Question 6

Bias & Fairness Tools

False Negative Rate (FNR) Parity

What the Metric Measures:

- **False Negative Rate (FNR):** The proportion of positive instances (e.g., qualified candidates) incorrectly classified as negative (e.g., rejected) by the model. FNR = FN / (TP + FN), where FN is false negatives and TP is true positives.
- **FNR Parity:** Ensures that the FNR is equal across different demographic groups (e.g., gender, race). It measures whether the model is equally likely to miss positive cases for all groups.

Why It's Important:

FNR Parity is crucial in various applications, especially those involving resource allocation, risk assessment, or opportunities. A disparity in FNR can lead to significant unfairness:

- **Missed Opportunities:** If the FNR is higher for a particular group, it means individuals in that group who are actually eligible or deserving are more likely to be denied a positive outcome (e.g., loan approval, job interview, medical treatment). This can perpetuate existing inequalities and create new disadvantages.
- **Disparate Impact:** In high-stakes scenarios like criminal justice (where a false negative might mean releasing a potentially dangerous individual), a difference in FNR across racial groups could have severe and disproportionate consequences.
- **Erosion of Trust:** When a system unfairly denies opportunities or resources to certain groups at a higher rate, it can erode trust in the system and lead to perceptions of bias and discrimination.

How a Model Might Fail This Metric:

- A model might fail FNR Parity due to several reasons, often stemming from biases in the training data or the model's design:
- **Imbalanced Training Data:** If the training dataset has an uneven representation of different groups and the positive outcome is less frequent for a particular group, the

model might learn to be more cautious in predicting the positive class for that group, leading to a higher FNR.

- **Feature Bias:** Certain features used by the model might be correlated with the sensitive attribute and also be predictive of the outcome. If these features capture societal biases, the model might inadvertently learn to make different types of errors for different groups. For example, if certain language patterns are more common in one group's applications and the model penalizes these patterns, it could lead to a higher FNR for that group.
- **Optimization for Overall Accuracy:** If the model is solely optimized for overall accuracy without considering fairness constraints, it might achieve high accuracy by performing well on the majority group while exhibiting significant disparities in error rates (like FNR) for minority groups.
- **Proxy Variables:** Features that are seemingly neutral but act as proxies for sensitive attributes can also lead to unfair outcomes. For instance, using zip code as a feature in a loan application model could indirectly capture racial or socioeconomic disparities, leading to different FNRs across groups residing in different zip codes.
- **Sampling Bias:** If the data collection process itself introduces biases, the resulting dataset might not accurately reflect the true distribution of the outcome across different groups, leading to a model that performs differently in terms of FNR.

Optional: Applying the Tool (Hypothetically with Demo Data)

This dataset might include features like credit score, income, loan amount, and sensitive attributes like race and gender, along with the actual outcome (loan default or no default).

1. **Load the Data:** The data would be loaded into the Aequitas tool.
2. **Specify Sensitive Attributes:** We would indicate "race" and "gender" as the sensitive attributes we want to analyze for bias.
3. **Specify Outcome Variable:** We would identify the "loan default" column as the outcome we are interested in (positive outcome being "no default" in this context, and a false negative would be predicting "no default" when the person actually defaulted).
4. **Run the Audit:** Aequitas would then calculate various bias metrics, including False Negative Rate Parity, for different subgroups within the sensitive attributes (e.g., FNR for White applicants vs. Black applicants, FNR for male applicants vs. female applicants).
5. **Interpret Results:** The tool would likely output tables or visualizations showing the FNR for each subgroup and the parity ratios (e.g., FNR for Group A / FNR for

Reference Group). If the parity ratio is significantly different from 1 (often with a defined threshold), it would indicate a violation of FNR Parity. For example, if the FNR for Black applicants is significantly higher than for White applicants, it suggests the model is disproportionately likely to incorrectly classify Black individuals as low-risk when they are actually high-risk.