Model Risk and Explainability

Theme II: Algorithmic Finance

Research Question: How can we ensure AI/ML models in finance are reliable, interpretable, and compliant with regulations?

PhD Seminar in Digital Finance

Model Risk Management: SR 11-7 Framework

Federal Reserve SR 11-7 (2011)
Model risk arises from:

1. **Specification error**: Wrong model
2. **Implementation error**: Coding bugs
3. **Misuse**: Wrong application

Three Lines of Defense

- 1st: Model developers
- 2nd: Model validation (independent)
- 3rd: Internal audit

Key Requirements
Model Inventory

- All models documented
- Risk-tiered classification
- Owner accountability

Validation Standards

- Conceptual soundness
- Ongoing monitoring
- Outcomes analysis
- Benchmarking

ML Challenge
How to validate black boxes?

Fed SR 11-7 (2011), "Guidance on Model Risk Management" – still the US standard

Concept Drift and Model Degradation

Types of Drift
Covariate Drift

$$P_{train}(X) \neq P_{test}(X)$$

Feature distributions change.
Concept Drift

$$P_{train}(Y|X) \neq P_{test}(Y|X)$$

Relationship changes.
Label Drift

$$P_{train}(Y) \neq P_{test}(Y)$$

Outcome base rate changes.

Detection Methods

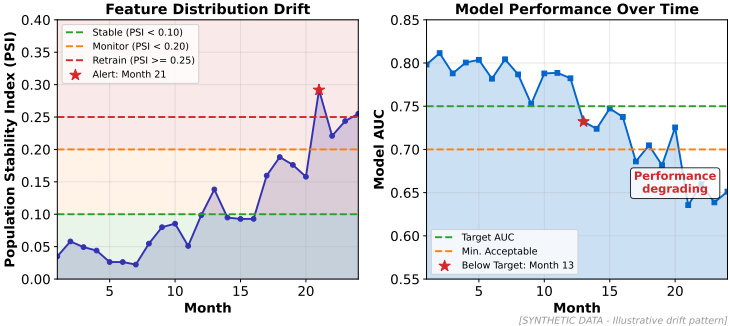
Method	Detects
KS statistic	Covariate drift
PSI	Distribution shift
Performance decay	Concept drift
Prediction stability	All types

Finance Examples

- COVID-19: Massive concept drift
- Interest rates: Covariate shift
- New products: Label drift

Population Stability Index (PSI) is standard for monitoring; threshold typically 0.25

Concept Drift Monitoring Dashboard



PSI and KS statistics detect distribution shifts; thresholds trigger retraining or investigation.

Explainability Methods: SHAP and LIME

SHAP (Shapley Values)
Attribution based on game theory:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

Properties

- Efficiency: $\sum_i \phi_i = f(x) - \mathbb{E}[f]$
- Symmetry: Equal contribution $\rightarrow$ equal attribution
- Null: Zero contribution $\rightarrow$ zero attribution

LIME (Local Interpretable)
Fit local linear model:

$$g(x') = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

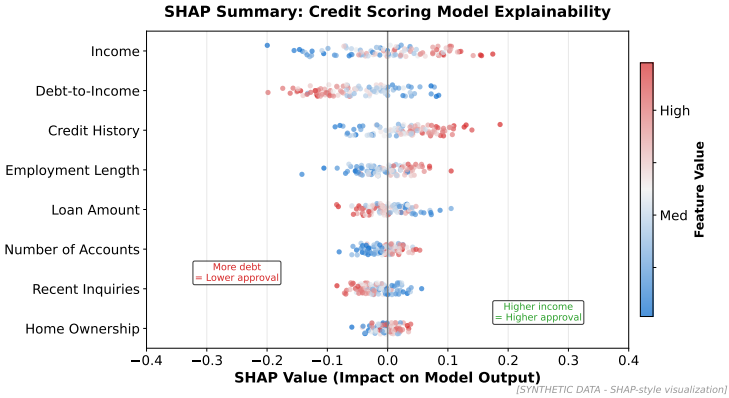
where π_x weights by proximity to x.

Comparison

	SHAP	LIME
Consistency	Yes	No
Speed	Slow	Fast
Local	Yes	Yes
Global	Yes	No

Lundberg & Lee (2017), "A Unified Approach to Interpreting Model Predictions"

SHAP Summary: Feature Impact Visualization



Beeswarm plots show feature importance and direction of impact for each prediction.

Rudin's Critique: Interpretable vs. Explainable

Rudin (2019) Argument
"Stop explaining black box models for high-stakes decisions."

Key Claims

1. Explanations can be unfaithful
2. Interpretable models often equally accurate
3. Explanations create false confidence

Evidence
COMPAS recidivism: Simple rules match accuracy of proprietary black box.

Counter-Arguments

1. Complex patterns require complex models
2. Interpretable models also have risks
3. Explanations better than nothing

Finance-Specific View
Pro-interpretable:

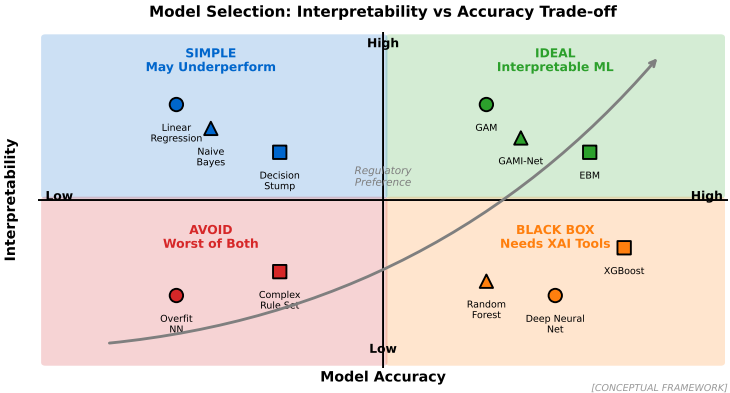
- Adverse action notices (ECOA)
- Regulatory examination
- Audit trails

Pro-complex:

- Fraud detection (adversarial)
- High-dimensional data
- Non-linear relationships

Rudin (2019), "Stop Explaining Black Box Models," Nature Machine Intelligence

The Interpretability-Accuracy Trade-Off



Rudin argues the top-right quadrant (interpretable + accurate) is more attainable than assumed.

Adverse Action Notices and Right to Explanation

US Legal Requirements
ECOA/Regulation B

- Must provide reasons for denial
- "Principal reasons" (4 factors)
- Consumer must understand

FCRA

- Disclose score factors
- Consumer can dispute
- Access to credit file

Implementation Challenges
SHAP-Based Reasons

- Top-k negative contributors
- But: May not be causal
- Consumer cannot always act on them

Example Problem
SHAP says: "ZIP code hurt your score."
Consumer: "I can't change my ZIP code."
Actionability Requirement
Explanations should suggest remediation.

CFPB guidance requires "actionable" adverse action reasons

EU AI Act: High-Risk AI Requirements

Credit Scoring = High-Risk
Article 6 Annex III classification.
Requirements (Article 9-15)

1. Risk management system
2. Data governance
3. Technical documentation
4. Record-keeping
5. Transparency
6. Human oversight
7. Accuracy, robustness

Practical Implications
Conformity Assessment

- Self-assessment (credit)
- Documentation package
- CE marking (for EU market)

Penalties
Up to 35M EUR or 7% global revenue.
Timeline

- 2024: Act enters force
- 2025: Prohibited practices
- 2026: High-risk obligations

EU AI Act (2024) – credit scoring explicitly listed in Annex III high-risk category

Open Research Questions

Publishable Research Directions

1. **Explanation Faithfulness**
   - RQ: Do post-hoc explanations accurately represent model behavior?
   - Method: Compare SHAP/LIME to true model mechanics (synthetic data)
   - Gap: Limited understanding of when explanations mislead
2. **Drift Detection for Financial Models**
   - RQ: What is the optimal monitoring strategy for concept drift?
   - Method: Simulation with known drift patterns, cost-benefit analysis
   - Gap: Industry heuristics (PSI) lack theoretical foundation
3. **Regulatory Compliance Costs**
   - RQ: What are the accuracy costs of interpretability requirements?
   - Method: Compare constrained vs. unconstrained models on real data
   - Gap: Regulators assume small cost; evidence needed

Model risk is practical research area with regulatory demand

Exercises

Mathematical
Derive Shapley values for 3-player game:

$$v(\{1\}) = 1, v(\{2\}) = 2$$

$$v(\{1,2\}) = 4, v(\{3\}) = 0$$

$$v(\{1,3\}) = 2, v(\{2,3\}) = 3$$

$$v(\{1,2,3\}) = 6$$

Compute ϕ_1, ϕ_2, ϕ_3.

Empirical
Using any classification model:

1. Train on lending data
2. Generate SHAP values
3. Create waterfall plot
4. Assess explanation quality

Package: SHAP (Python)

Research Proposal
Draft 1-page proposal:

- "EU AI Act Compliance Costs"
- Survey or archival method
- Sample selection
- Outcome variables

Due: Week 7 – Shapley calculation is tractable for small games

SHAP is standard tool; understanding Shapley theory helps interpret outputs

Required Readings

Core Papers (Read Before Class)

1. **Fed SR 11-7** (2011). "Guidance on Model Risk Management."
   - Focus: Sections II-IV (core framework)
2. **Rudin** (2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions." *Nature Machine Intelligence*, 1, 206-215.
   - Focus: Main argument, healthcare/criminal justice examples

Supplementary

- Lundberg & Lee (2017): SHAP methodology – NeurIPS
- FINMA Guidance 08/2024: Swiss AI governance
- EU AI Act (2024): Legal text, Annex III

SR 11-7 remains foundational; Rudin is provocative counterpoint to SHAP/LIME orthodoxy