1. Dataset Description

The dataset contains structured candidate profiles for hiring decisions. Features include:

- Numerical: Age, EducationLevel, ExperienceYears, PreviousCompanies, DistanceFromCompany, InterviewScore, SkillScore, PersonalityScore
- Categorical: Gender (sensitive), RecruitmentStrategy
- Target: HiringDecision (1 = Hire, 0 = No Hire)

To simulate bias, the training set was skewed with **80% male resumes**. Gender was numerically encoded (0 = Female, 1 = Male), and no label encoding was needed.

2. Model Architecture and Performance

I used a Logistic Regression classifier for its simplicity, interpretability, and strong performance on linearly separable data. All features were standardized using StandardScaler() to ensure equal contribution and stable convergence (mean = 0, std = 1).

Model Summary**:**

- Input: 10 standardized features
- Output: Binary classification (Hire / No Hire)
- Optimizer: LBFGS (default)
- Epoch limit: 1000 (max_iter)
- Baseline Accuracy: 0.858 (evaluated on unbiased test set)

The model was trained on a biased training set (80% male resumes) and tested on an untouched, representative dataset.

3. Fairness Analysis

The following group fairness metrics were used to analyze model behavior across gender:

- Demographic Parity Difference (DPD)
- Equal Opportunity Difference (EOD)
- Average Odds Difference (AOD)

Results Before Mitigation:

Metric	Value	Interpretation
DPD	0.040	Males and females hired at slightly different rates
EOD	0.097	Qualified males more likely to be hired than females
AOD	-0.059	Model moderately favors males

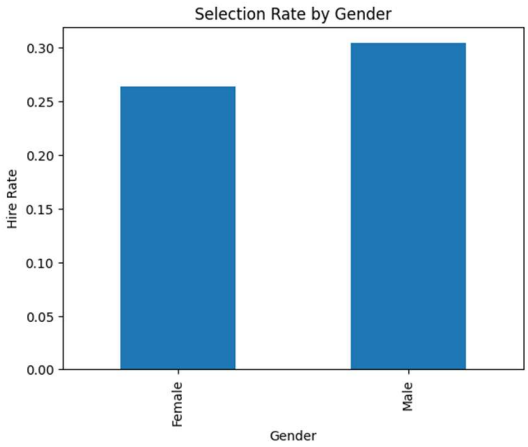

Figure 3-1: Prediction Distribution by Gender

4. Explainability Analysis (SHAP)

SHAP was used to explain 5 model predictions (3 Hire, 2 No-Hire).
Findings:

- EducationLevel, RecruitmentStrategy, and PersonalityScore were the top drivers
- Gender consistently appeared at the bottom of SHAP feature importance
- SHAP plots confirmed low direct impact of gender

5. Mitigation Methods and Tradeoffs

Two bias mitigation techniques were applied and compared:

5.1 Counterfactual Data Augmentation

Duplicate female training samples were made and flipped their gender, then retrained the model.

Bias slightly got worse, so this method was not effective in this scenario.

Metric	Value (aug)
Accuracy	0.855
DPD	0.049
EOD	0.102
AOD	-0.066

5.2 Reweighing (Manual)

I assigned higher weights to underrepresented (gender, label) groups and trained a weighted logistic regression.

This was the most effective method, improving all fairness metrics with minimal impact on accuracy.

Metric	Value (Reweighing)
Accuracy	0.855
DPD	0.005
EOD	0.055
AOD	-0.022

6. Conclusion

The original model showed moderate bias toward males, despite gender not being a dominant feature in SHAP explanations. Among the tested mitigation strategies, Reweighing was the most successful, reducing all fairness gaps significantly with less than 0.3% accuracy loss.