D1

When Explanations Change but Predictions Do Not: Unpacking Explanatory Drift in Financial AI Models

Abdolrahim Tahvildari (PhD) Faculty of Design & Creative Technologies

As financial institutions increasingly deploy “black-box” machine learning algorithms for automated credit scoring, keeping these systems transparent and trustworthy is essential (Pelosi et al., 2025). Explanation techniques such as SHapley Additive exPlanations (SHAP) are widely used to show which factors, for example income or credit history, most influenced a given decision, helping institutions meet this transparency requirement. However, a critical gap remains largely unaddressed: explanatory drift — the tendency for these explanations to shift over time even when nothing appears wrong with the model itself (Cossu et al., 2024; Muschalik et al., 2022).

This pilot study investigates a concerning paradox: an AI model’s explanations for its decisions can change substantially even while its predictive accuracy stays essentially the same. Using a Random Forest classifier trained on a well-known benchmark credit dataset, we simulated the kinds of changes real-world data undergoes over time, including shifts in applicant profiles, in how those characteristics relate to creditworthiness, and general data noise. After retraining the model on this altered data, we compared its explanations to the original model’s, testing both on an identical set of unseen applicants.

The retrained model’s predictive accuracy remained stable and comparable to the original. However, the explanations for why it made those predictions changed considerably: the features it identified as most important, and how strongly it weighted them, shifted noticeably between the two versions. In short, “Explanations Changed, Predictions Did Not.”

This matters because a model can appear stable and trustworthy on the surface — accurate and consistent — while the reasoning it offers regulators, auditors, and customers becomes unreliable. Our findings suggest that current approaches to monitoring AI drift, which typically focus only on predictive performance, are insufficient. We argue that explainability itself needs to be actively and continuously audited, not treated as a one-off compliance checkbox.

Keywords
Explainable Artificial Intelligence (XAI), Machine Learning Drift, SHapley Additive exPlanations (SHAP), Credit Risk Assessment, Regulatory Compliance

References

Cossu, A., Spinnato, F., Guidotti, R., & Bacciu, D. (2024). Drifting explanations in continual learning. *Neurocomputing, 597*, 127960. https://doi.org/10.1016/j.neucom.2024.127960

Muschalik, M., Fumagalli, F., Hammer, B., & Hüllermeier, E. (2022). Agnostic explanation of model change based on feature importance. *KI - Künstliche Intelligenz, 36*(3–4), 211–224. https://doi.org/10.1007/s13218-022-00766-6

Pelosi, D., Cacciagrano, D., & Piangerelli, M. (2025). Explainability and interpretability in concept and data drift: A systematic literature review. *Algorithms, 18*(7), Article 443. https://doi.org/10.3390/a18070443