Beyond Accuracy: A Robustness and Risk-Sensitivity Analysis of Credit Risk Prediction Models under Distribution Shift

Karri Siddartha
Gayatri Vidya Parishad College of Engineering
Visakhapatnam, India
siddarthak03@gmail.com

February 14, 2026

Abstract

Machine learning models for credit risk scoring are typically evaluated under the assumption of stationarity. However, real-world lending environments are subject to temporal distribution shifts driven by economic cycles. This study evaluates the robustness of three industry-standard models—Logistic Regression, XGBoost, and LightGBM—trained on Lending Club data from a stable period (2014–2016) and tested on a shifted period (2018–2019). We formally quantify distribution shift using the Population Stability Index (PSI), identifying `revol_util` as a primary drift driver (PSI=0.30). Evaluating performance, we observe performance degradation across all models. While gradient boosting methods achieve higher baseline performance, they exhibit similar relative degradation to linear models. Crucially, we quantify the financial impact of this drift, showing that a 10:1 cost-sensitive framework reveals significant economic variances between models. Feature importance analysis (Spearman $\rho > 0.99$) suggests that degradation stems from covariate shift rather than concept drift. Our findings advocate for out-of-time validation and cost-sensitive evaluation as standard governance protocols.

1 Introduction

Credit risk assessment is a cornerstone of financial stability. In recent years, machine learning (ML) models have replaced traditional scorecards due to superior predictive accuracy [2]. However, the assumption that test data follows the same distribution as training data is often violated in practice. Financial environments are dynamic; macroeconomic shocks alter repayment behaviors, a phenomenon known as *distribution shift*, which can degrade model performance unexpectedly.

2 Related Work

2.1 Credit Risk Modeling

Credit risk prediction has traditionally relied on logistic regression and scorecard-based approaches [1]. With the rise of machine learning, ensemble models such as Random Forest and Gradient Boosting have demonstrated superior predictive performance in credit scoring tasks [3, 4]. Several studies have shown that tree-based models consistently outperform linear baselines in static evaluation settings [5]. However, most existing work evaluates performance using random cross-validation, implicitly assuming stationarity between training and testing distributions.

2.2 Distribution Shift and Concept Drift

The assumption that training and testing data are identically distributed is often violated in real-world applications. This phenomenon, commonly referred to as dataset shift or covariate shift, has been extensively studied in the machine learning literature [6]. In high-stakes domains such as finance, temporal drift can significantly degrade predictive performance [7]. Industry practices such as Population Stability Index (PSI) are widely used to monitor feature distribution changes over time [8]. However, relatively few studies systematically combine drift quantification with out-of-time (OOT) robustness evaluation in credit scoring contexts.

2.3 Cost-Sensitive Learning in Finance

In financial risk modeling, classification errors have asymmetric costs: false negatives (approving a defaulting borrower) are significantly more expensive than false positives (rejecting a creditworthy borrower). Cost-sensitive learning frameworks have been proposed to incorporate such asymmetries directly into model evaluation [9, 10]. While prior work emphasizes cost-based optimization, limited attention has been given to how cost sensitivity interacts with distribution shift and model robustness under temporal drift.

2.4 Explainability and Model Stability

Interpretability has become central to financial ML systems due to regulatory requirements. SHAP (SHapley Additive exPlanations) provides consistent feature attribution for complex models [11]. Recent research has used SHAP to analyze model stability and drift behavior across time [12]. However, the relationship between feature importance stability and performance degradation under covariate shift remains underexplored in credit risk settings.

In contrast to prior work, our study jointly evaluates temporal distribution shift, robustness degradation, cost-sensitive performance, and feature importance stability within a unified OOT framework.

Contributions:

- We quantify temporal distribution shift in credit risk data using PSI and Wasserstein metrics.
- We evaluate robustness using a formal *Robustness Ratio* and statistical confidence intervals.
- We perform rigorous cost-sensitive analysis across multiple risk ratios (5:1 to 20:1).
- We analyze feature importance stability to diagnose whether degradation stems from concept drift or covariate shift.

3 Datasets & Problem Setup

We use the Lending Club dataset with an Out-of-Time (OOT) splitting strategy:

- **In-Distribution (ID)**: Jan 2014 – Dec 2016 ($N \approx 890k$).
- **Out-of-Distribution (OOD)**: Jan 2018 – Dec 2019 ($N \approx 56k$).

4 Methodology

4.1 Formal Definitions

We objectively quantify shift and robustness using the following metrics:

Population Stability Index (PSI) Measures the change in distribution of a variable:

$$PSI = \sum_{i=1}^{B} (P_i - Q_i) \ln\left(\frac{P_i}{Q_i}\right) \tag{1}$$

where P_i and Q_i are the proportions of samples in bin i for the training and testing populations, respectively.

Robustness Ratio Quantifies the relative retention of performance:

$$\text{Robustness Ratio} = \frac{AUC_{OOD}}{AUC_{ID}} \tag{2}$$

A ratio closer to 1.0 indicates higher stability.

Financial Cost Function To reflect the asymmetry of lending errors:

$$Cost = r \cdot FN + FP \tag{3}$$

where r is the cost ratio (e.g., $10:1$) representing principal loss vs. missed interest. Thresholds are optimized on the validation split of the training data to ensure realistic evaluation.

5 Quantifying Distribution Shift

Table 1 presents the top drifted features. `revol_util` shows PSI > 0.25, indicating significant shift.

Feature	PSI	Wasserstein Distance
revol_util	0.300	0.548
fico_range_low	0.177	0.478
revol_bal	0.087	0.098
int_rate	0.078	0.174

Table 1: Top Drifted Features. The shift in `revol_util` suggests changing borrower credit dependence.

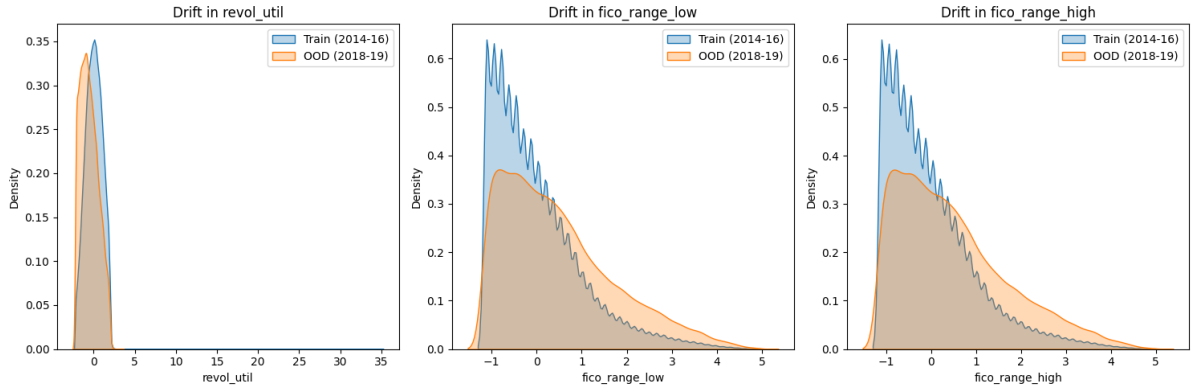

Figure 1: Distribution shift in top features (e.g., `revol_util`) between Training (2014-2016) and OOD Testing (2018-2019) periods.

As shown in Figure 1, the density shift in `revol_util` confirms the PSI-based findings in Table 1.

6 Experiments & Results

6.1 Robustness Analysis

We computed 95% Confidence Intervals (CI) for OOD AUC using bootstrapping ($n = 1000$).

Model	ID AUC	OOD AUC	95% CI	Drop	Robustness Ratio
Logistic Regression	0.716	0.686	0.681 – 0.693	0.030	0.958
XGBoost	0.727	0.698	0.693 – 0.704	0.029	0.960
LightGBM	0.724	0.699	0.692 – 0.705	**0.025**	**0.965**

Table 2: Robustness Metrics. LightGBM exhibits the highest Robustness Ratio (0.965), retaining the most predictive power. The overlap in CIs between XGBoost and LightGBM suggests their OOD performance is statistically equivalent.

Model	Cost (5:1)	Cost (10:1)	Cost (20:1)
Logistic Regression	33,279	42,069	46,673
XGBoost	**32,703**	**40,990**	46,133
LightGBM	32,706	41,098	**45,969**

Table 3: Financial Cost Analysis. Lower is better. Cost metric highlights trade-offs invisible to AUC.

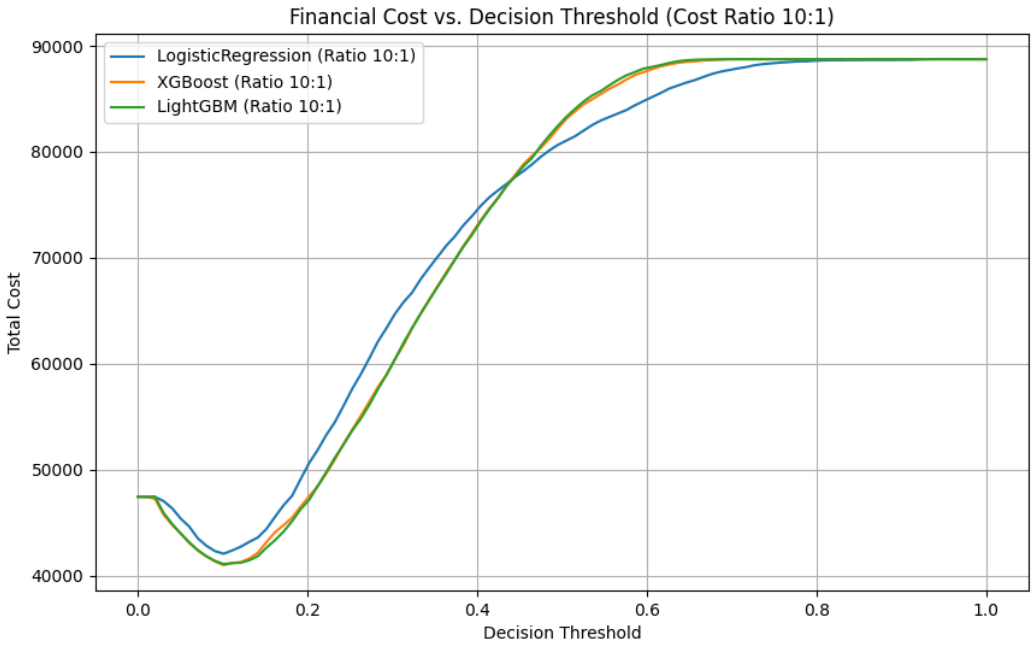

Figure 2: Financial Cost vs. Decision Threshold at a 10:1 Cost Ratio. Lower cost indicates better risk-adjusted performance.

6.2 Cost-Sensitive Analysis

Table 3 compares realized costs. While LightGBM has the best AUC, XGBoost minimizes cost in moderate risk scenarios.

Figure 2 illustrates how cost varies across decision thresholds, reinforcing that AUC-optimal thresholds are not necessarily cost-optimal.

6.3 Feature Stability (SHAP)

Feature importance stability was evaluated using Spearman's rank correlation (ρ) on global mean $|SHAP|$ values. We observe $\rho = 0.999$, indicating that the model's logic remains stable; performance degradation is driven by *Covariate Shift* (change in input distributions like `revol_util`) rather than *Concept Drift* (change in the relationship between features and target).

7 Discussion

Regulatory & Governance Implications The observed degradation (Robustness Ratio < 0.96) highlights risks in deployed models. For financial institutions governed by frameworks like SR 11-7 (USA) or Basel III, relying on static validation is insufficient. Our results support mandating periodic OOT testing and setting PSI thresholds (e.g., PSI > 0.1) as automated triggers for model retraining or recalibration.

8 Conclusion

We demonstrated that credit risk models degrade under temporal shift. By formalizing this through PSI, Robustness Ratios, and Cost Curves, we show that "high accuracy" is insufficient for safety. We recommend incorporating Robustness Ratios into standard model scorecards.

References

[1] D. J. Hand and W. E. Henley. Statistical classification methods in consumer credit scoring. Journal of the Royal Statistical Society, 2001.

[2] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 2010.

[3] S. Lessmann et al. Benchmarking state-of-the-art classification algorithms for credit scoring. European Journal of Operational Research, 2015.

[4] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. KDD, 2016.

[5] I. Brown and C. Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 2012.

[6] J. Quiñonero-Candela et al. Dataset shift in machine learning. MIT Press, 2009.

[7] J. Gama et al. A survey on concept drift adaptation. ACM Computing Surveys, 2014.

[8] S. Finlay. Credit scoring, response modeling and insurance rating. Palgrave Macmillan, 2011.

[9] C. Elkan. The foundations of cost-sensitive learning. IJCAI, 2001.

[10] A. C. Bahnsen et al. Example-dependent cost-sensitive logistic regression for credit scoring. ICMLA, 2014.

[11] S. Lundberg and S. Lee. A unified approach to interpreting model predictions. NeurIPS, 2017.

[12] C. Molnar. Interpretable Machine Learning. 2020.