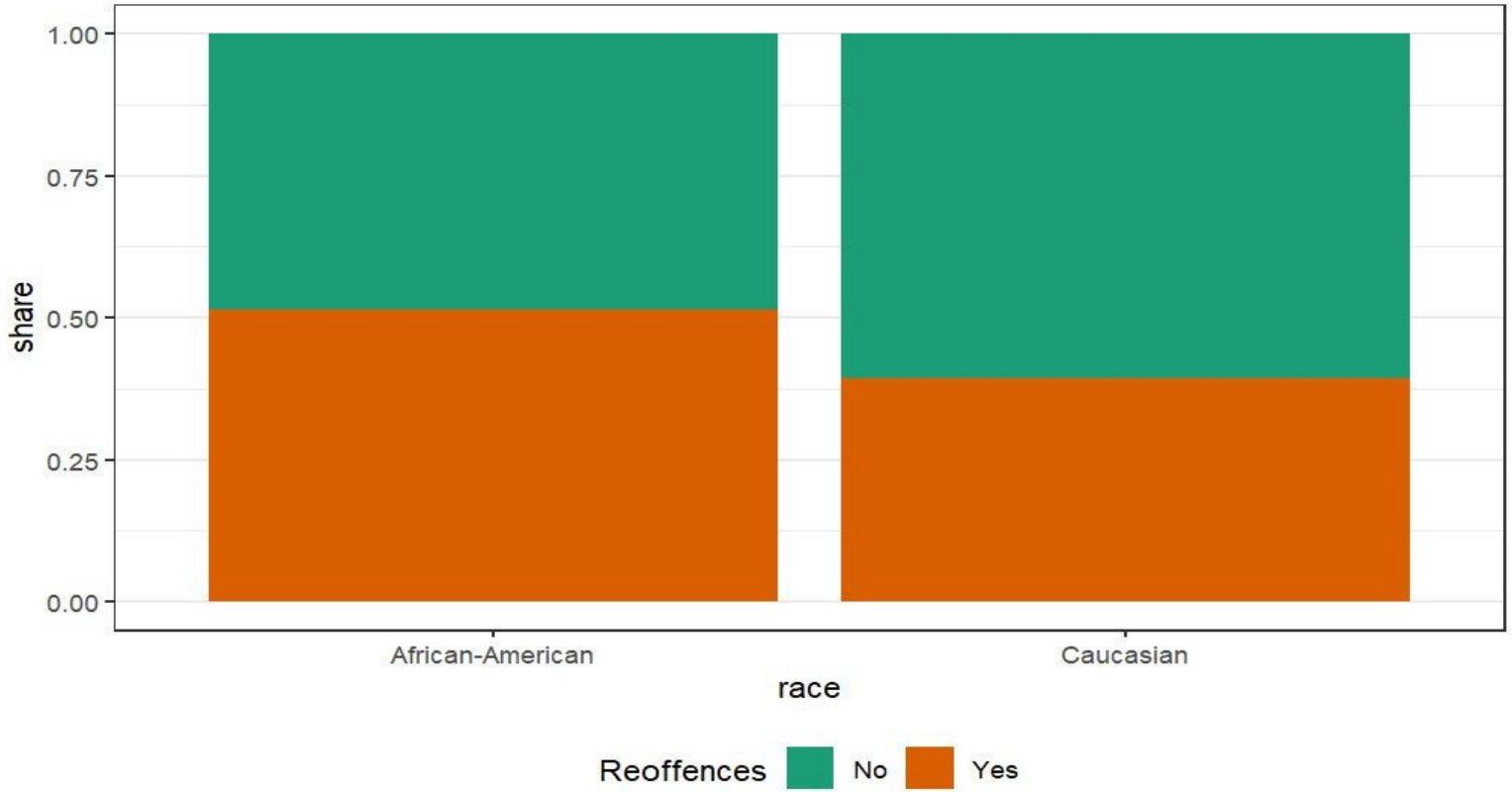
COMPAS dataset

Debiasing a biased world

Alex Matakos, Karen Hovhannisyan, and Theo Blauberg

starting point

- We want the recidivism in the next two years after release from prison.
- Our original model was trained using previous criminal records and some background information of the released inmates.
- We are concentrating on fairness at the group level, this way we can better understand broader context and potential impacts of our decisions and actions.
    - By not addressing group fairness the algorithms will perpetuate and in worst case amplify existing biases in the judicial system

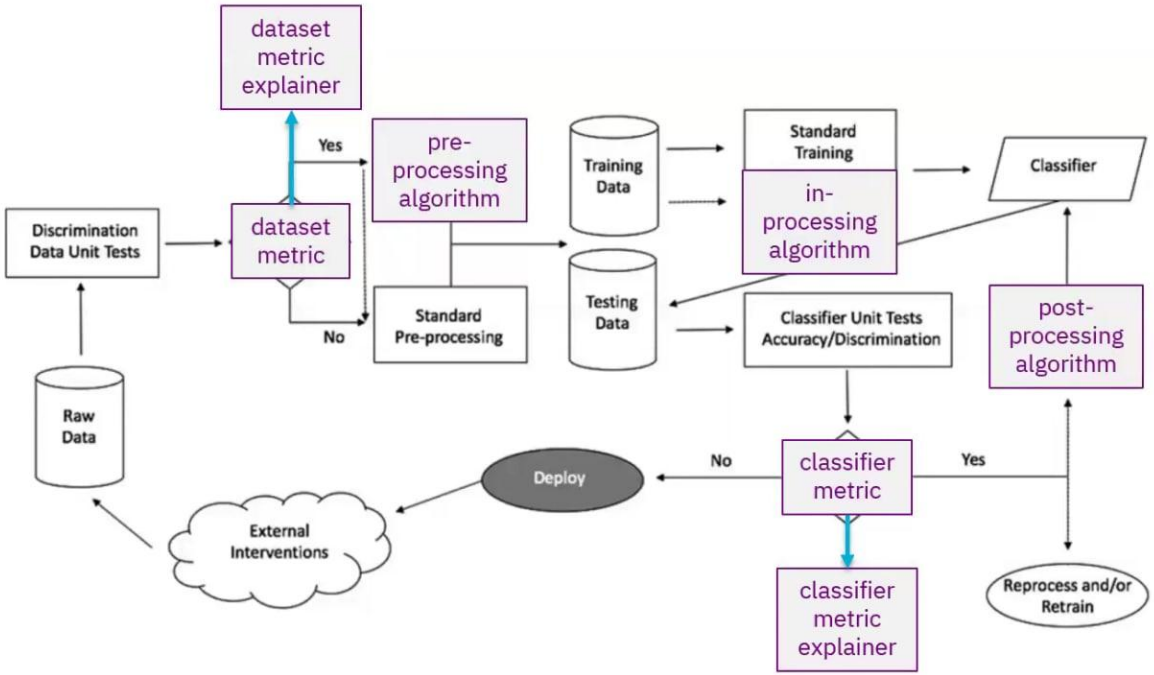


The different outcomes on reoffenses comparing Caucasian and African-Americans. We believe these are partly affected by the biased policing of minorities in the US.

a new model

- We constructed a new model that gained an accuracy of 88.4 % with using only two predictive variables -> the duration of the captivity and the charge degree
- The new model itself was a way to debiase the results. As the accuracy of the model was increased and the statistical parity difference was decreased.
    - There was still bias left in the model so we had to try out some additional debiasing methods.

Metrics, Algorithms, and Explainers



Bias mitigation can happen at three different steps. Preprocessing, inprocessing, or postprocessing. Fairness is a multifaceted, context-dependent social construct that defies simple definition.

debiasing strategies

Our plan is to use in conjunction two different preprocessing techniques and one postprocessing method.

Disparate impact remover

- editing feature values to increase group fairness while preserving within group rank ordering.

Reweighing

- adjusting the importance of different training examples in order to balance the data and improve fairness

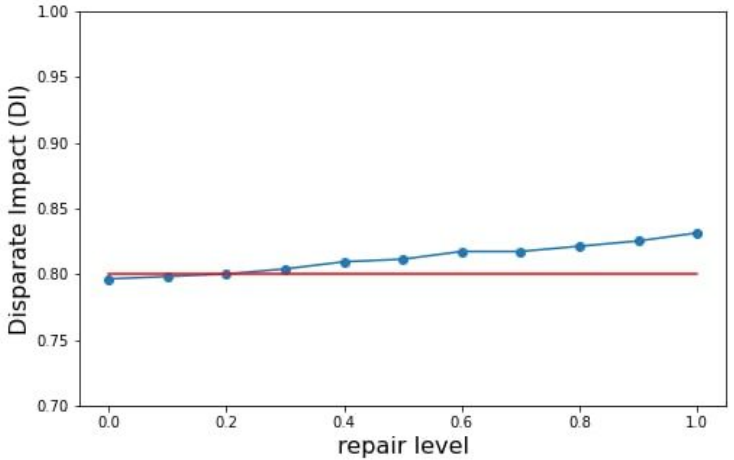
Reject option classification

- Changing the outcomes of the privileged and unprivileged to ensure fairness.

to implement this we will be using AIF360 package by Bellamy et al. (2019)

disparate impact remover

- In the first step we will use the disparate impact remover.
- Method was first used by Feldman et al. (2014)

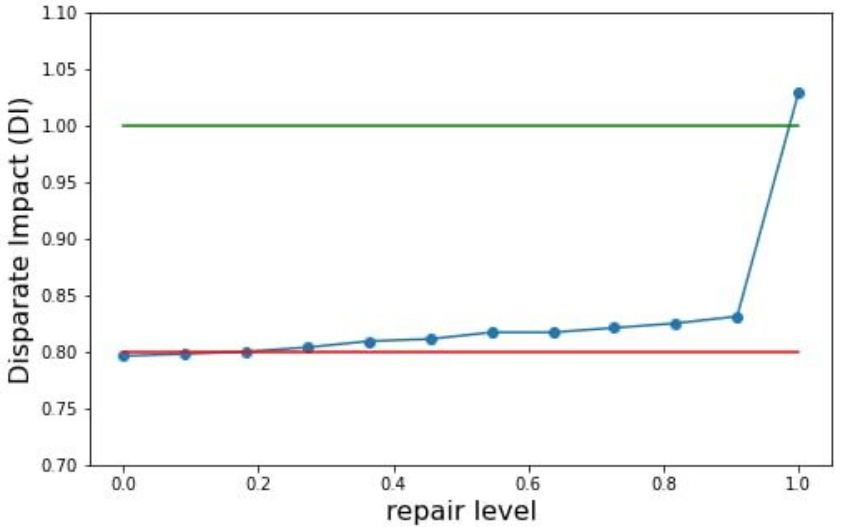


reweighing

- Second step of our debiasing pipeline is the reweighing of the importance of privileged and unprivileged groups
  - this is advantageous in improving the fairness of the predictions. (Kamiran, 2012b)
- Disparate impact was increased after reweighing to 1.02

reject option classification

- The last part of our debiasing pipeline is a reject option classifier
- This intervenes at the last stage when we already have probabilities from the logistic regression
- We find a probability interval where all labels within this are flipped
- Assumes biased decisions occur near decision boundary



Kamiran (2012a)

conclusion

- Having a less biased predictor can bring more justice to the judicial system
- With only reasonable impacts to the prediction accuracy can the predictions be debiased significantly
    - With the disparate impact remover proving to be especially impactful

references

- Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4-1.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268).
- Kamiran, Faisal, Asim Karim, and Xiangliang Zhang. "Decision theory for discrimination-aware classification." 2012 IEEE 12th International Conference on Data Mining. IEEE, 2012.
- Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." Knowledge and information systems 33.1 (2012): 1-33.