Examining the Impact of Bias Mitigation on Credit Scoring: Balancing Performance, Fairness, and Explainability

Abstract

Machine Learning models are increasingly used in high-stakes decision-making, necessitating that algorithmic decisioning systems are not only accurate but also explainable and guarantee fair treatment of affected subjects. Accordingly, regulatory frameworks, including the EU's AI Act, the GDPR framework, and the Basel Accord require financial institutions to achieve explainability and fairness targets.

Counterfactual explanations have emerged as a promising approach to address these requirements. They offer intuitive, user-centered justifications by illustrating how small input changes can lead to different model outcomes. Likewise, bias mitigation techniques improve group fairness and help financial institutions ensure compliance in lending operations. To that end, bias mitigation techniques alter a classifier's training data, the classifier itself, or the probability predictions it generates. Impacting the prediction model development process and potentially the underlying data, bias mitigation may substantially impact explanation methods, particularly counterfactual explanations. However, the effect of bias mitigation on explainability remains unclear.

Moreover, credit scoring systems must maintain high predictive performance, which some bias mitigation techniques may compromise. Given the complex interplay between accuracy, explainability, and fairness, this paper examines their dependency structure. We aim to clarify interactions and synergies among learning algorithms, bias mitigation techniques, and explanation methods to help decision-makers navigate the comprehensive option space and ultimately design powerful and compliant credit scoring systems.

Our empirical analysis embraces multiple bias mitigation techniques, spanning pre-, in-, and post-processing methods, established learning algorithms, and multiple real-world data sets with diverse characteristics. Unsurprisingly, no single bias mitigation method excels across all evaluation dimensions. Some methods improve fairness at the cost of accuracy or explanation quality, while others benefit predictive performance or explanation quality but may compromise model fairness. To further validate and generalize these findings, we plan to extend our evaluation to additional datasets and conduct robustness analyses to assess the stability of these trade-offs under varying conditions. Overall, our results highlight the need for joint evaluation frameworks that consider predictive performance, fairness, and explainability as interdependent components in trustworthy AI systems.

Authors & Affiliations

Shih-Chi Ma[1,2], Prof. Dr. Benjamin Fabian[2], Prof. Dr. Stefan Lessmann[1]
[1]Humboldt University of Berlin, Berlin, Germany. [2]Technical University of Applied Sciences Wildau, Wildau, Germany