

An Empirical Evaluation of Explainability Drift Under Post-Training Quantization of Machine Learning Models

Mahendra .¹, Aparna Khare¹, 

1. Ministry of Electronics and Information Technology, Government of India, National Informatics Centre, Lucknow, IND

Received: April 25, 2026 | Review began: May 21, 2026 | Review ended: May 27, 2026 | Published: June 04, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Post-training quantization (PTQ) is widely adopted to reduce the computational and memory footprint of machine learning models deployed in resource-constrained environments. While its impact on predictive performance has been extensively studied, its implications for model explainability remain insufficiently understood. Explanations are often assumed to remain reliable as long as predictive accuracy is preserved. This assumption, however, has rarely been examined empirically. In this work, we present a systematic empirical study of explainability stability under PTQ across three representative model families, namely, Logistic Regression, Random Forest, and XGBoost, using the IEEE-CIS Fraud Detection dataset. Explanation stability is evaluated using complementary global and local attribution metrics, including feature rank correlation, mean attribution drift, and instance-level cosine similarity of SHapley Additive exPlanations explanations. Models are trained once in full precision and subsequently quantized to INT8 and INT4 without retraining to isolate numerical precision effects. The results indicate that predictive performance and explanation stability are not inherently aligned. While some models retain stable predictive metrics under reduced precision, measurable attribution drift may still occur. In particular, aggressive quantization can induce substantial instability in explanation behavior even when aggregate performance remains unchanged. These findings present that explainability cannot be inferred solely from accuracy metrics and should be explicitly evaluated when deploying quantized models in high-stakes decision-making systems.

Categories: Evaluation Methods, Explainable AI, Machine Learning (ML)

Keywords: post-training quantization, explainability drift, quantized machine learning, explainability evaluation, explainable artificial intelligence, attribution stability, model compression, model reliability

Introduction

Post-training quantization (PTQ) has emerged as a widely used technique for reducing model size and improving deployment efficiency without retraining. By converting model parameters from full precision to lower numerical formats, PTQ enables reduced memory usage and faster inference while often preserving predictive performance. In practical settings where computational constraints, limited access to training data, or reproducibility requirements restrict the feasibility of quantization-aware training, PTQ presents a deployment-friendly alternative.

How to cite this article:

. M, Khare A (June 04, 2026) An Empirical Evaluation of Explainability Drift Under Post-Training Quantization of Machine Learning Models. Cureus J Comput Sci 3 : es44389-026-00100-3. DOI <https://doi.org/10.7759/s44389-026-00100-3>

Alongside, advances in explainable artificial intelligence have introduced several attribution-based methods to enhance model transparency. These methods are increasingly used in operational systems to support auditability and decision review.

Because attribution methods depend on the numerical values and structural characteristics of the trained model, modifications introduced through precision reduction may alter explanation behavior, even when aggregate predictive metrics remain stable. Despite extensive study of predictive degradation under quantization, comparatively limited empirical attention has been given to the stability of model explanations following numerical precision reduction. This gap is particularly relevant in impact-critical domains such as fraud detection, where attribution outputs may facilitate audit and compliance functions. In such settings, preservation of predictive accuracy alone may be insufficient if explanation behavior shifts under compression.

This study examines whether post-training quantization preserves explanation stability across model families of varying structural complexity, namely Logistic Regression, Random Forest, and XGBoost, under 8-bit Integer representation (INT8) and 4-bit Integer representation (INT4) precision levels. The empirical evaluation is conducted on the IEEE-CIS Fraud Detection dataset [1], a high-dimensional and highly imbalanced binary classification dataset. All models are trained once in full precision and subsequently quantized without retraining in order to isolate numerical precision effects and maintain reproducibility. Quantization-aware training has been intentionally excluded to focus specifically on compression-induced perturbations.

Related work

Foundational work with reference to PTQ has shown that reduced-precision inference can substantially decrease memory and computational requirements while maintaining competitive predictive performance across a range of architectures [2-4]. The majority of PTQ research evaluates quantization primarily through predictive accuracy, latency, and memory footprint, with limited attention to secondary properties such as model reliability or interpretability.

In parallel, the field of explainable artificial intelligence (XAI) has developed attribution-based methods, including SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), to enhance transparency of complex models [5-6]. Interpretability has become increasingly important in critical decision systems, where explanation outputs may support audit processes, regulatory compliance, and operational oversight [7]. Beyond interpretability itself, recent literature has also examined robustness and stability of explanations under input perturbations, parameter noise, and structural modifications, demonstrating that explanations or attribution outputs may vary under small changes to model parameters or inputs [8-11].

More broadly, prior work has demonstrated that predictive metrics alone may not fully characterize model behavior, particularly under distributional shifts or calibration effects [12]. However, existing studies typically analyze predictive robustness and explanation robustness separately, focusing either on performance stability under model perturbations [2-4] or on the sensitivity of attribution methods [8-11]. Systematic empirical investigation of how numerical precision reduction, specifically post-training quantization, affects attribution stability remains limited.

In particular, there is sparse empirical analysis of explanation drift in classical tabular models subjected to PTQ, despite their continued use in operational domains such as fraud detection [13-14]. Prior work has examined SHAP-based explanations for tree and linear models in tabular settings [5,15], as well as the effects of quantization on predictive performance [2-4], but the interaction between reduced numerical precision and SHAP-based attribution behavior remains underexplored.

This work contributes to this gap by providing a controlled empirical assessment of attribution stability across model families and quantization levels, with explicit comparison between predictive degradation and explanation drift under post-training quantization.

How to cite this article:

Research questions

The central objective of this study is to examine whether post-training quantization preserves explanation stability across model families of varying structural complexity. More specifically, this work investigates whether numerical precision reduction alters attribution behavior in ways that are consistent, proportional, or structurally divergent across linear and tree-based models.

In an attempt to address this objective in a systematic manner, the analysis is guided by two related investigative dimensions. First, the study evaluates the extent to which predictive performance degradation under reduced precision aligns with changes in attribution stability. This seeks to determine whether shifts in conventional performance metrics serve as reliable indicators of explanation drift. Second, the study considers whether explanation instability may arise even in cases where aggregate predictive metrics remain largely unchanged. This inquiry assesses the possibility that attribution geometry may shift independently of classification performance, thereby challenging the sufficiency of predictive validation alone when deploying compressed models.

Together, these research questions shape the experimental design, metric selection, and interpretive analysis presented in the subsequent sections.

Materials And Methods

Methodology

This study adopted a controlled empirical design to evaluate explainability stability under PTQ across multiple machine learning model families. The methodology was structured to isolate the effects of numerical precision reduction on model interpretability in a systematic manner and minimize the variations introduced by inconsistent preprocessing or training procedures.

A standardized experimental workflow was followed throughout the study. Initially, models were trained using full-precision 32-bit floating-point representation (FP32) representations under uniform preprocessing and training conditions. Subsequently, PTQ techniques were applied to generate reduced-precision model variants while preserving identical architectural and training configurations. This approach allowed the full-precision and quantized implementations to be compared under consistent conditions.

To maintain consistency across multiple experimental iterations, all workflows were implemented in Python using identical data preparation, inference, and explainability generation procedures for each model family. Reproducibility was further supported by keeping the evaluation conditions fixed and using standardized metric calculations throughout the study.

The methodological framework was designed to examine how quantization-induced numerical approximations affect the stability, consistency, and interpretability of explainability outputs. To assess these effects, comparative evaluation was carried out between the original and quantized models using multiple explainability-focused metrics and analytical observations.

Results And Discussion

Experimental setup

The overall experimental pipeline is illustrated in Figure 1. A single full-precision FP32 model was trained and subsequently quantized, post-training at different precision levels, in order to ensure that any observed differences arise solely from numerical precision reduction. All model variants were evaluated on a common test set using both predictive and attribution stability (explainability) metrics.

How to cite this article:

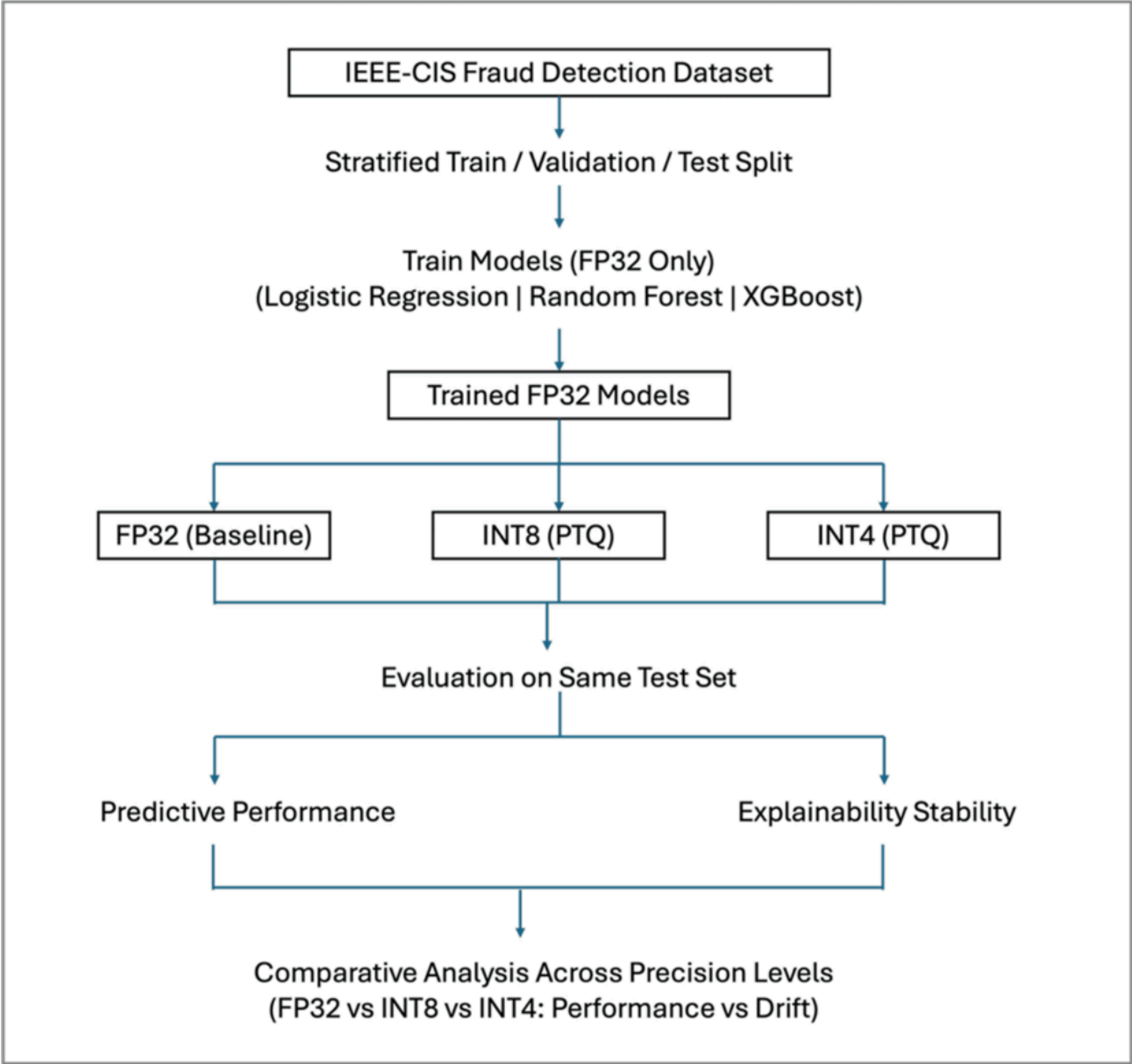


FIGURE 1: Experimental pipeline for evaluating explainability drift under PTQ.

A single FP32 model is trained and quantized without retraining and evaluated across precision levels using predictive and attribution stability metrics. PTQ, Post-training quantization; FP32, 32-bit Floating-point representation; INT8, 8-bit Integer representation; INT4, 4-bit Integer representation

Dataset

The empirical evaluation was conducted on the IEEE-CIS Fraud Detection dataset, which contains 420 engineered numerical and encoded categorical features derived from transactional metadata [1].

To preserve class distribution consistency across evaluation phases, a stratified train/validation/test split was performed. Following preprocessing, the dataset was partitioned into training, validation, and test sets containing 354,324, 118,108, and 118,108 samples, respectively. The fraud prevalence was noted to be approximately 3.5% across all splits, reflecting substantial class imbalance. All predictive and attribution metrics reported in this study were computed on the held-out test set.

How to cite this article:

The combination of high dimensionality and pronounced class imbalance makes this dataset well suited for evaluating both ranking-based performance metrics and the stability of feature attributions under numerical perturbation. While the imbalanced setting amplifies sensitivity to false-positive shifts, the large feature space enables meaningful assessment of attribution vector variation under reduced precision.

Model Families

Three model families were selected to represent increasing structural complexity and decision boundary characteristics. First, Logistic Regression, representing a linear decision surface directly parameterized by feature weights; second, Random Forest, representing a bagged ensemble of decision trees with non-linear threshold-based partitions; and third, XGBoost, representing gradient-boosted trees with additive structure and finely tuned split thresholds.

All models were trained once using full precision, that is, FP32. The trained models were then reused across quantization experiments without retraining. This design ensured that observed effects were attributable solely to numerical precision reduction rather than parameter re-optimization.

Quantization Approach

PTQ was applied at two precision levels, namely, INT8, resulting in moderate precision reduction commonly used in deployment settings, and INT4, resulting in aggressive precision reduction with greater expected numerical distortion.

For Logistic Regression and Random Forest, weight-level quantization was applied directly to model parameters. For XGBoost, due to tooling constraints and internal representation limitations, quantization was applied at the explanation layer rather than modifying internal tree thresholds. This approach isolated the effect of reduced precision on attribution computation while preserving the original decision structure.

Evaluation metrics

Predictive performance was evaluated using Precision, Recall, F1-score, receiver operating characteristic area under the curve (ROC-AUC), Precision-Recall area under the curve (PR-AUC), and confusion matrix analysis. Given the imbalanced nature of the dataset, metric selection was carefully considered: ROC-AUC provided threshold-independent discrimination assessment, while precision, recall, F1-score, and PR-AUC being particularly sensitive to false-positive behavior, were relevant under quantization-induced perturbations.

Precision, Recall, and F1-score quantified trade-offs in minority-class detection, while confusion matrices enabled direct examination of false-positive and false-negative shifts induced by quantization.

Explainability stability was assessed using SHAP-based attributions computed for both FP32 and quantized models. To quantify attribution drift, three complementary measures were employed. Global stability was measured using Spearman rank correlation between feature importance orderings across precision levels. Magnitude-level deviation was quantified using mean absolute attribution drift. Instance-level explanation consistency was evaluated using cosine similarity between local SHAP vectors. Together, these metrics captured both global feature ranking stability and local attribution alignment.

All evaluations were conducted on the identical held-out test set across precision levels to ensure direct comparability. No additional hyperparameter tuning or retraining was performed following quantization, thereby isolating numerical precision effects.

Empirical results

Predictive Performance

Table 1 summarizes the predictive performance of the full-precision (FP32) models on the test set.

Model	Logistic Regression	Random Forest	XGBoost
Precision	0.1256	0.1910	0.2518
Recall	0.6951	0.7398	0.8052
F1-Score	0.2127	0.3036	0.3836
ROC-AUC	0.8374	0.8906	0.9343
PR-AUC	0.2961	0.5212	0.6504

TABLE 1: FP32 baseline performance.

ROC-AUC, Receiver operating characteristic area under the curve; PR-AUC, Precision-Recall area under the curve; FP32, 32-bit Floating-point representation

XGBoost achieved the highest ROC-AUC and PR-AUC, indicating superior ranking performance under class imbalance. Random Forest demonstrated intermediate performance, while Logistic Regression yielded comparatively lower PR-AUC, reflecting reduced discrimination capacity relative to tree-based models. Given the low fraud prevalence (~3.5%), PR-AUC provided a more informative performance indicator than ROC-AUC. The baseline established a performance hierarchy consistent with increasing model complexity.

Table 2 presents predictive metrics under INT8 PTQ.

Model	Logistic Regression	Random Forest	XGBoost
Precision	0.1209	0.0491	0.2518
Recall	0.6682	0.9085	0.8052
F1-Score	0.2048	0.0931	0.3836
ROC-AUC	0.8235	0.7172	0.9343
PR-AUC	0.2040	0.1708	0.6504

TABLE 2: INT8 performance.

INT8, 8-bit Integer representation; ROC-AUC, Receiver operating characteristic area under the curve; PR-AUC, Precision-Recall area under the curve

How to cite this article:

Logistic Regression exhibited moderate degradation across most metrics, including the decline in PR-AUC. Random Forest showed substantial reduction in precision accompanied by increased recall, indicating a marked rise in false-positive predictions. ROC-AUC also decreased notably. In contrast, XGBoost metrics remained unchanged, as internal tree parameters were not quantized.

The results suggest that INT8 quantization introduced measurable predictive instability in threshold-based ensemble methods, while leaving explanation-layer quantized models unaffected in terms of classification performance.

Table 3 summarizes predictive performance under INT4 precision.

Model	Logistic Regression	Random Forest	XGBoost
Precision	0.0353	0.049	0.2518
Recall	0.9961	0.9107	0.8052
F1-Score	0.0683	0.0930	0.3836
ROC-AUC	0.8120	0.7163	0.9343
PR-AUC	0.1858	0.1693	0.6504

TABLE 3: INT4 performance.

INT4, 4-bit Integer representation; ROC-AUC, Receiver operating characteristic area under the curve; PR-AUC, Precision-Recall area under the curve

Logistic Regression exhibited a near-collapsed behavior under INT4, with recall approaching unity and precision flattening, indicating effective loss of decision boundary resolution. Random Forest maintained similarly unstable behavior observed under INT8, with persistently low precision and reduced ranking performance. XGBoost remained unchanged in predictive metrics due to explanation-layer-only quantization.

Overall, aggressive PTQ amplifies false-positive inflation in both linear and ensemble models when applied directly to model parameters.

Explainability Stability

Explainability stability was evaluated using global feature rank correlation, mean absolute attribution drift, and cosine similarity of local SHAP vectors.

Table 4 summarizes the Spearman rank correlation between FP32 and the quantized models, assessing the global feature rank stability.

Model	INT8	INT4
Logistic Regression	0.9977	0.8657
XGBoost	0.9517	0.6727
Random Forest	0.8551	0.8542

TABLE 4: Performance Spearman rank correlation (FP32 vs Quantized).

FP32, 32-bit Floating-point representation

Logistic Regression maintained near-perfect global rank stability under INT8, with moderate degradation under INT4. XGBoost exhibited greater rank reordering under INT4 relative to INT8. Random Forest showed moderate rank instability across both precision levels.

Table 5 summarizes the mean absolute attribution drift of the quantized models.

Model	INT8	INT4
Logistic Regression	0.000176	0.002633
XGBoost	0.000499	0.004596
Random Forest	0.000839	0.000849

TABLE 5: Mean absolute attribution drift.

INT8, 8-bit Integer representation; INT4, 4-bit Integer representation

Attribution magnitude drift increased under INT4 across models. XGBoost exhibited the largest magnitude shift under aggressive precision reduction, consistent with sensitivity at the attribution computation layer.

To assess local explanation similarity, Table 6 presents the SHAP cosine similarity for the quantized models.

How to cite this article:

Model	INT8	INT4
Logistic Regression	0.9973	0.8804
XGBoost	0.9986	0.8854
Random Forest	0.3715	0.3697

TABLE 6: Local SHAP cosine similarity.

INT8, 8-bit Integer representation; INT4, 4-bit Integer representation

Logistic Regression and XGBoost retained high local similarity under INT8. Under INT4, both exhibited moderate divergence. Random Forest displayed substantially lower local similarity under both INT8 and INT4, indicating significant instance-level explanation instability. Notably, explanation collapse in Random Forest occurred even in cases where recall increased, illustrating a decoupling between prediction metrics and explanation fidelity.

Explainability Drift

Predictive degradation for each model m under quantization level q was defined as:

$$\Delta PR_{AUC_{m,q}} = PR_{AUC_{m,q}} - PR_{AUC_{m,FP32}} \quad (1)$$

where $PR_{AUC_{m,FP32}}$ denoted the baseline full precision performance.

To quantify the stability of local explanations under quantization, explanation drift can be defined as the divergence between attribution vectors produced by the full-precision model and its quantized counterpart for identical inputs. For a model m under quantization level q , explanation drift will measure the geometric deviation between the corresponding SHAP attribution vectors and can be defined as:

$$\text{Drift}_{m,q} = 1 - \cos(\phi_{m,FP32}, \phi_{m,q}) \quad (2)$$

where $\phi_{m,FP32}$ and $\phi_{m,q}$ denote the local SHAP attribution vectors generated by the full-precision and quantized models, respectively, and $\cos(.)$ represents cosine similarity. Higher drift values indicate greater directional divergence between explanations.

Figure 2 presents a two-dimensional mapping of predictive degradation against local explanation drift to visualize the relationship between predictive degradation and explainability instability across model families and precision levels.

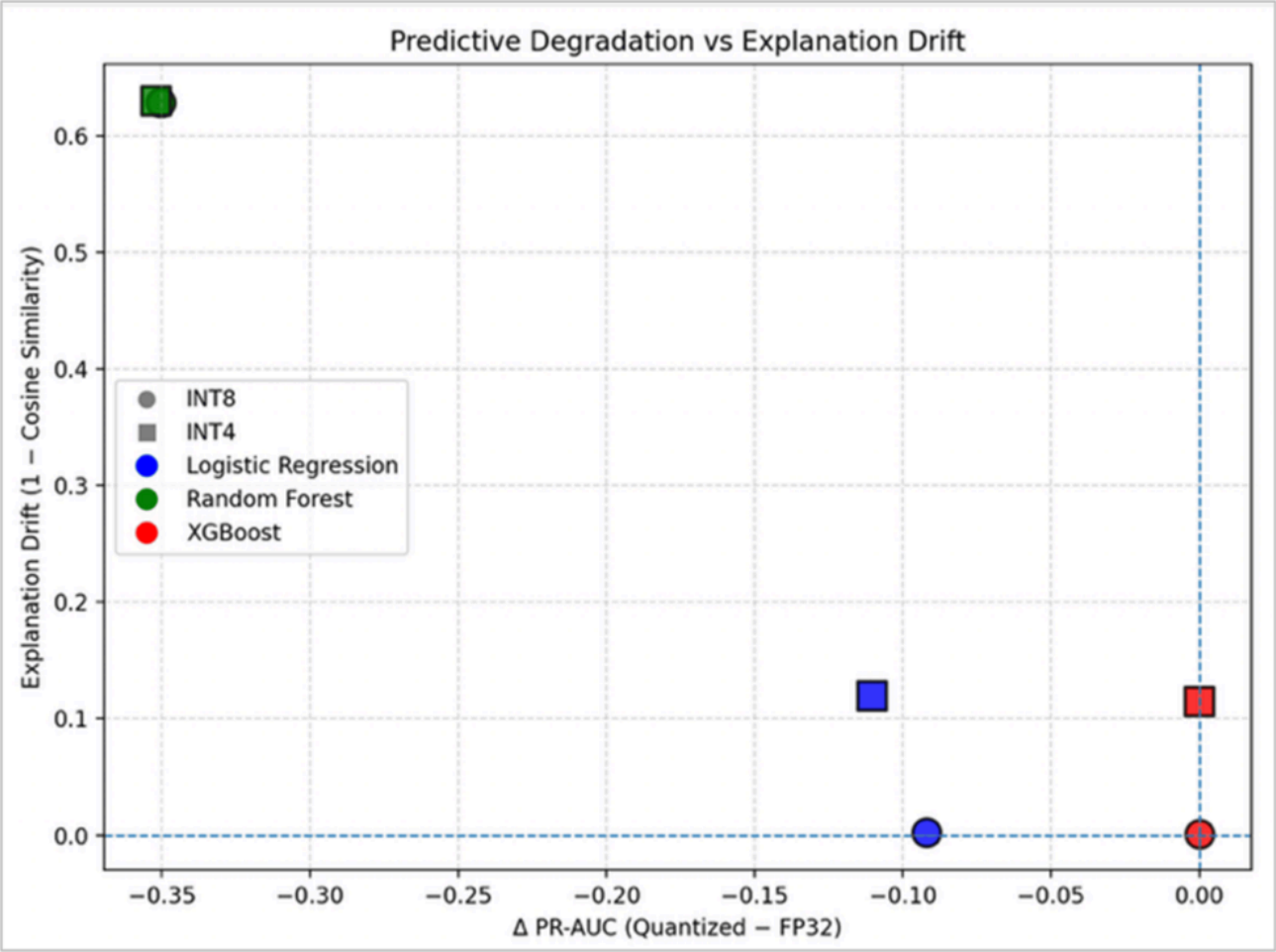


FIGURE 2: Predictive degradation vs explanation drift.
INT8, 8-bit Integer representation; INT4, 4-bit Integer representation; FP32, 32-bit Floating-point representation; PR-AUC, Precision-Recall area under the curve

Analysis and discussion

The empirical results from the study indicated that predictive performance and attribution stability were not inherently aligned under PTQ. It was observed that across model families, attribution drift increased disproportionately relative to changes in conventional predictive metrics. In several instances, nearly identical ROC-AUC scores were accompanied by noticeable changes in both global feature rankings and local attribution vectors, suggesting that standard validation metrics may not fully capture compression-induced changes in interpretability.

Tree-based ensembles exhibited systematic increases in false-positive rates under reduced precision, consistent with the sensitivity of threshold-based decision rules to numerical perturbations. Substantial instability was observed in local SHAP attributions under aggressive INT4 quantization, indicating structural changes in instance-level explanation patterns.

In contrast, Logistic Regression demonstrated a more gradual degradation pattern as precision decreased, reflecting direct weight perturbation effects on a linear decision surface. At INT4 precision, nearly uniform positive predictions reflected a collapse of margin resolution, that is, the model’s ability to distinguish between classes via its decision boundary or prediction scores, resulting in reduced differentiation in feature attributions.

The plot of predictive degradation vs explainability drift (Figure 1) reveals architecture-dependent clustering patterns rather than a uniform relationship between predictive degradation and explanation drift. Under reduced precision, models occupy distinct instability regimes, confirming that degradation in predictive metrics does not reliably predict corresponding changes in explanation fidelity. Notably, certain configurations exhibit stable predictive metrics alongside measurable explanation drift. This divergence indicates that attribution geometry may shift even when aggregate predictive integrity appears preserved. In operational contexts, these findings imply that it is crucial that compressed models should be evaluated along an independent axis of explanation stability in addition to predictive performance.

Limitations

The conclusions drawn from this study are subject to certain limitations. First, since the empirical evaluation has been conducted solely on the IEEE-CIS Fraud Detection dataset, the findings should not be interpreted as establishing general behavior across domains such as vision, text, or time-series tasks. Second, the quantization procedures implemented reflect a specific implementation of PTQ. Alternative calibration strategies, scaling methods, or hardware-aware quantization schemes may result in different predictive and attribution behaviors. Third, since the analysis does not include comparisons with quantization-aware training, the results isolate compression-induced effects under PTQ and do not address whether training-time adaptation mitigates explanation drift. Finally, explainability stability has been evaluated using SHAP-based attributions. Other attribution methods may exhibit different sensitivity to reduced precision.

Conclusions

This paper has examined the stability of model explanations under PTQ in a high-dimensional fraud detection setting. The study introduced a controlled empirical framework for measuring attribution drift across precision levels and model families and evaluated the relationship between predictive degradation and explanation instability. The results indicate that predictive stability and explanation stability are not inherently aligned and that compressed models can retain decision performance while exhibiting altered attribution structure. Consequently, within the scope of this study, the findings imply that evaluation of quantized models in decision-critical applications should not rely solely on predictive metrics. Attribution stability warrants independent assessment, and the divergence between global feature ranking and local explanation consistency highlights the need for multi-level explainability evaluation under compression.

Future work

This analysis was limited to PTQ on a single tabular dataset and SHAP-based explanations. Future work may extend this framework to additional datasets, neural architectures, alternative compression strategies including quantization-aware training. Additionally, cross-method analysis of explainability stability using other attribution methods remains an open direction.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Aparna Khare, Mahendra .

Acquisition, analysis, or interpretation of data: Aparna Khare, Mahendra .

Drafting of the manuscript: Aparna Khare

Critical review of the manuscript for important intellectual content: Aparna Khare, Mahendra .

How to cite this article:

. M, Khare A (June 04, 2026) An Empirical Evaluation of Explainability Drift Under Post-Training Quantization of Machine Learning Models. *Cureus J Comput Sci* 3 : es44389-026-00100-3. DOI <https://doi.org/10.7759/s44389-026-00100-3>

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data supporting the findings of this study are publicly available from the Kaggle repository (IEEE-CIS Fraud Detection dataset) at <https://www.kaggle.com/c/ieee-fraud-detection>. No new data were generated in this study.

Acknowledgements

The authors gratefully acknowledge the IEEE-CIS Fraud Detection dataset community and Kaggle for providing the data used in this study. The data used in this study are publicly available through the Kaggle platform.

References

1. IEEE-CIS Fraud Detection Dataset: Kaggle. (2019). Accessed: March 1, 2026: <https://kaggle.com/competitions/ieee-fraud-detection>.
2. Jacob B, Kligys S, Chen B, Zhu M, Tang M, Howard A: [Quantization and training of neural networks for efficient integer-arithmetic-only inference](#). Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 2704-2713. [10.1109/CVPR.2018.00286](#)
3. Nagel M, Baalen M, Blankevoort T, Welling M: [Data-free quantization through weight equalization and bias correction](#). Proceedings of IEEE/CVF International Conference on Computer Vision. 2019, 1325-1334. [10.1109/ICCV.2019.00141](#)
4. Li Y, Gong R, Tan X, et al.: [BRECQ: pushing the limit of post-training quantization by block reconstruction](#). International Conference on Learning Representations (ICLR). 2021, 1-16.
5. Lundberg SM, Lee SI: [A unified approach to interpreting model predictions](#). Advances in Neural Information Processing Systems (NeurIPS). 2017, 30:4766-4777. [10.48550/arXiv.1705.07874](#)
6. Ribeiro MT, Singh S, Guestrin C: [Why should I trust you? explaining the predictions of any classifier](#). Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 2016, 1135-1144. [10.1145/2939672.2939778](#)
7. Molnar C: [Interpretable Machine Learning: A Guide for Making Black Box Models Explainable](#). Lulu.com, Morrisville; 2020.
8. Kindermans PJ, Hooker S, Adebayo J, et al.: [The \(un\)reliability of saliency methods](#). 267-280. [10.1007/978-3-030-28954-6_14](#)
9. Ghorbani A, Abid A, Zou J: [Interpretation of neural networks is fragile](#). Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33:3681-3688. [10.1609/aaai.v33i01.33013681](#)
10. Yeh CK, Hsieh CY, Suggala AS, Inouye DI, Ravikumar P: [On the \(in\)fidelity and sensitivity of explanations](#). Advances in Neural Information Processing Systems (NeurIPS). 2019, 32:10967-10978.
11. Alvarez-Melis D, Jaakkola TS: [Towards robust interpretability with self-explaining neural networks](#). Advances in Neural Information Processing Systems (NeurIPS). 2018, 31:7786-7795.
12. Guo C, Pleiss G, Sun Y, Weinberger KQ: [On calibration of modern neural networks](#). Proceedings of the International Conference on Machine Learning (ICML). 2017, 70:1321-1330.
13. Chen T, Guestrin C: [XGBoost: a scalable tree boosting system](#). Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 2016, 785-794. [10.1145/2939672.2939785](#)
14. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N: [Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission](#). Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 2015, 1721-1730. [10.1145/2783258.2788613](#)

How to cite this article:

15. Lundberg SM, Erion G, Chen H, et al.: [From local explanations to global understanding with explainable AI for trees](#). Nature Machine Intelligence. 2020, 2:56-67. [10.1038/s42256-019-0138-9](#)