CEPAT Journal of Computer Engineering: Progress, Application and Technology
Vol. 2, No. 2, November 2026, pp. 84~93
ISSN: 2963-6728, DOI: https://doi.org/10.25124/cepat.v5i02.12096

Explainable Delivery Delay Prediction in E-Commerce Logistics

Veljko Spasojević[1], Jasmina Dončić[2], Milica Tufegdžić[1], Vladan Čolić[1], Marija Mojsilović[1]
[1]Department of Trstenik, Academy of Professional Studies Šumadija, Serbia
[2]Department of Kragujevac, Academy of Professional Studies Šumadija, Serbia

Article Info

Article history:

Received August 28, 2026
Revised October 17, 2026
Accepted October 19, 2026

Keywords:
SHAP
XGBoost
Counterfactual explanation
Cost-sensitive classification
Delivery risk
Model robustness
Interpretable machine learning

ABSTRACT

Delivery-delay prediction is most useful when operational users can understand both the risk signal and model failure conditions. Using the public DataCo Smart Supply Chain dataset, we reconstructed 180,519 line-item records into 65,752 unique orders, excluded post-shipment variables, and evaluated a fixed-configuration Extreme Gradient Boosting (XGBoost) classifier on 30 encoded pre-shipment features. On an untouched 20% test set, XGBoost achieved area under the receiver operating characteristic curve (AUC-ROC) 0.765, precision-recall AUC 0.833, Brier score 0.189, and late-class F1 0.690; Random Forest was closely competitive (AUC-ROC 0.762, F1 0.692). Grouped SHapley Additive exPlanations (SHAP) attributed 72.7% of explanation magnitude to shipping mode, payment type, and order hour. Standard Class exposed a major subgroup failure: recall was 0.045 and within-subgroup AUC-ROC was 0.515. Among 3,121 high-risk late orders, 96.7% crossed below 0.50 under at least one valid shipping-mode substitution, demonstrating predictive sensitivity rather than causal recourse. Across nine alternative false-negative/false-positive cost scenarios, validation-selected thresholds outperformed a flag-all-late baseline, although the advantage was sometimes small. The findings support leakage-aware diagnostic auditing rather than autonomous operational deployment.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Veljko Spasojević
Academy of Professional Studies Šumadija, Department of Trstenik
Trstenik, Serbia
Email: vspasojevic@asss.edu.rs

1. INTRODUCTION

Reliable delivery is a core requirement of e-commerce operations, but predictive accuracy alone does not determine whether a model is operationally useful. A decision-support model should distinguish late from non-late orders, communicate uncertainty, reveal the factors driving a warning, and expose conditions under which performance deteriorates. Müller et al. [1] show that delivery-delay research still emphasizes incremental predictive performance more often than practical applicability and integrated decision support. The resulting gap is particularly important in e-commerce, where missed delays and false alarms have different consequences and where aggregate performance can conceal failure for a specific shipping service.

Recent work spans several complementary strands. Ahmed et al. [2] combined SOM, ANN, and SHAP for interpretable supply-chain forecasting on DataCo; Lesmana et al. [3] applied XGBoost to e-commerce delivery-delay prediction; Katangoori [4] integrated delivery-risk classification with broader retail and logistics analytics; Garine and Chakrabortty [5] combined deep learning and policy optimization for order classification; and Mohammed et al. [6] compared machine-learning models for delivery-time prediction in an operational supply-chain setting. Forecasting and resilience studies further combine deep learning or machine learning with explainability and segmentation [7], [8], [9]. These studies demonstrate the predictive value of data-driven models but use different units of analysis, targets, feature sets, and validation protocols, which limits direct comparison of reported performance.

Explainability is increasingly treated as a decision-support requirement rather than an optional visualization. Olan et al. [10] examine XAI capabilities in supply-chain decision support; Dereci and Tuzkaya [11] connect local explanations with maintenance and spare-parts decisions; Sadeghi R. et al. [12] link XAI transparency with agile decision-making; Abdulrashid et al. [13] use SHAP in transport-logistics risk analysis; and Abdulla and Baryannis [14] integrate interpretable machine learning with supplier selection. Counterfactual surveys emphasize feasibility, robustness, and the distinction between model recourse and causal intervention [15], [16], while cost-sensitive classification shows that asymmetric and uncertain error costs can change the preferred decision rule [17]. Methodological work also shows that XGBoost performance depends materially on hyperparameter selection [18], and Shapley-value theory provides the basis for consistent feature-attribution analysis [19]. Together, these studies motivate an evaluation that combines prediction, explanation, calibration, subgroup diagnosis, and decision sensitivity rather than treating each component separately.

The DataCo Smart Supply Chain benchmark used here (Mendeley Data V5, doi: 10.17632/8gx2fvg2k6.5) has already appeared in recent machine-learning and XAI work [2], [4]; therefore, this study does not claim novelty from using DataCo, XGBoost, or SHAP alone. DataCo is organized at line-item level rather than one row per order, so a row-level split can place line items from one order on both sides of the learning boundary and give multi-item orders disproportionate weight. The difference between this research and previous research is that the present study reconstructs one order per sample, excludes realized-shipping information, retains an untouched test set, and evaluates the same pre-shipment model through reference baselines, probability calibration, SHAP stability, subgroup failure, constrained counterfactual sensitivity, cost scenarios, and chronological holdout. The novelty of this research is an integrated diagnostic framework in which these checks are evaluated together while maintaining explicit leakage and causal-interpretation guardrails.

The purpose of this study is to evaluate how reliably late-delivery risk can be predicted from information available at order placement and to determine how explanation stability, probability quality, shipping-service sensitivity, asymmetric costs, and subgroup behavior affect operational interpretation. The contributions are fourfold: (1) order-level DataCo reconstruction with explicit exclusion of outcome-adjacent post-shipment variables; (2) fixed-model predictive evaluation with reference baselines, precision-recall and calibration metrics, bootstrap uncertainty, and chronological sensitivity; (3) global, local, interaction, cross-market, and bootstrap-stability SHAP diagnostics with explicit non-causal interpretation; and (4) expanded shipping-service counterfactuals, validation-only cost-threshold sensitivity, and subgroup failure analysis. The study addresses four research questions: RQ1, how accurately and reliably can late-delivery risk be predicted from information available at order placement, relative to fixed baselines and under chronological holdout? RQ2, which feature groups dominate the model globally and locally, and are their rankings stable across markets and resamples? RQ3, how sensitive are high-risk predictions to valid within-dataset changes in shipping service? RQ4, how do asymmetric costs and shipping-mode subgroups change the apparent operational value of the classifier?

2. METHOD

Figure 1 presents the overall research workflow, from raw-data auditing and order-level reconstruction to leakage-safe model evaluation and the subsequent diagnostic and robustness analyses, while preserving the final 20% test set for untouched evaluation.

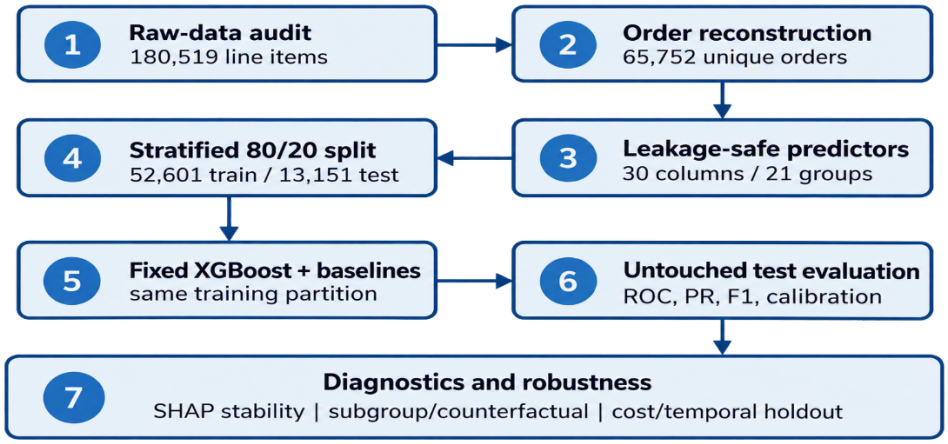


Figure 1. Research workflow for the leakage-safe evaluation framework. The final 20% test set is excluded from model fitting and validation-only threshold selection.

2.1. Dataset, Inclusion Criteria, and Order-Level Reconstruction

The public DataCo Smart Supply Chain dataset (Mendeley Data V5, doi: 10.17632/8gx2fvg2k6.5) contains 180,519 line-item records dated from 1 January 2015 to 31 January 2018, covering five market labels (Africa, Europe, LATAM [Latin America], Pacific Asia, and USCA [United States and Canada]) and 164 order countries. The binary target Late_delivery_risk is invariant within each Order Id. Grouping by Order Id produced 65,752 unique orders: 36,048 late (54.8%) and 29,704 on time or early (45.2%). An audit confirmed that the modeled order-level fields contained no missing values; therefore, all 65,752 reconstructed orders were retained and no final-model observation required deletion or imputation.

Order-level attributes that are constant within an order—shipping mode, payment type, customer segment, market, customer and order countries, customer coordinates, and order timestamp—were retained once. Line-item variables were aggregated into nine order-composition measures: item-line count, total quantity, total discount, mean discount rate, mean product price, total order value, total profit, mean profit ratio, and number of distinct product categories. The target and all retained categorical order attributes were checked for within-order consistency before reconstruction.

2.2. Leakage-Safe Predictors, Split, and Model Configuration

Only information available at or immediately after order placement was used. Post-shipment variables were excluded, including Days for shipping (real), Delivery Status, shipping date, and any feature derived from realized shipping time. Days for shipment (scheduled) was also omitted after verifying that it is deterministically mapped from Shipping Mode in DataCo: Same Day = 0, First Class = 1, Second Class = 2, and Standard Class = 4 scheduled days. Shipping mode was retained as a single service-speed code from 0 (Standard Class) to 3 (Same Day) to avoid duplicate representations of the same service configuration.

Payment type, customer segment, and market were one-hot encoded. The remaining conceptual predictors were same-country status; order month, day of week, hour, quarter, and weekend indicator; customer latitude and longitude; and the nine order-composition measures defined above. The resulting matrix contained 30 encoded columns representing 21 conceptual feature groups.

A stratified 80/20 split with random seed 42 produced 52,601 training orders and an untouched test set of 13,151 orders. XGBoost [18] was selected a priori as the primary model because gradient-boosted trees can represent nonlinear interactions in heterogeneous tabular predictors, return probabilistic risk scores, and support TreeSHAP decomposition. The fixed configuration was 300 estimators, maximum depth 6, learning rate 0.1, subsample 0.8, column subsampling 0.8, minimum child weight 3, random seed 42, and scale_pos_weight = 0.824 based only on the training labels. These values were retained unchanged from the submitted model to preserve comparability and avoid post-hoc tuning during revision; they are not claimed to be empirically optimal. Prior work shows that XGBoost tuning can materially affect performance [18].

For reference, three fixed untuned baselines were trained on the same 52,601-order training partition and evaluated on the same untouched test set: standardized Logistic Regression (C = 1.0, lbfgs, maximum 5,000 iterations), a Decision Tree with default complexity parameters and random seed 42, and a Random Forest with 300 trees, default tree-complexity parameters, and random seed 42. The same training-derived class weighting (positive-class weight 0.824; negative-class weight 1.0) was applied. These models are diagnostic comparators rather than optimized challengers, and no model or threshold was selected using the final test set.

The primary held-out evaluation at t = 0.50 retained accuracy, class-specific precision, recall, and F1 together with the area under the receiver operating characteristic curve (AUC-ROC). Precision-recall area under the curve (PR-AUC), average precision (AP), Brier score, and calibration error were added to assess ranking quality under class imbalance and the quality of predicted probabilities. PR-AUC was calculated by trapezoidal integration of the precision-recall curve; AP is reported alongside it because coarse probability outputs can make interpolated PR-AUC optimistic. Calibration was summarized by the Brier score, expected calibration error (ECE) over ten quantile bins, and reliability curves. The original XGBoost endpoint confidence intervals used 1,000 bootstrap resamples (seed 20260825); added model-comparison metrics and paired XGBoost-baseline differences used 500 bootstrap resamples.

Five-fold stratified cross-validation with shuffle and seed 42 was performed only within the 80% training partition; the fixed XGBoost configuration was refit in each fold and the class weight was recalculated from that fold's training labels. Fold-to-fold mean ± standard deviation was reported as a stability check. As an additional temporal sensitivity analysis, orders were sorted by timestamp, the earliest 80% (52,601 orders) were used for training, and the latest 20% (13,151 orders; 9 July 2017 to 31 January 2018) were held out chronologically and evaluated at t = 0.50 with the same fixed configuration.

The canonical analysis was run in Python 3.11.0 using pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, XGBoost 3.1.3, SHAP 0.50.0, SciPy 1.17.0, and Matplotlib 3.10.8.

2.3. Explainability and Operational Diagnostics

The held-out XGBoost evaluation at t = 0.50 remained the primary analysis. Baseline comparison, probability calibration, SHAP, interaction, counterfactual, cost, subgroup, market, grouped-SHAP bootstrap, and chronological-holdout analyses were secondary diagnostic or robustness checks. No null-hypothesis significance tests were used; uncertainty was reported through bootstrap confidence intervals and training-fold variability. Figure 1 summarizes the complete workflow and the separation between model development and final evaluation.

TreeSHAP was calculated on the untouched test set, following the broader Shapley-value feature-attribution framework summarized in [19]. Absolute contributions from one-hot variables were summed within conceptual feature groups before percentage importance was computed. Local explanations covered representative true-positive, true-negative, false-positive, and false-negative predictions, while pairwise interactions used the first 500 test observations. Explanation robustness was assessed by market-specific grouped SHAP with Spearman correlations of full rankings and by 200 bootstrap resamples (seed 7) of the grouped absolute-SHAP matrix to test preservation of the top-three set and order.

For the counterfactual diagnostic, every actual-late test order with submitted-model P(late) > 0.75 was evaluated rather than sampling ten cases. Each eligible order was rescored under the three alternative valid shipping modes while all other pre-shipment predictors were held fixed. Because shipping mode and scheduled days represent the same service choice, scheduled days was not manipulated separately. Among alternatives crossing below P(late) = 0.50, the smallest service-code change was selected, using predicted probability as a tie-breaker; if no alternative crossed the threshold, the lowest-risk alternative was retained. These values quantify predictive sensitivity only and are not causal treatment effects or service recommendations.

Cost analysis used illustrative cost units rather than empirically measured monetary values. The anchor scenario assigned false negative (FN) = 20, false positive (FP) = 4, true positive (TP) = 2, and true negative (TN) = 0, representing a setting in which a missed late delivery is substantially more costly than a false alert while a correctly flagged order can still incur intervention effort. Because exact operational costs are uncertain [17], sensitivity was evaluated over nine FN/FP combinations: FN ∈ {10, 20, 40} and FP ∈ {2, 4, 8}, with TP = 2 and TN = 0. For every scenario, the 52,601 training orders were split into 42,080 selector-training and 10,521 validation orders (seed 7); a temporary fixed-configuration XGBoost model selected the minimum-validation-cost threshold over t = 0.01-0.90. The final submitted model was then evaluated once on the untouched test set at the validation-selected threshold and compared with all-on-time and flag-all-late strategies.

The complete revision script reproducing the submitted XGBoost model identity, order reconstruction, fixed baselines, probability metrics, calibration, SHAP analyses, expanded counterfactuals, multi-scenario cost sensitivity, subgroup diagnostics, chronological sensitivity analysis, and figure generation is available from the corresponding author upon reasonable request.

3. RESULTS AND DISCUSSION

3.1. Predictive Performance, Baselines, and Calibration

On the 13,151-order untouched test set, XGBoost achieved AUC-ROC = 0.765, PR-AUC = 0.833, late-class F1 = 0.690, and Brier score = 0.189 (Table 1); the ten-bin ECE was 0.041. Training-partition cross-validation yielded AUC-ROC = 0.764 ± 0.006 and late-class F1 = 0.695 ± 0.005. The chronological holdout yielded AUC-ROC = 0.764, late-class F1 = 0.695, and late recall = 0.610. The PR-AUC and Brier intervals in Table 1 use 500 bootstrap resamples, while the remaining primary endpoint intervals retain the original 1,000-resample procedure. To contextualize the primary XGBoost results, Table 2 compares its held-out performance with the three fixed, untuned reference classifiers evaluated on the same untouched test set.

Table 1. Held-out XGBoost performance with bootstrap confidence intervals.

Metric	Estimate	95% bootstrap CI
Accuracy	0.709	0.702-0.717
AUC-ROC	0.765	0.758-0.773
PR-AUC	0.833	0.826-0.840
Brier score	0.189	0.185-0.192
Late precision	0.829	0.819-0.840
Late recall	0.591	0.581-0.603
Late F1	0.690	0.681-0.700
On-time precision	0.632	0.622-0.643
On-time recall	0.852	0.844-0.861
On-time F1	0.726	0.718-0.734

Table 2. Fixed untuned reference-model comparison on the same untouched test set.

Model	AUC-ROC	PR-AUC	AP	Brier	Late F1
XGBoost	0.765	0.833	0.833	0.189	0.690
Logistic Regression	0.710	0.724	0.724	0.217	0.653
Decision Tree	0.637	0.762	0.631	0.359	0.674
Random Forest	0.762	0.829	0.827	0.189	0.692

XGBoost had the highest AUC-ROC and PR-AUC in the fixed comparison, but Random Forest was closely competitive and had a marginally higher late-class F1. In paired bootstrap resampling, the XGBoost-Random Forest AUC-ROC difference was 0.0036 (95% CI -0.0014 to 0.0086), while the Brier-score and late-F1 differences also included zero; only the small PR-AUC advantage (0.0041, 95% CI 0.0009-0.0068) excluded zero. The comparison therefore does not support a broad superiority claim over Random Forest. The untuned Decision Tree produced highly discrete probabilities, so its AP = 0.631 is reported alongside the interpolation-sensitive PR-AUC = 0.762. Figure 2 shows the comparative ROC and precision-recall curves, calibration behavior, and the primary XGBoost confusion matrix.

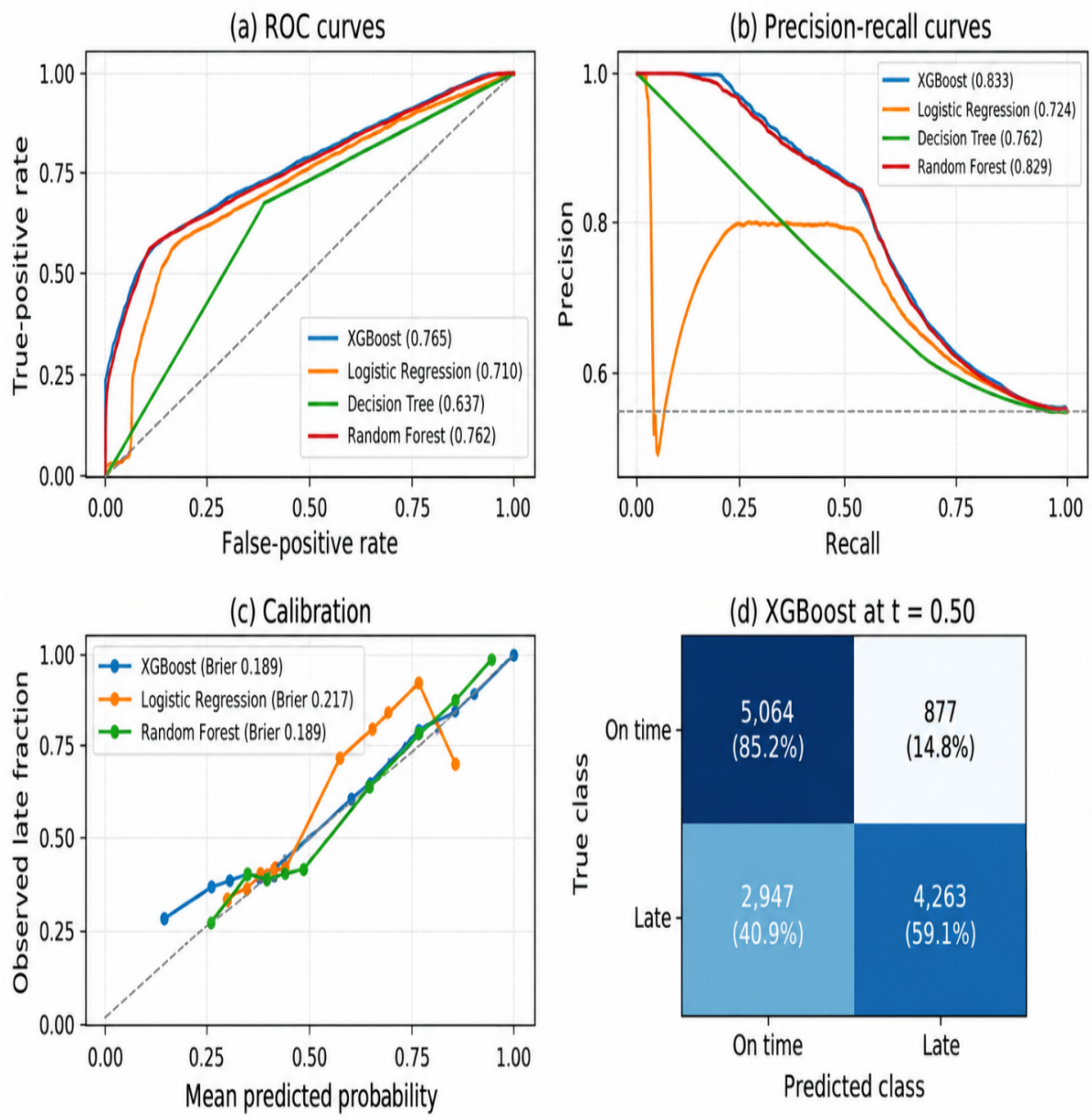


Figure 2. Held-out evaluation: (a) ROC curves, (b) precision-recall curves, (c) ten-bin probability calibration for models with non-degenerate probability scores, and (d) XGBoost confusion matrix at t = 0.50. Values in parentheses are AUC-ROC in (a), PR-AUC in (b), and Brier score in (c).

3.2. Exploratory SHAP Diagnostics

Grouped SHAP importance was concentrated in Shipping Mode (49.5%), Payment Type (15.7%), and Order Hour (7.4%); together these groups accounted for 72.7% of total grouped mean absolute SHAP magnitude. The next groups were Customer Latitude (2.8%), Total Profit (2.7%), Total Order Value (2.6%), and Mean Product Price (2.4%). Across 200 bootstrap resamples of the grouped SHAP matrix, the same three feature groups appeared in the same order in 100% of resamples. Figure 3(a) shows grouped importance, while Figure 3(b) shows the encoded-feature SHAP distribution for the leading features.

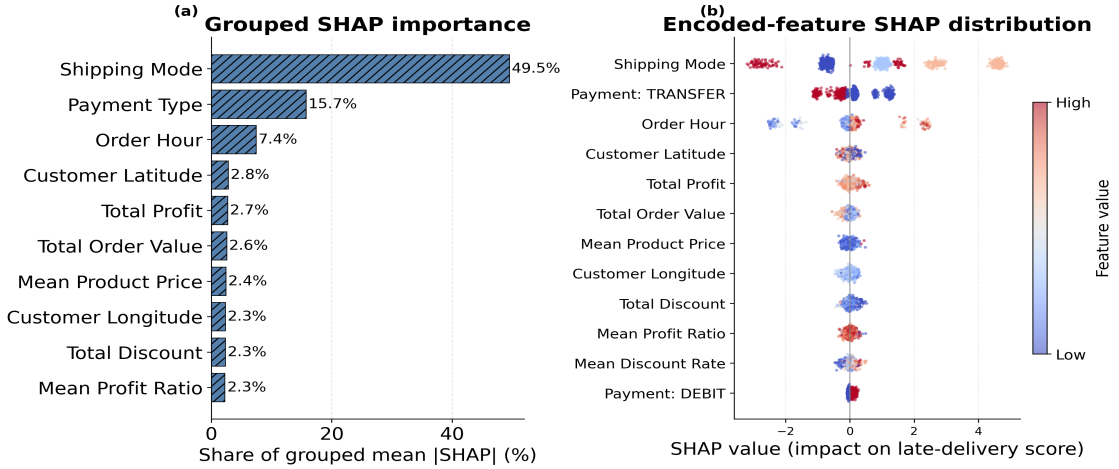


Figure 3. SHAP diagnostics on the untouched test set: (a) grouped mean absolute SHAP share and (b) encoded-feature SHAP distribution for the leading features.

On the fixed 500-order interaction subset, the three strongest grouped interactions were Shipping Mode × Payment Type (0.255), Shipping Mode × Order Hour (0.181), and Payment Type × Order Hour (0.044). Figure 4 shows redesigned local explanations for representative true-positive, true-negative, false-positive, and false-negative outcomes using the six largest absolute contributions in each case. Positive SHAP values increase the late-delivery score and negative values decrease it.

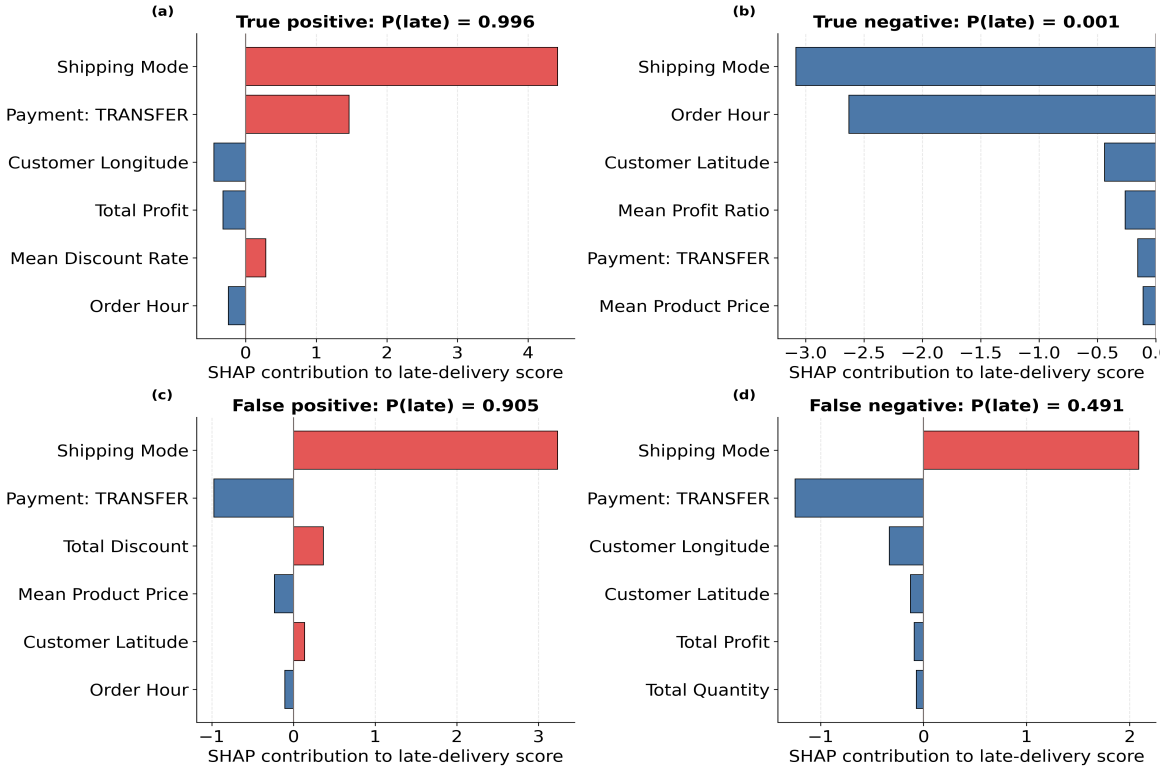


Figure 4. Representative local SHAP explanations on the untouched test set: (a) true positive, (b) true negative, (c) false positive, and (d) false negative. Red bars denote positive contributions that increase the late-delivery score; blue bars denote negative contributions that decrease it.

3.3. Exploratory Counterfactual Service Scenarios

Table 3. Representative expanded shipping-mode counterfactuals (pp = percentage points).

Case	Base P(late)	Current	Alternative	New P(late)	Change
1	0.750	Second	Standard	0.322	-42.9 pp
2	0.816	Second	Standard	0.363	-45.3 pp
3	0.991	Same Day	Standard	0.413	-57.8 pp
4	0.998	First	Standard	0.341	-65.6 pp
5	1.000	First	Standard	0.646	-35.4 pp

The expanded cohort contained 3,121 actual-late test orders with P(late) > 0.75. At least one valid shipping-mode substitution moved 3,017 cases (96.7%) below P(late) = 0.50; the mean and median reductions were 64.5 and 61.0 percentage points, respectively. The remaining 104 cases stayed above 0.50 under all valid alternatives. Table 3 includes representative cases across the high-risk range, including one non-crossing example. Many selected alternatives move toward Standard Class because the submitted model assigns that service a comparatively negative contribution in this historical benchmark. This pattern reinforces that the counterfactuals measure model sensitivity, not the causal effect or operational desirability of changing service.

3.4. Exploratory Cost Sensitivity

Table 4 summarizes the anchor cost scenario on the untouched test set, comparing the validation-selected XGBoost threshold with the default threshold and the all-on-time and flag-all-late baseline strategies.

Table 4. Anchor illustrative cost scenario on the untouched test set (cost units).

Strategy	Scenario cost	Reduction vs. all-on-time	Savings / 1,000
All on time	144,200	—	—
Flag all as late	38,184	73.5%	8,061
XGBoost (t = 0.50)	70,974	50.8%	5,568
XGBoost (validation t = 0.08)	36,676	74.6%	8,176

In the anchor scenario, validation selected t = 0.08 without consulting the final test set. Applied to the submitted model, this threshold produced test-set recall = 1.000 and cost 36,676, a 74.6% reduction relative to the all-on-time strategy and 1,508 total cost units (114.7 per 1,000 orders) lower than flag-all-late. The broader FN/FP grid selected thresholds from 0.08 to 0.54. The model was cheaper than flag-all-late in all nine scenarios, but the advantage ranged widely from 57.3 to 1,291.5 cost units per 1,000 orders. Figure 5 shows that the practical advantage is smallest when false positives are inexpensive and increases when unnecessary alerts are more costly; the cost analysis therefore supports threshold adaptation rather than a universal operating point.

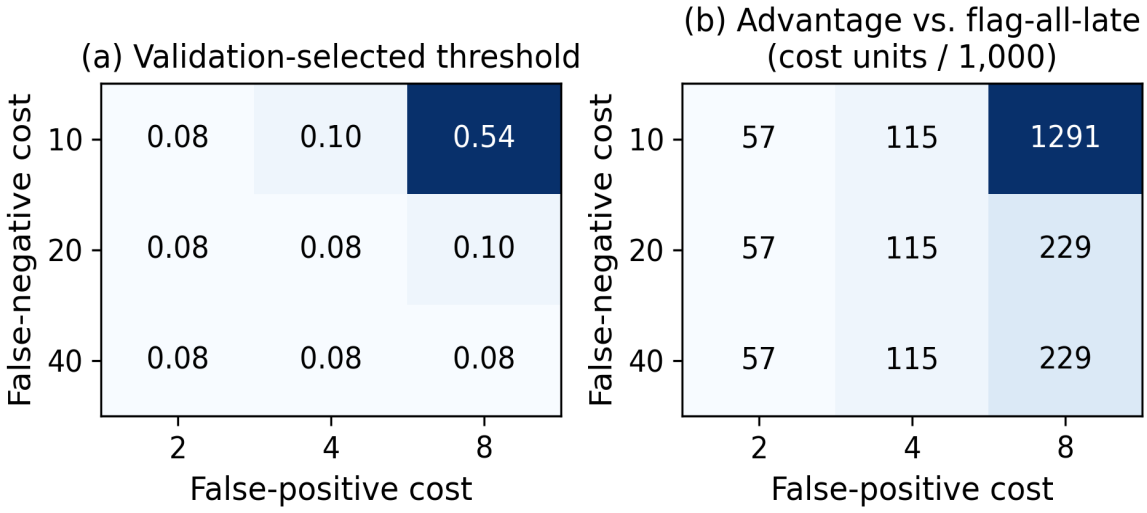


Figure 5. Multi-scenario cost sensitivity: (a) validation-selected classification thresholds and (b) test-set model advantage over flag-all-late across the FN/FP grid. TP = 2 and TN = 0 in all panels; all values are illustrative cost units.

3.5. Exploratory Subgroup and Cross-Market Diagnostics

Standard Class contained 7,947 test orders and exposed a failure that is not explained by the 0.50 threshold alone. Late-class recall was 0.045 and F1 was 0.081 at t = 0.50, while within-Standard discrimination was nearly absent (AUC-ROC = 0.515; AP = 0.394). Median P(late) was 0.338 for actually late Standard orders and 0.333 for on-time Standard orders. The mean Shipping Mode SHAP contribution among actual-late Standard orders was -0.744, indicating that the learned service-mode effect systematically reduced their scores. Lowering the threshold to the cost-selected t = 0.08 raised Standard recall to 1.000 only by flagging every Standard order (specificity = 0; precision = 0.385). Thus, threshold adjustment can recover sensitivity but does not restore within-subgroup discrimination. Recall for Second Class, First Class, and Same Day at t = 0.50 remained 0.986, 0.999, and 1.000, respectively (Table 5).

Table 5. Late-class performance by shipping mode on the held-out test set.

Shipping mode	Orders	Late %	Accuracy	Precision	Recall	F1
Standard Class	7,947	38.5%	0.606	0.393	0.045	0.081
Second Class	2,529	77.5%	0.767	0.775	0.986	0.868
First Class	1,951	95.2%	0.952	0.952	0.999	0.975
Same Day	724	46.5%	0.986	0.971	1.000	0.985

Across the five market labels, Shipping Mode accounted for 48.1%-50.5% of grouped SHAP magnitude, Payment Type for 14.6%-16.3%, and Order Hour for 7.0%-8.9%. Pairwise Spearman correlations between full market-specific feature rankings ranged from 0.917 to 0.996 (mean 0.966). The global top-three feature-group ordering was preserved in 100% of the 200 SHAP bootstrap resamples. Because subgroup sizes differ substantially, these subgroup performance estimates are descriptive and are not used for inferential claims.

3.6. Interpretation and Relation to Prior Work

RQ1 is answered by moderate rather than exceptional discrimination and by the absence of a clear aggregate winner over Random Forest. XGBoost achieved the highest AUC-ROC and PR-AUC, but Random Forest had nearly identical AUC-ROC and Brier score and a marginally higher late-class F1. The chronological holdout remained close to the random-split result, supporting internal temporal stability within DataCo. These estimates are lower than the high predictive performance reported in some recent DataCo and e-commerce studies [2], [3], [4], but this is not necessarily inconsistent: those studies differ in unit of analysis, target construction, feature availability, model optimization, and validation design. The present estimates are intentionally tied to one-order-per-sample, pre-shipment prediction with realized-shipping variables excluded, so direct numerical ranking against those studies would be inappropriate.

For RQ2, shipping mode dominates SHAP magnitude and remains the leading explanation group in every market label. This is broadly consistent with prior work that identifies transportation or service variables as important logistics predictors [6] and with studies using SHAP to expose operational drivers [2], [13]. However, the Standard Class diagnosis qualifies that consistency: a feature can be globally influential while also encoding benchmark-specific service policy that contributes to subgroup failure. In DataCo, shipping mode deterministically determines scheduled shipping duration and is associated with markedly different late rates, so its attribution is descriptive rather than causal. Payment type and order hour provide additional signal, while the high cross-market rank correlations and preservation of the top-three ordering across all 200 bootstrap resamples indicate stable explanations inside this dataset only.

RQ3 is strengthened by the expanded rather than case-sampled counterfactual evaluation. Among all 3,121 high-risk actual-late orders, 96.7% had at least one valid shipping-mode substitution that moved the model below 0.50, with a median risk reduction of 61.0 percentage points. The scale of this effect confirms that the classifier is highly sensitive to the service label, but the frequent movement toward Standard Class also demonstrates why model counterfactuals cannot be interpreted as operational interventions. Holding every other predictor fixed does not reproduce the logistics consequences of actually changing fulfillment service. This distinction is consistent with counterfactual literature emphasizing feasibility, robustness, and caution in causal interpretation [15], [16].

For RQ4, asymmetric costs materially change the preferred operating point, consistent with cost-sensitive classification research showing that unequal and uncertain error costs alter decision rules [17]. Validation-selected thresholds ranged from 0.08 to 0.54 across the nine FN/FP scenarios, and the model was cheaper than flag-all-late in every scenario. However, the advantage was only 57.3 cost units per 1,000 orders in the least favorable grid settings, so a lower scenario cost should not be confused with material operational benefit. The subgroup result is more consequential: aggregate AUC-ROC and F1 conceal near-random

discrimination for Standard Class, the largest test subgroup. Together, these findings show that deployment decisions require both cost assumptions and subgroup-specific validation.

3.7. Practical and Research Implications

For practice, the results support using pre-shipment models as triage and audit tools rather than autonomous dispatch rules. Organizations should assess probability calibration and select operating thresholds on current local data, compare candidate models against simple baselines, and gate deployment by shipping-service performance rather than aggregate AUC-ROC alone. The Standard Class result shows that lowering a threshold can increase recall while merely converting the subgroup into an all-flag policy. Counterfactual outputs should therefore be presented as predictive sensitivity scenarios unless a proposed service change is supported by causal or prospective operational evidence.

For research, the analysis highlights five reporting practices that improve the credibility of public logistics benchmarks: reconstruct the operational unit before splitting, isolate all decision rules from the final test set, report probability quality in addition to discrimination, inspect subgroup failure alongside aggregate metrics, and test whether explanations and performance persist under resampling and time ordering. The present analyses address these issues within DataCo; the next priority is independent validation on contemporary e-commerce platforms, carriers, and prospective operational data.

3.8. Limitations and Validity

Several limitations constrain interpretation. First, DataCo covers 2015-2018 and a single public platform. E-commerce fulfillment, carrier networks, service promises, and customer behavior have changed since that period, so the present probabilities, thresholds, and feature effects should not be transferred directly to contemporary operations. Second, the late label is structurally related to the promised shipping schedule, while shipping mode determines scheduled days in this dataset. This coupling is informative for prediction but limits causal interpretation of service effects. Third, XGBoost and the three reference baselines use fixed, untuned configurations. The comparison is deliberately controlled and test-set independent, but it does not establish the best achievable algorithm or hyperparameter setting.

Fourth, the chronological holdout is internal to the same 2015-2018 benchmark and therefore tests sensitivity to time ordering rather than transportability to a contemporary external population. Fifth, the expanded counterfactual analysis covers all high-risk late test cases but still changes one service label while holding correlated operational conditions fixed; it remains a predictive stress test, not a causal intervention study. The SHAP interaction analysis uses a fixed 500-order subset and does not quantify uncertainty for individual interaction estimates. Sixth, the cost values are illustrative units rather than measured business costs; the nine-scenario grid reduces dependence on one ratio but cannot replace organization-specific economic estimation. Finally, calibration and subgroup results are internal to DataCo and may drift under different carriers, product mixes, service policies, or time periods; subgroup estimates remain descriptive and are not used for inferential claims.

4. CONCLUSION

Using one order per sample and only information available before shipment, the fixed XGBoost model achieved AUC-ROC = 0.765, PR-AUC = 0.833, Brier score = 0.189, and late-class F1 = 0.690, while Random Forest showed closely comparable performance and the chronological holdout produced similar results within the historical DataCo benchmark. Explainability analysis identified shipping mode, payment type, and order hour as the dominant drivers of model predictions, with their leading importance remaining stable across markets and bootstrap resamples. At the same time, subgroup analysis revealed a critical weakness hidden by aggregate metrics, as Standard Class deliveries exhibited near-random within-subgroup discrimination (AUC-ROC = 0.515) and very low late-class recall at the default threshold. Expanded counterfactual evaluation further showed strong predictive sensitivity to shipping-mode substitutions, while cost-sensitivity analysis demonstrated that the preferred decision threshold depends substantially on the assumed balance between missed delays and false alerts. Taken together, these findings show that apparently acceptable aggregate predictive performance is not sufficient evidence of reliable operational usefulness. The main scientific contribution of this study is therefore not a new prediction algorithm, but an integrated and reproducible diagnostic framework that combines leakage-aware order-level reconstruction, fixed baseline comparison, discrimination and probability calibration, explainability stability analysis, subgroup failure diagnosis, expanded counterfactual sensitivity, cost-sensitive threshold evaluation, and chronological validation. Within the evaluated benchmark, this framework provides a systematic way to identify model weaknesses that would otherwise remain hidden before operational deployment, while also emphasizing that explainability outputs and counterfactual scenarios should be interpreted as diagnostic evidence rather than direct operational prescriptions. Because the analysis is based on the historical DataCo dataset from 2015–2018, the reported

probabilities, thresholds, and feature effects should not be assumed to transfer directly to contemporary e-commerce environments. Future work should therefore prioritize validation on independent and more recent datasets, incorporate carrier and real-time tracking information, estimate organization-specific operational costs, evaluate recalibration under distribution shift, and prospectively test whether suggested operational changes lead to measurable improvements in delivery performance.

ACKNOWLEDGEMENTS

The authors acknowledge the maintainers of the DataCo Smart Supply Chain dataset for making the data publicly available for research. This research received no external funding. The authors declare no conflict of interest.

REFERENCES

[1] N. Müller, P. Burggräf, F. Steinberg, C. R. Sauer, and M. Schütz, "An analytical review of predictive methods for delivery delays in supply chains," *Supply Chain Anal.*, vol. 11, p. 100130, Sep. 2025, doi: 10.1016/j.sca.2025.100130.

[2] K. R. Ahmed, M. E. Ansari, Md. N. Ahsan, A. Rohan, M. B. Uddin, and M. A. H. Rivin, "Deep learning framework for interpretable supply chain forecasting using SOM ANN and SHAP," *Sci. Rep.*, vol. 15, no. 1, p. 26355, Jul. 2025, doi: 10.1038/s41598-025-11510-z.

[3] Stevanus Putra Lesmana, Dina Hermawati, Maulina Mukaromah, Iqbal Ahmad Bukhari, and Norma Puspitasari, "Machine Learning Implementation for E-commerce Delivery Delay Prediction Using XGBoost Algorithm," *Green Eng. Int. J. Eng. Appl. Sci.*, vol. 2, no. 3, pp. 46–57, Jul. 2025, doi: 10.70062/greenengineering.v2i3.219.

[4] A. Katangoori, "An empirical analysis of data-driven supply chain optimization in retail and logistics," *Supply Chain Anal.*, vol. 14, p. 100206, Jun. 2026, doi: 10.1016/j.sca.2026.100206.

[5] R. Garine and R. K. Chakrabortty, "A deep learning and policy optimization approach for supply chain order classification," *Supply Chain Anal.*, vol. 12, p. 100166, Dec. 2025, doi: 10.1016/j.sca.2025.100166.

[6] A. Mohammed *et al.*, "Machine learning models integrated with strategic theories for predicting delivery time in supply chains," *Results Eng.*, vol. 30, p. 111069, Jun. 2026, doi: 10.1016/j.rineng.2026.111069.

[7] X. Li, V. Krivtsov, C. Pan, A. Nassehi, R. X. Gao, and D. Ivanov, "End-to-end supply chain resilience management using deep learning, survival analysis, and explainable artificial intelligence," *Int. J. Prod. Res.*, vol. 63, no. 3, pp. 1174–1202, Feb. 2025, doi: 10.1080/00207543.2024.2367685.

[8] S. K. Jauhar, S. Harinath, V. Krishnaswamy, and S. K. Paul, "Explainable artificial intelligence to improve the resilience of perishable product supply chains by leveraging customer characteristics," *Ann. Oper. Res.*, vol. 354, no. 1, pp. 103–142, Nov. 2025, doi: 10.1007/s10479-024-06348-z.

[9] M. Irhuma, A. Alzubi, T. Öz, and K. Iyiola, "Migrative armadillo optimization enabled a one-dimensional quantum convolutional neural network for supply chain demand forecasting," *PLOS ONE*, vol. 20, no. 3, p. e0318851, Mar. 2025, doi: 10.1371/journal.pone.0318851.

[10] F. Olan, K. Spanaki, W. Ahmed, and G. Zhao, "Enabling explainable artificial intelligence capabilities in supply chain decision support making," *Prod. Plan. Control*, vol. 36, no. 6, pp. 808–819, Apr. 2025, doi: 10.1080/09537287.2024.2313514.

[11] U. Dereci and G. Tuzkaya, "An explainable artificial intelligence model for predictive maintenance and spare parts optimization," *Supply Chain Anal.*, vol. 8, p. 100078, Dec. 2024, doi: 10.1016/j.sca.2024.100078.

[12] K. Sadeghi R., D. Ojha, P. Kaur, R. V. Mahto, and A. Dhir, "Explainable artificial intelligence and agile decision-making in supply chain cyber resilience," *Decis. Support Syst.*, vol. 180, p. 114194, May 2024, doi: 10.1016/j.dss.2024.114194.

[13] I. Abdulrashid, R. Zanjirani Farahani, S. Mammadov, M. Khalafalla, and W.-C. Chiang, "Explainable artificial intelligence in transport Logistics: Risk analysis for road accidents," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 186, p. 103563, Jun. 2024, doi: 10.1016/j.tre.2024.103563.

[14] A. Abdulla and G. Baryannis, "A hybrid multi-criteria decision-making and machine learning approach for explainable supplier selection," *Supply Chain Anal.*, vol. 7, p. 100074, Sep. 2024, doi: 10.1016/j.sca.2024.100074.

[15] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review," *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–42, Dec. 2024, doi: 10.1145/3677119.

[16] J. Jiang, F. Leofante, A. Rago, and F. Toni, "Robust Counterfactual Explanations in Machine Learning: A Survey," in *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, Jeju, South Korea: International Joint Conferences on Artificial Intelligence Organization, Aug. 2024, pp. 8086–8094. doi: 10.24963/ijcai.2024/894.

[17] V. Komisarenko and M. Kull, "Cost-sensitive classification with cost uncertainty: do we need surrogate losses?," *Mach. Learn.*, vol. 114, no. 6, p. 132, Jun. 2025, doi: 10.1007/s10994-024-06634-8.

[18] T. Lima Marinho, D. C. Do Nascimento, and B. A. Pimentel, "Optimization on selecting XGBOOST hyperparameters using meta-learning," *Expert Syst.*, vol. 41, no. 9, p. e13611, Sep. 2024, doi: 10.1111/exsy.13611.

[19] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence," *Auton. Intell. Syst.*, vol. 4, no. 1, p. 2, Feb. 2024, doi: 10.1007/s43684-023-00060-8.