SAPIENZA
UNIVERSITÀ DI ROMA

Bridging Black Boxes and Markets: Explainable Deep Learning for Financial Decision Making

Candidate: Fausto Zamparelli 1982378
Thesis Advisor: Prof. Novella Bartolini
Faculty of Information Engineering, Informatics and Statistics
Bachelor in Applied Computer Science and Artificial Intelligence
Academic Year 2024/25

Apple Card
Apple Card with in
cast care noxy.

DHH · Nov 7, 2019
@dhh · **Follow**

The @AppleCard is such a [redacted] sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

Steve Wozniak
@stevewoz · **Follow**

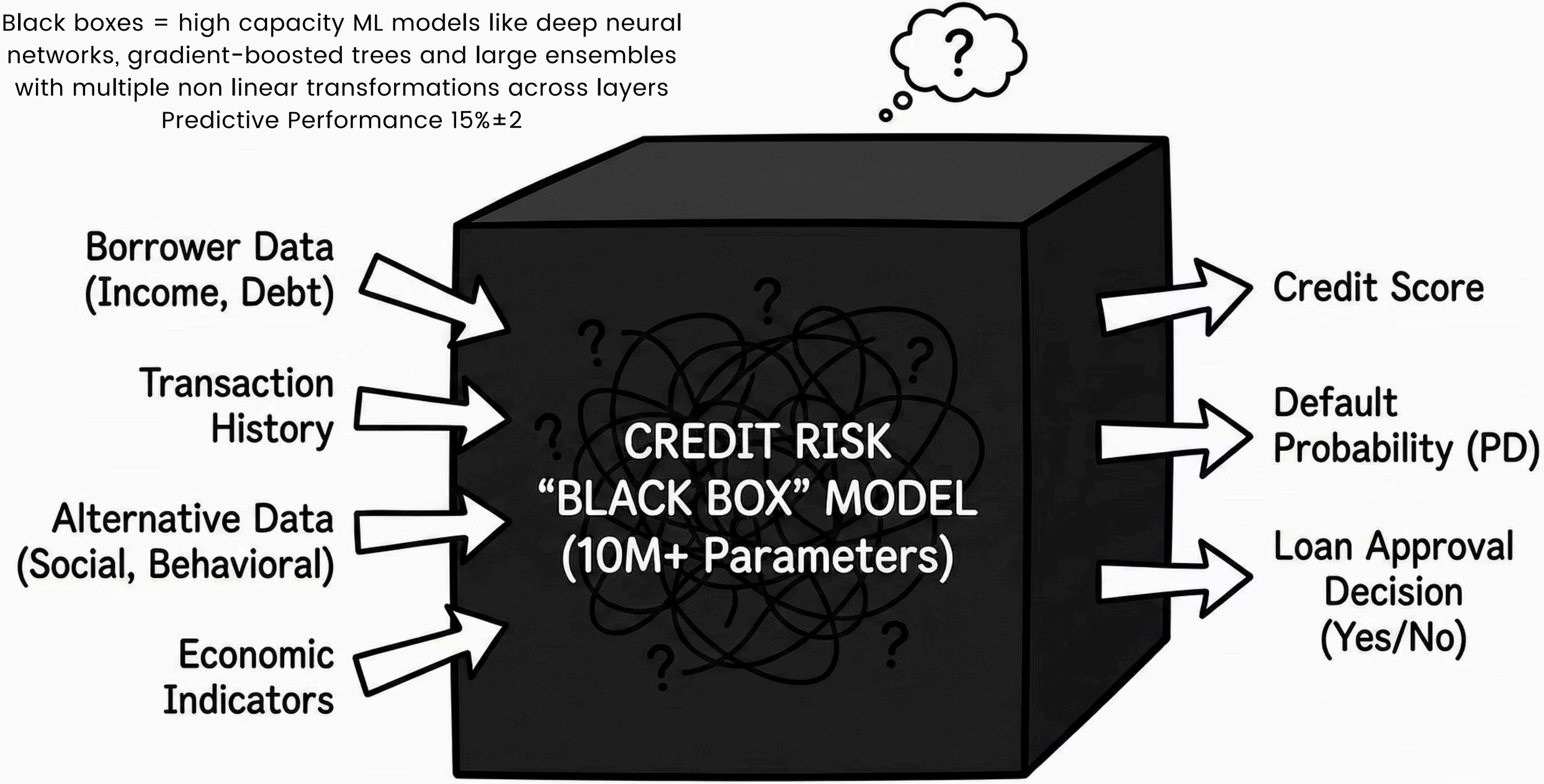
The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

1:51 AM · Nov 10, 2019

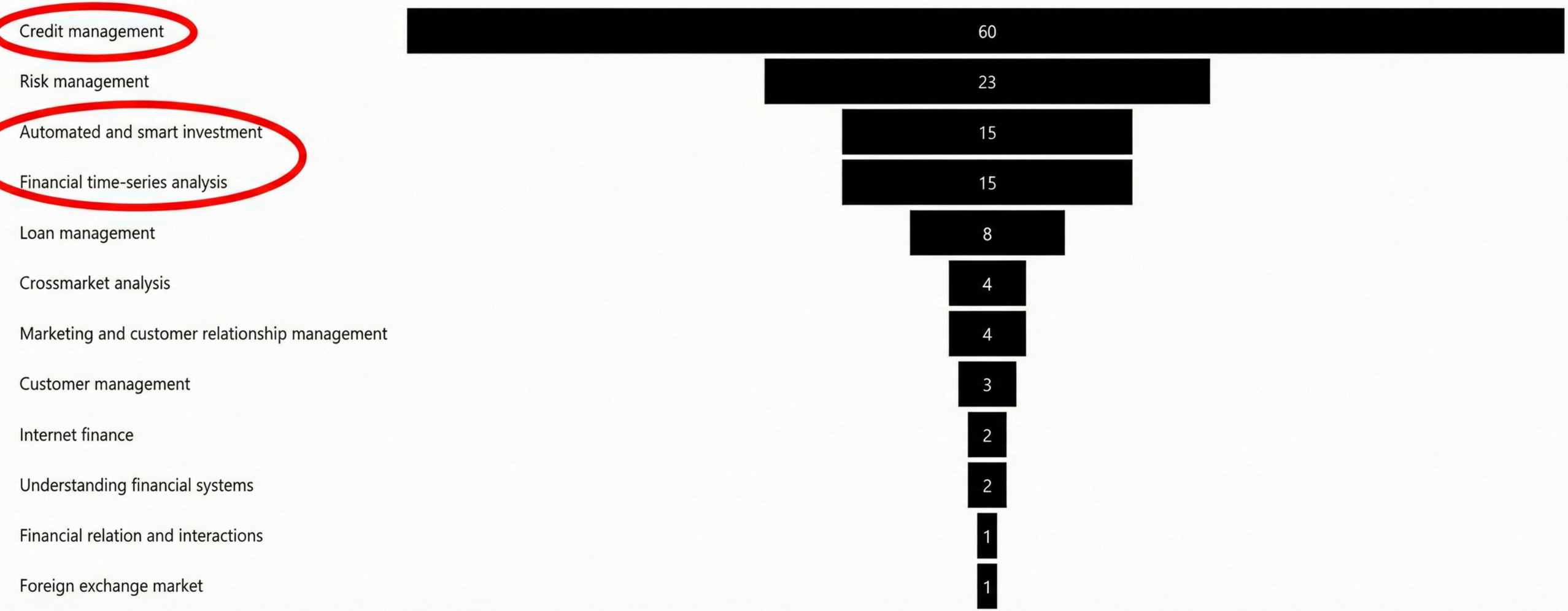
3.1K Copy link

Read 99 replies

Black boxes = high capacity ML models like deep neural networks, gradient-boosted trees and large ensembles with multiple non linear transformations across layers
Predictive Performance 15%±2



In my thesis I explore the tradeoff between performance and explainability focussing on Credit risk and Algorithmic Trading/Market forecasting.



CREDIT RISK:

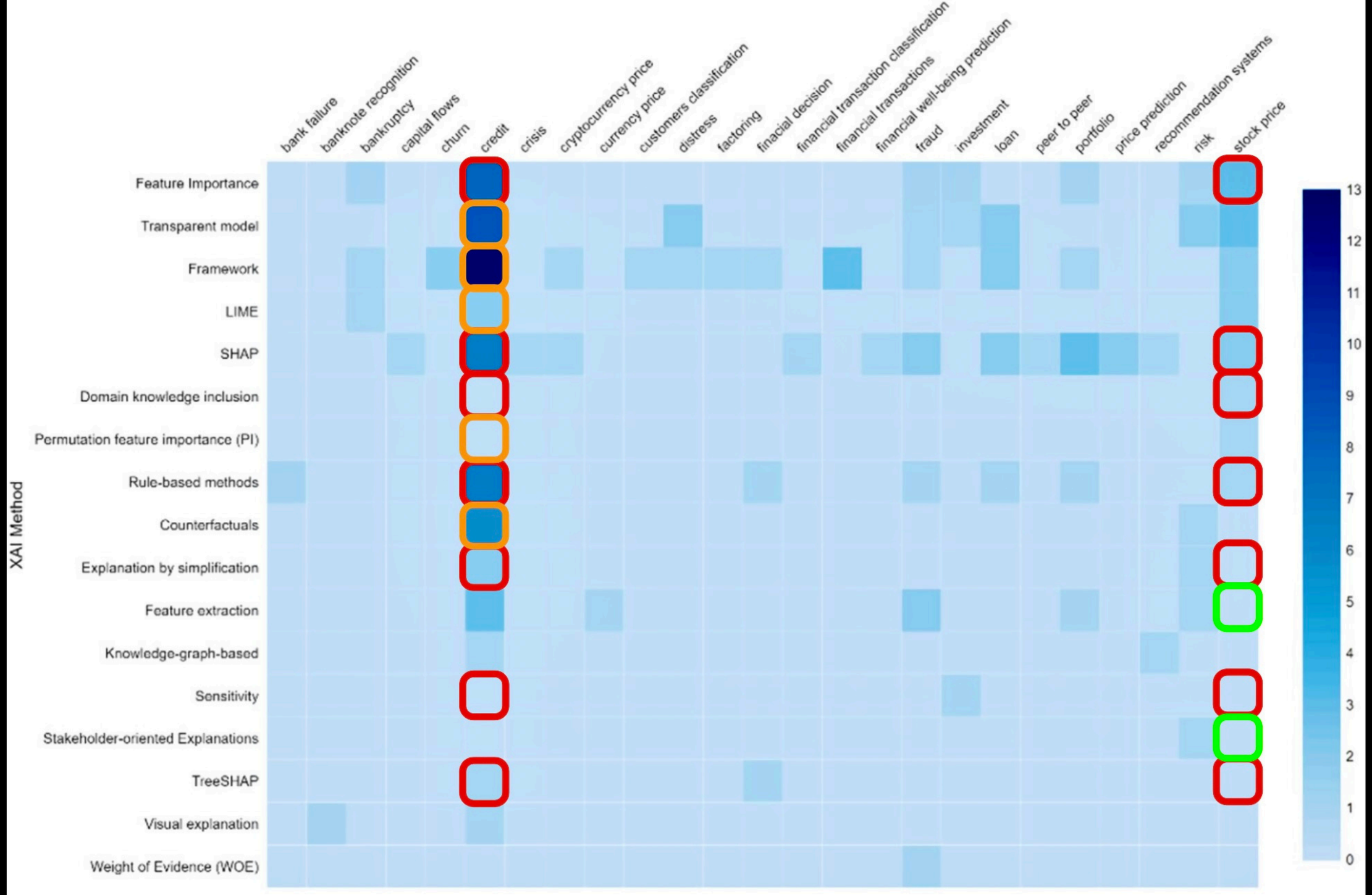
- EU: European Banking Authority Guidelines on PD/LGD (2017), Loan Origination (2020), and ML for IRB Models (2023).
- US: Equal Credit Opportunity Act (ECOA), Regulation B, and CFPB Circulars (2022–2024). What they ask for: These rules legally mandate interpretability. You cannot use a "black box."

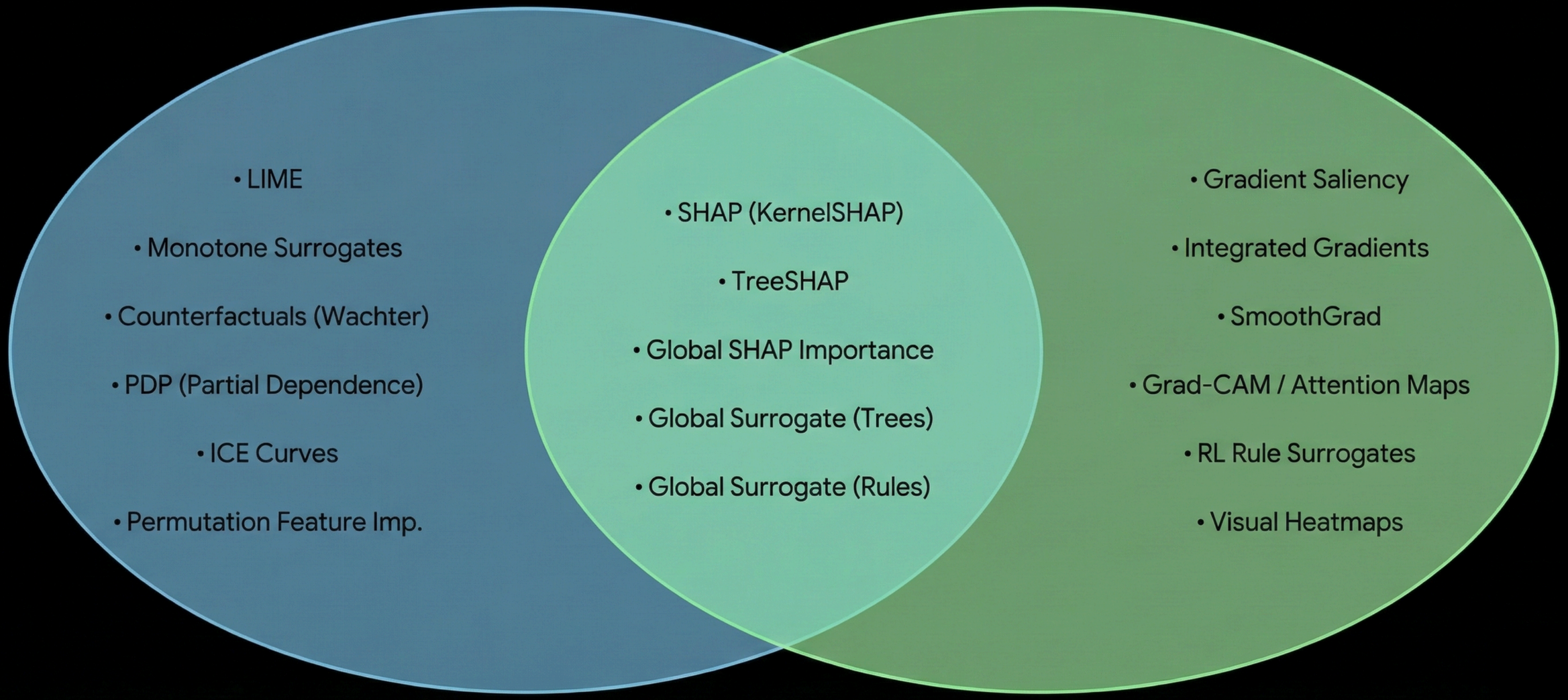
- TO Supervisors: You must demonstrate model stability, traceability, and governance to get capital approval.
- To Customers: You must provide "specific reasons" for any adverse action

ALGORITHMIC TRADING: Xai is an Internal Risk Tool

- Regulations: EU: MiFID II (Art 17), RTS 6, and ESMA Q&A. US: SEC Regulation SCI, Regulation ATS, and FINRA Notice 15-09.
- What they ask for: Pre-trade controls, kill switches, stress testing, and preventing "disorderly markets" (flash crashes).

XAI is discretionary (but recommended). It is used internally to debug strategies, understand P&L drivers, and give portfolio managers the confidence to turn the system on.


bank failure
banknote recognition
bankruptcy
capital flows
churn
credit
crisis
cryptocurrency price
currency price
customers classification
distress
factoring
finacial decision
financial transaction classification
financial transactions
financial well-being prediction
fraud
investment
loan
peer to peer
portfolio
price prediction
recommendation systems
risk
stock price
XAI Method
Feature Importance
Transparent model
Framework
LIME
SHAP
Domain knowledge inclusion
Permutation feature importance (PI)
Rule-based methods
Counterfactuals
Explanation by simplification
Feature extraction
Knowledge-graph-based
Sensitivity
Stakeholder-oriented Explanations
TreeSHAP
Visual explanation
Weight of Evidence (WOE)
13
12
11
10
9
8
7
6
5
4
3
2
1
0


Credit Risk
Trading
• LIME
• Monotone Surrogates
• Counterfactuals (Wachter)
• PDP (Partial Dependence)
• ICE Curves
• Permutation Feature Imp.
• SHAP (KernelSHAP)
• TreeSHAP
• Global SHAP Importance
• Global Surrogate (Trees)
• Global Surrogate (Rules)
• Gradient Saliency
• Integrated Gradients
• SmoothGrad
• Grad-CAM / Attention Maps
• RL Rule Surrogates
• Visual Heatmaps

CREDIT RISK

$$\mathrm{EL}_i = \mathrm{PD}_i \times \mathrm{LGD}_i \times \mathrm{EAD}_i$$

- LGD (loss given default) = fraction of the exposure that the bank expects to lose if default happens, after selling collateral, recoveries, it lives in [0,1].
- EAD (exposure at default) = how large the position is at the moment of default, in euro,
- PD = the probability that borrower i will default in a given horizon, typically 12 months.

€100k loan for a house (80k house + 20k renovations), already repaid 10k.
EAD = 90k,
LGD = (90k-60k)/90k = 0.33
PD = 0.03.
EL = 0.03 x 0.33 x 90000 = 900

$$\mathrm{EL}_{\mathrm{portfolio}} = \sum_i \mathrm{EL}_i.$$

→ Interest Rates (cost of funding, operating cost, profit)

→ Provisioning

→ Unexpected Loss

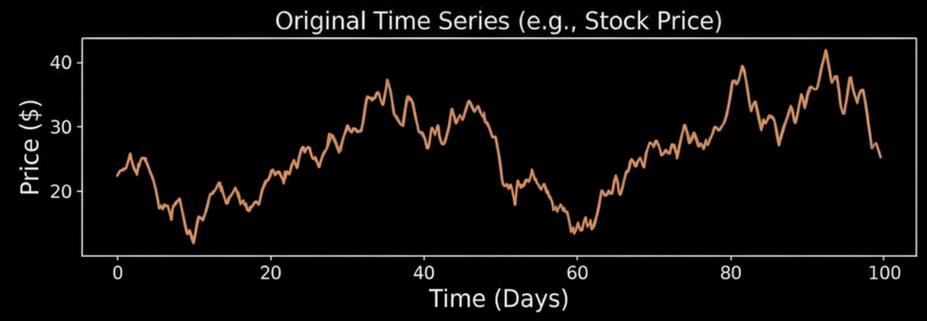
Through the rest of the thesis I mainly study the models that predict these three building blocks of expected loss. All of them are supervised learning models: you give the model past examples of borrowers, with their features and outcomes, and it learns a mapping from features to PD, LGD or EAD.

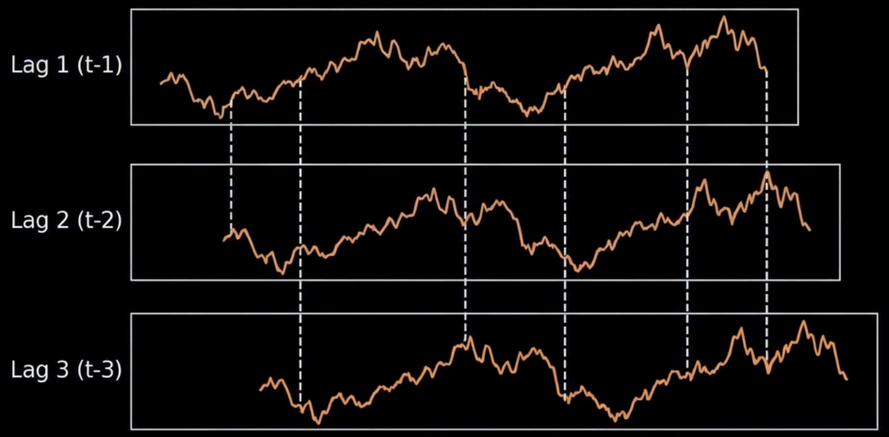
$$D = (x_i, y_i)_{i=1}^n$$

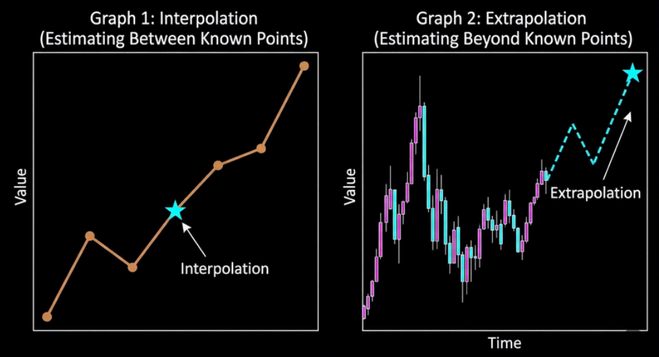
Credit Risk Data Structure: Classification & Regression Targets

	Contract Data			Financials		Behavioral		Macro		Prediction Targets		
Borrower ID	Product Type	Loan Amt (€)	LTV	DTI	Income (€)	Delinq (12m)	Score	Region	Unemp %	Default (Class)	Loss (Reg)	
B-1001	Credit Card	5,000	0.00	0.35	45,000	0	720	North	5.0%	0	0	
B-1002	Mortgage	250,000	0.85	0.42	85,000	0	780	Centre	6.0%	0	0	
x_i (Inputs) B-1003	Personal Loan	15,000	0.00	0.20	32,000	1	610	South	12.0%	1	8,500	y_i (Targets)
B-1004	Credit Card	2,000	0.00	0.60	28,000	2	540	North	5.0%	1	1,200	
B-1005	Mortgage	180,000	0.70	0.25	95,000	0	800	South	12.0%	0	0	

TIME SERIES







$$PnL = \sum_{t=0}^{T-1} q\Delta p_{t+1} - TC_T$$

- p_t: mid-price of the asset at time t.
- q_t: position (how many units you hold) between time t and t
- $\Delta p_{t+1} = p_{t+1} - p_t$: price change over $[t, t+1]$.
- TC_T: total transaction costs and market impact over the wh
  - Times: $t = 0, 1, 2$ (so $T = 2$).
  - Prices: $p_0 = 100, p_1 = 102, p_2 = 101$.
  - You buy 10 shares at $t = 0$ and hold them until $t = 2$. So:
    - $q_0 = 10$ between 0 and 1.
    - $q_1 = 10$ between 1 and 2.
  - Assume transaction costs $\mathrm{TC}_2 = 1$ (total cost of entering/exiting).

$$\mathrm{PnL}_2 = 10 \cdot 2 + 10 \cdot (-1) - 1 = 20 - 10 - 1 = 9.$$

Forecast-and-Act (SL)

Predict a future quantity:

$$f(x_t) \approx \mathbb{E}[Y_{t+H} \mid x_t].$$

- x_t: features at time t (window of past prices, indicators, etc.)
- Y_{t+H}: some future target at horizon $t + H$.

Typical choices for Y_{t+H}:

- Next return: $Y_{t+H} = \frac{p_{t+H} - p_t}{p_t}$.
- Binary direction: $Y_{t+H} = \mathbf{1}\{p_{t+H} > p_t\}$.
- Future volatility.
- Future order-flow imbalance.

Convert forecast into a position:

Once we have a forecast $f(x_t)$, we apply a trading rule, for example:

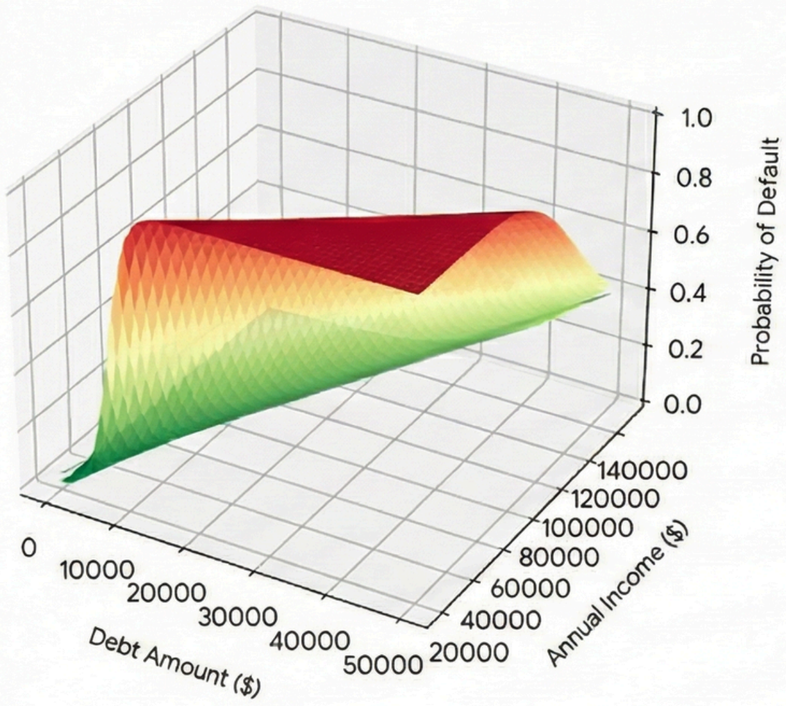
- Regression case (forecasting returns):
  - If $f(x_t) > \alpha$: go long.
  - If $f(x_t) < -\alpha$: go short.
  - Otherwise: flat.
- Probability case (forecasting up probability):
  - If $\mathbb{P}(\mathrm{up} \mid x_t) > 0.6$: long.
  - If $\mathbb{P}(\mathrm{up} \mid x_t) < 0.4$: short.

Direct Policy Learning (RL)

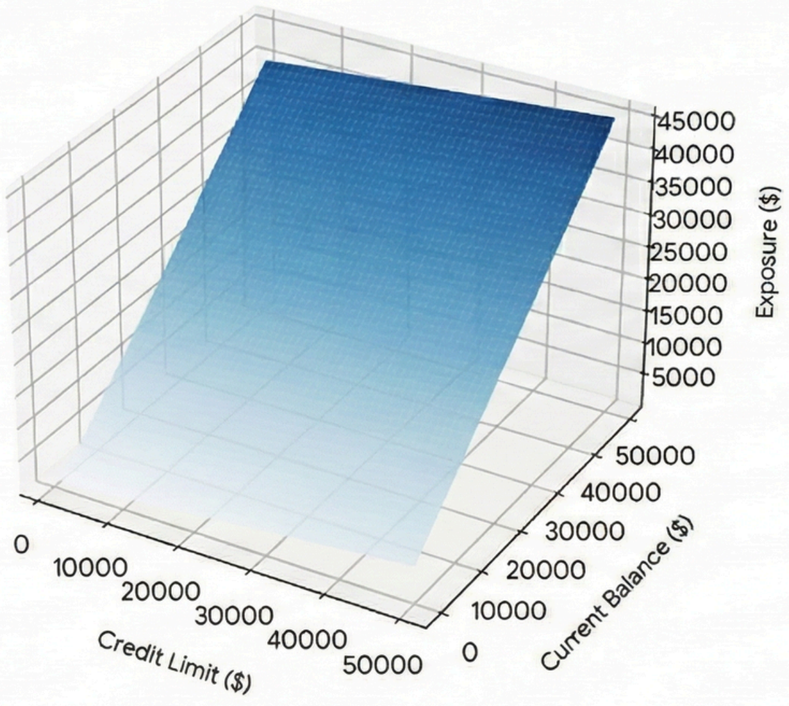
$$\max_{\pi} \mathbb{E}[\mathrm{PnL}_T^{\pi}].$$

Credit Risk Models: Classification vs. Regression Assumptions

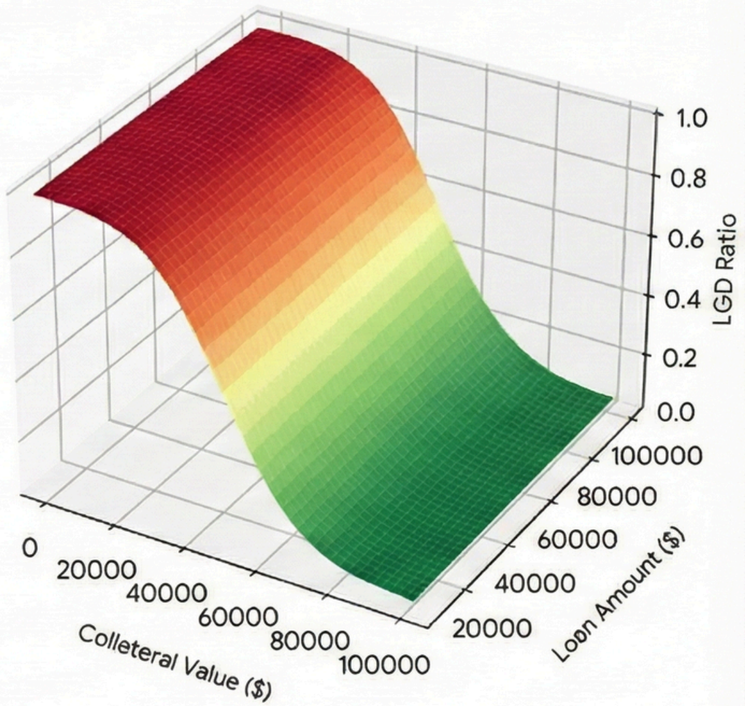
PD: Binary Classification (e.g., Logistic Regression)



EAD: Continuous Regression (e.g., Linear Regression)



LGD: Bounded Regression (e.g., Beta Regression)



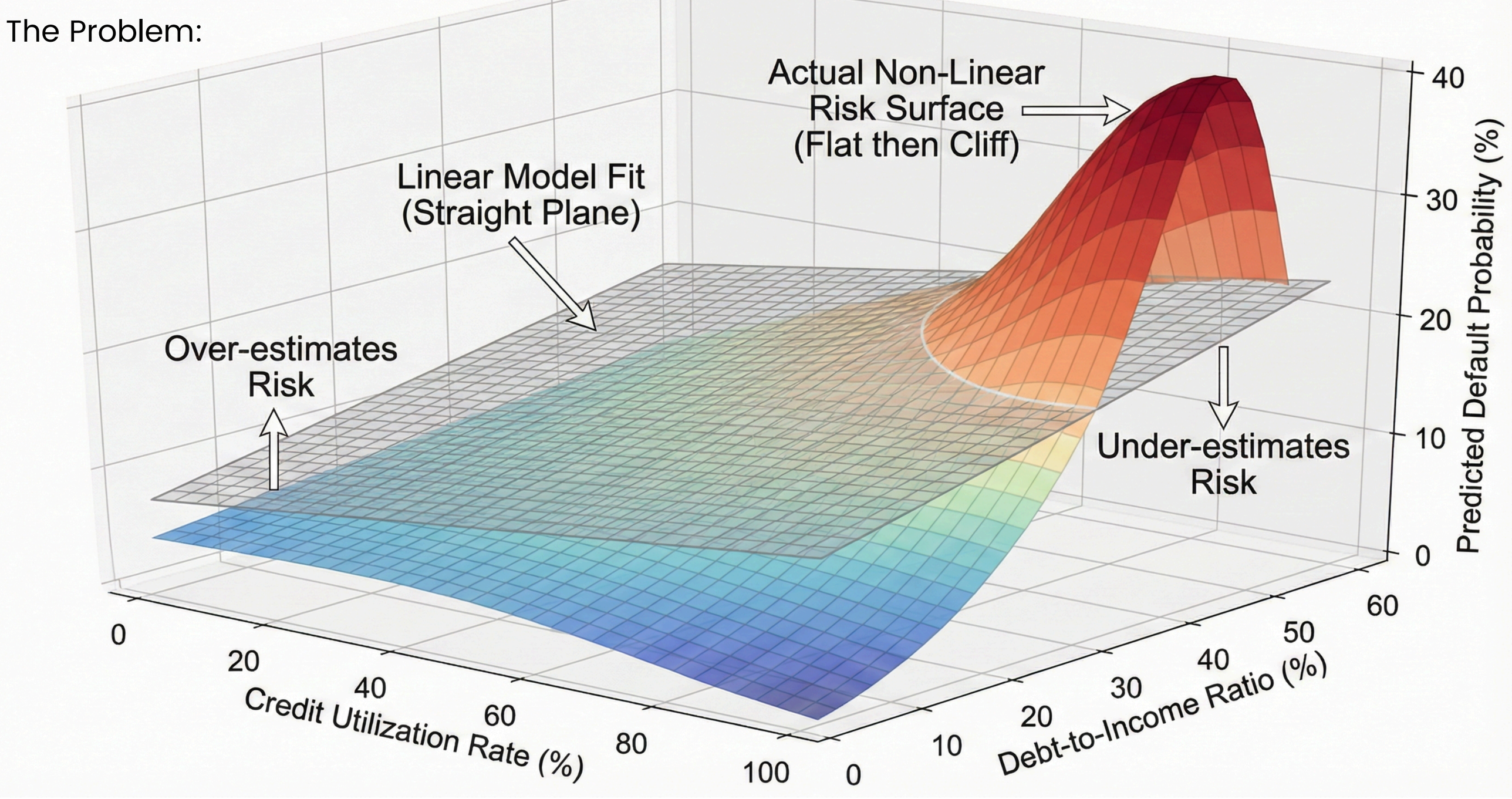
Note: These models operate on multidimensional feature spaces $f(X_1, X_2, \ldots, X_i)$ simultaneously.

$$\text{PD:} \quad y_i^{\text{PD}} \in \{0,1\}, \quad \hat{p}^{\text{PD}}(x_i) = f_{\text{PD}}(x_i), \qquad \hat{y}^{\text{PD}}(x_i) = \mathbf{1}\{\hat{p}^{\text{PD}}(x_i) \geq \tau\},$$

$$\text{LGD:} \quad y_i^{\text{LGD}} \in [0,1], \quad \hat{y}^{\text{LGD}}(x_i) = f_{\text{LGD}}(x_i),$$

$$\text{EAD:} \quad y_i^{\text{EAD}} \geq 0, \qquad \hat{y}^{\text{EAD}}(x_i) = f_{\text{EAD}}(x_i).$$

The Problem:



Solution: Challenger strategy (champion vs challenger) / segmentation

Supervised Learning Objective:

$$\hat{\theta} \;=\; \arg\min_{\theta} \; \frac{1}{n}\sum_{i=1}^{n} \ell(y_i,\; f_\theta(x_i)) \;+\; \lambda\,\Omega(\theta),$$

INSTRINSICALLY TRANSPARENT MODELS

Linear Regression

$$\hat{y}^{\mathrm{LGD}}(x) = \alpha_0 + \sum_{j=1}^{p} \alpha_j x_j,$$

$$\ell_{\mathrm{reg}}(y_i,\, \hat{y}_i) \;=\; (y_i - \hat{y}_i)^2, \quad \hat{y}_i = f_\theta(x_i),$$

Logistic Regression

$$\log \frac{\Pr(y^{\mathrm{PD}} = 1 \mid x)}{\Pr(y^{\mathrm{PD}} = 0 \mid x)} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

$$\hat{p}^{\mathrm{PD}}(x) = \sigma\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right), \quad \sigma(z) = \frac{1}{1+e^{-z}}.$$

$$\ell_{\mathrm{PD}}(y_i^{\mathrm{PD}},\, \hat{p}_i) \;=\; -\left(y_i^{\mathrm{PD}} \log \hat{p}_i \;+\; (1 - y_i^{\mathrm{PD}}) \log(1 - \hat{p}_i)\right), \quad \hat{p}_i = f_\theta(x_i).$$

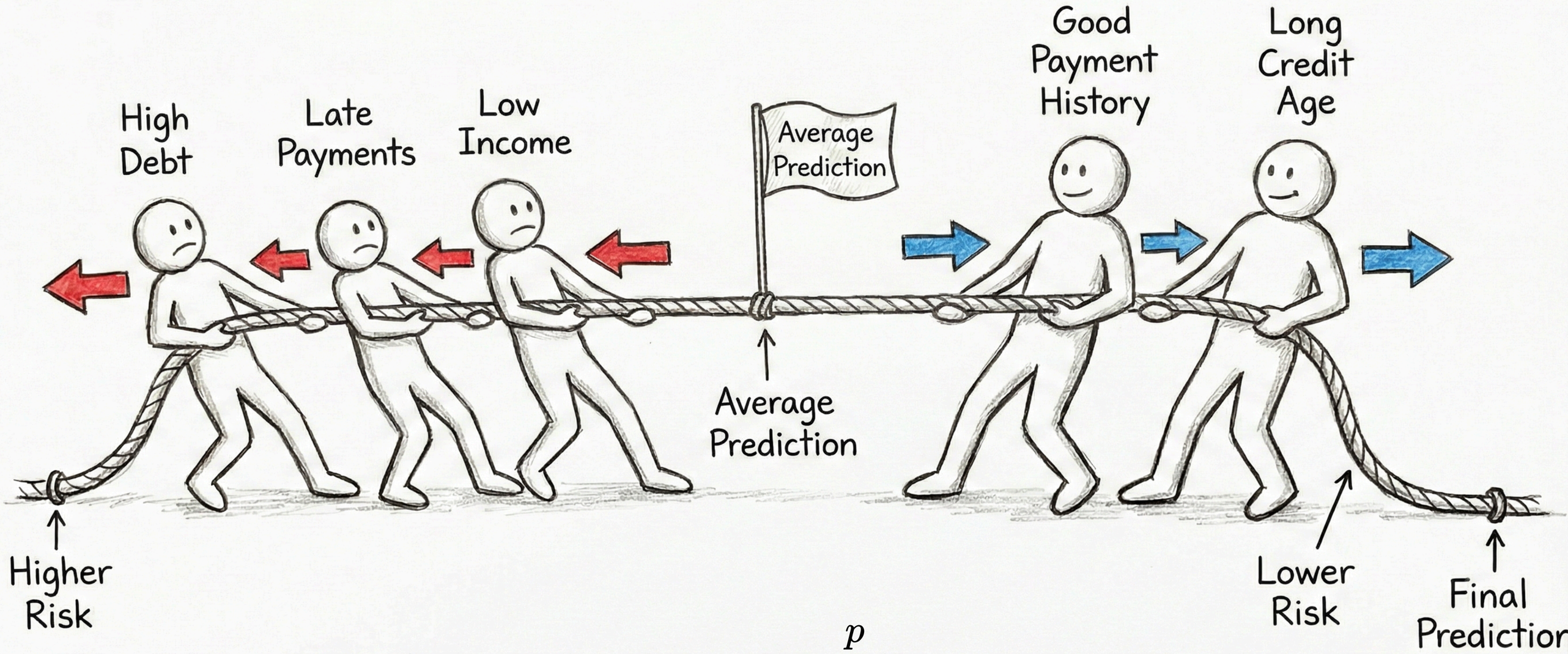
XAI METHODS

Intrinsic (Ante-hoc)
Post-hoc (Model Agnostic / Dependent)

- Global: “which features matter overall for the model” → sanity checks, checking bias against protected classes, guide feature selection (remove noise)
- Local: “which features matter the most for this single instance” → generate adverse action notice, help human underwriters review edge cases, what is the lest friction change a customer can do to flip its prediction (counterfactuals)

SHAP

SHAP



$$f(x) \approx \phi_0 + \sum_{j=1}^{p} \phi_j(x),$$

Let F = {1, ..., p} be the features set .

A permutation/coalition is one possible order in which features could enter the model

For a given permutation π and a feature j, define:

- The set of predecessors of j in that order:

$$\mathrm{pred}_\pi(j) = \{\text{features that appear before } j \text{ in } \pi\}$$

- The model's prediction when only those predecessors are known:

$$f_{\mathrm{pred}_n(j)}\left(x_{\mathrm{pred}_n(j)}\right)$$

- The mode's prediction when we also know feature j:

$$f_{\mathrm{pred}_\pi(\mathrm{j})\cup\{\mathrm{j}\}}\left(x_{\mathrm{pred}_\pi(\mathrm{j})\cup\{\mathrm{j}\}}\right)$$

The marginal contribution of feature j in this particular order is:

$$\Delta_j^\pi(x) = f_{\mathrm{pred}_\pi(j)\cup\{j\}}\left(x_{\mathrm{pred}_\pi(j)\cup\{j\}}\right) - f_{\mathrm{pred}_\pi(j)}\left(x_{\mathrm{pred}_\pi(j)}\right)$$

$$\phi_j(x) = \frac{1}{p!}\sum_\pi \Delta_j^\pi(x).$$

Example

- If we know nothing:
  $f_\emptyset = 0.02$ (baseline PD 2%)
- If we only know one feature:

$$f_{\{1\}}(x_1) = 0.03, \quad f_{\{2\}}(x_2) = 0.04, \quad f_{\{3\}}(x_3) = 0.06$$

- If we know two features:

$$f_{\{1,2\}}(x_1, x_2) = 0.05, \quad f_{\{1,3\}}(x_1, x_3) = 0.08, \quad f_{\{2,3\}}(x_2, x_3) = 0.09$$

- If we know all three:

$$f_{\{1,2,3\}}(x_1, x_2, x_3) = 0.10.$$

1. $\pi = (1, 2, 3)$
   $\mathrm{pred}_\pi(3) = \{1, 2\}$
   $\Delta_3^\pi = f_{\{1,2,3\}} - f_{\{1,2\}} = 0.10 - 0.05 = 0.05$
2. $\pi = (1, 3, 2)$
   $\mathrm{pred}_\pi(3) = \{1\}$
   $\Delta_3^\pi = f_{\{1,3\}} - f_{\{1\}} = 0.08 - 0.03 = 0.05$
3. $\pi = (2, 1, 3)$
   $\mathrm{pred}_\pi(3) = \{1, 2\}$
   $\Delta_3^\pi = f_{\{1,2,3\}} - f_{\{1,2\}} = 0.10 - 0.05 = 0.05$
4. $\pi = (2, 3, 1)$
   $\mathrm{pred}_\pi(3) = \{2\}$
   $\Delta_3^\pi = f_{\{2,3\}} - f_{\{2\}} = 0.09 - 0.04 = 0.05$
5. $\pi = (3, 1, 2)$
   $\mathrm{pred}_\pi(3) = \emptyset$
   $\Delta_3^\pi = f_{\{3\}} - f_\emptyset = 0.06 - 0.02 = 0.04$
6. $\pi = (3, 2, 1)$
   $\mathrm{pred}_\pi(3) = \emptyset$
   $\Delta_3^\pi = f_{\{3\}} - f_\emptyset = 0.06 - 0.02 = 0.04$

$$\phi_3(x) = \frac{1}{6}(0.05 + 0.05 + 0.05 + 0.05 + 0.04 + 0.04) = \frac{0.28}{6} \approx 0.0467.$$

$$\phi_1(x) \approx 0.0117, \quad \phi_2(x) \approx 0.0217, \quad \phi_3(x) \approx 0.0467.$$

$$\phi_0 + \phi_1(x) + \phi_2(x) + \phi_3(x) \approx 0.02 + 0.0117 + 0.0217 + 0.0467 \approx 0.10 = f(x),$$

Shortcomings:

- exponential coalitions (30 features → 1 billion)
- marginalization (100 sampes = 100x for a subset)

Solutions:

- TreeSHAP (for tabular data): exploits the tree structure to calculate exact values recursively. It pushes calculations down the nodes, reducing complexity from exponential to polynomial.

- DeepSHAP (for NN): adapts DeepLIFT to approximate Shapley values. It "backpropagates" feature contributions from the output layer to the input layer, similar to a gradient pass.

- LinearSHAP (for linear/logistic regression): exploits the additive model definition. It computes the exact attribution analytically (Coefficient x (Feature - Mean}) instantly without any sampling.

- From the table we see that credit risk is basically “won” by SHAP on tree models: TreeSHAP + global SHAP importance have the best mix of fidelity, interpretability, fairness and maturity, so they are the backbone for explaining credit scores.
- Counterfactuals are the best complement when we care about actionability (“what should the client change?”) and fairness checks.
- PDP/ICE and global surrogates (small trees/rules) are useful supporting tools for global understanding, but are secondary compared to SHAP + counterfactuals.

- In trading there is no single perfect method, but the best compromise comes from SHAP (on tree/GBM models) and global surrogate trees/rules: they give reasonably faithful and readable summaries of complex strategies.
- Gradient-based saliency, Integrated Gradients, Grad-CAM are promising for deep and RL models, but in the table they score lower on interpretability, actionability and maturity, so they are still mostly research tools rather than stable, regulatory-grade explanations.

Credit Risk

	Fidelity	Interpretability	Actionability	Fairness	Maturity
LIME	M	H	M	M	M
Monotone Surrogates	M	H	H	M	M
Counterfactuals (Wachter)	M	H	H	H	M
PDP (Partial Dependence)	M	H	M	H	H
ICE Curves	H	M	M	H	M
Permutation Feature Imp.	M	M	L	M	H
SHAP (KernelSHAP)	M	H	M	H	M
TreeSHAP	H	H	M	H	H
Global SHAP Importance	H	H	M	H	H
Global Surrogate (Trees)	M	H	M	M	M
Global Surrogate (Rules)	M	H	H	M	M

Algorithmic Trading

	Fidelity	Interpretability	Actionability	Fairness	Maturity
SHAP (KernelSHAP)	M	M	L	M	M
TreeSHAP	H	M	M	M	M
Global SHAP Importance	H	M	L	M	M
Global Surrogate (Trees)	M	H	M	M	M
Global Surrogate (Rules)	M	H	M	M	M
Gradient Saliency	M	L	L	L	L
Integrated Gradients	H	M	L	L	L
SmoothGrad	M	M	L	L	L
Grad-CAM / Attention Maps	M	M	L	L	L
RL Rule Surrogates	M	M	M	L	L
Visual Heatmaps	L	M	L	L	M

SAPIENZA
UNIVERSITÀ DI ROMA

Thank You For Your Time!