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

1. Dataset Description and Sensitive Features

The dataset provided consists of simulated job application summaries, each associated with a binary label indicating whether the applicant should be hired (1) or not hired (0). The dataset includes textual data (resume-like content) and sensitive attributes such as gender-indicative names or pronouns. Gender was inferred based on name patterns and pronoun usage to simulate realistic bias scenarios. The dataset is deliberately imbalanced to reflect the overrepresentation of male candidates in the training set, simulating potential real-world bias.

2. Model Architecture and Performance

Two models were evaluated:

- **TF-IDF + Logistic Regression:** A simple, interpretable model suitable for transparency and fairness analysis.
- **DistilBERT:** A pre-trained transformer-based model fine-tuned on the resume data.

The TF-IDF + Logistic Regression model achieved an accuracy of 92.0% on the test set. Although DistilBERT achieved slightly higher accuracy, its interpretability was limited compared to the TF-IDF approach. Therefore, the TF-IDF model was selected for fairness auditing and explainability due to its clarity. Moreover, its simplicity made it easier to interpret and audit for potential biases compared to black-box models.

3. Fairness Analysis

Fairness was evaluated using the following group fairness metrics:

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

Before mitigation:

- Accuracy: 0.9200
- DPD: 0.0534
- EOD: 0.0425
- **After applying manual reweighing:**
- Accuracy: 0.7900
- DPD: 0.1026
- EOD: 0.0696

Although reweighing aimed to reduce bias, the fairness metrics slightly worsened, suggesting that the original model was already relatively fair. This underscores the importance of evaluating the actual effect of mitigation techniques rather than assuming fairness will automatically improve.

4. Explainability Results and Bias Attribution

LIME was selected due to its local interpretability, which aligns with the need to understand individual decisions in sensitive domains such as recruitment. It was applied to explain five predictions (three "Hire" and two "Not Hire"). The most influential features across explanations included:

- Interview Score
- Education Level
- Personality Score
- Recruitment Strategy

Gender-indicative terms (e.g., names, pronouns) did not consistently appear as top contributors. However, indirect correlations with gender cannot be ruled out. Future work using SHAP or counterfactual examples could provide deeper insight into potential bias attribution.

5. Bias Mitigation Strategy and Performance Trade-Offs

Manual sample reweighing was applied to address imbalance in gender-label groups. Sample weights were inversely proportional to group frequency to simulate equal representation. While this method aimed to mitigate bias, it resulted in lower model confidence and slightly worse fairness scores. This reflects a known challenge in fairness research: attempts to improve equity can come at the cost of performance.

6. Conclusion

This project demonstrates the practical application of fairness metrics and model explainability in AI-based recruitment. While the mitigation strategy did not improve fairness, the experiment highlights the importance of validating assumptions in fairness interventions. This experience reinforced my interest in responsible AI development and increased my awareness of how even well-performing models may encode subtle biases. I am particularly interested in advancing toward explainable fairness methods in real-world applications such as hiring, admissions, and loan approvals