Social Science and Management
Print ISSN: 3007-6854
Online ISSN: 3007-6862
DOI:https://doi.org/ 10.61784/ssm4094

EXPLAINABLE SUPPLIER RISK PREDICTION FOR CRITICAL TELECOMMUNICATIONS INFRASTRUCTURE PROCUREMENT: AN EMPIRICAL XGBOOST–SHAP FRAMEWORK

MengLin Bian[1*], Ying Wang[2], YiZhen Lin[3]
[1]*University of Missouri, Champaign, IL, United States.*
[2]*Pepperdine University, Malibu, CA 90263, United States.*
[3]*Stanford University, Stanford, CA 94305, United States.*
[*]*Corresponding Author: MengLin Bian*

Abstract: Telecommunications networks are critical national infrastructure, yet their equipment supply base is highly concentrated, so a single supplier failure can propagate into a national service outage. Machine-learning supplier-risk models are increasingly proposed for procurement screening, but they are almost always evaluated on discrimination alone: nobody asks whether the resulting explanations are directionally sensible, whether they survive a model refit, or whether the model actually reduces procurement cost. We present CTI-SRP, a two-tier explainable framework that decomposes supplier risk into financial-viability risk and operational delivery risk, couples a monotonicity-constrained XGBoost learner to exact TreeSHAP attribution, and adds an audit layer with three new explanation-quality metrics — Explanation Coherence Rate (ECR), Local Coherence Rate (LCR) and Monotonicity Violation Rate (MVR) — plus an Explanation Stability Index (ESI) and a cost-sensitive decision layer. Experiments use two real datasets (6,819 listed firms with 95 financial ratios; 180,519 order lines) and a transparently specified synthetic telecom cohort in which the true drivers, signs and interactions are known by construction. Encoding 46 accounting-theory sign priors as monotone constraints raises ECR from 0.870 to 0.978 and drives MVR from 0.345 to exactly zero at no cost in AUC (0.9547→0.9573). TreeSHAP nearly doubles top-5 driver stability over gain importance (Jaccard 0.446 vs 0.247). A leakage audit shows that the near-perfect delivery-risk accuracy reported in prior work collapses from AUC 0.992 to 0.733 once post-shipment features are removed. On the synthetic cohort the pipeline recovers all 16 true drivers with zero false drivers, all three true interactions in the top three, and a 0.953 correlation between true coefficient magnitude and mean |SHAP|. A criticality-weighted composite index captures 62.7% of realised loss within a 10% audit budget, versus 48.2% for probability ranking alone.
Keywords: Supplier risk prediction; Explainable artificial intelligence; XGBoost; SHAP; Monotonicity constraints; Critical infrastructure procurement; Telecommunications supply chain; Cost-sensitive learning

1 INTRODUCTION

Mobile and fixed telecommunications networks are the substrate on which energy grids, payment systems, emergency services and public administration now depend. They are formally designated critical national infrastructure in most jurisdictions, and the security of their supply chains has become an explicit object of state policy: between 2018 and 2021 a series of governments restricted or excluded specific equipment vendors from their national networks, and the resulting policy debate has been analysed as a shift from path-dependent procurement towards risk-based governance [1]. The technical threat surface has been catalogued in parallel by sector agencies [2-3].

What makes telecommunications procurement unusual is the combination of three properties. First, extreme supplier concentration: the radio access and core network equipment markets are effectively oligopolies, so the loss of one qualified supplier removes a large fraction of available capacity rather than a marginal share. Second, very long qualification and lead times, which means substitution cannot be improvised after a disruption is observed. Third, an asymmetric loss function: admitting a supplier that later fails can cost orders of magnitude more than unnecessarily screening out a supplier that would have performed, because the downside is a service outage on critical infrastructure rather than a procurement inefficiency. Together these imply that supplier risk must be assessed before award, from information available at award time, and that the assessment must be defensible to an auditor or regulator.

Machine learning has been applied extensively to supply chain and supplier risk, and gradient-boosted trees in particular dominate tabular risk problems [4-12]. Post-hoc attribution, especially SHAP, is now the default answer to the resulting opacity [13-14]. Yet the literature that combines the two is dominated by a single evaluation pattern: report AUC, report a SHAP summary plot, conclude that the model is explainable. We argue that this pattern leaves four substantive gaps.

Gap 1: explanations are never audited for directional coherence. A model can attain excellent discrimination while implying, on some region of the input space, that a higher debt ratio lowers insolvency risk. Such an artefact is invisible in an aggregate importance plot but fatal in a procurement memo, because a buyer who acts on it will make the supply base worse. The monotonic-classification literature has long argued that domain sign knowledge should be enforced rather than hoped for, but this has not been connected to the auditability of SHAP explanations [15-17].

Gap 2: explanation stability is not measured. Attribution methods are known to be fragile [18-19]. If the top drivers change every time the model is refit on a resample, the explanation cannot support a decision that must be justified months later. Practitioners nonetheless choose between SHAP, split-gain and permutation importance on convenience rather than on evidence about which ranking is reproducible.
Gap 3: benchmark results are contaminated by target leakage. The most widely used public supply-chain dataset carries a late-delivery label that is a deterministic function of realised shipping duration [20]. Models that retain that column report accuracies above 0.98, which is arithmetic rather than prediction [21]. Reported performance therefore cannot be used to calibrate expectations for a pre-award screening system.
Gap 4: discrimination is reported instead of decision value. AUC is threshold-free and cost-blind. For critical infrastructure the quantity of interest is whether a screening policy reduces expected loss relative to the best policy that ignores the model, and over what range of cost asymmetries it does so [22-23].
This paper addresses all four gaps within one framework, CTI-SRP (Critical Telecommunications Infrastructure Supplier Risk Prediction). Our contributions are:
1) A two-tier risk decomposition. We separate supplier financial-viability risk from operational delivery risk, fit a tier-specific model to each, and fuse them with a criticality-weighted severity term into a Critical-Infrastructure Supplier Risk Index (CISRI). An ablation shows each block carries non-redundant signal, with the financial tier worth 7.1 AUC points and the operational block 11.8.
2) Auditable explanations via domain-constrained boosting. We elicit sign priors for 46 of 94 financial ratios from accounting theory, impose them as monotone constraints, and introduce three metrics — ECR, LCR and MVR — that quantify whether the resulting explanations agree with those priors globally, locally and at the level of the decision rule. Constraints raise ECR by 10.9 points, raise LCR by 6.3 points, and eliminate monotonicity violations entirely, while AUC does not degrade.
3) An Explanation Stability Index. ESI measures the agreement of global driver rankings across bootstrap refits. TreeSHAP and split-gain importance are statistically indistinguishable over the full ranking, but TreeSHAP is far more stable at the head of the ranking (Jaccard@5 of 0.446 versus 0.247), which is precisely the part that reaches a decision document.
4) A leakage-free pre-award protocol and a quantified leakage audit. We specify which features are admissible at award time, adopt an out-of-time split, and show that the same learner scores AUC 0.733 under the clean protocol versus 0.992 with one post-shipment column restored.
5) Ground-truth validation of the explanation pipeline. Because no public dataset exposes telecom-specific procurement attributes, we specify a synthetic cohort whose coefficients, signs, interactions and null features are known. This lets us verify — impossibly on real data — that the pipeline recovers the true causal structure, and it supplies the cost-sensitive analysis with a fused telecom-attribute setting.
All code, generated data and result tables are released with the paper. We are explicit throughout about which findings rest on real data and which on the synthetic cohort.

2 RELATED WORK

2.1 Machine Learning for Supplier and Supply-Chain Risk

Supply chain risk management has an extensive pre-machine-learning tradition [4]. The learning-based strand was consolidated by Baryannis et al., who both surveyed the intersection of artificial intelligence and supply chain risk and, in the work closest to ours in spirit, made the accuracy–interpretability trade-off the explicit object of study for delivery-delay prediction [5-6]. Brintrup et al. predicted supplier disruptions in complex asset manufacturing and showed that engineered agility features carried most of the signal [7]. Network-structural approaches infer missing supply relations, first with classical learners and later with graph neural networks [8, 24]. Text-driven approaches mine news feeds for early warning [25]. Recent work continues to decompose risk factors for supply-chain risk assessment with tree ensembles [26]. Two limitations recur. Interpretability, where treated at all, is treated as a model-class choice (use a decision tree instead of an SVM) rather than as a measurable property of the explanations a model produces. And evaluation is discrimination-centric, so the cost structure that makes critical-infrastructure procurement distinctive never enters the objective.

2.2 Financial Distress as a Supplier-Risk Channel

A supplier that becomes insolvent stops delivering, so corporate failure prediction is a direct component of supplier risk. The field runs from discriminant and logit models to modern ensembles, which consistently outperform their statistical predecessors [27-29]. We build the financial tier on the Taiwan Economic Journal panel introduced by Liang et al., which pairs 95 financial ratios with bankruptcies defined by Taiwan Stock Exchange regulation [30]. Two properties make this panel apt here rather than merely convenient: Taiwan region's listed universe is dominated by information and communications technology hardware manufacturers, which is the actual upstream population for telecom equipment; and the ratio set is large enough that sign priors can be elicited for a substantial subset from accounting theory alone.

2.3 Explainability, Monotonicity and Cost-Sensitive Evaluation

Post-hoc explanation is dominated by LIME and SHAP, with TreeSHAP giving exact Shapley values for tree ensembles in polynomial time [13-14, 31]. The critical literature is equally developed: Rudin argues that high-stakes decisions should use inherently interpretable models [32]; broader surveys map the field and its evaluation problems [33-34]; attribution methods can be adversarially manipulated and are sensitive to feature dependence [18-19]. Monotonic classification offers a constructive middle path, embedding domain sign knowledge directly in the hypothesis class [15]; it has been applied to credit rating with constrained SVMs and, more recently, to credit scoring with monotone neural additive models [16-17]. Cost-sensitive learning provides the decision-theoretic apparatus — optimal thresholds from a cost matrix and cost curves for visualising performance across cost ratios — while precision–recall analysis is preferred to ROC under strong imbalance [22-23, 35]. Our contribution is to compose these three strands: we use monotone constraints not primarily to improve accuracy but to make SHAP explanations auditable, we measure that auditability, and we evaluate the whole pipeline by expected procurement cost.

2.4 Positioning

To our knowledge, no prior work (i) quantifies the coherence of SHAP explanations against elicited domain sign priors, (ii) measures the refit stability of competing global importance measures on a supplier-risk task, (iii) quantifies the leakage in the standard public delivery-risk benchmark, or (iv) validates a supplier-risk explanation pipeline against a known generative ground truth. CTI-SRP does all four.

3 THE CTI-SRP FRAMEWORK

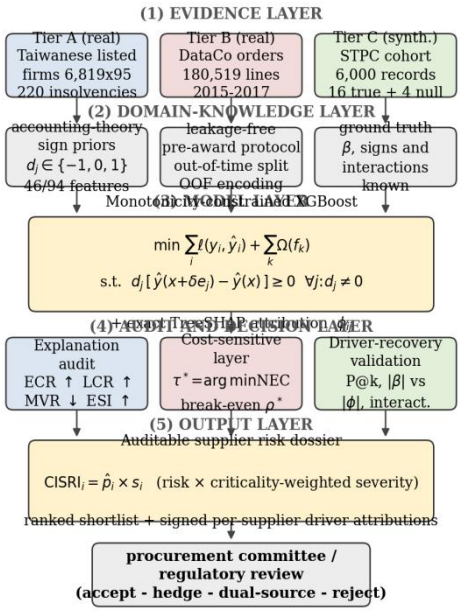

Figure 1 The CTI-SRP Framework

Note: Evidence from two real panels and one transparently specified synthetic cohort is combined with elicited domain knowledge (sign priors, an admissibility protocol, and known ground truth), fitted with a monotonicity-constrained boosted ensemble, and then passed through three audit modules before a criticality-weighted dossier reaches the procurement committee.

3.1 Two-Tier Risk Decomposition

A supplier fails to deliver for reasons that live on different time scales and in different data, see Figure 1. We therefore model two tiers. Tier A estimates financial-viability risk, the probability that supplier i becomes financially unable to perform within the contract horizon, from firm-level accounting information. Tier B estimates operational delivery risk, the probability that a specific committed order breaches its service-level agreement, from order- and lane-level information available at order placement. The two are deliberately not pooled: they have different units of observation (firm-year versus order line), different base rates (3.2% versus 55%), and different decision consequences (disqualification versus hedging).

Tier C fuses them. For supplier i with fused risk estimate pi and severity si we define the Critical-Infrastructure Supplier Risk Index

$$\text{CISRI}_i = \hat{p}_i \times s_i, \quad s_i = v_i \cdot c_i \tag{1}$$

where vi is contract value and ci $\in \{1,2,3\}$ is the criticality tier of the sourced component. CISRI is an expected-loss ranking rather than a probability ranking; Section 5.8 shows this matters substantially when the audit budget is small.

3.2 Monotonicity-Constrained Gradient Boosting

Both tiers use gradient-boosted decision trees [9-10]. The ensemble predicts ŷi = σ(∑k fk(xi)) and is fitted by minimising

$$\mathcal{L} = \sum_i \ell(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{2}$$

with T the leaf count and w the leaf weights. Class imbalance is handled by scaling positive-class gradients by the negative-to-positive ratio in the training fold.

The novel component is the constraint set. For each feature j we elicit a sign prior dj ∈ {−1, 0, +1}, where +1 means that a larger value can only increase risk, −1 that it can only decrease it, and 0 that theory is silent. We then require

$$d_j \cdot \left[\hat{y}\left(x + \delta e_j\right) - \hat{y}(x)\right] \geq 0 \quad \forall \delta > 0, \forall j : d_j \neq 0 \tag{3}$$

which the boosting implementation enforces exactly during split selection by rejecting splits whose child weights would violate the required ordering. Priors are elicited from accounting theory, not from the data: profitability, liquidity, cash-generation and capitalisation ratios receive dj = −1; leverage, borrowing-dependency, contingent-liability and expense-intensity ratios receive +1. This yields 46 constrained features out of 94. Critically, (3) constrains the decision rule, and therefore constrains every explanation derived from it, which is what makes the audit metrics of Section 3.4 achievable rather than aspirational.

3.3 Attribution

We use exact TreeSHAP, which assigns to feature j of instance i the Shapley value ϕij satisfying local accuracy, missingness and consistency [13-14]. Global importance is the mean absolute attribution, Ij = n^{-1} ∑i |ϕij|. Pairwise structure is read from SHAP interaction values, whose off-diagonal mean absolute magnitude scores each feature pair.

3.4 Explanation-Quality Metrics

Let Φ be the attribution matrix on a held-out set, D = {j : dj ≠ 0} the constrained set, and ρj the Spearman correlation between x·j and ϕ·j. We define three metrics.

Explanation Coherence Rate (ECR) is the share of theory-covered features whose global attribution direction matches the prior:

$$\mathrm{ECR} = |D|^{-1} \sum_{j \in D} \mathbb{1}\left[\mathrm{sgn}\left(\rho_j\right) = d_j \wedge \left|\rho_j\right| \geq \varepsilon\right] \tag{4}$$

with ε = 0.05 excluding numerically null relations. ECR is computed over the whole covered set, not the top few features, so it cannot be inflated by a model that gets the headline drivers right and the tail wrong.

Local Coherence Rate (LCR) is a finer, instance-level version. Writing ĥj for the median of feature j,

$$\mathrm{LCR} = N^{-1} \sum_{j \in D} \sum_i \mathbb{1}\left[\mathrm{sgn}\left(\phi_{ij}\right) = d_j \cdot \mathrm{sgn}\left(x_{ij} - \hat{h}_j\right)\right] \tag{5}$$

where N counts the non-degenerate (instance, feature) pairs. LCR asks whether the attribution shown to a buyer for a specific supplier points the right way, which is the form in which an explanation is actually consumed. It cannot reach 1 even for a perfectly monotone model, because Shapley attributions are baseline-relative and interaction-loaded; it is a comparative measure.

Monotonicity Violation Rate (MVR) audits the decision rule rather than the attribution. For each constrained feature we sweep a 12-point quantile grid, average the model output over a fixed sample at each grid value to obtain a partial-dependence profile gj(·), and count adverse steps:

$$\mathrm{MVR} = |D|^{-1} \sum_{j \in D} \frac{\#\{m : \mathrm{sgn}\, \Delta g_j(m) = -d_j\}}{\#\{m : \Delta g_j(m) \neq 0\}} \tag{6}$$

MVR = 0 is the property a procurement auditor needs: no region of the input space in which improving a ratio is predicted to worsen the supplier.

Explanation Stability Index (ESI) asks whether the driver ranking is reproducible. We draw B bootstrap resamples of the training fold, refit, and recompute the global ranking r(b) on the fixed test set. Then

$$\mathrm{ESI} = \binom{B}{2}^{-1} \sum_{a<b} \tau\left(r^{(a)}, r^{(b)}\right) \tag{7}$$

with τ Kendall's rank correlation over all features. Because a decision memo cites only a handful of drivers, we also report Jaccard overlap of the top-k sets for k ∈ {5, 15}. ESI is computed identically for TreeSHAP, split-gain and permutation importance, which makes those three directly comparable for the first time on this task.

3.5 Cost-Sensitive Decision Layer

Let CFN be the loss from admitting a supplier that fails and CFP the loss from rejecting one that would have performed, and ρ = CFN/CFP. The expected cost of threshold τ is

$$\mathrm{EC}(\tau) = n^{-1}[C_{\mathrm{FN}} \cdot \mathrm{FN}(\tau) + C_{\mathrm{FP}} \cdot \mathrm{FP}(\tau)] \tag{8}$$

A model is only useful if it beats the best policy that ignores it — admit everyone or reject everyone — whose cost is min(CFNπ, CFP(1−π)) at base rate π. We therefore report the normalised expected cost

$$\mathrm{NEC} = \frac{\mathrm{EC}(\tau^*)}{\min\left(C_{\mathrm{FN}} \cdot \pi\,,\, C_{\mathrm{FP}} \cdot (1-\pi)\right)} \tag{9}$$

where τ^* is chosen on the validation fold only. NEC < 1 means the screening policy adds value; NEC = 1 means the optimal policy degenerates to the trivial one. This yields a diagnostic we use throughout: the break-even cost ratio $\rho^* = \min\{\rho : \mathrm{NEC}(\rho) \geq 1\}$, the asymmetry beyond which a tier stops paying for itself.

4 EXPERIMENTAL SETUP

4.1 Datasets

Table 1 Datasets

Property	Tier A	Tier B	Tier C
Source	TEJ [30]	DataCo [20]	STPC (ours)
Nature	real	real	synthetic
Units	6,819 firms	180,519 orders	6,000 suppliers
Features	95	24	20
Positive rate	3.23%	54.8%	15.2%
Period	1999–2009	2015–2017	—
Split	60/20/20	out-of-time	56/14/30
ρ (default)	20	3	20

Tier A uses the Taiwan Economic Journal panel of 6,819 listed companies described by 95 financial ratios, of which 220 (3.23%) entered bankruptcy as defined by Taiwan Stock Exchange regulation [30]. One constant-valued indicator is dropped, leaving 94 features. We use a stratified 60/20/20 train/validation/test split, see Table 1.

Tier B uses the DataCo order-level supply-chain dataset: 180,519 order lines from January 2015 to September 2017 with a binary late-delivery-risk label [20]. We report two cohorts: the full dataset, temporally subsampled to 56,055 lines for tractability on a single-core budget, and a technology/electronics hardware sub-cohort of 6,401 lines that is the closest available proxy for equipment procurement. We state plainly that this is a distribution-and-retail operation, not telecom procurement; it is used as a large real cross-domain benchmark for the operational tier and for the leakage audit, and no telecom-specific claim is drawn from it.

4.2 Leakage-Free Pre-Award Protocol

The Tier B label is, by construction, an indicator that realised shipping duration exceeded the scheduled window. Any model retaining realised shipping days, delivery status, shipping date or order status is reading the answer [21]. We exclude all four and every derivative, keeping only information available when the order is placed: scheduled shipping window, quantity, order value, discount rate, unit price, profit ratio, shipping-mode indicators, payment-type indicators, customer-segment indicators, calendar features and destination coordinates. Lane, product-category, market and department history enter as target encodings fitted on the training window only, with five-fold out-of-fold encoding inside that window and additive smoothing (prior weight 20) to prevent both temporal and within-fold leakage. The split is out-of-time — train before July 2016, validate July–December 2016, test 2017 onwards — which mirrors deployment and prevents a model from interpolating within a period it has already seen.

4.3 Synthetic Telecom Procurement Cohort

Table 2 STPC True Structural Coefficients

Attribute (block)	β	d_j
financial distress score (F)	+1.30	+1
on-time delivery rate 24m (O)	−0.95	−1
lead-time coeff. of variation (O)	+0.80	+1
geopolitical exposure (G)	+0.70	+1
lead-time mean days (O)	+0.60	+1
capacity headroom (O)	−0.60	−1
segment HHI concentration (G)	+0.55	+1
security assurance level (S)	−0.55	−1
defect rate ppm, log (O)	+0.50	+1
single-source flag (G)	+0.45	+1
component criticality (C)	+0.45	+1
export-control exposure (G)	+0.40	+1
cyber incidents 36m (S)	+0.40	+1
sub-tier visibility depth (S)	−0.35	−1
ESG compliance score (C)	−0.35	−1
buyer revenue share (C)	+0.30	+1
noise_1 … noise_4	0.00	0

Note: Interactions: single-source × geopolitical (0.65); lead-time CV × criticality (0.55); financial distress × buyer revenue share (0.50).

No public dataset exposes the attributes that actually distinguish telecom procurement risk. We therefore specify a generative process for a Synthetic Telecom Procurement Cohort (STPC) of 6,000 supplier records over 16 informative attributes spanning five blocks — financial, operational, geopolitical, security and commercial — plus four standard-normal null features. Coefficients are given in Table 2. Continuous attributes are standardised, defect rate enters in logs, the latent log-odds adds three interaction terms with weights 0.65, 0.55 and 0.50 plus Gaussian noise of scale 0.35, and the intercept is set so the disruption rate is 12%. Severity is drawn as lognormal contract value times criticality tier and is never used as a predictor.

Two design choices matter. First, the cohort is coupled to real data: the financial-distress attribute is resampled from the out-of-sample predictions of the fitted Tier A model on the real firm panel, so the financial tier's empirical distribution propagates into the fused setting. Second, the ground truth is known: we know which features matter, with what sign, in what interaction, and which are pure noise. This permits the driver-recovery validation of Section 5.4, which is not possible on any real dataset. The cohort is explicitly synthetic; it is used to validate the pipeline and to explore the telecom-attribute space, never to make an empirical claim about real suppliers.

4.4 Models, Tuning and Metrics

We compare nine baselines — logistic regression, Gaussian naïve Bayes, decision tree, random forest, extremely randomised trees, AdaBoost, gradient boosting, a two-layer MLP, and unconstrained XGBoost — against the constrained model. Linear and neural baselines are standardised; all class-weighted learners use balanced weights. XGBoost hyperparameters come from random search over depth, learning rate, ensemble size, subsampling, column sampling, minimum child weight, L2 penalty and split-gain threshold, selected on the validation fold (20 draws for Tier A; 25 with 40-round early stopping for Tier B). The constrained model reuses the unconstrained model's selected configuration exactly, so the comparison isolates the effect of the constraints rather than of tuning. The Tier A selection was depth 5, learning rate 0.02, 700 trees, subsample 0.8, column sample 0.6, minimum child weight 1, $\lambda = 5.0$, $\gamma = 0.5$.

We report AUC with 95% percentile bootstrap intervals, PR-AUC, the Kolmogorov–Smirnov statistic, F1, Matthews correlation, geometric mean of sensitivity and specificity, the Brier score, and NEC. Model comparisons use a paired bootstrap on the AUC difference (2,000 resamples for Tier A). All randomness is seeded at 42.

5 RESULTS AND DISCUSSION

5.1 Tier A: Financial-Viability Risk

Table 3 Tier A Test Performance (Real Firm Panel, $\rho = 20$)

Model	AUC	PR-AUC	KS	MCC	Brier	NEC
Gaussian NB	.623	.046	.286	.034	.783	1.315
Decision tree	.828	.306	.630	.361	.085	.449
MLP	.835	.257	.572	.253	.029	.533
Logistic reg.	.913	.317	.734	.288	.091	.420
AdaBoost	.942	.430	.742	.323	.117	.356
Extra trees	.944	.518	.734	.323	.022	.366
GBDT	.948	.552	.752	.350	.021	.342
Random forest	.951	.460	.778	.367	.022	.361
XGBoost	.955	.540	.774	.408	.024	.330
XGBoost-Mono (ours)	.957	.576	.780	.405	.025	.332

Note: Best value per column in bold row where applicable. Thresholds tuned on the validation fold only.

Table 3 reports Tier A. The constrained model attains the best AUC (0.957, 95% CI [0.933, 0.976]), the best PR-AUC (0.576) and the best KS (0.780), with NEC 0.332 — a 66.8% reduction in expected procurement cost relative to the best model-free policy. Ranking by NEC rather than AUC reshuffles the middle of the table (AdaBoost overtakes extra trees and random forest despite a lower AUC), which is a small but concrete demonstration that discrimination and decision value are not interchangeable.

Table 4 Paired Bootstrap Tests, XGBoost-Mono vs. Baselines

Baseline	ΔAUC	p
Gaussian NB	+.335	<.001
Decision tree	+.130	<.001
MLP	+.123	<.001
Logistic regression	+.044	.001
AdaBoost	+.015	.004
Extra trees	+.014	.015
GBDT	+.010	.154
Random forest	+.006	.266
XGBoost	+.003	.254

We are deliberate about what the accuracy comparison does and does not show. Table 4 gives paired bootstrap tests. The constrained model is significantly better than logistic regression (ΔAUC +0.044, p = 0.001), extra trees (+0.014, p = 0.015), AdaBoost (+0.015, p = 0.004) and the three weak baselines (p < 0.001). It is not significantly better than unconstrained XGBoost (+0.003, p = 0.254), GBDT (+0.010, p = 0.154) or random forest (+0.006, p = 0.266). This is the honest and, we would argue, the interesting result: on a panel of this size the strong tree ensembles are statistically indistinguishable, so the case for the constrained model cannot rest on accuracy, see Figure 2. It rests on the fact that the constraints buy a large improvement in explanation quality at no accuracy cost — the point of Section 5.5.

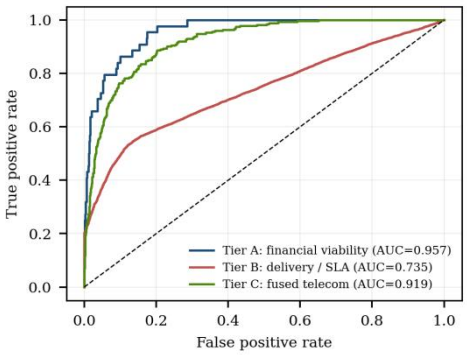

Figure 2 Test ROC Curves for the Constrained Model on All Three Tiers

Note: The spread reflects genuine differences in task difficulty, not in model capacity: Tier B is intrinsically hard once post-shipment information is removed.

5.2 Tier B and the Leakage Audit

Under the leakage-free protocol the operational tier is much harder than the literature suggests. On the technology sub-cohort the constrained model reaches AUC 0.742 and PR-AUC 0.824; on the full cohort 0.733 and 0.808. All learners cluster between 0.726 and 0.751, with a pruned decision tree nominally highest on the sub-cohort (0.751) — an unsurprising outcome given roughly 3,200 training rows and 24 features, and one we do not interpret as evidence of anything.

Table 5 explains why prior figures are so much higher. Restoring a single column — realised shipping days — to an otherwise identical pipeline moves AUC from 0.733 to 0.992, PR-AUC from 0.808 to 0.993, KS from 0.409 to 0.938 and NEC from 1.000 to 0.062. The 25.9-point AUC gap is entirely an artefact: that column is one of the two quantities whose comparison defines the label, so the model is not predicting a delay but reporting one. Any pre-award screening system calibrated on such figures would be badly mis-specified, and any benchmark comparison that mixes leaky and clean pipelines is meaningless. We therefore treat AUC ≈ 0.73 as the realistic ceiling for this feature set and this label, see Figure 3.

Table 5 Leakage Audit on the Operational Tier

Metric	Leakage-free	Leaky
AUC	.733	.992
PR-AUC	.808	.993
KS	.409	.938
F1	.706	.975
MCC	.000	.944
NEC (ρ = 3)	1.000	.062

Note: Identical learner, split and encodings; the leaky column is realised shipping days.

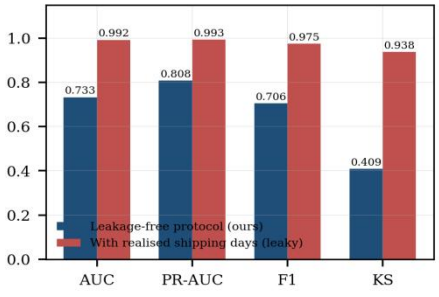

Figure 3 The Leakage Artefact

Note: Restoring one post-shipment column inflates every metric; the apparent skill is arithmetic, not prediction.

A second, less obvious finding follows from the cost analysis. With a base rate of 54.6% and ρ = 3, the model-free policy of hedging every shipment is already near-optimal, so NEC ≈ 1.000: the operational tier has almost no

standalone decision value at that asymmetry. Section 5.7 quantifies this as a break-even ratio of ρ* = 2. This is not a failure of the model but a property of the problem — when breaches are common and their consequences severe, universal hedging dominates selective hedging — and it is precisely the argument for fusing the operational tier with a rare-event tier rather than deploying it alone.

5.3 Tier C: Value of the Two-Tier Architecture

On the fused telecom cohort the constrained model reaches AUC 0.919 and the best PR-AUC of any learner (0.703), against an estimated Bayes-optimal AUC of 0.943 given the injected noise — so the pipeline recovers roughly 97% of the attainable discrimination. The MLP is nominally highest on AUC (0.924), which is expected: the generative process is a noisy generalised linear model with three multiplicative terms, a functional form smooth learners fit naturally. The constrained ensemble nonetheless wins on PR-AUC, the metric that matters at a 15% base rate.

Table 6 Tier C Architecture Ablation

Variant	AUC	PR-AUC	NEC
Full (two-tier)	.919	.703	.519
w/o security block	.904	.647	.541
w/o geopolitical block	.855	.538	.713
w/o financial tier	.848	.524	.673
w/o operational block	.801	.476	.930

Table 6 ablates the architecture. Removing the financial tier costs 7.1 AUC points and 17.9 PR-AUC points; removing the operational block costs 11.8 and 22.8, and pushes NEC to 0.930, close to worthlessness; the geopolitical block is worth 6.5 AUC points, the security block 1.6. Each block therefore carries non-redundant signal, and the two-tier decomposition of Section 3.1 is empirically justified rather than merely tidy. The ordering is also managerially informative: the two blocks that dominate — operational history and financial viability — are the two that a buyer can actually monitor continuously, whereas the geopolitical block, though substantial, is largely exogenous, see Figure 4.

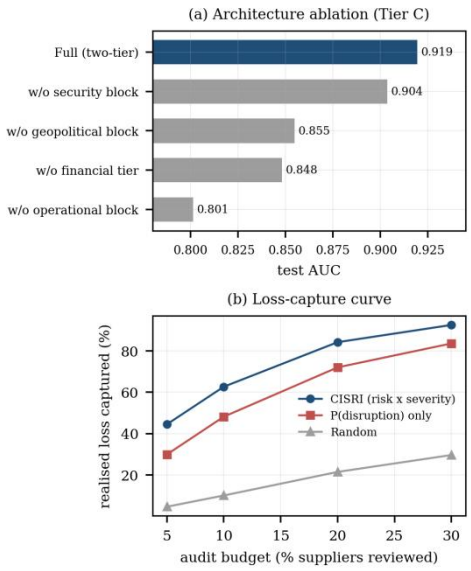

Figure 4 (a) Architecture Ablation; (b) Loss-Capture Curves

Note: (a) Every block carries non-redundant signal, with the operational and financial blocks dominating. (b) Ranking suppliers by CISRI (risk × severity) dominates ranking by risk probability alone at every audit budget, with the gap widest where budgets are tightest.

5.4 Ground-Truth Driver and Interaction Recovery

Table 7 Ground-Truth Driver Recovery (Tier C, k = 16)

Method	P@16	R@16	False drv.	Best null rank
TreeSHAP	1.000	1.000	0	17
XGBoost gain	1.000	1.000	0	17
Permutation	1.000	1.000	0	17

Because the STPC ground truth is known, we can ask directly whether the pipeline discovers it. Table 7 reports driver recovery at k = 16, the true number of informative attributes. All three importance measures achieve precision, recall and F1 of 1.000 with zero false drivers: every one of the four injected null features is ranked 17th or worse, with mean

null rank 18.5. The pipeline does not manufacture spurious risk drivers, which is the minimum an auditable system must guarantee.

Recovery of the ordering is stronger still. The correlation between true coefficient magnitude $|\beta|$ and mean $|SHAP|$ is $r = 0.953$ ($p = 1.1\times10^{-8}$) with Spearman $\rho = 0.893$ ($p = 3.2\times10^{-6}$). TreeSHAP is therefore not merely selecting the right features but ranking them in close proportion to their true structural weight — which is what licenses using a SHAP ranking as the basis for prioritising mitigation spend.

Table 8 SHAP Interaction Screening (Tier C)

Feature pair	$\lvert\phi_{ij}\rvert$	True?
single-source × geopolitical	.099	yes
lead-time CV × criticality	.082	yes
fin. distress × buyer rev. share	.063	yes
fin. distress × on-time rate	.062	no
fin. distress × lead-time CV	.051	no

Table 8 reports interaction screening. The three highest-scoring pairs are exactly the three interactions written into the generative process, in the correct order of magnitude, and the first false positive (0.062) falls only marginally below the weakest true interaction (0.063). Interaction detection is therefore reliable at the top of the ranking but not sharply separated immediately below it — a caution worth stating for practitioners who might read the fourth-ranked pair as a finding.

The recovered interactions are substantively interpretable and align with procurement doctrine. Single-sourcing amplifies geopolitical exposure: neither is dangerous alone, but a sole-source supplier in an exposed jurisdiction is far riskier than the sum of the two. Lead-time volatility matters more for high-criticality components, because there is no buffer to absorb it. And financial distress compounds with buyer-revenue dependence, since a distressed supplier that depends on this buyer is simultaneously more likely to fail and more likely to fail in a way that cannot be replaced. These are exactly the conjunctions that a marginal-effects model would miss.

5.5 Explanation Coherence: The Central Result

Table 9 Explanation-Quality Audit

Setting / model	AUC	ECR	LCR	MVR
Tier A unconstrained	.9547	.870	.614	.345
Tier A constrained (ours)	.9573	.978	.676	.000
Tier C unconstrained	.9169	1.000	.871	.091
Tier C constrained (ours)	.9195	1.000	.893	.000

Note: Tier A audit covers 46 theory-constrained features (45 for MVR) and 60,518 instance-feature pairs. Higher is better for ECR and LCR; lower for MVR.

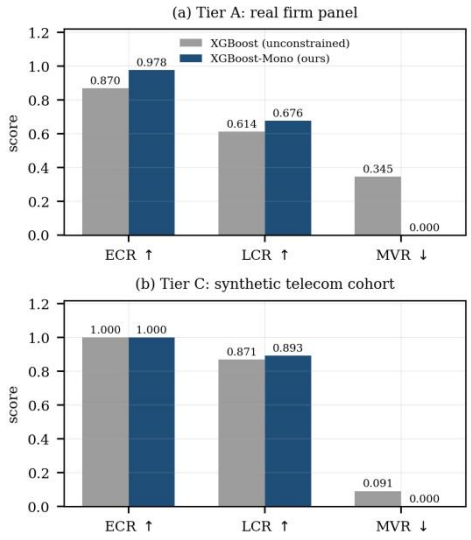

Figure 5 Explanation-Quality Audit

Note: On real data the unconstrained model violates elicited sign priors on 13% of covered features and 34.5% of decision-rule steps; constraints remove violations entirely at no accuracy cost. On the synthetic cohort ECR saturates because the true structure is monotone by construction, and MVR remains the discriminating metric.

Table 9 and Figure 5 give what we regard as the paper's central finding. On the real firm panel, the unconstrained model — despite an excellent AUC of 0.955 — gets the global direction wrong for 13.0% of the features on which accounting theory is unambiguous (ECR 0.870), and its decision rule moves the wrong way on 34.5% of partial-dependence steps

(MVR 0.345). In other words, a buyer who reads its explanations and acts on them would, on roughly a third of the audited transitions, be told that improving a ratio worsens the supplier. Imposing the sign priors raises ECR to 0.978 (+10.9 points), raises instance-level LCR from 0.614 to 0.676 (+6.3 points over 60,518 attributions), and eliminates monotonicity violations entirely (MVR exactly 0.000). AUC does not fall; it rises marginally, from 0.9547 to 0.9573, and PR-AUC rises from 0.540 to 0.576.

The synthetic panel serves as a control. There ECR saturates at 1.000 for both models, exactly as it should: the generative structure is monotone by construction, so a well-fitted model has little opportunity to violate the priors globally. MVR still separates the two (0.091 versus 0.000) and LCR still improves (0.871 to 0.893). The comparison of the two panels is informative in itself — ECR discriminates when the underlying relationships are noisy, high-dimensional and partly non-monotone, which is the real-data regime, while MVR discriminates in both. Practitioners auditing a model should report both.

Why does enforcing 46 constraints not cost accuracy? The constraints are not arbitrary regularisation: they encode relationships that genuinely hold in the data-generating process. Restricting the hypothesis class to functions consistent with them removes capacity that the unconstrained learner was spending on noise, which is why the effect on held-out discrimination is slightly positive rather than negative. This is the practical case for monotone constraints in procurement: they are close to free, and they convert an explanation from a plausible narrative into an auditable one, see Figure 6.

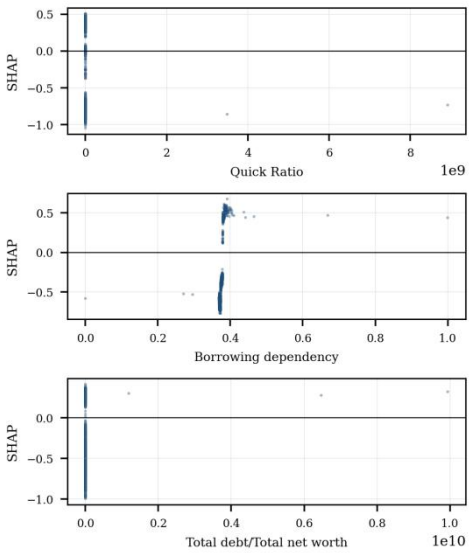

Figure 6 SHAP Dependence for the Three Strongest Tier A Drivers under the Constrained Model
Note: Attribution is monotone in the feature throughout its range and signed as accounting theory requires: liquidity reduces risk, borrowing dependency and leverage increase it.

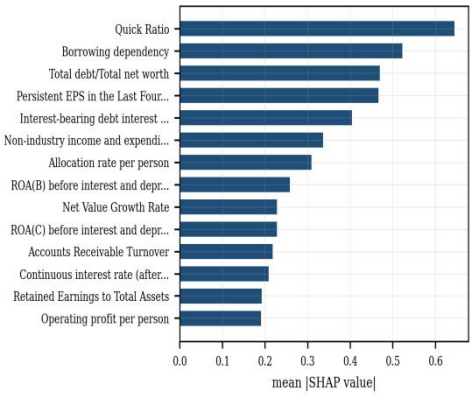

Figure 7 Global Driver Importance, Tier A Constrained Model (Top 14 by Mean |SHAP|)

The recovered driver set (Figure 7) is economically coherent. Quick ratio dominates (mean |SHAP| 0.645), followed by borrowing dependency (0.522), total-debt-to-net-worth (0.469) and persistent four-quarter EPS (0.466). Signed dependence is uniformly correct: liquidity and earnings persistence reduce risk (Spearman −0.49 and −0.92 respectively), leverage and borrowing dependency increase it (+0.92 and +0.79). Notably, split-gain importance ranks quick ratio only 10th while ranking it 1st by SHAP — a discrepancy that matters, since gain counts how often a feature was split on rather than how much it moves predictions.

5.6 Explanation Stability

Table 10 Explanation Stability Index (Tier A, B = 14, 91 Pairs)

Method	ESI (τ)	SD	J@5	J@15
TreeSHAP	.469	.069	.446	.445
XGBoost gain	.480	.054	.247	.487
Permutation	.405	.070	.339	.475

Table 10 and Figure 8 report refit stability. Over the full 94-feature ranking the three methods are close (τ between 0.405 and 0.480), and split-gain is nominally highest. The picture inverts at the head of the ranking, which is the only part that reaches a procurement memo: TreeSHAP retains 44.6% of its top-5 drivers across refits versus 24.7% for split-gain — an 81% relative improvement — with permutation importance in between at 33.9%. Split-gain's high full-ranking τ is driven by consistent ordering in the long, low-importance tail; its top-5 set is the least reproducible of the three.

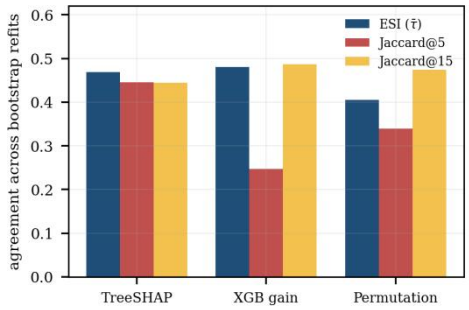

Figure 8 Refit Stability

Note: Full-ranking Kendall τ is similar across methods, but TreeSHAP's top-5 driver set is far more reproducible — the property that matters for a decision that must be defended later.

Two implications follow. First, the widespread practice of reporting split-gain importance because it is free with the model is not defensible when the output will be cited in a decision record. Second, the absolute numbers are sobering: even TreeSHAP retains under half its top-5 set across bootstrap refits on a panel with 220 positive cases. This is a property of the data, not of the attribution method — with a rare outcome and correlated ratios many features are near-substitutes — but it means single-run SHAP rankings should be reported with stability intervals. We recommend that ESI be computed and disclosed alongside any supplier-risk explanation intended for audit.

5.7 Cost-Sensitive Decision Value

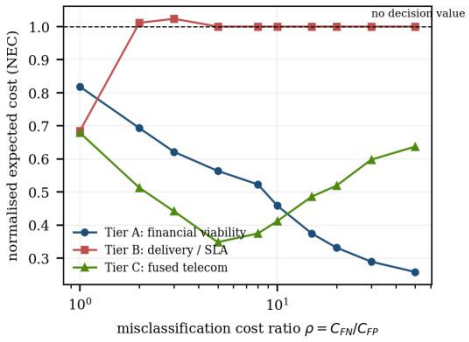

Figure 9 Normalised Expected Cost Against Cost Asymmetry

Note: Below the dashed line the model beats the best model-free policy. The financial tier improves monotonically with asymmetry; the operational tier breaks even at ρ* = 2; the fused telecom setting is optimal in the mid range.

Figure 9 sweeps ρ from 1 to 50 with τ* re-optimised on validation at each point. The three tiers behave qualitatively differently, and the differences are managerially meaningful.

The financial tier improves monotonically with asymmetry, from NEC 0.818 at ρ = 1 to 0.258 at ρ = 50 — a 74% expected-cost reduction. Rare-event screening is exactly the regime in which a model earns its keep: the base rate is 3.2%, so the trivial policy of admitting everyone is already cheap, and the model's value comes from finding the few failures before they are admitted. There is no break-even point below ρ = 50.

The operational tier behaves in the opposite way, and this is the most policy-relevant negative result in the paper. NEC is 0.685 at ρ = 1 but exceeds 1 from ρ = 2 onward: ρ* = 2. With a 55% breach base rate, once a breach is even twice as costly as an unnecessary hedge, the optimal policy is to hedge everything, and selective screening cannot improve on it.

A procurement organisation that deployed an SLA-breach classifier as a standalone gate on critical infrastructure would therefore be spending model-development budget to reproduce a decision it could reach by policy. The correct use of the operational tier is as an input to a fused index and to capacity planning, not as a gate.

The fused telecom setting is intermediate and non-monotone, with NEC minimised at 0.348 near $\rho = 5$ and rising to 0.637 at $\rho = 50$. The interior optimum reflects the 15% base rate: at low asymmetry the model discriminates usefully, while at very high asymmetry the trivial policy again becomes competitive. The practical reading is that a fused index is most valuable at moderate asymmetries, and that organisations should estimate their own ρ before assuming a screening model will pay for itself.

5.8 CISRI and Audit-Budget Efficiency

Risk probability is not the quantity a procurement committee should rank on, because a likely failure of a commodity part matters less than an unlikely failure of a single-sourced core component. Figure 4(b) evaluates the CISRI ranking of (1) against realised loss. At a 5% audit budget CISRI captures 44.4% of realised loss versus 29.8% for probability ranking and 4.5% for random selection; at 10%, 62.7% versus 48.2% and 10.0%; at 20%, 84.3% versus 72.1%. The advantage is largest where budgets are tightest — a 49% relative improvement at 5% coverage, narrowing to 17% at 20% coverage — which is the regime real supplier-audit programmes operate in. Because severity is observable ex ante from contract value and criticality tier, this improvement requires no additional modelling, only the decision to rank on expected loss.

5.9 Implications for Practice

Four recommendations follow directly from the results. Elicit and enforce sign priors. Constraints cost nothing measurable in accuracy and are the only mechanism we tested that guarantees MVR = 0, which is the property an auditor will actually test. Audit explanations, do not merely display them. ECR, LCR and MVR are cheap to compute and would have flagged the unconstrained model as unsuitable for procurement use despite its excellent AUC. Report stability. A driver ranking cited in a decision record should carry an ESI figure. Estimate ρ before building. The break-even analysis shows that whether a screening model can help at all depends jointly on base rate and cost asymmetry, and can be determined before any model is trained.

5.10 Threats to Validity

We state the limitations plainly. (i) Domain gap. No real telecom procurement risk dataset is publicly available. Tier A is a proxy population justified by the ICT-manufacturing composition of Taiwan's listed universe but not telecom-specific; Tier B is a distribution operation, used only as a cross-domain benchmark and for the leakage audit. Telecom-specific attributes are studied only in the synthetic cohort. (ii) Synthetic ground truth. The recovery results in Section 5.4 validate the pipeline against a generative process we specified; they show the method works when the truth has that form, not that real supplier risk has that form. (iii) Vintage. The financial panel covers 1999–2009 and predates the current geopolitical regime; the absolute risk levels it implies should not be transported. (iv) Prior elicitation. Sign priors were assigned by the authors from accounting theory rather than by a panel of procurement experts; a mis-specified prior would be enforced as confidently as a correct one, so elicitation should be documented and contestable in deployment. (v) Compute budget. Single-core constraints limited random search to 20–25 draws, ESI to 14 bootstrap refits and the full Tier B cohort to a temporally stratified 56,055-line subsample. Larger budgets would tighten intervals but are unlikely to reverse the qualitative findings, all of which are large in magnitude.

6 CONCLUSION

We presented CTI-SRP, a two-tier explainable framework for supplier risk in critical telecommunications infrastructure procurement, and evaluated it on two real datasets and one transparently specified synthetic cohort. Three findings stand out. Encoding 46 accounting-theory sign priors as monotone constraints raises explanation coherence from 0.870 to 0.978 and eliminates decision-rule monotonicity violations entirely, at no cost in discrimination — explanations can be made auditable for free. TreeSHAP is substantially more stable than split-gain importance where it matters most, at the head of the driver ranking (Jaccard@5 0.446 versus 0.247), though even it retains under half its top-5 set across refits, so stability should be disclosed rather than assumed. And the standard public delivery-risk benchmark is contaminated: the same learner scores 0.992 with a post-shipment column and 0.733 without it, so the near-perfect accuracies in the literature are arithmetic rather than prediction.

Beyond accuracy, the cost analysis shows that whether a supplier-risk model adds decision value at all depends jointly on base rate and cost asymmetry: rare-event financial screening improves monotonically with asymmetry (NEC 0.258 at $\rho = 50$), whereas high-base-rate delivery screening breaks even at $\rho^* = 2$ and should not be deployed as a standalone gate. Ranking on expected loss rather than probability captures 62.7% of realised loss within a 10% audit budget against 48.2% for probability alone, at no modelling cost.

Future work should pursue three directions: eliciting sign priors from procurement expert panels with documented disagreement rather than from a single analyst; extending the framework to multi-tier supply networks where graph structure carries risk that firm-level features cannot [8, 24]; and validating the audit metrics against how procurement

professionals actually assess explanation trustworthiness, since ECR, LCR, MVR and ESI are constructed to be normatively sensible but have not yet been shown to predict human trust calibration.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Radu R, Amon C. The governance of 5G infrastructure: between path dependency and risk-based approaches. Journal of Cybersecurity, 2021, 7(1): tyab017.
[2] European Union Agency for Cybersecurity (ENISA). ENISA Threat Landscape for 5G Networks. Athens: ENISA, 2020.
[3] Enduring Security Framework. Potential Threat Vectors to 5G Infrastructure. NSA, CISA and ODNI, 2021.
[4] Ho W, Zheng T, Yildiz H, et al. Supply chain risk management: a literature review. International Journal of Production Research, 2015, 53(16): 5031-5069.
[5] Baryannis G, Validi S, Dani S, et al. Supply chain risk management and artificial intelligence: state of the art and future research directions. International Journal of Production Research, 2019, 57(7): 2179-2202.
[6] Baryannis G, Dani S, Antoniou G. Predicting supply chain risks using machine learning: the trade-off between performance and interpretability. Future Generation Computer Systems, 2019, 101: 993-1004.
[7] Brintrup A, Pak J, Ratiney D, et al. Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing. International Journal of Production Research, 2020, 58(11): 3330-3341.
[8] Kosasih E E, Brintrup A. A machine learning approach for predicting hidden links in supply chain with graph neural networks. International Journal of Production Research, 2022, 60(17): 5380-5393.
[9] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.
[10] Friedman J H. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 2001, 29(5): 1189-1232.
[11] Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data?. In: Advances in Neural Information Processing Systems, 2022, 35: 507-520.
[12] Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Information Fusion, 2022, 81: 84-90.
[13] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, 2017, 30: 4765-4774.
[14] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2(1): 56-67.
[15] Cano J-R, Gutiérrez P A, Krawczyk B, et al. Monotonic classification: an overview on algorithms, performance measures and data sets. Neurocomputing, 2019, 341: 168-182.
[16] Chen C-C, Li S-T. Credit rating with a monotonicity-constrained support vector machine model. Expert Systems with Applications, 2014, 41(16): 7235-7247.
[17] Chen D, Ye W. Monotonic neural additive models: pursuing regulated machine learning models for credit scoring. In: Proceedings of the 3rd ACM International Conference AI in Finance, 2022: 70-78.
[18] Slack D, Hilgard S, Jia E, et al. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of AAAI/ACM Conference AI, Ethics, and Society, 2020: 180-186.
[19] Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artificial Intelligence, 2021, 298: 103502.
[20] Constante F, Silva F, Pereira A. DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS. Mendeley Data, V5, 2019. DOI: 10.17632/8gx2fvg2k6.5.
[21] Kaufman S, Rosset S, Perlich C, et al. Leakage in data mining: formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 2012, 6(4): Art. 15.
[22] Elkan C. The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2001: 973-978.
[23] Drummond C, Holte R C. Cost curves: an improved method for visualizing classifier performance. Machine Learning, 2006, 65(1): 95-130.
[24] Brintrup A, Wichmann P, Woodall P, et al. Predicting hidden links in supply networks. Complexity, 2018, 2018: Art. ID 9847925.
[25] Handfield R B, Sun H, Rothenberg L. Assessing supply chain risk for apparel production in low cost countries using newsfeed analysis. Supply Chain Management: An International Journal, 2020, 25(6): 803-821.
[26] Burstein G, Zuckerman I. Deconstructing risk factors for predicting risk assessment in supply chains using machine learning. Journal of Risk and Financial Management, 2023, 16(2): Art. 97.
[27] Altman E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 1968, 23(4): 589-609.
[28] Ohlson J A. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 1980, 18(1): 109-131.

[29] Barboza F, Kimura H, Altman E. Machine learning models and bankruptcy prediction. Expert Systems with Applications, 2017, 83: 405-417.
[30] Liang D, Lu C-C, Tsai C-F, et al. Financial ratios and corporate governance indicators in bankruptcy prediction: a comprehensive study. European Journal of Operational Research, 2016, 252(2): 561-572.
[31] Ribeiro M T, Singh S, Guestrin C. '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, 2016: 1135-1144.
[32] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1(5): 206-215.
[33] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82-115.
[34] Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd ed. 2022.
[35] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 2015, 10(3): e0118432.