Bias Detection and Explainability in AI Job Screening Models

Executive Summary Report

1. Dataset and Model Overview

Dataset: 1,500 job candidate records with 10 structured features including Age, Gender (sensitive attribute), Education, Experience, and performance scores (Interview, Skill, Personality). Training data was intentionally imbalanced (67% male, 33% female) to simulate real-world bias.

Model: Logistic Regression with four variants-Baseline, Standardized, Class-Balanced, and Reweighing approaches for bias mitigation.

2. Performance Results

Model	Accuracy	Precision	Recall
Baseline	86.3%	81.1%	71.2%
Standardized	86.7%	82.0%	71.8%
Reweighing	84.4%	72.0%	81.5%

Key Finding: Standardized model achieved the highest accuracy, while reweighing improved fairness with minimal accuracy loss.

3. Bias Detection Results

Fairness Metrics Analysis

Metric	Baseline Model	Reweighing Model
Demographic Parity	Female: 25.5%, Male: 28.7%	Female: 35.0%, Male: 35.0%
Equal Opportunity	Female: 70.3%, Male: 72.0%	Female: 80.0%, Male: 81.0%
Average Odds Difference	0.012	0.027

Critical Findings:

- **3.2% gender gap** in hiring predictions detected in baseline model
- **Complete elimination** of demographic parity gap through reweighing
- **Successful bias mitigation** with acceptable fairness metrics
-

4. Explainability Analysis

Feature Importance (SHAP Analysis)

1. **SkillScore** (34.2%) - Primary decision driver
2. **PersonalityScore** (28.9%) - Strong influence
3. **InterviewScore** (19.8%) - Moderate impact
4. **ExperienceYears** (8.7%) - Secondary factor
5. **Gender** (4.5%) - Minimal direct influence

Sample Predictions Analysis

- **Hire Decisions**: Driven by high skill scores (97+) and personality fit (95+)
- **No-Hire Decisions**: Low skill scores (32) and poor personality fit (37) were decisive factors

5. Bias Mitigation Trade-offs

Reweighing Strategy Results

Fairness Improvements:

- Eliminated 3.2% gender hiring gap
- Increased overall hiring rate from 27% to 35%
- Improved equal opportunity metrics

Performance Trade-offs:

- Accuracy decreased by 2.3 percentage points
- Precision reduced by 10 percentage points
- Recall improved by 9.7 percentage points