

Article

A Governance-Oriented Evaluation of Explainable AI in Business Analytics

Seyed Mohammad Vahedi ¹, Adebayo Adewumi Abayomi-Alli ², Olusola Oluwakemi Abayomi-Alli ² and Pavel Stefanovič ^{2,*}

¹ Department of Information Technology, Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania; seyedmohammad.vahedi@stud.vilniustech.lt

² Department of Information Systems, Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania; adebayo-adewumi.abayomi-alli@vilniustech.lt (A.A.A.-A.); olusola-oluwakemi.abayomi-alli@vilniustech.lt (O.O.A.-A.)

* Correspondence: pavel.stefanovic@vilniustech.lt

Abstract

Artificial intelligence (AI) is increasingly embedded in analytics-driven decision systems operating in high-stakes environments, where predictive performance alone may not ensure robustness, traceability, or readiness for governance. While explainable artificial intelligence (XAI) is commonly treated as an interpretability tool, its role as a measurable diagnostic component remains underexplored. This study evaluates explainability beyond predictive accuracy using a controlled dual-pipeline design with identical data, model, and validation settings. Using the UCI Default of Credit Card Clients dataset, an XGBoost model achieved strong predictive performance ($AUC \approx 0.78$; $AP \approx 0.56$) while maintaining high decision stability under retraining (agreement $\approx 99.49\%$). Global explanations were highly reproducible across runs (Spearman $\rho \approx 0.99$), and entropy-based local explanation analysis revealed substantially higher attribution dispersion in false-negative cases (odds ratio ≈ 6.58), linking explanatory uncertainty to misclassification-prone regions. The findings demonstrate that explainability diagnostics can reveal measurable patterns of stability, reproducibility, and uncertainty that are not captured by predictive metrics alone. The study advances a governance-oriented perspective in which explainability serves as a reproducible diagnostic layer that supports monitoring, validation, and audit-oriented assessment in analytics-driven decision systems.

Keywords: governance-oriented XAI; decision stability; explainability; auditability; uncertainty diagnostics; SHAP; business analytics



Academic Editor: Paweł Weichbroth

Received: 12 May 2026

Revised: 12 June 2026

Accepted: 17 June 2026

Published: 27 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Artificial intelligence (AI) has become a core component of business analytics, supporting high-stakes decisions in domains such as credit risk evaluation, auditing, and supply chain management. In these settings, predictive accuracy alone is no longer sufficient. Decisions must also be transparent, reproducible, and accountable to satisfy operational and regulatory expectations [1–3]. This requirement has increased the relevance of explainable artificial intelligence (XAI), which aims to make complex model behavior interpretable and usable within organizational decision processes [4,5]. Prior research has extensively examined explanation techniques, fidelity metrics, and benchmarking performance [6–8]. However, explainability is still often treated as a descriptive or trust-enhancing feature rather than as a measurable diagnostic component of predictive decision systems. When

explainability is treated solely as a descriptive or trust-enhancing feature, organizations may overlook instability in model decisions, fail to identify uncertainty-prone prediction regions, and lack reproducible evidence required for validation, auditing, and governance processes [9–11]. As AI systems become increasingly embedded in high-stakes decision environments, these limitations can obscure operational risks and reduce organizational visibility into model behavior, even when conventional predictive performance metrics appear stable [2,11]. This leaves an important gap: the extent to which explainability can support stability assessment, reproducibility, uncertainty detection, and audit-oriented monitoring remains insufficiently operationalized.

This gap is particularly relevant because explanation methods can be sensitive to retraining conditions, perturbations, and data characteristics [9–12]. In operational settings, small probability shifts near decision thresholds may produce materially different outcomes, even when aggregate predictive performance appears stable. Such instability can become governance-relevant when model decisions and attribution patterns must be documented, reviewed, and defended within accountable AI workflows [13,14].

To address this gap, this study evaluates explainability beyond predictive accuracy under controlled retraining conditions. A dual-pipeline design is employed in which the predictive model, dataset, validation protocol, and decision threshold remain fixed, while explainability is introduced as an additional diagnostic layer. The explainability-augmented pipeline incorporates SHAP-based explanations, multi-seed retraining diagnostics, entropy-based uncertainty analysis, and structured reproducibility artifacts. Using repeated retraining and cross-validated evaluation, the study examines four governance-relevant dimensions: decision stability, reproducibility of explanations, uncertainty-sensitive misclassification behavior, and traceability-oriented diagnostic evidence. By treating explainability as a diagnostic component rather than a predictive enhancement mechanism, the study makes three contributions. First, it introduces a measurable framework for evaluating explainability that goes beyond predictive accuracy alone, assessing stability, reproducibility, and uncertainty-sensitive behavior. Second, it provides empirical evidence linking variability in retraining and entropy-based attribution dispersion to boundary-sensitive decision risks. Third, it advances a governance-oriented evaluation perspective in which explainability serves as a reproducible diagnostic layer that supports monitoring, validation, and audit-oriented assessment in analytics-driven decision systems.

2. Analytical Background and Research Gap

2.1. Explainability Beyond Predictive Performance

XAI has evolved from a technical concern into a central component of predictive systems operating in high-stakes environments. Early research primarily emphasized the trade-off between interpretability and predictive performance, often favoring inherently interpretable models for sensitive applications [3]. Later studies clarified related concepts such as interpretability, explainability, transparency, and trustworthiness within organizational AI settings [1,4]. More recent work positions explainability beyond transparency alone, linking it to monitoring, validation, operational decision support, and the adoption of analytical outputs within organizational decision processes [2,5,15–18].

At the same time, the growing adoption of complex machine learning models has intensified concerns regarding the reliability of explanations themselves. Existing studies show that explanation methods are sensitive to retraining conditions, model initialization, feature dependencies, and distributional perturbations [9,12,19]. Similar findings have been reported for SHAP and LIME, where attribution patterns may vary despite relatively stable predictive performance [10,20,21]. Consequently, explainability cannot be treated

solely as a static interpretability layer, as instability near decision thresholds can lead to materially different operational outcomes.

Recent reviews further indicate that XAI evaluation remains heavily centered on fidelity, faithfulness, and interpretability metrics [6–8]. Although these metrics provide important insight into explanation quality, they only partially capture how explainability functions within operational decision infrastructures. In practice, predictive systems increasingly require monitoring, reproducibility assessment, and audit readiness. As the authors of [11,22] argue, responsible AI governance depends on mechanisms that support traceability and structured oversight across the AI lifecycle. Under such conditions, explanation instability becomes not only a technical limitation but also a governance concern.

This perspective motivates a shift from viewing explainability primarily as a trust-enhancing feature toward treating it as a measurable diagnostic layer within predictive systems. Recent studies increasingly emphasize that explainability should support monitoring, traceability, and robustness assessment rather than function solely as a post hoc interpretability mechanism [11,23]. From this viewpoint, explainability diagnostics become valuable not only for interpreting model behavior but also for exposing instability, uncertainty, and reproducibility patterns that remain invisible in prediction-only pipelines [9,10].

2.2. Governance-Oriented Diagnostic Perspective

As explainability becomes increasingly integrated into decision infrastructures, the evaluation of XAI systems extends beyond interpretability alone. In high-stakes environments, predictive systems are expected to support reproducibility, traceability, and continuous monitoring across deployment and retraining cycles [11,24]. Under such conditions, explanation outputs increasingly serve as diagnostic signals that support validation, review, and audit processes within operational decision systems [25,26].

Recent governance-oriented research further suggests that responsible AI depends not only on transparency mechanisms but also on the ability to evaluate model behavior under changing operational conditions [14,22,27]. This shift has important implications for XAI evaluation because predictive stability and explanatory stability do not necessarily evolve together. A model may preserve relatively stable aggregate performance while simultaneously producing inconsistent local explanations or unstable behavior near classification thresholds. Earlier work by Alvarez-Melis and Jaakkola demonstrated that interpretability methods can exhibit substantial instability under perturbations. At the same time, later studies have reported similar sensitivity to retraining and changes in feature distributions [10,21]. Consequently, explanation robustness increasingly emerges as a system-level reliability concern rather than merely a technical property of interpretability algorithms.

Within this context, governance-oriented diagnostics can be interpreted as measurable indicators of monitoring visibility. Decision stability reflects whether model outputs remain consistent across retraining conditions, whereas structural stability captures whether dominant explanatory drivers remain reproducible across repeated runs [10,28]. At the local level, dispersion in attribution patterns may reveal uncertainty-sensitive regions where predictions become less structurally concentrated. Recent studies further emphasize that explanation quality should be evaluated not only by interpretability but also by robustness, reproducibility, and sensitivity to perturbations [6,29]. From this perspective, explainability becomes part of a broader monitoring infrastructure that can expose instability, limitations in reproducibility, and boundary-sensitive behavior in predictive pipelines [11,23].

In practice, these diagnostics support validation, compliance, auditing, and operational oversight processes by providing reproducible evidence regarding system behavior under repeated evaluation conditions. Similarly, refs. [13,30] emphasize the importance of traceable documentation and auditable reporting structures in AI-enabled decision systems.

Accordingly, the value of explainability emerges not only from making predictions understandable but also from enabling measurable inspection of operational model behavior.

To operationalize governance-oriented explainability within a measurable evaluation framework, the study defines the diagnostic constructs summarized in Table 1.

Table 1. Operationalization of governance-oriented diagnostic constructs in explainable analytics.

Construct	Operational Definition	Operationalization	Metric
Decision Stability	Consistency of model decisions across repeated retraining conditions	Compare decision outputs generated across multiple retrained models under fixed validation settings	Agreement rate, flip rate, probability variance
Structural Stability of Explanations	Reproducibility of dominant explanatory patterns across retraining runs	Compare global SHAP importance rankings and dominant feature sets across repeated executions	Spearman rank correlation, Top-K Jaccard similarity
Explanation Dispersion	Distribution spread of feature attributions within local explanations	Compute entropy over normalized absolute SHAP attribution values for individual predictions	Shannon entropy of SHAP attributions
Traceability and Governance Readiness	Ability of the analytics pipeline to generate reproducible and reviewable diagnostic artifacts	Store and version predictions, explanation outputs, retraining configurations, and validation artifacts across experimental runs	Versioned diagnostic and reproducibility artifacts

2.3. Research Gap and Research Questions

Existing explainable AI research generally follows three directions: improving interpretability techniques and evaluation metrics [6–8]; investigating the robustness of explanations under perturbations and retraining conditions [9,10,19]; and examining governance-related issues such as accountability, traceability, and responsible AI oversight [22,24]. Although these directions collectively advance the understanding of explainability, they are rarely integrated within a unified evaluation framework.

As a result, limited attention has been given to explainability as an operational diagnostic layer under repeated retraining conditions. This limitation becomes particularly important in high-stakes analytics environments where small variations in probability near decision thresholds can produce materially different operational outcomes. Although recent studies increasingly acknowledge the instability of explanations, fewer works examine whether explainability diagnostics can systematically reveal uncertainty-sensitive behavior, reproducibility limitations, and boundary-sensitive decision regions within predictive pipelines [10,29]. Consequently, the governance implications of explanation stability remain insufficiently operationalized in empirical evaluations of XAI.

Despite growing interest in explanation robustness, reproducibility, and AI governance, these research streams have largely evolved in parallel [9–11,22]. Existing stability studies primarily evaluate the robustness of explanations as a technical property of interpretability methods [9,10,19], whereas governance-oriented studies focus on accountability, oversight, and organizational frameworks [11,22,24]. Consequently, limited attention has been given to how explainability-derived diagnostics can be systematically operationalized as measurable evidence for assessing decision stability, explanation reproducibility, uncertainty-sensitive behavior, and monitoring and auditing requirements in AI-enabled decision systems.

Building upon this gap, this study contributes to the literature by integrating previously disconnected streams of research within a unified evaluation perspective. Specifically,

it operationalizes explainability-derived diagnostics as measurable governance-relevant indicators and evaluates their ability to support monitoring, validation, and audit-oriented assessment in AI-enabled decision systems. By examining these diagnostics under controlled retraining conditions, the study extends explainability beyond its traditional role in interpretability and explores its value as a source of governance-relevant evidence.

To address this gap, this study adopts a dual-pipeline perspective in which predictive conditions remain fixed while explainability is introduced as an additional diagnostic layer. As illustrated in Figure 1, the distinction between the two pipelines lies not in predictive capability but in the integration of monitoring-oriented explainability components designed to expose stability, reproducibility, uncertainty-sensitive behavior, and traceability patterns under repeated retraining. Rather than evaluating explainability through predictive improvement, the study examines whether these diagnostics provide measurable insight into the operational behavior of AI-enabled decision systems.

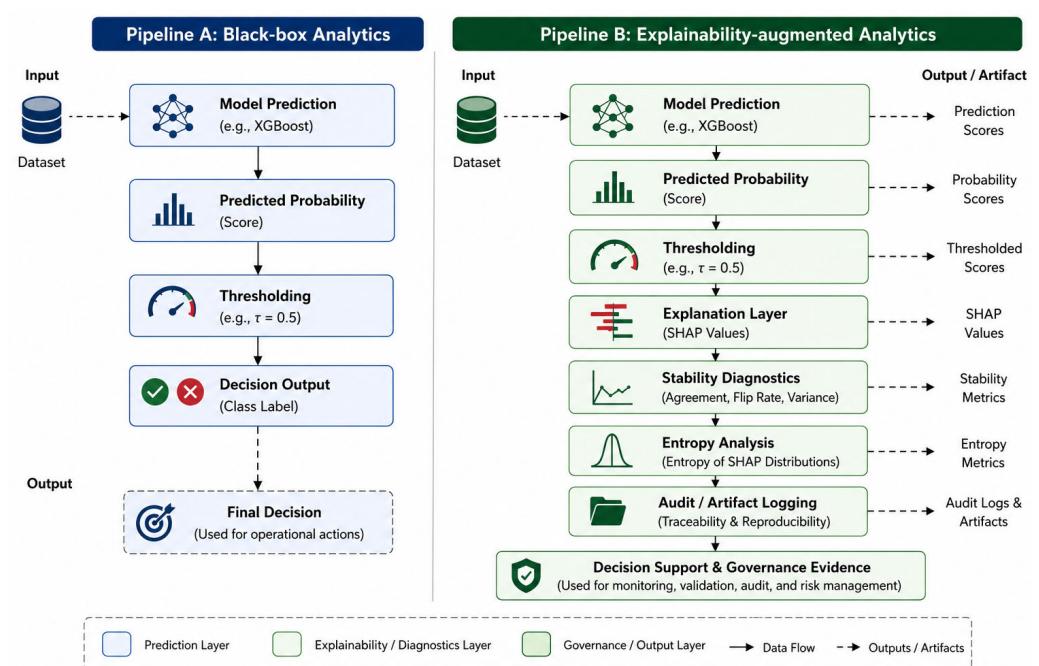


Figure 1. Dual-pipeline framework from prediction to governance evidence.

Accordingly, the study investigates the following research questions:

RQ1. To what extent do model decisions remain stable across repeated retraining conditions in analytics-driven decision systems?

RQ2. To what degree do global explanation structures remain reproducible across retrained models?

RQ3. How does dispersion in local explanation patterns relate to uncertainty-sensitive behavior near decision boundaries?

RQ4. How can explainability diagnostics support traceability, monitoring, and governance-oriented assessment in AI-enabled decision systems?

3. Methodology and Experimental Design

3.1. Experimental Design and Reproducibility Protocol

The experimental framework was designed to evaluate explainability as a diagnostic layer under fixed predictive conditions. Rather than comparing predictive architectures or optimizing classification performance, the study examines whether explanation-based diagnostics expose stability, reproducibility, and uncertainty-sensitive behavior across

repeated retraining conditions. Accordingly, all experiments were conducted in a fixed evaluation environment, with the dataset, validation protocol, preprocessing pipeline, operating threshold, and model configuration remaining unchanged throughout the study.

As illustrated in Figure 1, the experimental design follows a dual-pipeline structure. The first pipeline operates as a conventional prediction-only workflow, generating classification outputs without additional explainability diagnostics. The second pipeline preserves identical predictive conditions while incorporating monitoring-oriented explainability components, including retraining-stability analysis, global-explanation reproducibility assessment, entropy-based local-attribution analysis, and structured traceability artifacts. This design enables diagnostic behavior to be evaluated independently from predictive optimization.

The study employs stratified five-fold cross-validation with fixed data partitions across all experimental runs. Out-of-fold (OOF) predictions were retained throughout the evaluation process to prevent information leakage and maintain a consistent basis for downstream analysis. Because the objective is to examine retraining variability rather than data partition variability, identical train-validation splits were preserved across repeated executions. Under this setup, only internal model randomness changes between runs, allowing variability in predictions and explanations to be attributed to retraining dynamics rather than differences in data composition.

For all predictive, stability, and governance-oriented diagnostics, evaluation was performed using OOF predictions generated under the fixed five-fold cross-validation protocol. In this setting, each observation is assigned to the test partition exactly once and therefore contributes a single unbiased prediction from a model trained without access to that observation. Consequently, the OOF prediction matrix preserves a one-to-one correspondence between observations and predictions, resulting in a total sample count equal to the dataset size ($N = 29,965$). This differs from fold-level aggregation procedures that may count observations multiple times across validation runs. The use of OOF predictions ensures that agreement rates, flip rates, probability variance estimates, entropy diagnostics, and confusion-matrix statistics are computed at the observation level while avoiding information leakage and maintaining statistical independence between training and evaluation samples.

Retraining variability was examined using a multi-seed protocol comprising 10 independent executions under identical validation settings. Ten independent retraining seeds {1–10} were selected to balance statistical robustness and computational feasibility. Repeated retraining under different random initializations enables the assessment of prediction and explanation stability beyond the variability of a single model execution. Using 10 seeds allows agreement rates, prediction variability, and explanation reproducibility to be evaluated across multiple independent runs while maintaining manageable computational requirements. Each retraining cycle preserved the same preprocessing pipeline, hyperparameter configuration, and operating threshold while varying only the internal random initialization conditions of the predictive model. The operational classification threshold was fixed at $\tau = 0.5$ across all experiments to preserve comparability across retraining runs and prevent recalibration of the threshold from influencing stability-related diagnostics.

To strengthen reproducibility and auditability, the experimental workflow incorporated structured artifact logging across all retraining executions. Prediction outputs, explanation artifacts, retraining configurations, validation splits, and seed-specific evaluation results were systematically preserved throughout the study. This design supports deterministic reconstruction of the analytical environment and enables evaluation of diagnostic evidence within a traceable and reproducible framework.

3.2. Data Collection and Preprocessing

The empirical evaluation was conducted using the UCI Default of Credit Card Clients dataset, which contains 30,000 customer records and 23 explanatory variables related to demographic attributes, credit limits, repayment history, bill statements, and payment behavior. The target variable indicates whether a client defaults in the following month. The dataset combines structured tabular characteristics with high-stakes decision implications, class imbalance, and heterogeneous financial behavior patterns, making it suitable for examining decision stability, explanation reproducibility, and uncertainty-sensitive attribution behavior in predictive systems where traceability and auditability are operationally relevant.

The study intentionally uses a single dataset to isolate variability in retraining and explanation behavior under fixed experimental conditions. Since the contribution lies in evaluating explainability diagnostics rather than benchmarking predictive performance across domains, maintaining a stable analytical environment is essential. A fixed dataset allows observed variation in predictions and explanations to be attributed more directly to retraining dynamics rather than differences in domain structure or dataset-specific noise. Similar controlled evaluation strategies are commonly adopted in explanation stability research to assess robustness and reproducibility across repeated experimental settings [9,10,21]. A minimal preprocessing protocol was applied to preserve the dataset's original structure. Duplicate records were removed, resulting in 29,965 observations. No missing values were present, and no resampling, synthetic balancing, or feature engineering was introduced. Class imbalance was intentionally retained because it reflects operational characteristics of credit-risk decision settings and provides a meaningful context for evaluating threshold behavior, false-negative cases, and uncertainty-sensitive explanation patterns. All preprocessing choices remained fixed across validation folds and retraining executions.

3.3. Experimental Design and Model Development

Three supervised classifiers were evaluated under identical validation conditions: Logistic Regression, LightGBM, and XGBoost. Logistic Regression served as a transparent linear baseline, while LightGBM and XGBoost represented tree-based ensemble approaches commonly used in structured tabular prediction tasks. Model selection was based on threshold-independent out-of-fold evaluation, with particular attention to AUC and average precision in the presence of class imbalance. XGBoost achieved the strongest overall predictive performance and was therefore selected as the reference model for the downstream stability and explainability analyses.

The use of XGBoost is also methodologically appropriate for this study's objectives. Tree-based ensemble models are widely used in decision support and risk analytics because they capture non-linear relationships while remaining compatible with feature attribution analysis [2,16]. XGBoost supports efficient integration with SHAP-based explanations, making it suitable for repeated evaluation of global explanation reproducibility and local attribution dispersion under multi-seed retraining conditions.

To preserve comparability across all experiments, the final modeling configuration remained fixed throughout the study. Hyperparameter settings, validation splits, preprocessing decisions, retraining seeds, and the operational threshold were held constant across repeated executions. Consequently, observed differences in decisions, explanations, and entropy-based attribution behavior can be interpreted within a stable modeling environment in which retraining randomness is the primary source of variation.

3.4. Stability and Explainability Diagnostics

The diagnostic layer of the study assesses whether predictive decisions and explanation structures remain stable under repeated retraining. In line with the constructs defined in Table 1, the analysis focuses on four dimensions: decision stability, structural stability of global explanations, dispersion in local explanations, and traceability of diagnostic artifacts. Rather than treating explainability as a performance-enhancing mechanism, the study examines whether explanation-based diagnostics provide measurable evidence about reproducibility, instability, and uncertainty-sensitive behavior within the fixed modeling environment [4,11].

Decision stability was assessed by comparing binary decisions and predicted probabilities across the ten retrained XGBoost models. For each observation i , the predicted probability under the seed s is denoted by $\hat{p}_{i,s}$, while the corresponding binary decision is denoted by $\hat{y}_{i,s}$. The modal decision across retraining runs is represented by $\tilde{y}_i$. Instance-level agreement was computed as

$$A_i = \frac{1}{S} \sum_{s=1}^S 1(\hat{y}_{i,s} = \tilde{y}_i) \quad (1)$$

where S denotes the number of retraining seeds. The flip rate was defined as $F_i = 1 - A_i$. Probability variance was further computed as

$$Var_i = \frac{1}{S-1} \sum_{s=1}^S (\hat{p}_{i,s} - \bar{p}_i)^2 \quad (2)$$

where $\bar{p}_i$ is the mean predicted probability for observation i , across retraining runs.

The agreement rate and probability variance metrics were adopted as operational measures of prediction stability across retraining runs. Similar stability and consistency analyses have been reported in prior studies on model reproducibility and prediction consistency under repeated training conditions [10,28].

Structural stability of explanations was evaluated using SHAP-based global attribution patterns. SHAP was selected because it provides a theoretically grounded and widely adopted feature-attribution framework compatible with tree-based ensemble models. Since the objective is not to compare explanation methods but to evaluate the diagnostic behavior of a fixed explainability layer under repeated retraining, SHAP provides an appropriate basis for reproducibility analysis [4,20,31]. For each retraining run, global feature importance was computed by aggregating mean absolute SHAP values across out-of-fold validation observations. The resulting feature rankings were compared across seeds using Spearman rank correlation and Top-K Jaccard similarity:

$$J(K_a, K_b) = \frac{|K_a \cap K_b|}{|K_a \cup K_b|} \quad (3)$$

where K_a and K_b denote the Top-K feature sets obtained from two retraining runs. These diagnostics assess whether the dominant explanatory drivers remain reproducible across repeated model generations.

Local explanation behavior was examined through entropy-based attribution dispersion. For each observation, absolute SHAP values were normalized to form a local attribution distribution:

$$q_{i,j} = \frac{|\phi_{i,j}|}{\sum_{k=1}^M |\phi_{i,k}|} \quad (4)$$

where $\phi_{i,j}$ denotes the SHAP value of the feature j for observation i , and M is the number of features. Shannon entropy was then computed as follows:

$$H_i = -\sum_{j=1}^M q_{i,j} \log(q_{i,j}) \quad (5)$$

Higher entropy indicates a more dispersed attribution structure, whereas lower entropy indicates that the explanation is concentrated around fewer dominant features. Entropy was used as a structural dispersion indicator to examine whether misclassified observations, particularly false negatives, exhibit different attribution patterns from correctly classified cases. This design enables examination of uncertainty-sensitive prediction regions through the structural behavior of local explanations rather than predictive confidence alone [11,29].

To evaluate whether entropy patterns differed across classification outcomes, observations were grouped by confusion-matrix categories at a fixed operational threshold. Pairwise statistical comparisons were then used to examine differences in entropy distributions across outcome groups. In addition, logistic modeling was applied to assess the association between attribution entropy and false-negative outcomes. This analysis was treated as associative rather than causal. It was intended to determine whether local explanation dispersion provides a useful diagnostic signal for identifying error-prone or boundary-sensitive cases [10,21].

Finally, all diagnostic outputs were preserved as artifacts of reproducibility. These included predicted probabilities, binary decisions, SHAP attribution outputs, entropy values, feature-ranking summaries, retraining configurations, validation partitions, and seed-specific results. This structured logging supports deterministic reconstruction of the analytical workflow and enables inspection of diagnostic evidence beyond a single model execution. Accordingly, the explainability layer is evaluated not only as an interpretability mechanism but also as a reproducible source of evidence for monitoring, validation, and audit-oriented assessment [12,30].

4. Results

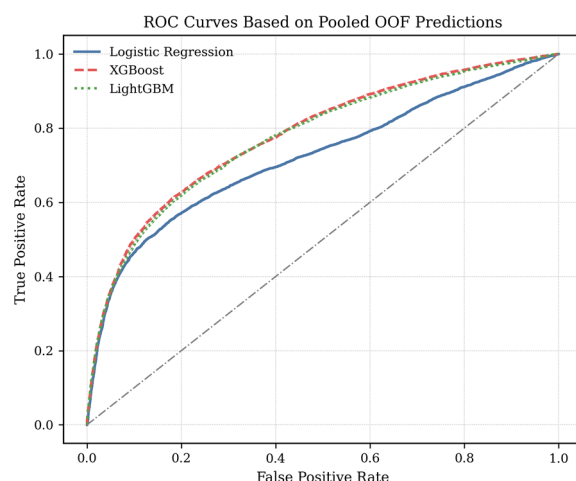
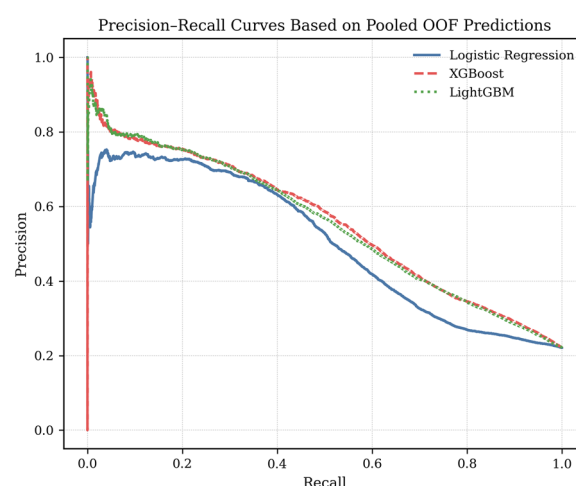
This section presents the empirical findings of the governance-oriented XAI evaluation pipeline. Explainability is evaluated as a structural property, with reproducibility under retraining treated as an operational requirement in high-stakes decision systems. All metrics are computed using OOF predictions under a fixed cross-validation protocol, with stability assessed across multiple seeds at a fixed threshold ($\tau = 0.5$). Results are organized into four analytical layers: (1) predictive performance and model selection, (2) threshold behavior, (3) decision robustness under retraining, and (4) explanation reproducibility. All results are fully traceable to versioned artifacts.

4.1. Predictive Performance (OOF) and Model Selection

Three models (Logistic Regression, LightGBM, and XGBoost) were evaluated under identical cross-validation conditions using pooled OOF predictions. XGBoost achieved the strongest performance ($AUC = 0.7814 \pm 0.0096$; $AP = 0.5579 \pm 0.0128$), followed by LightGBM ($AUC = 0.7762$) and Logistic Regression ($AUC = 0.7223$). The comparative performance of the evaluated models is summarized in Table 2. Figure 2 illustrates the ROC curves based on pooled OOF predictions, while Figure 3 presents the corresponding precision-recall curves highlighting performance differences under class imbalance. Accordingly, XGBoost is selected as the reference model for subsequent analyses. The predictive model is fixed in all experiments to isolate the contribution of explainability diagnostics.

Table 2. Summary of the model’s performance.

Model	AUC (Mean $\pm$ Std)	AP (Mean $\pm$ Std)
Logistic Regression	0.7223 $\pm$ 0.0147	0.5042 $\pm$ 0.0134
LightGBM	0.7762 $\pm$ 0.0091	0.5533 $\pm$ 0.0133
XGBoost	0.7814 $\pm$ 0.0096	0.5579 $\pm$ 0.0128

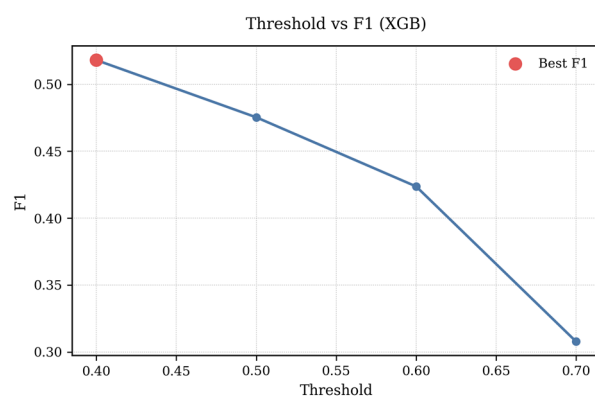
**Figure 2.** ROC curves (pooled OOF) across Logistic Regression, XGBoost, and LightGBM.**Figure 3.** Precision–recall curves (pooled OOF) across Logistic Regression, XGBoost, and LightGBM.

4.2. Threshold Behavior and Confusion Matrix Analysis

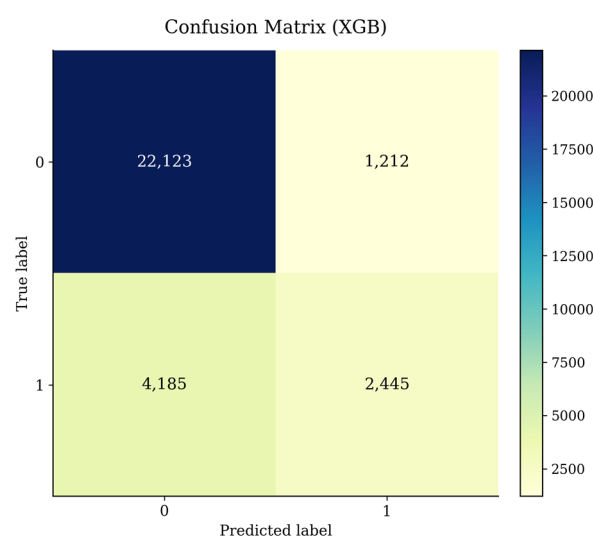
Although model selection is based on threshold-independent metrics (AUC and AP), deployment requires a fixed decision threshold. Using pooled OOF predictions, the highest F1-score is achieved at $\tau = 0.4$ ($F1 = 0.5181$), reflecting a recall-oriented regime. Increasing the threshold improves precision at the expense of recall. Although the highest F1-score was observed at $\tau = 0.4$, a fixed threshold of $\tau = 0.5$ was retained throughout the subsequent analyses to ensure consistency across retraining experiments. Using a common decision threshold enabled evaluation of stability, reproducibility, and attribution-based measures under a fixed operational regime, thereby avoiding the confounding effects of threshold recalibration across repeated retraining experiments. At this threshold, performance is precision = 0.6687 (± 0.0087), recall = 0.3688 (± 0.0133), and $F1 = 0.4753$ (± 0.0112), corresponding to a precision-oriented regime under class imbalance. As shown in Table 3, performance varies across decision thresholds, while Figure 4 illustrates the corresponding trend in F1-score.

Table 3. Threshold sensitivity analysis (XGBoost).

Threshold	Precision	Recall	F1-Score
0.4	0.6236	0.4431	0.5181
0.5	0.6686	0.3688	0.4754
0.6	0.7081	0.3023	0.4237
0.7	0.7562	0.1932	0.3078

**Figure 4.** Variation of F1-score across decision thresholds for the selected XGBoost model.

At $\tau = 0.5$, the confusion matrix (Figure 5) shows 22,123 true negatives, 1212 false positives, 4185 false negatives, and 2445 true positives.

**Figure 5.** Confusion matrix (XGBoost, $\tau = 0.5$).

The model exhibits high specificity, correctly identifying most non-default cases. However, false negatives exceed true positives, indicating a conservative, precision-oriented decision boundary. This asymmetry motivates further analysis to determine whether misclassified instances, particularly false negatives, exhibit distinct explanatory patterns.

4.3. Decision Stability Under Retraining

Decision stability is evaluated by retraining the selected XGBoost model 10 times with different seeds under a fixed 5-fold cross-validation protocol. OOF probabilities are aggregated per instance, and decisions are derived using a fixed threshold ($\tau = 0.5$). The model demonstrates high reproducibility ($n = 29,965$), with a mean agreement rate of 0.9949 and a flip rate of 0.0051. Only 2.17% of instances exhibit at least one decision flip, indicating

that instability is localized to a small subset of boundary-sensitive cases. The mean PD variance is 0.000197 (median = 8.47×10^{-5}), indicating that the probability estimates are highly concentrated. Most observations exhibit near-zero variability, with a sparse tail corresponding to boundary-sensitive cases, consistent with observed decision flips. Figure 6 illustrates the distribution of the variance of predicted probabilities, highlighting strong concentration near zero and a sparse tail associated with boundary-sensitive instances.

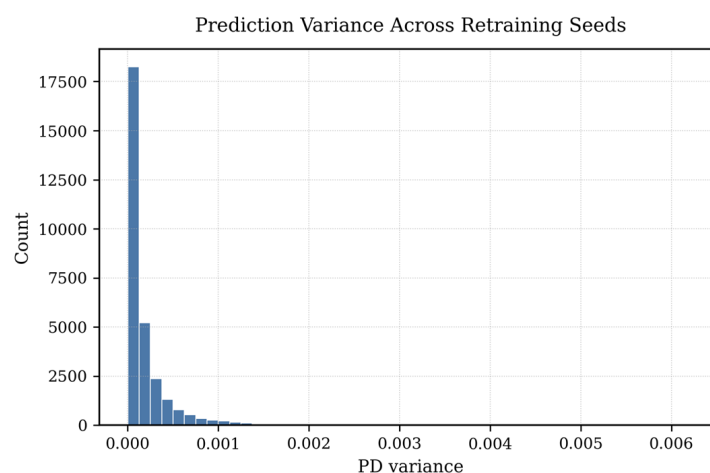


Figure 6. Distribution of predicted probability variance (XGBoost, 10 seeds).

As shown in Figure 7, decision flip rates remain consistent across folds, indicating that specific validation partitions do not drive retraining sensitivity.

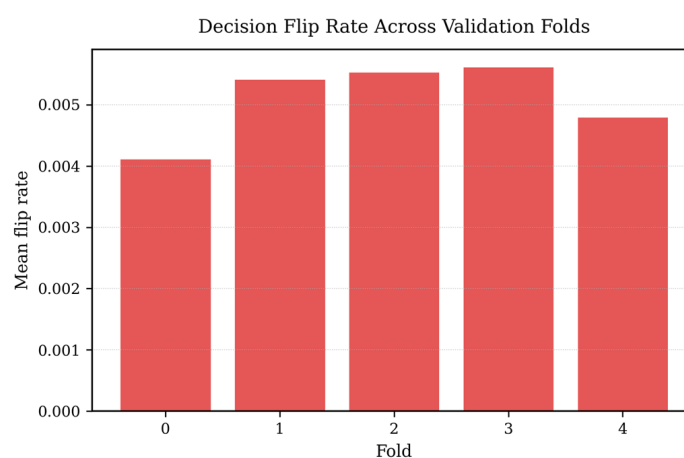


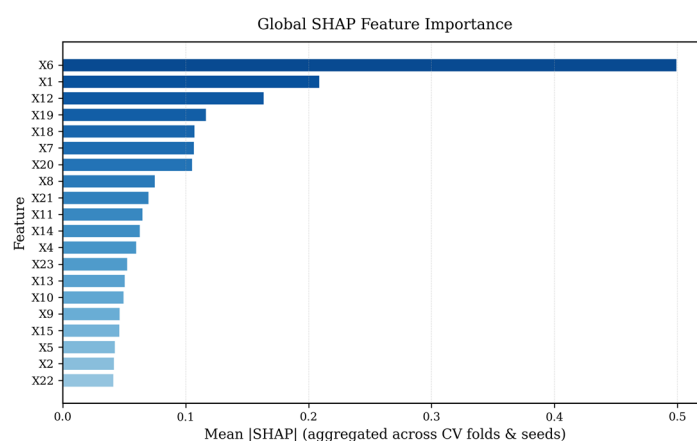
Figure 7. Mean decision flip rate across folds (XGBoost, 10 seeds).

4.4. Global Explanation Reproducibility

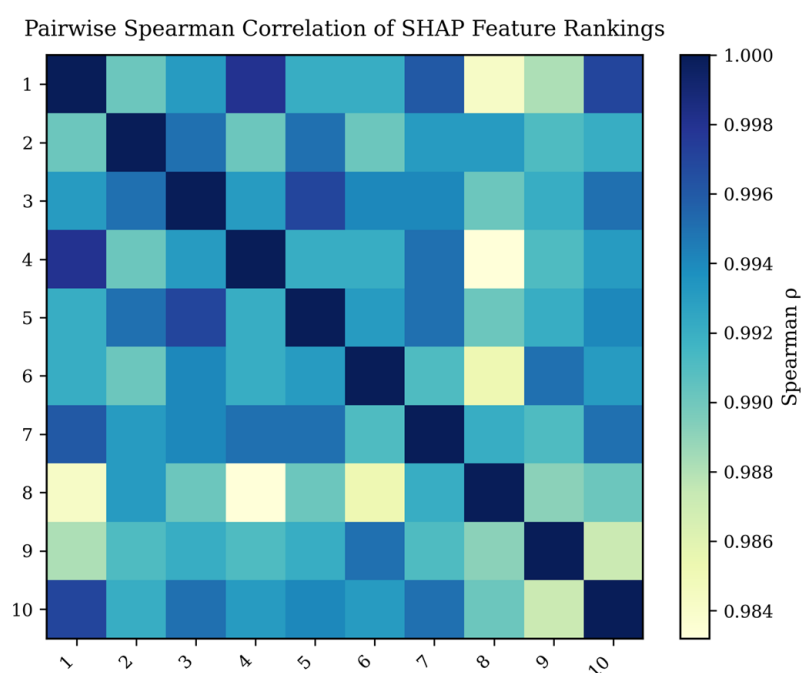
Global explanation stability is assessed using SHAP-based feature importance across 10 seeds and 5 cross-validation folds. Stability is evaluated along three dimensions: attribution magnitude, rank consistency, and driver-set overlap. Global importance is computed as the mean absolute SHAP values aggregated across folds and seeds. A consistent hierarchy of dominant drivers is observed. A small subset of features accounts for most of the explanatory weight, with Feature X6 showing clear dominance, whereas features such as X3 (EDUCATION) and X16 (BILL_AMT5) contribute negligibly. As shown in Table 4 and Figure 8, a consistent hierarchy of dominant features emerges, with a small subset capturing most of the explanatory weight.

Table 4. Global feature importance.

Feature	Mean SHAP	Feature	Mean SHAP	Feature	Mean SHAP
X6	0.4994	X8	0.0751	X10	0.0497
X1	0.2090	X21	0.0701	X9	0.0465
X12	0.1637	X11	0.0651	X15	0.0462
X19	0.1167	X14	0.0628	X5	0.0426
X18	0.1074	X4	0.0600	X2	0.0419
X7	0.1069	X23	0.0528	X22	0.0413
X20	0.1056	X13	0.0507		

**Figure 8.** Global feature importance based on mean absolute SHAP values (XGBoost).

This concentration pattern remains stable under retraining. Pairwise Spearman correlations show near-perfect rank agreement (mean = 0.9921 ± 0.0032) as illustrated in Figure 9, and Top-10 feature sets exhibit strong overlap (Jaccard = 0.9354 ± 0.0880), as shown in Figure 10.

**Figure 9.** Pairwise Spearman rank correlation of global explanations.

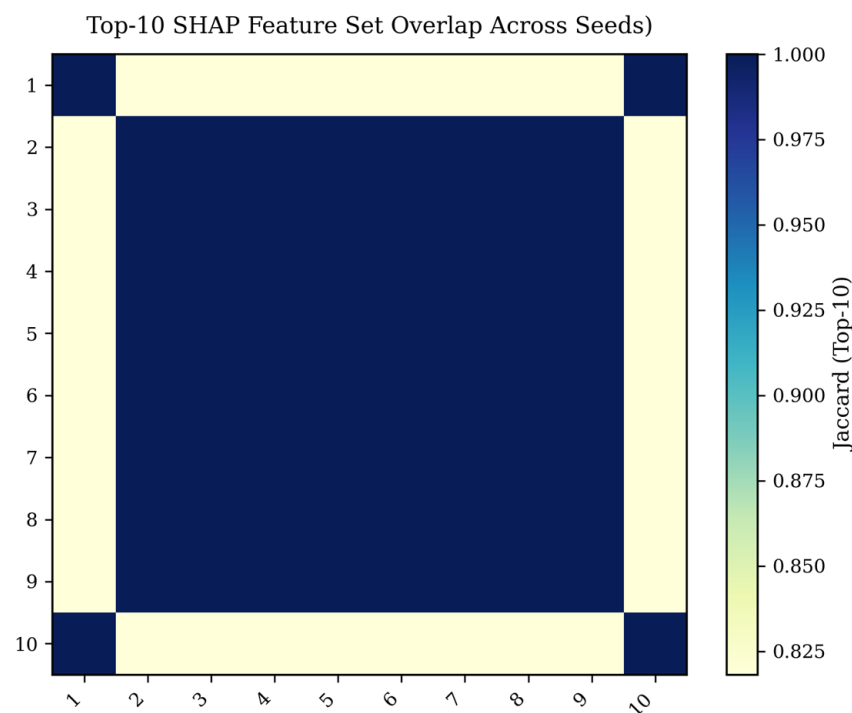


Figure 10. Top-K feature set overlap (Jaccard similarity).

These results show that both the importance and identity of key drivers remain stable under retraining. This structural stability supports the use of global explanations as reliable, reproducible summaries of model behavior, thereby enhancing the auditability, validation, and defensibility of documentation in high-stakes settings.

4.5. Local Explanation Entropy and Error Diagnostics

Shannon entropy is computed over normalized absolute SHAP values to quantify dispersion in local explanations. Across OOF observations, the mean entropy is 2.5205, indicating moderate dispersion. Entropy is further stratified by confusion-matrix outcomes at the fixed threshold ($\tau = 0.5$) using OOF rows aggregated across seeds ($N \times 10 = 299,650$).

Two consistent patterns emerge: false negatives exhibit the highest entropy (more diffuse attributions); in contrast, true positives show the lowest (more concentrated contributions). These patterns are summarized in Table 5 and illustrated in Figure 11. This suggests that misclassified cases, particularly missed defaults, are associated with higher explanatory dispersion, whereas correctly identified positives exhibit more focused attribution patterns.

Table 5. Mean entropy by outcome.

Outcome	Count	Mean Entropy	Std
FN	41,827	2.593991	0.173476
TN	221,465	2.573077	0.151498
FP	11,885	2.161743	0.396083
TP	24,473	2.092756	0.357963

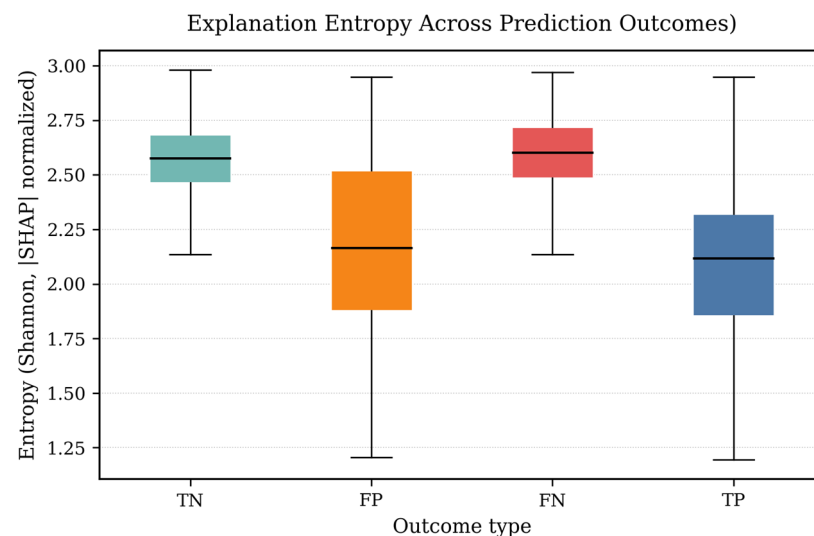


Figure 11. Local explanation entropy by outcome ($\tau = 0.5$).

4.5.1. Pairwise Statistical Comparisons

Non-parametric tests confirm significant differences in entropy across outcome categories. As summarized in Table 6, the strongest separation is observed between false negatives and true positives (2.594 vs. 2.093, $p < 0.001$, rank-biserial = -0.785), with smaller but significant differences across other pairs. These results indicate that entropy systematically varies by outcome type.

Table 6. Pairwise Entropy Differences Across Outcome Categories.

Comparison	Mean Entropy (Group 1)	Mean Entropy (Group 2)	Effect Size (Rank-Biserial)	<i>p</i> -Value
FN vs. TP	2.594	2.093	-0.785	<0.001
FN vs. TN	2.594	2.573	-0.094	<0.001
FP vs. TP	2.162	2.093	-0.106	<0.001

4.5.2. Logistic Modeling of FN Risk via Entropy

A logistic model was used to examine the association between entropy and false-negative (FN) outcomes. Higher entropy was positively associated with FN cases (coefficient = 1.8841, OR ≈ 6.58 , $p < 0.001$), suggesting that observations with more diffuse attribution patterns were substantially more likely to be misclassified as negatives. This relationship should be interpreted as associative rather than causal and may reflect structural dispersion in explanations linked to boundary-sensitive predictions. From a governance perspective, elevated entropy may signal uncertainty, helping to identify cases that warrant additional monitoring and review.

5. Discussion and Conclusions

This study evaluates whether explainability diagnostics provide measurable value beyond predictive accuracy under controlled retraining conditions. The findings show that predictive performance alone does not fully capture the operational reliability of analytics-driven decision systems. Although the model demonstrates stable aggregate performance, residual variability persists in boundary-sensitive cases. Within this context, explainability provides visibility into where instability emerges across retraining conditions. Stable global explanations provide consistent representations of model behavior, while

localized dispersion in feature attributions highlights uncertainty-prone decisions that may otherwise remain undetected.

From a governance perspective, these findings extend beyond interpretability alone. Decision stability becomes observable through agreement rates, flip rates, and variance in probability across retraining runs. Reproducible explanation structures support traceability and strengthen the defensibility of model-based decisions. At the same time, entropy-based dispersion signals provide a practical mechanism for identifying high-risk cases, particularly those associated with false negatives. Structured diagnostic artifacts further support documentation, validation, and audit readiness within AI-enabled decision workflows. Taken together, the results position explainability not merely as a post hoc interpretability tool but as a diagnostic layer capable of supporting monitoring, validation, and governance-oriented assessment without modifying predictive outputs.

This study contributes to XAI and analytics research by repositioning explainability as a governance-relevant property of predictive systems rather than a purely interpretive feature. Prior research has primarily emphasized interpretability techniques and fidelity metrics, often treating explanations as static artifacts [4,6,8,29]. In contrast, this study evaluates explainability through reproducibility, stability, and uncertainty-sensitive behavior under controlled retraining conditions through the following main contributions:

1. Decision stability is operationalized through agreement rates, flip rates, and probability variance under repeated retraining conditions, providing measurable indicators of retraining robustness.
2. Entropy-based attribution dispersion is systematically associated with false-negative outcomes, enabling uncertainty-sensitive regions to be identified without implying causal relationships.
3. The study isolates diagnostic effects from predictive optimization by holding the predictive model constant and introducing explainability as an additional diagnostic layer.

Collectively, these findings support a governance-oriented evaluation perspective in which explainability functions as a reproducible diagnostic component within predictive decision systems. The findings have practical implications for organizations operating analytics-driven decision systems in high-stakes environments. Predictive performance alone may be insufficient for evaluating operational robustness or audit readiness. Decision stability metrics make the sensitivity to retraining observable. Indicators such as flip rates and probability variance enable continuous monitoring of decision consistency without altering the predictive model. Because instability is concentrated in a relatively small subset of boundary-sensitive cases, monitoring efforts can be targeted to high-risk observations rather than applied uniformly across all predictions. Reproducible explanation structures further support audit-oriented evaluation. Consistent feature rankings and dominant explanatory drivers provide traceable representations of model behavior that can assist validation and review processes [29,30]. In addition, entropy-based diagnostics offer a practical uncertainty signal that may help identify misclassification-prone cases for additional inspection. Overall, the findings suggest that explainability diagnostics can support monitoring and oversight functions within predictive decision-making infrastructures beyond what interpretability alone can provide.

These governance-oriented diagnostics also provide practical value for different organizational stakeholders. Agreement and flip-rate measures can help auditors and model validation teams assess whether decision outcomes remain stable across retraining cycles and model updates. Global explanation reproducibility provides evidence of the consistency of feature-importance patterns and can help validate whether the model's reasoning remains structurally stable over time. Entropy-based diagnostics provide an

additional uncertainty signal by highlighting cases with diffuse attribution patterns and an increased risk of false negatives. Such observations may be useful to risk managers and AI governance teams when prioritizing monitoring activities, targeted reviews, and human oversight. Together, these indicators provide traceable evidence that can support validation, audit preparation, model risk management, and ongoing governance processes beyond traditional predictive performance assessment.

This study has several limitations that should be considered when interpreting the findings. First, the empirical evaluation was conducted using a single benchmark dataset under controlled experimental conditions. Although this design was intentionally adopted to isolate retraining-related variability and explanation behavior, it may limit the direct generalizability of the results to other domains, datasets, and decision contexts. Future studies should evaluate the proposed framework across additional datasets and application settings to further assess its external validity. Second, the dataset's original class imbalance was intentionally preserved to reflect realistic credit-risk conditions and support the analysis of threshold behavior and false-negative outcomes. However, different imbalance ratios may influence the behavior of stability and explanation diagnostics, warranting further investigation. Third, the experimental setting does not incorporate real-world deployment drift, evolving data distributions, or temporal changes that may affect model behavior in operational environments. Evaluating governance-oriented diagnostics under such conditions remains an important direction for future research. Finally, the study focuses exclusively on SHAP-based explanations and does not compare alternative approaches to explanation, such as LIME or counterfactual explanations. In addition, fairness-related diagnostics and subgroup-specific stability analyses were outside the scope of the present work. Future research may extend the proposed framework by examining whether the observed stability, reproducibility, and entropy-related patterns remain consistent across different explanation methods, demographic subgroups, and deployment scenarios.

Despite these limitations, the study provides an empirically grounded evaluation of explainability from a governance-oriented perspective. The findings demonstrate that explainability diagnostics can expose measurable patterns of stability, reproducibility, and uncertainty that remain difficult to observe through predictive performance metrics alone. Taken together, this work advances a governance-oriented evaluation framework in which explainability is treated not merely as a post hoc interpretability feature but as a measurable, reproducible diagnostic component of predictive decision-making infrastructures.

Author Contributions: Conceptualization, S.V. and A.A.A.-A.; validation, S.V. and A.A.A.-A.; formal analysis, S.V.; data curation, S.V.; writing—original draft preparation, S.V., A.A.A.-A., O.O.A.-A. and P.S.; writing—review and editing, A.A.A.-A., O.O.A.-A. and P.S.; visualization, S.V.; supervision, A.A.A.-A. and P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at <https://doi.org/10.24432/C55S3H>.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Belle, V.; Papantonis, I. Principles and Practice of Explainable Machine Learning. *Front. Big Data* **2021**, *4*, 688969. [[CrossRef](#)] [[PubMed](#)]
2. Coussement, K.; Abedin, M.Z.; Kraus, M.; Maldonado, S.; Topuz, K. Explainable AI for enhanced decision-making. *Decis. Support Syst.* **2024**, *184*, 114276. [[CrossRef](#)]
3. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [[CrossRef](#)] [[PubMed](#)]
4. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [[CrossRef](#)]
5. Mohamed, A.; Abdelqader, K.; Shaalan, K. Explainable Artificial Intelligence: A systematic Review of Progress and Challenges. *Intell. Syst. Appl.* **2025**, *28*, 200595. [[CrossRef](#)]
6. Kadir, M.A.; Mosavi, A.; Sonntag, D. Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In Proceedings of the 2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES), Nairobi, Kenya, 26–28 July 2023; pp. 000111–000124. [[CrossRef](#)]
7. Le, P.Q.; Nauta, M.; Nguyen, V.B.; Pathak, S.; Schlötterer, J.; Seifert, C. Benchmarking eXplainable AI—A Survey on Available Toolkits and Open Challenges. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; pp. 6665–6673. [[CrossRef](#)] [[PubMed](#)]
8. Miró-Nicolau, M.; Jaume-i-Capó, A.; Moyà-Alcover, G. A comprehensive study on fidelity metrics for XAI. *Inf. Process. Manag.* **2025**, *62*, 103900. [[CrossRef](#)]
9. Alvarez-Melis, D.; Jaakkola, T.S. On the Robustness of Interpretability Methods. *arXiv* **2018**, arXiv:1806.08049. [[CrossRef](#)]
10. Papagiannidis, E.; Mikalef, P.; Conboy, K. Responsible artificial intelligence governance: A review and research framework. *J. Strateg. Inf. Syst.* **2025**, *34*, 101885. [[CrossRef](#)]
11. Ballegeer, M.; Bogaert, M.; Benoit, D.F. Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring. *Eur. J. Oper. Res.* **2025**, *326*, 630–640. [[CrossRef](#)]
12. Hu, X.; Zhu, M.; Feng, Z.; Stanković, L. Manifold-based Shapley explanations for high-dimensional correlated features. *Neural Netw.* **2024**, *180*, 106634. [[CrossRef](#)] [[PubMed](#)]
13. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 220–229. [[CrossRef](#)]
14. Kalra, V. Explainable AI (XAI) Governance in CCaaS Platforms: A Trust, Compliance, and Accountability Model for Enterprise Contact Centers. *J. Inf. Syst. Eng. Manag.* **2025**, *10*, 1300–1308. [[CrossRef](#)]
15. Chang, T.-S.; Bau, D.-Y. Explainable artificial intelligence (XAI) in business management research: A success/failure system perspective. *J. Electron. Bus. Digit. Econ.* **2025**, *4*, 36–53. [[CrossRef](#)]
16. Kostopoulos, G.; Davrazos, G.; Kotsiantis, S. Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. *Electronics* **2024**, *13*, 2842. [[CrossRef](#)]
17. Olan, F.; Spanaki, K.; Ahmed, W.; Zhao, G. Enabling explainable artificial intelligence capabilities in supply chain decision support making. *Prod. Plan. Control.* **2025**, *36*, 808–819. [[CrossRef](#)]
18. Sadeghi, R.K.; Ojha, D.; Kaur, P.; Mahto, R.V.; Dhir, A. Explainable artificial intelligence and agile decision-making in supply chain cyber resilience. *Decis. Support Syst.* **2024**, *180*, 114194. [[CrossRef](#)]
19. Burger, C.; Chen, L.; Le, T. “Are Your Explanations Reliable?” Investigating the Stability of LIME in Explaining Text Classifiers by Marrying XAI and Adversarial Attack. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 12831–12844. [[CrossRef](#)]
20. Hermosilla, P.; Berrios, S.; Allende-Cid, H. Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models. *Appl. Sci.* **2025**, *15*, 7329. [[CrossRef](#)]
21. Yuan, H.; Liu, M.; Kang, L.; Miao, C.; Wu, Y. An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. *arXiv* **2023**, arXiv:2204.11351. [[CrossRef](#)]
22. Jafari Meimandi, K.; Reuel, A.; Aranguiz-Dias, G.; Rahama, H.; Ayadi, A.-E.; Boullier, X.; Verdo, J.; Montanie, L.; Kochenderfer, M. An Adaptive Responsible AI Governance Framework for Decentralized Organizations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Madrid, Spain, 20–22 October 2025; Volume 8, pp. 1324–1334. [[CrossRef](#)]
23. Zhu, R.; Christensen, C.; Zarrin, B.; Bækgaard, P.; Alstrøm, T. Towards Trustworthy AI in Demand Planning: Defining Explainability for Supply Chain Management. In Proceedings of the 17th International Conference on Agents and Artificial Intelligence, Porto, Portugal, 23–25 February 2025; Volume 3, pp. 1245–1253. [[CrossRef](#)]
24. Camilleri, M.A. Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert Syst.* **2024**, *41*, e13406. [[CrossRef](#)]

25. Gerlach, J.; Hoppe, P.; Jagels, S.; Licker, L.; Breitner, M.H. Decision support for efficient XAI services—A morphological analysis, business model archetypes, and a decision tree. *Electron. Mark.* **2022**, *32*, 2139–2158. [[CrossRef](#)]
26. Zhang, C.; Cho, S.; Vasarhelyi, M. Explainable Artificial Intelligence (XAI) in auditing. *Int. J. Account. Inf. Syst.* **2022**, *46*, 100572. [[CrossRef](#)]
27. Kuehnert, B.; Kim, R.; Forlizzi, J.; Heidari, H. The “Who”, “What”, and “How” of Responsible AI Governance: A Systematic Review and Meta-Analysis of (Actor, Stage)-Specific Tools. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 23–26 June 2025; pp. 2991–3005. [[CrossRef](#)]
28. Gawantka, F.; Just, F.; Savelyeva, M.; Wappler, M.; Lässig, J. A Novel Metric for Evaluating the Stability of XAI Explanations. *Adv. Sci. Technol. Eng. Syst. J.* **2024**, *9*, 133–142. [[CrossRef](#)]
29. Zheng, X.; Shirani, F.; Chen, Z.; Lin, C.; Cheng, W.; Guo, W.; Luo, D. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. *arXiv* **2025**, arXiv:2410.02970. [[CrossRef](#)]
30. Sandu, I.; Wiersma, M.; Manichand, D. Time to audit your AI algorithms. *Maandbl. Voor Account. En. Bedrijfsecon.* **2022**, *96*, 253–265. [[CrossRef](#)]
31. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Volume 2016, pp. 1135–1144. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.