# Lesson 2.2: Credit Scoring — From FICO to ML

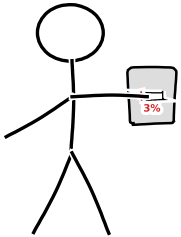
## Module 2: The Access Problem

Prof. Dr. Joerg Osterrieder

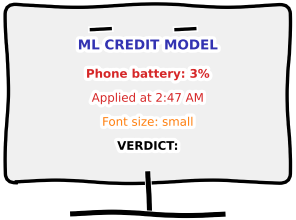
Digital Finance — BSc Course

# How Machines Judge Your Creditworthiness

"*I have a stable job, savings, and zero debt!*"




**"Battery at 3% suggests poor planning skills. DECLINED."**




*(Alternative data: great for inclusion, terrifying for privacy)*

**When your credit score depends on charging habits**

# Learning Objectives

After completing this lesson, you will be able to:

1. **Explain** the purpose of a credit scorecard and how FICO-style scoring works [Understand]
2. **Calculate** Weight of Evidence (WoE) and Information Value (IV) for a feature [Apply]
3. **Interpret** a logistic regression PD model and its ROC curve [Apply]
4. **Compare** traditional scorecards with gradient boosting and neural-network approaches [Analyze]
5. **Evaluate** the trade-offs of alternative data for thin-file borrowers [Evaluate]

**Bloom's levels covered:** Understand, Apply, Analyze, Evaluate

---

**Objectives follow Bloom's taxonomy: Understand → Apply → Analyze → Evaluate.**

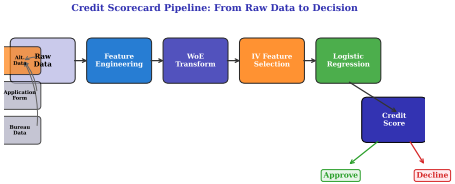
# Bridge: From Exclusion to Scoring

**Last lesson (M2L1):**

- 1.4 billion adults remain unbanked globally
- Exclusion stems from data gaps—no credit history, no formal records
- Traditional banks require documentation that many cannot provide

**This lesson:**

- Credit scoring **determines who gets access**
- If the model is wrong, creditworthy people are excluded
- ML and alternative data can fill the data gap—but at what cost?




The credit scoring pipeline: from raw data to approve/decline.

**Exclusion stems from data gaps. Credit scoring determines who gets in.**

# What Is Credit Scoring?

## Definition: Credit Score

A **credit score** is a numerical summary of a borrower's predicted creditworthiness, derived from historical data. It compresses complex financial behaviour into a single number used for binary decisions: **approve** or **decline**.

**Key characteristics:**

- **Predictive:** Estimates the *probability of default* (PD) within a defined time horizon (typically 12 months)
- **Ordinal:** Higher score $\Rightarrow$ lower default risk (or vice versa, depending on convention)
- **Standardised:** Enables consistent lending decisions across millions of applications
- **Regulatory:** Basel II/III require banks to estimate PD, LGD, and EAD for capital adequacy

**Central tension:** A score that is too conservative excludes good borrowers; one that is too permissive increases losses.

---

**Credit scoring converts messy human financial behaviour into a single, decision-ready number.**

# The FICO Paradigm

**FICO scores (300–850)** have dominated US consumer credit since 1989:

| Factor | Weight | What It Captures |
|---|---|---|
| Payment history | 35% | On-time vs. late payments, delinquencies |
| Amounts owed | 30% | Credit utilisation ratio |
| Length of history | 15% | Age of oldest and average accounts |
| New credit | 10% | Recent inquiries and new accounts |
| Credit mix | 10% | Diversity of credit types (cards, mortgage, auto) |

**Strengths:** Transparent, well-understood, decades of back-testing.
**Weaknesses:** Requires an existing credit file—useless for thin-file or no-file borrowers (the "Catch-22" of credit).

**An estimated 45 million Americans are "credit invisible"—they have no FICO score at all (CFPB, 2015).**

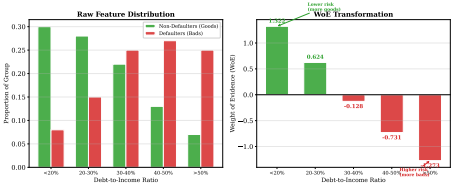
# How a Traditional Scorecard Works

A **credit scorecard** maps borrower characteristics into points:

1. **Bin** each feature (e.g., age: 18–25, 26–35, ...)
2. Compute **Weight of Evidence (WoE)** per bin
3. Select features using **Information Value (IV)**
4. Fit a **logistic regression** on WoE-transformed features
5. Convert regression coefficients into **scorecard points**

**Why logistic regression?**

- Interpretable: each feature's contribution is transparent
- Monotonic: higher income ⇒ better score (if engineered correctly)
- Regulators can audit every coefficient




WoE transforms a raw feature into a measure of default-separation power.

**Traditional scorecards are built for interpretability first, accuracy second—a regulatory requirement.**

# Weight of Evidence (WoE)

## Definition: Weight of Evidence

For a binned feature, the WoE of bin $i$ measures how strongly that bin separates "goods" (non-defaulters) from "bads" (defaulters):

$$\text{WoE}_i = \ln\left(\frac{\text{Distribution of Goods}_i}{\text{Distribution of Bads}_i}\right)$$

where $\text{Distribution of Goods}_i = \frac{\#\text{Goods in bin } i}{\text{Total Goods}}$ and analogously for Bads.

**Interpretation:**

- WoE $> 0$: bin has proportionally **more goods** than bads $\Rightarrow$ lower risk
- WoE $< 0$: bin has proportionally **more bads** than goods $\Rightarrow$ higher risk
- WoE $= 0$: bin's goods/bads ratio equals the population ratio $\Rightarrow$ no discrimination

**Worked example:** If bin "Age 26–35" contains 25% of all goods and 15% of all bads:

$$\text{WoE} = \ln\left(\frac{0.25}{0.15}\right) = \ln(1.667) = 0.511$$

This positive WoE indicates lower default risk for this age group.

**WoE transforms categorical and binned features into a continuous, log-odds scale suitable for logistic regression.**

# Information Value (IV): Feature Selection

## Definition: Information Value

The **Information Value (IV)** of a feature quantifies its overall power to separate goods from bads across all bins:

$$\text{IV} = \sum_{i=1}^{n} (\text{Distr. Goods}_i - \text{Distr. Bads}_i) \times \text{WoE}_i$$

**IV interpretation rules (Siddiqi, 2006):**

| IV Range | Predictive Power |
|---|---|
| $< 0.02$ | Useless — do not include |
| $0.02 - 0.10$ | Weak predictor |
| $0.10 - 0.30$ | Medium predictor |
| $0.30 - 0.50$ | Strong predictor |
| $> 0.50$ | Suspicious — check for data leakage |

**Example:** If a feature has 3 bins with WoE values of 0.5, $-0.2$, 0.3 and proportions yielding ($p_{\text{good}} - p_{\text{bad}}$) of 0.05, 0.03, 0.04, then $\text{IV} = 0.05 \times 0.5 + 0.03 \times (-0.2) + 0.04 \times 0.3 = 0.025 + (-0.006) + 0.012 = 0.031$. This indicates weak predictive power ($\text{IV} < 0.10$).

**Warning:** $\text{IV} > 0.50$ often signals that the feature contains information from the future (e.g., the outcome itself) rather than genuine predictive power.

**IV is the industry-standard metric for feature selection in credit scoring — it guides which variables enter the scorecard.**

# PD, LGD, EAD: The Basel Framework

**Basel II/III** requires banks to estimate three risk parameters for every exposure:

| Parameter | Meaning | Typical Range |
|---|---|---|
| **Probability of Default (PD)** | Likelihood the borrower defaults within 12 months | 0.03%–20% |
| **Loss Given Default (LGD)** | Fraction of exposure lost if default occurs | 10%–60% |
| **Exposure at Default (EAD)** | Outstanding amount at the time of default | Varies |

**Expected Loss (EL):**

$$\text{EL} = \text{PD} \times \text{LGD} \times \text{EAD}$$

**Example:** A \$10,000 personal loan with PD $= 5\%$, LGD $= 40\%$:

$$\text{EL} = 0.05 \times 0.40 \times \$10{,}000 = \$200$$

The bank must hold capital proportional to the **unexpected loss** (beyond EL), determined by the Basel Internal Ratings-Based (IRB) formula.

**Credit scoring primarily targets PD estimation; LGD and EAD are often modelled separately.**

# Logistic Regression for PD Estimation

**Logistic regression** is the workhorse of traditional credit scoring:

$$\text{PD}(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

where $x_1, \ldots, x_k$ are WoE-transformed features and $\beta_i$ are fitted coefficients. **Why logistic regression for credit scoring?**

- Output is bounded between 0 and 1 — directly interpretable as a probability
- **Coefficients are additive** in log-odds space, enabling scorecard point systems
- Easily auditable: regulators can inspect each $\beta_i$
- Well-calibrated when the training data is representative

**From PD to score:**

$$\text{Score} = A - B \times \ln\left(\frac{\text{PD}}{1 - \text{PD}}\right)$$

where $A$ and $B$ are scaling constants chosen so that, e.g., a score of 600 corresponds to 50:1 odds of non-default, and every 20 points doubles the odds ("points to double the odds" or PDO).

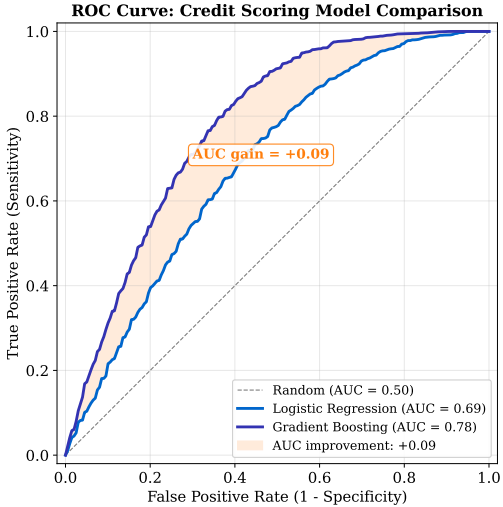
**The log-odds linearity of logistic regression is what makes traditional scorecard construction possible.**

# Measuring Discrimination: The ROC Curve

**How well does the model separate goods from bads?**

- The **ROC curve** plots True Positive Rate (sensitivity) against False Positive Rate (1 − specificity) at every score threshold
- **AUC** (Area Under the Curve): summary metric
  - AUC = 0.50: random guessing
  - AUC = 0.70–0.80: acceptable
  - AUC = 0.80–0.90: good
  - AUC > 0.90: excellent (check for leakage)
- Industry standard for credit scoring: **AUC 0.70–0.80** on out-of-time validation

**Note:** AUC measures *ranking* ability, not calibration. A model can rank well but assign wrong absolute PDs.



# Gini Coefficient and KS Statistic

**Gini coefficient:**

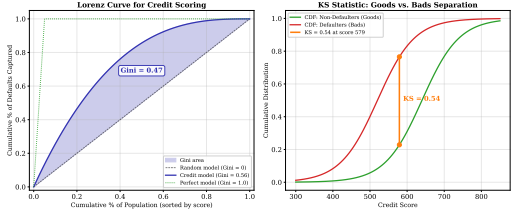
$$\text{Gini} = 2 \times \text{AUC} - 1$$

- Gini $= 0$: no discrimination
- Gini $= 1$: perfect discrimination
- Industry benchmark: Gini 0.40–0.60 for consumer lending

**Example:** If AUC $= 0.78$, then Gini $= 2 \times 0.78 - 1 = 0.56$. A Gini of 0.56 indicates good discriminatory power.

**Kolmogorov–Smirnov (KS) statistic:**

- Maximum vertical distance between the cumulative distribution of goods and bads
- KS $= \max_s |F_{\text{good}}(s) - F_{\text{bad}}(s)|$
- KS $> 0.40$: strong separation
- Identifies the **optimal cut-off score**




The Gini coefficient equals twice the area between the Lorenz curve and the diagonal.

**Gini and KS are the two most-reported discrimination metrics in credit risk model validation.**

# Beyond Logistic Regression: ML Enters Credit Scoring

**Why move beyond logistic regression?**

- Logistic regression assumes **linear** relationships in log-odds space
- Real-world credit risk involves **non-linear interactions**: e.g., high income + high debt behaves differently from high income alone
- ML models can capture these interactions automatically

**ML models used in credit scoring:**

| Model | Advantage | Limitation |
|---|---|---|
| Gradient Boosting (XGBoost, LightGBM) | Highest accuracy, handles missing data | Black-box |
| Random Forest | Robust, less overfitting | Lower accuracy than boosting |
| Neural Networks | Flexible, handles unstructured data | Requires large data, opaque |
| Support Vector Machines | Good for high-dimensional data | Poor interpretability |

**The accuracy–interpretability trade-off:** ML models typically improve AUC by 2–5 percentage points over logistic regression, but at the cost of transparency.

**In FinTech lending, gradient boosting has become the de facto standard for PD estimation.**

# Gradient Boosting for Credit Scoring

**How gradient boosting works (simplified):**

1. Start with a simple prediction (e.g., the base default rate)
2. Fit a shallow decision tree to the **residual errors**
3. Add the tree's predictions (scaled by a learning rate) to the current model
4. Repeat for hundreds of iterations
5. Final prediction = sum of all trees

**Why it dominates credit scoring competitions:**

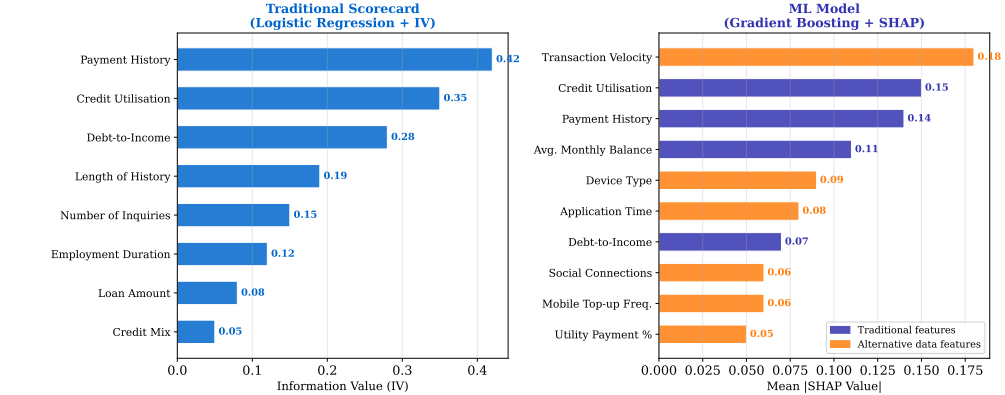
- Automatically captures **non-linear effects** and **feature interactions**
- Handles **missing values** natively (no imputation needed)
- Built-in **regularisation** (max depth, learning rate, subsampling)
- Extremely fast training with LightGBM or XGBoost

**Key hyperparameters for credit scoring:**

- `max_depth`: 3–6 (shallow trees prevent overfitting)
- `learning_rate`: 0.01–0.1 (slower learning = better generalisation)
- `n_estimators`: 100–1000 (more trees with lower learning rate)

**XGBoost won the 2015 KDD Cup credit scoring challenge and remains the industry benchmark.**

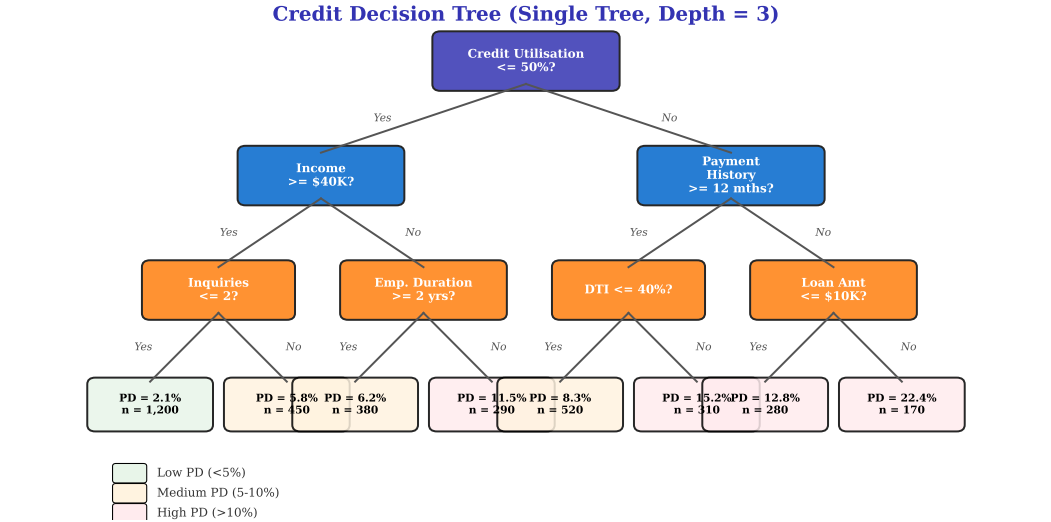
# Feature Importance: Traditional vs. ML




- **What you see:** Two horizontal bar charts—left shows 8 traditional features (IV scores), right shows 10 ML features (SHapley Additive exPlanations (SHAP) values) with orange bars highlighting alternative data
- **Key pattern:** ML models surface new features (transaction velocity, device type, mobile top-ups) that traditional scorecards ignore
- **Takeaway:** Alternative data enables scoring for thin-file borrowers but raises privacy and fairness questions

**ML models often surface features (e.g., transaction velocity, device type) that traditional scorecards ignore.**

# Visualising Credit Decisions: A Single Tree



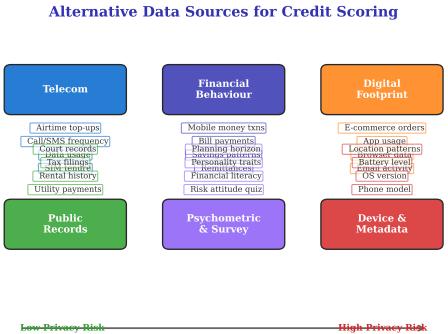

- Each node splits on one feature (e.g., income > $40K?)

# Alternative Data: Scoring the "Unscorable"

**Thin-file borrowers** have too little credit history for a traditional score. Alternative data fills the gap:

- **Telecom data:** Mobile top-up patterns, call/SMS frequency
- **Utility payments:** Electricity, water, rent payment history
- **Digital footprint:** E-commerce transactions, app usage
- **Psychometric tests:** Behavioural questionnaires
- **Social media:** Network connections, posting patterns
- **Device metadata:** Phone model, OS, battery level at application time

**Promise:** Enables credit for 1.4B unbanked.
**Risk:** Privacy invasion, proxy discrimination, regulatory uncertainty.




**Alternative data can extend credit to thin-file borrowers, but raises profound questions about privacy and fairness.**

# Alternative Data in Practice

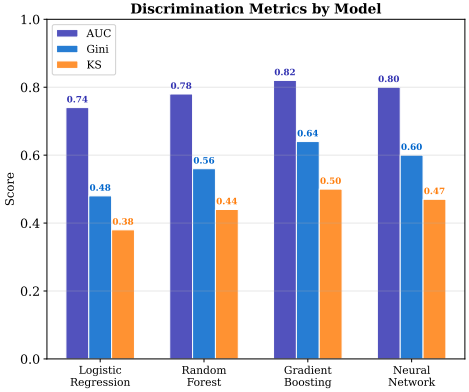
| Company / Region | Data Used | Result | Concern |
|---|---|---|---|
| M-Shwari (Kenya) | Mobile money history, air-time | 60%+ of borrowers are first-time | Over-indebtedness |
| Lenddo (Philippines) | Social media, smartphone data | Gini $\approx$ 0.50 for thin-file | Privacy, bias |
| FICO XD (US) | Utility, telecom, public records | Scores 15M "unscorable" Americans | Accuracy gap |
| Ant Group (China) | Alipay transactions, Sesame Credit | 500M+ scored | Social control |

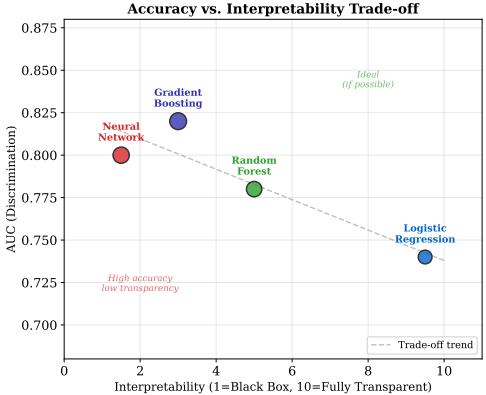
**Key trade-offs:**

- **Inclusion vs. privacy:** More data $\Rightarrow$ better scoring, but deeper surveillance
- **Accuracy vs. fairness:** Features that predict well may correlate with protected characteristics (race, gender, geography)
- **Innovation vs. regulation:** Regulators struggle to keep pace with novel data sources

**Alternative data is not inherently good or bad—the question is governance and consent.**

# Model Performance: Logistic Regression vs. ML



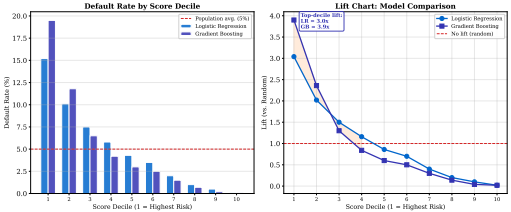

**ML models typically gain 2–5 AUC points over logistic regression, with larger gains on alternative-data-rich populations.**

# Lift Chart: Where Better Models Matter Most

A **lift chart** shows how much better the model is than random selection at each decile:

- **Top decile lift:** How many times more defaults are captured in the riskiest 10% vs. random
- Lift $> 3.0$ in top decile: good model
- ML models often show the biggest lift improvement in the **middle deciles**—precisely where approve/decline decisions are most uncertain

**Business impact:** A 10% lift improvement in the cut-off region can translate to millions in reduced losses or additional approved loans.

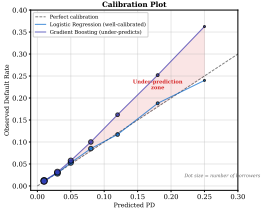


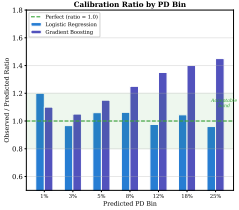

**Lift charts reveal where a model adds practical value—not just overall accuracy, but decision-boundary accuracy.**

# Calibration: Do Predicted PDs Match Reality?

**Discrimination $\neq$ calibration.** A model can rank borrowers correctly but assign wrong absolute probabilities. **What is calibration?**

- If the model says PD = 5%, then 5% of those borrowers should actually default
- A **calibration plot** graphs predicted PD (x-axis) vs. observed default rate (y-axis)
- Perfect calibration $\Rightarrow$ points lie on the 45-degree diagonal




**Why calibration matters:**

- **Pricing:** Loan interest rates are set based on PD
- **Capital:** Basel requires accurate PD for capital adequacy
- **Provisioning:** Expected loss reserves depend on PD accuracy

**A well-calibrated model is essential for pricing, capital computation, and regulatory compliance.**

# Fairness and Bias in Credit Scoring

**Why fairness matters:**

- Credit scores directly determine access to housing, education, and entrepreneurship
- Historical data **embeds past discrimination**: if certain groups were denied credit historically, the model learns to deny them again
- Protected characteristics: race, gender, age, ethnicity, religion

**Types of bias in credit models:**

| Bias Type | Description |
|---|---|
| Historical bias | Training data reflects past discriminatory practices |
| Proxy discrimination | Zip code, university, or job title correlates with race/ethnicity |
| Measurement bias | Alternative data may be less available for disadvantaged groups |
| Feedback loops | Denied applicants never generate repayment data, reinforcing exclusion |

**Regulatory context:** US Equal Credit Opportunity Act (ECOA) and EU AI Act classify credit scoring as "high-risk" AI requiring explainability and fairness testing.

**Algorithmic fairness in credit scoring is not just an ethics question—it is a legal requirement.**

# Explainability: Why Was I Declined?

**Adverse action notices** (US ECOA / FCRA): When a credit application is declined, the lender **must** provide specific reasons. **Explainability techniques for ML models:**

| Method | Type | Explanation Provided |
|---|---|---|
| Logistic regression | Inherent | "Your debt-to-income ratio was too high" |
| SHAP values | Post-hoc | Feature-level contribution to individual prediction |
| LIME | Post-hoc | Local linear approximation around the decision point |
| Partial dependence | Post-hoc | Average effect of one feature across the population |
| Reject inference | Model-specific | Adjusts for missing data on declined applicants |

**The practical solution:** Many FinTechs use a **two-model approach**:

1. **ML model** makes the actual decision (higher accuracy)
2. **Logistic surrogate** generates human-readable explanations

**Explainability is not optional—regulators and consumers have a legal right to understand credit decisions.**

# Model Validation: Ensuring the Score Works

**A credit scoring model must be validated on multiple dimensions:**

| Dimension | What Is Tested | Key Metric |
|---|---|---|
| Discrimination | Can the model rank-order risk? | AUC, Gini, KS |
| Calibration | Do predicted PDs match observed defaults? | Hosmer–Lemeshow test |
| Stability | Does the model perform consistently over time? | PSI (Population Stability Index) |
| Concentration | Is the score distribution reasonable? | Herfindahl Index |
| Back-testing | Does it predict well on historical out-of-time data? | Actual vs. predicted default |
| Stress testing | Does performance hold under adverse scenarios? | PD under stress |

**Validation frequency:**

- **At deployment:** Full validation (discrimination, calibration, stability)
- **Quarterly:** Monitoring reports (PSI, actual vs. predicted)
- **Annually:** Comprehensive review with model risk management

**Model validation is continuous, not a one-time event—models degrade as populations and economies change.**

## Definition: Population Stability Index (PSI)

PSI measures how much the score distribution of a new population has shifted from the development sample:

$$\text{PSI} = \sum_{i=1}^{n} (\text{Actual}_i - \text{Expected}_i) \times \ln\left(\frac{\text{Actual}_i}{\text{Expected}_i}\right)$$

**PSI interpretation:**

| PSI | Action |
|---|---|
| $< 0.10$ | No significant shift — monitor |
| $0.10 - 0.25$ | Moderate shift — investigate |
| $> 0.25$ | Significant shift — model rebuild likely needed |

**Example:** If actual proportion $= 0.12$ and expected $= 0.10$ for one bin, contribution $= (0.12 - 0.10) \times \ln(0.12/0.10) = 0.02 \times 0.182 = 0.0036$. Sum across all bins; PSI $< 0.10$ suggests stable population.

**Common causes of drift:**

- Economic changes (recession, pandemic)
- Policy changes (new lending criteria attract different applicants)
- Data quality issues (vendor changes, missing fields)

**COVID-19 caused PSI $> 0.25$ for many consumer credit models, triggering widespread model recalibration.**

# The FinTech Credit Scoring Stack

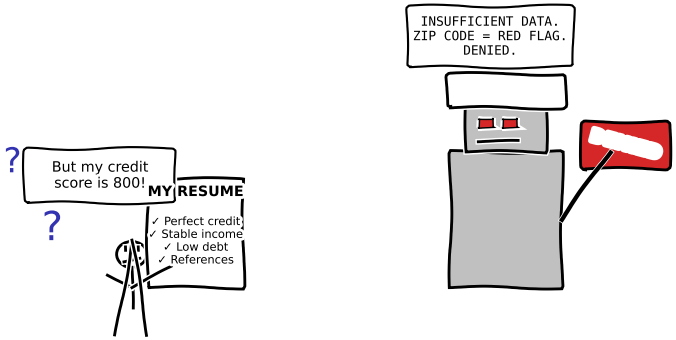
**How FinTech lenders build scoring differently from traditional banks:**

| Component | Traditional Bank | FinTech Lender |
|---|---|---|
| Data | Bureau data (credit file) | Bureau + alternative data |
| Features | 15–30 engineered features | 200–2000 auto-generated features |
| Model | Logistic regression | Gradient boosting / ensemble |
| Scoring speed | Batch (hours) | Real-time ($<$1 second) |
| Decisioning | Human review for borderline | Fully automated |
| Monitoring | Quarterly reports | Real-time dashboards |
| Update cycle | Annual rebuild | Continuous retraining |

**Result:** FinTechs can approve loans in **minutes** instead of days, extend credit to thin-file borrowers, and adapt to market changes faster—but face greater model risk and regulatory scrutiny.

**The FinTech scoring stack trades interpretability and stability for speed, inclusion, and accuracy.**

# One Last Thought...




**Credit scoring: When algorithms say "no" and even they cannot explain why.**

**Sometimes the best way to remember a concept is to laugh about it.**

# Key Takeaways

1. A **credit score** compresses borrower risk into a single number predicting probability of default (PD)
2. Traditional **FICO-style scorecards** use WoE, IV, and logistic regression—transparent but limited to credit-bureau data
3. **Basel II/III** requires banks to estimate PD, LGD, and EAD; credit scoring primarily targets PD
4. **Discrimination metrics** (AUC/Gini, KS) measure ranking power; **calibration** checks that predicted PDs are accurate
5. **Gradient boosting** outperforms logistic regression by 2–5 AUC points but sacrifices interpretability
6. **Alternative data** (telecom, utility, digital footprint) can score thin-file borrowers but introduces privacy and fairness risks
7. **Explainability** is legally required (ECOA, FCRA, EU AI Act); SHAP and surrogate models bridge the gap
8. **Model validation** is continuous: PSI monitors drift, and models must be rebuilt when populations shift

---

**Credit scoring is where financial inclusion meets machine learning meets regulation—getting it right matters.**

# Summary / Next Lesson Preview

**This lesson:** We traced credit scoring from FICO-style scorecards through logistic regression to ML models, examined alternative data for thin-file borrowers, and explored the fairness–accuracy–interpretability triangle.

**Key vocabulary:**

- Credit scorecard
- Weight of Evidence (WoE)
- Information Value (IV)
- PD, LGD, EAD
- Logistic regression
- Gradient boosting
- ROC / AUC / Gini / KS
- Alternative data
- Thin-file borrower
- Calibration
- PSI (Population Stability Index)
- Adverse action notice

**Next lesson (M2L3):** *Microfinance and Digital Lending* — We will examine how FinTech lenders operationalise the credit scoring models from this lesson to deliver micro-loans in minutes, exploring the business models, unit economics, and regulatory challenges of digital lending platforms.

**Review: Can you explain the difference between discrimination (AUC) and calibration, and why both matter?**