A Benchmark Study on the Stability of Interpretability in Credit Scoring

Abstract

Machine learning (ML) models have significantly advanced credit scoring by outperforming traditional statistical approaches in predictive accuracy. However, their inherent opacity poses challenges for regulatory compliance, stakeholder trust, and deployment in high-stakes financial applications. While the importance of interpretability is widely recognized, the stability of interpretability-how consistently models explain their predictions-remains underexplored and lacks standardized evaluation methodologies. This paper introduces a comprehensive benchmarking framework that extends the foundational work of Bart Baesens by integrating both global and local interpretability stability assessments into the model evaluation process. The framework leverages a popular and well know post hoc explanation methods, such as SHAP, and incorporates novel quantitative stability metrics-including the Sequential Rank Agreement (SRA), Coefficient of Variation (CV), and the Stability Measure for Local Interpretability (SMLI) to assess the robustness of feature importance rankings under data perturbations. Empirical validation on both synthetic and real-world credit scoring datasets demonstrates that predictive performance alone is insufficient for model reliability. Our findings contribute to the field of Explainable AI (XAI) by offering a rigorous and reproducible methodology for evaluating model transparency and stability, thereby guiding the development of trustworthy, interpretable, and regulatory-compliant ML systems in financial services.

Authors & Affiliations

Sepúlveda E.[1], V., Baesens B.[1], Verdonck T.[1]
[1]KU Leuven