Evaluating Fairness of Ranking Algorithms

Amy Greenwald, Yitian Cao, Zainab Iftikhar

Bryn Mawr College, Brown University

MOTIVATION

- Investigating how hiring (and ranking algorithms in general) are biased and what are the effective ways to mitigate the bias.
- Experimenting the efficacy of removing gender, race, and class identifiers to generate fair ranking.

BACKGROUND

- Existing hiring algorithms that companies claim to be "unbiased" oftentimes only try to meet the Equal Employment Opportunity Commission (EEOC) basic requirements.
- Even when a hiring algorithm is "good enough" for EEOC standards, its interaction with humans such as hiring managers still encourages discriminatory actions.
- Two assessments of discrimination: [1]
  - *disparate treatment*
  - *disparate impact ("4/5" rule)*
- Two general categories of current approaches to mitigating bias in ranking algorithms:
  - *in-processing: data cleaning -> ranking (normally done WITHOUT machine learning)*
  - *post-processing: data cleaning -> ranking -> evaluating -> reranking (normally done WITH machine learning, and evaluating and reranking could happen multiple times)*

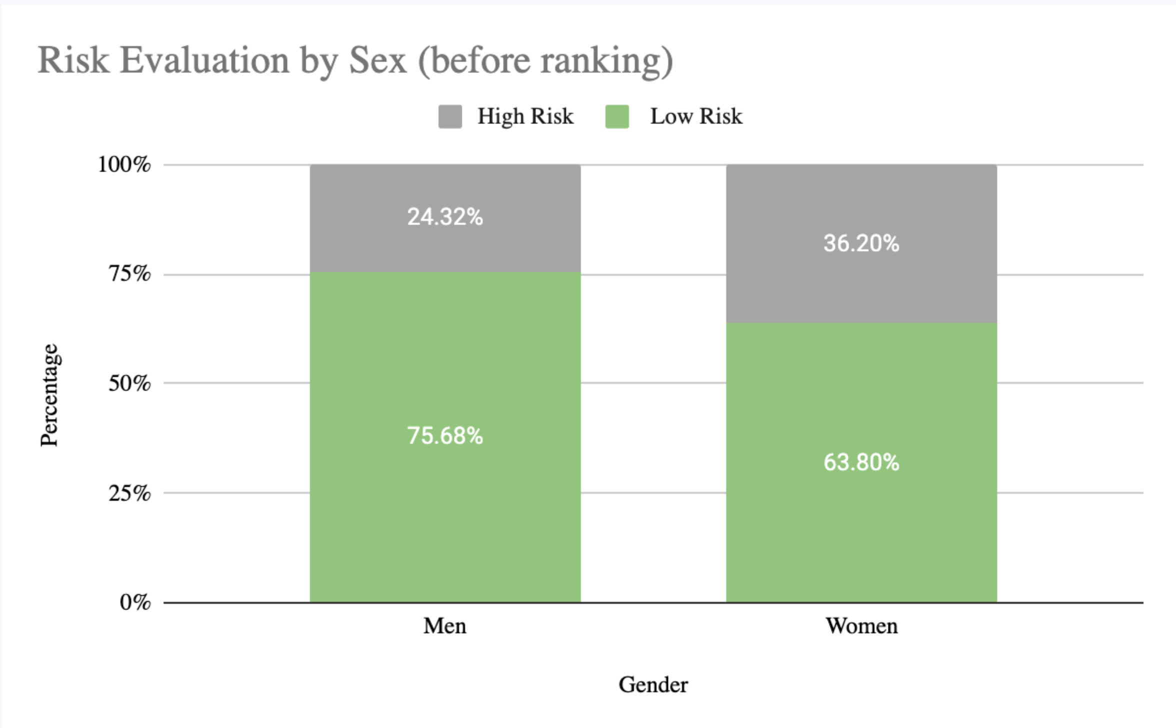
METHODOLOGY

- Understand, then Compare and Contrast various types of ranking algorithms.
- Experiment with a particular algorithm
  - *Themis-ml* [2]
  - *a fairness-aware *post-processing* machine learning algorithm*
- Four Training Models [2] (protected attribute = gender; training data = German Credit Score):
  1. ***Baseline (B):*** *classifier trained on all available input variables, including protected attributes.*
  2. ***Remove Protected Attribute (RPA):*** *classifier where input variables do not contain protected attributes.*
  3. ***Reject-Option Classification (ROC):*** *classifier using the reject-option classification method.*
  4. ***Additive Counterfactually Fair Model (ACF):*** *classifier using the additive counterfactually fair method.*

- Evaluate fairness by comparing the percentage of men and women classified as low-risk for a loan.
- Evaluate utility effectiveness by checking if the AUC value remains the same.

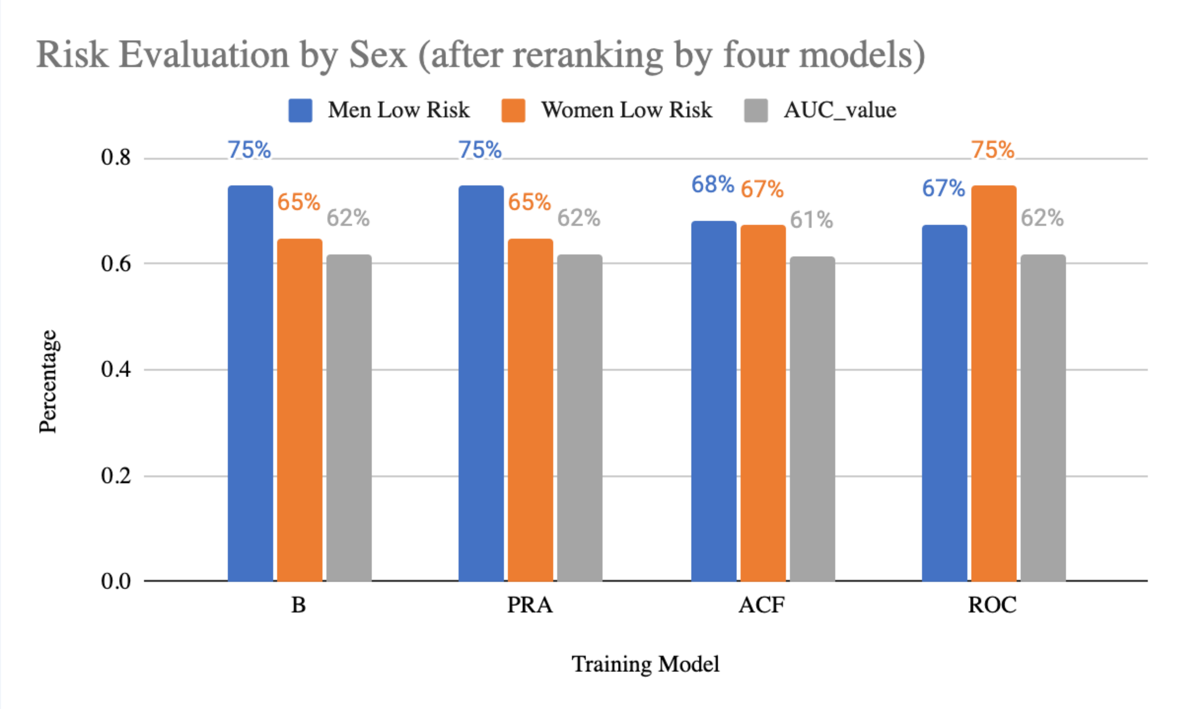
FINDINGS

- Raw Data



* Men (unprotected group) are 12% more likely to be labeled as low risk than Women (protected group).

- Reranking Result



- For PRA and B, there is no noticeable change in distribution between the two gender groups.
- For ACF, the difference between the two gender groups is significantly decreased by 11%.
- 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 & EVALUATION

- Simply removing the identifiers related to certain attributes (e.g. gender, race, or class) can not improve the fairness of the ranking result.
- This is still a simple data set that produces binary classifications. We should deploy real-life evaluation on the algorithms to see if the algorithms can achieve better representation for the marginalized group.
- Future work should also focus on the social and systemic dimensions for ranking or hiring algorithms to be in place.

REFERENCE

1. Manish Raghavan, Solon Barocas, Jon Kleinberg, Karen Levy, "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices", https://dl.acm.org/doi/abs/10.1145/3351095.3372828
2. Sahin Cem Geyik, Stuart Ambler, Krishnaram Kenthapadi, "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search", https://arxiv.org/abs/1905.01989.
3. Tom Sühr, Sophie Hilgard, Himabindu Lakkaraju, "Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring", https://dl.acm.org/doi/abs/10.1145/3461702.3462602

ACKONWLEDGEMENT

Thanks to Brown ExploreCSR program and my mentor Zainab Iftikhar for guiding this project.