Fairness and Bias Auditing for Financial ML Models

Introduction

Machine learning models are widely used in finance for credit scoring and loan approvals. While they promise objectivity and efficiency, they may inadvertently propagate historical biases, leading to systemic discrimination against protected groups (e.g., racial minorities or women).

Sources of Algorithmic Bias

- Historical data reflecting societal inequities.
- Feature choices correlating with protected attributes.
- Objectives focused solely on accuracy without fairness constraints.

Fairness Metrics

- **Demographic Parity**:
$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1).$$
- **Equal Opportunity**:
$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1).$$
- **Disparate Impact**:
$$\frac{P(\hat{Y} = 1 \mid A = 1)}{P(\hat{Y} = 1 \mid A = 0)} \geq 0.8.$$

Bias Auditing Algorithm

1. Train a predictive model on historical data.

2. Evaluate group-wise metrics: accuracy, true positive rate (TPR), false positive rate (FPR).
3. Compute fairness metrics: demographic parity difference, equal opportunity difference, disparate impact.
4. Flag potential bias if metrics exceed set thresholds.
5. Optionally, apply mitigation (e.g., reweighting, adversarial debiasing).

Code Outline (Python)

- Generate synthetic data with a protected attribute.
- Train logistic regression model.
- Compute overall and group-wise metrics.
- Calculate disparate impact and flag bias.
- Include advanced metrics: bootstrap CIs, Wasserstein distance, counterfactual tests, reweighting mitigation.

Example Output:

```
Overall Accuracy: 0.73

Group-wise metrics:
   group  accuracy   TPR  positive_rate
0      0      0.73  0.62           0.37
1      1      0.72  0.67           0.37

Disparate Impact: 1.01
 No major disparate impact detected.
```

Formal Theoretical Guarantees

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

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

Theorem 3 (Adversarial Debiasing Bound). *When minimizing adversarial loss with parameter λ, we have*

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

Conclusion

Combining rigorous fairness metrics, empirical code tests, and theoretical guarantees ensures that financial ML models are both accurate and equitable. Regular auditing with these frameworks supports compliance, reduces risk, and promotes trust.

Links to Code and Proofs

- Code Repository
- Formal Theoretical Proof Document