Interactive Human-Centric Bias Mitigation

Inge Vejsbjerg, Elizabeth M. Daly, Rahul Nair, Svetoslav Nizhnichenkov

IBM Research, Ireland
ingevejs@ie.ibm.com, elizabeth.daly@ie.ibm.com, rahul.nair@ie.ibm.com, svetoslav.nizhnichenkov@ibm.com

Abstract

Bias mitigation algorithms differ in their definition of bias and how they achieve that objective. Bias mitigation algorithms impact different cohorts differently and allowing end-users and data scientists to understand the impact of these differences in order to make informed choices is a relatively unexplored domain. This demonstration presents an interactive bias mitigation pipeline that allows users to understand the cohorts impacted by their algorithm choice and provide feedback in order to provide a bias mitigated pipeline that most aligns with their goals.

Introduction

Advances in the area of bias mitigation have led to many algorithms and toolkits being made available to data scientists. However, a recent study of data scientists using these toolkits stated that one of the most important features was the importance of the "ability to adapt to a context-specific use case and data" (Lee and Singh 2021). This is a challenging problem when faced with the fact that fairness is problematic to prescriptively define given it can be multi-dimensional and context-dependent (Grgic-Hlaca et al. 2018). Studies of public perception of algorithmic decision-making have repeatedly shown gaps between the consumer/users' perspectives on fairness and the mathematical definitions (Lee and Baykal 2017; Srivastava, Heidari, and Krause 2019; Woodruff et al. 2018). As a result, there is a significant risk that a data scientist, when given the requirement to ensure bias mitigation is part of an AI solution, may select a strategy without understanding the differences or the impact of their decision. The goal of this demo is to provide a human-centric approach to allow a user understand the impact of applying different bias mitigation algorithms, and guide them through some options and alternatives to elicit what fairness means to them in the context of their data and use case. In this way we aim to support the user to select a bias mitigation solution most suited to their needs.

System Overview

The overall pipeline is shown in figure 1. **Problem Specification:** Our solution allows a user upload a dataset to be analyzed. The user specifies the mitigation strategies to be considered, the favorable outcome, the protected attributes and optionally the privileged cohorts. If the privileged cohort is not provided then it is automatically inferred using MDSS (Zhang and Neill 2017). The user can also specify the optimization function to be used when performing the AutoML pipeline search. The solution optimizes for both balanced accuracy and disparate impact by default, however this is something that can be adjusted.

Model Alternative Generation: We leverage a combination of open source resources in order to generate model

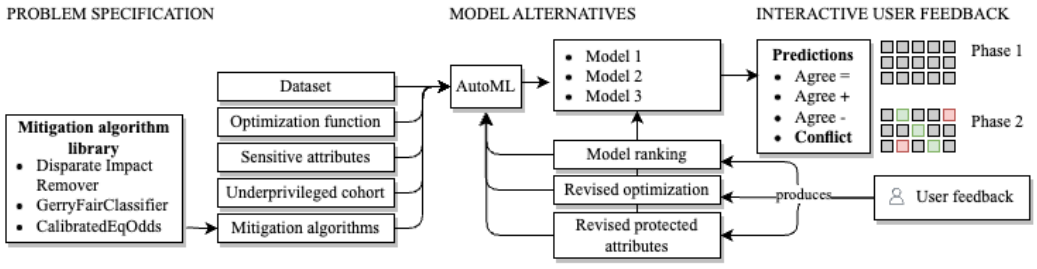
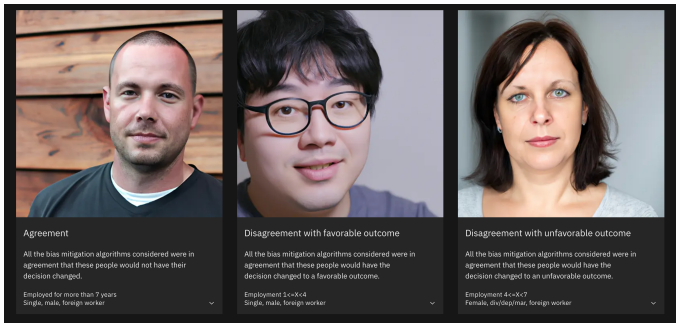
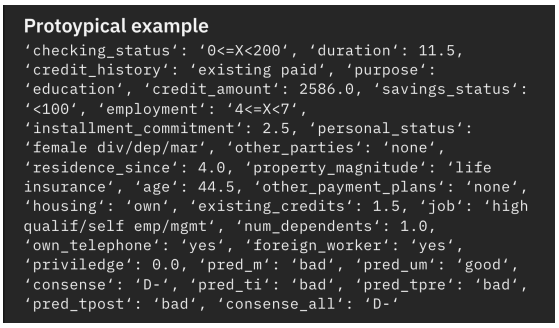


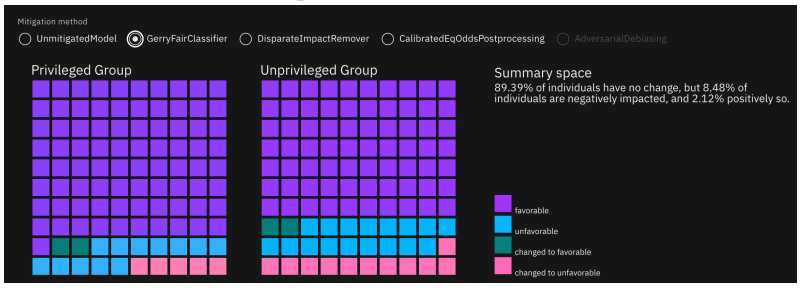
Figure 1: Overview of pipeline



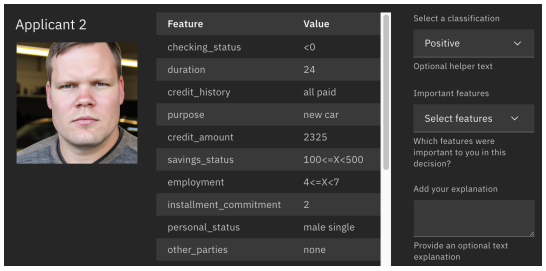
(a) Impacted Cohorts



(b) Prototypical Person



(c) Model Alternatives



(d) Feedback

Figure 2: Bias mitigation visuals

variations. The AIX360 toolkit was used to provide the bias mitigation algorithms (Bellamy et al. 2018). We use the LALE opensource framework to search over possible ML pipelines taking into account different bias mitigation strategies optimizing for the input objective (Hirzel et al. 2022). By default the user can select accuracy however the framework also supports multi-objective optimization. As a result, the solution allows the user to refine the objective metric if the resulting solutions do not align with their definition of fairness. Our solution generates a single optimized ML pipeline for each user-selected possible algorithm. The resulting output is then separated into four cohorts where **all algorithms agree**:

- the entry will not have their label adjusted and therefore is **not impacted** by the bias mitigation process.
- the label will be **changed to the positive class** and so will benefit from the algorithm.
- the label will be **changed to the negative class** meaning they could be harmed by bias mitigation.

Finally the cohort where the **algorithms disagree** meaning the choice of algorithm will impact these people differently.

Interactive User Feedback: At each phase the user may give feedback, however one challenge when collecting human feedback is that the burden on the user can impact accuracy due to labeller fatigue (Cakmak, Chao, and Thomaz 2010). Agan et. al. showed that models trained on data collected where users exhibit more automaticity behaviors are more likely to demonstrate bias (Agan et al. 2023). This is due to users engaging their 'system 1' behavior where biases are more likely to emerge compared to when their decisions are more deliberate and engaging their 'system 2'. As a result, any human-in-the-loop solution involving bias mitigation must consider carefully how to present information to end-users in order to have them deliberate on their judgments. In order to encourage users to reflect on their feedback our novel solution has incorporated generative AI to create a synthetic representation of the users represented in the data. In phase 1 the user is shown a prototypical persona of each cohort reflecting where all mitigation algorithms make the same decision. The prototypical features are selected as the most likely attributes to be observed in each impacted cohort and used as prompts to OpenAPI to generate each image representing the prototypical person. If the user feedback shows a level of disagreement above a threshold the system can recommend revising either the metrics they are optimizing for in the model search or they can reconsider which attributes they consider protected and the resulting under privileged class. Assuming the user agrees with the examples contained in the cohorts they can move on to explore the variations between the algorithms. The visualization Figure 2 c) shows the user what proportion of the population are either left unchanged or impacted by the algorithm and finally Figure 2 d) shows where the user can give feedback annotating the expected outcome of representative samples where the algorithms conflict. This information can then be used in two ways: 1) to recommend the algorithm the user's choice most aligns with, or 2) it can be leveraged as an additional input to the optimization function when performing the model pipeline search. One advantage of human-in-the-loop decision making in the context of trust is that the decision makers' choices can be logged and made available for scrutiny or as part of a governance process (Arnold et al. 2019).

Acknowledgements

This work was funded by the European Union's Horizon Europe research and innovation programme under grant agreement no. 101070568.

References

Agan, A. Y.; Davenport, D.; Ludwig, J.; and Mullainathan, S. 2023. Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Technical report, National Bureau of Economic Research.

Arnold, M.; Bellamy, R. K.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K. N.; Olteanu, A.; Piorkowski, D.; et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5): 6–1.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Cakmak, M.; Chao, C.; and Thomaz, A. L. 2010. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, 2(2): 108–118.

Grgic-Hlaca, N.; Redmiles, E. M.; Gummadi, K. P.; and Weller, A. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*, 903–912.

Hirzel, M.; Kate, K.; Ram, P.; Shinnar, A.; and Tsay, J. 2022. Gradual AutoML using Lale. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4794–4795.

Lee, M. K.; and Baykal, S. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, 1035–1048.

Lee, M. S. A.; and Singh, J. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–13.

Srivastava, M.; Heidari, H.; and Krause, A. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2459–2468.

Woodruff, A.; Fox, S. E.; Rousso-Schindler, S.; and Warshaw, J. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–14.

Zhang, Z.; and Neill, D. B. 2017. Identifying Significant Predictive Bias in Classifiers. arXiv:1611.08292.