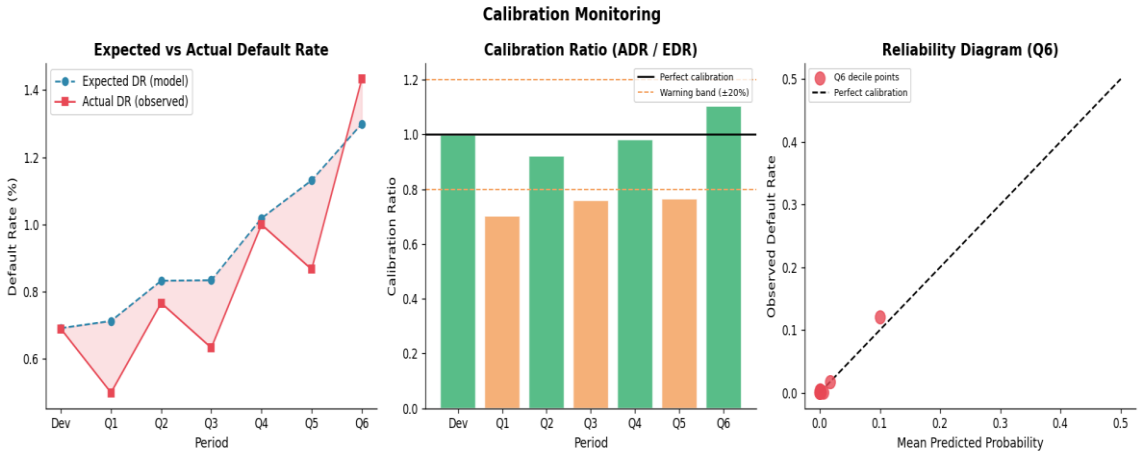

Fig 4: Calibration Monitoring: (Left) Expected vs. Actual Default Rate over Time; (Centre) Calibration Ratio with ±20% Warning Bands; (Right) Reliability Diagram for Q6 Showing Decile-Level Calibration.

5.5. Statistical Significance Testing

Two-sample KS tests for distributional equivalence confirm the CSI findings (Fig. 5). Zero features exhibit statistically significant drift at Q1. By Q2, seven of eight features are significant ($p < 0.05$), and all eight features are significant for Q3–Q6. This rapid onset of statistical significance at Q2 contrasts with the PSI score-level stability,

confirming that input-level monitoring via CSI and KS testing provides earlier warning than score-level PSI alone.

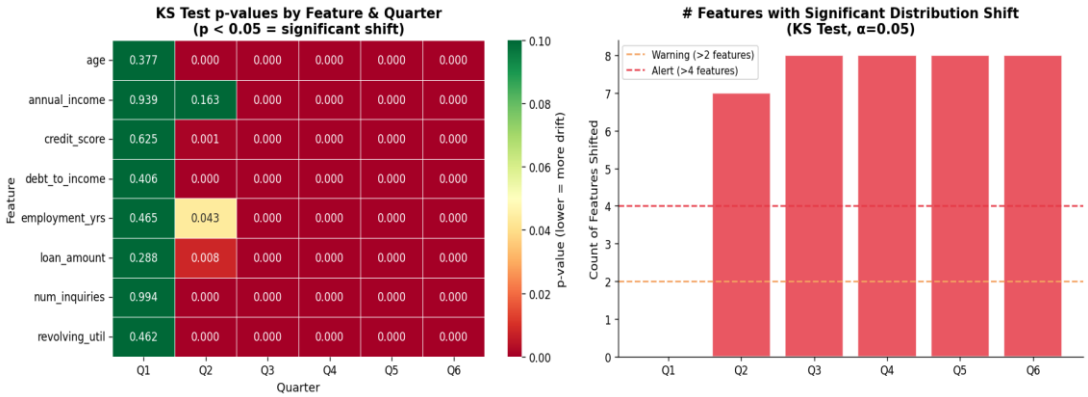

Fig 5: (Left) KS Test P-Values for Each Feature by Quarter. Green Cells Indicate P ≥ 0.05 (No Significant Shift); Red Cells Indicate P < 0.05 (Significant Shift). (Right) Count of Features with Significant Distributional Shift per Quarter.

5.6. Metric Complementarity and the Case for Multi-Dimensional Monitoring

Table 7 synthesises the overall RAG status across all four monitoring layers and six quarters, constituting the ICRMA traffic-light dashboard as described in Section IV.

Table 7: ICRMA Traffic-Light Dashboard Quarterly Summary

Monitoring Layer	Q1	Q2	Q3	Q4	Q5	Q6
Score Stability (PSI)	✓ GREEN	✓ GREEN	✓ GREEN	✓ GREEN	✓ GREEN	✓ GREEN
Input Stability (CSI)	✓ GREEN	⚠ AMBER	⚠ AMBER	● RED	● RED	● RED
Discriminatory Power	✓ GREEN	✓ GREEN	✓ GREEN	✓ GREEN	✓ GREEN	✓ GREEN
Calibration	⚠ AMBER	✓ GREEN	⚠ AMBER	✓ GREEN	⚠ AMBER	✓ GREEN

Traffic-Light Model Monitoring Dashboard
(Credit Risk Scorecard — Quarterly Review)

Quarter	PSI	AUROC	Gini	KS	CR
Q1	0.0024 GREEN	0.9450 GREEN	0.8899 GREEN	0.8121 GREEN	0.7012 RED
Q2	0.0026 GREEN	0.9273 GREEN	0.8545 GREEN	0.7935 GREEN	0.9205 GREEN
Q3	0.0157 GREEN	0.9192 GREEN	0.8383 GREEN	0.7079 GREEN	0.7592 RED
Q4	0.0211 GREEN	0.9043 GREEN	0.8087 GREEN	0.6714 GREEN	0.9816 GREEN
Q5	0.0244 GREEN	0.8722 GREEN	0.7443 GREEN	0.6515 GREEN	0.7658 RED
Q6	0.0565 GREEN	0.9390 GREEN	0.8780 GREEN	0.7874 GREEN	1.1025 AMBER

Fig 6: ICRMA Traffic-Light Dashboard Rendered from Executed Monitoring Notebook. Colour Coding: Green = No Action, Amber = Review, Red = Escalate.

The dashboard reveals a critical empirical finding: no single monitoring layer correctly characterises the model's status in all quarters. An institution monitoring only discriminatory power (the most common single-metric approach) would observe ALL GREEN across all six quarters and would miss the systematic calibration deterioration and severe feature drift present from Q2 onwards. Conversely, an institution monitoring only CSI would trigger escalation from Q4, potentially unnecessary if discriminatory power is maintained and calibration is periodically corrected through PD scaling.

This non-redundancy across dimensions directly supports the position that evaluation beyond AUROC is necessary, and formalises that argument within a governance architecture with defined thresholds and escalation procedures.

6. Practical Implementation Considerations

6.1. Data Infrastructure Requirements

Effective model monitoring requires access to three data streams: (i) model input features at time of scoring (preserved in a model input audit log); (ii) model output scores; and (iii) observed outcome data, typically available with a 12–24 month lag for consumer credit. PSI and CSI can be computed at origination frequency without waiting for outcomes; discriminatory power and calibration require outcome data and are thus computed at vintage maturity. Infrastructure to join all three streams at account level is a prerequisite.

6.2. Threshold Calibration and Multiple Testing

The thresholds in Table I represent industry conventions [9], [5]. In practice, institutions may combine statistical thresholds with expert judgment and business impact considerations. Institutions should consider calibrating thresholds through bootstrap simulation of expected metric variance under the null hypothesis, setting thresholds at the 95th or 99th percentile of the null distribution. When monitoring p features simultaneously, the family-wise error rate at $\alpha = 0.05$ for $p = 8$ is $1 - 0.95^8 \approx 0.34$. Benjamini–Hochberg correction is recommended for formal statistical tests, while PSI/CSI thresholds should be treated as practical effect-size benchmarks rather than hypothesis tests.

6.3. Extension to Complex Models

The monitoring metrics discussed are applicable to any scoring model producing continuous risk estimates. For gradient boosting and neural network models, discriminatory power and calibration monitoring are straightforward. CSI attribution is more complex, and SHAP value drift analysis [23] may be required to attribute score instability to specific features. Log loss monitoring is particularly valuable for neural network models where calibration may degrade without rank-order deterioration.

6.4. Limitations

Our simulation uses logistic regression and synthetic data, which, while enabling controlled comparison, may not capture seasonal effects, vintage effects, or correlations introduced by macroeconomic cycles. The alternating calibration pattern observed may be an artefact of the synthetic drift schedule. Real portfolio monitoring may exhibit more persistent calibration regimes, particularly during economic downturns. Therefore, the results should be interpreted as illustrative of metric complementarity rather than directly representative of production environments.

7. Conclusion

This paper has presented the Integrated Credit Risk Monitoring Architecture (ICRMA), a multi-dimensional framework for the ongoing monitoring of credit risk models. By formalising PSI, CSI, AUROC, Gini, KS statistic, and calibration ratio within a unified mathematical framework and linking them to a governance traffic-light system, ICRMA aims to provide both theoretical rigour and operational practicality.

Empirical validation on synthetic retail lending data with programmatically controlled drift produces three primary findings. First, a logistic regression scorecard with development AUROC = 0.9359 maintains GREEN discriminatory power across all six quarters despite severe CSI drift (num_inquiries CSI = 1.036, debt-to-income CSI = 0.946 at Q6), confirming that rank-order metrics alone are insufficient. Second, calibration monitoring identifies systematic miscalibration in Q1, Q3, and Q5, with calibration ratios of 0.701, 0.759, and 0.766, materially relevant for IFRS 9 expected credit loss (ECL) provisioning, that are completely invisible to discriminatory power metrics. Third, KS feature testing identifies significant drift for seven of eight features as early as Q2, while score-level PSI remains GREEN, confirming that CSI provides earlier and more granular warning than PSI.

These findings empirically substantiate the broader position that evaluation beyond AUROC is operationally necessary, and extend that work by embedding the complementary metrics within a governance architecture with defined thresholds, escalation protocols, and regulatory alignment. Future work will examine the extension of ICRMA to ensemble and deep learning credit models, the integration of SHAP-based feature attribution with CSI, and the development of adaptive monitoring intervals driven by real-time drift detection algorithms.

References

[1] Basel Committee on Banking Supervision, "International Convergence of Capital Measurement and Capital Standards (Basel II)," Bank for International Settlements, June 2006.

[2] Basel Committee on Banking Supervision, "Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems," Bank for International Settlements, Dec. 2010.

[3] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," Pattern Recognition, vol. 45, no. 1, pp. 521–530, Jan. 2012, doi:10.1016/j.patcog.2011.06.019.

[4] M. Baena-García, J. del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavaldà, and R. Morales-Bueno, "Early drift detection method," in Proc. 4th Int. Workshop on Knowledge Discovery from Data Streams, 2006.

[5] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, "Supervisory Guidance on Model Risk Management," SR Letter 11-7 / OCC Bulletin 2011-12, Apr. 2011.

[6] European Banking Authority, "Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures," EBA/GL/2017/16, Nov. 2017.

[7] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, vol. 7, no. 2, pp. 179–188, 1936.

[8] W. H. Beaver, "Financial ratios as predictors of failure," Journal of Accounting Research, vol. 4, pp. 71–111, 1966.

[9] N. Siddiqi, Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Hoboken, NJ: Wiley, 2006.

[10] R. Anderson, The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. Oxford, U.K.: Oxford Univ. Press, 2007.

[11] D. Tasche, "Validation of Internal Rating Systems and PD Estimates," in The Analytics of Risk Model Validation, G. N. Christodoulakis and S. Satchell, Eds. London, U.K.: Elsevier, 2008, pp. 169–196.

[12] B. Engelmann, E. Hayden, and D. Tasche, "Measuring the Discriminative Power of Rating Systems," Bundesbank Discussion Paper Series 2, No. 01/2003, 2003.

[13] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: A review," Journal of the Royal Statistical Society: Series A, vol. 160, no. 3, pp. 523–541, 1997, doi:10.1111/j.1467-985X.1997.00078.x

[14] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," Journal of Statistical Planning and Inference, vol. 90, no. 2, pp. 227–244, Oct. 2000.

[15] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., Dataset Shift in Machine Learning. Cambridge, MA: MIT Press, 2009.

[16] S. Kullback and R. A. Leibler, "On information and sufficiency," Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, 1951.

[17] Basel Committee on Banking Supervision, "Studies on the Validation of Internal Rating Systems," Working Paper No. 14, May 2005.

[18] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed. New York, NY: Wiley, 2000.

[19] S. P. Pathi, "Model Evaluation Beyond AUC: A Comparative Study of Somers' D, Log Loss, Population Stability Index (PSI), and Kolmogorov–Smirnov (KS) Statistic in Credit Risk and Healthcare Prediction Models," IJETCSIT, pp. 106–111, Oct. 2025, doi: 10.63282/3050-9246/ICRTCSIT-113.

[20] Basel Committee on Banking Supervision, "Principles for Effective Risk Data Aggregation and Risk Reporting (BCBS 239)," Bank for International Settlements, Jan. 2013.

[21] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Computing Surveys, vol. 46, no. 4, pp. 1–37, Mar. 2014, doi: 10.1145/2523813.

[22] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Machine Learning (ICML), Bonn, Germany, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

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