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

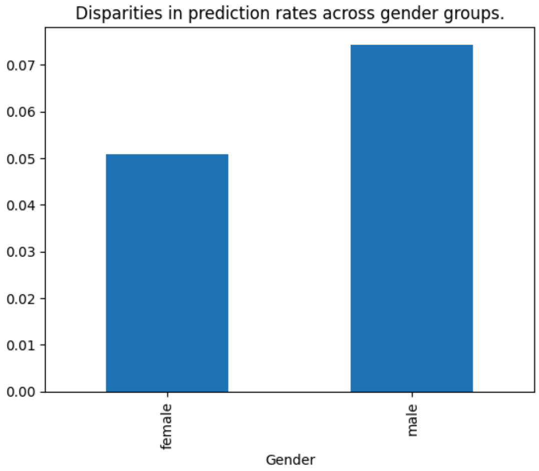
Text Data → TF-IDF Vectorization → Logistic Regression → Binary Prediction]
Training split: 80% male, 20% female (intentional imbalance)

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

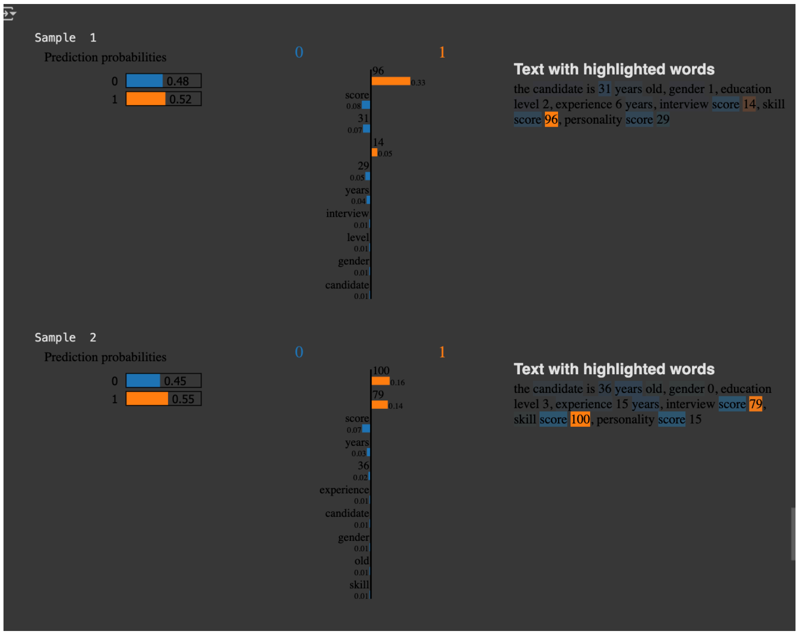


V. Explainability results and discussion

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

- Samples 1 and 2 (**Hire**)

We can see that the strongest positive influence on predicting a hire. Other features, such as the interview Score and Age had smaller effects, while gender had a minimal impact. This suggests that the model focuses more on technical qualifications rather than demographic factors. However, certain numerical tokens (like high scores) heavily influence decisions, which may reflect bias toward extremes.



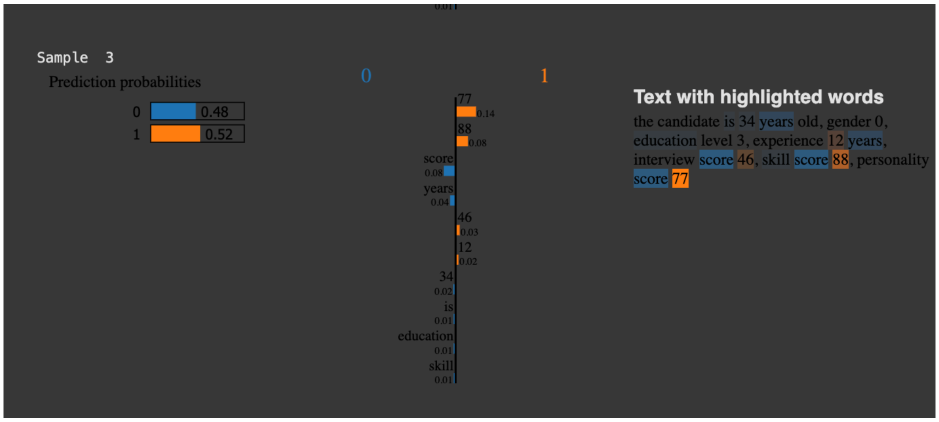
Text with highlighted words

the candidate is 31 years old, gender 1, education level 2, experience 6 years, interview score 14, skill score 96, personality score 29

Text with highlighted words

the candidate is 36 years old, gender 0, education level 3, experience 15 years, interview score 79, skill score 100, personality score 15

- Samples 3 (**Hire):** The model hired this candidate mainly due to strong soft and technical skills, despite moderate interview performance. Again, gender had little influence, supporting consistency in the model's focus on qualifications.

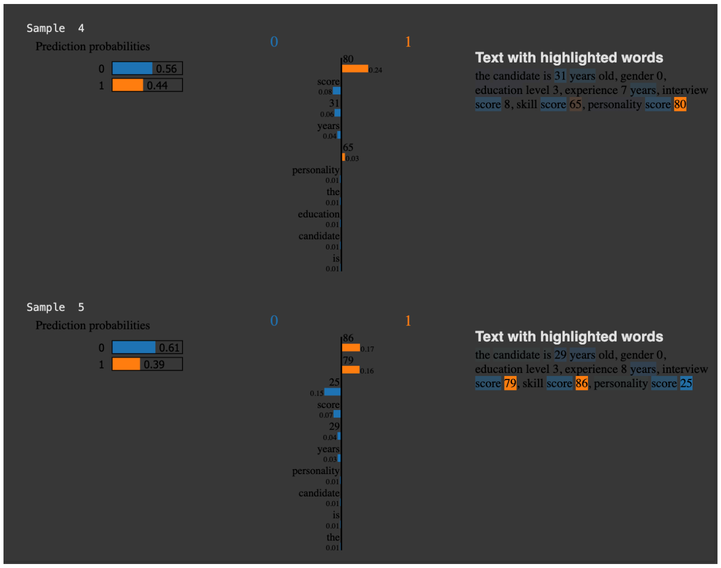


Text with highlighted words

the candidate is 34 years old, gender 0, education level 3, experience 12 years, interview score 46, skill score 88, personality score 77

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

Despite strong personality (80) or skill (86), these candidates were not hired. LIME shows that positive traits like high personality and skill scores contributed toward hiring,but they were outweighed by negative influences like lower interview score in sample 4 or personality score in Sample 5.



Sample 4

Prediction probabilities

Text with highlighted words

the candidate is 31 years old, gender 0, education level 3, experience 7 years, interview score 8, skill score 65, personality score 80

Sample 5

Prediction probabilities

Text with highlighted words

the candidate is 29 years old, gender 0, education level 3, experience 8 years, interview score 79, skill score 86, personality score 25

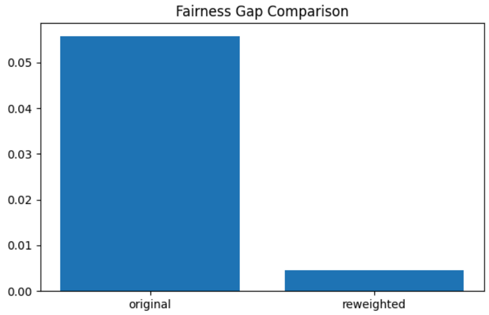
VI. Mitigation results and tradeoffs

Metric	Original	Mitigated(Reweighing)	Improvement
Accuracy	0.675	0.720	+6.7%
Male Hire Rate	16.9%	20.9%	+4.0%
Female Hire Rate	11.3%	20.5%	+9.2%
Fairness Gap	5.6%	0.5%	91% reduction

Accuracy improved while dramatically reducing bias

Still slight male preference (0.5% gap)

Suggested next step: Remove explicit gender markers



Fairness Gap Comparison
original
reweighted