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

48-Hour Challenge – Bias Detection & Explainability in AI

Prepared by: Rana Helal

---

1. Dataset Description

The dataset includes 1,500 job applicant profiles, each represented with 11 structured features. These include:

- **Numerical Features:**
  Age, ExperienceYears, PreviousCompanies, DistanceFromCompany, InterviewScore, SkillScore, PersonalityScore
- **Categorical Features:**
  Gender (0 = Female, 1 = Male),
  EducationLevel (1 to 4),
  RecruitmentStrategy (1 to 3)
- **Target Column:**
  HiringDecision (0 = Not Hire, 1 = Hire)

There were no missing values, and all columns were appropriately typed. To simulate bias, the training dataset was artificially imbalanced to include significantly fewer female candidates (only 30% of female data used for training).

---

2. Modeling Approach

Data Preparation:

- The dataset was split into training and testing subsets (70% / 30%).
- Numerical features were scaled using StandardScaler.

Classifier:

- A Logistic Regression model was trained using scikit-learn.
- Model: LogisticRegression(max_iter=1000)

Performance:

- **Accuracy:** 84.7%
- **Precision (Hire):** 78%
- **Recall (Hire):** 72%

- **F1-Score (Hire)**: 75%

Although overall accuracy is high, lower performance on the "Hire" class indicates an imbalance and potential unfairness.

---

3. Fairness Analysis

Group fairness metrics were calculated using the Fairlearn library. The sensitive attribute chosen was Gender.

Metric	Value
Accuracy (Female)	83.4%
Accuracy (Male)	86.0%
Demographic Parity Difference	0.104

A demographic parity difference of 0.104 indicates that gender bias exists in the model's predictions.

---

4. Explainability with SHAP

The SHAP (SHapley Additive Explanations) framework was applied to understand the feature contributions to individual predictions.

- Five samples were selected: three predicted as “Hire” and two as “Not Hire”.
- Key influencing features included InterviewScore, SkillScore, and PersonalityScore.
- Gender did not appear as a top feature, but indirect influence through correlated attributes cannot be ruled out.

SHAP visualizations helped to understand how individual features pushed predictions toward “Hire” or “Not Hire”.

---

5. Bias Mitigation

To reduce bias, a constrained optimization approach was used:

- **Method**: ExponentiatedGradient (with DemographicParity constraint) from Fairlearn
- **Post-Mitigation Performance**:
  - Accuracy: 86.3%
  - Demographic Parity Difference: 0.088

This shows that fairness improved without harming (and even slightly improving) overall model accuracy.

---

6. Summary and Conclusions

This challenge demonstrated a real-world application of fairness auditing in recruitment models. The workflow included:

- Identifying bias caused by training data imbalance
- Measuring fairness with standard metrics
- Explaining predictions using SHAP
- Applying bias mitigation techniques

The results confirm that fairness improvements are achievable alongside strong performance, provided proper auditing and modeling strategies are followed.

Additional Notes:

- Full implementation is available in a well-documented Python notebook.
- Visualizations include correlation heatmaps, distribution plots, and SHAP graphs.
- Code is modular and reproducible, designed for Colab or local execution