Model Drift Detection in Credit Risk Prediction

An Empirical Study of Data Drift, Concept Drift, and Model Degradation

Asini Susanya Karunarathna

https://github.com/asinisusanya/ml-drift-detection

Abstract

Machine learning models deployed in real world decision making systems are vulnerable to performance degradation due to evolving data distributions. In domains such as credit risk assessment, such failures may remain undetected until significant financial consequences occur. This study investigates the impact of model drift on a credit risk prediction model by simulating data drift and concept drift scenarios. A baseline LightGBM classifier is trained on reference data and evaluated under increasing drift severity. Statistical drift detection techniques, including the Kolmogorov–Smirnov test and Population Stability Index, are applied to identify distributional changes. Results demonstrate that data drift alone does not necessarily degrade model performance, whereas concept drift leads to a severe collapse in predictive accuracy and ROC-AUC. Explainable AI using SHAP is employed to identify drift-sensitive features. The study focuses on diagnosis and analysis rather than mitigation, emphasizing the importance of drift monitoring in reliable machine learning deployment.

---

1 Introduction

Machine learning models are increasingly used in operational decision systems such as credit scoring, fraud detection, and loan approval. These models are typically trained on historical data under the assumption that future data follows a similar distribution. However, real-world environments are dynamic, with changes in economic conditions, customer behavior, and institutional policies.

Such changes lead to *model drift*, where the statistical properties of incoming data differ from those observed during training. If left unmonitored, drift may silently degrade model performance, resulting in unreliable or biased decisions.

This study aims to empirically analyze how different forms of drift affect a credit risk prediction model. Rather than proposing corrective solutions, the focus is on detecting, analyzing, and explaining drift-induced failures. In practice, many deployed machine learning systems rely on static validation metrics and periodic retraining schedules. This work highlights why such approaches are insufficient without explicit drift monitoring, particularly in high-risk domains such as credit scoring.

1.1 Objectives

- Simulate realistic data drift and concept drift scenarios.
- Detect drift using statistical hypothesis testing.
- Evaluate model robustness under increasing drift severity.
- Explain drift-induced failure using model explainability techniques.

1.2 Scope

This work focuses exclusively on drift detection and performance analysis. No mitigation, retraining, or adaptation strategies are implemented. Potential solutions are discussed only as future work.

2 Dataset Description

The experiments are conducted using the German Credit Risk dataset, which contains demographic, financial, and loan-related attributes. Each record represents a loan applicant. The dataset used in this study is publicly available and was obtained from the German Credit Risk dataset:

`German Credit Risk dataset`

2.1 Features

- **Numeric Features:** Credit amount, loan duration, age.
- **Categorical Features:** Saving accounts, checking account status, housing, purpose of loan.

2.2 Target Variable

Since an explicit default label was unavailable, a synthetic binary target variable was constructed using domain-inspired rules. Applicants with a combination of high credit amount, long loan duration, and weak savings status were classified as high-risk. The resulting

class distribution is moderately imbalanced, reflecting realistic credit risk settings.

3 Methodology

This study follows a structured experimental pipeline consisting of data preparation, baseline model training, drift simulation, drift detection, and performance evaluation.

3.1 Experimental Pipeline

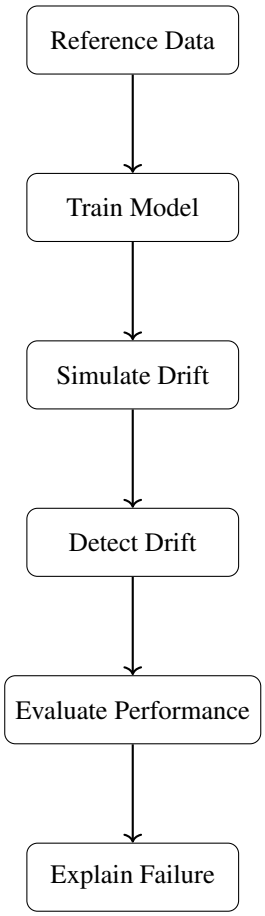

Figure 1: Vertical experimental pipeline for model drift analysis

Figure 1 illustrates the sequential workflow adopted in this study, from reference data preparation to drift explanation.

3.2 Baseline Model

A LightGBM classifier was trained on the reference dataset. Categorical features were encoded using one-hot encoding, while numeric features were passed directly to the model. Baseline accuracy and ROC-AUC were recorded to establish a performance benchmark.

3.3 Drift Simulation

To emulate real-world deployment conditions, synthetic drift was introduced at three levels:

- **Mild Drift:** Small shifts in numeric feature distributions.
- **Moderate Drift:** Larger numeric shifts and categorical redistribution.
- **Severe Drift:** Data drift combined with concept drift through altered target logic.

Concept drift in the severe scenario represents a policy or economic regime change affecting the relationship between features and default risk.

3.4 Drift Detection

Statistical drift detection techniques were used to compare reference data with drifted datasets. The Kolmogorov–Smirnov (KS) test was applied to numeric features, while the Population Stability Index (PSI) was used for categorical features.

3.5 Performance Evaluation

Model performance was evaluated under each drift scenario using accuracy and ROC-AUC. This step connects statistical drift to its practical impact on predictive reliability.

4 Results

4.1 Model Performance under Drift

Table 1: Model performance across drift scenarios

Drift Scenario	Accuracy	ROC-AUC
Reference	1.000	1.000
Mild Drift	0.968	0.991
Moderate Drift	0.980	0.995
Severe (Concept Drift)	0.235	0.209

These results indicate that predictive performance remains stable under data drift but deteriorates sharply when concept drift is introduced.

4.2 Statistical Drift Detection Summary

Table 2: Summary of detected drift across features

Drift Scenario	KS-Test Detections	PSI Detections
Mild Drift	2	0
Moderate Drift	3	0
Severe Drift	3	0

The absence of categorical drift detections suggests that numeric features are the primary drivers of observed distributional change.

4.3 Visualization of Drift and Performance

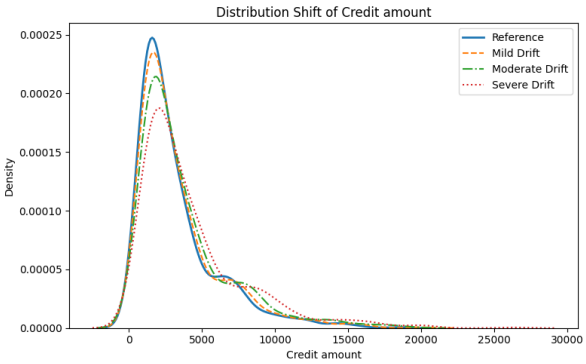

Figure 2: Distribution shift of credit amount under increasing drift severity

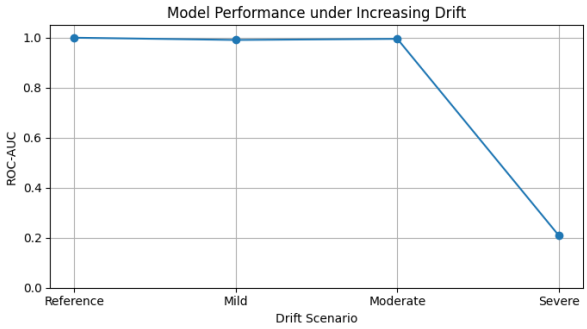

Figure 3: ROC-AUC under increasing drift severity

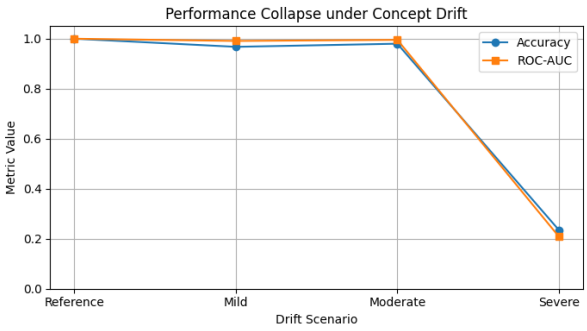

Figure 4: Accuracy and ROC-AUC collapse under severe concept drift

4.4 Explainability Analysis

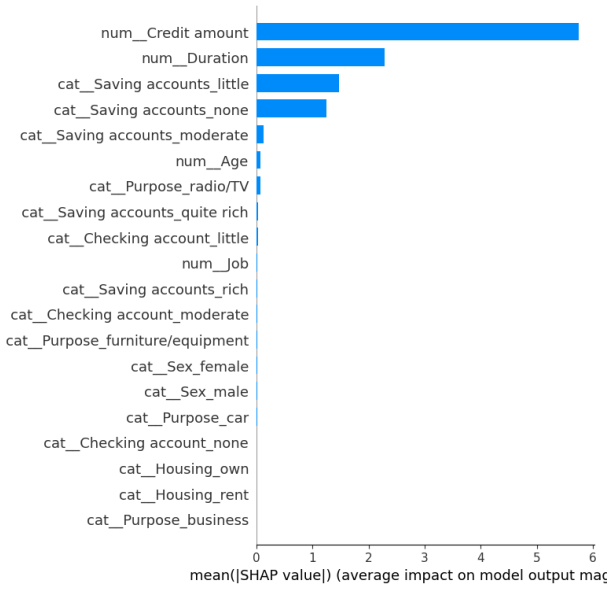

Figure 5: Global SHAP feature importance for the baseline model

SHAP analysis indicates that credit amount and loan duration are the most influential features. These features were also the most affected by drift, explaining the observed performance collapse.

5 Discussion

The results demonstrate that statistically detectable data drift does not necessarily imply immediate model failure. Despite significant distributional shifts, the model remained robust under mild and moderate drift scenarios. In contrast, concept drift resulted in a severe degradation of predictive performance. This highlights that changes in the underlying relationship between features and the target variable pose a greater risk than data drift alone.
Explainability analysis further confirms that drift-sensitive features align with the most influential predictors, emphasizing the importance of feature-level monitoring.

6 Limitations and Future Work

This study relies on synthetic drift scenarios, which may not fully capture real-world complexity. No adaptive retraining or mitigation strategies were implemented.
Future work may explore online drift detection, automated retraining pipelines, human-in-the-loop validation, and governance-aware monitoring frameworks.

7 Conclusion

This work provides an empirical analysis of model drift in a credit risk prediction system. By integrating statistical drift detection, performance evaluation, and explainable AI, the study demonstrates that data drift alone does not guarantee failure, whereas concept drift poses a critical threat. The findings reinforce the necessity of continuous drift monitoring for reliable machine learning deployment.