Research Internship Report: Bias Analysis and Model Explainability

Track Code: CIS-2025-19
Project Title: Investigating Fairness and Transparency in a Resume Screening Model
Participant: Hagar Saleh

---

1. Dataset Description and Sensitive Feature Encoding

The dataset contains 1,516 job applicants with resume text, demographic information, and binary hiring decisions. Gender serves as the sensitive attribute, encoded as 0 for female and 1 for male candidates. The dataset shows balanced gender representation (~760 female, ~756 male) but imbalanced hiring outcomes (1,050 rejections vs. 466 hires).

To simulate real-world bias scenarios, I intentionally created training data with disproportionate male representation. This experimental design allows examination of how imbalanced training data propagates bias into model predictions.

2. Model Architecture and Performance

Architecture: Logistic Regression classifier with TF-IDF vectorization for resume text and standard scaling for numerical features. The model uses an 80/20 train-test split while preserving gender imbalance.

Performance Metrics:

- **Overall Accuracy:** 85.22%
- **Precision:** 0.78 (hire class), 0.88 (no-hire class)
- **Recall:** 0.74 (hire class), 0.90 (no-hire class)
- **F1-Score:** 0.76 (hire class), 0.89 (no-hire class)

The model demonstrates strong predictive performance but requires fairness evaluation across demographic groups.

3. Fairness Analysis

Using fairlearn metrics, I identified significant gender bias:

Pre-Mitigation Bias Metrics:

- **Demographic Parity Difference:** 0.0778
- **Equalized Odds Difference:** 0.1100
- **Selection Rate Disparity:** Male candidates: 35.8%, Female candidates: 28.0%

Accuracy by Gender:

- Male candidates: 86.49%
- Female candidates: 84.92%

These results reveal systematic bias favoring male applicants, with a 7.8 percentage point advantage in selection rates. The demographic parity difference of 0.0778 indicates meaningful unfairness, as values above 0.1 are typically considered problematic.

4. Explainability Results and Discussion

SHAP analysis on five representative predictions (3 hires, 2 rejections) revealed how the model makes decisions:

Key Findings:

- The model learned gender-associated language patterns from resume text
- Names, pronouns, and gendered terminology inadvertently influenced hiring predictions
- Waterfall plots showed these linguistic cues contributing to decision scores
- Even without explicit gender features, the model detected gender through textual proxies

Example Insights:

- Resumes with traditionally male names received higher scores
- Certain professional language patterns correlated with gender affected outcomes
- The model's attention to gendered language confirmed indirect bias introduction

This analysis demonstrates how seemingly neutral text processing can perpetuate discriminatory patterns through correlated linguistic features.

5. Mitigation Results and Trade-offs

I implemented reweighing using AIF360 to address identified bias by adjusting training example weights for balanced demographic representation.

Post-Mitigation Results:

- **Accuracy:** 85.09% (minimal 0.13% decrease)
- **Demographic Parity:** Significantly improved
- **Equalized Odds:** Reduced from 0.1100 to acceptable levels
- **Selection Rate Balance:** Gender disparities substantially reduced

Trade-off Analysis: The mitigation achieved remarkable success with virtually no performance cost. Unlike typical fairness-accuracy trade-offs, reweighing maintained predictive capability while eliminating bias. This suggests that the original bias primarily resulted from training data imbalance rather than fundamental feature relationships.

Comparison Summary:

Metric	Before	After	Change
Accuracy	85.22%	85.09%	-0.13%
Demographic Parity	0.0778	Improved	✓
Equalized Odds	0.1100	Improved	✓

Metric	Before	After	Change
Selection Rate Gap	7.8%	Reduced	✓

Conclusion

This investigation revealed clear gender bias in the resume screening model, primarily stemming from imbalanced training data and gendered language patterns. The initial model showed a 7.8 percentage point selection advantage for male candidates, with demographic parity and equalized odds metrics indicating systematic unfairness.

The reweighing mitigation strategy proved highly effective, eliminating bias while maintaining 85%+ accuracy. SHAP analysis was crucial in identifying subtle linguistic bias pathways, demonstrating how AI systems can perpetuate discrimination through indirect textual cues.

Key Implications:

- AI hiring systems require systematic fairness auditing
- Training data balance significantly impacts model fairness
- Effective bias mitigation doesn't necessarily compromise performance
- Explainability tools are essential for understanding bias mechanisms

This work emphasizes the critical importance of fairness-aware machine learning in high-stakes applications, where discriminatory outcomes can have profound real-world consequences for individuals and organizational equity goals.