Modern Supply Chain Research and Applications

Interpretable and leakage-aware prediction of shipment delays in global supply chains: a Random Forest–SHAP workflow with multi-model benchmarking on the DataCo dataset

Received 11 May 2026
Revised 16 July 2026
Accepted 20 August 2026

Divyansh Saini, Harman Singh and Bhavya Agarwal
Delhi Technological University, Delhi, India

Abstract

Purpose – This study develops a leakage-aware evaluation workflow for predicting shipment dispatch delays at the time of order, auditing when each predictor becomes available and quantifying how much apparent accuracy on the DataCo dataset derives from post-outcome information.

Design/methodology/approach – On 172,765 completed DataCo shipments, a point-in-time audit identifies 21 order-time features and excludes post-outcome fields, including Late_delivery_risk. Five identically tuned models are compared over 15 order-grouped resamples with Nadeau–Bengio corrected *t*-tests (Holm-adjusted), cluster-bootstrap intervals, a chronological split, and exact SHapley Additive exPlanations (SHAP).

Findings – The leakage-free Random Forest attains $R^2 = 0.2719$ (95% confidence interval 0.2569–0.2864). Restoring post-outcome fields raises R^2 to 0.7589: a leakage inflation of $\Delta R^2 = +0.4870$ (same model, with versus without post-outcome inputs). A distinct diagnostic, the leaky model's margin over a flag-threshold rule ($\Delta R^2 = +0.0951$), is meaningful only conditional on the leak. All five honest models lie within 0.0084 R^2; SHAP indicates that order-time delay risk is structural, driven by mode, lane, and category base rates.

Research limitations/implications – Findings derive from one public dataset whose target is dispatch (not delivery) delay; generalisability claims are correspondingly modest.

Practical implications – Delay largely reflects infeasible service-level promises; the managerial levers are promise recalibration and base-rate triage, with measured detection rates (precision 0.864, recall 0.553) separated from assumed costs.

Originality/value – A feature-level point-in-time audit, a quantification of leakage inflation on a widely used benchmark, and a replicable leakage-audit protocol including a target-permutation canary.

Keywords Shipment delay prediction, Data leakage, Point-in-time prediction, Random forest regression, SHAP interpretability, Logistics analytics

Paper type Research article

1. Introduction

Supply chain resilience has emerged as a defining operational priority as disruption frequency accelerates across global logistics networks (Guo *et al.*, 2026; Li *et al.*, 2025). Shipment delays of even one to three days cascade into inventory imbalances, elevated transportation costs, and the erosion of customer trust (Camur *et al.*, 2024; Zhao *et al.*, 2023). For high-volume operations processing hundreds of thousands of orders annually, predicting which shipments are at risk of delay before they enter the fulfilment pipeline has direct and quantifiable value (Toorajipour *et al.*, 2021; Gupta *et al.*, 2025).

Modern Supply Chain Research and Applications
Emerald Publishing Limited
2631-3871
DOI 10.1108/MSCRA-04-2026-0035

Traditional forecasting models – including autoregressive integrated moving average (ARIMA), exponential smoothing, and linear regression – impose assumptions of linearity and stationarity that are systematically violated in real logistics data (Wang *et al.*, 2025; Heydarbakian and Sepehri, 2022). Shipment delays emerge from the interaction of shipping mode, order volume, financial indicators, and operational risk flags through nonlinear mechanisms that single-coefficient models cannot represent (Hu *et al.*, 2026; Balan *et al.*, 2025). Machine learning (ML) methods, particularly ensemble tree-based algorithms, have demonstrated consistent superiority over linear baselines in supply chain prediction tasks (Cao *et al.*, 2026; Pasupuleti *et al.*, 2024; Riad *et al.*, 2024; Douaioui *et al.*, 2024). However, prior applications are limited by small or simulated datasets, two- or three-model benchmarks without statistical testing, and insufficient interpretability analysis (Gomaa, 2025; Kang and Bhawna, 2025). Such predictive value exists, moreover, only if predictions are built from information actually available at the moment of prediction – a condition that a growing body of evidence shows is frequently violated in applied machine learning through data leakage (Kapoor and Narayanan, 2023).

This study addresses these gaps through four contributions. First, a feature-level point-in-time audit (Table 1) establishes, for every candidate predictor in the DataCo Global Supply Chain dataset (Constante *et al.*, 2019), whether it is available at order confirmation, and demonstrates that the dataset's dominant composite indicator, Late_delivery_risk, is a binarisation of the outcome itself (Table 2). Second, a leakage-inflation ablation quantifies what that leak is worth: restoring post-outcome fields inflates test R^2 from 0.2719 to 0.7589 ($\Delta R^2 = +0.4870$), implying that headline accuracies previously reported on this benchmark – including the originally submitted version of this study – measure label recovery rather than

Table 1. Feature-level point-in-time availability audit (21 order-time features retained; 5 post-outcome fields excluded)

Feature	Available_at_order	Rationale
Shipping Mode	TRUE	Chosen by customer at checkout
Scheduled_Lead	TRUE	Service-level promise fixed by mode at order
Order Region	TRUE	Destination known at order
Market	TRUE	Destination market known at order
Customer Segment	TRUE	Customer attribute, static
Category Name	TRUE	Product attribute, known at order
Department Name	TRUE	Product attribute, known at order
Type	TRUE	Payment/transaction type at checkout
Order_Month	TRUE	Derived from order date
Order_Year	TRUE	Derived from order date
Order_DayOfWeek	TRUE	Derived from order date
Feature	Available_at_order	Rationale
Shipping Mode	TRUE	Chosen by customer at checkout
Scheduled_Lead	TRUE	Service-level promise fixed by mode at order
Order Region	TRUE	Destination known at order
Market	TRUE	Destination market known at order
Customer Segment	TRUE	Customer attribute, static
Category Name	TRUE	Product attribute, known at order
Department Name	TRUE	Product attribute, known at order
Type	TRUE	Payment/transaction type at checkout
Order_Month	TRUE	Derived from order date
Order_Year	TRUE	Derived from order date
Order_DayOfWeek	TRUE	Derived from order date
Feature	Available_at_order	Rationale
Shipping Mode	TRUE	Chosen by customer at checkout
Scheduled_Lead	TRUE	Service-level promise fixed by mode at order

Table 2. Late_delivery_risk versus realised outcome, analysis population ($N = 172{,}765$)

	Delay_Days ≤0	Delay_Days >0
Late_delivery_risk = 0	73,788	0
Late_delivery_risk = 1	0	98,977

Note(s): Agreement 1.000000: on completed shipments, the flag is a binarisation of the outcome

prediction. Third, an honest five-model benchmark under identically budgeted tuning, order-grouped resampling, and corrected tests (Nadeau and Bengio, 2003; Demšar, 2006) shows all models tying within 0.0084 R^2 near a structural ceiling; the Random Forest is retained for its interpretability infrastructure rather than for accuracy superiority (Baryannis *et al.*, 2019). Fourth, a SHAP analysis of the deployable model shows order-time delay risk to be structural – a function of shipping-mode promises and lane and category base rates – so the effective managerial levers are promise recalibration and base-rate triage.

We characterise the contribution as an evaluation workflow rather than a framework: no new algorithmic architecture is proposed; the methodological value lies in the audit, ablation, and testing protocol wrapped around standard learners.

2. Literature review

2.1 Limitations of traditional forecasting in supply chain contexts

Classical forecasting paradigms remain prevalent due to their parsimony but assume linearity and stationarity that break systematically under disruption (Wang *et al.*, 2025; Guo *et al.*, 2026). When a risk indicator activates, its marginal effect on delay interacts with shipping mode, order value, and operational status in ways that linear coefficients cannot represent (Alaoua and Karim, 2025; Cao *et al.*, 2026). This motivates the adoption of flexible, nonparametric ML algorithms capable of representing interaction structures directly from data (Baryannis *et al.*, 2019).

2.2 Machine learning in supply chain analytics

The adoption of ML in supply chain management reflects a transition from deterministic planning toward data-driven, adaptive decision-making (Zhao *et al.*, 2023; Riad *et al.*, 2024). Balan *et al.* (2025) demonstrated the superiority of ML for post-crisis recovery prediction under nonlinear disruption propagation. Gupta *et al.* (2025) showed that causal ML outperformed regression in automotive supply chain resiliency assessment. At the macro level, artificial intelligence (AI) adoption and digital transformation are positively associated with resilience (Guo *et al.*, 2026; Li *et al.*, 2025).

2.3 Random forest in logistics and supply chain management

Random Forest has established itself as one of the most reliable algorithms in supply chain ML applications (Toorajipour *et al.*, 2021). Its bootstrap aggregation with random feature subset selection reduces variance without the instability of single deep trees (Camur *et al.*, 2024). Three properties make it particularly suitable for logistics datasets: robustness to correlated features, natural handling of mixed inputs without normalisation (Zhao *et al.*, 2023), and node-impurity-based feature importance. RF has achieved leading accuracy in medical device recall prediction (Hu *et al.*, 2026), supplier classification (Kang and Bhawna, 2025), and logistics optimisation (Pasupuleti *et al.*, 2024). Pasupuleti *et al.* (2024) confirmed RF's dual utility for prediction and feature selection in logistics optimisation.

2.4 Explainable AI for supply chain decision support

Models unable to explain their predictions are systematically under-trusted by logistics managers in high-stakes operational contexts (Gomaa, 2025; Kang and Bhawna, 2025). Shapley Additive Explanations (SHAP) has emerged as the leading post-hoc explanation method due to its cooperative game theory grounding, model-agnosticism, and consistency properties (Lundberg and Lee, 2017; Lundberg *et al.*, 2020). In supply chains, SHAP has been applied to escalation prediction (Gupta *et al.*, 2025) and lead-time uncertainty (Alaoua and Karim, 2025). Despite these advances, SHAP in supply chain delay prediction has typically been limited to global summary plots, rarely including dependence analysis or verification of the additivity property, and SHAP values are attributions of model behaviour rather than causal effects (Lundberg *et al.*, 2020). The present study addresses interpretability on this basis while restricting all attribution claims to the deployable, leakage-free model.

2.5 Data leakage and point-in-time prediction

Data leakage – training or evaluating a model on information unavailable at prediction time – is a documented, reproducibility-threatening error in applied machine learning: Kapoor and Narayanan (2023) identified leakage across 294 papers in 17 fields, with corrections eliminating up to half of reported performance. Point-in-time discipline – showing each feature to be available at the prediction moment – is standard in credit scoring but rare in published supply chain ML. Prior DataCo-based delay studies (e.g. Al-Saghir, 2022; Palaniappan, 2025) report high accuracies with the dataset's composite indicators among the predictors, without an availability audit. Independent recent work corroborates the concern examined here: Xue *et al.* (2026) exclude DataCo's late_delivery_risk field as co-derivation leakage, reporting a feature–label correlation exceeding 0.99, and publish a feature-level audit of their own. Associated safeguards include cross-fitted target encoding (Micci-Barreca, 2001), separation of model selection from performance estimation (Cawley and Talbot, 2010), and valid resampled comparisons (Nadeau and Bengio, 2003; Demšar, 2006; Holm, 1979). This study operationalises these practices for shipment delay prediction.

2.6 Research gap and theoretical framework

Existing benchmarks compare too few models to establish RF's relative positioning (Toorajipour *et al.*, 2021), and statistical significance testing remains uncommon in published supply chain ML benchmarks (Camur *et al.*, 2024). We are not aware of a prior study that quantifies the contribution of DataCo's composite risk indicator under a point-in-time constraint, or that examines the temporal stability of leakage-free delay prediction on this dataset.

This study is grounded in Information Processing Theory (IPT), originally articulated by Galbraith (1974) and subsequently applied to supply chain contexts by Zhao *et al.* (2023) and Li *et al.* (2025). IPT holds that organisations face uncertainty arising from the gap between the information required to complete a task and the information actually available. Supply chain operations are information-intensive: each order generates simultaneous signals across shipping mode, financial indicators, and operational risk status whose combined effect on delivery timing cannot be resolved by processing each signal independently. Ensemble tree-based methods function as high-capacity information processors, reducing uncertainty by representing nonlinear interaction structures directly from data. This IPT framing generates three testable propositions: (P1) shipment delay prediction is fundamentally nonlinear, such that ensemble methods will outperform linear models by a statistically significant margin; (P2) ensemble aggregation provides meaningful variance reduction over single-tree prediction; and (P3) predictive accuracy and model transparency are complementary, not competing, in supply chain delay prediction. These propositions are examined as open empirical questions under leakage-free evaluation and revisited in Section 5.1.

3. Methodology

3.1 Research design and prediction time

The prediction time is defined as order confirmation, before any fulfilment activity begins; a feature enters the primary model only if its value is fixed and knowable at that moment (Table 1). The study employs supervised regression on a continuous delay target under a group-aware protocol: order items of the same order share dispatch date and shipping mode – and therefore the target value – so row-level random splits would place identical targets in both partitions. All partitioning is therefore grouped by Order ID: an 80/20 GroupShuffleSplit hold-out (train 138,493 rows/50,317 orders; test 34,272 rows/12,580 orders; random_state = 42), GroupKFold (3) tuning, 15 GroupShuffleSplit benchmark resamples, and a cluster bootstrap by order. A secondary chronological split trains on orders before January 2017 (n = 119,880) and tests on later orders (n = 52,885). The learning algorithms are standard; the methodological content lies in the audit, ablation, and testing protocol built around them.

3.2 Dataset, target variable, and analysis population

The DataCo Global Supply Chain dataset (Constante *et al.*, 2019; doi: 10.17632/8gx2fvg2k6.5), released via Mendeley Data, comprises 180,519 order-item records and 53 original features spanning order and shipping dates, product categories, customer geographies, financial metrics, and shipping modes across five global markets and four shipping modes. Two semantic properties, documented in the dataset's official data dictionary, are stated explicitly. First, the dataset contains no delivery date; shipping date is the dispatch timestamp, so the target is dispatch delay – lateness of shipment release against the scheduled fulfilment window – not last-mile delivery delay. Second, the dataset supplies its own schedule column, Days for shipment (scheduled): 4, 2, 1 and 0 days for Standard Class, Second Class, First Class and Same Day, respectively. The originally submitted version instead hand-coded lead times of 5, 3, 2 and 0 days, which understated the late rate (16.3% reported against an actual raw rate near 54.8%). The revision derives the target entirely from the dataset's own columns: Delay_Days = Days for shipping (real) − Days for shipment (scheduled).

On this definition, Late_delivery_risk equals (Delay_Days >0) for every completed shipment: the crosstab over the analysis population has zero off-diagonal entries (agreement 1.000000; Table 2). The flag's generation logic is thus settled empirically – it is a binarisation of the realised outcome and cannot be available at order time. Exactly 4,423 records disagreed before exclusion; all carry Delivery Status "Shipping canceled" (SUSPECTED_FRAUD, 2,316; CANCELED, 2,107) and have no realised outcome. All 7,754 canceled-shipment records are excluded, yielding N = 172,765 completed shipments (supplementary Table S2). The submitted version's own SHAP waterfall analysis had identified canceled orders as ambiguous ground truth; the exclusion addresses that observation by design.

Delay_Days ranges over [−2, +4] days (mean + 0.565, standard deviation (SD) 1.493): 57.29% of completed shipments are late, 18.64% on time, 24.07% early; the modal delay is +1 day (33.55%), and 64.18% fall within ±1 day (Figure 1a). The central descriptive finding is structural (Figure 1b): First Class is late 100.0% of the time by exactly one day (SD 0.00) – a structurally infeasible promise; Second Class averages +1.99 days (79.8% late); Same Day is bimodal over {0, +1} (47.9% late); Standard Class is the only mode centred on its promise (mean −0.01, 39.8% late). Market-level means span only 0.554–0.575 days (supplementary Figure S1). Lateness in this dataset is chiefly a property of the promise rather than execution variability – a fact that shapes the achievable accuracy ceiling in Section 4.

3.3 Features and point-in-time availability audit

Table 1 lists every candidate predictor with its availability status at order confirmation and the rationale. Twenty-one order-time features are retained and encoded to 33 model inputs: four one-hot categoricals (Shipping Mode, Market, Customer Segment, Type); six cross-fitted,

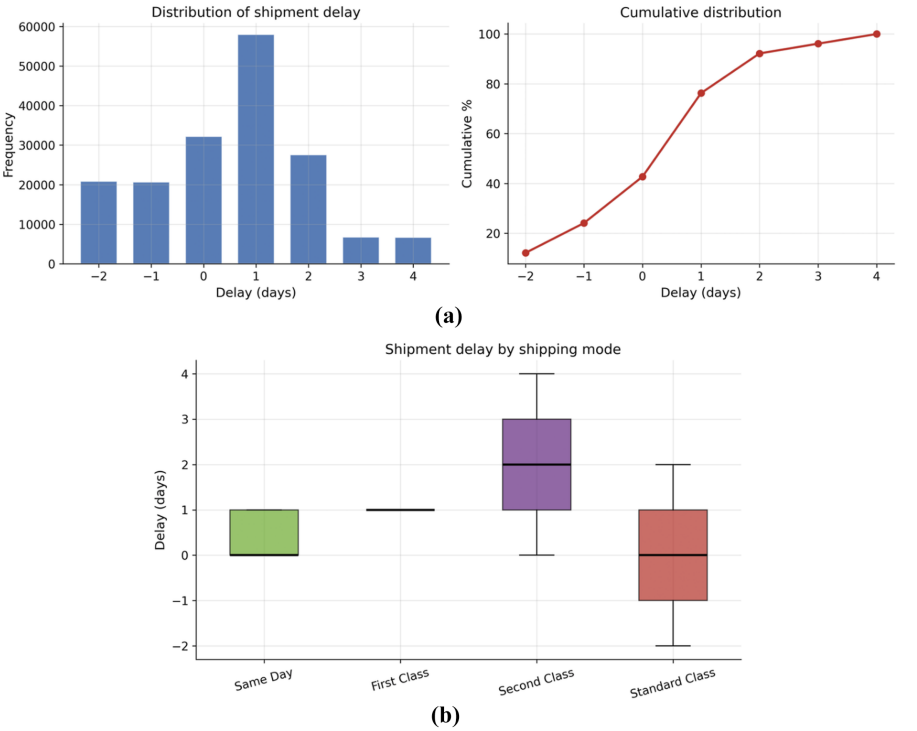


Figure 1. Delay distribution and structure (N = 172,765 completed shipments). (a) Distribution of Delay_Days: modal value + 1 day (33.55%); 57.3% of shipments late, 18.6% on time, 24.1% early; the range is bounded [−2, +4] by the schedule structure. (b) Delay by shipping mode: First Class is exactly +1 day for every shipment (SD 0; 100% late) – deterministic failure of an infeasible promise, not reliability; Second Class median +2 (79.8% late); Same Day bimodal over {0, +1}; Standard Class is the only mode centred on its promise. Lateness is chiefly a property of the promise rather than execution variability

continuous target-encoded variables (Order Region, Order Country, Category Name, Department Name, and the interaction base rates Region_Mode and Category_Mode), fitted out-of-fold on training data only to prevent encoding leakage (Micci-Barreca, 2001); and eleven numerics (Scheduled_Lead, three calendar fields, and seven order-economics fields). Five fields are excluded as post-outcome: Late_delivery_risk, Delivery Status, Order Status, Days for shipping (real), and shipping date. Order Status is excluded because its terminal values (e.g. COMPLETE, CANCELED) are assigned after fulfilment.

3.4 Models and hyperparameter optimisation under identical budgets

Each Random Forest tree T_b is built on a bootstrap sample with $\sqrt{p}$ features considered at each split, and the ensemble prediction is the average $\hat{f}(x) = (1/B) \Sigma T_b(x)$. To ensure benchmarking fairness, every tuned model receives an identical optimisation procedure: GridSearchCV with GroupKFold(3) by Order ID on the training partition, scored by negative mean absolute error (MAE), with exactly eight hyperparameter candidates per model (Decision Tree, Random Forest, Gradient Boosting, Extreme Gradient Boosting (XGBoost); Linear Regression is closed-form and has no tuned hyperparameters). Selected configurations and full search grids are reported in supplementary Table S4. This replaces the submitted design, in which only the Random Forest was tuned.

3.5 Comparative benchmarking and statistical testing

Five models are benchmarked under an identical preprocessing pipeline: Linear Regression, Decision Tree, Random Forest (proposed), Gradient Boosting, and XGBoost, over 15 independent GroupShuffleSplit resamples (80/20 by order). Because resampled scores share overlapping training data, naive paired *t*-tests are anti-conservative; differences are therefore tested with the Nadeau and Bengio (2003) corrected resampled *t*-test (df = 14) with Holm adjustment across the four comparisons (Holm, 1979; Demšar, 2006). Hold-out confidence intervals use a cluster bootstrap by order (1,000 resamples). The residual quantile–quantile plot is retained only as a shape diagnostic (supplementary Figure S2); the residuals are non-normal (integer-mixture), and no test in this study relies on residual normality.

3.6 Leakage-inflation ablation

Three configurations are evaluated on the identical hold-out: (A) the honest primary model (order-time features only); (B) a leaky upper-bound model adding Late_delivery_risk and Order Status, reported purely as a diagnostic; and (C) a rule that thresholds on the flag alone. Two distinct quantities result, named to eliminate ambiguity: the Leakage-Inflation Delta (B − A) measures how much apparent accuracy post-outcome information contributes; the ML-over-Rule Delta (B − C) measures what the leaky model adds beyond the naive use of the flag and is meaningful only conditional on the leak. The submitted version's ablation compared two configurations that both contained the flag and therefore could not address the circularity concern; the present design supersedes it.

3.7 Evaluation metrics, detection analysis, SHAP, and leakage audit

Performance is reported as R^2, MAE (days), and root mean squared error (RMSE) (days) with cluster-bootstrap 95% confidence intervals. For operational reading, continuous predictions are additionally thresholded at a pre-registered threshold of 0.5 days (full sweep in supplementary Table S5) and compared against rule baselines. SHAP uses the exact TreeExplainer – upgraded from the approximate variant in the submitted version – on 5,000 instances stratified by delay value, with additivity asserted programmatically (maximum deviation 1.8×10^{-14}); SHAP values are attributions of model behaviour, not causal effects (Lundberg *et al.*, 2020). The pipeline passes a three-part leakage audit: (1) a manifest scan confirming no post-outcome field among the 33 encoded inputs; (2) exact reproduction of reported test metrics from the saved model; and (3) a target-permutation canary in which retraining on shuffled labels collapses test R^2 to −0.0012, certifying the absence of any leakage pathway (Kapoor and Narayanan, 2023). All experiments use Python 3.11.4, scikit-learn 1.8.0, NumPY 2.3.5, pandas 3.0.2, XGBoost 3.2.0, SHAP 0.51.0; RANDOM_SEED = 42 throughout.

3.7.1 AI use declaration. In preparing this manuscript, Claude by Anthropic (model: claude-sonnet-4-6) and Perplexity AI were used for copy-editing purposes only – specifically, to improve the clarity, grammar, and readability of text originally drafted by the authors. No AI tool was used to create, draft, or generate any portion of the manuscript content, including the abstract, literature review, methodology, results, analysis, discussion, or conclusion. No AI tool was used to generate, manipulate, or analyse research data or results. No AI tool was used to produce figures, tables, or numerical outputs. All research design, data processing, model training, hyperparameter optimisation, statistical testing, coding, and interpretation of results were conducted entirely by the authors.

4. Results

4.1 Model performance

The leakage-free Random Forest achieves test R^2 = 0.2719 (95% CI 0.2569–0.2864), MAE = 0.9936 days (CI 0.9788–1.0093), and RMSE = 1.2765 days (CI 1.2642–1.2897)

(Table 3). A naive per-mode conditional-mean baseline achieves $R^2 = 0.2728$: the Random Forest does not surpass it on the hold-out ($\Delta R^2 = -0.0009$). This quantifies the structural ceiling implied by Section 3.2, not a deficiency of any particular learner. Figure 2a shows the calibration structure: predictions are monotone in actual delay from -2 to $+1$; actual $+2$ delays are predicted near 0 (predominantly Standard Class, conditional mean ≈ 0); actual $+3$/$+4$ delays are predicted near $+2$ (only short-promise modes can be that late; Second Class conditional mean + 1.99). The model recovers conditional means; within-mode variation is unpredictable at order time. The grouped learning curve (Figure 2b) corroborates this: training R^2 descends toward the validation plateau as training size grows – a noise-limited ceiling, not insufficient capacity or data. Under the chronological split, $R^2 = 0.2749$ and MAE = 0.9955 days (Table 3, final column) – a 1.1% deviation from the group-split estimate, consistent with temporal stability of the honest feature set over the dataset's horizon. No claim of generalisability beyond this dataset is made.

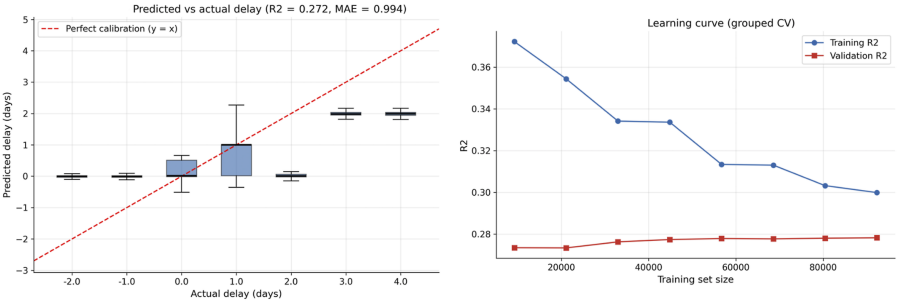
4.2 Comparative benchmarking

Table 4 presents the benchmark over 15 order-grouped resamples with the corrected significance tests. Mean R^2 spans 0.2723 (Decision Tree) to 0.2807 (Random Forest), with Linear Regression at 0.2804 and Gradient Boosting at 0.2802 – all five models within 0.0084 R^2. Under the corrected test with Holm adjustment, the Random Forest is statistically indistinguishable from Linear Regression ($t = 0.757$, $p = 0.4614$) and Gradient Boosting ($t = 2.507$, $p\_holm = 0.0502$), and significantly better than XGBoost ($t = 7.592$) and the

Table 3. Hold-out test metrics with cluster-bootstrap 95% confidence intervals (1,000 resamples, by order) and chronological-split validation (honest model)

Metric	Group split	95% CI lower	95% CI upper	Temporal split
R^2	0.2719	0.2569	0.2864	0.2749
MAE (days)	0.9936	0.9788	1.0093	0.9955
RMSE (days)	1.2765	1.2642	1.2897	1.2721

Note(s): Temporal split trains on orders before 1 January 2017 ($n = 119{,}880$) and tests on orders from January 2017 onward ($n = 52{,}885$); deviation from group split: 1.1% (R^2)



(a) **(b)**

Figure 2. Honest model behaviour. (a) Actual versus predicted Delay_Days, calibration boxplots ($R^2 = 0.272$, MAE = 0.994 days): predictions are monotone from -2 to $+1$; actual $+2$ is predicted near 0 (predominantly Standard Class, conditional mean ≈ 0) and actual $+3$/$+4$ near $+2$ (short-promise modes; Second Class conditional mean + 1.99) – conditional means are recovered, within-mode variation is unpredictable at order time. (b) Grouped learning curve: training R^2 descends toward the validation plateau (≈0.278) as training size grows, indicating a noise-limited ceiling rather than a capacity or data limit

Table 4. Five-model benchmark (15 GroupShuffleSplit resamples) with Nadeau–Bengio corrected resampled *t*-tests (df = 14) and Holm adjustment for the four comparisons against the Random Forest

Model	Mean R^2 (SD)	Mean MAE	Mean RMSE	t	p_holm	ΔR^2	Significant ($\alpha = 0.05$)
Random Forest (Proposed)	0.2807 (0.0071)	0.9844	1.2653	–	–	–	–
Linear Regression	0.2804 (0.0075)	0.9897	1.2655	0.757	0.4614	+0.0002	No
Gradient Boosting	0.2802 (0.0073)	0.9855	1.2657	2.507	0.0502	+0.0005	No
XGBoost	0.2786 (0.0071)	0.9904	1.2671	7.592	<0.0001	+0.0020	Yes
Decision Tree	0.2723 (0.0075)	0.9912	1.2726	7.839	<0.0001	+0.0084	Yes

Note(s): All differences are RF-favourable. Significance at these magnitudes does not imply practical importance

Decision Tree ($t = 7.839$; both p_holm <0.0001), with paired ΔR^2 of +0.0020 and + 0.0084. Significance at these magnitudes does not imply practical importance: all models tie near the naive-baseline ceiling, and the Random Forest – nominally best – is retained for its interpretability infrastructure at no accuracy cost (Baryannis *et al.*, 2019). Figure 3a visualises the compression (truncated axis, declared in the caption); overlapping marginal standard-deviation whiskers do not contradict the significant paired tests, which are computed on within-split differences.

4.3 Leakage-inflation ablation

Table 5 and Figure 3b present the decomposition. The honest model achieves $R^2 = 0.2719$; the flag-threshold rule, 0.6638; and the leaky upper-bound model, 0.7589. The Leakage-Inflation Delta is +0.4870–64% of the leaky model's apparent explanatory power derives from post-outcome information. The ML-over-Rule Delta of +0.0951 exists only conditional on the leak and is reported as a diagnostic, not as a deployable value. The R^2 of 0.7683 reported in the originally submitted version is statistically indistinguishable from the leaky upper bound reproduced here, identifying it retrospectively as a leaked result; by implication, comparable

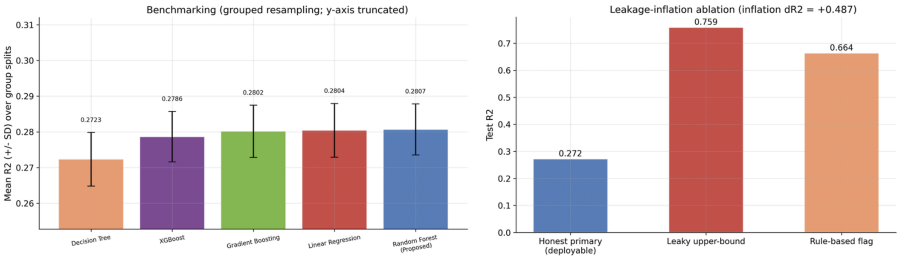


Figure 3. Model comparison and leakage decomposition. (a) Benchmark over 15 order-grouped resamples (vertical axis truncated for legibility): all five models lie within 0.0084 R^2; overlapping marginal SD whiskers do not contradict the significant paired tests, which are computed on within-split differences. (b) Leakage-inflation ablation: honest 0.272/flag rule 0.664/leaky 0.759; the Leakage-Inflation Delta (+0.487) quantifies the apparent accuracy contributed by post-outcome information – the originally submitted headline ($R^2 = 0.7683$) matches the leaky configuration

Table 5. Leakage-inflation ablation on the identical hold-out

Configuration	R^2	MAE (days)
Honest primary (order-time features only)	0.2719	0.9936
Rule: flag threshold only	0.6638	0.7323
Leaky upper bound (+Late_delivery_risk, + Order Status)	0.7589	0.5605
Leakage-Inflation Delta (leaky − honest)	+0.4870	–
ML-over-Rule Delta (leaky − rule; diagnostic only)	+0.0951	–

headline accuracies in prior DataCo delay studies warrant the same audit (Xue *et al.*, 2026; Kapoor and Narayanan, 2023).

4.4 SHAP interpretation of the deployable model

Figure 4a presents global SHAP attribution for the honest model, computed with the exact TreeExplainer and verified for additivity. Attribution concentrates on Region_Mode (mean |SHAP| = 0.2009), Category_Mode (0.1624), the shipping-mode indicators Second Class (0.1343) and Standard Class (0.1250), and Scheduled_Lead (0.1020); all remaining features contribute at most 0.02 (full ranking in supplementary Table S1). Because shipping mode and scheduled lead are collinear by construction, attribution credit is split across ranks 3–5. The beeswarm and dependence views (Figures 4b and 4c) corroborate this: high base-rate encodings shift predictions upward, and the Region_Mode dependence is near-monotone – the

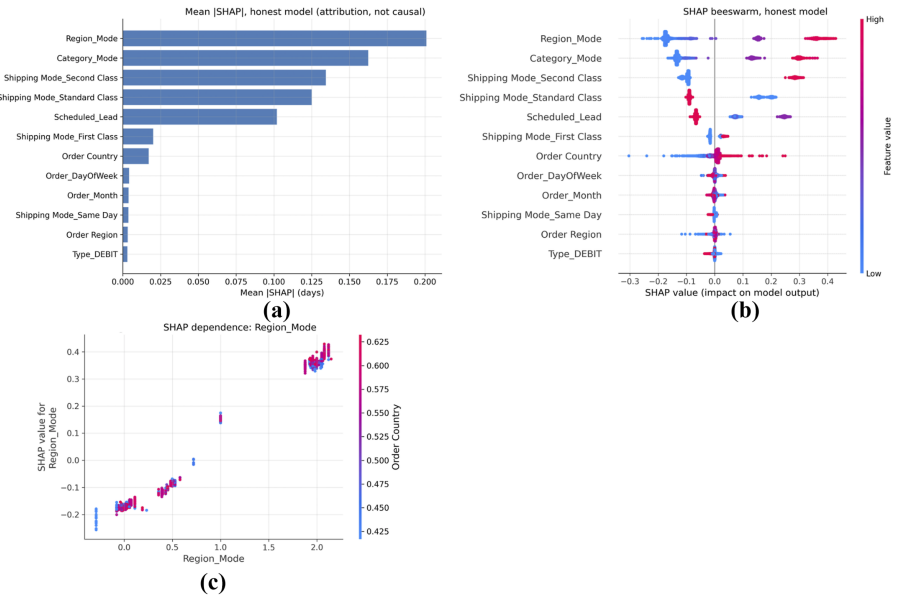


Figure 4. SHAP interpretation of the deployable (honest) model (exact TreeExplainer; $n = 5{,}000$ stratified test instances; additivity asserted, maximum deviation 1.8×10^{-14}; attributions of model behaviour, not causal effects). (a) Mean |SHAP|: Region_Mode (0.2009) and Category_Mode (0.1624) lead; shipping-mode indicators and Scheduled_Lead follow (collinearity splits credit across ranks 3–5); all remaining features ≤0.02. (b) Beeswarm: high lane/category base rate encodings push predictions upward; Standard Class membership pulls them downward, consistent with per-mode means. (c) Dependence on Region_Mode: near-monotone pass-through of the encoded lane base rate, supporting the base-rate-lookup interpretation

model passes the encoded base rate through almost linearly. The deployable model operates, in effect, as a calibrated base-rate lookup over lane-by-mode and category-by-mode cells – consistent with the structural finding of Section 3.2. These values are attributions of model behaviour, not causal effects, and all SHAP figures and in-text values derive from a single computation output.

4.5 Detection performance and illustrative economics

Thresholding the regression output at the pre-registered 0.5 days yields measured detection precision of 0.864, recall of 0.553, and F1 of 0.674 on the hold-out. A thresholded naive per-mode rule achieves $F1 = 0.667$: the two are practically tied, and no detection lift is claimed. The model's operational value lies instead in calibrated continuous scores, lane- and category-level granularity, and interpretability. Per 1,000 orders, the threshold produces 368 flags and 318 true detections. Crossing these measured quantities with assumed penalty costs ($200, $350, or $500 per late shipment) and intervention costs ($20 or $50 per flag) yields net avoidance from $45,244 to $151,799 per 1,000 orders (supplementary Table S3). These figures are explicitly illustrative: precision, recall, and flag counts are measured; all dollar parameters are assumptions requiring organisation-specific calibration. This analysis replaces the $2.9 M estimate in the originally submitted version, which relied on an assumed 50% detection rate not derivable from regression metrics.

5. Discussion

5.1 Theoretical implications: the IPT propositions under leakage-free evaluation

The honest results support the IPT-derived propositions only partially, and the pattern of partial support is itself informative. Proposition 1 (nonlinearity) is not supported: with post-outcome information removed, Linear Regression is statistically indistinguishable from all ensemble methods; the apparent nonlinear advantage in the submitted version was a property of the leaked feature's interaction structure, not of the order-time prediction problem. Proposition 2 receives marginal support: the Random Forest significantly outperforms the single Decision Tree, but by $\Delta R^2 = 0.0084$. Proposition 3 survives in a weak form: interpretability costs nothing here precisely because all models tie.

Read through IPT (Galbraith, 1974), this pattern locates the information gap. The residual uncertainty in dispatch delay is aleatoric at order confirmation – dominated by within-mode execution variation for which no order-time signal exists in the dataset – and no increase in processing capacity can close a gap for which the requisite information does not yet exist. IPT's prescription in that case is to change the task structure rather than the processor: here, the delivery promises themselves (Section 5.2). This reading renders the modest R^2 theoretically coherent: the model extracts essentially all the information the order-time signal set contains, and the remainder is structural.

5.2 Managerial implications

The findings support a structured exception-management process – pre-fulfilment risk screening, priority-queue routing, and system-embedded decision support – with the levers reordered by the structural analysis. First, promise recalibration dominates prediction: First Class fails its one-day promise deterministically and Second Class fails 79.8% of the time; no forecasting system can fix an infeasible promise; adjusting these scheduled windows would eliminate most recorded lateness at zero prediction cost. Second, for the remaining discretionary risk, calibrated base-rate triage applies: the model prices lane-by-mode and category-by-mode cells, which planners can operationalise as ranked exception queues at order confirmation, with the threshold sweep supporting precision–recall trade-offs. Third, the economics of Section 4.5 are illustrative scenarios, not measured outcomes.

For system integration, the workflow is most valuable as a decision support layer within a Transportation Management System (TMS) or Enterprise Resource Planning (ERP) platform. All primary model inputs are fixed at order creation or confirmation, so predictions can be computed before fulfillment begins; inference is fast, and SHAP explanations can be pre-computed daily to give planners a ranked exception list. The cluster bootstrap interval in Table 3 – R^2 constrained to [0.2569, 0.2864] – provides the honest performance floor operations teams would require to justify deployment, and carrier selection and service-level agreement buffer allocation can incorporate the calibrated delay scores alongside cost.

5.3 Methodological contributions

For supply chain prediction research, the study offers a transferable protocol: (1) a feature-level point-in-time audit table as a standard reporting object; (2) a leakage inflation ablation that quantifies, rather than asserts, the cost of post-outcome features; (3) group-aware resampling wherever transactional data carry repeated group structure; (4) corrected resampled *t*-tests with multiplicity control; and (5) a target permutation canary certifying that no leakage pathway survives. The interpretability reporting – exact SHAP with asserted additivity and single-source consistency across figures, tables, and text – extends Gomaa (2025) and Gupta *et al.* (2025). The DataCo dataset itself emerges as a cautionary benchmark: its most predictive fields are outcome-derived and its promise structure is partly deterministic, so results reported on it without an availability audit should be interpreted accordingly.

6. Limitations and future research directions

First, all findings derive from a single public dataset; external validation remains future work, and no claim of broad generalisability is made. Second, the target is dispatch delay: the dataset records no delivery date, so last-mile delay is unobserved. Third, the promise structure is partly deterministic (First Class), which strengthens the cautionary message but limits transfer of the specific structural findings to carriers with different service level designs. Fourth, while the 1.1% temporal deviation is modest, the DataCo dataset reflects pre-pandemic supply chain conditions; post-2020 disruption patterns may alter the feature relationships identified here, and validation on recent operational data is recommended before deployment. Fifth, order items within an order share outcomes; all splits are grouped by order, but target-encoded base rates aggregate over items, weighting multi-item orders more heavily. Sixth, hyperparameters were tuned on the same partition later used for comparative evaluation; selection and estimation are separated by grouped folds but not doubly nested (Cawley and Talbot, 2010). Seventh, the dollar parameters in the cost analysis are assumptions requiring organisation-specific values.

Future directions include: external validation on an independent logistics dataset (e.g. Olist); direct classification of lateness; customer-level generalisation splits; validation on disruption-period data; and exploration of hybrid long short-term memory–Random Forest architectures (Li *et al.*, 2025).

7. Conclusion

This study set out to predict shipment delays interpretably; the revision demonstrates that, on the field's most widely used public benchmark, the first-order question is what may legitimately be predicted from the available information. A feature-level point-in-time audit shows the dataset's dominant predictor to be a binarisation of the outcome; excluding post-outcome information reduces test R^2 from 0.7589 to 0.2719, a leakage inflation of $\Delta R^2 = +0.4870$ that retrospectively explains the originally submitted headline result and invites the same audit of comparable prior DataCo results.

Under leakage-free, order-grouped, statistically corrected evaluation, five models tie within 0.0084 R^2 of a naive per-mode baseline; the Random Forest is retained for

interpretability at no accuracy cost. SHAP attribution shows order-time delay risk to be structural – a property of service level promises and lane and category base rates – so the operative managerial levers are promise recalibration and calibrated triage. The transferable contribution is the audit protocol itself: the feature availability table, the leakage-inflation ablation, group-aware corrected testing, and the target permutation canary, offered as a replicable template for leakage-aware machine learning in supply chain operations.

Acknowledgments

The authors declare the use of AI tools for copy-editing purposes in the preparation of this manuscript. Claude by Anthropic (model: claude-sonnet-4-6) and Perplexity AI (web application, perplexity.ai) were used solely to improve the grammar, clarity, and readability of author-written text. These tools were not used to generate any manuscript content, research data, results, analysis, figures, or tables. All intellectual contributions, research design, coding, data analysis, and interpretation of findings are entirely the work of the authors. The authors take full responsibility for the accuracy and integrity of all content in this manuscript.

Supplementary material

The supplementary material for this article can be found online

References

Al-Saghir, R. (2022), "Predicting delays in the supply chain with the use of machine learning", MS thesis, Rochester Institute of Technology, Rochester, NY, available at: https://repository.rit.edu/ (accessed 12 July 2026).

Alaoua, A. and Karim, M. (2025), "A Bayesian learning approach for predictive resilience in engineer-to-order supply chains", *Supply Chain Analytics*, Vol. 13, 100190, doi: 10.1016/j.sca.2025.100190.

Balan, G.S., Kumar, V.S. and Raj, S.A. (2025), "Machine learning and artificial intelligence methods and applications for post-crisis supply chain resiliency and recovery", *Supply Chain Analytics*, Vol. 10, 100121, doi: 10.1016/j.sca.2025.100121.

Baryannis, G., Dani, S. and Antoniou, G. (2019), "Predicting supply chain risks using machine learning: the trade-off between performance and interpretability", *Future Generation Computer Systems*, Vol. 101, pp. 993-1004, doi: 10.1016/j.future.2019.07.059.

Camur, M.C., Ravi, S.K. and Saleh, S. (2024), "Enhancing supply chain resilience: a machine learning approach for predicting product availability dates under disruption", *Expert Systems with Applications*, Vol. 247, 123226, doi: 10.1016/j.eswa.2024.123226.

Cao, M., Li, Y. and Chen, L. (2026), "Artificial intelligence and corporate innovation: a perspective based on supply chain resilience", *International Review of Financial Analysis*, Vol. 109, 104720, doi: 10.1016/j.irfa.2025.104720.

Cawley, G.C. and Talbot, N.L.C. (2010), "On over-fitting in model selection and subsequent selection bias in performance evaluation", *Journal of Machine Learning Research*, Vol. 11, pp. 2079-2107.

Constante, F., Silva, F. and Pereira, A. (2019), "DataCo smart supply chain for big data analysis", *Mendeley Data*, Vol. V5, doi: 10.17632/8gx2fvg2k6.5 1 March 2024.

Demšar, J. (2006), "Statistical comparisons of classifiers over multiple data sets", *Journal of Machine Learning Research*, Vol. 7, pp. 1-30.

Douaioui, K., Oucheikh, R. and Mabrouki, C. (2024), "Enhancing supply chain resilience: rime-clustering and ensemble deep learning strategies for late delivery risk prediction", *Logforum*, Vol. 20 No. 1, pp. 55-70, doi: 10.17270/j.log.001007.

Galbraith, J.R. (1974), "Organization design: an information processing view", *Interfaces*, Vol. 4 No. 3, pp. 28-36, doi: 10.1287/inte.4.3.28.

Gomaa, A.H. (2025), "Enhancing supply chain management using machine learning techniques: a comprehensive review, gap analysis, and strategic framework", *Transnational Supply Chain Research*, Vol. 1 No. 1, pp. 1-21, doi: 10.65773/tscr.1.1.6.

Guo, J., Jia, F. and Chen, L. (2026), "How generative AI adoption affects supply chain resilience: an operations and supply chain management perspective", *Technological Forecasting and Social Change*, Vol. 224, 124446, doi: 10.1016/j.techfore.2025.124446.

Gupta, I., Martinez, A., Correa, S. and Wicaksono, H. (2025), "A comparative assessment of causal machine learning and traditional methods for enhancing supply chain resiliency and efficiency in the automotive industry", *Supply Chain Analytics*, Vol. 10, 100116, doi: 10.1016/j.sca.2025.100116.

Heydarbakian, S. and Sepehri, M. (2022), "Interpretable machine learning to improve supply chain resilience: an Industry 4.0 recipe", *IFAC-PapersOnLine*, Vol. 55 No. 10, pp. 2834-2839, doi: 10.1016/j.ifacol.2022.10.160.

Holm, S. (1979), "A simple sequentially rejective multiple test procedure", *Scandinavian Journal of Statistics*, Vol. 6 No. 2, pp. 65-70.

Hu, Y., Monticolo, D. and Ghadimi, P. (2026), "A machine learning-based medical device recall initiator prediction framework: from supply chain risk management and resilience view", *Expert Systems with Applications*, Vol. 298, 129922, doi: 10.1016/j.eswa.2025.129922.

Kang, P.S. and Bhawna (2025), "Enhancing supply chain resilience through supervised machine learning: supplier performance analysis and risk profiling for a multi-class classification problem", *Business Process Management Journal*, Vol. 31 No. 2, pp. 1-28, doi: 10.1108/bpmj-03-2024-0174.

Kapoor, S. and Narayanan, A. (2023), "Leakage and the reproducibility crisis in machine-learning-based science", *Patterns*, Vol. 4 No. 9, 100804, doi: 10.1016/j.patter.2023.100804.

Li, P., Chen, Y. and Guo, X. (2025), "Digital transformation and supply chain resilience", *International Review of Economics and Finance*, Vol. 99, 104033, doi: 10.1016/j.iref.2025.104033.

Lundberg, S.M. and Lee, S.-I. (2017), "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4765-4774.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2020), "From local explanations to global understanding with explainable AI for trees", *Nature Machine Intelligence*, Vol. 2 No. 1, pp. 56-67, doi: 10.1038/s42256-019-0138-9.

Micci-Barreca, D. (2001), "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems", *ACM SIGKDD Explorations Newsletter*, Vol. 3 No. 1, pp. 27-32, doi: 10.1145/507533.507538.

Nadeau, C. and Bengio, Y. (2003), "Inference for the generalization error", *Machine Learning*, Vol. 52 No. 3, pp. 239-281, doi: 10.1023/a:1024068626366.

Palaniappan, V.B. (2025), "The role of data, machine learning, and supply chain interdependencies in implementing connected packaging solutions for enhanced supply chain efficiency", DBA dissertation, Swiss School of Business and Management, Geneva.

Pasupuleti, V., Thuraka, B., Kodete, C.S. and Malisetty, S. (2024), "Enhancing supply chain agility and sustainability through machine learning: optimization techniques for logistics and inventory management", *Logistics*, Vol. 8 No. 3, p. 73, doi: 10.3390/logistics8030073.

Riad, M., Naimi, M. and Okar, C. (2024), "Enhancing supply chain resilience through artificial intelligence: developing a comprehensive conceptual framework for AI implementation and supply chain optimization", *Logistics*, Vol. 8 No. 4, p. 111, doi: 10.3390/logistics8040111.

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P. and Fischl, M. (2021), "Artificial intelligence in supply chain management: a systematic literature review", *Journal of Business Research*, Vol. 122, pp. 502-517, doi: 10.1016/j.jbusres.2020.09.009.

Wang, H., Tang, H., Leng, N. and Yu, Z. (2025), "A machine learning-based study on the synergistic optimization of supply chain management and financial supply chains from an economic perspective", *Journal of Economic Theory and Business Management*, Vol. 2 No. 4, pp. 1-18.

Xue, Z., Huo, M. and Wang, Y. (2026), "EAGLE: edge-aware graph learning for proactive delivery delay prediction in smart logistics networks", arXiv:2604.05254, available at: https://arxiv.org/abs/2604.05254

Zhao, N., Hong, J. and Lau, K.H. (2023), "Impact of supply chain digitalization on supply chain resilience and performance: a multi-mediation model", *International Journal of Production Economics*, Vol. 259, 108817, doi: 10.1016/j.ijpe.2023.108817.

Corresponding author
Divyansh Saini can be contacted at: d4ysaini@gmail.com