Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 3, No. 1, 2026

Robust Explainable Supplier Risk Prediction under Imperfect Domain Knowledge: Empirically Vetted Monotonic XGBoost with TreeSHAP

Jiawen Xu, Chenhao Liu *

College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China

Abstract: Monotonic constraints are attractive for supplier-risk models because they can force predictions to respect domain knowledge, but a wrong expert sign is then enforced everywhere. This paper studies that overlooked failure mode and proposes empirically vetted monotonic XGBoost (EV-Mono), in which domain experts still supply candidate directions while training data are allowed only to veto, never reverse, a candidate constraint. A training-only bootstrap partial-dependence gate is combined with exact TreeSHAP, a prior-misspecification stress test (PMST), and a new Explanation Intervention Consistency (XIC) metric that checks whether risk-reducing feature interventions move the corresponding attribution in the expected direction. Controlled experiments use 12,000 fully synthetic supplier records with 18 known monotone risk drivers, six null variables, nonlinear main effects and sign-preserving interactions; an adverse covariate-shift test contains 3,000 additional records. With correct priors, EV-Mono attains AUC 0.947, zero monotonicity violations, XIC 1.000, and places all 18 informative features in the top 18 TreeSHAP ranking. Under 40% deliberately corrupted priors, hard monotonic XGBoost falls to AUC 0.932, explanation-coherence rate 0.611 and monotonicity-violation rate 0.343; EV-Mono rejects all seven wrong signs and retains AUC 0.946, coherence 1.000 and violation rate 0.091. The study shows that domain constraints should be auditable inputs rather than unquestioned rules.

Keywords: Supplier risk, explainable artificial intelligence, monotonic XGBoost, TreeSHAP, domain knowledge, prior misspecification, trustworthy machine learning.

1. Introduction

Supplier qualification increasingly depends on heterogeneous evidence: financial ratios, delivery history, capacity signals, cyber-security indicators, single-source exposure and geopolitical concentration. The resulting decision is asymmetric. A false negative can propagate into missed deliveries, production interruption or infrastructure downtime, while a false positive can unnecessarily reduce competition. Supply-chain risk research therefore emphasizes both anticipatory assessment and the integration of multiple risk channels [2-5], and tree ensembles are a natural fit for the nonlinear interactions present in tabular procurement data [6-8].

The difficulty is not only predictive. A procurement analyst must be able to defend why a supplier was escalated and whether a proposed mitigation would move risk in the expected direction. SHAP has become a standard attribution method for tree ensembles because it decomposes a prediction into feature contributions [9], [10]. However, post-hoc explanations inherit the behavior of the fitted model. If the learner has discovered a locally perverse relation—for example, larger debt appearing to reduce failure risk—the explanation can faithfully describe a model that is directionally implausible.

Monotonic constraints offer a direct remedy by restricting the hypothesis class so that designated features can move the prediction in only one direction [11], [12]. Recent supplier-risk work has shown the practical value of coupling monotonic XGBoost with TreeSHAP and explanation-quality audits, providing a strong foundation for auditable procurement screening [1]. Yet the constraint itself creates a second-order risk: the model treats an incorrect sign prior with exactly the same certainty as a correct one. In real organizations, priors may be elicited from different functions, change with contract context, or encode proxies whose semantics are not as clean as an accounting identity.

This paper shifts the research question from "Do monotonic constraints improve explanation coherence?" to "What happens when some of the domain constraints are wrong, and can the system detect that before deployment?" The proposed EV-Mono architecture preserves expert ownership of direction: data are not allowed to invent or flip a monotonic sign. Instead, a training-only bootstrap diagnostic can veto a candidate sign whose observed partial-dependence direction is repeatedly inconsistent with the stated prior. This asymmetric design is intended for governance settings where empirical evidence may challenge domain knowledge but should not silently overwrite it.

The contributions are fourfold. First, we introduce an empirically vetted monotonic XGBoost procedure that screens candidate sign constraints through bootstrap partial-dependence agreement. Second, we propose PMST, a nested stress protocol that deliberately corrupts 0–40% of otherwise valid priors and compares hard enforcement with vetted enforcement under paired random seeds. Third, we introduce XIC, an explanation test based on risk-reducing feature interventions rather than global ranking alone. Fourth, we release a fully specified synthetic supplier benchmark and all code, allowing every reported number to be regenerated and the ground truth of both informative and null features to be checked directly. The central result is not an accuracy win on clean data; it is graceful degradation when domain knowledge is imperfect.

2. Related Work

2.1. Supplier and Financial Risk Prediction

Supplier-risk analytics spans expert assessment, conventional statistical models and increasingly machine-learning systems. Reviews emphasize that supply-chain risk is multi-dimensional and that risk identification must connect operational evidence to managerial decisions [2], [3]. Baryannis et al. explicitly study the performance–interpretability trade-off for supply-chain risk prediction [4], while Brintrup et al. demonstrate the value of data-driven supplier-disruption modeling in complex asset manufacturing [5]. Text-derived early-warning signals provide another complementary channel [28], and supplier-development research shows why the buyer–supplier relationship itself can affect performance outcomes [30-36].

Financial distress is a direct supplier-risk mechanism. Classical work established ratio-based bankruptcy prediction [20], [21], and later machine-learning studies report material gains from nonlinear models [22], [23]. The Taiwanese bankruptcy corpus distributed through UCI contains 6,819 instances and 95 variables [24], and subsequent work has examined class imbalance on that dataset [29]. These resources are valuable predictive benchmarks, but they do not reveal an objective ground truth for whether a particular expert monotonic prior is correct in every modeled context. Because the present question is prior misspecification itself, the main experiment requires a controlled data-generating process in which the sign of each driver is known by construction.

2.2. Explainability, Monotonicity and Reliability

TreeSHAP provides exact Shapley-value attribution for tree ensembles and supports both local explanations and aggregate importance analysis [9], [10]. LIME [26] is another influential local surrogate approach. At the same time, high-stakes explainability research cautions that a plausible explanation is not automatically a faithful or safe one [15], [16]; explanation methods can be attacked [13], and feature dependence can alter Shapley-based interpretations [14]. These findings motivate evaluation of explanations as objects with measurable reliability rather than decorative plots.

Monotonic classification incorporates directional knowledge into model construction [11]. Applications in credit risk demonstrate the appeal of constraining a learner when institutional knowledge provides a defensible ordering [12]. XGBoost [6], built on gradient-boosted trees [7], is especially practical because monotonic constraints are supported directly while preserving nonlinear interactions. Yet standard evaluations usually assume that the supplied directions are correct. That assumption is precisely what PMST relaxes.

Two additional evaluation issues matter here. First, class imbalance makes PR-AUC informative alongside ROC-AUC [17]. Second, probability quality matters when risk scores are converted into interventions; calibration can differ substantially across algorithms [27]. We therefore report Brier score and expected calibration error in addition to discrimination. Finally, all screening diagnostics are fit on the training fold only, reflecting the broader requirement to prevent information leakage into model development [19].

3. Problem Formulation and Method

3.1. Supplier-Risk Model with Candidate Sign Priors

Let x_i in [0, 1]^p denote normalized pre-decision supplier indicators and y_i in {0,1} denote a future disruption event. The unconstrained boosted model estimates p_i = P(y_i=1|x_i). For a subset D of features, a domain specialist supplies d_j in {-1, +1}. A positive sign means increasing x_j should not reduce predicted risk; a negative sign means increasing x_j should not increase predicted risk. Hard monotonic XGBoost enforces each supplied d_j during tree construction. This is auditable when d_j is correct, but an erroneous d_j becomes a global structural error rather than a local estimation error.

We assume that expert knowledge is authoritative as a proposal but not infallible as a constraint. Consequently, empirical evidence is used only to accept or reject a candidate direction. It is never used to change +1 into -1 or vice versa. Rejected features are left unconstrained and can be returned to the expert panel for review.

3.2. Empirically Vetted Monotonic XGBoost (EV-Mono)

For each candidate feature j, EV-Mono fits B unconstrained XGBoost models on bootstrap resamples of the training fold. For each refit, a partial-dependence profile is evaluated at G quantiles. Let Delta_{jbg} be the change in mean predicted risk between adjacent grid points for bootstrap b. The prior-consistency score is

$$A_j = (1 / N_j) \sum_{b,g} I[d_j * Delta_{jbg} > 0]$$

Where N_j counts non-negligible adjacent changes. The candidate constraint is retained if A_j >= tau. In the reported experiment, B=3, G=7 and tau=0.72. These values were fixed before inspecting test performance. A score below tau is a veto, not evidence for the opposite sign. All vetting uses the training set; the validation set is reserved for the classification threshold and the test set is untouched until final evaluation.

This procedure separates three governance roles. Domain specialists define the semantic direction. Bootstrap evidence measures whether the available training distribution repeatedly supports that direction. XGBoost then enforces only the vetted subset. The design therefore prevents the most problematic form of automated “knowledge correction,” in which a noisy dataset silently becomes the source of policy.

3.3. Prior Misspecification Stress Test (PMST)

PMST measures robustness to incorrect domain knowledge directly. Starting from the known true sign map, a fixed random ordering of the 18 informative features is generated. The first k signs are inverted to create nested corruption levels of 0%, 10%, 20%, 30% and 40% (0, 2, 4, 5 and 7 wrong signs after rounding). At each level, hard monotonic XGBoost enforces all candidate signs, while EV-Mono first applies the training-only veto gate. Competing models use the same train/validation/test split and, within a corruption level, the same XGBoost random seed, so the comparison isolates the constraint set rather than stochastic training variation.

3.4. Explanation Audits

Three explanation metrics are used. Explanation Coherence Rate (ECR) computes the Spearman correlation

between each feature and its held-out TreeSHAP attribution; a feature is coherent when the correlation has the true sign and absolute magnitude at least 0.05. Monotonicity Violation Rate (MVR) directly audits the prediction rule by evaluating nine-point partial-dependence curves and measuring the fraction of non-negligible adjacent steps that move against the true sign. ECR tests attribution direction; MVR tests the learned function [36-38].

XIC is designed to be more local. For each held-out supplier and constrained feature, we apply a small risk-reducing intervention of 0.08 on the normalized scale: an adverse feature is decreased and a protective feature is increased, with clipping to [0, 1]. We then recompute exact TreeSHAP and check whether the intervened feature's attribution to raw log-odds fails to increase. XIC is the fraction of valid interventions satisfying that inequality. XIC does not claim causal effect; it tests internal consistency between the displayed attribution and the model's own prescribed risk direction.

4. Experimental Design

4.1. Controlled Supplier Benchmark

The reported benchmark is fully synthetic and contains no real supplier, company or procurement record. This choice is methodological rather than cosmetic: evaluating false domain priors requires knowing which priors are false [39]. We generate 12,000 observations with 24 normalized features. Eighteen features are informative and assigned known monotone signs; six are independent null variables. Five correlated latent blocks represent financial, operational, structural/geopolitical, security and commercial conditions. Each observed feature is a noisy sigmoid transformation of its block, creating realistic correlation without making variables deterministic.

Table I. Controlled feature blocks and true directions

Block	Variables and risk direction
Financial (7)	5 protective (-); 2 adverse (+)
Operat. (5)	3 protective (-); 2 adverse (+)
Structural (2)	2 adverse (+)
Security (3)	2 protective (-); 1 adverse (+)
Comm. (1)	1 adverse (+)
Null (6)	6 no-effect variables

The latent log-odds combines nonlinear but monotone main effects whose absolute coefficients range from 0.45 to 1.20 with four sign-preserving interactions: single-source exposure x geopolitical exposure, lead-time variability x low capacity headroom, debt-to-assets x buyer-revenue dependence, and cyber incidents x low security assurance. Gaussian noise with standard deviation 0.25 is added before Bernoulli sampling. The intercept is calibrated to a 15% target event rate; the realized full-sample rate is 15.1%.

Data are split 60/20/20 into training, validation and test sets: 7,200/2,400/2,400 observations. A separate 3,000-record shift set increases latent financial, operational and geopolitical stress; its realized event rate is 27.1%. The shift set is not used for training or threshold selection. The complete generated datasets accompany the code, so the reported experiment can be checked at record level.

4.2. Models, Metrics and Reproducibility

We compare class-weighted logistic regression, random forest [25], unconstrained XGBoost, hard monotonic XGBoost, and EV-Mono. XGBoost uses 420 trees, depth 4, learning rate 0.035, subsampling 0.85 and column sampling 0.86; positive gradients are scaled by the training class ratio. The same configuration is used for unconstrained, hard-monotonic and vetted-monotonic models to isolate the role of constraints. Exact TreeSHAP values are obtained through XGBoost contribution prediction. The classification threshold maximizes F1 on the validation fold only.

Predictive metrics are ROC-AUC, PR-AUC, Brier score, 10-bin expected calibration error (ECE), F1 and Matthews correlation coefficient. For the central AUC comparisons we additionally report 2,000-resample paired bootstrap intervals on the fixed test set. Explanation metrics are ECR, MVR, XIC and precision at 18 (P@18) against the known informative-feature set. Finally, we perform four paired bootstrap refits for a descriptive Jaccard stability audit of the top-5 and top-10 SHAP feature sets. The random seed is 42 throughout. No test result is used to tune the vetting threshold or model hyperparameters.

5. Results

5.1. Clean-Prior Prediction and Covariate Shift

Table II. Test Performance with Correct Domain Priors

Model	AUC	PR	Brier	ECE	S-AUC
Logistic regression	0.948	0.810	0.094	0.133	0.935
Random forest	0.940	0.784	0.066	0.048	0.926
XGBoost	0.944	0.796	0.076	0.077	0.933
Hard-Mono	0.947	0.807	0.082	0.098	0.934
EV-Mono	0.947	0.807	0.082	0.098	0.934

Table II gives a deliberately non-triumphal result. Logistic regression is marginally best in clean discrimination (AUC 0.948, PR-AUC 0.810), while random forest has the best Brier score (0.066) and ECE (0.048). EV-Mono reaches AUC 0.947 and PR-AUC 0.807. Thus the proposed method is not justified by claiming universal predictive superiority. Its value is that the accuracy is competitive while the model obeys explicit knowledge constraints and can reject bad ones.

Compared with unconstrained XGBoost, clean EV-Mono improves AUC from 0.9441 to 0.9465. The paired bootstrap difference is 0.00247 with a 95% interval of [0.00053, 0.00461]. Under the adverse shift, EV-Mono AUC decreases to 0.9339; logistic regression is 0.9353. The shift experiment therefore shows reasonable ranking persistence, not immunity to distribution change. Calibration also worsens for every model under shift, which is important if procurement actions depend on absolute probabilities rather than ranking.

5.2. Explanation Reliability with Correct Priors

Table III. Explanation audit on the clean test set

Model	ECR	MVR	XIC	P@18	Nulls
XGBoost	1.000	0.125	0.831	1.000	0
Hard-Mono	1.000	0.000	1.000	1.000	0
EV-Mono	1.000	0.000	1.000	1.000	0

All three tree models recover the correct global attribution direction for the 18 known drivers (ECR 1.000), so ECR alone is not discriminating in this deliberately monotone data-generating process. MVR and XIC reveal the difference. Unconstrained XGBoost violates the true monotonic direction on 12.5% of audited partial-dependence steps and achieves XIC 0.831. Hard-Mono and EV-Mono reduce MVR to zero and XIC to a perfect 1.000. All methods place the 18 informative features in the top 18, with no null feature promoted into that set.

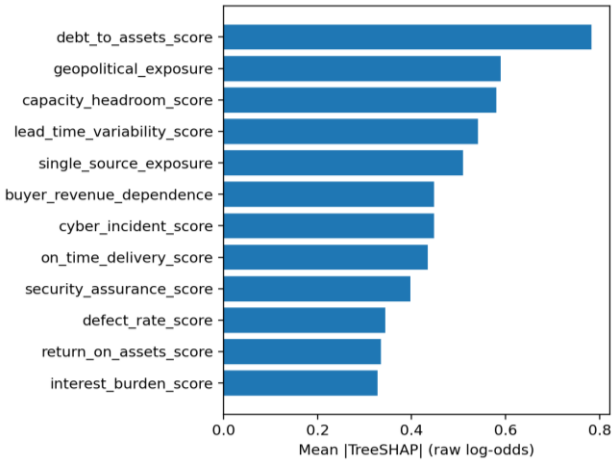

Figure 1. Top 12 global TreeSHAP drivers for EV-Mono on the clean test set.

The leading SHAP features include debt-to-assets, geopolitical exposure, capacity headroom, lead-time variability and single-source exposure. Because the data-generating process includes nonlinear interactions, this ranking need not reproduce the order of the main-effect coefficients; it is an attribution ranking for the fitted model, not an estimator of structural coefficients. The more important ground-truth check is that every informative variable precedes every null variable.

5.3. Robustness to Wrong Domain Priors

Table IV. Prior-misspecification stress test

Corrupt.	Model	AUC	PR	ECR	MVR	Wrong
0%	Hard-Mono	0.947	0.807	1.000	0.000	0/18
0%	EV-Mono	0.947	0.807	1.000	0.000	0/18
10%	Hard-Mono	0.944	0.798	0.889	0.094	2/18
10%	EV-Mono	0.946	0.802	1.000	0.043	0/16
20%	Hard-Mono	0.943	0.795	0.778	0.190	4/18
20%	EV-Mono	0.946	0.802	1.000	0.063	0/14
30%	Hard-Mono	0.938	0.779	0.722	0.235	5/18
30%	EV-Mono	0.944	0.799	1.000	0.070	0/13
40%	Hard-Mono	0.932	0.753	0.611	0.343	7/18
40%	EV-Mono	0.946	0.801	1.000	0.091	0/11

Table IV is the central experiment. Hard enforcement degrades almost monotonically as the prior map is corrupted. At 40% corruption, seven incorrect directions remain embedded in the model: AUC falls from 0.9465 to 0.9323, PR-AUC from 0.8074 to 0.7530, ECR from 1.000 to 0.611, and MVR rises from 0 to 0.343. The problem is therefore not merely a small accuracy penalty; the explanations systematically communicate the wrong domain direction for a substantial subset of features.

EV-Mono removes every deliberately flipped constraint at every tested corruption level while retaining all unflipped constraints. At 40% corruption it enforces 11 correct signs and zero wrong signs, producing AUC 0.9457, PR-AUC 0.8010, ECR 1.000 and MVR 0.091. Relative to hard enforcement, the paired AUC advantage is 0.01340 with a 95% bootstrap interval of $[0.00781, 0.01904]$. The residual MVR is expected because seven rejected variables are left unconstrained rather than forcibly corrected. This is an intentional governance trade-off: uncertainty becomes model flexibility instead of confidently encoded misinformation.

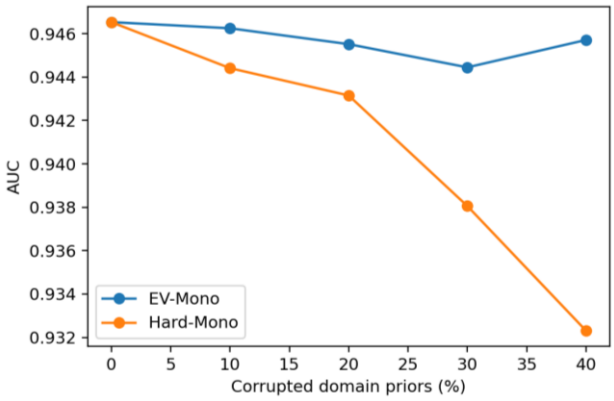

Figure 2. AUC under nested corruption of candidate monotonic priors. Paired seeds are used within each corruption level.

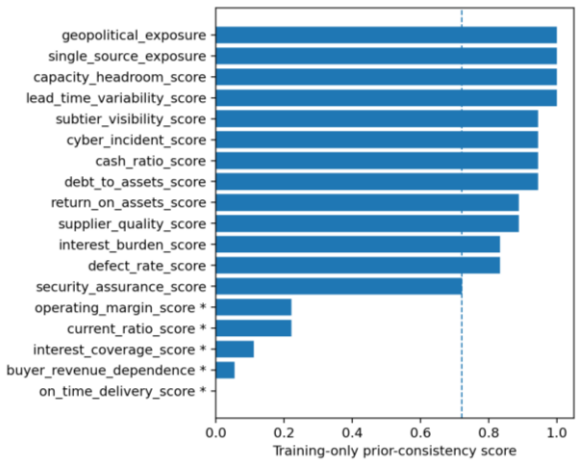

Figure 3. Training-only vetting scores at 30% corruption. An asterisk marks a deliberately flipped prior; the dashed line is tau = 0.72.

Figure 3 makes the diagnostic mechanism visible. At 30% corruption, the five flipped priors receive consistency scores from 0.000 to 0.222 and are vetoed, while the 13 correct priors score at least 0.722 and are retained. The gate therefore separates incorrect from correct signs in this controlled setting without consulting test labels and without inventing alternative signs. This clean separation should not be presumed in real procurement data; rather, it demonstrates the intended behavior when the training distribution contains sufficient directional evidence.

5.4. Sensitivity to the Vetting Threshold

Table V. Threshold sensitivity at 40% prior corruption

tau	Enf.	Wrong	AUC	PR	MVR
0.60	11	0	0.946	0.801	0.091
0.65	11	0	0.946	0.801	0.091
0.72	11	0	0.946	0.801	0.091
0.80	10	0	0.946	0.799	0.105
0.90	7	0	0.945	0.795	0.112

The separation is not an artifact of a single threshold. At 40% corruption, tau values from 0.60 through 0.72 retain the same 11 correct constraints and reject all seven wrong ones. Raising tau to 0.80 leaves 10 constraints and changes AUC only from 0.9457 to 0.9455; tau=0.90 becomes noticeably more conservative, retaining seven constraints with AUC 0.9445 and MVR 0.112. Thus the principal robustness result holds over a useful threshold interval, while excessively aggressive vetoing gradually gives up some monotonic coverage.

This sensitivity analysis also clarifies how tau should be governed in practice. A lower threshold risks admitting disputed directions when evidence is noisy, whereas a high threshold can remove valid protections and increase local violations because more features revert to the unconstrained learner. The appropriate value is therefore a policy parameter that should be documented and stress-tested rather than selected solely for test-set accuracy.

5.5. Refit Stability

Table VI. Descriptive tree-shap refit stability (b = 4)

Model	J@5	SD	J@10	SD
XGBoost	0.587	0.123	0.798	0.124
Hard-Mono	0.508	0.123	0.798	0.124
EV-Mono	0.508	0.123	0.798	0.124

The stability audit is intentionally treated as descriptive because only four paired bootstrap refits are used. Unconstrained XGBoost has Jaccard@5 0.587 versus 0.508 for both hard monotonic and EV-Mono models; all three have Jaccard@10 0.798. Since all clean candidate priors pass vetting, hard monotonic and EV-Mono are the same hypothesis class and consequently have the same paired stability result. This finding prevents an unwarranted claim that monotonicity automatically stabilizes feature rankings. Constraint correctness and explanation stability are separate properties.

6. Discussion

6.1. What the Results Change

The main implication is procedural: a monotonic prior should be treated as a governed model input, not a theorem. Hard constraints are most compelling exactly when the model will be audited, but that is also where an incorrect prior is most damaging because the error is systematic and difficult to discover from a conventional performance table. PMST turns that hidden assumption into a measurable robustness property. A model card can report not only which directions are enforced, but how strongly training-only evidence supports them and how performance changes when a plausible subset is withdrawn or challenged.

EV-Mono deliberately gives data less authority than a fully automated sign-learning scheme. The training distribution may be biased, temporally narrow or affected by historical procurement policy. Automatically flipping a disputed sign would therefore convert observational association into policy. Veto-only vetting is more conservative: it says “do not hard-code this direction yet.” A rejected prior can trigger expert review, segmentation by commodity type, or collection of additional evidence.

This design directly extends an important limitation identified in the audit-oriented supplier-risk framework of Bian et al., which explicitly notes that a misspecified sign prior would otherwise be enforced as confidently as a correct one and calls for documented, contestable elicitation [1]. The present results operationalize that concern as a measurable pre-deployment gate and a stress-test protocol. The relationship is complementary: monotonic constraints provide the auditability mechanism; empirical vetting provides robustness to imperfect elicitation.

6.2. Interpretation of XIC and SHAP

XIC is useful because a global SHAP ranking can look sensible while a procurement officer receives a locally confusing recommendation. By perturbing one feature in a risk-reducing direction and re-evaluating that feature’s attribution, XIC asks whether the displayed explanation behaves consistently with the model’s stated policy. It is still an explanation audit, not a causal test. Correlated supplier

variables can move together in reality, and changing one normalized feature while holding the rest fixed may be operationally infeasible. XIC should therefore be read as a model-consistency diagnostic that complements, rather than replaces, scenario analysis and causal domain reasoning.

The clean experiment also shows why several metrics are needed. ECR saturates at 1.000 even for unconstrained XGBoost because the dominant global relationships are learned correctly. MVR detects local directional violations, and XIC detects attribution-level inconsistencies. Conversely, the stability test shows that perfect monotonic compliance does not imply perfectly reproducible feature ranks. A procurement explanation can be directionally safe yet unstable in which of several correlated variables appears in the top five.

6.3. Limits to External Validity

The strongest limitation is the synthetic benchmark. The experiment establishes method behavior under a transparent generative process; it does not establish the prevalence of incorrect sign priors in any real supplier population. Feature names are procurement-motivated, but their distributions are normalized abstractions rather than measurements from identifiable firms. The controlled setting is necessary for the ground-truth stress test, but external validation on longitudinal supplier data is required before operational deployment.

Second, the vetting threshold tau=0.72 and B=3 bootstrap models are pragmatic compute choices, not universal constants. If features are sparse, heavily confounded or weakly identified, correct priors may receive low agreement scores and be left unconstrained. Third, the shift experiment changes latent block means while preserving the data-generating mechanism; structural concept drift could be more severe. Fourth, the models are evaluated at the supplier-record level and do not represent network propagation, sub-tier dependencies or contract-specific loss. Fifth, probability calibration is not optimized; random forest achieves better Brier and ECE on the clean set, suggesting that a deployed probability-based decision layer should calibrate the final model on a separate fold.

Finally, the bootstrap stability study is small and should not support significance claims. It is included because instability is a separate audit dimension and because the paired design reveals that our constraint-vetting contribution should not be conflated with ranking stability. Future work should use larger bootstrap budgets, expert panels that record disagreement, real pre-award supplier outcomes, temporal splits and commodity-specific priors.

7. Conclusion

This paper addresses a failure mode that becomes more important as domain-constrained machine learning moves into procurement governance: domain knowledge can be wrong. Hard monotonic XGBoost makes explanations directionally auditable when priors are correct, but PMST shows that the same mechanism can encode systematic error when priors are corrupted. EV-Mono preserves the governance value of expert-supplied signs while allowing training evidence only to veto unsupported constraints.

On the controlled 12,000-record benchmark, EV-Mono is competitive rather than dominant on clean prediction, but it eliminates monotonicity violations and achieves perfect XIC. More importantly, at 40% prior corruption it rejects all seven wrong signs and retains AUC 0.946 and ECR 1.000, whereas hard monotonic enforcement falls to AUC 0.932 and ECR 0.611. The practical recommendation is simple: document sign priors, vet them on training data, stress-test misspecification, and report explanation audits separately from predictive performance. In high-stakes supplier screening, trustworthy constraints should be contestable by design.

Data and Code Availability

The complete Python code, generated benchmark datasets, model-output tables, figures, metadata, bootstrap AUC results and a reference-verification log accompany this manuscript. The data are synthetic and generated by the released code with seed 42; no row represents a real supplier or company.

References

[1] Zhang, S., & Qiu, L. (2024). Data-driven assessment of pavement performance and resilience under extreme weather using the Long-Term Pavement Performance Program. AI and Data Science Journal, 1(1), 57–66.

[2] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. Symmetry, 17(4), 610.

[3] Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear complexity state space models for cryptocurrency volatility prediction. IEEE Open Journal of the Computer Society, 7, 226–243.

[4] Wang, J., Liu, J., Zheng, W., & Ge, Y. (2025). Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions. IEEE Access.

[5] Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access, 12, 116546–116557.

[6] Zhang, H. (2025). Reinforcement learning approaches for layout optimization in electronic design automation with electromagnetic compatibility constraints. Frontiers in Robotics and Automation, 2(2), 77–93.

[7] Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. Symmetry, 17(3), 341.

[8] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

[9] Chen, J., Wang, M., & Sun, T. (2025). Intelligent tax systems and the role of natural language processing in regulatory interpretation. American Journal of Machine Learning, 6(4), 74–94.

[10] Chen, Z., Liu, J., & Chen, J. (2025). Machine learning methods for financial forecasting in enterprise planning: Transitioning from rule-based models to predictive analytics. Frontiers in Artificial Intelligence Research, 2(3), 541–564.

[11] Chen, J., Liu, J., Liang, Y., & Zhou, M. (2026). KE-MLLM: A knowledge-enhanced multi-sensor learning framework for explainable fake review detection. Applied Sciences, 16(6), 2909.

[12] Yue, X., Yang, J., & Liu, W. (2026). FraudDebate-Agent: A multi-agent LLM framework with an evidence-based debate mechanism for financial statement fraud detection. Mathematics, 14(15), 2695.

[13] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[14] Zi, B. (2024). Cloud-native distributed systems for real-time payment intelligence. AI and Data Science Journal, 1(1), 51–56.

[15] Zhang, S., & Qiu, L. (2023). An interpretable, uncertainty-aware framework for bridge deterioration prediction and risk-based maintenance prioritization. Academic Journal of Architecture and Civil Engineering, 1(2), 21–28.

[16] Jin, J., Xing, S., Ji, E., & Liu, W. (2025). Xgate: Explainable reinforcement learning for transparent and trustworthy API traffic management in IoT sensor networks. Sensors, 25(7), 2183.

[17] Xing, S., Wang, Y., & Liu, W. (2025). Multi-dimensional anomaly detection and fault localization in microservice architectures: A dual-channel deep learning approach with causal inference for intelligent sensing. Sensors, 25(11), 3396.

[18] Xing, S., & Wang, Y. (2025). Proactive data placement in heterogeneous storage systems via predictive multi-objective reinforcement learning. IEEE Access, 13, 117986–117998.

[19] Ji, E., Wang, Y., Xing, S., & Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. IEEE Access.

[20] Xing, S., & Wang, Y. (2025). Cross-modal attention networks for multi-modal anomaly detection in system software. IEEE Open Journal of the Computer Society.

[21] Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. Sensors, 24(24), 7949.

[22] Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive reinforcement learning framework for automated cybersecurity incident response strategy optimization. Applied Sciences, 15(2), 951.

[23] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974–195985.

[24] Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. PLOS ONE, 20(10), e0332640.

[25] Wang, M., Zhang, X., Yang, Y., & Wang, J. (2025). Explainable machine learning in risk management: Balancing accuracy and interpretability. Journal of Financial Risk Management, 14(3), 185–198.

[26] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501–520.

[27] Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1–13.

[28] Bian, M., Li, P., Lin, Y., Wang, Y., & Teng, D. (2026). Assessing green public supply chains resilience under carbon neutrality goals: A multi-agent reinforcement learning framework. IEEE Access.

[29] Qin, Y., Rhee, M., Teng, D., Zhang, C., & Zi, B. (2026). Neuro-symbolic constraint verification for LLM-driven internet finance transaction execution. Scientific Reports.

[30] Rhee, M., Zou, J., Mo, T., Teng, D., & Yang, J. S. (2026). Enhancing web search agents with self-play contrastive fine-tuning. IEEE Access.

[31] Wang, B., Zhao, W., & Wang, Z. (2024). DAS-GNN: A scalable disagreement-aware graph neural framework for anomaly detection in large-scale networks. AI and Data Science Journal, 1(1), 67–75.

[32] Lin, Y., Wang, Y., & Bian, M. (2024). XOR-EW: An explainable and cost-aware artificial intelligence framework for operational risk early warning in financial institutions. AI and Data Science Journal, 1(1), 94–102.

[33] Zhang, C., & Wang, Y. (2024). A data-driven multi-objective production scheduling framework for battery pack manufacturing under material arrival uncertainty. AI and Data Science Journal, 1(1), 86–93.

[34] Rhee, M., Zou, J., & Qin, Y. (2024). CLARITY: A closed-loop framework for duplicate-aware failure prediction and budgeted automated remediation in cloud infrastructure. AI and Data Science Journal, 1(1), 76–85.

[35] Wang, Y., & Zhang, C. (2023). AERIS: Adaptive executor-scoped resource management for privacy-preserving data processing in Apache Spark. Eurasia Journal of Science and Technology, 5(2), 25–33.

[36] Qin, Y., Zou, J., & Rhee, M. (2022). An AI-driven intelligent transaction execution framework for internet finance platforms: Cost-sensitive real-time integrity interdiction. Eurasia Journal of Science and Technology, 4(2), 36–44.

[37] Zou, J., Qin, Y., & Rhee, M. (2024). Beyond model scale: A task-aware reliability benchmark for large language models in clinical decision support. Innovation and Technology Studies, 1(1), 39–47.

[38] Bian, M., Wang, Y., & Lin, Y. (2024). Explainable supplier risk prediction for critical telecommunications infrastructure procurement: An empirical XGBoost–SHAP framework. Social Science and Management, 1(1), 64–77.

[39] Jiao, Y., & Liang, Y. (2024). An integrated process-mining and explainable machine-learning framework for bottleneck and internal-control risk detection in the multi-entity financial close. Social Science and Management, 1(3), 96–107.