arXiv:2608.09433v1 [cs.LG] 10 Aug 2026

How Simple Can It Get? From Interpretable Equations to Readable Rules for Financial Decision Making

Adia Lumadjeng✉
a.c.lumadjeng@uva.nl
University of Amsterdam
Amsterdam, the Netherlands

Ilker Birbil
s.i.birbil@uva.nl
University of Amsterdam
Amsterdam, the Netherlands

Erman Acar
e.acar@uva.nl
University of Amsterdam
Amsterdam, the Netherlands

Abstract

In regulated domains such as finance, a model that cannot be explained cannot be deployed, yet many interpretable classifiers defeat their own purpose by producing formulas with dozens of features that no regulator could read. We take the reverse direction. Starting from an interpretable classifier expressed as a single equation over the input features, we progressively simplify it into more readable forms, including a pruned monomial, a directional if–then rule, and the integer scorecards and tallies that finance already deploys. Because the equation is itself the predictive model rather than a post-hoc explanation we can directly quantify what is lost under each simplification. Across four financial datasets, we find that pruning is nearly free and that fidelity can erode faster than predictive performance, allowing simpler rules to remain effective classifiers without faithfully reproducing the original model. A human assessment shows that simplification improves perceived readability, while preferences for different representations vary by professional background. Beyond measuring these losses empirically, we show that some can be anticipated from the original model: we derive a bound on the change caused by pruning and predict how faithfully a rule retaining only the direction of each feature's effect preserves the original ranking.

CCS Concepts

• **Computing methodologies** → *Supervised learning*; **Machine learning**; • **Applied computing** → *Business rules*; Business process management.

1 Introduction

When a lender rejects an application or a payment system freezes a transaction, someone is eventually owed a reason. In financial decision-making this is not only good practice but often a regulatory requirement: decisions must be explainable to a customer, an auditor, or a supervisor in terms a person can follow. Yet producing a useful explanation remains difficult. Credit default and fraud are typically imbalanced classification problems, where the events of interest form a small minority of observations. At the same time, interpretability is often treated as a structural property of a model, while the practical question is whether the resulting explanation can actually be understood and used by a human decision maker.

Interpretable-by-design classifiers address part of this problem by making the reasoning accessible from the model itself. One such classifier is ECSEL [16], which learns a *signomial* equation, i.e., a sum of power-law terms, that serves simultaneously as classifier and an explanation. In this work, we focus on its simplest single-term form: a *monomial*, consisting of a product of input features raised to fitted exponents, and passed through a sigmoid to obtain the predicted probability. This monomial is the *interpretable* equation we study throughout this paper. However, transparency by design does not guarantee readability in practice. Trained to predict loan default on a dataset, ECSEL produces the equation in Figure 1(a): a product over all forty-two features. Nothing in it is hidden, yet the resulting explanation is cumbersome to inspect and communicate. The very structure that makes the model transparent can therefore become a source of complexity when the equation grows large.

Figure 1(b) shows the same learned model after simplification into a directional rule over seven retained features. The resulting representation is more compact, but achieves this by discarding information: most features are removed, and the retained exponent magnitudes are replaced by their directions. The two panels therefore expose a tension that model transparency alone does not resolve: making an explanation easier to read requires deciding which parts of an already-interpretable model can be removed.

Learned model

$$\text{risk}(x) = 5.91 \cdot x_{\text{int_received}}^{1.19} x_{\text{last_pymt}}^{-0.85} x_{\text{time_pymnt}}^{0.86} \cdot \ldots \text{(38 more features)}$$

(a) The fitted monomial: a product over all 42 features.

Readable rule

Predict *default* when **total interest received** and **time since last payment** are **high**, **total payment** and **last payment** are **low**,

(b) The same model as a readable rule over seven features.

Figure 1: The same loan-default classifier as the full 42-feature monomial (a) and a simplified seven-feature directional rule (b).

The interpretability literature has largely approached this problem in the forward direction: constructing models that remain transparent while retaining predictive performance [12, 15]. We study the reverse direction. Given an already-interpretable model, *how far can its representation be simplified while preserving predictive performance and fidelity to the original model, while becoming easier for people to understand and use?* This distinction is particularly relevant for interpretable-by-design approaches such as ECSEL. The original work evaluates the fitted model at full size while using pruning to produce more legible displayed equations [16]. It does not evaluate how these simplified representations perform as classifiers or how faithfully they preserve the behavior of the fitted model. We make this transition explicit. Starting from a fitted single-term monomial, we construct controlled simplifications that alter different components of its learned structure: pruning weakly

contributing features, reducing exponent magnitudes to their directions, and replacing continuous feature values by binary conditions. This yields representations ranging from a pruned monomial to a sign-only directional rule and point-based scorecard and tally. Each representation is evaluated as the classifier it displays: the rule shown to the reader is also the rule whose predictions are evaluated. This lets us distinguish its predictive performance, its fidelity to the original model, and its perceived readability and practical preference in our human assessment.

Our contributions are as follows:

(1) **Controlled simplification of an interpretable classifier.** We derive readable rules directly from a learned monomial, with each representation preserving a different subset of the information encoded in the original interpretable equation.
(2) **Quantifying the cost of simplification.** Across financial datasets, we evaluate each derived representation as a classifier and against the original monomial, measuring predictive performance, rank fidelity, and calibration to identify which information can be removed at little cost.
(3) **Predicting fidelity before simplification.** For the directional rule, we derive a model-based prediction of rank fidelity from the fitted exponents and feature covariance, allowing the expected effect of sign-flattening to be assessed before evaluating the resulting rule.
(4) **Assessing human readability.** We conduct a human assessment with finance/risk professionals and AI/ML researchers to determine whether simplification translates into perceived ease and practical preference, and how these differ across professional backgrounds.

Finally, we compare post-hoc pruning with sparsity imposed during training through iterative hard thresholding. This serves as a robustness check on the extracted structure rather than a separate contribution: agreement between the two routes provides complementary evidence that the retained features are not solely an artifact of post-hoc pruning.

2 Related Work

Explainable artificial intelligence has traditionally addressed the lack of transparency of complex machine learning models through post-hoc explanations. Methods such as LIME [18] and SHAP [17] are widely used model-agnostic methods that explain individual predictions without modifying the underlying classifier. In credit scoring, Chen et al. [5] find that their explanations become less stable as class imbalance increases. Because these explanations are produced post hoc, the explanation remains separate from the predictive model itself.

This limitation has motivated a complementary line of research on *inherently interpretable* models, in which the predictor is designed to be transparent. Examples include generalized additive models [15], Bayesian Rule Lists [12], Optimal Classification Trees [2], and CORELS [1], with recent work extending interpretable trees to credit scoring [20]. Survey work distinguishes such ante-hoc approaches from post-hoc explanations [8]. Structural transparency, however, does not ensure readability: rule lists, trees, and additive models become harder to follow as they grow in size, depth, or interaction complexity. ECSEL [16] belongs to this family, representing the classifier through signomial score functions whose exponents admit interpretations. It faces the same tension, since the equation presented to a reader is pruned while the reported model is not.

Credit scoring provides a long-standing setting in which compact model representations are valued. Traditional scorecard development commonly bins characteristics, supporting model comprehensibility and review [19]. Modern approaches learn sparse integer scoring systems, including SLIM [21], RiskSLIM [22], and FasterRisk [14], while Chi and Birbil [6] directly optimize such scores for decision net benefit. Structured lending models have also combined interpretable feature-group scores into an overall risk estimate [4]. Model format and complexity can affect users' accuracy, response time, and confidence [10]. Whereas these approaches learn compact representations directly from data, we derive them from an already fitted interpretable model. ECSEL [16] provides this starting point through its equation-based representation.

3 Approach

We first introduce the monomial classifier that serves as the starting point of our study, then derive a set of readable representations by systematically simplifying the information it contains.

We consider binary classification on tabular data, with features $x \in \mathbb{R}^m$ and labels $y \in \{0, 1\}$, where $y = 1$ denotes the minority risk class (e.g., *default* or *fraud*). Given training data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$, we seek a scoring function that ranks positive instances above negatives and labels them after applying a decision threshold.

The Model. We build on ECSEL [16], which learns a signomial equation as a joint classifier and explanation. A signomial is a sum of K monomial terms, each a product of the features raised to fitted exponents. Because the exponents are real-valued, every feature is scaled to a fixed positive range $[\ell, u]$ with $0 < \ell$ before fitting; this scaling is fit on the training split alone. The model computes a single logit

$$z(x) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{m} x_j^{\beta_{kj}}, \qquad P(y = 1 \mid x) = \sigma\big(z(x)\big), \tag{1}$$

where $\alpha_k \in \mathbb{R}$ are term coefficients, $\beta_{kj} \in \mathbb{R}$ are the fitted exponents, and $\sigma(z) = 1/(1+e^{-z})$ is the sigmoid. All tasks here are binary, so a single logit z and a sigmoid suffice; the multiclass form is not needed, but is given in [16]. Parameters $\{\alpha_k, \beta_{kj}\}$ are learned by gradient descent, with sparsity regularization applied to the exponents. The number of terms K controls model expressivity, ranging from a single monomial (K=1) to richer additive combinations of them ($K > 1$).

The Single Monomial. We restrict the general ECSEL formulation to its simplest case, K=1, yielding:

$$z(x) = \alpha \prod_{j=1}^{m} x_j^{\beta_j} \quad \xrightarrow{\log|\cdot|} \quad \log|z(x)| = \log|\alpha| + \sum_{j=1}^{m} \beta_j \log x_j. \tag{2}$$

In log-space, this monomial is linear: $|\beta_j|$ determines the feature j's strength on the log-score, while the signs of β_j and α determine its direction. For $\alpha > 0$, positive exponents increase predicted risk and negative exponents decrease it; these directions reverse for $\alpha < 0$.

A monomial explanation can therefore be viewed as containing three interpretable components: its *support*, which determines

which features participate; its *direction*, determined by the sign of the exponents together with the sign of α; and its *magnitude*, given by the absolute exponent values. These components provide the basis for the controlled simplifications studied in Section 3.1: each readable representation removes or preserves a different subset of the information contained in the original model.

Why the Monomial. We focus on ECSEL because it is an equation-learning method in which the learned expression serves explicitly as both the classifier and its explanation [16]. Its closed-form equations also support interpretability properties including global feature behavior, decision-boundary analysis, and local feature attribution. ECSEL therefore provides a natural setting for studying the gap between formal interpretability and practical readability, and for measuring what is preserved as an interpretable equation is progressively simplified.

Within ECSEL, we study the single monomial ($K = 1$), its simplest equation form. This isolates the effect of simplifying the learned representation without introducing the additional structure of multiple terms. The same perspective may extend to other interpretable models, while multi-term signomials provide a natural next step within equation learning.

3.1 Controlled Simplification of the Monomial

A learned monomial contains several kinds of information that contribute differently to its interpretation. Throughout this paper, we distinguish four components:

(1) **Feature support:** which features participate in the rule.
(2) **Feature values:** whether features are represented by their continuous values or by binary conditions.
(3) **Direction:** whether increasing a feature increases or decreases the risk, determined by the sign of its exponent.
(4) **Magnitude:** how strongly each feature influences the prediction, determined by the absolute value of its exponent.

These components can be simplified independently, letting us isolate the cost of each simplification rather than treating readability as a single notion. Table 1 summarizes the information retained by each representation.

Table 1: Information retained by each representation.

Form	Support	Values	Direction	Magnitude
Full monomial	All	Cont.	✓	Exact
Pruned monomial	Reduced	Cont.	✓	Exact
Directional rule	Reduced	Cont.	✓	–
Scorecard	Reduced	Binary	✓	Quantized
Tally	Reduced	Binary	✓	–

3.1.1 Controlling Feature Support: the Pruned Monomial. The first simplification reduces the feature support while leaving the remaining components unchanged: we retain r features and set the remaining exponents to zero, so the pruned form is still a monomial over continuous values, with the same directions and the same exact magnitudes for the retained features. Only the number of participating features changes (Table 1).

This raises the question of which features to drop. We prune by exponent magnitude, keeping the largest $|\beta_j|$. In the log-space form of Eq. 2, each feature contributes an additive term $\beta_j \log x_j$, so pruning by $|\beta_j|$ removes the terms with the smallest coefficients in absolute value. Because the features are scaled to a bounded positive range, each deleted term can be bounded in terms of its exponent, yielding a bound on the change in the log-score.

Remark 1 (Worst-case perturbation under pruning). *Let S denote the set of features removed by pruning. Define the full and pruned log-scores as*

$$s_{\text{full}}(x) := \log|\alpha| + \sum_{j=1}^{m} \beta_j \log x_j, \quad s_{\text{pruned}}(x) := \log|\alpha| + \sum_{j \notin S} \beta_j \log x_j.$$

With features scaled to $[\ell, u]$, $0 < \ell$, and $L := \max\{|\log \ell|, |\log u|\}$, the log-scores before and after pruning satisfy

$$\left|s_{\text{full}}(x) - s_{\text{pruned}}(x)\right| = \left|\sum_{j \in S} \beta_j \log x_j\right| \le L \sum_{j \in S} |\beta_j|, \qquad \text{for every } x.$$

The bound follows from the triangle inequality and $|\log x_j| \le L$. Its content is a selection criterion rather than a performance guarantee: the right-hand side depends on the discarded exponents only through $\sum_{j \in S} |\beta_j|$, so among all ways of discarding $m - r$ features, dropping the smallest magnitudes minimizes the worst-case perturbation of the log-score. This motivates exponent magnitude as the pruning criterion, but does not guarantee preservation of predictive performance or ranking. We therefore choose r on validation data as the smallest value whose PR-AUC remains within a fixed tolerance of the full monomial (see Section 4).

3.1.2 Binarizing Feature Values: Scorecard and Tally Rules. The second simplification changes how the retained features are represented. The pruned monomial preserves the continuous feature values, whereas scorecards replace each feature with a binary risk condition. Continuous measurements are therefore reduced to a single yes/no statement, while the feature support and the direction of each effect are preserved. This follows the logic of traditional credit and clinical scoring systems, where an applicant accumulates points or meets conditions until a decision threshold is reached.

Each retained feature is converted into a binary condition with cutpoint c_j, set to the feature's training-set median (the midpoint for binary features). The signs of β_j and α determine whether larger or smaller values indicate greater risk, and thus whether the condition is $x_j \ge c_j$ or $x_j \le c_j$. These thresholds are fixed rather than optimized to isolate the effect of replacing continuous values by binary conditions rather than learning a new scorecard. Cutpoint selection is a separate scorecard-modelling choice in credit risk [19].

From these binary conditions we derive two additive representations. The first is the *tally rule*, which treats every condition equally and counts how many are satisfied. Let $\mathbf{1}_j(x)$ denote the indicator that instance x satisfies condition j. The tally rule is given by

$$T(x) = \sum_{j=1}^{r} \mathbf{1}_j(x), \qquad \text{flag when } T(x) \ge t. \tag{3}$$

where t is the threshold on the number of satisfied conditions.

The second is the *scorecard*, which retains a coarse notion of feature importance by assigning each condition an integer point value. Exponent magnitudes are rescaled so that the strongest retained

condition receives $P_{\max}$ points, with every retained condition worth at least one:

$$p_j = \max\left\{1,\ \text{round}\left(P_{\max}\frac{|\beta_j|}{\max_k |\beta_k|}\right)\right\}. \tag{4}$$

The floor keeps every feature in the support visible on the card: without it, a condition whose exponent is small relative to the largest would be assigned zero points and effectively removed from the scorecard. The resulting score is

$$V(x) = \sum_{j=1}^{r} p_j\, \mathbf{1}_j(x), \qquad \text{flag when } V(x) \geq v. \tag{5}$$

where v is the threshold on the total number of points.

The two representations differ only in how much magnitude information they retain: the tally is the scorecard with every point value set to one. The scorecard preserves relative importance approximately through integer points, whereas the tally removes it altogether. Both, however, discard the continuous feature values that remain available to the pruned monomial.

Example 1 (Scoring an Applicant.). *Suppose pruning leaves the three-feature monomial*

$$z(x) = c\, x_{\text{utilization}}^{+1.4} x_{\text{months_employed}}^{-0.9} x_{\text{age}}^{-0.4}, \qquad c > 0, \tag{6}$$

for predicting credit-card default. The signs give three conditions: high utilization, low employment, and low age. Consider an applicant with high utilization, low employment, and not low age. The corresponding indicators are therefore

$$(\mathbf{1}_{\text{utilization}}, \mathbf{1}_{\text{months_employed}}, \mathbf{1}_{\text{age}}) = (1, 1, 0).$$

With $P_{\max} = 5$, rescaling the exponent magnitudes $(1.4, 0.9, 0.4)$ relative to the largest gives scorecard points $(5, 3, 1)$. The two representations therefore score the applicant

$$T(x) = 1 + 1 + 0 = 2 \geq t = 2, \tag{7}$$

$$V(x) = 5(1) + 3(1) + 1(0) = 8 \geq v = 6. \tag{8}$$

Both rules therefore flag the applicant as risky. They need not agree in general: the tally counts every satisfied condition equally, whereas the scorecard weights them by their relative importance.

3.1.3 Removing Effect Magnitudes: The Directional Rule. The final simplification removes only the effect magnitudes while leaving the remaining components unchanged. Unlike the scorecard and tally, it preserves the continuous feature values and their multiplicative combination, but discards the relative strength of each feature. This is achieved by replacing every exponent with its sign, giving the directional log-score

$$s_{\text{dir}}(x) \ = \ \sum_{j=1}^{r} \text{sign}(\beta_j)\ \log x_j. \tag{9}$$

Viewed in log-space, the directional rule replaces the coefficient vector β with its sign vector. The feature support and directions are preserved, and only the coefficient magnitudes are removed. The intercept $\log|\alpha|$ is irrelevant to the ranking and therefore omitted. This is the most aggressive of our simplifications, and the one whose cost is least predictable from the form alone. Intuitively, if all retained features contribute equally, replacing their exponent magnitudes by one should have little effect on the ranking, whereas if a few features dominate the monomial, flattening the exponent magnitudes may substantially alter it.

The following proposition characterizes the similarity between the full and directional scores after exponent flattening. This similarity can be expressed as a Pearson correlation computed from the fitted exponents and feature covariance. Our fidelity measure, however, is the tie-corrected Kendall's τ_b [13]. Under elliptical feature distributions, Greiner's classical relation [9] links Pearson correlation to Kendall's τ. We therefore use this relation to predict rank fidelity, noting that τ and τ_b can differ when ties are frequent.

Proposition 3.1 (Fidelity after exponent flattening). *Write $w = (\log x_1, \ldots, \log x_m)$ and let Σ denote its covariance. Up to constants, the full and directional log-scores are the linear functions $s_{\text{full}} = \beta^\top w$ and $s_{\text{dir}} = d^\top w$, where $d_j = \text{sign}(\beta_j)$ on the retained set S, $|S| = r$, and $d_j = 0$ otherwise. Their correlation is*

$$\rho \ = \ \frac{\beta^\top \Sigma\, d}{\sqrt{\beta^\top \Sigma\, \beta}\,\sqrt{d^\top \Sigma\, d}}, \tag{10}$$

which holds without distributional assumptions. If w is elliptically distributed, Greiner's relation [9] links this correlation to the rank fidelity between the two scores:

$$\tau \ = \ \tfrac{2}{\pi}\arcsin(\rho). \tag{11}$$

Proof sketch. For any fixed vectors a, b,

$$\text{Cov}(a^\top w, b^\top w) = a^\top \Sigma b, \qquad \text{Var}(a^\top w) = a^\top \Sigma a.$$

Substituting $a = \beta$ and $b = d$ into the definition of the Pearson correlation between s_{full} and s_{dir} yields (10), which requires only that the second moments of w exist. Since $(s_{\text{full}}, s_{\text{dir}})$ is a linear transformation of w, it is elliptically distributed whenever w is. Equation (11) therefore follows directly from Greiner's classical relation. □

Equation (10) makes this intuition precise: ρ is high when replacing the fitted exponents by their signs leaves the two score directions well aligned under the observed feature covariance, and low when exponent magnitudes or feature dependencies make that flattening consequential. The proposition therefore yields a prediction of directional-rule fidelity directly after training: β and d follow from the fitted exponents and Σ from the training data. No held-out data or evaluation of the directional rule is required. Section 5.3 compares the resulting predicted Kendall's τ with observed fidelity.

3.2 Sparsity During Training

Every simplification so far starts from a fitted monomial with all m exponents non-zero. This is not incidental: ECSEL regularizes the exponents with an ℓ_1 penalty, which shrinks small coefficients but under gradient-based optimization does not reliably drive them to exact zeros, so a compact rule always requires a separate reduction step afterwards. As an alternative, we use iterative hard thresholding (IHT) [3], which retains only the s largest exponents after each gradient step. This provides an independent route to sparsity for assessing whether the extracted structure depends on post-hoc pruning.

3.3 Human Assessment of Readable Forms

The simplifications are intended to produce representations that are not only formally simpler, but easier to understand and use. We therefore conduct an anonymous human assessment using the Loan and FraudEcom classifiers. For each dataset, participants are presented with the pruned monomial, directional rule, scorecard, and tally derived from the fitted classifier, without information about predictive performance. They rate each representation's ease of understanding on a five-point Likert scale and select the representation they would prefer for practical use. This allows us to distinguish perceived readability from practical preference and examine whether these judgments differ by professional background.

4 Experimental Setup

Datasets. We evaluate on four public financial classification datasets, summarized in Table 2. They cover consumer credit default (Loan [11], Default [24]) and fraud detection (FraudEcom [23], Creditcard [7]), range over an order of magnitude in size, and span positive-class rates from 0.17% to 22%. The datasets also differ in feature type: Loan and Default contain named continuous attributes, FraudEcom combines continuous variables with small integer counts, and Creditcard consists of anonymized principal components. As discussed in Section 5, these differences influence which simplifications remain effective.

Table 2: Financial datasets, ordered by class imbalance; positive-class rate is the PR-AUC base rate.

Dataset	n	Features	Pos. rate	Domain
Creditcard	284,807	30	0.17%	card fraud
FraudEcom	151,112	6	9.4%	account fraud
Loan	395,492	42	10.1%	loan default
Default	30,000	26	22.1%	credit default

Preprocessing and evaluation. Each dataset is split into training, validation, and test sets (60/20/20, stratified by label), repeated over five random seeds. Following Lumadjeng et al. [16], features are scaled to $[0.01, 10.01]$, with the scaler fit on the training split only. Hyperparameters were optimized using Optuna on the training split, with the best configuration selected by cross-validation. Pruning level and all decision thresholds are chosen on the validation split. The test split is used only for the final evaluation. Unless stated otherwise, reported results are means over the five seeds.

Models and readable forms. Our model is the ECSEL monomial. From each fitted monomial we derive the four readable forms of Section 3.1: the pruned monomial, directional rule, scorecard, and tally, and evaluate each as a classifier in its own right. The pruning level is selected as the smallest number of retained features r whose validation PR-AUC is within 0.01 of the full monomial. We additionally compare post-hoc pruning with an in-training sparse variant based on IHT. Throughout, scorecards use a maximum of five points ($P_{\max} = 5$), while scorecard and tally thresholds are selected on the validation split.

Metrics. Because all datasets are imbalanced, we use the area under the precision–recall curve (PR-AUC) as the primary performance metric and report the corresponding base rate for reference. To measure how faithfully a simplified form reproduces the full monomial, we report Kendall's τ_b between their scores,

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_s)(C + D + T_f)}}, \tag{12}$$

where C and D denote the number of concordant and discordant pairs, and T_s and T_f account for ties in the simplified and full scores, respectively. For the directional rule, we additionally compare the observed τ_b with the value τ_{pred} predicted by Proposition 3.1, computed solely from the fitted exponents and the training data.

Finally, we assess whether the scores produced by each representation can be converted into reliable probability estimates. Since the directional rule, scorecard, and tally do not natively output probabilities, we calibrate each representation on the validation set using isotonic regression and report its expected calibration error (ECE) on the test set,

$$\text{ECE} = \sum_{b=1}^{B} \frac{|b|}{n} \left|\text{acc}(b) - \text{conf}(b)\right|, \tag{13}$$

where the test set is partitioned into B equal-mass bins by predicted probability, $|b|$ denotes the number of instances in bin b, $\text{acc}(b)$ its observed positive rate, and $\text{conf}(b)$ its mean predicted probability. ECE therefore assesses the reliability of the recalibrated probabilities. Unless stated otherwise, all reported metrics are averaged over the five random seeds.

Human assessment. Of 36 respondents, 34 were included in the subgroup analysis: 14 with a finance/risk background (including three also reporting AI/ML research) and 20 AI/ML researchers. Two respondents reporting neither background were excluded.

5 Results

We first quantify the cost of each controlled simplification and examine when different representations succeed or fail across datasets. We next evaluate the prediction of Proposition 3.1, assess the readability of the resulting explanations in a human study, and finally compare post-hoc pruning with in-training sparsity.

Figure 2 shows the resulting representations for Loan and FraudEcom; the remaining datasets undergo the same controlled simplifications but are omitted for brevity. These are also the representations presented to participants in the human assessment.

5.1 The Cost of Controlled Simplification

Table 3 reports predictive performance, fidelity, and recalibrated ECE for the controlled simplifications of Section 3.1. PR-AUC base rates are 0.101 (Loan), 0.094 (FraudEcom), 0.221 (Default), and 0.002 (Creditcard). Across five seeds, variability is generally small for the full and pruned monomials but larger for some aggressive simplifications. We discuss several patterns below.

Pruning is nearly free. Pruning selected under the validation tolerance generalizes with little predictive loss on the held-out test sets. The largest reduction in PR-AUC is 0.008 (FraudEcom), while Loan and Creditcard lose only 0.001 and 0.002, respectively.

(a) Loan

Pruned monomial

$$z_{\text{default}}(x) \propto x_{\text{int_received}}^{+1.92} x_{\text{total_pymnt}}^{-1.77} x_{\text{mths_since_pymnt}}^{+0.83} x_{\text{last_pymnt}}^{-0.76} \times x_{\text{annual_inc}}^{+0.22} x_{\text{mths_since_issue}}^{-0.18} x_{\text{out_prncp}}^{-0.12}.$$

Directional rule

Predict *default* when **total interest received** is **high**, **total payment** is **low**, **time since last payment** is **high**, **last payment** is **low**, **income** is **high**, **time since issue** is **low**, and **outstanding principal** is **low**.

Scorecard

Add 5 points for **total interest received** **high**; 5 for **total payment** **low**; 2 for **time since last payment** **high**; 2 for **last payment** **low**; 1 for **income** **high**; 1 for **time since issue** **low**; and 1 for **outstanding principal** **low**.
Flag as *default* when the total score is at least 11 out of 17.

Tally

Flag as *default* when at least 6 of the 7 conditions hold: **total interest received** **high**; **total payment** **low**; **time since last payment** **high**; **last payment** **low**; **income** **high**; **time since issue** **low**; and **outstanding principal** **low**.

(b) FraudEcom

Pruned monomial

$$z_{\text{fraud}}(x) \propto x_{\text{ip_shared_count}}^{+0.56} x_{\text{device_shared_count}}^{+0.27} x_{\text{sex_bin}}^{+0.01} x_{\text{purchase_value}}^{-0.01}.$$

Directional rule

Predict *fraud* when **IP shared count** is **high**, **device shared count** is **high**, the **sex indicator** is 1, and **purchase value** is **low**.

Scorecard

Add 5 points for **IP shared count** **high**; 2 for **device shared count** **high**; 1 for **sex indicator** = 1; and 1 for **purchase value** **low**.
Flag as *fraud* when the total score is at least 8 out of 9.

Tally

Flag as *fraud* when all 4 conditions hold: **IP shared count** **high**; **device shared count** **high**; **sex indicator** = 1; and **purchase value** **low**.

Figure 2: Controlled simplifications of representative Loan and FraudEcom classifiers (seed 45). Each panel derives four representations from the same retained features.

Table 3: Predictive performance, fidelity, and recalibrated ECE (mean over five seeds).

Dataset	Metric	Full	Pruned	Directional	Scorecard	Tally
Loan	PR-AUC	0.886	0.885	0.642	0.329	0.300
	τ_b	1.000	0.846	0.568	0.206	0.196
	ECE	0.001	0.001	0.001	0.002	0.002
FraudEcom	PR-AUC	0.647	0.647	0.639	0.333	0.291
	τ_b	1.000	0.845	0.831	0.670	0.656
	ECE	0.004	0.004	0.004	0.004	0.004
Default	PR-AUC	0.380	0.377	0.381	0.303	0.334
	τ_b	1.000	0.820	0.744	0.375	0.530
	ECE	0.014	0.016	0.016	0.014	0.016
Creditcard	PR-AUC	0.677	0.675	0.621	0.018	0.017
	τ_b	1.000	0.698	0.614	0.259	0.406
	ECE	0.000	0.000	0.000	0.000	0.000

Fidelity remains high, with Kendall's τ_b between 0.70 and 0.85. Thus, substantial reductions in feature support can be achieved without materially degrading held-out predictive performance.

Binarizing feature values is costly. On Loan, PR-AUC falls from 0.885 for the pruned monomial to 0.329 for the scorecard, and 0.300 for the tally. Similar reductions occur on Creditcard, where both point-based forms approach the dataset base rate. Unlike the other datasets, however, the Creditcard features are anonymized principal components rather than meaningful financial variables. Thresholding these latent components discards much of the information contained in their continuous values, making them particularly unsuitable for point-based rules. We return to this observation in the next subsection. The cost of binarizing feature values is therefore substantially larger than that of removing effect magnitudes, suggesting that, on these datasets, the continuous feature values carry more predictive information than the precise exponent magnitudes.

Predictive performance and fidelity capture different properties. The directional rule illustrates the distinction between predictive performance and fidelity. On Default, it essentially matches the predictive performance of the full monomial (PR-AUC 0.381 versus 0.380) despite a lower rank fidelity ($\tau_b = 0.744$). Conversely, on Loan, fidelity drops to 0.568 while predictive performance remains well above the scorecard and tally. Fidelity can therefore erode faster than predictive performance, while the simplified rule remains an effective classifier.

Simplified scores can be reliably calibrated. From Table 3, we see that after validation-set isotonic recalibration, ECE remains low across representations and datasets, showing that simplified scores can yield reliable probability estimates despite losses in ranking.

5.2 When Do Different Simplifications Succeed?

The results of Table 3 show that no representation is uniformly best. Instead, the success of a simplification depends on whether the information it removes is predictive for the underlying dataset.

The two classifiers illustrate contrasting outcomes. On Loan, pruning reduces 42 features to seven with virtually no predictive loss, while removing effect magnitudes and especially binarizing continuous values is costly. In contrast, FraudEcom retains four features, two of which dominate the others by almost two orders of magnitude. Removing effect magnitudes therefore has little effect (PR-AUC 0.647 to 0.639), making the directional rule a simpler

yet faithful summary. The remaining datasets reinforce this variation: on Default, the directional rule matches the full monomial closely (PR-AUC 0.381 versus 0.380) despite retaining ten to thirteen features, suggesting that useful signal is distributed across many variables. On Creditcard, binarizing the anonymized components reduces the point-based forms to near-base-rate performance, while the directional rule remains stronger. Thus, simplification is not a single ladder of increasingly weaker models: its cost depends on where the monomial's predictive information resides.

5.3 Predicting Fidelity Before Simplification

Proposition 3.1 predicts the rank fidelity of the directional rule from the learned monomial and training data, without constructing or evaluating the simplified rule. Table 4 compares predicted and observed Kendall rank fidelity across five seeds.

Table 4: Predicted and observed rank fidelity across five seeds.

Dataset	ρ	τ_{pred}	τ_b	MAE
Loan	0.732 ± 0.260	0.578 ± 0.270	0.568 ± 0.248	0.024
Default	0.911 ± 0.027	0.731 ± 0.040	0.744 ± 0.039	0.019
Creditcard	0.821 ± 0.018	0.614 ± 0.019	0.614 ± 0.026	0.008
FraudEcom	0.906 ± 0.134	0.759 ± 0.165	0.831 ± 0.238	0.236

On Loan, Default, and Creditcard, the prediction is consistently accurate. Across the fifteen fitted models, the mean absolute error is 0.017, with a maximum error of 0.039. The prediction remains accurate over observed fidelities ranging from 0.25 to 0.83, providing empirical support for Proposition 3.1 on these datasets.

FraudEcom is the exception, with a substantially larger MAE of 0.236. The failure is driven by variation across fits: in three seeds, $\rho \approx 0.96$ corresponds to observed $\tau_b \approx 0.98$, whereas in another $\rho = 0.973$ but $\tau_b = 0.440$. The dominant features are low-cardinality count variables, producing many tied scores. Pearson correlation is insensitive to this tie structure, whereas Kendall's τ_b accounts for it. Consequently, the Greiner conversion becomes inaccurate, reflecting a failure of its distributional assumptions rather than of the correlation formula itself.

Thus, the fidelity of the directional rule is generally predictable immediately after training. Among the simplifications considered here, it is the only representation whose expected agreement with the original classifier can be estimated before the simplified rule is produced.

5.4 Human Readability and Preference

Participants evaluated the Loan and FraudEcom representations shown in Figure 2. Objective comprehension was high across all four forms, with 92–100% of responses correctly identifying how a stated feature condition affected the prediction. The main differences therefore concern perceived ease and practical preference rather than basic understanding. All three simplified rules were perceived as easier to understand than the pruned monomial on both datasets.

Aggregated across both datasets, finance/risk respondents selected the directional rule in 57% of assessments, compared with 29% for the point-based forms. AI/ML researchers showed the opposite pattern, selecting point-based forms in 72% of assessments

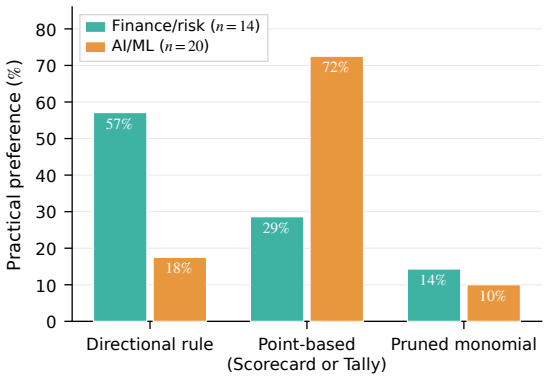

Figure 3: Practical preference by professional background, aggregated across Loan and FraudEcom.

and the directional rule in 18%. The pruned monomial was rarely preferred by either group (14% and 10%, respectively). Thus, while simplification improves perceived ease, the representation preferred in practice differs by professional background.

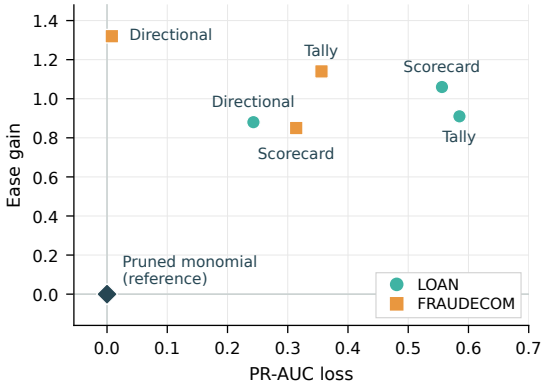

Figure 4: Gain in mean perceived ease versus loss in PR-AUC, both relative to the pruned monomial.

Preference alone, however, does not capture what is preserved by a simplification. Figure 4 plots, relative to the pruned monomial, the increase in mean perceived ease against the loss in PR-AUC. The directional rule gains substantially in perceived ease at lower predictive cost than the point-based forms. This is particularly pronounced on FraudEcom, where PR-AUC changes only from 0.647 to 0.639. Participants were not shown predictive performance, so this is an ex-post comparison rather than a trade-off they were asked to make.

5.5 In-Training Sparsity Recovers Similar Rules

Finally, we examine whether the extracted rules depend on the post-hoc pruning procedure itself. Table 5 compares post-hoc pruning with an in-training sparse variant based on iterative hard thresholding (IHT), reporting the number of retained features r, their Jaccard overlap, and predictive performance. Across all datasets, both approaches retain nearly identical feature sets and achieve

comparable predictive performance. Feature selection is identical on Loan, FraudEcom, and Creditcard, and differs by only two of twelve features on Default, suggesting that the extracted rules primarily reflect learned data structure rather than pruning artifacts.

Table 5: Post-hoc pruning versus in-training sparsity (IHT).

Dataset	r_{prune}	r_{IHT}	Jaccard	PR-AUC$_{\text{prune}}$	PR-AUC$_{\text{IHT}}$
Loan	6	6	1.00	0.885	0.889
FraudEcom	3	3	1.00	0.647	0.651
Default	12	12	0.83	0.377	0.381
Creditcard	3	3	1.00	0.675	0.691

6 Limitations and Future Work

Our study is deliberately restricted to single monomials ($K = 1$) and four financial decision-making datasets, and the findings may not extend directly to richer models or other domains. The readable forms likewise use fixed simplifications, such as median binarization, by design: our aim is to isolate the cost of removing specific information rather than optimize each representation for predictive performance. The results therefore characterize the cost of the proposed simplifications, not the best achievable performance of each rule form. Our human assessment is exploratory, with 36 respondents, and measures stated preferences rather than use in actual decisions. Future work could extend controlled simplification to multi-term signomials ($K > 1$), other inherently interpretable models, and larger decision-based human evaluations.

7 Conclusion

In financial decision making, an interpretable model can still be too complex to communicate. We studied this gap by progressively reducing learned interpretable equations into pruned equations, directional rules, scorecards, and tallies, and measuring predictive performance, fidelity, and human perception at each stage. The reductions are not points on a single accuracy–readability tradeoff. Removing weakly contributing features is nearly free, while removing effect magnitudes is often cheap and its effect on fidelity can be anticipated from the fitted model. Binarizing continuous feature values to derive point-based forms incurs the largest predictive losses. Importantly, predictive performance and fidelity can diverge: a simpler rule may remain an effective classifier without faithfully reproducing the original model. The human assessment confirms that simplification improves perceived ease, while practical preferences in our sample vary by professional background. Readable explanations are therefore better treated as measurable reductions of an interpretable model than assumed faithful because they are easier to read.

Ethics and Privacy Statement

We use public financial benchmark datasets without personally identifiable information. While simplification may improve transparency, it can discard information relevant to accurate and equitable decisions and should therefore complement, not replace, domain validation, fairness assessment, and regulatory review.

References

[1] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. 2018. Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research* 18, 234 (2018), 1–78.

[2] Dimitris Bertsimas and Jack Dunn. 2017. Optimal Classification Trees. *Machine Learning* 106, 7 (2017), 1039–1082. doi:10.1007/s10994-017-5633-9

[3] Thomas Blumensath and Mike E. Davies. 2008. Iterative Thresholding for Sparse Approximations. *Journal of Fourier Analysis and Applications* 14, 5 (2008), 629–654. doi:10.1007/s00041-008-9035-z

[4] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2022. A Holistic Approach to Interpretability in Financial Lending: Models, Visualizations, and Summary-Explanations. *Decision Support Systems* 152 (2022), 113647. doi:10.1016/j.dss.2021.113647

[5] Yujia Chen, Raffaella Calabrese, and Belen Martin-Barragan. 2024. Interpretable Machine Learning for Imbalanced Credit Scoring Datasets. *European Journal of Operational Research* 312 (2024), 357–372. doi:10.1016/j.ejor.2023.06.036

[6] Wenhao Chi and Ş. İlker Birbil. 2026. Learning an Interpretable Risk Scoring System for Maximizing Decision Net Benefit. arXiv:2604.04241 [cs.LG]

[7] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. 2015. Calibrating Probability with Undersampling for Unbalanced Classification. In *2015 IEEE Symposium Series on Computational Intelligence*. IEEE.

[8] Antonio Di Marino, Vincenzo Bevilacqua, Angelo Ciaramella, Ivanoe De Falco, and Giovanna Sannino. 2025. Ante-Hoc Methods for Interpretable Deep Models: A Survey. *ACM Comput. Surv.* 57, 10 (2025). doi:10.1145/3728637

[9] R. Greiner. 1909. Über das Fehlersystem der Kollektivmaßlehre. *Zeitschrift für Mathematik und Physik* 57 (1909), 121–158.

[10] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (2011), 141–154. doi:10.1016/j.dss.2010.12.003

[11] KA-KA-shi. 2020. Lending Club Loan Data. Kaggle. Accessed: 2026-08-09. https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv

[12] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[13] Filip Lindskog, Alexander McNeil, and Uwe Schmock. [n. d.]. Kendall's tau for elliptical distributions. ([n. d.]).

[14] Jiachang Liu, Chudi Zhong, Boxuan Li, Margo Seltzer, and Cynthia Rudin. 2022. FasterRisk: Fast and Accurate Interpretable Risk Scores. In *Advances in Neural Information Processing Systems*, Vol. 35.

[15] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 623–631. doi:10.1145/2487575.2487579

[16] Adia C. Lumadjeng, Ilker Birbil, and Erman Acar. 2026. ECSEL: Explainable Classification via Signomial Equation Learning. In *Forty-third International Conference on Machine Learning*.

[17] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30. 4768–4777.

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144. doi:10.1145/2939672.2939778

[19] Gero Szepannek. 2022. An Overview on the Landscape of R Packages for Open Source Scorecard Modelling. *Risks* 10, 3 (2022), 67. doi:10.3390/risks10030067

[20] Jiancheng Tu and Zhibin Wu. 2025. Inherently interpretable machine learning for credit scoring: Optimal classification tree with hyperplane splits. *European Journal of Operational Research* 322, 2 (2025), 647–664. doi:10.1016/j.ejor.2024.10.046

[21] Berk Ustun and Cynthia Rudin. 2016. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning* 102, 3 (2016), 349–391. doi:10.1007/s10994-015-5528-6

[22] Berk Ustun and Cynthia Rudin. 2019. Learning Optimized Risk Scores. *Journal of Machine Learning Research* 20, 150 (2019), 1–75.

[23] vbinh002. [n. d.]. Fraud Ecommerce. Kaggle. Accessed: 2026-08-09. https://www.kaggle.com/datasets/vbinh002/fraud-ecommerce

[24] I-Cheng Yeh and Che-hui Lien. 2009. The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Systems with Applications* 36, 2 (2009), 2473–2480. doi:10.1016/j.eswa.2007.12.020

Received 2 August 2026