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

Bassel Shaheen

July 7, 2025

1. Introduction

AI is transforming hiring—but also risks amplifying social biases. Even when gender is excluded from training, correlated features may leak bias. This project addresses:

- Detecting gender bias
- Explaining model decisions
- Mitigating unfairness

Our work is rooted in fairness research [1, 2] and reflects real-world stakes in responsible AI.

2. Dataset Description

We analyzed **1,501** job candidates with 11 features. The target variable:

`HiringDecision`: 1 = Hire, 0 = No Hire

Sensitive attribute:

- `Gender` (0 = Female, 1 = Male)

Sample rows:

Age	Gender	EduLvl	ExpYrs	Distance	Hire
26	1	2	0	26.78	1
39	1	4	12	25.86	1
48	0	2	3	9.92	0

Plots showed:

- Imbalanced gender distribution
- Higher hiring rates for males

3. Model and Methods

We implemented:

- **Model:** Logistic Regression (scikit-learn)

- **Split:** 70/30, stratified by gender

Gender was excluded from training features but retained for fairness checks.

Baseline performance:

- Accuracy ≈ 78%
- Balanced precision/recall

4. Fairness Analysis

Fairness metrics measured:

- **Demographic Parity (DP) difference:** ≈ 0.17
- **Equal Opportunity (EO) difference:** ≈ 0.14
- **Average Odds Difference (AOD):** ≈ 0.15

Interpretation: Males were more likely to be predicted as hires, indicating bias.

5. Explainability

Using SHAP, top predictive drivers were:

- SkillScore
- InterviewScore
- PersonalityScore

Key insight: PersonalityScore showed partial gender correlation—posing a proxy risk for bias.

6. Bias Mitigation

We applied **reweighing**:

- Assigned inverse weights based on gender proportions
- Retrained the model using weighted samples

Post-mitigation metrics:

- DP reduced to ≈ 0.08
- EO reduced to ≈ 0.06
- AOD reduced to ≈ 0.07
- Accuracy dipped by ≈ 2–3%

Tradeoff: Slight performance drop for significant fairness gain.

7. Task Alignment

Challenge requirements achieved:

- Binary classifier trained for hiring
- Fairness metrics calculated (DP, EO, AOD)
- SHAP explainability performed
- Bias mitigation applied
- Pre/post mitigation comparison documented

Conclusion: Excluding gender isn't enough. Fairness metrics and explainability are essential tools for surfacing hidden biases. Mitigation strategies like reweighing are crucial for deploying fair, responsible AI in hiring.

Code Repository: Github Repository

References

[1] S. Barocas and A. Selbst, *Big Data's Disparate Impact*. California Law Review, 2016.

[2] N. Mehrabi et al., *A Survey on Bias and Fairness in Machine Learning*. ACM Computing Surveys, 2021.