Algorithmic Credit and Fairness

Theme II: Algorithmic Finance

Research Question: Can algorithmic credit scoring be both accurate and fair, and what trade-offs are inherent?

PhD Seminar in Digital Finance

The Evolution of Credit Scoring

Traditional Scoring (FICO)
Logistic regression on 5 factors:

- Payment history (35%)
- Amounts owed (30%)
- Length of credit history (15%)
- Credit mix (10%)
- New credit (10%)

Score: 300-850 (interpretable)

ML-Based Scoring
Features: 1,000+ variables

- Transaction patterns
- Social connections
- Device/behavioral data
- Alternative data (rent, utilities)

Performance Gains

Model	AUC	KS
Logistic	0.72	0.35
XGBoost	0.78	0.42
Neural Net	0.80	0.45

Hurlin, Perignon & Saurin (2024), "The Fairness of Credit Scoring Models," Management Science

Formal Definitions of Fairness

Group Fairness Metrics
Let $A \in \{0, 1\}$ = protected attribute, Y = outcome, $\hat{Y}$ = prediction.
Demographic Parity

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

Equalized Odds

$$P(\hat{Y} = 1|Y = y, A = 0) = P(\hat{Y} = 1|Y = y, A = 1)$$

for all $y \in \{0, 1\}$
Calibration

$$P(Y = 1|\hat{Y} = s, A = 0) = P(Y = 1|\hat{Y} = s, A = 1)$$

Intuitive Meaning
Demographic Parity
Equal approval rates across groups.
Equalized Odds
Equal TPR and FPR across groups.
Calibration
Score means the same thing for all groups: "700 = 5% default risk" for everyone.
Individual Fairness
Similar individuals treated similarly:

$$d(i, j) \leq \epsilon \Rightarrow |s_i - s_j| \leq \delta$$

These metrics capture different aspects of fairness; no single metric is universally best

The Impossibility Theorem

Kleinberg, Mullainathan & Raghavan (2017)

Proposition (Impossibility)

If base rates differ across groups:

$$P(Y = 1|A = 0) \neq P(Y = 1|A = 1)$$

then no predictor can satisfy both:

1. *Calibration*
2. *Equal false positive/negative rates*

except in trivial cases.

Proof Sketch

Calibration implies:

$$P(Y = 1|\hat{Y} = s) = s$$

for all groups.

Equal FPR implies:

$$\frac{\sum_s (1-s) n_{0s}}{\sum_s n_{0s}} = \frac{\sum_s (1-s) n_{1s}}{\sum_s n_{1s}}$$

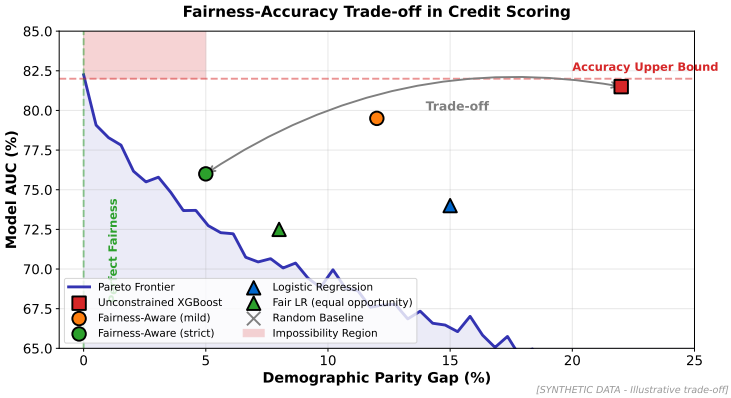
With different base rates, score distributions must differ, making equality impossible.

Implication

Society must choose which fairness criterion to prioritize.

Kleinberg, Mullainathan & Raghavan (2017), "Inherent Trade-Offs in Fair Risk Assessment," ITCS

The Fairness-Accuracy Pareto Frontier



The impossibility theorem manifests as a trade-off curve; policy chooses the operating point.

Empirical Evidence: ML Fairness in Practice

Hurlin, Perignon & Saurin (2024)
Dataset: 500K+ consumer loans (France)

Key Findings

	Logistic	XGBoost
AUC	0.71	0.76
Demographic parity gap	4.2%	6.8%
Equalized odds gap	3.1%	5.2%
Calibration error	1.8%	1.2%

Trade-Off Observed
Better accuracy → worse group fairness.

Sources of Disparity

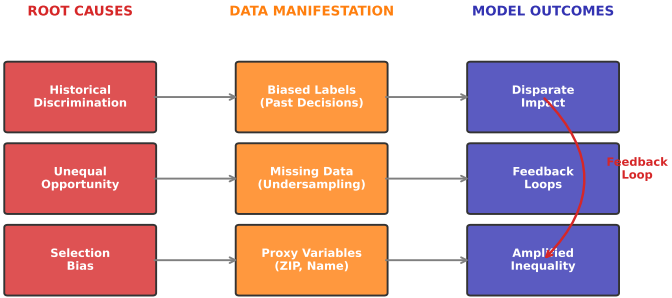
1. **Historical bias**: Past lending discrimination
2. **Proxy variables**: ZIP code, job type
3. **Measurement error**: Thin files for minorities
4. **Feedback loops**: Denial → no credit history

Alternative Data
Mixed evidence:

- Rent, utilities: May help thin-file
- Social media: Raises privacy concerns
- Education: Potentially discriminatory

ML models amplify existing biases unless explicitly constrained

Sources of Algorithmic Bias in Credit



Kleinberg, Mullainathan & Raghavan (2017): Bias propagates through the ML pipeline

[CONCEPTUAL FRAMEWORK]

Bias flows from historical data through proxy variables and feedback loops into disparate outcomes.

Fair Machine Learning Methods

Pre-Processing

- Remove protected attributes
- Reweighting samples
- Learn fair representations

Problem: Proxy variables remain.

In-Processing

Add fairness constraint to loss:

$$\min_{\theta} L(\theta) + \lambda \cdot \text{FairnessViolation}(\theta)$$

Example: Equalized odds regularizer.

Post-Processing

Adjust thresholds per group:

$$\hat{Y}_a = \mathbf{1}[s > \tau_a]$$

Choose τ_a to equalize FPR or TPR.

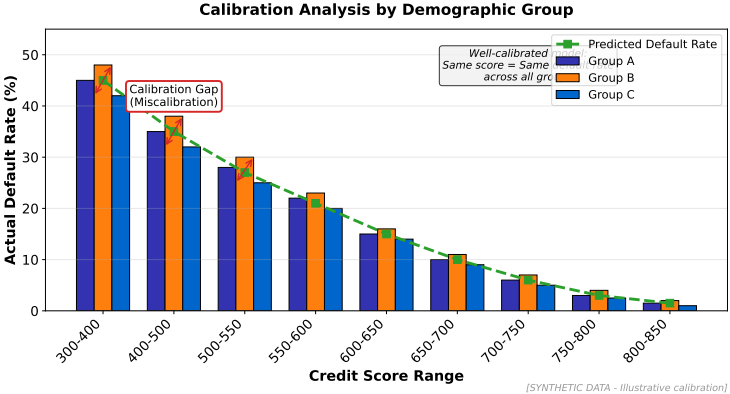
Cost-Benefit Analysis

Method	Accuracy	Fairness
Baseline	100%	0%
Threshold adj.	98%	80%
Regularization	95%	90%
Aggressive	90%	95%

Key Insight: Pareto frontier exists.

arXiv Dec 2024, "An experimental study on fairness-aware ML for credit scoring"

Calibration Analysis: Default Rates by Score and Group



Well-calibrated models have similar default rates at each score level across demographic groups.

Regulatory Landscape: ECOA, FCRA, EU AI Act

US Framework

ECOA (1974)

- Prohibits discrimination on protected characteristics
- Disparate impact doctrine applies
- Adverse action notices required

FCRA (1970)

- Right to know factors
- Accuracy requirements
- Dispute resolution

EU Framework

EU AI Act (2024)

Credit scoring = High-risk AI

- Conformity assessment
- Technical documentation
- Human oversight
- Bias testing requirements

FINMA Guidance 08/2024

Switzerland-specific:

- Model risk management
- Explainability requirements
- Board-level accountability

FINMA Guidance 08/2024, "Governance and Risk Management when using AI" (Dec 2024)

Open Research Questions

Publishable Research Directions

1. **Dynamic Fairness**
   - RQ: How do fair lending policies affect credit access over time?
   - Method: Agent-based model with feedback loops
   - Gap: Most fairness work is static; dynamics understudied
2. **Fairness-Accuracy Frontier Estimation**
   - RQ: What is the empirical cost of fairness constraints?
   - Method: Multi-objective optimization on real lending data
   - Gap: Theoretical trade-offs known; magnitudes not
3. **Alternative Data and Inclusion**
   - RQ: Does alternative data expand or restrict credit access?
   - Method: Natural experiment (fintech entry, regulation)
   - Gap: Marketing claims vs. rigorous evidence

Fairness in finance is a policy-relevant research area with limited causal evidence

Exercises

Mathematical
Prove the impossibility result:
Given:

- Calibration: $P(Y = 1|\hat{Y} = s, A = a) = s$
- Different base rates

Show that equal FPR across groups is impossible.

Empirical
Using UCI German Credit:

1. Train logistic, XGBoost
2. Compute AUC, demographic parity
3. Apply threshold adjustment
4. Plot fairness-accuracy frontier

Data: UCI ML Repository

Research Proposal
Draft 1-page proposal:

- "Fintech Lending and Disparate Impact"
- Natural experiment (if available)
- Outcome variables
- Data requirements

Due: Week 6 – Mathematical proof is good preparation for fairness discussions

German Credit is classic fairness benchmark; real lending data harder to access

Required Readings

Core Papers (Read Before Class)

1. **Kleinberg, Mullainathan & Raghavan** (2017). "Inherent Trade-Offs in the Fair Determination of Risk Scores." *ITCS*.
   - Focus: Theorem 1 (impossibility), Sections 2-3
2. **Hurlin, Perignon & Saurin** (2024). "The Fairness of Credit Scoring Models." *Management Science*.
   - Focus: Empirical methodology, Tables 2-4

Supplementary

- Corbett-Davies & Goel (2018): Fairness survey – accessible introduction
- FINMA Guidance 08/2024: Swiss regulatory perspective
- Barocas & Selbst (2016): Legal foundations – law review

Kleinberg et al. is mathematically elegant; Hurlin et al. provides empirical grounding