When Better Prediction Is Invalid: Decision-Time Leakage and Finite-Capacity Policy Value in Delivery-Risk Modelling

Meenakshi Yadav
School of Computer Applications & IT, Lords University, Alwar, Rajasthan 301028, India

Gurjeet Singh, PhD
School of Computer Applications & IT, Lords University, Alwar, Rajasthan 301028, India

ABSTRACT
Late-delivery classifiers are actionable only when every input exists before intervention. This study treats feature availability as part of the decision problem and measures how crossing that boundary inflates both predictive metrics and apparent finite-capacity policy value. Using 180,519 DataCo orders, five cumulative boundaries are audited: B0 has 14 strict order-time fields; B1 adds five valid order-placement attributes; B2-B4 deliberately add dispatch, realised-transit, and delivery-outcome information as leakage controls. CatBoost ROC-AUC rises from 0.7468 at B0 to 0.7954 at B1, then to 0.9965 at B2 and 1.0000 at B4. At 20% intervention capacity and a 3:1 missed-delay:false-alarm cost ratio, B0 reduces decision cost by 32.53%, B1 by 34.11%, and invalid B2-B4 by 36.47%, with normalized decision regret collapsing to zero. The valid B1 gain survives five-fold, forward-time, and bootstrap checks. The contribution is a reproducible leakage-to-policy-value audit linking temporal feature validity to discrimination, calibration, and resource-constrained decision value rather than proposing a new classifier or causal intervention model.

General Terms
Machine Learning, Decision Support, Supply Chain Analytics, Model Validation

Keywords
Delivery risk, data leakage, decision-time validity, supply chain, CatBoost, policy value, resource constraint.

1. INTRODUCTION
Late-delivery prediction is useful only if risk is identified early enough to allocate review or expediting capacity. Supply-chain machine-learning research therefore faces an implementation question beyond accuracy: whether the data used by a model actually exist when the operational decision must be made [1, 9]. A feature can be highly predictive yet invalid for prospective action if it is populated only after dispatch or delivery.

The DataCo benchmark makes this boundary visible because a completed transaction row can contain order, dispatch, transit, and delivery information together. Retrospective classification can legitimately use later fields, but a pre-shipment intervention model cannot. Leakage is thus not only a data-mining problem [7]; in event-driven supply chains it is a decision-time validity problem that can also contaminate the business case built from model scores.

This paper constructs five cumulative feature boundaries on the same orders. B0 and B1 are valid at order placement; B2-B4 deliberately cross the timeline and serve only as diagnostic controls. RQ1 measures predictive change within and beyond the valid boundary, RQ2 tests whether the B1 gain survives resampling, time, and model-family checks, and RQ3 measures how the same boundary changes a fixed-capacity prioritisation policy. The novelty is the linkage from feature-time validity to downstream policy-value inflation and normalized regret, not a new classifier, calibration method, or causal expediting policy [10].

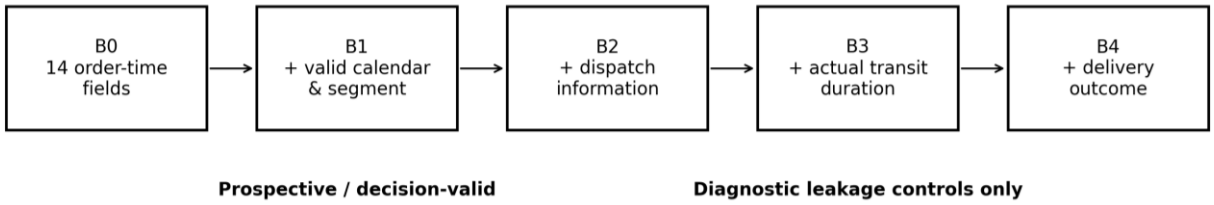

Figure 1. Cumulative decision-time boundaries. B2-B4 are diagnostic leakage controls, not deployable models.

2. RELATED WORK AND RESEARCH GAP
2.1 Positioning and decision-validity gap
Recent supply-chain and delivery-risk work emphasizes predictive performance, interpretability, and implementation readiness, while complex benchmark studies show that very high classification metrics are technically attainable [2, 9]. The relevant comparison for an early-warning system, however, is not numerical prestige but whether each feature precedes the intervention deadline. Data leakage formalizes the general validity problem; here it is made operational by tagging the earliest business event at which each input becomes available [7]. Ranking and calibration remain important once the information boundary is valid, and action-oriented learning further asks whether scores improve a feasible decision [8]. Yet neither calibration nor optimization repairs a temporal violation. The policy experiment therefore separates prediction from causal treatment effects: it evaluates whether scarce review capacity is allocated to orders that become late, not whether an intervention is proven to prevent lateness [10]. With leaked outcome-proximal features, the ranking can converge toward a same-capacity oracle, making the apparent policy benefit a retrospective upper bound rather than deployable value.

3. METHODOLOGY
3.1 Data and target
The DataCo Smart Supply Chain benchmark contains 180,519

orders and 53 fields [4]. The binary Late_delivery_risk target has 98,977 positives (54.83%) and 81,542 negatives, so ROC-AUC is interpreted without a severe rare-event correction while PR-AUC is also reported.

3.2 Decision-time feature boundaries

Table 1. Cumulative feature boundaries used in the decision-time audit.

Boundary	Decision-time details
B0	Order placement | 14 fields | Valid Submitted order-time operational and commercial fields
B1	Order placement | 19 fields | Valid Customer segment + order year/month/day-of-week/hour
B2	Actual dispatch | 24 fields | Invalid for pre-shipment action Shipping timestamp components + elapsed time to dispatch
B3	Post-transport duration | 25 fields | Invalid Actual days for shipping
B4	Delivery/post-outcome | 26 fields | Invalid Final Delivery Status

The same modelling family, row splits, and evaluation metrics are used across B0-B4, so changes can be attributed to information timing. B2-B4 are intentionally invalid for pre-shipment action and are included only to measure retrospective advantage.

3.3 Predictive model and validation

CatBoost is the principal learner, using deterministic categorical encoding, 300 iterations, depth 7, learning rate 0.06, and seed 42. The primary evaluation is an 80/20 stratified holdout. B0 and B1 are additionally compared on five shared stratified folds, a forward 2018 test, and a 500-resample paired bootstrap; logistic regression, random forest, XGBoost, and LightGBM provide model-family checks of the valid enrichment.

3.4 Predictive metrics

ROC-AUC measures ranking, PR-AUC emphasizes the late class, Brier score measures probabilistic accuracy, and F1 summarizes threshold performance [3, 6]. Because the policy ranks probabilities under a capacity limit, discrimination and probability quality are more relevant than raw accuracy.

3.5 Fixed-capacity policy-value audit

For each predicted late-delivery probability and capacity share kappa, the policy selects the K = floor(kappa N) highest-risk orders. A selected on-time order incurs false-alarm cost and an unselected late order incurs missed-delay cost. This is a ranking loss under a fixed hypothetical action budget, not a causal model of expediting effectiveness [5].

$$J(\pi) = CFN \times FN_{unselected}(\pi) + CFP \times FP_{selected}(\pi) \quad (1)$$

$$\text{Policy reduction} = 100\,[J(\text{no action}) - J(\pi)] / J(\text{no action}) \quad (2)$$

A same-capacity oracle selects realised late orders first. Normalized decision regret (NDR) compares each policy with that oracle and the no-action baseline:

$$NDR(\pi) = [J(\pi) - J(\text{oracle})] / [J(\text{no action}) - J(\text{oracle})] \quad (3)$$

NDR = 0 equals same-capacity oracle performance and NDR = 1 equals the no-action baseline. The primary audit uses 20% capacity and a 3:1 missed-delay:false-alarm ratio; all costs are normalized decision units, not audited transport expenditure.

4. RESULTS

4.1 Predictive performance across decision-time boundaries

B0 achieves ROC-AUC 0.7468. Adding valid order-time enrichment in B1 raises ROC-AUC to 0.7954 and lowers Brier score from 0.1931 to 0.1780. Once dispatch information enters at B2, ROC-AUC jumps to 0.9965; B4 reaches 1.0000. These near-perfect values are diagnostic evidence of outcome proximity, not superior deployable intelligence.

Table 2. Predictive performance as later event information is admitted.

Boundary	Fields	ROC-AUC	PR-AUC	Brier	F1
B0	14	0.7468	0.8187	0.1931	0.6702
B1	19	0.7954	0.8513	0.1780	0.7004
B2	24	0.9965	0.9970	0.0193	0.9774
B3	25	0.9965	0.9970	0.0194	0.9774
B4	26	1.0000	1.0000	0.0000	1.0000

4.2 Validation of the enriched order-time boundary

Table 3. Validation of the valid B1 order-time enrichment relative to B0.

Validation	B0 ROC-AUC	B1 ROC-AUC	Difference	Evidence
Five-fold mean	0.7447	0.7930	+0.0483	One-sided Wilcoxon p=0.03125
2018 forward test	0.7426	0.7723	+0.0297	n_test=2123
Paired holdout bootstrap	0.7468	0.7954	+0.0486	95% CI 0.0457-0.0514; 100% positive

The B1 improvement survives every validation check: the five-fold mean gain is 0.0483 with all fold differences positive (one-sided Wilcoxon p=0.03125), the 2018 forward gain is 0.0297, and the paired holdout bootstrap gives a 95% B1-B0 AUC difference interval of about 0.0457-0.0514.

4.3 Leakage propagates into policy value

Table 4. Leakage-to-policy-value results at 20% capacity and a 3:1 missed-delay:false-alarm ratio.

Boundary	Precision @K	Recall late	Policy cost	Reduction	NDR
B0	0.919	0.335	40072	32.53%	0.108
B1	0.952	0.347	39128	34.11%	0.065
B2	1.000	0.365	37728	36.47%	0.000
B3	1.000	0.365	37728	36.47%	0.000
B4	1.000	0.365	37728	36.47%	0.000

At identical 20% workload, B0 reduces the no-action cost by 32.53% and B1 by 34.11%. Invalid B2-B4 reach 36.47% with NDR = 0 because all selected orders are late, reproducing the same-capacity outcome-aware oracle. The 3.95 percentage-point apparent gain from B0 to the invalid boundary is therefore policy-value leakage rather than deployable business value.

4.4 Capacity, calibration and model-family checks

Capacity changes the trade-off: under the stored 2:1 sensitivity, B1 precision falls from 1.000 at 10% capacity to 0.761 at 50%, while late-order recall rises from 0.182 to 0.694. Fixed K therefore keeps workload explicit and prevents high missed-delay costs from degenerating into an 'intervene on almost everyone' threshold policy.

Calibration is tested rather than assumed to help: the raw B0 CatBoost Brier score is about 0.193, while post-hoc calibration is neutral or worse in the stored audit. The valid B1 enrichment also generalizes across nonlinear models: random forest, XGBoost, LightGBM, and CatBoost gain about 0.045-0.054 ROC-AUC, whereas logistic regression is essentially unchanged. This supports an information-boundary effect without claiming one boosting package is universally best.

5. DISCUSSION

The findings change how late-delivery model quality should be interpreted. A moderate ROC-AUC from a strict prospective feature set can be more operationally credible than near-perfect discrimination obtained after the event timeline has been crossed. B1 also shows that leakage control need not imply feature minimalism: legitimate order-time enrichment improves discrimination across resampling, time, and nonlinear model families.

The policy experiment adds an economic reason for temporal discipline. Under finite capacity, ranking determines which orders consume scarce attention; leaked outcome-proximal variables can therefore improve precision@K and drive regret toward zero, creating a falsely compelling allocation case. Leakage-to-policy-value auditing should accompany predictive metrics whenever scores feed prioritisation, triage, or exception management.

The study does not show that expediting prevents lateness. Causal effectiveness requires treatment assignment and counterfactual evidence [10]. The present policy is a prioritisation/review abstraction, so a live programme would still need intervention cost, efficacy, operational constraints, and side effects.

5.1 Governance and design implications

A credible delivery-risk DSS should expose the decision timestamp, a feature-availability contract, a validated risk estimate, and the workload/cost rule that converts scores into actions. Each feature should be tagged with the earliest event at which it becomes reliable; monitoring should track discrimination, calibration, selection rate, and realised precision among reviewed orders. The sequence is: validate the information boundary first, then calibrate and optimize, and only then estimate causal intervention effects where required.

5.2 Limitations and external validity

DataCo is a public benchmark rather than a prospectively instrumented enterprise intervention study; the 2018 forward set is small, category handling may differ in production, and the top-K policy assumes equal review capacity with normalized costs. External validation should use event-time-stamped enterprise data, a pre-registered feature-availability policy, heterogeneous intervention costs, and observed outcomes. A natural extension is to combine leakage-safe prediction with causal treatment-effect estimation and constrained allocation.

6. CONCLUSION

Decision-time leakage is both a predictive-validity and a policy-value problem. On 180,519 DataCo orders, valid order-time enrichment raises CatBoost ROC-AUC from 0.7468 to 0.7954 and remains positive under cross-validation, a forward test, and paired bootstrap. Adding dispatch or outcome information raises ROC-AUC to 0.9965-1.0000 but is invalid for pre-shipment action; at 20% capacity the same leakage makes the ranking appear oracle-optimal and inflates apparent cost reduction. The paper's contribution is a reproducible audit that propagates feature timing through predictive metrics into finite-capacity decision value. For supply-chain AI, 'when was this feature known?' must be answered before interpreting accuracy or policy profitability.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING

No external funding is reported for this study.

DATA AVAILABILITY

The DataCo Smart Supply Chain dataset is publicly distributed through Mendeley Data. The audit used the supplied DataCoSupplyChainDataset.csv file (SHA-256: fa6d022ed437155e1a2f0378710602848703c8a7f203f7ff5d77805bf8480aa6). Reproduction scripts, boundary definitions and holdout predictions are retained in the study reproducibility package and can accompany the journal submission or a public repository deposit.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Meenakshi Yadav: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft. Gurjeet Singh: Supervision, Methodology, Validation, Writing - review and editing.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During preparation of this manuscript, the authors used OpenAI ChatGPT for language drafting, structural editing and formatting support. The authors reviewed and verified the analyses, numerical results, citations and conclusions and take full responsibility for the content of the manuscript.

7. REFERENCES

[1] Baryannis, G., Dani, S., & Antoniou, G. (2019). Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. Future Generation Computer Systems, 101, 993-1004. https://doi.org/10.1016/j.future.2019.07.059

[2] Bassiouni, M.M., Chakrabortty, R.K., Sallam, K.M., & Hussain, O.K. (2024). Deep learning approaches to identify order status in a complex supply chain. Expert Systems with Applications, 250, 123947. https://doi.org/10.1016/j.eswa.2024.123947

[3] Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3.

[4] Constante, F., Silva, F., & Pereira, A. (2019). DataCo Smart Supply Chain for Big Data Analysis. Mendeley Data.

[5] Elkan, C. (2001). The foundations of cost-sensitive learning. Proceedings of the 17th International Joint Conference on Artificial Intelligence, 973-978.

[6] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

[7] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery

from Data, 6(4), Article 15.

[8] Mandi, J., Kotary, J., Berden, S., Mulamba, M., Bucarey, V., Guns, T., & Fioretto, F. (2024). Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. Journal of Artificial Intelligence Research, 80, 1623-1701.

[9] Müller, N., Burggraef, P., Steinberg, F., Sauer, R., & Schütz, M. (2025). An analytical review of predictive methods for delivery delays in supply chains. Supply Chain Analytics, 11, 100130. https://doi.org/10.1016/j.sca.2025.100130.

[10] Wyrembek, M., Baryannis, G., & Brintrup, A. (2025). Causal machine learning for supply chain risk prediction and intervention planning. International Journal of Production Research, 63(15), 5629-5648. https://doi.org/10.1080/00207543.2025.2458121