Bias Detection And Explainability In AI Models
CIS-2025-19
Summer 2025

Uncovering Bias and Explaining Decisions in a Text-Based Job Screening Model

Name : Rana Saad Ibrahim Mahdy
Zewail City of Science, Technology, and Innovation.

DATE
Jul 06th, 2025

I. Objective

Analyze gender bias in hiring predictions and implement bias mitigation

II. Dataset Overview

- 1500 resume samples with 11 features
- Gender distribution: 49.2% male, 50.8% female
- HiringDecision: 31% Hire (1), 69% Not Hire (0)(**Target Imbalance**)
- Text feature collected from: Age, Gender, EducationLevel, ExperienceYears, InterviewScore, SkillScore, PersonalityScore
-

Sensitive Feature Encoding

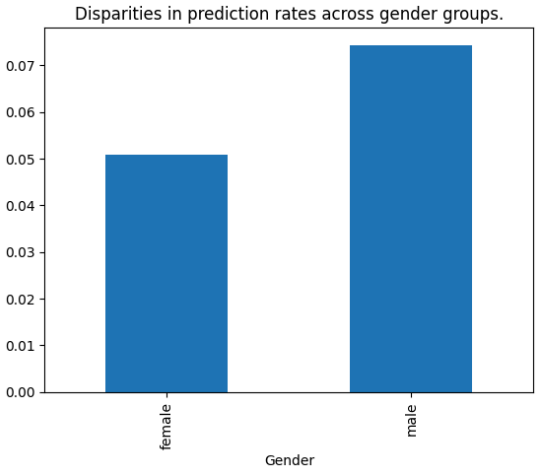
- Gender: Binary (0=Female, 1=Male)
- Text: Explicit gender terms (e.g., "gender 1") to enable bias measurement while testing mitigation

III. Model architecture and performance

Text Data → TF-IDF Vectorization → Logistic Regression → Binary Prediction]
Training split: 80% male, 20% female (intentional imbalance)

Performance:

- train accuracy: 73 %
- test accuracy: 68.6 %



IV. Fairness analysis

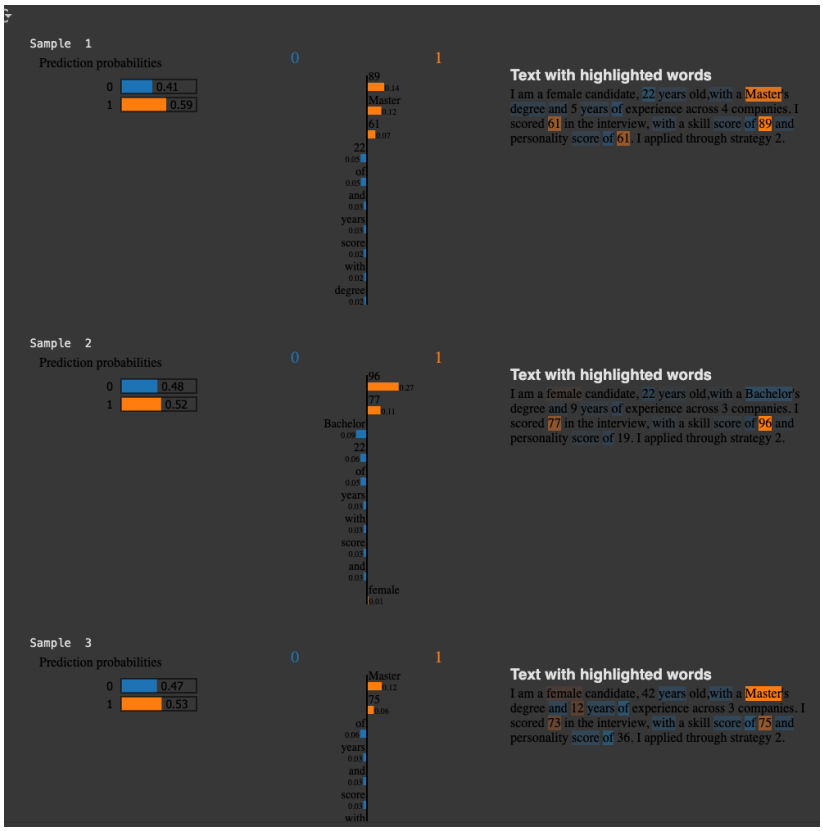
- demographic parity difference: 0.024
- equal opportunity difference: 0.037
- average odds difference: 0.022

V. Explainability results and discussion

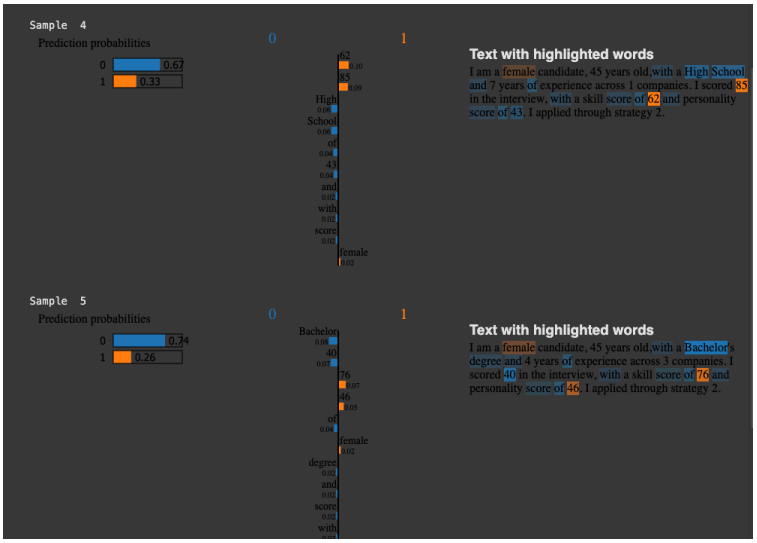
5 model predictions (3 Hire, 2 No-Hire)

- Samples 1,2,3 (**Hire**)

The LIME explanations for the three samples show that high skill and interview scores (“score of 96”, “score of 89”) and higher education levels (“Master’s”, “Bachelor’s”) were the strongest positive contributors to Hire predictions. Gender-indicative words like “female” appeared with negligible weights, suggesting limited direct gender influence in these cases. The model appears to rely primarily on performance-related features rather than demographic terms. This highlights that although gender terms are present in text, their impact on these predictions was minimal compared to qualifications and scores.



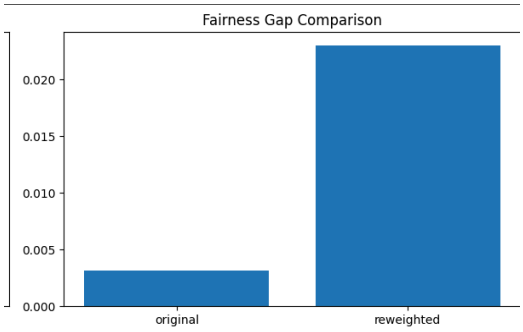
- Samples 4,5 (**No-Hire**)



The most influential negative factors were lower education levels ("High School" in sample 4, "Bachelor's" in sample 5) and lower interview and skill scores. Although the term "female" appeared in both texts, its contribution was minimal compared to performance-related features. This suggests the model prioritized qualifications and scores more than gender terms when predicting non-hiring decisions in these examples.

VI. Mitigation results and tradeoffs

Metric	Original	Mitigated(Reweighing)
Accuracy	68.77%	68.11%
Male Hire Rate	8.78%	14.86%
Female Hire Rate	7.84%	11.11%



Accuracy dropped slightly, which is an acceptable trade-off in fairness-aware machine learning. Male and Female hire rates both increased under the reweighed model. However, the gap between male and female hire rates increased, suggesting that reweighing alone did not reduce the disparity and, in fact, slightly exaggerated the difference. This highlights a key challenge in fair ML, which is that mitigation strategies can involve trade-offs, and their effectiveness depends on the nature of the bias.