Uncovering Bias and Explaining Decisions in a Job Screening Model

Dataset description:

The dataset contained structured job application information for candidates. It included numerical features such as: Age, Experience Years, Previous Companies. In addition, two categorical features were present: Education Level and Recruitment Strategy.
The **target variable was Hiring Decision** encoded as 1 for hired and 0 for not hired, and the **sensitive feature was Gender** encoded as 1 for a man and 0 for a woman.
Before model training, categorical features were one-hot encoded, and numerical features were standardized using z-score scaling. The dataset was intentionally split to simulate bias by creating a training set with 80% males and 20% females, while the test set had more balanced representation.

Model architecture and performance:

A binary classification model to predict the Hiring Decision based on structured features was trained. For simplicity and interpretability, Logistic Regression was used.
To simulate real-world imbalance and study bias, the training data was constructed with an 80% male and 20% female ratio, while the test set had a more balanced gender distribution.
All numerical features were standardized using z-score normalization, and categorical features (excluding gender) were one-hot encoded.
The model achieved the following performance on the test set:

- Accuracy: 87.3%
- Precision (Hire): 0.81
- Recall (Hire): 0.78
- F1-score (Hire): 0.80

These results indicate strong overall performance, though some differences were observed between the two classes in recall, suggesting a potential fairness issue which was investigated further.

Fairness Metrics:

To assess the fairness of the model, we computed three key group fairness metrics based on the sensitive attribute Gender:

Metric	Male (1)	Female (0)
Demographic Parity	.304	.303
Equal Opportunity (TPR)	.826	.768
Average Odds Difference	.038	

The Demographic Parity shows balanced prediction rates across gender groups, meaning the model predicts “Hire” at nearly the same rate for males and females. However, the Equal Opportunity score reveals that the model is better at identifying truly qualified male applicants (TPR = 82.6%) compared to female ones (TPR = 76.8%).
The Average Odds Difference, which combines both TPR and FPR differences, was 0.038 — indicating a mild bias that may affect group-level fairness.

- ## Explainability results and discussion:

To interpret individual model decisions, we used LIME to explain 5 predictions from the test set — 3 classified as “Hire” and 2 as “No-Hire.”
Across all five explanations, the most influential features were:

1. Recruitment Strategy (usually a strong negative factor)
2. Education Level and Education Level (positive or negative depending on presence)
3. Interview Score, Skill Score, Personality Score, and Experience Years (contextual impact)

In each case, **Recruitment Strategy** and **Education Level** had the greatest influence on the model’s decisions. Notably, the sensitive attribute **Gender did not appear** in any explanation, suggesting that the model did not rely on gender when making predictions for individual applicants. These results show that while group-level fairness metrics revealed a slight imbalance, the model’s individual decisions were explainable and largely based on job-relevant criteria, supporting transparency and interpretability.

- ## Mitigation results and tradeoffs:

To reduce group-level bias, we applied Reweighing which is a preprocessing method from the AIF360 toolkit that adjusts instance weights during training to reduce the effect of sensitive attributes like Gender. After retraining the model with these weights, we observed slightly improved fairness metrics:

1. Demographic Parity changed from 0.304 vs 0.303 to 0.297 vs 0.310
2. Equal Opportunity (TPR) gap narrowed from 5.8% to 3.1%
3. Average Odds Difference dropped from 0.038 to 0.028

These results indicate a mild but measurable improvement in fairness. Importantly, this was achieved without significant loss in model performance, preserving overall accuracy. This tradeoff demonstrates how simple bias mitigation strategies can be effective in practice and highlights the importance of evaluating both individual and group-level fairness in AI models.