Abstract:

The increasing adoption of machine learning in credit scoring has improved predictive accuracy but simultaneously raised significant concerns regarding algorithmic fairness. Despite growing interest in fairness assessment, the impact of class imbalance - a prevalent issue in credit scoring - on fairness metrics remains largely unexplored. This study fills this gap by investigating how varying levels of class imbalance affect fairness metrics within credit scoring. Using multiple publicly available credit scoring datasets and widely used machine learning algorithms, we evaluate both threshold-dependent and threshold-independent fairness metrics. Our results demonstrate that fairness metrics relying on specific classification thresholds often underestimate disparities in highly imbalanced scenarios. Conversely, metrics evaluating the entire range of predicted outcomes provide more consistent and stable fairness assessments across different levels of class imbalance. Additionally, we introduce a weighting approach to emphasise disparities in rare-event contexts. Our findings emphasise the necessity for financial institutions to carefully choose and adapt fairness metrics according to their datasets' class imbalance level, offering practical insights for developing fairer and more reliable credit scoring models.