Coursework Report: Automated Interview Screening

Student Name: Xiaoxin Deng
CID: 06046817

1. Task 1: Feature Selection and Measurement Critique

In this task, we critically evaluate the selected features through the lens of *measurement theory*, focusing on the reliability and validity of each attribute. Since the dataset involves hiring decisions that directly affect individuals' access to opportunities, it is essential to assess whether the chosen features constitute appropriate and fair measurements of the underlying constructs they are intended to represent.

Reliability of measurements. Reliability concerns the consistency of a measurement under repeated or similar conditions. Several attributes in the dataset, such as `Age` and `Hours Per Week`, are relatively reliable, as they are objectively defined and unlikely to vary substantially if measured again. Demographic attributes including `Sex`, `Race`, and `Place Of Birth` are also generally stable, although their reliability may be affected by misreporting or ambiguous category definitions.

In contrast, features derived from human judgement—namely `interview score` and `cv assessment score`—are less reliable. These scores depend on the subjective evaluations of interviewers or assessors and may vary significantly across evaluators, contexts, or time. As a result, the same candidate could plausibly receive different scores under identical conditions, indicating limited measurement reliability.

Categorical socio-economic features such as `Workclass`, `Education`, `Occupation`, `Marital Status`, and `Relationship` are typically stable at the time of collection, but their reliability depends on accurate self-reporting and consistent categorization. Ambiguity in category boundaries or inconsistent reporting can introduce noise into the data.

Validity of measurements. Validity refers to whether a feature accurately captures the construct it is intended to measure. Several features in the dataset raise important validity concerns. For example, `Education` and `Occupation` are often used as indicators of skill or competence, yet they also reflect access to educational opportunities, labour market structures, and historical inequalities. As such, they may be only weakly valid measures of intrinsic ability or future job performance.

Similarly, `Hours_Per_Week` measures time spent working rather than productivity or effectiveness, limiting its validity as a proxy for candidate quality. Human-judgement-based features such as `interview_score` and `cv_assessment_score` are intended to summarise candidate suitability, but they may instead capture assessor preferences, cultural norms, or implicit biases. This raises the risk that these scores encode subjective or socially conditioned factors rather than job-relevant competence.

The target variable, `prior_hiring_decision`, also warrants scrutiny. As a historical hiring outcome, it reflects past human decisions and organisational practices, which may themselves be biased. Consequently, it is an imperfect proxy for true candidate suitability and may propagate existing unfairness when used as a ground-truth label.

Proxy measurements and sensitive attributes. Several features in the dataset may function as *proxy measurements* for sensitive attributes. For instance, `Place_Of_Birth` and `Race` are closely related and may encode similar demographic information. Even if explicit sensitive attributes were removed, correlated features such as `Education`, `Occupation`, or linguistic

signals in CV assessments could allow the model to infer group membership indirectly. This undermines the notion of fairness through unawareness and highlights how discrimination can arise even without explicit use of protected attributes.

Summary. Overall, the dataset contains a mixture of relatively reliable but potentially invalid measurements, as well as subjective features with limited reliability. Many attributes are only indirect proxies for the constructs of interest and may encode historical or societal biases. These measurement issues imply that unfairness may originate from the data itself, independently of the learning algorithm, underscoring the importance of careful feature selection and critical evaluation in fairness-sensitive applications such as hiring.

2. Task 2: Fairness Metric Selection

Equalized Opportunity is chosen as the primary metric. Because it requires equal true positive rate among 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. 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—false negatives among qualified individuals—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 represents a 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.

In summary, Equalized Opportunity is chosen because it directly addresses unfair denial of opportunities to qualified candidates and provides a clear, interpretable way to reason about group-wise error distributions in a high-stakes decision-making setting.

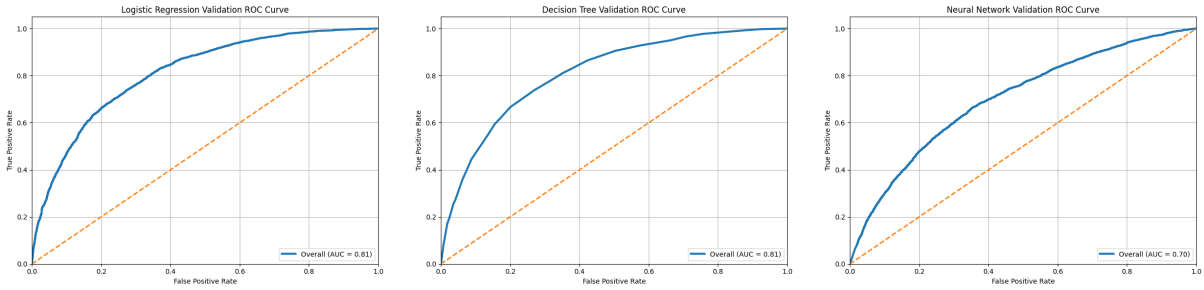
3. Task 3: Baseline Model Comparison

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 discriminative ability. While Neural Network has the lowest effectiveness in terms of accuracy (0.65) and AUC (0.70). However, it achieves the best fairness metric value (Equalized Opportunity difference 0.12), indicating that the Neural Network is the fairest model among the three 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	0.16
Decision Tree	0.74	0.81	0.15
Neural Network	0.65	0.70	0.12

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

Configuration	Accuracy	AUC	Fairness Metric Value
Baseline (No Mitigation)			
Pre-processing (Re-weighing)			
Post-processing (Thresholding)			

Table 2: Impact of Mitigation Strategies on Performance and Fairness

5. Task 5: Reflection and Analysis

Final Recommendation

Legitimacy

Resource Constraints (Top-N Selection)