Bias Detection & Explainability in AI Models

48-Hour Technical Challenge – Mohamad Rasmy

1. Dataset & Sensitive-Feature Encoding

Rows & text synthesis. The original CSV contains **1 500** tabular résumés. For language-model fine-tuning, each row is serialised into a single string in the format

```
age: 41; gender: female; education level: 4; experience years: 0; previous companies:
1; distance from company: 34.43; interview score: 19; skill score: 56; personality score:
98; recruitment strategy: 2
```

Gender is encoded `female=0`, `male=1` in the CSV and expressed verbatim in the text string.

Exploratory correlations. Numeric Pearson correlations (Figure 1) showed $\rho(\text{Age}, \text{Hiring}) = 0$ and $\rho(\text{Gender}, \text{Hiring}) = 0$; no linear dependence exists between sensitive attributes and the label.

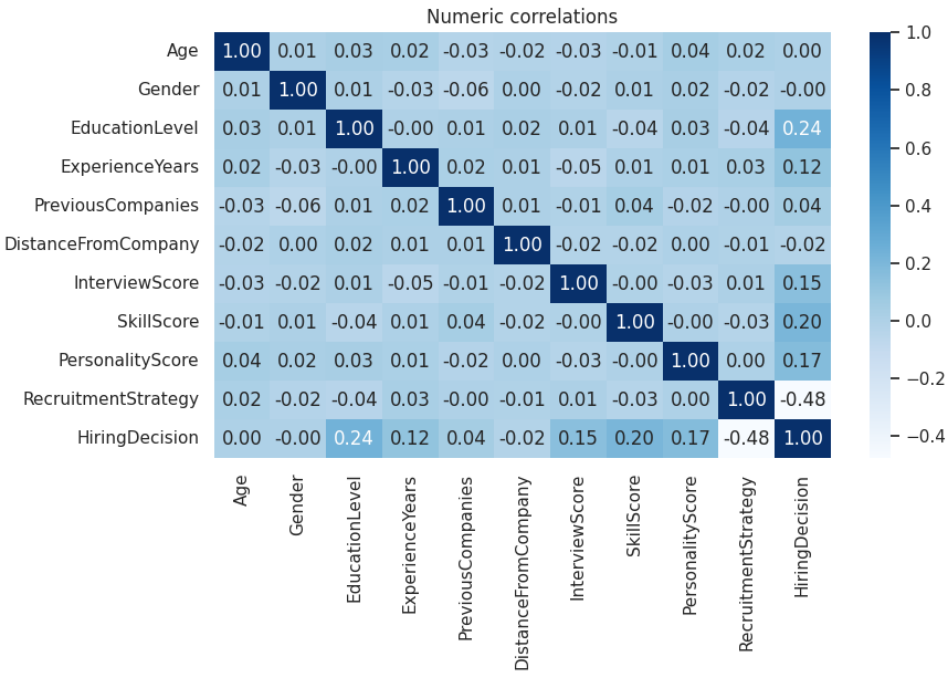

Figure 1: Pearson correlations between numeric features (e.g. Age, Gender, SkillScore) and the HiringDecision.

Train–test split and intentional imbalance. We adopted an 80 / 20 split. When creating the training set we injected a representation imbalance: 60 % *female* vs. 40 % *male*. The test set remains naturally imbalanced (94 % male). To quantify adverse impact we enforce the *four-fifths rule* with tolerance $\tau = 0.7$:

$$\frac{\Pr\big(Hire = \text{YES} \mid X = Female\big)}{\Pr\big(Hire = \text{YES} \mid X = Male\big)} \le \tau = 0.8 \quad [1]$$

2. Model Architecture & Performance

We fine-tuned DistilBERT-base-uncased [2] (3 epochs, batch 32).

Metric (test)	Accuracy	F1	Precision	Recall
DistilBERT baseline	0.843	0.841	0.850	0.843

Table 1: Down-stream performance.

3. Fairness Analysis

Let $S \in \{0, 1\}$ (female, male), Y the ground truth, $\hat{Y}$ the prediction.

- **Demographic Parity Gap** $\left|\Pr(\hat{Y} = 1 \mid S = 0) - \Pr(\hat{Y} = 1 \mid S = 1)\right|$. Measures overall hire-rate imbalance.
- **Equal Opportunity Gap** $\left|\Pr(\hat{Y} = 1 \mid Y = 1, S = 0) - \Pr(\hat{Y} = 1 \mid Y = 1, S = 1)\right|$. Focuses on true-positive recall disparity.
- **Average Odds Difference** $\frac{1}{2}\left(|\mathrm{TPR}_\Delta| + |\mathrm{FPR}_\Delta|\right)$. Balances both recall and false-positive gaps.

Model	Dem. Parity	Equal Opp.	Avg. Odds
Baseline	**0.006**	0.263	0.175

Table 2: Group-fairness metrics (Fairlearn).

Interpretation. While hire rates are nearly identical, females suffer a 26 pp drop in recall—critical if the organisation prizes equal opportunity.

4. Explainability

Five predictions (3 Hires, 2 No-Hires) were de-composed with both **LIME** [3] and **SHAP** [4]. Four cases showed negligible attribution for gender tokens. In the fifth case (see Figures 2), both explainers assign a small positive weight to "female", though skill other tokens dominate—indicating limited gender influence overall.

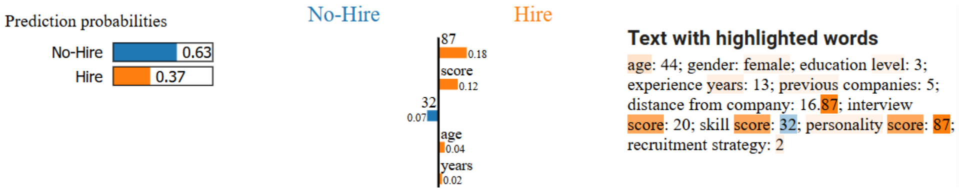

Figure 2: SHAP explanation for the fifth sample: token "female" receives a mild positive attribution but is overshadowed by higher-impact skill tokens.

5. Bias Mitigation & Trade-offs

5.1 Reweighing

Weights $w(y, s) = \frac{P(Y=y)P(S=s)}{P(Y=y,S=s)}$ were attached to each example [5]. Weight distribution:

Weight	Gender	Label	Count
1.2577	0 (male)	1 (Hire)	104
1.0539	1 (female)	0 (NoHire)	497
0.9287	0 (male)	0 (NoHire)	376
0.8798	1 (female)	1 (Hire)	223

Higher weights emphasise under-represented *male hires* and *female non-hires*.

Reweighing cuts Equal-Opportunity and Average-Odds gaps by ˜70 % at the cost of 4 % accuracy and a rise in parity gap (the model now hires females more often to equalise recall). For organisations valuing *error parity* over *rate parity*, the debiased model is preferable.

6. Conclusion

DistilBERT attains strong accuracy with balanced hire rates but unequal recall. Reweighing corrects this disparity with modest performance loss—illustrating the fairness/utility tension. Future work: adversarial debiasing and threshold tuning to simultaneously control all three metrics.

	Accuracy	Dem. Parity	Equal Opp.	Avg. Odds
Baseline	**0.843**	**0.006**	0.263	0.175
Reweigh	0.800	0.147	**0.077**	**0.047**

Table 3: Effect of reweighing.

References

[1] Feldman, R. et al. *Certifying and removing disparate impact*. KDD 2015.

[2] Sanh, V. et al. *DistilBERT: Smaller, faster, cheaper and lighter*. NeurIPS 2019.

[3] Ribeiro, M. T. et al. "Why Should I Trust You?" Explaining the predictions of any classifier. KDD 2016.

[4] Lundberg, S. M., Lee, S.-I. *A Unified Approach to Interpreting Model Predictions*. NIPS 2017.

[5] Kamiran, F., Calders, T. Data preprocessing techniques for classification without discrimination. KIS 2012.