Fairness by Design: Transparent Credit Decisions through Semi-Structured Neural Models

Abstract

Today, high-stakes decisions are routinely automated by financial institutions, making it increasingly important to assess not only how fair or morally sound these decisions are, but also how transparently institutions can explain them to affected customers. Acknowledging this importance, regulatory bodies are placing growing demands on machine learning-based credit scoring models to ensure fairness and interpretability, though without prescribing specific methodologies.

A substantial body of research has emerged in recent years to address fairness in algorithmic decision-making, with methods typically classified as pre-processing, in-processing, or post-processing, depending on the stage at which fairness interventions are applied. While many existing approaches recognise the trade-off between predictive accuracy and fairness, pre- and post-processing methods do not guarantee optimal performance, as they do not directly optimise this trade-off during training, limiting their effectiveness.

In this work, we propose an in-processing method that simultaneously addresses three key requirements for a compliant, high-performing model: interpretability of covariate effects, flexibility in the modelling approach, and fairness.

Our method builds on the semi-structured distributional regression framework, integrating deep neural networks with an orthogonalisation step to disentangle structured, interpretable effects from complex, high-level interactions among covariates or unstructured data (e.g. text and images). This enables transparent modelling of main effects while retaining the ability to capture high-level interactions that traditional interpretable models typically miss. For example, our approach can clearly show how individual factors, such as income or employment status, directly influence credit decisions while separately handling how changes in income affect default risk differently depending on an individual's age group or occupation.

To incorporate fairness into our approach, we formulate a multi-objective loss function that explicitly balances predictive performance and fairness criteria. The optimisation is guided by various fairness definitions, including group fairness, individual fairness, and performance consistency. Our findings demonstrate the promise of semi-structured approaches for achieving equitable, interpretable, and high-performing credit risk assessments.

Authors & Affiliations

Dr Victor Medina-Olivares[1], Prof. Stefan Lessmann[2], Prof. Jonathan Crook[1]
[1]The University of Edinburgh, Edinburgh, United Kingdom. [2]Humboldt University of Berlin, Berlin, Germany