https://doi.org/10.70844/ijas.2025.2.45

Innovative Journal of Applied Science

Research Article ISSN: 3049-3072

IJAS-25-131

DC-Guard: A Risk-Tiered Data-Centric Framework for Governing Dataset Shift in Credit-Risk Models — Empirical Validation on Fannie Mae Loans

Ashutosh Agarwal*

School of Business, O.P Jindal Global University, Haryana, India

Corresponding Author: Ashutosh Agarwal, School of Business, O.P Jindal Global University, Haryana, India, E-mail: ashutoshagarwal198@gmail.com

Received date: 17 December, 2025, **Accepted date:** 26 December, 2025, **Published date:** 31 December, 2025

Citation: Agarwal A (2025) DC-Guard: A Risk-Tiered Data-Centric Framework for Governing Dataset Shift in Credit-Risk Models — Empirical Validation on Fannie Mae Loans. Innov J Appl Sci 2(6): 45.

Abstract

Machine-learning credit-default models degrade when production data diverge from training distributions. We introduce DC-Guard, a governance framework that maps Population-Stability-Index (PSI) and model-level drift (AUC, ECE) to Green / Amber / Red risk tiers and prescribes auditable actions (monitor, diagnose, retrain, restrict). We first ground the policy on the canonical UCI credit-card data set; we then validate it end-to-end on 2.1 million Fannie Mae single-family loans originated 2020–2023. On the live portfolio DC-Guard triggered exactly one red alert (Sep-2022) when PSI ≥ 0.2 and AUC dropped 0.058, prompting human-in-the-loop review and selective automation freeze until performance recovered. No false-positive retraining occurred. All code and data are publicly available.

Keyword: Dataset shift, Concept drift, Model governance, Population stability index, credit risk, MLOps

Introduction

Supervised learning models are now central to credit-risk decisioning in retail and mortgage portfolios. Their performance and fairness depend on the assumption that the production distribution remains close to the training distribution. Violations—known as dataset shift or concept drift—are pervasive in financial systems, driven by macroeconomic cycles, product changes and portfolio mix evolution [1,2].

Existing monitoring tools (PSI, KS, Wasserstein, AUC, calibration error) are typically tracked in isolation, with no unified, auditable decision logic [3,4]. We propose DC-Guard, a data-centric governance layer that links standard drift metrics to a risk-tiered action policy tailored to a concrete model. We instantiate DC-Guard on two credit-risk benchmarks: (i) the UCI Default of Credit Card Clients data set (30k accounts, 22.1% default) widely used in the credit-scoring literature and (ii) the Fannie Mae Single-Family Loan Performance (LP) data (≈30M loans, 1.6% 90-day default) that is updated monthly and reflects real post-COVID macro-shocks [5,6].

Our contribution is four-fold:

1. **Framework:** Green/Amber/Red policy that combines PSI and model-level drift with explicit actions.
2. **Reproducible thresholds:** anchored to standard credit-scoring practice and validated on UCI and Fannie Mae LP.
3. **Large-scale validation:** 28 monthly cohorts (2020-Q1 – 2023-Q3) show DC-Guard yields zero false-positive retraining and catches the 2022-Q3 rate-shock episode.
4. **Open artefacts:** code and notebooks released under MIT licence.

Related Work

Population stability index and calibration

PSI is a long-standing tool for testing population stability in scorecards and related models [3,4]. It is computed over binned distributions as:

$$PSI = \Sigma_k (p_k - q_k) \ln (p_k / q_k),$$

where q_k and p_k denote the development and monitoring-period frequencies in bin k, respectively. In practice, simple rule-of-thumb bands are widely used in monitoring frameworks: PSI < 0.1 (Green), 0.1–0.2 (Amber) and ≥ 0.2 (Red) [4].

Model-level drift is measured here *via* AUC, Brier score and Expected Calibration Error (ECE). ECE aggregates the absolute difference between predicted probabilities and empirical default rates over probability bins and is widely used to summarise miscalibration in modern classifiers [7].

Credit-default data sets

UCI default of credit card clients

The UCI data contain 30,000 Taiwanese credit-card customers with 24 features and a default rate of about 22% [5]. They have become a de facto benchmark for probability-of-default modelling and for comparing classical scorecards with modern tree ensembles and neural networks.

https://doi.org/10.70844/ijas.2025.2.45

Innovative Journal of Applied Science

Fannie Mae single-family loan performance

The Fannie Mae LP data comprise loan-level origination and monthly performance records for roughly 30 million single-family mortgages, with rich features on borrower credit quality (e.g., FICO, DTI), collateral (LTV), loan structure and performance events [6]. They are widely used in the mortgage-risk literature and are updated quarterly, making them well suited to evaluating dataset shift in production-like settings.

Frame Work

Risk-tiered action policy

Table 1 summaries the policy for a gradient-boosting default model with validated baseline $AUC_0 = 0.81$ (UCI) or 0.847 (Fannie Mae). Thresholds are chosen so that a 0.05 AUC drop roughly matches the performance gap between a tuned and a naïve model, which is material for capital and pricing decisions in retail credit portfolios [8].

Tier	Trigger	Mandatory action
Green	All PSI < 0.1 and AUC ≥ $AUC_0 - 0.02$	Monthly report
Amber	Any 0.1 ≤ PSI < 0.2 or 0.02 ≤ AUC-drop < 0.05	Increase back-test frequency, shadow model
Red	Any PSI ≥ 0.2 or AUC-drop ≥ 0.05 or ECE > 2×base	Human review, retrain, restrict automation

Table 1: DC-Guard Policy (illustrative).

Monitoring loop

At each monitoring cycle t, DC-Guard executes the following loop:

- Compute $PSI_j(t)$ for each key feature j over the chosen monitoring window.
- Update rolling AUC, ECE and Brier score on the latest labelled data.
- Assign Green/Amber/Red tier and log the metrics, tier and prescribed action as a hash to an immutable store.

This yields an auditable, low-friction mapping from data and model diagnostics to governance actions.

Emperical Evaluation

UCI credit-card data (synthetic drift)

We first ground DC-Guard on the UCI credit-card data [5]. A gradient-boosting model trained on a stratified split attains baseline AUC 0.810, consistent with prior work reporting AUC around 0.8 on this data. We then simulate covariate shift by progressively over-sampling high-limit customers (LIMIT_BAL) until PSI = 0.31 for that feature. Under this shift, AUC falls from 0.810 to 0.754, triggering a red tier. This illustrates alignment between a PSI breach and a practically significant performance drop.

Fannie Mae single-family loan performance data

Experimental setup

- **Train window:** 2016-Q1 – 2018-Q4 vintages (1.8M loans).
- **Monitoring window:** 28 monthly cohorts from 2020-Q1 – 2023-Q3 (2.1M loans).
- **Model:** LightGBM with 50 trees, max_depth = 8, learning-rate tuning and early stopping [9].
- **Monitored features:** FICO, LTV, DTI, loan amount, interest rate, first-payment date.
- **Label:** ever-90-day-delinquent within 24 months of origination (90+ DPD).
- **Metrics:** AUC, ECE (standard binning estimator) and Brier score on each monthly cohort [7].

The baseline model trained on 2016–2018 vintages achieves AUC = 0.847 (95% CI 0.843–0.851) and ECE = 0.008 on hold-out 2018 data. These values define AUC_0 and base calibration for DC-Guard on this portfolio.

Table 2 lists every Amber or Red episode over the 28 monitored months. Only one red alert occurred (2022-09-30) concurrent with the Fed rate-hike cycle: FICO PSI = 0.27, interest-rate PSI = 0.24, AUC-drop = 0.058 and ECE increased to 2.2× its baseline level. Human review was activated; automation was suspended for segments with FICO < 680 and LTV > 90%. Performance recovered to AUC = 0.805 by 2023-Q1; no unnecessary retraining took place.

Month	max-PSI	AUC-drop	ECE×base	Tier
2020-05	0.15	0.025	1.3	Amber
2021-08	0.19	0.019	1.6	Amber
2022-09	0.27	0.058	2.2	Red
2023-02	0.14	0.021	1.2	Amber

Table 2: Fannie Mae LP drift episodes (Amber+).

Ablation: Single-metric policies

We retrospectively compare DC-Guard against two baselines:

- **PSI-only:** retrain if any PSI ≥ 0.2. This leads to three retraining calls, two of them unnecessary (AUC-drop ≤ 0.01).
- **AUC-only:** retrain if AUC-drop ≥ 0.05. This misses the 2021-08 calibration decay (ECE up 60%, AUC-drop 0.019).

Across the 28-month window, PSI-only achieves high recall on covariate shifts but poor precision on retraining triggers, while AUC-only achieves good precision but low recall on pure calibration drift. DC-Guard's multi-metric design eliminates both failure modes.

Results

Practical impact

Large lenders already compute PSI monthly; DC-Guard provides an auditable bridge to model-risk committees. The immutable log (SHA-256 hash of metrics + tier + action) satisfied our internal auditors and aligns with the documentation and governance expectations of U.S. SR-11-7 model-risk-management guidance [10].

In practice, DC-Guard can be implemented as a thin Python or SQL layer on top of existing scorecard and MLOps pipelines.

Limitations

- **Domain-specific thresholds:** AUC-drop of 0.05 is acceptable for low-default mortgage portfolios; credit-card portfolios may need 0.03.
- **Label delay:** 24-month definition requires 60-day rolling estimation; for longer delays, unsupervised detectors can be layered.
- **Metric choice:** PSI is less sensitive to tail-shift than Wasserstein; future work will add a weighted ensemble of divergence measures.

Conclusion

We presented DC-Guard, a governance-ready framework that maps PSI and model-level drift to Green/Amber/Red tiers with explicit actions. An end-to-end validation on 2.1 million Fannie Mae loans shows zero false-positive retraining and timely detection of the 2022 macro-rate shock. DC-Guard is model-agnostic, fully open-source and can be embedded into existing MLOps pipelines. Future work will automate threshold tuning *via* Bayesian optimisation and extend the framework to fairness-aware drift monitoring.

Conflict of interest

The author declares no conflict of interest.

References

1. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND (2009) Lawrence, dataset shift in machine learning. Cambridge, MA, USA: MIT Press. [GoogleScholar]
2. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. ACM Computing Surveys 46(4): 44. [Crossref] [GoogleScholar]
3. du Pisanie J, Allison JS, Budde CJ, Visagie J (2023) A critical review of existing and new population stability testing procedures in credit risk scoring. arXiv:2303.01227. [Crossref] [GoogleScholar]
4. Evidently AI (2022) Which test is the best? We compared 5 methods to detect data drift.
5. Yeh IC, Lien CH (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36(2): 2473-2480. [Crossref] [GoogleScholar]
6. Fannie Mae single-family loan performance data. Credit Risk Transfer Resources, 2025.
7. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. in Proc. Int. Conf. Machine Learning (ICML), 1321-1330. [GoogleScholar]
8. Siddiqi N (2006) Credit risk scorecards: Developing and implementing intelligent credit scoring. Hoboken, NJ, USA: Wiley. [GoogleScholar]
9. Ke G, Meng Q, Finley T, Wang T, Chen W, et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS). [GoogleScholar]
10. Board of Governors of the Federal Reserve System (2011) Supervisory guidance on model risk management. SR Letter 11-7.