Coursework Report: Automated Interview Screening

Student Name: Xiaoxin Deng
CID: 06046817

1. Task 1: Feature Selection and Measurement Critique

Reliability of measurements. Reliability concerns the consistency of a measurement under repeated or similar conditions. The following attributes in the dataset, `Age`, `Sex`, `Race`, `Place Of Birth`, `Education`, `Workclass`, `Occupation`, `Marital Status`, and `Relationship` are generally reliable, as they are objectively defined and unlikely to vary substantially if measured again.

On the other hand, `interview score`, `cv assessment score` are less reliable due to their subjective nature. Different interviewers may assign different scores to the same candidate, leading to variability in these measurements. Similarly, `Hours Per Week` may fluctuate based on temporary circumstances, making it less reliable as a stable indicator of work commitment.

Validity of measurements. Validity refers to whether a feature accurately captures the construct it is intended to measure, in this case, a candidate's suitability for employment.

`interview scores, cv assessment scores, Occupation` have the highest validity, as they are directly related to a candidate's qualifications and job experience.

`Education, Workclass, Hours Per Week, Marital Status` have moderate validity. While they provide some information about a candidate's background and stability, they are indirect indicators and may not fully capture job-relevant skills or potential. For example, higher education does not always equate to better job performance and personal ability. These attributes may also act as proxy measurements for socio-economic status or access to opportunity, rather than directly measuring competence.

`Age, Race, Sex, Place Of Birth, Relationship` are invalid in this case, as they do not inherently reflect a candidate's job performance or potential. Their inclusion risks introducing bias and discrimination, as they may correlate with societal prejudices rather than merit. In particular, such features function as proxy variables that capture structural or demographic characteristics instead of job-relevant ability, and therefore lack construct validity even if they appear statistically predictive.

Based on these considerations, only job-relevant and valid features such as interview scores, CV assessment scores, and occupation are retained for modelling, while demographic and proxy features should be excluded from the screening model.

2. Task 2: Fairness Metric Selection

Equalized Opportunity is chosen as the primary metric. Because it requires parity in true positive rates across demographic groups. It is well suited to the hiring context because it directly targets allocative harm. Allocative harm arises when qualified individuals from certain groups are systematically denied beneficial outcomes. By enforcing parity in true positive rates, Equalized Opportunity ensures that candidates who are qualified according to the historical decision label have an equal chance of being correctly selected, regardless of group membership. While these labels may themselves reflect historical bias, Equalized Opportunity nevertheless provides a principled constraint on error disparities conditional on the available ground truth. In this way, the metric prevents situations where one group experiences a higher rate of false negatives (qualified candidates being rejected), which would correspond to unfairly withholding opportunities.

Compared to Demographic Parity, which equalizes overall selection rates across groups, Equalized Opportunity more explicitly addresses the distribution of errors. Demographic Parity does not condition on qualification and may therefore encourage accepting unqualified candidates from some groups or rejecting qualified candidates from others in order to match overall rates. In contrast, Equalized Opportunity focuses on a specific and ethically salient error type, making it more appropriate for merit-based decision-making scenarios such as hiring.

Equalized Odds extends this idea by additionally requiring parity in false positive rates, thereby equalizing both types of classification errors across groups. However, this constraint is often difficult to satisfy in practice and can significantly reduce predictive performance, especially when base rates differ between groups.

Given these trade-offs, Equalized Opportunity is chosen due to its principled balance. It mitigates allocative harm by controlling the most harmful error type in this context, while still allowing flexibility in other parts of the error distribution.

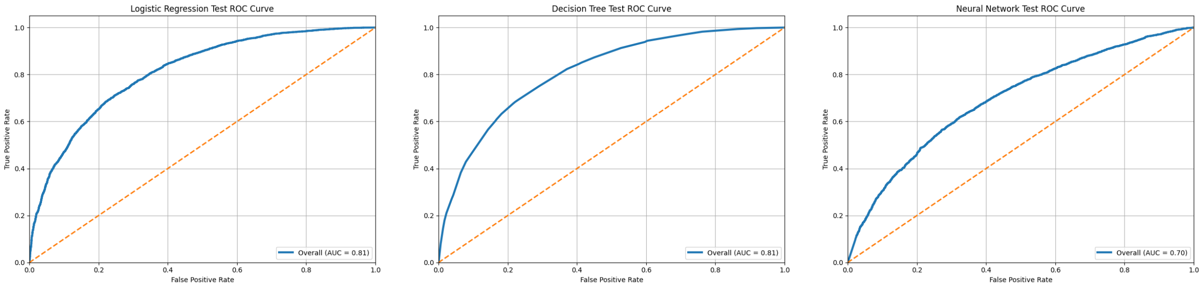
3. Task 3: Baseline Model Comparison

Logistic Regression and the Decision Tree achieve comparable performance, with accuracies of 0.73 and 0.74 respectively, and identical AUC values of 0.81, indicating similar threshold-independent discriminative ability. While Neural Network has the lowest effectiveness in terms of accuracy (0.65) and AUC (0.70). However, it achieves the best Equalized Opportunity difference fairness metric 0.08, indicating that the Neural Network achieves the lowest Equalized Opportunity difference among the baseline models by sacrificing its accuracy and AUC.

This highlights a clear performance-fairness trade-off: models with stronger predictive performance tend to exhibit greater disparity across demographic groups, whereas the Neural Network sacrifices accuracy and discriminative power in exchange for improved fairness.

Model	**Accuracy**	**AUC**	**Fairness Metric Value**
Logistic Regression	0.73	0.81	0.17
Decision Tree	0.74	0.81	0.15
Neural Network	0.65	0.70	0.08

Table 1: Performance and Fairness Comparison of Baseline Models



(a) Logistic Regression (b) Decision Tree (c) Neural Network

Figure 1: Side-by-side ROC curve comparison for baseline models.

4. Task 4: Mitigating Bias

The most effectiveness model from Task 3 is Decision Tree, which achieves the highest accuracy (0.74) and AUC (0.81), while maintaining a reasonable fairness metric value (0.15). We apply two mitigation strategies: pre-processing via re-weighing and post-processing via thresholding. For clarity, the post-processing thresholding method is applied to the re-weighed Decision Tree model, rather than directly to the original baseline, in order to examine the combined effect of pre- and post-processing on fairness. All metrics are evaluated on the test set.

The pre-processing re-weighing method slightly decreases the accuracy to 0.73 and AUC to 0.80, while increasing the fairness metric value to 0.18. This is as expected, since re-weighing adjusts the training data distribution to reduce bias, it does not explicitly enforce conditional fairness constraints such as Equalized Opportunity. As a result, particularly for non-smooth learners like decision trees, re-weighing can increase group-conditional error disparities, leading to a higher EO difference on the test set.

The post-processing thresholding method maintains the accuracy at 0.73, AUC is the same as re-weighing decision tree model since post-processing operates on decision thresholds rather than the underlying score distribution, while significantly reducing the fairness metric value to 0.01, demonstrating a substantial improvement in fairness with minimal impact on accuracy.

These changes indicate that thresholding particularly reduces disparities in false negative rates between demographic subpopulations.

Configuration	Accuracy	AUC	Fairness Metric Value
Baseline (No Mitigation)	0.74	0.81	0.15
Pre-processing (Re-weighing)	0.73	0.80	0.17
Post-processing (Thresholding)	0.72	0.80	0.01

Table 2: Impact of Mitigation Strategies on Performance and Fairness

5. Task 5: Reflection and Analysis

Final Recommendation

Based on the results from Table 1, the Decision Tree model is recommended for deployment for its low Equalized Opportunity difference and high accuracy compare to other baseline models.

According to results from Table 2, Re-weighing pre-processing is not recommended due to its higher Equalized Opportunity difference and lower accuracy. Thresholding post-processing is preferred for its significant fairness improvement with minimal accuracy loss.

However, this recommendation is conditional and applies only within the narrow experimental framing of this coursework. Even with post-processing mitigation, the model should not be treated as an autonomous decision-maker, but at most as a decision-support tool subject to human oversight.

Legitimacy

Fairness metrics measure statistical parity, but do not capture the full ethical complexity of hiring decisions. Moreover, the model is learning to decide based on historical hiring decisions, which may themselves be biased or flawed. Thus, even a fair model may perpetuate existing injustices.

In particular, post-processing thresholding introduces explicit group-dependent decision rules, meaning that individuals with identical model scores may receive different outcomes based solely on group membership. While this may improve group-level fairness metrics, it raises concerns about individual fairness and transparency, reinforcing the risk that the system presents an appearance of objectivity over a fundamentally value-laden decision process.

Therefore, any deployment must be situated within a broader institutional commitment to equity, transparency, and accountability, rather than relying solely on algorithmic fairness as a justification for automated decision-making.

Resource Constraints (Top-N Selection)

In a Top-N selection scenario, the model must rank candidates and select the top 100 based on predicted scores, rather than applying a fixed threshold to classify candidates as hired or not. In this case, the thresholding approach is no longer applicable, as the decision boundary is dynamic and depends on the distribution of scores across all candidates. This introduces additional complexity in ensuring fairness, as the selection process must consider not only individual scores but also group-level representation within the top N.

Crucially, Top-N selection is a strategy selecting one candidate necessarily excludes another. This violates the assumption of independent decisions underlying standard fairness metrics such as Equalized Opportunity. Even if Equalized Opportunity is satisfied under binary thresholding, it does not necessarily hold when decisions are made through ranking, as small score differences near the cut-off can disproportionately affect which groups are represented among the top N.

This highlights that fairness guarantees are highly dependent on deployment context, and that fairness metrics computed under idealized threshold-based assumptions may fail to capture harms in realistic, resource-constrained decision processes.