Bias in ‘fair’ hiring algorithms: A Fairness Analysis

Yitian Cao[1], Zainab Iftikhar[2], Amy Greenwald[2]
[1]Bryn Mawr College, Department of Computer Science; [2]Brown University, Department of Computer Science

Introduction

- Existing hiring algorithms claim to be "unbiased" but often focus on meeting basic Equal Employment Opportunity Commission (EEOC) requirements.
- Despite meeting standards, these algorithms may still exhibit discriminatory behavior with hiring managers.
- We investigate inherent biases in hiring algorithms, examining the efficacy of mitigating bias by removing gender, race, and class identifiers from the ranking process.
- Two forms of discrimination, disparate treatment and disparate impact, are assessed using the "4/5" rule[1].
- Current approaches to mitigating bias in ranking algorithms: in-processing:
  - in-processing (data cleaning -> ranking) without ML
  - post-processing (data cleaning -> ranking -> evaluation -> reranking) with ML, allowing multiple iterations.

Methodology

- We evaluated four ranking algorithms, a specific focus on Themis-ml[2], a fairness-aware post-processing machine learning algorithm.
- The four training models are evaluated using Themis-ml, employing a protected attribute (gender) and training data from the German Credit Score dataset.
- We first evaluated fairness by comparing the percentage of men and women classified as low-risk for a loan and then calculated utility effectiveness by checking if the AUC value remains the same

Model Architecture

- Models include Baseline (B), Remove Protected Attribute (RPA), Reject-Option Classification (ROC), and Additive Counterfactually Fair Model (ACF) classifiers
  - Baseline (B): classifier trained on all available input variables, including protected attributes.
  - Remove Protected Attribute (RPA): classifier where input variables do not contain protected attributes.
  - Reject-Option Classification (ROC): classifier using the reject-option classification method.
  - Additive Counterfactually Fair Model (ACF): classifier using the additive counterfactually fair method

Risk Evaluation (Pre-Ranking)

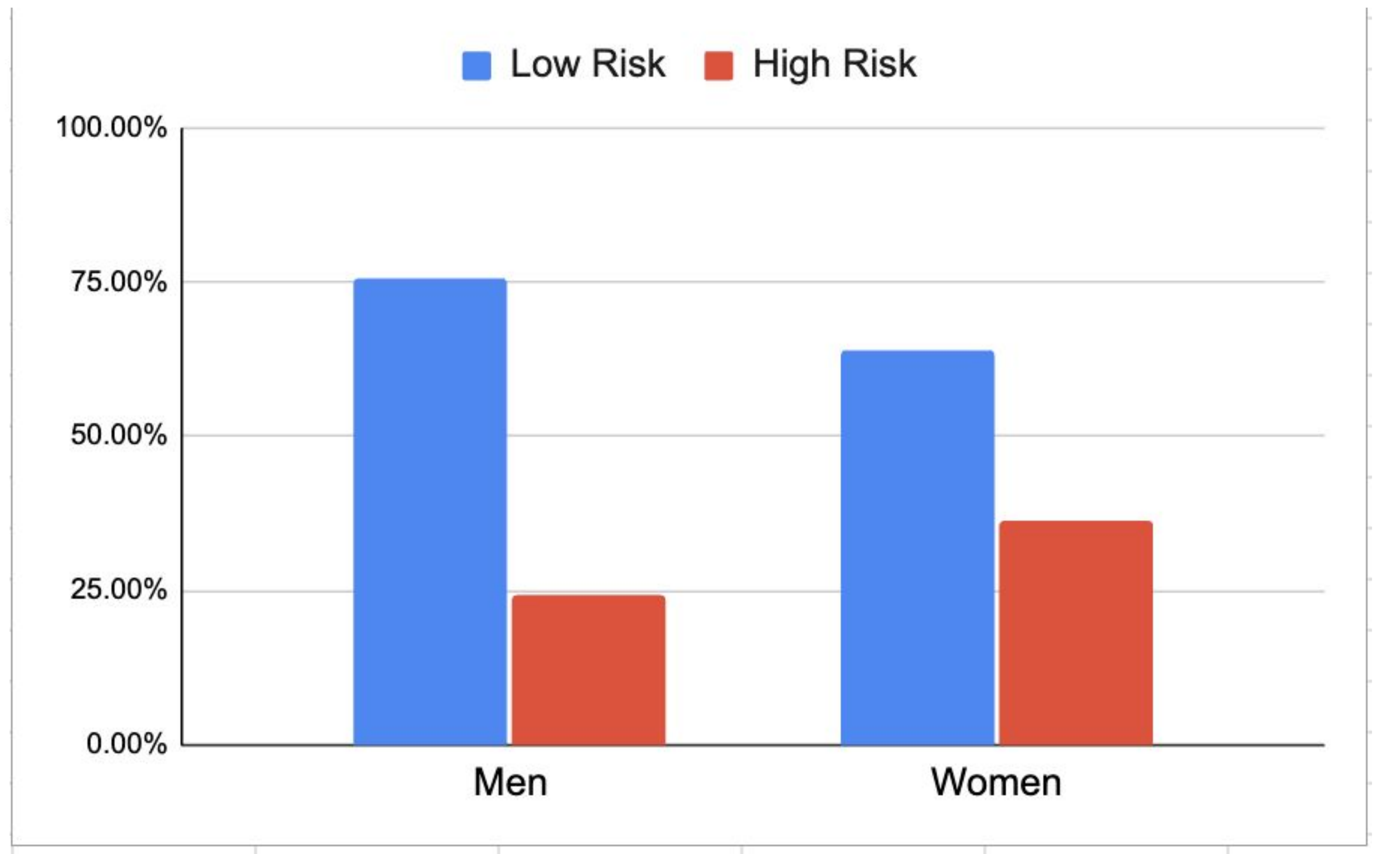

Figure 1: Our results showed that men (unprotected group) are 12% more likely to be labeled as low risk.

Risk Evaluation (Post-Ranking)

Model	Men (High Risk)	Women (High Risk)	AUC
Baseline	25%	35%	62
RPA	25%	35%	62
ROC	**32%**	**33%**	61
ACF	**33%**	**25%**	62

*For Baseline and RPA, there is no noticeable change in distribution between the two gender groups. However, the difference between the two gender groups is significantly decreased by 11% in ACF model. For ROC, surprisingly, women are more likely to be labeled as low risk, and the difference between the two groups is -8%. All four training models maintain the utility AUC value around 62%

Conclusion

- Takeaways:
  - Identifiers related to certain attributes (e.g. gender, race, or class) are not a good indicator of the presence of biases in hiring algorithms. Removing them do not increase the fairness of the ranking result.
- Limitations:
  - Controlled experiment
  - Training dataset was limited
- Future work:
  - Focus on the social and systemic dimensions for ranking algorithms for marginalized groups.
  - Real-life evaluations to achieve better representation for marginalized groups.
  - Multi-modal modeling[3]

References

[1] Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020, January). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 469-481).
[2] Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019, July). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining (pp. 2221-2231).

Acknowledgements

Thank you to Brown exploreCSR program and my direct mentor Zainab Iftikhar.