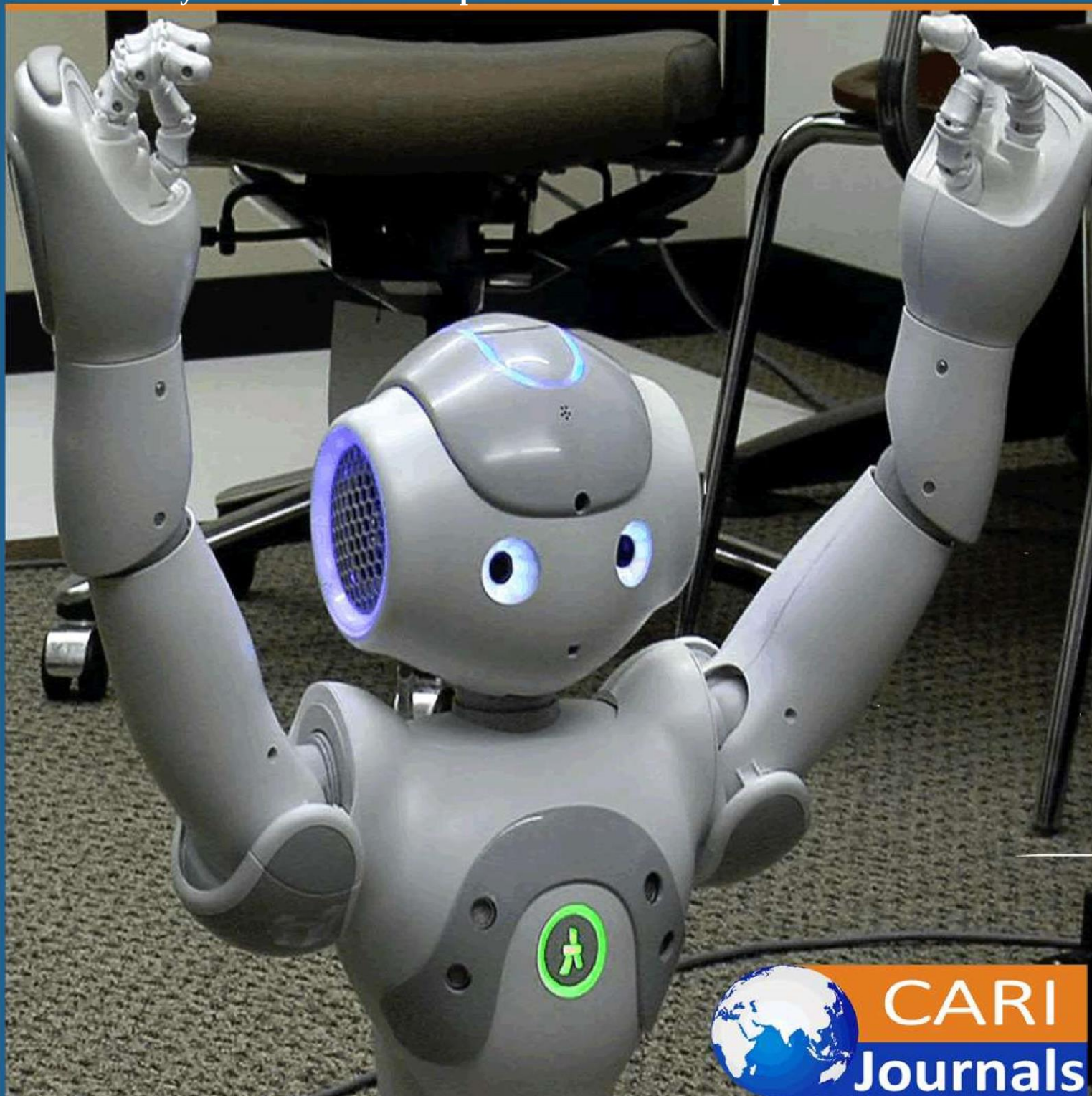
International Journal of
Computing and
Engineering
(IJCE)

Predicting Departure Delays in Time Critical Transportation: A Study of E-Commerce Dispatch and School Bus Operations



CARI Journals

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 8, Issue No. 4, pp. 1 - 31, 2026
www.carijournals.org

Predicting Departure Delays in Time Critical Transportation: A Study of E-Commerce Dispatch and School Bus Operations

Aditya Kumar Sharma
University of Washington, United States
https://orcid.org/0009-0006-0810-9395

Accepted: 4th Oct, 2026, Received in Revised Form: 20th Oct, 2026, Published: 3rd Nov, 2026

ABSTRACT

Purpose: This synthetic simulation study verifies a protocol for evaluating departure-delay prediction in stylized parcel dispatch and school bus operations. It tests whether information available before scheduled departure can rank lateness risk and support alerts within limited human review capacity.

Methodology: The benchmark retains the prior five-seed experiment and adds twenty-seed Monte Carlo analysis, paired model comparisons, formal calibration measures, fixed-workload alert curves, delayed and missing observations, and one-factor sensitivity analyses. Each primary seed generates 9,000 stylized parcels and 9,000 stylized school-bus departures across 300 abstract operating days, with snapshots at 90, 60, and 30 minutes before schedule. The generator creates 90,000 distinct departures in the legacy five-seed analysis and 360,000 in the expanded analysis. No Chewy, carrier, school district, driver, employee, customer, student, route, or shipment records are used.

Findings: In the shifted synthetic test period, late-departure prevalence was 19.1% (parcel) and 12.8% (school). At 60 minutes, the multivariable logistic model achieved mean average precision of 0.741 and 0.580, compared with 0.518 and 0.487 for the slack-only model; mean Brier scores were 0.088 and 0.079. Ranking was similar to one fixed gradient-boosting configuration. The shift worsened calibration and increased alert burden. At 10% review capacity, the model captured 43.8% and 48.1% of late parcel and school departures, respectively.

Unique Contribution to Theory, Practice and Policy: The study formalizes departure readiness, information timing, calibration, and human review capacity within one evaluation protocol. It gives practitioners reproducible tests for alert performance and offers policymakers criteria for local validation and safety review. The synthetic results verify protocol behavior under the specified generator but do not establish real-world effectiveness or deployment readiness.

Keywords: *Departure Delay Prediction, E-Commerce Logistics, School Bus Operations, Explainable Machine Learning, Synthetic Simulation*

1 INTRODUCTION

Transportation schedules create commitments that are simple to state and difficult to protect under variability. A movement can leave on time only when the work, vehicle, personnel, information, and release conditions needed for that movement converge sufficiently early. A late departure may originate in an upstream workload interruption, an unavailable vehicle, an unresolved exception, a delayed handoff, a data feed that is not current, or a response that arrives after the remaining recovery window. A conventional end of shift report records the failure after much of the opportunity to act has passed. The scientific question in this paper is narrower: can pre departure information be used to estimate departure delay risk at several fixed warning horizons?

The first application setting is e commerce dispatch. Fulfillment work must connect to scheduled transportation capacity. Orders can be picked, packed, and available in a facility while still missing the transportation connection that moves them toward the customer. Shipment cutoff terminology is not uniform across organizations: a cutoff may describe a shipment tender, trailer close, carrier handoff, or vehicle departure. A credible study must define the event instead of treating all of these milestones as interchangeable. Research on cutoff based shipment promises and real time batch dispatch shows why fulfillment timing and transportation deadlines should be considered together (Chewy Inc., 2026; Gautam & Geunes, 2024; Mohring et al., 2024).

Pet product distribution provides a consequential service context but not a source of data for this paper. Chewy publicly describes a network that uses national, regional, and local logistics providers and identifies shipping performance, delivery experience, and shipping expense as material business considerations (Chewy Inc., 2026). The second setting is school bus operations. An assigned route requires an appropriate vehicle, an available and qualified driver, completion of required checks, and sufficient preparation time before the approved route can begin. School transportation research emphasizes route construction, time windows, vehicle use, stochastic travel time, and on time arrival reliability (Babaei & Rajabi Bahaabadi, 2019; Park & Kim, 2010; Sun et al., 2018). A departure warning system would complement, rather than replace, route planning and safety controls. A depot departure model cannot by itself establish that a student arrives on time, that a stop is served safely, or that maximum ride time and accessibility requirements are satisfied.

This revision responds without replacing the original experiment. It retains the original default generator, original five seed results, original 90, 60, and 30 minute horizons, original feature representation, original transfer analysis, original forecast noise stress test, and original equations. It adds a twenty seed expansion, paired replicate differences, calibration intercept and slope, expected and maximum calibration error, log loss, complete horizon reporting, fixed capacity alert curves, explicit observation degradation, one factor sensitivity diagnostics, coefficient tables, worked score decompositions, and automated integrity checks. It also makes the manuscript's evidence boundary explicit in the title, abstract, methods, results, and conclusion.

The paper has four objectives. The first is to state an outcome specific, time aware framework that can be evaluated in both domains without conflating departure with arrival or delivery. The second

is to test whether a multivariable readiness representation contains signal beyond a single schedule slack summary under an author specified generator. The third is to show how calibration, distribution shift, data quality, and limited attention alter the operational meaning of an alert score. The fourth is to identify which elements can responsibly transfer between parcel and school transportation and which require local data and governance.

2 LITERATURE AND RESEARCH POSITIONING

2.1 Fulfillment Deadlines and Dispatch Decisions

Mohring et al. study cutoff based shipment promises for stochastic warehouse fulfillment and show why a promised service level depends on the relationship between processing uncertainty and a fixed deadline (Mohring et al., 2024). Gautam and Geunes analyze real time fulfillment policies and when a dispatch batch should be released (Gautam & Geunes, 2024). These studies focus on promise design and dispatch control. The present study conditions on a scheduled departure and asks whether its delay risk can be estimated before the event using a small operational feature set. It therefore sits between upstream fulfillment control and downstream transportation execution.

The distinction limits the claims that can be made. The framework does not optimize carrier routes, redesign a cutoff, select a trailer, or determine a shipment promise. It produces a risk estimate for an already defined departure boundary. A future study could connect the score to a constrained scheduling or recovery optimizer, but that would be a new decision problem requiring action feasibility, resource competition, and causal evaluation.

2.2 School Bus Routing and Uncertainty

Park and Kim review the school bus routing problem and describe its connections to stop assignment, fleet use, scheduling, and operational constraints (Park & Kim, 2010). Sun et al. represent stochastic and time dependent transportation conditions in school bus routing (Sun et al., 2018). Babaei and Rajabi Bahaabadi explicitly consider stochastic travel times and on time arrival reliability (Babaei & Rajabi Bahaabadi, 2019). These works support the importance of time windows and uncertainty in student transportation but do not imply that a parcel readiness model is a complete school bus solution.

The proposed school application concerns departure readiness at a depot or another consistently defined route origin. It does not determine a safe route, assign a student to a vehicle, authorize a substitute driver, alter an approved stop, or trade away accessibility or maximum ride time to improve punctuality. Arrival reliability requires route geography, traffic, stop sequence, boarding and alighting, and downstream schedule slack. Those outcomes should be modeled separately or linked through a validated route level model.

2.3 Explainable and Dynamic Transportation Prediction

Transportation prediction literature includes flight delay prediction using heterogeneous information and explicit explanation analysis (Pineda Jaramillo et al., 2024), as well as bus delay

propagation work that emphasizes dependencies along a route (Zhang et al., 2025). These studies illustrate that feature attribution and causal explanation are not interchangeable. The present paper uses logistic regression as a transparent model whose log odds decomposition can be reconstructed exactly within the model space. It does not call the coefficients causal effects.

Rudin argues that interpretable models should be considered directly for high stakes decisions rather than relying automatically on post hoc explanations of black boxes (Rudin, 2019). Gradient boosting is included as a nonlinear comparator following the general boosting formulation of Friedman (Friedman, 2001). SHAP is discussed as a possible future attribution tool (Lundberg & Lee, 2017), but SHAP explanations are not computed here. The revision therefore uses the narrower claim explainable by construction for the logistic model and demonstrates it with coefficients and worked synthetic alerts.

Dynamic prediction also raises the problem of information availability. A predictor is not time aware merely because a timestamp appears in a table. The value must have been available to the decision process before the prediction time. Leakage reviews emphasize that backdated or post outcome variables can make machine learning studies appear stronger than they are (Kapoor & Narayanan, 2023). The revised simulation adds a separate observation layer to test the effect of missing, stale, and delayed values, while retaining a clear statement that the layer is illustrative rather than an operational data architecture.

2.4 Probability Calibration, Shift, and Decision Value

Niculescu Mizil and Caruana distinguish ranking performance from the quality of estimated probabilities (Niculescu-Mizil & Caruana, 2005). Kull et al. develop beta calibration as an alternative to logistic calibration (Kull et al., 2017), while Guo et al. demonstrate that accurate classifiers can still be poorly calibrated (Guo et al., 2017). Van Calster et al. emphasize calibration in the large, calibration slope, and flexible calibration curves when a model is externally evaluated (Van Calster et al., 2019). The revised paper reports numerical calibration intercept, slope, expected calibration error, maximum calibration error, Brier score, and log loss, alongside reliability curves.

Concept drift and domain adaptation work explain why a model that performs in one period or domain can degrade when the data distribution or feature outcome relationship changes (Ben-David et al., 2010; Gama et al., 2014). Ovadia et al. evaluate uncertainty under dataset shift (Ovadia et al., 2019). The benchmark's deliberate shift is not a realistic forecast of seasonality; it is a controlled demonstration of why a frozen threshold can generate more alerts and less reliable probabilities after conditions change.

A model can rank risk well and still be operationally unusable if the alert queue exceeds available attention. Decision curve analysis formalizes the idea that the value of a prediction depends on the relative consequences of true and false actions (Vickers & Elkin, 2006). This paper uses fixed workload evaluation as the primary operationally readable analysis and retains a conceptual

expected utility equation for future intervention studies. It does not claim that the chosen 5, 10, 15, or 20 percent capacities represent any real dispatch team.

2.5 Human Attention and Simulation Verification

Repeated or low value alerts can overload decision makers. Ancker et al. show that workload, complexity, and repeated alerts affect alert response in a clinical decision support system (Ancker et al., 2017). The domain differs from transportation, but the human factors principle is relevant: false alerts and repeated notifications are part of system performance, not merely user inconvenience. The revised analysis therefore collapses repeated snapshots to event level when reporting all horizon alert burden and reports fixed capacity curves.

Simulation studies also require verification of code and careful statements about validity. Sargent distinguishes verification of implementation from validation of whether a model is an adequate representation for its intended use (Sargent, 2013). The present work performs software integrity checks, preserves the generator formulas, records environment versions, and reports sensitivity results. It does not claim external validation because no observed operational data were used.

3 OPERATIONAL SCOPE AND PREDICTION TARGETS

3.1 Departure as the Unit of Analysis

A departure event is one planned movement associated with a scheduled time, a defined operational boundary, and an actual time at which that boundary is crossed. In parcel dispatch, the boundary could be a confirmed gate exit, a trailer departure, a carrier tender, or another event selected before label creation. In school transportation, it could be depot exit for the start of an assigned route. The boundary must be consistent within the evaluation. A completed pick, trailer close, driver check in, carrier arrival, vehicle movement, and final delivery are related events, not interchangeable labels.

All timestamps in a field study should be stored with a time zone or UTC offset and an availability timestamp. If a schedule is displayed in local time, daylight saving transitions and ambiguous repeated times must be resolved before modeling. Schedule revisions require the schedule version known at the prediction time, its effective timestamp, and the original commitment if that commitment is part of the service definition. Repeatedly moving a deadline after an operation is at risk can make failure look like punctuality.

The synthetic benchmark normalizes scheduled departure to time zero and uses a five minute lateness tolerance. This is an illustrative sensitivity point, not a carrier agreement, Chewy standard, school district policy, or safety rule.

$$t_{i,h} = s_i - h, \qquad Y_i = 1\{a_i > s_i + \delta_d\}, p_{i,h,d} = Pr\left(Y_i = 1 \mid F_{t_{i,h}}, d\right). \qquad (1)$$

3.2 Distinct Service Outcomes

Table 1 distinguishes the operational outcomes that the two applications must not be conflated.

Table 1: Distinct Transportation Outcomes

Outcome	Parcel application	School application
Prerequisite readiness	Departure specific orders or required work ready for a defined handoff	Vehicle, driver, and approved preparation ready for release
Departure lateness	Vehicle or defined dispatch boundary crossed later than scheduled	Bus or defined route origin crossed later than scheduled
Intermediate performance	Tender, transfer, or downstream connection achieved	Pickup and subsequent route commitments achieved
Recipient service	Delivery against the promised service window	School arrival and completion of approved student service

The synthetic endpoint is departure lateness only. The other rows require separate observations.

An on time parcel departure can still lead to a late delivery, and a late departure can sometimes be recovered downstream. A bus leaving its depot on time can still encounter congestion, extended boarding, or a route disruption. Conversely, a modest departure delay need not produce a late school arrival when the schedule contains slack. The experiment measures departure lateness only.

3.3 Incomplete, Canceled, and Reassigned Events

Canceled movements, unobserved actual departures, duplicate movements, and reassigned vehicles or drivers must not be automatically labeled on time or dropped without accounting. A field study could use a separate cancellation outcome, a composite service failure endpoint, inverse probability weighting, or a time to event formulation. The benchmark assumes that every synthetic event eventually departs and therefore does not estimate censoring or cancellation bias.

4 FRAMEWORK AND INFORMATION ARCHITECTURE

4.1 From Latent State to Operational Decision

The revised framework has six layers. First, a latent operational state represents readiness and disruptions. Second, operational events and forecasts are generated from that state. Third, an observation layer determines what is delayed, stale, missing, or timestamp noisy by the prediction time. Fourth, feature construction derives component readiness forecasts, slack, incident indicators, pressure, and freshness measures without using future information. Fifth, a calibrated risk model scores the departure. Sixth, a human decision process determines whether an alert is acknowledged and whether a permitted action is taken. The benchmark evaluates the first five layers and reports alert workload; it does not claim that an intervention improved an outcome.

This separation is important for both domains. A machine learning model can be mathematically correct and still fail if a scan arrives too late, a driver check in is missing, an event definition changes, or the alert queue exceeds the responsible person's capacity. A production system should retain event time, availability time, source, version, missingness, model version, threshold, alert status, acknowledgment, action, and final outcome.

4.2 Readiness Components and Schedule Slack

The shared representation uses forecasts of when critical prerequisites will be ready. For an e commerce departure, upstream work can summarize picking, packing, staging, labeling, documentation, and other departure specific requirements. For a school bus, preparation can summarize vehicle checks, fueling or charging readiness, required equipment, and other approved pre departure steps. Vehicle readiness and personnel readiness remain separate. A release forecast represents the remaining activities between the latest prerequisite and the defined departure boundary.

$$\hat{u}_i(t) = max\{\hat{R}_i(t), \hat{V}_i(t), \hat{C}_i(t)\} + \hat{L}_i(t), \quad S_i(t) = s_i - \hat{u}_i(t). \qquad (2)$$

Positive slack in the formulation indicates a predicted margin before the scheduled departure; negative slack indicates that predicted feasible completion extends beyond the schedule. The maximum assumes that prerequisites must all be ready before release. Real operations may permit overlapping work, parallel teams, or a sequence of dependent tasks, so the field implementation must use the actual process graph.

Table 2 specifies candidate field-study features, their units, and their application-specific meanings.

Table 2: Feature Specification for a Field Implementation

Feature	Unit and construction	Parcel meaning	School meaning
Work forecast	Minutes relative to scheduled departure	Forecast completion of departure specific fulfillment work	Forecast completion of approved trip preparation
Vehicle forecast	Minutes relative to schedule	Forecast availability of assigned transport capacity	Forecast availability of a suitable assigned bus
Personnel forecast	Minutes relative to schedule	Forecast readiness of required release personnel	Forecast readiness of an assigned qualified driver
Release forecast	Nonnegative minutes	Remaining final loading and release activity	Remaining final dispatch preparation

Feature	Unit and construction	Parcel meaning	School meaning
Slack	Scheduled time minus predicted feasible time	Margin to carrier or dispatch departure	Margin to route origin departure
Incident	Verified binary status at prediction time	Documented exception affecting departure	Documented trip or depot exception
Pressure	Defined site or depot workload index	Shared facility congestion or workload	Shared depot or resource pressure
Information age	Minutes since relevant update	Freshness of readiness information	Freshness of readiness information
Missingness and delay	Indicator and minutes	Feed availability and stale event status	Feed availability and stale event status

The benchmark uses stylized numeric counterparts. These are candidate fields, not claims about available Chewy or school systems.

Workload to readiness conversion requires its own validation. Remaining units, effective processing rates, forecast interruption time, and queue conditions may be used only if they are available at the prediction time. A rate calculated from activity after the target departure is leakage. A rate estimated from the preceding window is available but can be unstable when order mix, staffing, equipment, traffic, or weather changes.

4.3 Explicit Observation and Data Quality Layer

The original benchmark supplied noisy forecasts directly. The revision adds a controlled observation layer to make the information boundary concrete. A forecast can be available but stale, delayed by a feed, missing, or accompanied by timestamp jitter. Missingness remains visible through indicators and the model uses training period median imputation within a pipeline. The layer also adds a freshness penalty and a stale count. These choices are not estimates of a Chewy or school system; they are stress parameters that allow the paper to test whether the framework survives imperfect access.

In a field study, missingness must be interpreted rather than simply imputed. A missing driver check in can mean that a driver is absent, the feed is delayed, the device is offline, or the event definition is inconsistent. A missing carrier scan can have a similar ambiguity. The system should distinguish unknown from verified negative status and should route critical uncertainty for human review. Duplicate events, corrected timestamps, clock drift, outages, schedule changes, and late arriving messages require explicit audit fields.

4.4 Feature Construction and Exclusion Audit

The feature audit should record a business definition, unit, source owner, update frequency, event time, availability delay, missing value rule, allowed range, transformation, and known relationship to the outcome. Final completed volume, post departure reason codes, actual carrier arrival recorded later, end of shift labor totals, final delivery status, and any action taken after the prediction are excluded unless a real system proves they were available in time. Personally identifying customer or student fields are not required for the core departure risk task.

5 PREDICTION, CALIBRATION, AND DECISION EVALUATION

5.1 Calibration and Probability Quality

The calibration partition is distinct from the model-fitting and test partitions. Calibration in the large is estimated as the intercept offset with the calibration slope fixed at one. The calibration slope is estimated by regressing the outcome on the logit of the predicted probability; ideal values are zero for the intercept and one for the slope. Brier score and log loss measure overall probability error. For continuity with the original analysis, expected calibration error is the probability-weighted absolute difference between mean prediction and observed event frequency across ten fixed probability bins, and maximum calibration error is the largest such difference.

Because fixed probability bins can contain few late events at the 12.8% and 19.1% test prevalence, ECE and MCE are treated as secondary summaries. Calibration is interpreted primarily from reliability plots using ten equal-frequency bins, which place similar numbers of predictions in each bin, and a smooth calibration curve that avoids dependence on bin boundaries. Curves are reported by domain and model with variation across replicates. All calibration results are descriptive under the specified synthetic generator and shifted test period; they do not establish future calibration or the causal effect of an untested intervention.

$$p_i^{cal} = \frac{1}{1+e^{-(\alpha+\gamma\eta_i)}}, logit\left(p_i^{cal}\right) = \alpha + \gamma\beta_0 + \sum_{j=1}^{p} \gamma\beta_j z_{ij}. \qquad (4)$$

5.2 Explainability by Model Decomposition

For a calibrated logistic model, each standardized feature contributes its calibrated coefficient multiplied by its standardized value to the calibrated log odds. The contribution is exact within the fitted model space. It is not an additive percentage point change in risk, and it is not a causal effect. Slack and its component forecasts are correlated by construction. A field explanation should therefore show the underlying forecast times, source freshness, and prerequisite margins together with the grouped contributions.

The revision reports coefficient intervals across seeds and three worked synthetic examples for each domain. Two examples are high score alerts that are late in the generated test record, and one is a high score false alert. The false alert is included because an explanation must support review

of failure modes rather than display only successful cases. Dispatcher comprehension, explanation stability, and agreement between model reasons and actual operational causes remain untested.

5.3 Alerts, Workload, and Repeated Horizons

The legacy threshold is the 80th percentile of calibration probabilities for each model, domain, and horizon. It produces approximately 20 percent alerts during calibration but is not a hard capacity rule after shift. The test alert rate is reported rather than hidden. Fixed workload analysis selects the highest scoring 5, 10, 15, or 20 percent of test departures and reports recall, precision, and false alerts. This evaluates a capacity constraint without using test labels to choose who is alerted.

A departure can receive alerts at 90, 60, and 30 minutes. Snapshot level counts treat those observations as correlated because they share one outcome. For all horizon coverage and burden, alerts are collapsed to one event level: the event is counted once if any horizon alerts, and the earliest nominal warning is the maximum horizon that alerted. Mean alerts per event and mean alerts per alerted event are also reported.

5.4 Conceptual Intervention Value

Prediction is not intervention. A score can prioritize attention, but it does not specify which recovery action is feasible or whether an action changes the outcome. A future decision layer can compare expected avoidable loss with action cost and limited resource capacity.

$$U_i = p_i q_i L_i - c_i, \qquad \max_{u_i \in \{0,1\}} \sum_i u_i U_i \quad s.t. \quad \sum_i u_i r_i \le B. \qquad (5)$$

In this paper the utility expression remains a decision theoretic specification, not an estimated result. The experiment does not assign actions, estimate intervention effectiveness, model harm to another departure, or report savings. The appropriate next study is a human in the loop or controlled shadow and intervention evaluation with approved actions, response time, shared resource constraints, safety rules, and action logging.

6 SYNTHETIC STUDY DESIGN

6.1 Evidence Boundary and Simulation Purpose

The study is an executed statistical simulation. Domain names identify stylized parameterizations and not observed organizations. The generator is not a calibrated digital twin, discrete event recreation of a fulfillment center, or validated school bus depot simulator. Its purpose is to test whether the code, metrics, and evaluation protocol behave as intended under a transparent set of assumptions.

6.2 Baseline Data Generating Mechanism

Scheduled departure is normalized to time zero. Three latent readiness times represent upstream work, vehicle availability, and personnel readiness. A nonnegative release duration represents final

activity after the latest prerequisite. A known incident contributes to vehicle availability and additional release delay. An unobserved shock contributes a positive residual. The actual departure is constrained to be nonnegative, and the late outcome is actual departure greater than the five minute tolerance.

$$a_i = max\{0, max(R_i, V_i, C_i) + L_i + \kappa_d I_i + \varepsilon_i\}, \qquad Y_i = 1\{a_i > 5\}. \qquad (6)$$

The parcel parameterization has greater upstream work variability and a longer release duration. The school parameterization has relatively more personnel variability and a shorter release duration. These choices create different illustrative dependency structures and are not estimates. Forecasts are noisy measurements of latent readiness. Their error decreases at shorter horizons and includes an event shared component across horizons, creating correlated snapshots. The generator therefore builds in the directional mechanism behind H2; the result is a check of the implementation, not a discovery about real operations.

After day 220, known incident probability increases from 0.08 to 0.16 in the default configuration, the unobserved shock probability increases from 0.04 to 0.09, parcel work readiness shifts three minutes later, and school personnel readiness shifts three minutes later. The shift is deliberate and its magnitude is varied in sensitivity analysis. It is not a measured seasonal or company effect.

Table A1 lists the author-specified generator parameters used in the synthetic benchmark.

Table A1: Author-Specified Generator Parameters

Quantity	Parcel parameterization	School parameterization	Purpose
Work readiness R	−37 + 12z1 + 6P + 3J	−32 + 6z1 + 2P	Different upstream variability
Vehicle readiness V	−31 + 9z2 + 3P + 14I	−27 + 8z2 + 4P + 12I	Incident sensitivity
Personnel readiness C	−37 + 7z3 + 2P	−25 + 10z3 + 3P + 3J	Different personnel variability
Release duration L	max(3, 15 + 3z4 + 2max(P,0))	max(2, 8 + 2z4 + max(P,0))	Final release work
Additional incident effect	9I	5I	Known disruption consequence
Default late tolerance	5 minutes	5 minutes	Illustrative label definition
Default shift	Day 221 onward	Day 221 onward	Controlled test distribution shift

All quantities are illustrative assumptions, not estimates of real operations. z variables are standard normal; P is operating pressure; I is a known incident; J is the shift indicator.

The complete formulas, random seed handling, and observation layer are in the executable supplement. The generator preserves event level common pressure and horizon correlated forecast errors. The supplement also labels all generated files as synthetic.

6.3 Partitions, Sample Sizes, and Software

Table 3 summarizes the executed primary experiment settings and partition sizes.

Table 3: Executed Primary Experiment Settings

Setting	Specification
Legacy and expanded seeds	Legacy 20261001–20261005; expanded 20261001–20261020
Events	9,000 parcel and 9,000 school events per seed; 360,000 distinct departures expanded
Prediction horizons	90, 60, and 30 minutes before scheduled departure
Training calibration testing	Days 1–180; 181–220; 221–300; 5,400 / 1,200 / 2,400 events per domain, seed, horizon
Positive outcome	Actual departure more than five minutes after schedule
Models	Prevalence; slack only logistic; multivariable logistic; fixed gradient boosting
Calibration	Separate one dimensional logistic map on the calibration partition
Alert threshold	80th percentile of calibration probabilities, frozen for test
Uncertainty	Seed mean, sample SD, and 2.5th–97.5th seed percentile interval

All primary inputs and outputs are synthetic. The test prediction export contains 288,000 snapshots in the expanded analysis.

The legacy five seed analysis contains 90,000 distinct synthetic departures and 270,000 snapshots. The expanded twenty seed analysis contains 360,000 distinct departures and 1,080,000 snapshots before restricting to the test prediction export. For each domain, seed, and horizon, there are 5,400 training events, 1,200 calibration events, and 2,400 test events. No event crosses partitions.

6.4 Sensitivity Design

One factor at a time diagnostics vary lateness tolerance at 0, 5, and 10 minutes; shift strength at 0, 1, and 1.5 times the default; forecast noise at 0.75, 1, and 1.25 times the default; forecast error correlation at 0, 0.65, and 0.9; release time variability at 0.5, 1, and 1.5; operating day pressure at 0, 1, and 1.5; and missingness at 0, 0.10, and 0.25 within the observation layer. These levels are stress values, not estimated confidence ranges. The grid uses three seeds and fewer events per day

to keep the diagnostic computation tractable. It is therefore a robustness screen rather than a global sensitivity analysis.

The observation scenario uses ten percent missingness, a three minute mean feed delay, ten percent stale values, and one minute timestamp jitter. It is compared with clean feature access at the 60 minute horizon. Forecast error stress also retains the prior first seed analysis in which 0, 5, or 10 additional minutes of Gaussian error are added to the three readiness forecasts at test time while the model remains frozen.

6.5 Statistical Analysis

The expanded analysis reports mean and sample standard deviation across 20 seeds, together with 2.5th and 97.5th seed percentile intervals. These are Monte Carlo intervals over the chosen synthetic seeds, not confidence intervals about a real transportation population. Paired differences compare models within the same seed, domain, and horizon. The five seed analysis is retained as a reconciliation record and should be read as a minimal demonstration, not robust uncertainty quantification. Horizon specific measures use one row per event. All horizon measures are collapsed to event level before calculating alert burden.

$$BS = \frac{1}{n}\sum_{i=1}^{n} (p_i - Y_i)^2. \qquad (7)$$

7 RESULTS OF THE SYNTHETIC STUDY

7.1 Event Prevalence and Evidence Boundary

The expanded test late event prevalence averages 19.1 percent in the parcel parameterization and 12.8 percent in the school parameterization. These values are consequences of the selected latent distributions and post day 220 shift. They are not operational benchmarks. The more important evidence boundary is that every displayed result is a recovery result under the generator. A model cannot recover a relationship that was not supplied, and a favorable result cannot confirm that the supplied relationship exists in practice.

7.2 Primary 60 Minute Comparison

Table 4 reports the expanded synthetic test performance at the primary 60-minute horizon.

Table 4: Expanded Synthetic Test Performance at 60 Minutes

Domain	Model	AP mean [seed interval]	ROC AUC	Brier	Log loss	ECE	Recall %	Precision %	Alert %	False alerts per 100
Parcel	Prevalence	0.191 [0.176, 0.202]	0.500 [0.500, 0.500]	0.160 [0.148, 0.168]	0.510 [0.477, 0.531]	0.073 [0.049, 0.087]	—	—	—	—
Parcel	Slack only	0.518 [0.470, 0.569]	0.796 [0.781, 0.812]	0.127 [0.120, 0.133]	0.408 [0.387, 0.428]	0.056 [0.041, 0.071]	58.3	45.4	24.6	13.4
Parcel	Multivariable logistic	0.741 [0.715, 0.767]	0.892 [0.883, 0.906]	0.088 [0.083, 0.093]	0.300 [0.282, 0.312]	0.028 [0.017, 0.040]	79.9	53.9	28.4	13.1
Parcel	Fixed boosting	0.735 [0.708, 0.763]	0.889 [0.880, 0.904]	0.089 [0.083, 0.093]	0.303 [0.281, 0.317]	0.029 [0.020, 0.038]	79.2	53.6	28.3	13.2
School	Prevalence	0.128 [0.115, 0.148]	0.500 [0.500, 0.500]	0.116 [0.104, 0.133]	0.410 [0.373, 0.464]	0.064 [0.050, 0.084]	—	—	—	—
School	Slack only	0.487 [0.439, 0.525]	0.828 [0.812, 0.846]	0.090 [0.081, 0.104]	0.310 [0.279, 0.355]	0.046 [0.025, 0.071]	71.0	34.0	26.8	17.7
School	Multivariable logistic	0.580 [0.526, 0.638]	0.861 [0.838, 0.876]	0.079 [0.072, 0.091]	0.275 [0.250, 0.308]	0.031 [0.017, 0.056]	80.2	34.0	30.3	20.0
School	Fixed boosting	0.561 [0.508, 0.624]	0.855 [0.836, 0.869]	0.081 [0.074, 0.093]	0.281 [0.258, 0.318]	0.035 [0.020, 0.061]	78.8	33.8	30.0	19.9

Intervals are 2.5th–97.5th percentiles across 20 synthetic seeds. They are not confidence intervals for real operations. All models use chronological fitting; calibration and test outcomes remain separated.

At 60 minutes, the multivariable logistic model has higher mean average precision than slack only by 0.223 in parcel and 0.093 in school. Paired seed intervals for these differences remain above zero in the expanded baseline analysis, but that statement applies only to the generator. Brier score is lower for the multivariable model by 0.038 and 0.011, respectively. ROC AUC, log loss, calibration intercept, calibration slope, ECE, and MCE are included in the table and supplement rather than omitted as in the prior version.

The logistic model is close to the fixed boosting comparator. At 60 minutes, the paired average precision difference between logistic and boosting is 0.006 in parcel and 0.019 in school. The replicate interval crosses zero for at least some comparisons, which is consistent with a cautious conclusion: this low dimensional synthetic generator does not establish that one model family is generally superior. A fully tuned comparator and local field baselines remain necessary for a deployment claim.

7.3 Warning Horizon

Table 5 compares model performance across the three warning horizons.

Table 5: Performance by Warning Horizon

Domain	Minutes	Logistic AP	Slack AP	Logistic AUC	Logistic Brier	Logistic log loss	Logistic ECE
Parcel	90	0.701 [0.664, 0.732]	0.418 [0.365, 0.465]	0.875 [0.865, 0.887]	0.096 [0.089, 0.100]	0.321 [0.305, 0.336]	0.030 [0.016, 0.039]
Parcel	60	0.741 [0.715, 0.767]	0.518 [0.470, 0.569]	0.892 [0.883, 0.906]	0.088 [0.083, 0.093]	0.300 [0.282, 0.312]	0.028 [0.017, 0.040]
Parcel	30	0.795 [0.771, 0.819]	0.646 [0.603, 0.688]	0.914 [0.903, 0.930]	0.077 [0.073, 0.082]	0.267 [0.249, 0.287]	0.024 [0.014, 0.032]
School	90	0.513 [0.442, 0.567]	0.382 [0.324, 0.430]	0.837 [0.808, 0.853]	0.085 [0.076, 0.096]	0.295 [0.268, 0.324]	0.035 [0.016, 0.056]
School	60	0.580 [0.526, 0.638]	0.487 [0.439, 0.525]	0.861 [0.838, 0.876]	0.079 [0.072, 0.091]	0.275 [0.250, 0.308]	0.031 [0.017, 0.056]
School	30	0.675 [0.624, 0.713]	0.611 [0.571, 0.664]	0.887 [0.866, 0.905]	0.069 [0.062, 0.079]	0.247 [0.219, 0.275]	0.027 [0.016, 0.050]

Values are means with seed percentile intervals. The generator supplies lower forecast noise at shorter horizons, so the horizon pattern is conditional on that assumption.

Figure 1 visualizes the synthetic ranking pattern across the three warning horizons.

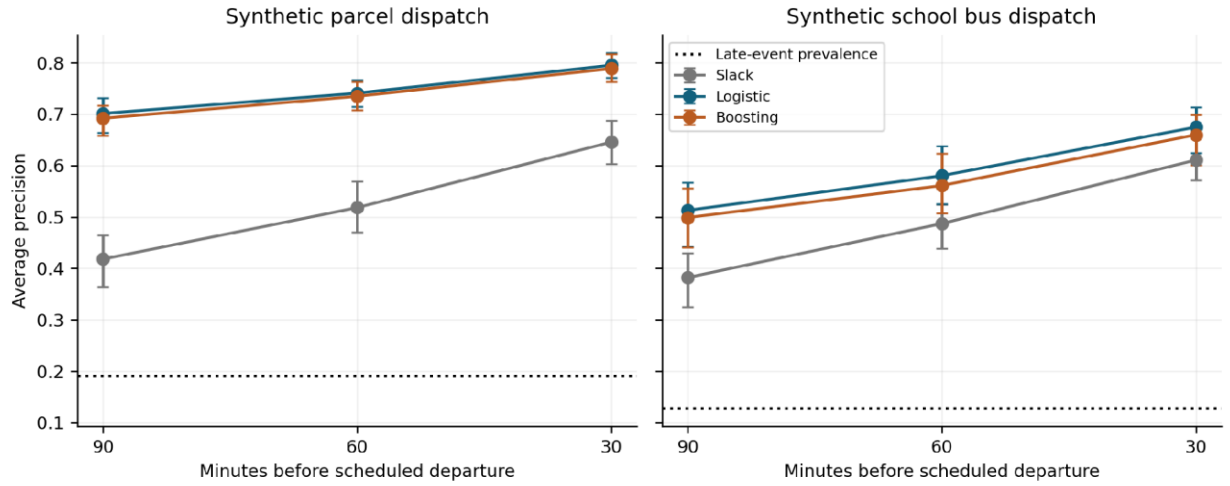


Figure 1: Synthetic Ranking Performance by Warning Horizon. Points show twenty seed means and bars show 2.5th to 97.5th seed intervals. The dotted line shows mean shifted test prevalence. Forecast noise decreases by construction at shorter horizons.

The figure should be read as a generator-conditional comparison of ranking performance and seed variation, not as an estimate of real transportation performance.

7.4 Calibration and Distribution Shift

Table 6 reports calibration diagnostics for the 60-minute predictions.

Table 6: Calibration Diagnostics at 60 Minutes

Domain	Model	Calibration intercept	Calibration slope	ECE	MCE	Brier	Log loss
Parcel	Slack only	0.518 [0.368, 0.665]	0.927 [0.768, 1.111]	0.056 [0.041, 0.071]	0.208 [0.099, 0.445]	0.127 [0.120, 0.133]	0.408 [0.387, 0.428]
Parcel	Multivariable logistic	0.257 [-0.016, 0.429]	0.911 [0.726, 1.100]	0.028 [0.017, 0.040]	0.139 [0.079, 0.210]	0.088 [0.083, 0.093]	0.300 [0.282, 0.312]
Parcel	Fixed boosting	0.274 [0.003, 0.426]	0.921 [0.768, 1.113]	0.029 [0.020, 0.038]	0.119 [0.081, 0.168]	0.089 [0.083, 0.093]	0.303 [0.281, 0.317]
School	Slack only	0.633 [0.277, 1.012]	0.915 [0.802, 1.034]	0.046 [0.025, 0.071]	0.248 [0.097, 0.379]	0.090 [0.081, 0.104]	0.310 [0.279, 0.355]

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 8, Issue No. 4, pp. 1 - 31, 2026
www.carijournals.org

Domain	Model	Calibration intercept	Calibration slope	ECE	MCE	Brier	Log loss
School	Multivariable logistic	0.410 [0.005, 0.824]	0.879 [0.782, 1.017]	0.031 [0.017, 0.056]	0.190 [0.082, 0.338]	0.079 [0.072, 0.091]	0.275 [0.250, 0.308]
School	Fixed boosting	0.471 [0.055, 0.906]	0.894 [0.797, 1.031]	0.035 [0.020, 0.061]	0.180 [0.098, 0.337]	0.081 [0.074, 0.093]	0.281 [0.258, 0.318]

Ideal calibration intercept is 0 and slope is 1. ECE and MCE use ten fixed probability bins. Constant prevalence slope is not identifiable.

The multivariable logistic model's expanded 60 minute ECE averages 0.028 in parcel and 0.031 in school. The seed ranges and MCE values are reported in the table and supplement. The reliability figure is descriptive and contains seed variation; it is not a confidence band for real operations.

Figure 2 visualizes the descriptive calibration behavior of the 60-minute predictions.

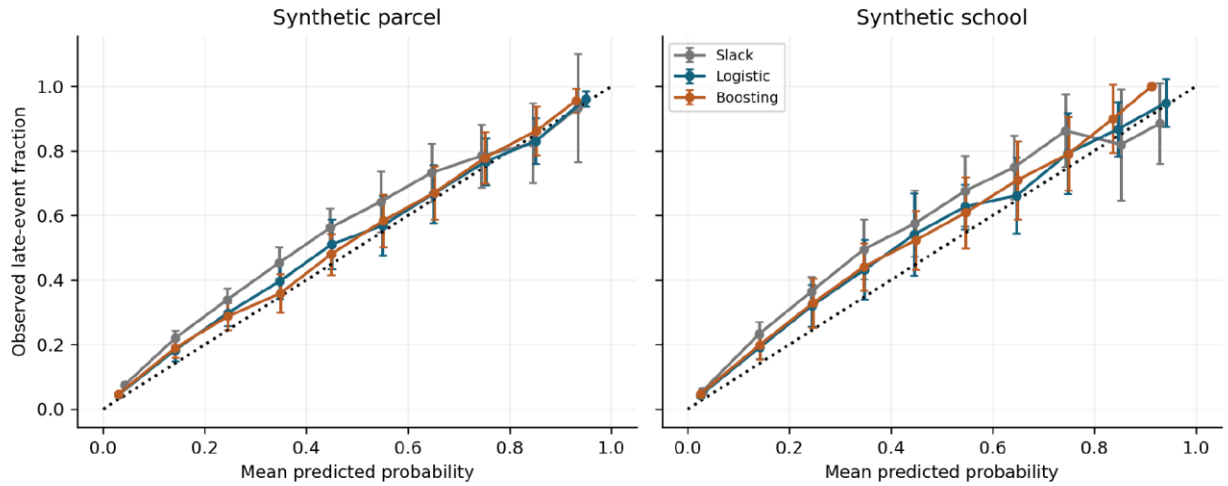


Figure 2: Descriptive 60 Minute Reliability Curves. Points show mean probability and observed event fractions by fixed bins; vertical bars show variation across seeds. The diagonal represents perfect calibration.

The curves compare predicted probabilities with observed event fractions in fixed bins; they are descriptive and do not establish external calibration.

The calibration threshold targets approximately 20 percent alerts in the calibration partition. In the shifted test partition, the mean logistic alert rate is 28.4 percent in parcel and 30.3 percent in school. The shift therefore creates additional workload. A system that reported only recall could conceal this failure; the revised evaluation reports calibration and false alert burden together.

7.5 Fixed Workload Analysis

Table 7 reports performance when review capacity is fixed rather than selected by a probability threshold.

Table 7: Fixed Workload Performance at 60 Minutes

Domain	Review capacity	Recall % / precision % by model	Logistic false alerts per 100
Parcel	5.0	sla 18.6 / 71.1 | log 24.4 / 93.0 | boo 24.1 / 92.0	0.3
Parcel	10.0	sla 32.1 / 61.4 | log 43.8 / 83.6 | boo 43.7 / 83.4	1.6
Parcel	15.0	sla 42.8 / 54.6 | log 58.6 / 74.6 | boo 58.5 / 74.5	3.8
Parcel	20.0	sla 51.6 / 49.3 | log 68.9 / 65.8 | boo 68.4 / 65.3	6.8
School	5.0	sla 25.2 / 64.4 | log 28.9 / 74.0 | boo 28.1 / 71.8	1.3
School	10.0	sla 41.5 / 53.1 | log 48.1 / 61.6 | boo 47.0 / 60.2	3.8
School	15.0	sla 52.7 / 45.0 | log 60.5 / 51.7 | boo 60.2 / 51.4	7.2
School	20.0	sla 61.5 / 39.3 | log 69.1 / 44.3 | boo 68.3 / 43.8	11.1

Capacity selects the highest scoring test departures; labels are used only to evaluate the selected set. The archive reports all three comparators; abbreviations identify slack, logistic, and boosting.

At a fixed ten percent review capacity, the model captures 43.8 percent of late parcel departures with mean precision 83.6 percent. The corresponding school values are 48.1 and 61.6 percent. At five percent capacity, recall is lower but precision is higher; at 20 percent capacity, recall increases while false alerts increase. This is a ranking and capacity analysis, not a demonstration that an alerted departure could have been saved.

Figure 3 visualizes the recall and precision trade-off as review capacity changes.

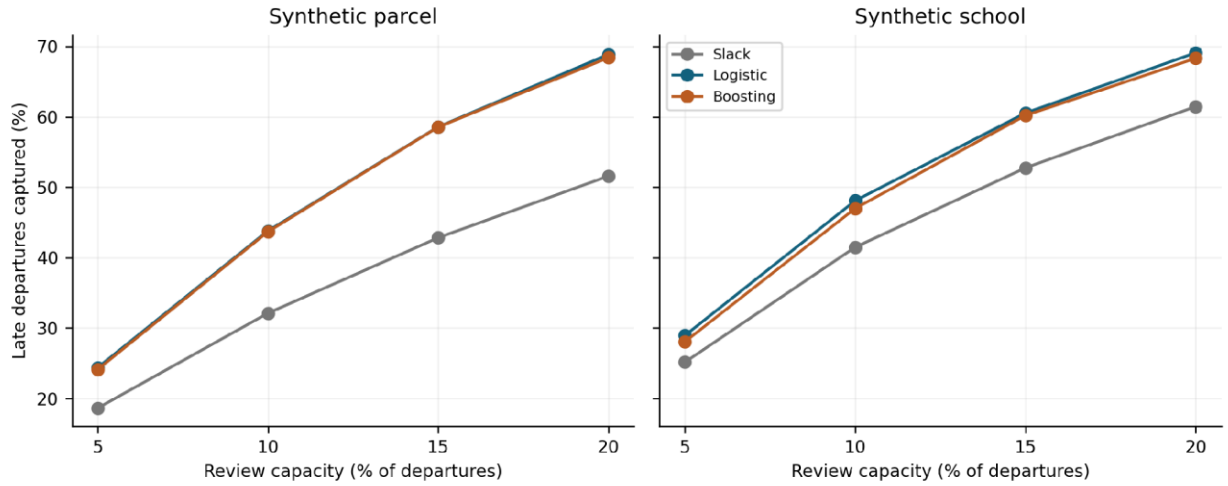


Figure 3: Fixed Workload Alert Analysis at 60 Minutes. Curves show the fraction of late departures captured as the highest scoring 5, 10, 15, or 20 percent are reviewed. This is not an intervention effect.

The curves show the capacity trade-off under ranking-based review and do not measure an intervention effect, alert acknowledgment, or operator response.

7.6 Event Level Coverage Across All Horizons

Table 8 summarizes event-level alert coverage after repeated horizon snapshots are collapsed to one departure.

Table 8: Event-Level Coverage Across All Three Horizons

Domain	Model	Late events alerted %	Mean earliest warning min	On time events alerted per 100	Mean alerts per event	Mean alerts per alerted event
Parcel	Slack only	81.2	74.0	25.8	0.73	1.77
Parcel	Multivariable logistic	88.4	84.7	20.6	0.86	2.28
Parcel	Fixed boosting	88.9	84.2	21.3	0.85	2.23
School	Slack only	86.3	78.9	31.2	0.81	1.91
School	Multivariable logistic	88.6	84.4	30.6	0.91	2.18

Domain	Model	Late events alerted %	Mean earliest warning min	On time events alerted per 100	Mean alerts per event	Mean alerts per alerted event
School	Fixed boosting	88.6	84.0	31.5	0.90	2.09

Alerts are collapsed to one event before calculating burden. Warning time is nominal and does not subtract response or mobilization delay.

When the three snapshots are collapsed to one event, the logistic model alerts at least once for approximately 88.4 percent of late parcel events and 88.6 percent of late school events. The mean earliest nominal warning is 84.7 and 84.4 minutes. On time events receiving at least one alert are 20.6 and 30.6 per 100 departures. A mean of 0.86 and 0.91 alerts per event illustrates why repeated alerts require suppression or acknowledgment rules.

These quantities use the maximum alerting horizon as the earliest warning and do not subtract feed, computation, acknowledgment, mobilization, or travel time. A nominal 90 minute warning may provide much less useful time in a real operation.

7.7 Observation Layer

Table 9 reports the effect of the controlled observation-layer stress conditions.

Table 9: Observation-Layer Stress at 60 Minutes

Domain	Condition	AP	ROC AUC	Brier	Log loss	ECE	Recall %	Precision %
Parcel	Clean feature access	0.740 [0.715, 0.767]	0.891 [0.885, 0.895]	0.087 [0.083, 0.091]	0.299 [0.287, 0.311]	0.031 [0.021, 0.041]	78.6	53.9
Parcel	Delayed missing stale	0.717 [0.694, 0.746]	0.881 [0.872, 0.887]	0.092 [0.089, 0.095]	0.312 [0.301, 0.317]	0.034 [0.025, 0.045]	77.0	53.1
School	Clean feature access	0.571 [0.521, 0.638]	0.861 [0.845, 0.873]	0.076 [0.074, 0.078]	0.266 [0.259, 0.274]	0.024 [0.016, 0.038]	80.8	33.2
School	Delayed missing stale	0.564 [0.508, 0.627]	0.857 [0.839, 0.870]	0.077 [0.075, 0.080]	0.269 [0.263, 0.280]	0.026 [0.019, 0.041]	79.5	33.1

The degraded condition uses ten percent missingness, three minute mean delay, ten percent stale values, and one minute timestamp jitter. These are controlled assumptions, not operational estimates.

Adding delayed, stale, missing, and timestamp noisy observations reduces 60 minute mean average precision from 0.740 to 0.717 in parcel and from 0.571 to 0.564 in school. The mean ECE and Brier score also change. The effect is modest in the chosen scenario because explicit missing indicators and imputation preserve some information; more severe missingness reduces ranking performance in the sensitivity grid. This experiment supports the methodological point that information availability should be treated as a modeled layer, while not validating the selected delay or missingness rates.

7.8 Sensitivity to Generator Assumptions

Table 10 presents the one-factor diagnostic sensitivity screen for the AP difference.

Table 10: One-Factor Sensitivity of AP Difference

Factor	Low	Default	High	Low parcel / school	Default parcel / school	High parcel / school
late tolerance	0	5	10	0.162 / 0.073	0.227 / 0.098	0.243 / 0.101
shift strength	0	1	1.5	0.246 / 0.128	0.227 / 0.098	0.197 / 0.078
forecast noise	0.75	1	1.25	0.178 / 0.073	0.227 / 0.098	0.269 / 0.123
forecast correlation	0	0.65	0.9	0.204 / 0.081	0.227 / 0.098	0.236 / 0.111
release variability	0.5	1	1.5	0.251 / 0.102	0.227 / 0.098	0.202 / 0.087
operating day pressure	0	1	1.5	0.242 / 0.090	0.227 / 0.098	0.206 / 0.102
missingness	0	0.1	0.25	0.225 / 0.090	0.212 / 0.075	0.171 / 0.056

Each cell reports multivariable logistic AP minus slack AP. Three seeds and 15 events per day were used for diagnostics. Values are not population uncertainty ranges.

The sensitivity screen shows that the direction and magnitude of the multivariable advantage vary with assumptions. At the default tolerance, the parcel AP difference is about 0.227; it changes when forecast noise, correlated error, shift strength, pressure dependence, release variability, lateness tolerance, or missingness changes. The school differences are smaller and more variable. The broad conclusion is not that one setting is correct, but that the apparent effect size is

conditional on the generator. Sensitivity output is therefore part of the result rather than an optional appendix.

Forecast noise, missingness, and shift severity affect both ranking and calibration. The expanded paper does not select only favorable values. It reports the full three level diagnostics in the supplement and displays a parcel heatmap of paired AP differences.

Figure 4 visualizes the diagnostic sensitivity of the parcel AP difference to selected generator assumptions.

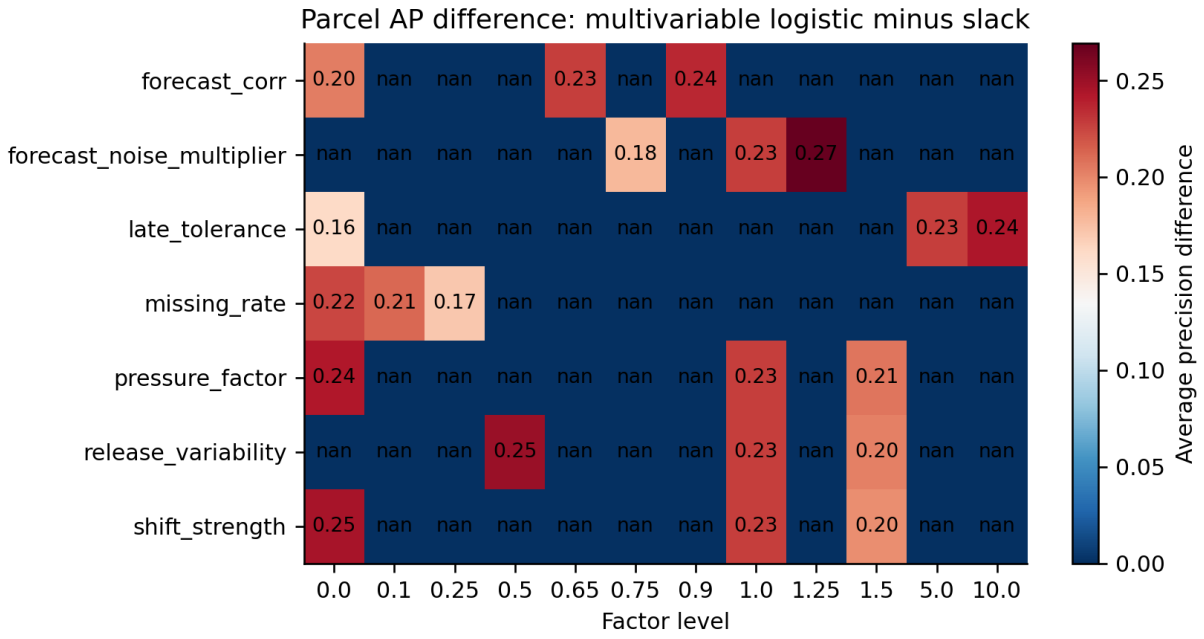


Figure 4: One Factor Sensitivity Screen for the Parcel AP Difference Between Multivariable Logistic and Slack Only Models. Values are diagnostic means over three seeds.

The screen shows how the logistic advantage over the slack-only comparator changes under selected assumptions; it is diagnostic rather than inferential.

7.9 Explainable Logistic Scores

Table 11 reports the calibrated standardized logistic coefficients for the 60-minute models.

Table 11: Calibrated Standardized Logistic Coefficients at 60 Minutes

Domain	Feature	Mean coefficient	Seed percentile interval
Parcel	age	-0.112	[-0.220, -0.024]
Parcel	crew_forecast	-0.095	[-0.178, -0.006]
Parcel	incident	0.937	[0.781, 1.141]

Domain	Feature	Mean coefficient	Seed percentile interval
Parcel	pressure	0.559	[0.450, 0.697]
Parcel	release_forecast	0.176	[0.076, 0.315]
Parcel	slack	-0.991	[-1.158, -0.854]
Parcel	vehicle_forecast	0.181	[0.071, 0.304]
Parcel	work_forecast	0.272	[0.179, 0.390]
School	age	-0.128	[-0.271, 0.012]
School	crew_forecast	0.243	[0.100, 0.422]
School	incident	0.570	[0.515, 0.666]
School	pressure	0.390	[0.228, 0.508]
School	release_forecast	0.005	[-0.186, 0.143]
School	slack	-1.260	[-1.427, -1.104]
School	vehicle_forecast	0.017	[-0.103, 0.167]
School	work_forecast	-0.041	[-0.149, 0.036]

Coefficients are conditional terms in a correlated synthetic feature representation. They are not causal effects or independent physical sensitivities.

The 60 minute calibrated standardized coefficients show that slack is strongly negative in both parameterizations, while incident and pressure are positive. Component forecast coefficients vary in sign and magnitude because the component forecasts and slack are correlated. A coefficient should not be read as an independent physical effect or a causal action recommendation. It is a conditional term in a synthetic logistic score.

Table 12 gives worked synthetic examples of the logistic score decomposition.

Table 12: Worked Synthetic Logistic Explanations

Domain	Example	Late	Predicted probability	Slack	Incident	Largest log odds contributions
Parcel	high score alert	1	0.998	-45.9	1	slack +3.83; incident +3.11; pressure +0.88
Parcel	high score alert	1	0.987	-37.0	1	slack +3.18; incident +3.11; vehicle_forecast +0.53

Domain	Example	Late	Predicted probability	Slack	Incident	Largest log odds contributions
Parcel	high score false alert	0	0.957	-19.3	1	incident +3.11; slack +1.88; pressure +0.96
School	high score alert	1	0.994	-29.2	1	slack +5.14; incident +1.96; pressure +1.05
School	high score alert	1	0.971	-20.3	1	slack +3.95; incident +1.96; pressure +0.87
School	high score false alert	0	0.897	-14.6	1	slack +3.19; incident +1.96; crew_forecast +0.65

The exact contribution vector is in explanation_examples.csv. A high score false alert is intentionally included in each domain.

The table displays two high score alerts and one high score false alert for each domain. The exact log odds identity is reconstructed from the standardized values, fitted coefficients, calibration slope, and calibrated intercept. In a real workflow, an alert narrative could translate the top contributions into operational language such as predicted work readiness beyond the departure commitment, a verified incident, or stale information. The model should also show the source timestamps and permitted response options. No dispatcher comprehension test was performed.

Figure 5 visualizes the exact calibrated log-odds contributions for one illustrative high-score parcel.

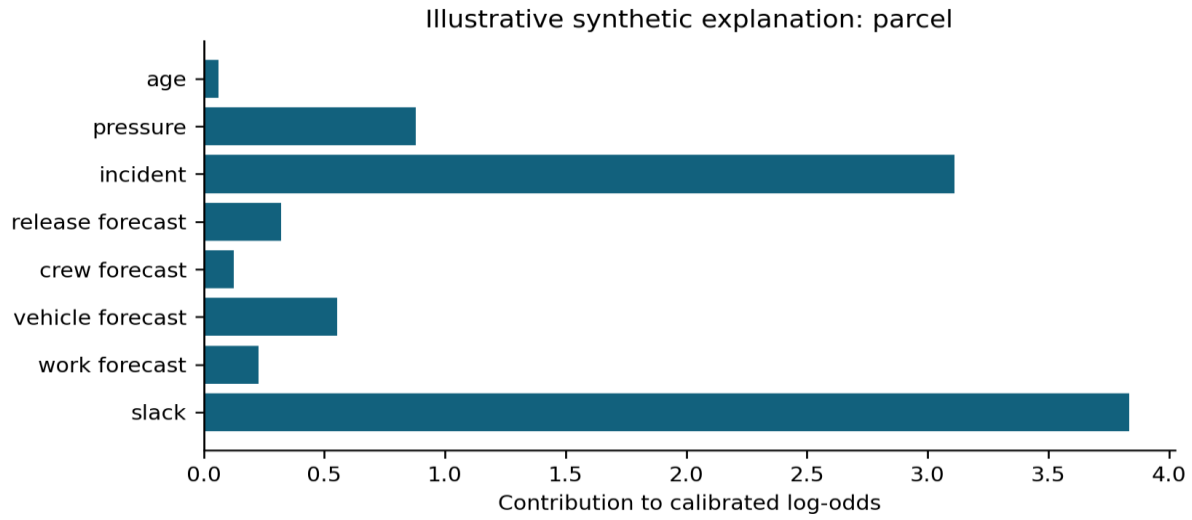


Figure 5: Exact Calibrated Log Odds Contributions for One High Score Synthetic Parcel Example. Positive bars increase the model score within the model space; they are not causal effects.

The bars reconstruct one model-space contribution identity; positive contributions increase the fitted score but are not causal effects.

7.10 Cross Domain Transfer

Table 13 compares cross-domain transfer with local estimation at 60 minutes.

Table 13: Synthetic Transfer and Local Estimation at 60 Minutes

Target or direction	Estimation	AP	Brier	Log loss	Recall %	Precision %	Alert %
Parcel to school	Source calibration	0.508 [0.459, 0.569]	0.091 [0.083, 0.104]	0.313 [0.286, 0.346]	58.6	39.1	19.3
Parcel to school	Target recalibration	0.508 [0.459, 0.569]	0.088 [0.081, 0.102]	0.305 [0.282, 0.341]	74.0	31.8	29.9
School	Local training	0.580 [0.526, 0.638]	0.079 [0.072, 0.091]	0.275 [0.250, 0.308]	80.2	34.0	30.3
School parcel to	Source calibration	0.689 [0.653, 0.719]	0.107 [0.099, 0.116]	0.364 [0.341, 0.396]	79.2	46.9	32.6
School parcel to	Target recalibration	0.689 [0.653, 0.719]	0.099 [0.094, 0.105]	0.331 [0.313, 0.348]	73.4	52.9	26.6
Parcel	Local training	0.741 [0.715, 0.767]	0.088 [0.083, 0.093]	0.300 [0.282, 0.312]	79.9	53.9	28.4

Target recalibration uses labeled target calibration events. It changes probability quality and threshold but cannot recover source model ranking information.

Applying the parcel fitted score to school events gives mean average precision of 0.508 versus 0.580 for local school estimation. Applying the school score to parcel events gives 0.689 versus 0.741 for local parcel estimation. Target recalibration changes Brier score and the threshold because it uses labeled target calibration data, but it leaves average precision unchanged because a monotone logistic calibration map preserves ranking. The comparison supports transferring feature definitions, code, and evaluation practice rather than treating fitted coefficients as portable.

8 APPLICATION TO E COMMERCE AND STUDENT TRANSPORTATION

8.1 Fulfillment to Carrier Handoff

The practical parcel implementation point is the interface between work completion and the transportation commitment. The operational technology overlap includes data pipelines, schedule versioning, event driven status, time series forecasts, exception rules, dashboards, alert routing, and decision logs. These systems can connect picking, packing, staging, loading, and handoff readiness to the scheduled transportation window. They should not be described as route or fleet control if the responsible manager does not control carrier routing, driver assignment, or network dispatch. The model supports an upstream handoff decision; it does not imply control of the entire carrier network.

Carrier cost and customer experience are downstream hypotheses. Missed departures can plausibly produce recovery handling, additional transportation, rework, or replacement expense, and a reliable handoff can plausibly improve the probability of meeting a delivery commitment. The present paper does not estimate shipping cost, carrier savings, replacement shipments, support contacts, customer satisfaction, or animal health. Prescription and temperature sensitive products would require a separate study with appropriate release, handling, compliance, and service definitions; they are not modeled here.

8.2 School Bus Departure and Downstream Arrival

The analogous school pilot would focus on depot departure or another defined route origin. Candidate inputs include vehicle readiness, driver availability and qualification status, completion of required checks, status of the preceding assignment, and approved preparation milestones. Alert routing should identify the responsible dispatcher and only permitted actions. Safety, capacity, accessibility, inspection, and communication requirements remain constraints, not variables to trade away for punctuality.

The school arrival outcome needs route geography, traffic, stop sequence, boarding and alighting, and schedule slack. A departure model can be linked to an arrival model but cannot substitute for one. Congestion can affect both vehicle readiness and route travel time, so a field study should not add independent travel time assumptions without modeling their dependence. A model must not recommend skipping a student, advancing a pickup without notice, exceeding capacity, using an unsuitable vehicle, or bypassing an inspection to improve a delay metric.

10 DISCUSSION AND LIMITATIONS

10.1 What the Simulation Supports

Within the author specified baseline generator, retaining readiness components and contextual features provides more ranking information than compressing them into one slack value. The logistic model is close to a fixed boosting comparator, which supports maintaining a transparent model in the candidate set rather than assuming that more complexity is automatically necessary.

Shorter horizons perform better in the generator because later forecasts are less noisy. A deliberate distribution shift worsens calibration and expands the alert queue. Missing, stale, and delayed observations reduce performance. Direct coefficient transfer loses ranking relative to local estimation.

These are implementation and protocol findings. They do not show that readiness components are predictive in Chewy or school operations, that a model can be deployed safely, that a carrier will depart earlier, that a bus will arrive earlier, or that a pet parent or student will experience a benefit.

10.2 Three Validity Questions

Internal validity of the simulation asks whether the implementation matches the stated generator. The supplement checks event counts, partition disjointness, finite probabilities, fixed seeds, and legacy numerical reconciliation. Statistical validity asks whether models are compared under chronological separation, whether calibration is evaluated separately, whether repeated snapshots are handled correctly, and whether uncertainty is reported in paired form. The revised paper improves those elements but remains limited by the finite number of seeds and the diagnostic rather than inferential nature of the sensitivity grid.

External validity asks whether the generator, features, missingness, and operating constraints represent real transportation systems. The answer is not established here. The generator omits detailed queueing, route geography, concurrent resource competition, cancellations, schedule changes, human adaptation, driver or vehicle reassignment, database outages, sensor semantics, carrier tender rules, passenger boarding, and downstream recovery. Its parameter values are illustrative assumptions. A richer simulation can test more mechanisms, but only authorized observed data can establish whether the mechanisms and scales match a real operation.

10.3 Interpretation of Explanations

The logistic decomposition is transparent but not causal. Slack and readiness component values are correlated; a positive contribution from a feature does not mean that changing the feature alone would reduce delay. A feature can also be a proxy for a policy that has already intervened. Explanation quality depends on event definitions, data freshness, and operational meaning. A production interface should group correlated prerequisites, display source ages, disclose missingness, and identify which response options are actually permitted.

10.4 Alert Burden and Human Response

The original calibration threshold creates approximately 20 percent alerts only in calibration. Shifted test conditions can increase that rate. Fixed workload curves reveal the precision and recall trade off under explicit review capacity, but they do not measure alert fatigue, acknowledgment, triage time, or action effectiveness. Repeated alerts at three horizons may be useful for escalation or may be redundant. Those alternatives require a human in the loop evaluation rather than a further change to the classifier alone.

10.5 Limitations That Remain After Revision

The twenty seed expansion improves Monte Carlo characterization but does not turn synthetic intervals into population confidence intervals. The one factor sensitivity grid does not explore all interactions among assumptions. The observation layer models missingness, staleness, delay, and timestamp noise with simple distributions; it does not recreate a database, scanner, telematics, or dispatch system. The transfer study uses the same synthetic feature schema and related latent structure, making it more favorable than a real cross domain test. Target recalibration uses labeled target data and is therefore supervised adaptation, not zero shot transfer.

The paper reports no intervention simulation because a parameterized action effect would be another unvalidated assumption and could be mistaken for evidence of benefit. The utility equation is retained as a protocol specification. A future study should estimate or design the intervention process, including the possibility that rescuing one departure harms another. No causal benefit, cost saving, emission reduction, school lateness reduction, customer improvement, pet health improvement, or safety improvement is claimed.

11 CONCLUSION

The revised evidence shows that, under one author specified generator, a multivariable logistic model can outperform a calibrated slack only model, remain close to one fixed boosting comparator, and provide a reconstructable score decomposition. The expanded analysis also shows why those results must be qualified: performance varies by horizon and assumptions, calibration and alert burden change after distribution shift, imperfect observation degrades inputs, repeated snapshots create event level workload, and fitted coefficients do not transfer perfectly between stylized domains.

The practical contribution is a reproducible starting point for operational research, not a validated production system. A responsible next step is an authorized field study with stable endpoint definitions, data availability audits, chronological validation, fixed workload evaluation, shadow deployment, human factors review, and a separate controlled assessment of permitted interventions. Until that work is completed, the numerical results should be interpreted only as behavior of the supplied synthetic generator.

REFERENCES

Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., & Kaushal, R. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making, 17*, 36. https://doi.org/10.1186/s12911-017-0430-8

Babaei, M., & Rajabi Bahaabadi, M. (2019). School bus routing and scheduling with stochastic time-dependent travel times considering on-time arrival reliability. *Computers & Industrial Engineering, 138*, 106125. https://doi.org/10.1016/j.cie.2019.106125

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning, 79*, 151–175. https://doi.org/10.1007/s10994-009-5152-4

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78(1)*, 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

Chewy Inc. (2026). Annual report on Form 10-K for the fiscal year ended February 1, 2026. Filed March 25, 2026. Stable source

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29(5)*, 1189–1232. https://doi.org/10.1214/aos/1013203451

Gama, J., Zliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46(4)*, Article 44. https://doi.org/10.1145/2523813

Gautam, N., & Geunes, J. (2024). Analysis of real-time order fulfillment policies: When to dispatch a batch. *Service Science, 16(2)*, 85–106. https://doi.org/10.1287/serv.2022.0042

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of Machine Learning Research, 70*

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine learning-based science. *Patterns, 4(9)*, 100804. https://doi.org/10.1016/j.patter.2023.100804

Kull, M., Silva Filho, T., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. *Proceedings of Machine Learning Research, 54*

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference*

on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Mohring, U., Jacobi, C., Furmans, K., & Stolletz, R. (2024). Managing cutoff-based shipment promises for order fulfilment processes in warehousing. *OR Spectrum*, *46*, 513–543. https://doi.org/10.1007/s00291-023-00739-7

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625–632. https://doi.org/10.1145/1102351.1102430

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Snoek, J., & Lakshminarayanan, B. (2019). Can you trust your model uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems, 32*.

Park, J., & Kim, B. I. (2010). The school bus routing problem: A review. *European Journal of Operational Research, 202(2)*, 311–319. https://doi.org/10.1016/j.ejor.2009.05.017

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. Stable source

Pineda Jaramillo, J., Munoz, C., Mesa Arango, R., Gonzalez Calderon, C., & Lange, A. (2024). Integrating multiple data sources for improved flight delay prediction using explainable machine learning. *Research in Transportation Business & Management*, *56*, 101161. https://doi.org/10.1016/j.rtbm.2024.101161

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10(3)*, e0118432. https://doi.org/10.1371/journal.pone.0118432

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation, 7(1)*, 12–24. https://doi.org/10.1057/jos.2012.20

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems, 28*

Sun, S., Duan, Z., & Xu, Q. (2018). School bus routing problem in the stochastic and time-dependent transportation network. *PLOS ONE, 13(8)*, e0202618. https://doi.org/10.1371/journal.pone.0202618

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, 230. https://doi.org/10.1186/s12916-019-1466-7

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26(6)*, 565–574. https://doi.org/10.1177/0272989X06295361

Zhang, Q., Ma, Z., & Cui, Z. (2025). A causality-based explainable AI method for bus delay propagation analysis. *Communications in Transportation Research, 5(2)*, 100178. https://doi.org/10.1016/j.commtr.2025.100178

2026 by the Authors. This Article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/)