# Alternative Credit Scoring: The Black Box Paradox

The best predictors of default are the ones no regulator can audit – so who decides what is fair?

Digital Finance

# Why Can Your Phone's Data Predict Your Creditworthiness Better Than Your Credit Score?

**The Black Box Paradox**

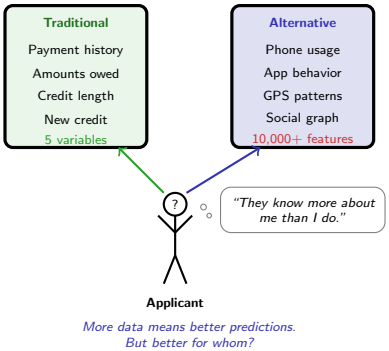
Traditional credit scores use five variables: payment history, amounts owed, length of history, new credit, and credit mix. They work – but they exclude two billion adults who have never had a formal loan.

**What alternative data promises:**

- Mobile phone usage patterns predict repayment behavior
- Utility and rent payments reveal financial discipline
- Social media and browsing data correlate with default risk
- Psychometric tests capture risk attitudes directly

**What alternative data conceals:**

- Nobody knows *why* phone data predicts default
- Correlations may encode race, gender, or income proxies
- Models that work today may fail when borrowers adapt
- Regulators cannot audit what they cannot explain




**Alternative credit scoring can include the excluded – but it predicts default using data that nobody fully understands, including the people being scored.**

# What If You Were Denied a Loan Because of How You Hold Your Phone?

## Reflection Prompt

You apply for a small personal loan. You have a stable job, pay your rent on time, and have never missed a payment. The lender uses an AI-powered credit model. Your application is declined.

**The reason given:** "Your application did not meet our scoring criteria." No further explanation.

What you do not know is that the model considered:

- How fast you scroll through the loan terms (too fast = impulsive)
- Your phone's battery level at application time (low = disorganized)
- The number of contacts in your phone (few = weak social network)
- Whether you type in ALL CAPS (correlated with higher default rates)
- Your GPS data showing you live near a payday lender

Each of these variables has a statistically significant correlation with default. None of them are *causes* of default. And at least two of them – neighborhood and social network size – are proxies for race and income.

**The question is not whether these features predict default.** They do. The question is whether a society should allow decisions about credit access to be made by features the applicant cannot see, understand, or contest.

---

**Statistical correlation is not the same as causation, fairness, or justice – but algorithms cannot tell the difference.**

# What Data Goes Into an Alternative Credit Score – and What Should Be Off-Limits?

| Data Type | Example Features | Predictive Power | Fairness Risk |
|---|---|---|---|
| Traditional | Payment history, balances, credit length | High (AUC 0.70–0.75) | Low |
| Financial | Utility, rent, telecom payments | Moderate–High | Low–Medium |
| Behavioral | App usage, typing speed, scroll patterns | High (AUC 0.80+) | High |
| Social/Network | Contact list, social media activity | Moderate | Very High |
| Geospatial | GPS location, neighborhood data | Moderate | Very High |
| Psychometric | Risk-attitude questionnaires | Moderate | Medium |

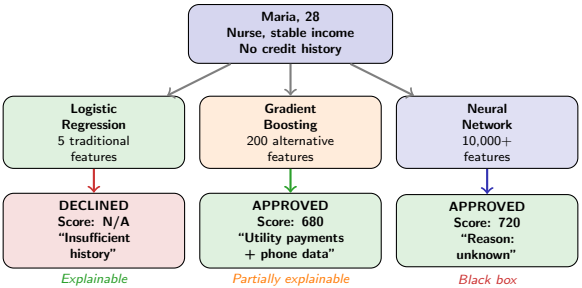
**The pattern to notice**
As you move down the table, predictive power often increases – but so does fairness risk:

- **Traditional data** is limited but auditable. Regulators understand every variable.
- **Financial behavior** (utility, rent) extends coverage to thin-file borrowers without adding opacity.
- **Behavioral data** is where the paradox bites: typing speed genuinely predicts default, but no one can explain *why* – and it may proxy for age or disability.
- **Social and geospatial data** are the most controversial: your neighborhood and your friends predict your creditworthiness, but using them encodes structural inequality.

**The regulatory question:** Where do you draw the line between inclusion and intrusion?

**More data improves accuracy but increases the risk of encoding discrimination – the fairness-accuracy frontier is the central tension of alternative credit scoring.**

# Follow One Loan Application Through Three Scoring Models – and Get Three Different Answers



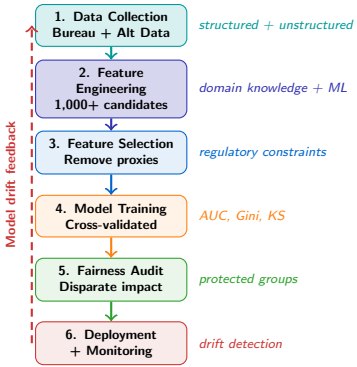

**Three models, three outcomes**

- **Logistic regression** uses only traditional credit bureau data. Maria has no credit file, so the model cannot score her at all. She is "credit invisible" – excluded by design, not by risk.
- **Gradient boosting** adds utility payments, phone top-ups, and app behavior. Maria's consistent rent and phone payments earn her a passing score. The model can partially explain why.
- **Neural network** ingests thousands of features including social graph and device metadata. It scores Maria even higher – but cannot explain which features drove the decision or whether any are discriminatory.

**The dilemma:** The most accurate model is the least explainable. The most explainable model excludes the people who need credit most. Which model should a regulator approve?

**Same applicant, three models, three outcomes. The model that includes Maria is the one no regulator can fully audit.**

# How Do You Build a Credit Scoring Pipeline from Raw Data to Default Probability?




**Six stages, two tension points**

- **Data collection** merges bureau records with alternative sources. Quality varies wildly – phone metadata is noisy, utility records are sparse.
- **Feature engineering** transforms raw data into model inputs. A single phone record can yield hundreds of features: call frequency, contact diversity, top-up regularity.
- **Feature selection** is where fairness enters. Features correlated with protected attributes (race, gender) must be identified and removed – or the model encodes discrimination.
- **Model training** optimizes for predictive accuracy. Cross-validation prevents overfitting, but does not prevent bias.
- **Fairness audit** tests whether approval rates differ across demographic groups. If disparate impact exceeds regulatory thresholds, the model fails.
- **Deployment** is not the end. Models drift as borrower behavior changes – continuous monitoring detects degradation before it causes harm.

**A credit scoring pipeline is not just a model – it is a system where every stage introduces accuracy-fairness trade-offs that compound downstream.**

# What Goes Wrong When the Training Data Reflects Historical Discrimination?

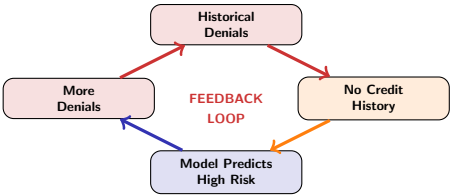
**When the Past Poisons the Future**
Machine learning models learn from historical data. If that history includes decades of discriminatory lending, the model does not correct the bias – it automates it.

**How bias enters the pipeline:**

- **Label bias:** Historical defaults reflect who was *given* loans, not who *could* repay. Groups denied credit have no repayment records – absence of data is not absence of creditworthiness.
- **Feature bias:** ZIP code, school attended, and employer are proxies for race and class. Removing "race" as a feature does not remove racial bias if correlated proxies remain.
- **Feedback loops:** Denied applicants cannot build credit history, making future denials more likely. The model creates the outcomes it predicts.

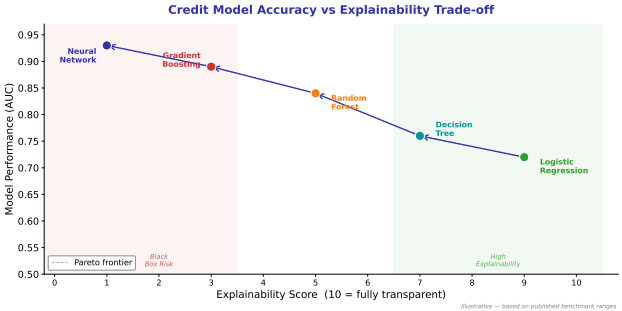
**Real-world consequences:**

- Apple Card investigation: women received lower limits than men with identical financial profiles
- Upstart audit: alternative data models approved more minorities – but at higher interest rates




**Algorithmic fairness requires breaking the feedback loops that encode historical discrimination – not just removing protected attributes.**

# Where Does Model Complexity Help and Where Does It Hurt?



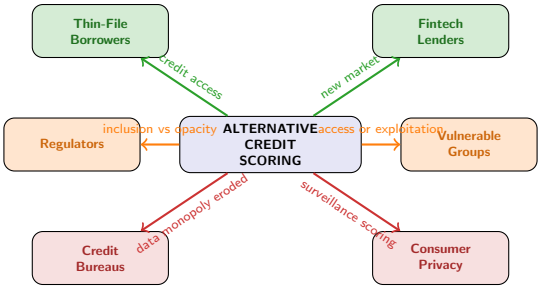

**Reading the trade-off frontier**

- **Logistic regression** is the regulatory baseline: fully explainable, every coefficient has a sign and magnitude a human can interpret. But its AUC ceiling is low – it misses nonlinear patterns in the data.
- **Decision trees** add interpretability through visual rules but sacrifice some accuracy and are prone to overfitting.
- **Random forests and gradient boosting** achieve significantly higher AUC by capturing interactions between features – but explaining any single prediction requires post-hoc tools like SHAP.
- **Neural networks** push accuracy highest but are effectively black boxes. No regulator can audit the internal weights of a deep network.

**The regulatory dilemma:** Moving right on this chart means more people get fair scores. Moving left means regulators can audit the model. You cannot maximize both.

**Illustrative AUC ranges based on published benchmark studies. The Pareto frontier between accuracy and explainability is the defining constraint of modern credit scoring.**

# Who Benefits from Alternative Data – Thin-File Borrowers or Predatory Lenders?




**Winners**

+ **Thin-file borrowers:** People without credit history gain access to loans for the first time. This is the strongest ethical argument for alternative data.
+ **Fintech lenders:** Alternative data lets fintechs underwrite populations that traditional banks cannot reach – a new market with less competition.

**Losers**

– **Credit bureaus:** Their data monopoly erodes as lenders bypass bureau scores entirely.
– **Consumer privacy:** Scoring based on phone and social data turns everyday behavior into a credit input without informed consent.

**Mixed impact**

~ **Regulators:** More inclusion is good, but opaque models make supervision nearly impossible.
~ **Vulnerable groups:** Gain access – but may face higher rates if models price in their neighborhood or social network.

**Alternative credit scoring is not neutral – it opens doors for the excluded but also creates new forms of surveillance and discrimination.**

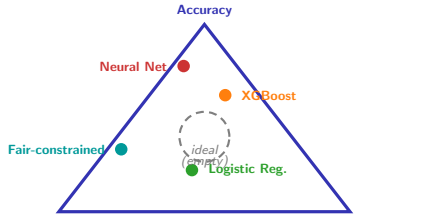
# The Scoring Evaluation Framework: Accuracy, Fairness, Explainability – Pick Two?

**The Credit Scoring Trilemma**
When evaluating any credit scoring model, assess three dimensions. The uncomfortable truth: optimizing two typically requires sacrificing the third.

1. **Accuracy: Does it predict default well?**
   Measured by AUC, Gini coefficient, KS statistic. Higher accuracy means fewer bad loans and more creditworthy borrowers approved. But accuracy on *whom*? A model can be highly accurate overall while systematically mispricing minorities.

2. **Fairness: Does it treat groups equitably?**
   Measured by disparate impact ratio, equalized odds, calibration across groups. A fair model approves protected groups at rates proportional to their true creditworthiness – not their historical treatment.

3. **Explainability: Can a human audit it?**
   Measured by number of features, availability of feature attributions (SHAP, LIME), compliance with adverse action notice requirements (ECOA, GDPR Art. 22).




*Every model lives somewhere in this triangle – none live in the center.*

**The framework:** Score each model 1–10 on all three. No model scores above 8 on all three simultaneously.

**The trilemma is real: no existing model maximizes accuracy, fairness, and explainability simultaneously. Every deployment is a policy choice about which dimension to sacrifice.**

# Your Challenge: Audit a Credit Model for Disparate Impact

## Mini-Challenge (15 minutes)

A fintech lender uses a gradient boosting model with 200 features (including phone metadata and utility payments) to score thin-file borrowers. The overall AUC is 0.85. The lender claims the model "expands access to underserved communities." A regulator asks you to audit it.

**Your deliverable:** Apply the trilemma framework:

1. **Accuracy audit**
   - Is 0.85 AUC measured on all applicants or only approved ones?
   - How does accuracy differ across demographic groups?
   - Is the model calibrated – does a "70% default probability" actually mean 70% of those borrowers default?
2. **Fairness audit**
   - Compute the disparate impact ratio: approval rate of protected group divided by approval rate of majority group. Is it above 0.8?
   - Do any of the 200 features proxy for race, gender, or age?
   - Does the model assign higher interest rates to borrowers in minority neighborhoods, even after controlling for risk?
3. **Explainability audit**
   - Can the lender provide an adverse action notice explaining *why* a specific applicant was denied?
   - Are SHAP values available for individual predictions?

**Conclude:** Would you approve this model for production use?

**Auditing a credit model is not a technical exercise – it is a policy judgment about how much accuracy a society is willing to trade for fairness and transparency.**