Mitigating Bias in Ai: Challenges, Ethical Implications, and Fairness-Driven Solutions

Keithu soldra
Lovely Professional University, Bangalore, India

Abstract:

The use of Artificial Intelligence (AI) in hiring provides scalability and efficiency but poses serious ethical issues because of the risk of reinforcing past biases and uninterpretability. This research explores the twofold challenge of reducing gender bias in AI-driven hiring systems and making decision-making transparent. A simulated hiring scenario was created with synthetically generated data infused with gender-based bias. A baseline logistic regression model was trained to mimic standard practices, against which a debiased model trained with the ExponentiatedGradient algorithm under the Fairlearn framework was compared, having been optimized under the Equalized Odds fairness constraint.

In order to improve the interpretability of models and build confidence among stakeholders, the research included Explainable AI (XAI) methods—LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations)—for interpreting and visualizing feature importance both prior and post debiasing. Fairness measures with evaluation metrics like Demographic Parity Difference and Equalized Odds Difference showed significant improvements: Equalized Odds Difference decreased more than 50%, and demographic parity also improved. The trade-off was that model accuracy slightly decreased, indicating the compromise between fairness and performance.

Interpretability analysis revealed that debiasing not only decreased discriminatory feature influence but also shifted model decision logic away from biased indicators, as evidenced by SHAP and LIME results. The results highlight the importance of merging algorithmic fairness interventions with strong explainability tools to provide accountable and equitable AI deployments.

This study provides an replicable and scalable model of ethical AI application in recruitment to show how transparency and fairness could be combined with automated decision systems. The model has wider potential applications in other high-stakes areas like healthcare and finance where fairness and responsibility are paramount.

Keywords:

Gender Bias, Fairness in AI, Equalized Odds, Explainable AI, XAI, SHAP, LIME, Debiasing Algorithms, Ethical AI Recruitment.