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

1. Dataset Description & Sensitive Feature Encoding

The dataset includes numeric features representing candidate demographics, qualifications, and assessments:

- Age, Gender (0 = Female, 1 = Male), EducationI, Experience, PreviousC, DistanceF, InterviewS, SkillScore, Personality, Recruitme.
- Target: Hiring (1 = Hire, 0 = Not Hire)
- Sensitive attribute: Gender

To simulate real-world bias, the training data was intentionally imbalanced with ~80% male candidates and 20% female.

2. Model Architecture & Performance

Model: Logistic Regression

- Input: All numeric features
- Output: Binary classification (Hire or Not Hire)
- Baseline Accuracy on test data: 85%
- Gender was included as a feature in the baseline model

3. Fairness Analysis (with Plots & Metrics)

Metrics Used:

- Demographic Parity: % of positive predictions across gender
- Equal Opportunity (TPR): True positives across gender
- False Positive Rate (FPR)
- Average Odds Difference (AOD): mean of TPR and FPR differences between gender groups

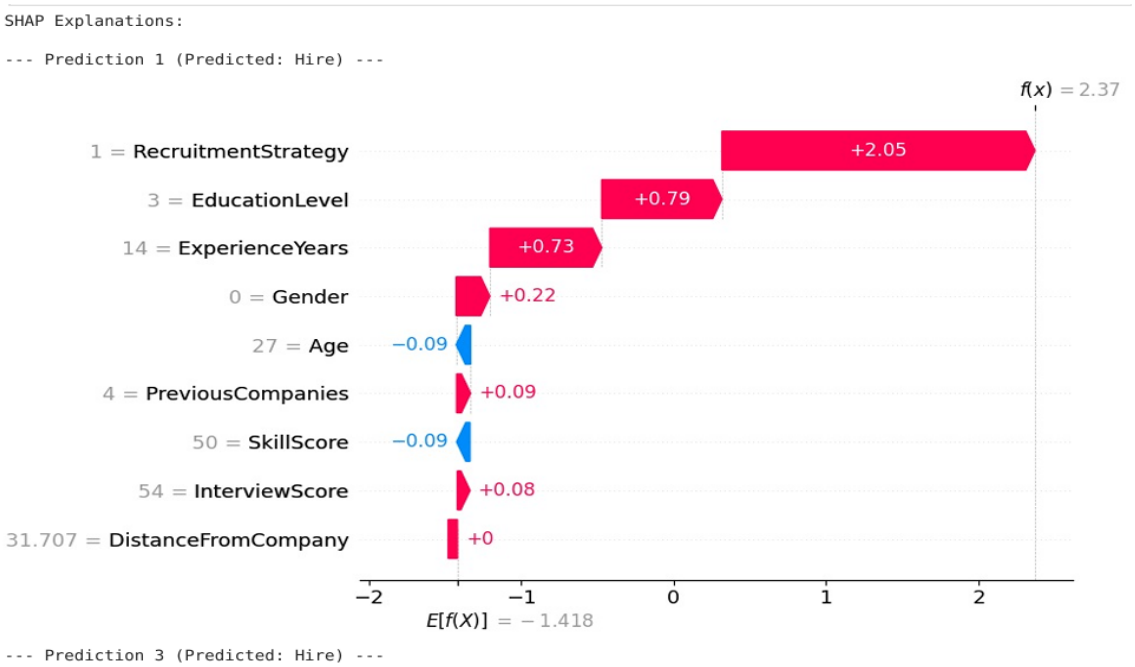
Baseline Results:

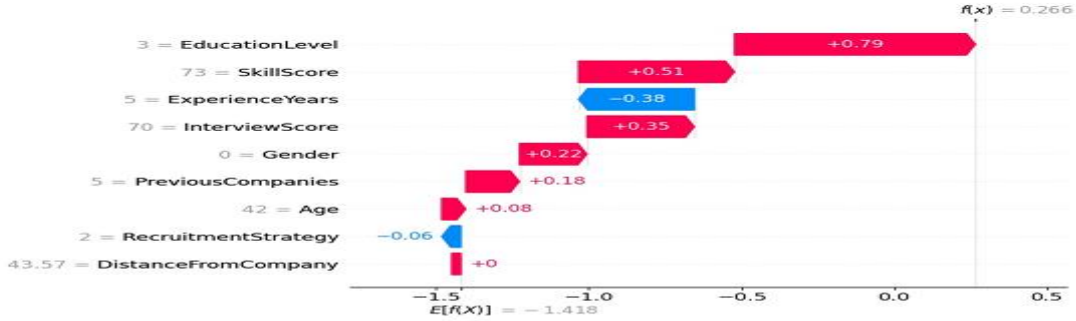
Metric	Female	Male
TPR	72.0%	64.9%
FPR	10.1%	3.6%
Demographic Parity	29.7%	18.9%
Average Odds Difference	0.068	

Observation: The model favors female candidates slightly more (higher TPR & FPR), possibly due to overcompensation from imbalanced training.

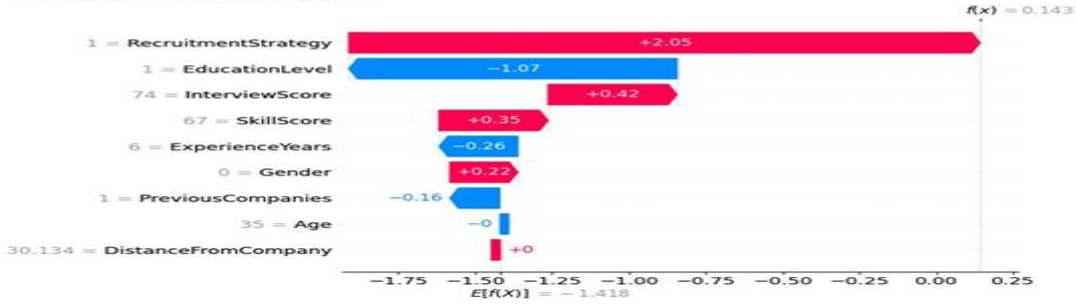
4. Explainability Results & Discussion

SHAP was used to explain 5 predictions (3 Hire, 2 No-Hire).
In the baseline model:
- Gender often had significant influence on "Hire" predictions.
- Example: For some candidates, a male gender label reduced the predicted probability.
After mitigation (removing gender), SHAP showed that skills, personality, and interview scores dominated prediction influence.

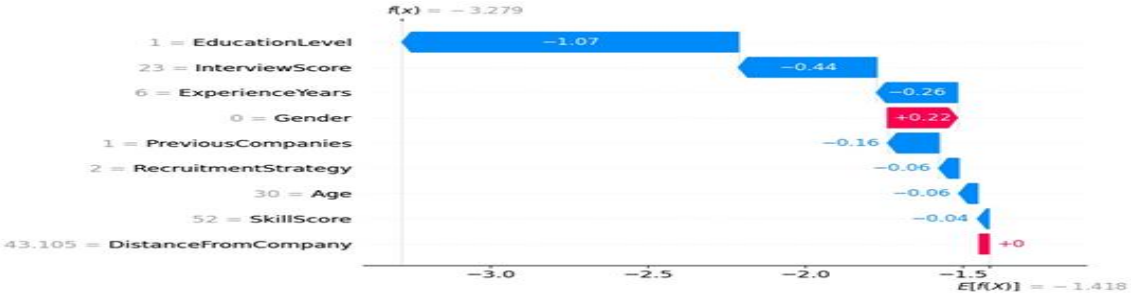




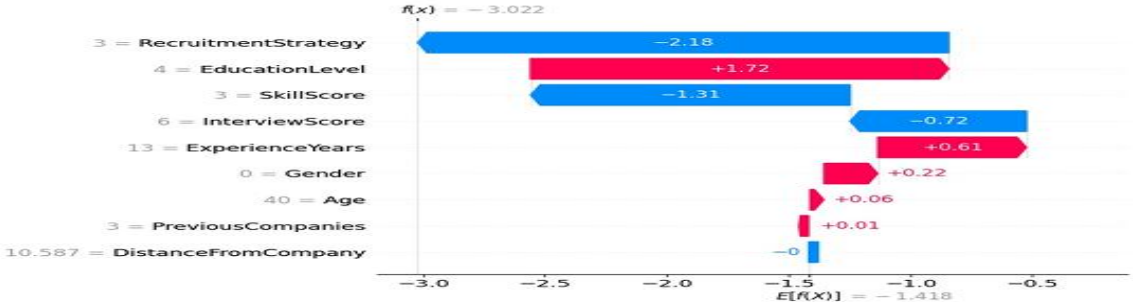
--- Prediction 20 (Predicted: Hire) ---



--- Prediction 0 (Predicted: No Hire) ---



--- Prediction 2 (Predicted: No Hire) ---



5. Mitigation Results & Tradeoffs

Method Used:
- Feature Removal (exclude Gender during training)

Post-Mitigation Results:

Metric	Female	Male
TPR	66.8%	64.9%
FPR	7.7%	3.6%
Demographic Parity	26.4%	18.9%
Average Odds Difference	0.030	
Accuracy	85%	

Outcome:
- Fairness improved substantially, especially in AOD (cut by more than half), without reducing performance.

Tradeoff:
- While fairness improved, proxy bias may still remain in other features correlated with gender.