Fairness and Bias Auditing in Financial ML Models

Introduction

Machine learning models in finance, particularly in credit scoring and loan approvals, risk propagating or amplifying historical biases. Such biases can lead to unfair denials of credit or unfavorable terms for certain groups, raising ethical and legal concerns.

Fairness Metrics

- **Demographic Parity**: Ensures decision outcomes are independent of the protected attribute.

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1).$$

- **Equal Opportunity**: Requires equal true positive rates across groups.

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1).$$

- **Disparate Impact**: Ratio of positive outcome rates should be within an acceptable range (commonly 0.8 to 1.25).

Bias Auditing Algorithm

1. Train a predictive model using historical data.
2. Evaluate group-wise metrics (accuracy, true positive rate, false positive rate).
3. Compute fairness metrics: demographic parity difference, equal opportunity difference, disparate impact.
4. Flag potential bias if metrics exceed pre-defined thresholds.

Formal Theoretical Guarantees

Theorem 1 (Bootstrap Consistency)**.** Under mild regularity conditions, bootstrap confidence intervals for fairness metrics (e.g., TPR) are asymptotically consistent and converge to the true distribution quantiles.

Theorem 2 (Fairness via Reweighting)**.** If sample weights satisfy integral equality constraints across groups, then group-level expected predictions can be balanced asymptotically.

Theorem 3 (Adversarial Debiasing Bound)**.** When minimizing adversarial loss with parameter λ, the difference in group outcome probabilities satisfies:

$$\sup_{A} \left| P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = a') \right| \leq \frac{1}{\lambda}.$$

Conclusion

Integrating fairness metrics and formal theoretical safeguards is essential in developing financial models that are both accurate and equitable. Regular audits and mitigation strategies ensure compliance with ethical standards and promote trust in automated decision systems.