Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring

Abstract

This study examines the stability of post-hoc explanations when applied to instance-dependent cost-sensitive (IDCS) classifiers in credit scoring. These classifiers optimize for loan-specific misclassification costs in order to enhance cost-efficiency compared to traditional models. However, despite increasing regulatory demands for transparency, the impact of IDCS classifiers on explanation stability remains unexplored. Given that IDCS models alter learning dynamics in ways that fundamentally impact interpretability methods, this is a critical gap that our study addresses.

We assess the stability of SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) across IDCS and traditional classifiers using four public credit scoring datasets. First, we benchmark predictive and cost-sensitive performance, introducing relative Average Expected Cost (relAEC) as a new cross-dataset comparable metric. We then evaluate explanation stability through the coefficient of variation and sequential rank agreement of feature importance rankings, incorporating class imbalance effects via controlled resampling.

Our findings reveal a critical trade-off: while IDCS classifiers improve cost-efficiency, they significantly compromise the stability of SHAP and LIME explanations, especially in imbalanced datasets. This instability raises concerns about the reliability of post-hoc interpretability techniques when applied to IDCS models in credit scoring—an area where consistent explanations are essential for regulatory compliance and stakeholder trust. Our results underscore the need for either more robust explainability methods tailored to IDCS classifiers or IDCS models that better balance cost-efficiency and interpretability.

Authors & Affiliations

Matteo Ballegeer[1,2], Dr. Matthias Bogaert[1,2], Prof. Dr. Dries Benoit[1,2]
[1]Ghent University, Ghent, Belgium. [2]FlandersMake@UGent---Corelab CVAMO, Ghent, Belgium