ISSN: 2663-5798 || Arab Journal for Scientific Publishing

الإصدار التاسع – العدد الثامن والثمانون
تاريخ الإصدار: 2 – شباط – 2026م
www.ajsp.net

"Real-Time Data Quality Indicators for Monitoring Predictive Power Degradation in Machine Learning Credit Scoring Models: An Extended DAMA Framework Perspective"

Researcher:

Mohammed Kamel AbdulRahim Asaad

Abdul Latif Jameel United Finance (ALJUF), Jeddah, Saudi Arabia

Certification: Master Certified Data Management Professional (CDMP)

1. Abstract

Machine-learning (ML) credit scoring models may lose predictive power after deployment due to changes in applicant populations, macroeconomic conditions, data pipelines, and labeling dynamics. In regulated financial institutions, delayed detection of performance degradation increases model risk and can translate into credit losses and governance findings. This study proposes a tool-agnostic, multi-frequency monitoring control system that integrates Data Quality Indicators (DQIs) with Statistical Process Control (SPC) to provide early warning signals across weekly data integrity checks, monthly stability monitoring, and quarterly predictive power assessment. The framework is positioned as a design-science governance artifact that operationalizes "predictive quality" as a measurable aspect of the DAMA-DMBOK Reasonableness dimension, linking input-data health to downstream model behavior. We formalize indicator definitions and provide full pseudocode for computation of missingness control charts, outlier drift checks, Population Stability Index (PSI), Characteristic Stability Index (CSI), Kolmogorov-Smirnov (KS), AUC/Gini, and an example dynamic delinquency-rate (DDR) diagnostic. To improve statistical rigor, we recommend reporting uncertainty for discriminatory metrics using bootstrap confidence intervals and calibrating thresholds through baseline sampling rather than fixed rules. The manuscript demonstrates the monitoring workflow using transparent synthetic aggregates to illustrate numerical traceability without exposing proprietary data, and it provides an escalation playbook for triage, investigation, and model action.

2. Keywords: Credit scoring, Model monitoring, Data quality indicators, And drift detection, Statistical process control, Population stability index

3. Introduction

Machine learning (ML) credit scoring models are deployed to automate risk ranking and decisioning at scale. Once deployed, their predictive ability can deteriorate due to changes in applicant populations, macroeconomic conditions, product strategy, data pipelines, and labeling dynamics (concept drift) (Gama et al., 2014; Quiñonero-Candela et al., 2009). In regulated financial institutions, delayed detection of performance degradation may translate into immediate credit losses and increased model risk.

Model risk management guidance emphasizes ongoing monitoring and appropriate change control for models used in decisioning, including performance tracking and identification of material deterioration (Board of Governors of the Federal Reserve System, 2011; Office of the Comptroller of the Currency, 2021; Office of the Comptroller of the Currency, 2025). In practice, discriminatory power measures (AUC/Gini, KS) and stability measures (PSI/CSI) are widely used (Siddiqi, 2006; du Pisanie et al., 2020; Taplin & Hunt, 2019; Potgieter et al., 2023), but they are frequently applied as isolated checks rather than an integrated control system.

4. Study Problem

Standard industry practice typically relies on periodic, retrospective validations often conducted annually which fail to identify performance decay as it happens. This creates a "blind spot" where a model may continue to drive decisions while its predictive power is actively degrading, leading to immediate and potentially massive financial losses. The problem is further compounded by the traditional focus of Data Quality activities on input data (completeness, uniqueness) rather than predictive quality. As established in the DAMA-DMBOK (2017), processes require Statistical Process Control (SPC) to remain stable (Board of Governors of the Federal Reserve System, 2011). In the context of credit scoring, the "maturing effect" of a portfolio means that risk accumulates over time (DAMA International, 2017). If newer cohorts reach elevated bad rates faster than historical benchmarks a trend detectable through Dynamic Delinquency Rate (DDR) analysis the model's predictive integrity is compromised (DAMA International, 2017). Therefore, the core problem is: How can financial institutions transition from reactive, annual reviews to a proactive, real-time Data Quality Indicator (DQI) framework that utilizes SPC to detect predictive power degradation?

5. Study Hypotheses

- H1: A framework of automated DQIs derived from model development metrics can proactively identify the real-time degradation of a model's predictive power before significant financial losses occur.
- H2: The application of Statistical Process Control (SPC) thresholds (Green, Amber, and Red zones) effectively distinguishes between expected population fluctuations and "special cause" structural model failure (Siddiqi, 2006).
- H3: Materializing the Reasonableness dimension of data quality through stability indices (PSI/CSI) provides an earlier warning signal of model decay than traditional dimensions like completeness or accuracy (Board of Governors of the Federal Reserve System, 2011; Siddiqi, 2006).

6. Study Objectives

- Identify and operationalize key metrics from the development phase such as Gini, KS, and PSI into automated Data Quality Indicators (DQIs) (DAMA International, 2017; Siddiqi, 2006).
- Integrate Statistical Process Control (SPC) methodologies to establish mathematical tolerances and control charts that distinguish between expected variation and "special cause" model degradation.
- Establish a multi-frequency monitoring schedule (daily, weekly, monthly, quarterly) for different DQIs to provide the earliest possible warning of performance decay.
- Demonstrate universal applicability across diverse machine learning architectures, including Logistic Regression, Random Forests, and Gradient Boosting.

7. Study Significance

This study contributes to the field of data management by extending the traditional DAMA Data Quality dimensions. It moves beyond the core six dimensions such as completeness and accuracy to materialize "predictive quality" within the Reasonableness dimension. While banks and financial institutions typically focus on standard input quality, they often overlook reasonableness; this research establishes it as a core requirement for model-driven organizations to ensure that model outputs and population distributions remain logically aligned with the development baseline and real-world risk patterns (DAMA International, 2017; Siddiqi, 2006).

8. Study Limits

Thematic Scope: Limited to Machine Learning credit scoring models (e.g., Logistic Regression, Random Forest, XGBoost) in financial services.

Methodological Scope: The demonstration of the DQI framework utilizes synthetic datasets modeled after real-world credit risk patterns to ensure replicability without compromising proprietary data.

Dimensional Scope: Primary focus on the Reasonableness and Predictive Quality dimensions.

Temporal Scope: Assumes a typical credit risk environment with an observation lag of 6 to 12 months for outcome crystallization (DAMA International, 2017).

8.1 Limitations and Future Work:

The main limitation is empirical: the demonstration uses synthetic data for transparency. While this supports reproducibility of computations and framework mechanics, it does not substitute for validation on real portfolios or publicly available credit datasets. A second limitation is operational: outcome labels may crystallize with a lag (typically 6–12 months), so quarterly discriminatory metrics must be interpreted with sampling uncertainty and cohort maturity effects. Future work should replicate the indicator suite on open or anonymized datasets where label timing can be approximated; compare the proposed SPC-based monitoring against alternative drift tests and change-point methods; evaluate alert trade-offs (false alarms vs. late

detection) under governance constraints; and study robustness across model families, products, and macro-regimes. Providing an open reference implementation and full synthetic artifacts would further strengthen reproducibility and adoption.

9. Key Terms and Definitions

Predictive Power: The ability of a credit scoring model to accurately distinguish between "Good" and "Bad" accounts, commonly measured using AUC and derived measures such as Gini. (Fawcett, 2006; Siddiqi, 2006).

Reasonableness: A DAMA data quality dimension assessing whether values are logical and consistent with business context and historical baselines; in this study it is operationalized via stability and predictive quality indicators. (DAMA International, 2017).

Credit Scoring Model: A statistically sound and objective decision-support tool designed to predict the likelihood of a specific credit event such as default, delinquency, or bankruptcy occurring within a predefined observation window. In the context of this study, these models leverage advanced machine learning architectures (e.g., Logistic Regression, Random Forests, XGBoost) to transform complex, multi-dimensional input data into a single numerical score. This score represents the creditworthiness of an applicant and serves as the primary mechanism for automated credit decisioning, risk-based pricing, and portfolio management within financial institutions (DAMA International, 2017).

- Data Quality Indicator (DQI): An operational metric derived from observed data and/or model outputs that signals potential deterioration in data fitness-for-use for the scoring process.
- Predictive Quality: An operational sub-construct of DAMA's Reasonableness dimension defined as the degree to which model outputs and their relationship to outcomes remain statistically consistent with development-period behavior under stable process conditions.
- Population Stability Index (PSI): A dimensionless index measuring distribution shift in the model score (or another binned variable) between baseline and monitoring windows.
- Characteristic Stability Index (CSI): A PSI-like index computed for a specific input feature to quantify feature-level distribution shift between baseline and monitoring windows.
- Kolmogorov-Smirnov (KS) statistic: The maximum separation between cumulative score distributions for goods vs. bads; used as a discriminatory power measure in credit scoring.
- Area under the ROC Curve (AUC) and Gini: Discriminatory power measures where Gini is commonly computed as Gini = 2·AUC − 1.. (Fawcett, 2006; Hanley & McNeil, 1982).
- Statistical Process Control (SPC): A control method that uses baseline-calibrated limits to distinguish common-cause variation from special-cause signals in monitored indicators.

10. Theoretical Framework and Previous Studies

10.1 Background and Related Work

This work intersects three literatures: (i) credit risk model governance and monitoring, (ii) data quality management (DAMA), and (iii) drift detection / performance monitoring in ML systems.

10.2 Credit risk model governance and monitoring

Model risk management standards emphasize ongoing monitoring as part of the model lifecycle (Board of Governors of the Federal Reserve System, 2011; Office of the Comptroller of the Currency, 2021; Office of the Comptroller of the Currency, 2025). Within credit scoring, discriminatory power measures (AUC/Gini, KS) and stability indices (PSI/CSI) are widely used as monitoring signals (Siddiqi, 2006; du Pisanie et al., 2023; Taplin & Hunt, 2019; du Pisanie et al., 2020; Potgieter et

al., 2023). Recent statistical work reviews population stability testing procedures and discusses limitations and alternatives to PSI-like indices (du Pisanie et al., 2023; du Pisanie et al., 2020; Potgieter et al., 2023; Taplin & Hunt, 2019).

10.3 DAMA data quality and the need to represent predictive behavior

DAMA-DMBOK positions data quality as fitness-for-use and emphasizes measurable dimensions (e.g., accuracy, completeness, timeliness, consistency, uniqueness, validity) and the operational need for controllable metrics (DAMA International, 2017). In ML scoring, "fitness for use" includes not only input conformance but also the stability of model outputs and their relationship to outcomes; this motivates formalizing Predictive Quality as a governance construct.

10.4 Drift and monitoring in ML systems

In ML, distribution shift (covariate drift), prior probability shift, and concept drift are well-studied phenomena (Gama et al., 2014; Quiñonero-Candela et al., 2009). Production ML guidance recommends layered monitoring (data validation, drift detection, performance tracking) with alerting and operational response playbooks (Breck et al., 2017). The credit scoring context adds complications: delayed labels and regulatory expectations around auditability and conservative change control.

10.5 Extended DAMA Framing: Defining Predictive Quality

The manuscript uses DAMA's data quality dimensions as a governance vocabulary and proposes an explicit extension suited for ML credit scoring oversight.

10.5.1 Definition

Definition (Predictive Quality): For a deployed scoring model M, Predictive Quality is the degree to which the joint behavior of model outputs and observed outcomes remains statistically consistent with a reference (development or last revalidation) period, within tolerated process variation. Predictive Quality is assessed through a bounded set of indicators that jointly capture (a) stability of score distributions, (b) stability of key input feature distributions, and (c) discriminatory power of the score with respect to an outcome proxy.

10.5.2 positioning inside DAMA

Predictive Quality is treated as an explicit sub-construct of Reasonableness, because Reasonableness captures whether values "make sense" in business/statistical context. Here, the model's score and its relationship to outcomes must remain reasonable relative to a baseline regime. Predictive Quality does not replace DAMA's core dimensions; it operationalizes Reasonableness for model outputs.

10.5.3 mapping indicators to DAMA constructs

Table 1. Mapping of proposed indicators to DAMA Data Quality constructs (extended Reasonableness via Predictive Quality).

Indicator family	Examples	Primary DAMA linkage
Input integrity (weekly)	Missing rate drift; outlier drift	Completeness, Validity, Accuracy
Stability indices (monthly)	PSI (scores); CSI (features)	Consistency, Reasonableness
Discriminatory power (quarterly)	KS; AUC/Gini; scoreband bad rates	Reasonableness (Predictive Quality)
Outcome dynamics (monthly/quarterly)	DDR; vintage curves	Reasonableness, Timeliness

10.6 DQI Control System Design

The framework is model-family agnostic and tool-agnostic. It requires: (i) model scores, (ii) a selected set of monitored input features, and (iii) an outcome proxy available with acceptable delay (e.g., 30+ DPD).

10.6.1 multi-frequency monitoring schedule

Table 2. Proposed multi-frequency schedule.

Frequency	Indicator scope	Primary purpose
Daily/near real-time (optional)	Pipeline health & freshness checks	Detect ETL/job failures quickly
Weekly	Input integrity + outliers	Detect special-cause data issues early
Monthly	Population/feature stability	Detect structural drift in distributions
Quarterly	Discriminatory power + outcome dynamics	Confirm performance deterioration using outcome proxies

10.6.2 Statistical Process Control (SPC) calibration

Rather than relying on fixed thresholds, the framework calibrates control limits from a baseline window (development or last revalidation period). For an indicator X(t), baseline mean μ and standard deviation σ are computed (or robust alternatives), and control limits are defined as (Montgomery, 2013):

$CL = \mu$

$UCL = \mu + k \cdot \sigma$

$LCL = \mu - k \cdot \sigma$

where k is typically 3 for Shewhart charts.

For indicators with non-normal distributions, quantile-based limits or transformations may be used. This revision also reports bootstrap confidence intervals for AUC and KS to quantify sampling uncertainty.

10.7 Indicator Definitions

10.7.1 Missingness drift (weekly)

For a monitored variable x_j in week t, the missing rate MR_j(t) is:

MR_j(t) = (# records where x_j is NULL/invalid in week t) / (total # records in week t).

10.7.2 Outlier drift (weekly)

Outlier drift targets silent pipeline failures (unit errors, truncation, and extraction changes). For a numeric variable x with baseline mean μ_x and standard deviation σ_x (or robust alternatives), a z-score is:

$z_x(t) = (\bar{x}(t) - \mu_x) / \sigma_x$, where $\bar{x}(t)$ is the weekly mean (or median).

Outlier rules can be expressed via SPC limits on z_x(t), or via IQR-based fences. Thresholds should be calibrated and justified in context.

10.7.3 PSI and CSI (monthly)

PSI compares score distributions in monitoring vs. baseline across bins; CSI applies the same index to a feature distribution. For bins i=1..k with baseline proportions e_i and monitoring proportions a_i:

$$PSI = \Sigma_i (a_i - e_i) \cdot \ln(a_i / e_i)$$

$$CSI = \Sigma_i (a_i - e_i) \cdot \ln(a_i / e_i)$$

PSI/CSI are dimensionless indices (not percentages). Rule-of-thumb interpretation bands are common in scorecard monitoring (e.g., <~0.1 small shift; ~0.1–0.25 moderate; >~0.25 material), but should be treated as heuristics and calibrated to each portfolio's baseline and risk appetite (Siddiqi, 2006; du Pisanie et al., 2023; du Pisanie et al., 2020; Potgieter et al., 2023; Taplin & Hunt, 2019).

10.7.4 KS statistic and AUC/Gini (quarterly)

KS is the maximum absolute difference between cumulative distributions of scores for goods vs. bads. AUC is the area under the ROC curve; Gini is commonly computed as Gini = 2·AUC − 1 (Hand & Till, 2001).

10.7.5 Outcome dynamics: scoreband bad rates and DDR

When full default labels are delayed, early delinquency proxies (e.g., 30+ DPD within a horizon) can be used to estimate discriminatory power and to track dynamic delinquency rate (DDR) trends by vintage or origination cohort.

11. Methodology

This paper follows a design-science structure: the primary artifact is a monitoring framework (control system) rather than a novel classifier. The empirical component is a reproducible synthetic demonstration intended to show how indicators can be computed, controlled with SPC, and interpreted together. Results are therefore presented as a worked example and not as universal empirical claims.

11.1 Synthetic data design and reproducibility

To ensure auditability without exposing proprietary portfolios, baseline and monitoring datasets are generated with explicit distributions over: (i) score bands, (ii) a monitored feature (financing term; iii) a numeric credit variable (installment amount; iv) missingness in a key ratio (DTI), and (v) a binary outcome proxy. Appendices provide aggregate artifacts and pseudocode sufficient to reproduce the reported indicators.

11.2 External validation protocol using public credit datasets (optional)

To strengthen external validity beyond the synthetic demonstration, the same monitoring design can be replicated on public credit datasets that contain borrower characteristics and binary default outcomes. Suitable examples include benchmark datasets used in credit scoring research (e.g., the German Credit dataset family and real-world benchmarking studies), provided the analyst performs appropriate preprocessing and documents train/validation splits. The objective is not to optimize predictive performance, but to verify that the proposed multi-frequency DQI signal chain (weekly integrity → monthly stability → quarterly predictive power) behaves consistently under real data-generating processes.

A practical replication protocol is: (1) build a baseline scorecard/ML model on a development window; (2) create sequential monitoring windows (e.g., monthly) by time-slicing or by simulating production drift through covariate perturbations; (3) compute PSI/CSI on stable feature bins derived from development; (4) compute quarterly discriminatory power (KS, AUC/Gini) on outcomes, with confidence intervals; and (5) operationalize action thresholds via SPC control limits calibrated on development variability.

This replication design aligns with established credit scoring benchmarking literature and supports comparative analysis across portfolios and model families without disclosing proprietary institutional data (Baesens et al., 2003; Lessmann et al., 2015).

11.3 Statistical uncertainty (bootstrap)

For performance indicators computed from finite samples (AUC, KS), this revision reports 95% bootstrap confidence intervals to avoid overconfident point-estimate claims. Bootstrapping resamples observations with replacement and recomputes the metric to estimate sampling variability.

12. Study Tool and Implementation Guidance

The framework can be implemented as (i) a code-first monitoring service (e.g., Python/R jobs scheduled via orchestration) or (ii) configuration inside a data quality platform. To preserve scientific neutrality, formulas, pseudocode, and table-level artifacts are provided so the indicator suite can be implemented in any environment.

12.1 Reference architecture (conceptual)

A minimal architecture consists of: (1) data ingestion from scoring events and outcome systems, (2) a metric computation layer that produces weekly/monthly/quarterly indicator tables, (3) an alerting layer implementing SPC rules and escalation playbooks, and (4) dashboards for visualization and audit trails.

12.2 Pseudocode (excerpt)

Inputs:

baseline_window, monitoring_window

monitored_features, score, outcome_proxy

Weekly:

compute MR_j(t) for each feature j

compute outlier statistics for numeric features

apply SPC rules -> alerts

Monthly:

compute PSI (scores) using fixed bins

compute CSI (feature_j) for key features

apply thresholds / SPC -> drift alerts

Quarterly:

compute KS and AUC/Gini using outcome_proxy

compute scoreband bad rates and DDR trend signals

bootstrap CI for AUC and KS (optional)

apply escalation policy if deterioration is statistically and operationally material

13. Results

13.1 Baseline vs. monitoring distributions

Table 3. Score band distribution used for PSI computation (baseline vs. monitoring).

Score band	Baseline %	Monitoring %
300-500	10.0	20.0
501-650	30.0	35.0
651-750	45.0	30.0
751-900	15.0	15.0

PSI (scores) = 0.138.

Table 4. Financing term distribution used for CSI computation (baseline vs. monitoring).

Term (months)	Baseline %	Monitoring %
12.0	60.0	10.0
24.0	25.0	20.0
36.0	10.0	30.0
48.0	5.0	40.0

CSI (financing term) = 1.855.

13.2 Weekly integrity indicators

Figure 1 presents a weekly missing-rate control chart for the DTI variable. Control limits are calibrated from the baseline window; weeks exceeding UCL are treated as special-cause signals requiring investigation.

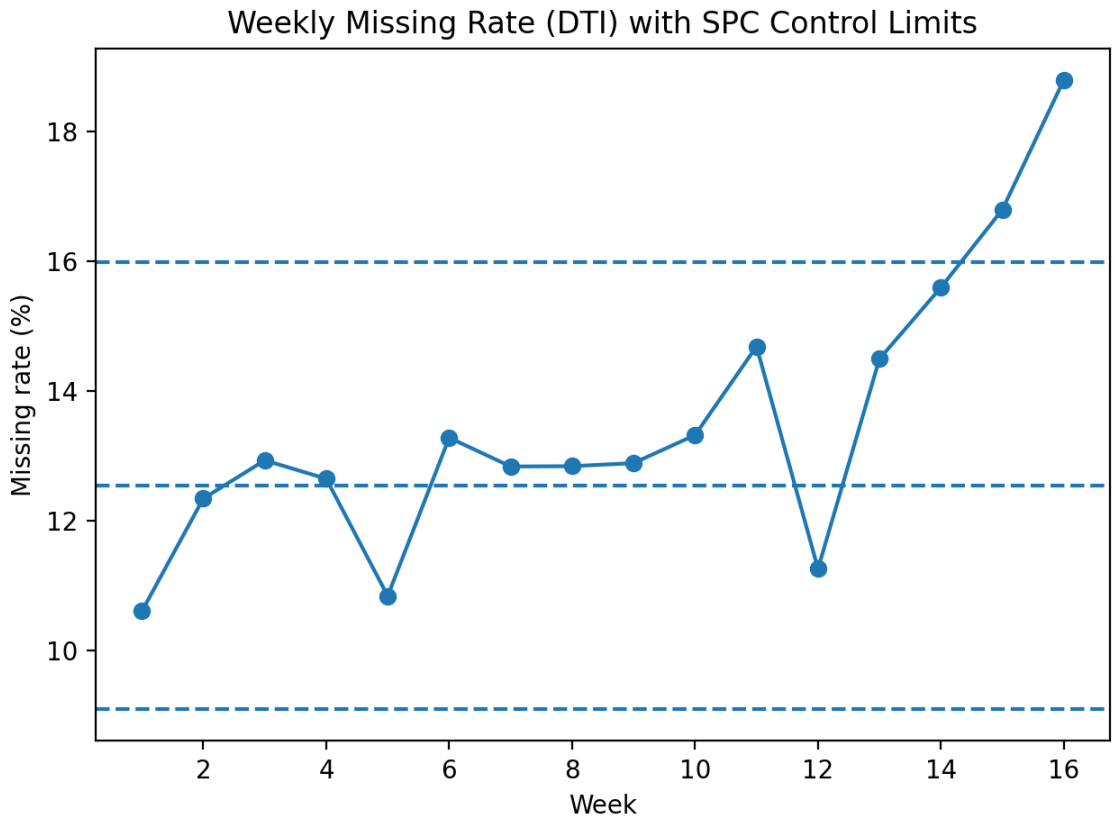

Figure 1. Weekly missing rate (DTI) with baseline-calibrated SPC limits.

Figure 2 shows installment amount distributions with monitoring-period outliers. Relative to the baseline mean and standard deviation, a representative extreme value of 6,500 yields a z-score of approximately 6.0, indicating atypical behavior that warrants upstream validation.

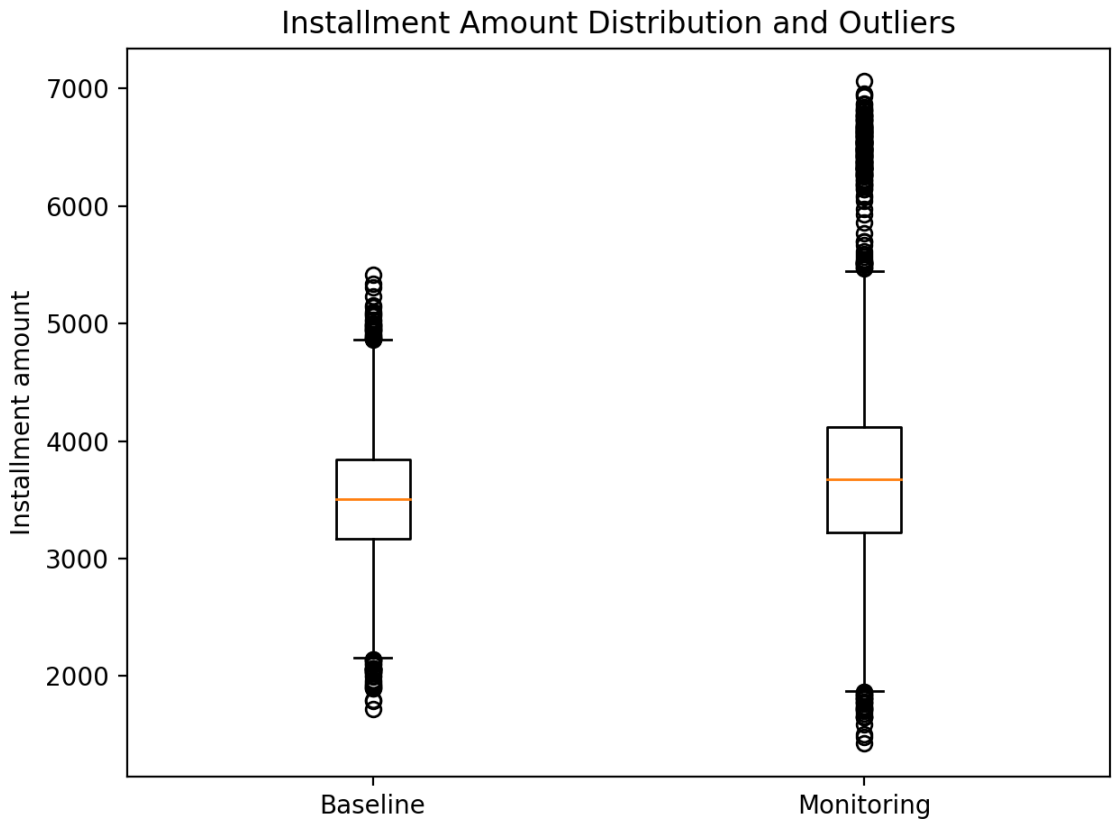

Figure 2. Installment amount distribution (baseline vs. monitoring) highlighting outliers.

13.3 Monthly stability indices

Figure 3 and Figure 4 visualize the distribution shifts underlying PSI and CSI.

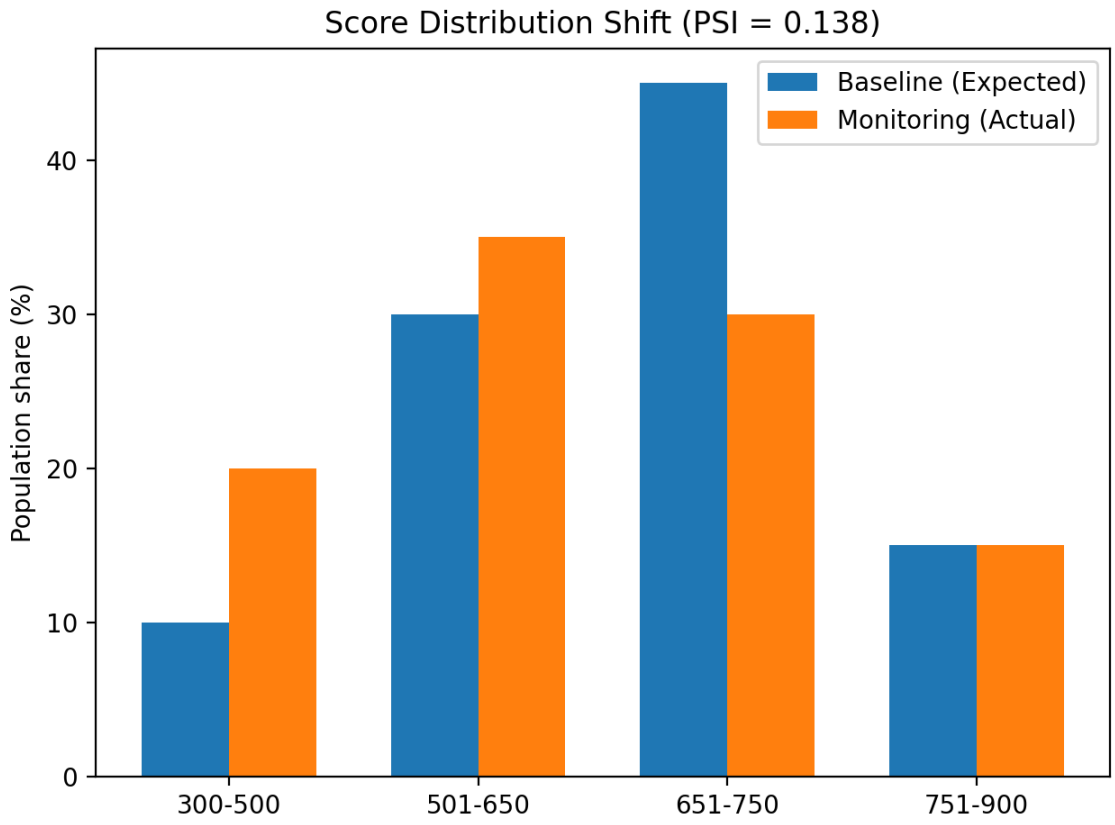

Figure 3. Score distribution shift used to compute PSI.

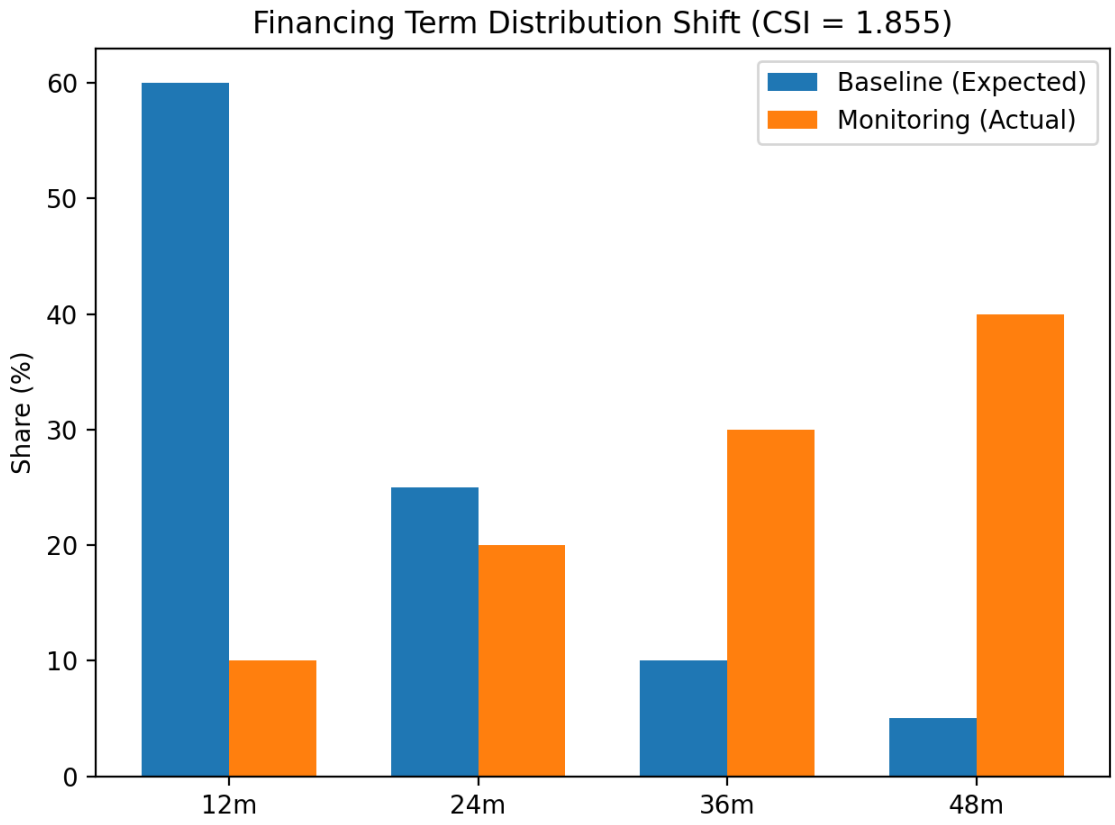

Figure 4. Financing term distribution shift used to compute CSI.

13.4 Quarterly discriminatory power (AUC/Gini, KS)

Table 5. Quarterly discriminatory power indicators (AUC/Gini and KS): baseline vs. monitoring.

Metric	Baseline	Monitoring
AUC	0.743	0.634
Gini	0.485	0.267
KS	0.386	0.237
Bad rate	0.188	0.310

Table 6. Bootstrap confidence intervals for AUC and KS (baseline vs. monitoring).

Metric	Baseline	Monitoring
AUC (95% CI)	0.743 [0.729, 0.755]	0.634 [0.622, 0.645]
KS (95% CI)	0.387 [0.365, 0.412]	0.238 [0.219, 0.258]

Figure 5 shows KS curves for baseline and monitoring.

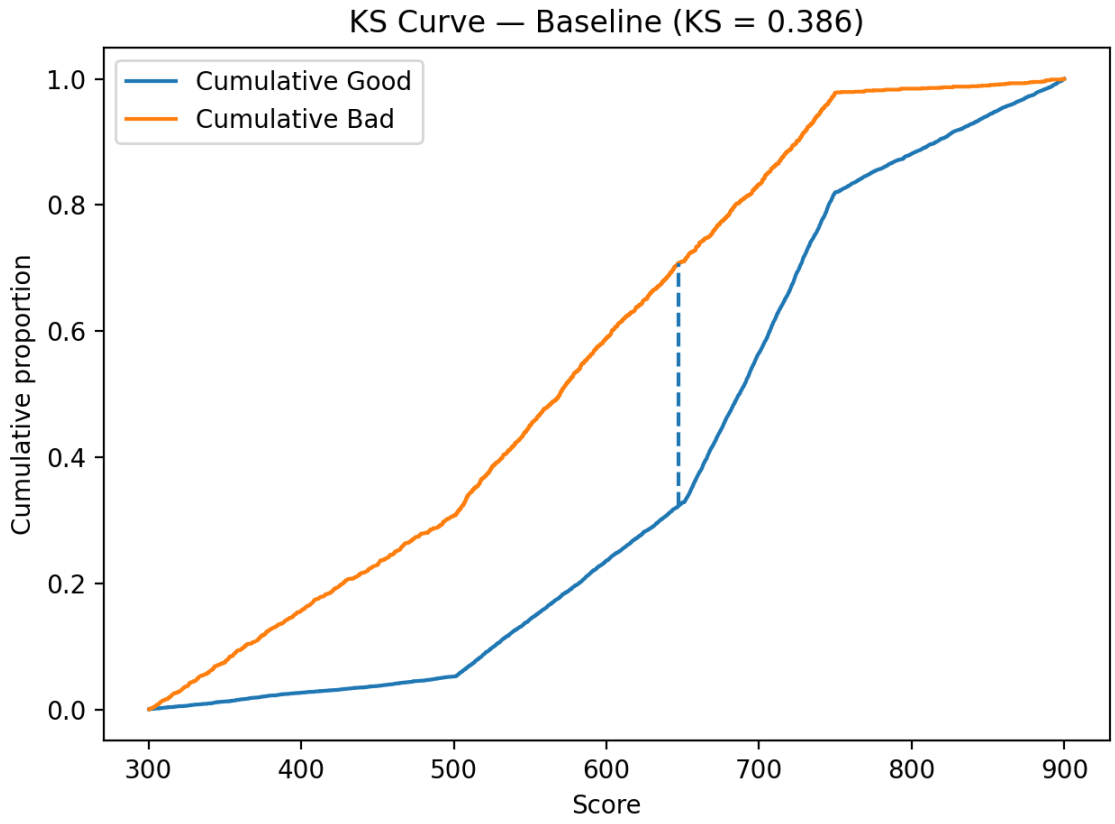

Figure 5a. KS curve - baseline period.

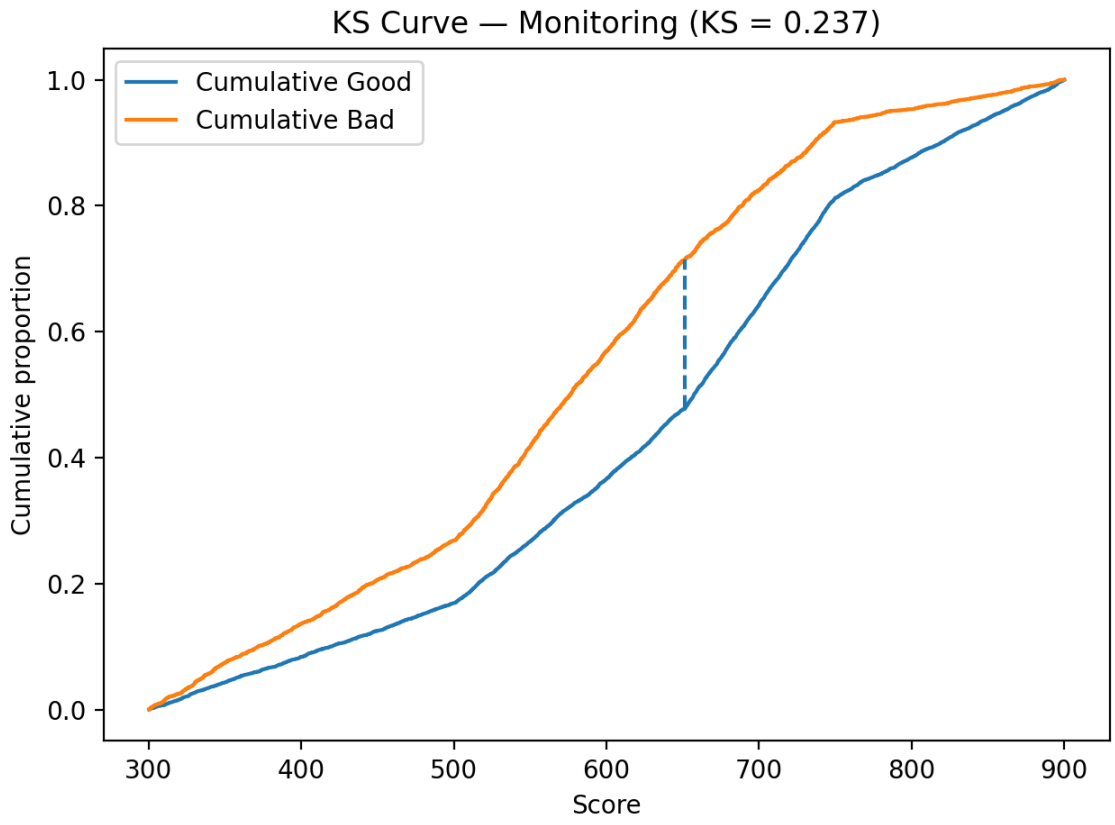

Figure 5b. KS curve - monitoring period.

Figure 6 shows ROC curves for baseline vs. monitoring.

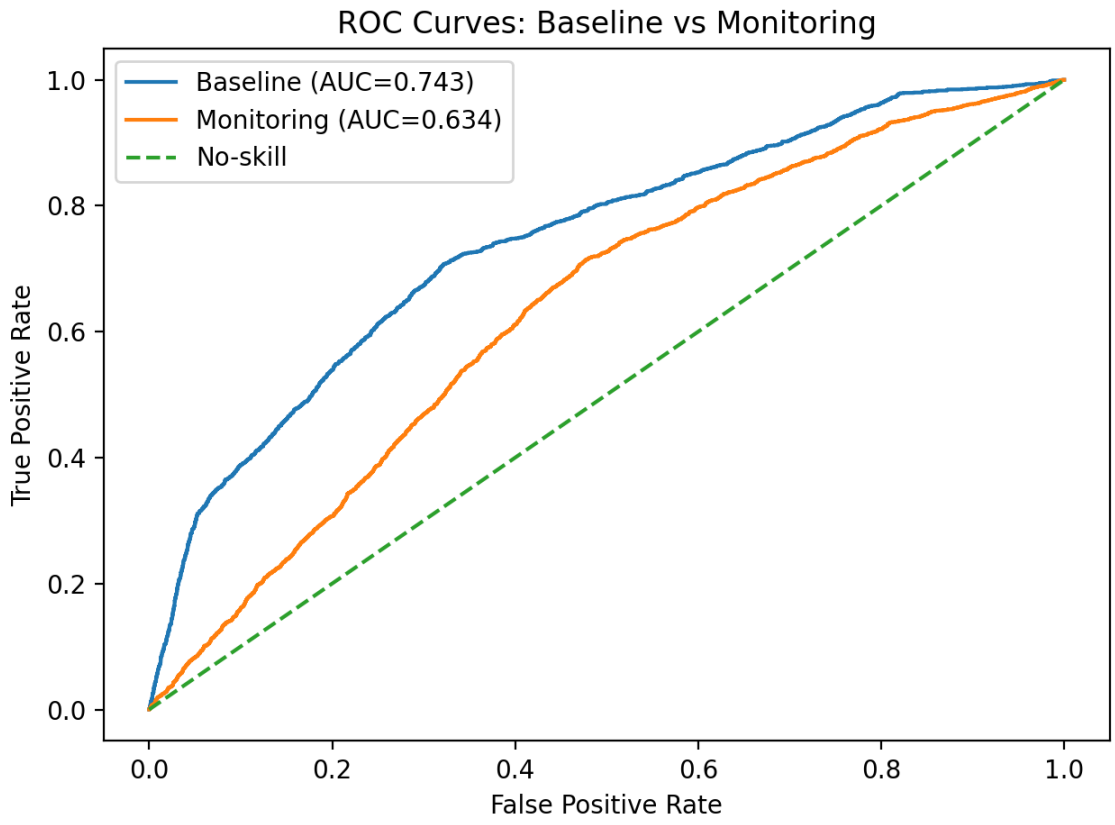

Figure 6. ROC curves - baseline vs. monitoring.

13.6 Integrated interpretation (hypothesis-consistent, not causal proof)

The indicator suite supports an operational narrative: weekly integrity signals can indicate pipeline changes; monthly stability indices can indicate structural population drift; and quarterly performance proxies confirm whether discriminatory power has degraded. This synthetic demonstration is constructed to illustrate this escalation path; it does not establish causality. In real portfolios, alternative explanations and confounding factors must be considered (policy changes, macro shocks, label-delay artifacts, and sampling effects).

14. Recommendations

Based on the proposed fr amework and the demonstration results, it is recommended that institutions operating credit scoring models adopt a multi-frequency monitoring approach that combines (1) weekly data integrity checks (missingness and outliers), (2) monthly stability monitoring (PSI for scores and CSI for key features), and (3) quarterly performance verification (AUC/Gini and KS using an outcome proxy such as 30+ DPD). Organizations should calibrate control limits using a baseline reference window, document thresholds and escalation rules as part of model governance, and maintain an auditable trail of indicator values, alerts, investigations, and corrective actions. When sustained drift is detected, the response should follow a tiered playbook: first validate data pipeline integrity, then investigate drift drivers by segment and policy changes, apply temporary compensating controls if needed, and finally initiate formal recalibration, retraining, or revalidation based on materiality and risk appetite.

15. Conclusion

This manuscript proposes a real-time monitoring control system for ML credit scoring models by integrating data integrity, stability, and performance indicators into an extended DAMA framework. The conceptual step is formalizing Predictive Quality as an operational sub-construct of Reasonableness, assessed through a bounded suite of indicators governed by SPC. By providing formal definitions, pseudocode, and traceable synthetic artifacts (tables and figures), the work offers an auditable blueprint for continuous model oversight.

Appendices

Appendix A. Synthetic aggregate artifacts

A1. Score band proportions used for PSI

A2. Financing term proportions used for CSI

Appendix B. Full pseudocode for indicator computation

```
Function PSI (expected [], actual []):
 eps = 1e-9
 total = 0
 For i in 1.k:
   e = expected[i] + eps
   a = actual[i] + eps
   total += (a - e) * ln (a / e)
 Return total
Function KS (scores [], labels []):
 Sort observations by score ascending
 Compute cumulative proportions of good and bad
 Return max absolute difference
Function BootstrapCI (metric_fn, data, B=500):
 Repeat b=1.B:
   sample with replacement
   compute metric_fn(sample)
 Return quantiles at 2.5% and 97.5% and the mean
```

Appendix C. Monitoring Runbook and Alert Escalation Matrix

Table C1. Alert escalation matrix (example)

Table 7 provides an example operational response playbook that maps indicator breaches to governance actions, escalation paths, and accountable owners.

Table 7. Example operational playbook for DQI alerts and governance actions.

Alert Level	Trigger (examples)	Required Evidence	Action / Owner
Green	Within SPC limits; PSI < 0.10; KS/Gini within baseline CI	Dashboard snapshot; last 4 periods trend; run logs	Continue monitoring; record as normal (Model Owner)
Amber	Single special-cause breach; PSI 0.10–0.25; KS decline but CI overlaps baseline	Segment drill-down; lineage check; pipeline change log	Triage + root cause; open ticket (Data Eng + Model Risk)
Red	Repeated breaches; PSI ≥ 0.25; KS/Gini below baseline with non-overlapping CI; DDR spike	Full diagnostic pack; labeling review; challenger comparison if available	Escalate to governance; compensating controls; recalibration/retrain (Model Governance)

Note: Threshold bands are illustrative and should be calibrated using the institution's baseline distribution and risk appetite. Discriminatory metrics should be evaluated with uncertainty (e.g., bootstrap confidence intervals) to reduce false alerts caused by sampling noise.

References:

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627–635.

Board of Governors of the Federal Reserve System, & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR 11-7; OCC 2011-12).

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. In 2017 IEEE International Conference on Big Data (BigData) (pp. 1123–1132). IEEE.

DAMA International. (2017). The DAMA guide to the data management body of knowledge (DAMA-DMBOK2) (2nd ed.). Technics Publications.

du Pisanie, N., Kleynhans, E. P. J., & Steenkamp, A. (2020). On testing the hypothesis of population stability for credit risk scorecards. ORiON, 36(2), 111–132.

du Pisanie, N., Kleynhans, E. P. J., & Steenkamp, A. (2023). Population stability testing procedures: A critical review. arXiv.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

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

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 45(2), 171–186.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136.

Montgomery, D. C. (2013). Introduction to statistical quality control (7th ed.). Wiley.

Office of the Comptroller of the Currency. (2021). Comptroller's handbook: Model risk management (Version 1.0, August 2021).

Office of the Comptroller of the Currency. (2025). Model risk management: Clarification for community banks (OCC Bulletin 2025-26).

Potgieter, S., van Zyl, L., Schutte, K., & Lombard, A. (2023). The population resemblance statistic: A novel chi-square measure of fit for banking. arXiv.

Siddiqi, N. (2006). Credit risk scorecards: Developing and implementing intelligent credit scoring. Wiley.

Taplin, R., & Hunt, J. (2019). A population accuracy index for measuring the stability of a model. Risks, 7(2), 53.

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. Signal Processing, 167, 107299.

ISSN: 2663-5798 || Arab Journal for Scientific Publishing

الإصدار التاسع – العدد الثامن والثمانون
تاريخ الإصدار: 2 – شباط – 2026م
www.ajsp.net

"مؤشرات جودة البيانات في الزمن الحقيقي لمراقبة تدهور القوة التنبؤية في نماذج تعلم الآلة لتقييم الجدارة الائتمانية: منظور ممتد لإطار DAMA"

إعداد الباحث:

محمد كامل عبد الرحيم اسعد

عبد اللطيف جميل المتحدة للتمويل (ALJUF)، جدة، المملكة العربية السعودية

الشهادة :محترف إدارة بيانات معتمد رئيسي (Master CDMP)

ملخص الدراسة:

تهدف هذه الدراسة إلى تطوير إطار عملي ومنهجي لمراقبة تدهور الأداء التنبؤي لنماذج تعلم الآلة المستخدمة في تقييم مخاطر الائتمان، وذلك عبر دمج مؤشرات جودة البيانات ومؤشرات استقرار التوزيعات مع مؤشرات الأداء التنبؤي ضمن منظومة مراقبة متعددة الترددات (أسبوعية/شهرية/ربع سنوية). تنطلق الدراسة من منظور إدارة البيانات وفق DAMA، وتُعرّف مفهوم "الجودة التنبؤية" بوصفه امتداداً تطبيقياً ضمن بُعد "المعقولية" (Reasonableness)، بحيث تصبح جودة البيانات مرتبطة بصورة مباشرة بقدرة النموذج على التمييز والترتيب الصحيح للمخاطر عبر الزمن.

يقترح الإطار مجموعة "مؤشرات جودة بيانات تشغيلية" لرصد الانحرافات المبكرة في مدخلات النموذج (مثل معدلات النقص، القيم الشاذة، وانحرافات التوزيعات)، ثم يربطها لاحقاً بمؤشرات الاستقرار والتحول السكاني، وصولاً إلى مؤشرات التدهور التنبؤي (مثل KS وGini وAUC). كما يتضمن الإطار آلية تفسير متكاملة تربط إشارات الانحراف المبكرة بتغيرات الاستقرار الهيكلي ثم تدهور القوة التنبؤية، بما يساعد فرق الحوكمة وإدارة المخاطر على الانتقال من المراجعات السنوية التفاعلية إلى رقابة استباقية شبه فورية، وتفعيل إجراءات المعالجة المناسبة قبل وقوع خسائر أو قرارات ائتمانية غير دقيقة.

الكلمات الإفتتاحية :تقييم الجدارة الائتمانية، مراقبة النماذج، مؤشرات جودة البيانات، اكتشاف الانجراف، الرقابة الإحصائية على العمليات، مؤشر استقرار السكان(PSI) .