

Original Article

Continuous Evaluation and Human Override in Automated Retail Replenishment

***Sachidananda Singh**

Senior Engineering Manager, Wayfair, USA.

Abstract:

Human review can correct information gaps in automated replenishment, but it also consumes time and can delay orders. This study evaluates a budget-constrained review framework in which a logistic score ranks candidate decisions for exposure to demand, supply or inventory-record disruptions. A simulation uses 1,000 M5 item-store sales series for evaluation and 500 separate series for scorer fitting over 16 weekly review cycles. All policies use the same established count-forecasting rule, updated from their own observed sales. Reviewers receive imperfect side information, may approve an order unchanged, and incur explicit review costs and delay. Five disruption draws and three reviewer seeds per draw support paired comparisons of automation, random review, risk-based review and full review. At a 10% budget, risk routing has exposure precision 23.37% against 11.03% for random review, but average cost is 0.54% higher and fill improves by only 0.067 percentage points. Full review has lower cost under the main assumptions, while increasing the review charge reverses that advantage in a sensitivity scenario. The framework records forecasts, decisions, review actions and outcomes for continuing evaluation, but the fitted scorer remains fixed within each evaluation run. The findings concern the specified simulation and do not establish the effectiveness of real planners or a general advantage of human override.

Keywords:

Automated Replenishment, Human Review, Inventory Control, Decision Monitoring, Logistic Regression, Retail Simulation, Lost Sales.

Article History:

Received: 04.08.2024

Revised: 06.09.2024

Accepted: 17.09.2024

Published: 24.09.2024

1. Introduction

Automated replenishment links sales forecasts to purchase orders. Its scale makes universal manual inspection difficult, while local information about promotions, delivery problems or unavailable shelf stock may remain outside the forecasting system. A central operational question is therefore which decisions deserve review when planner capacity is limited. Inventory records and physical stock can differ materially [1]. Inspection and replenishment consequently interact: correcting a record changes the inventory position used by the ordering policy [2]. Retail ordering studies also show that departures from automated recommendations can reflect operational conditions omitted from the system [3]. These observations motivate review, but do not imply that every adjustment is beneficial. Evidence on judgmental forecasting adjustments shows substantial variation, with some changes improving accuracy and others harming it [4].

Review has at least three consequences. It can reveal information, it can alter the order quantity, and it can postpone release of the order. Its value depends on the combination. A policy that identifies more disrupted decisions may still increase total cost if review



is slow, the information is unhelpful, or the adjustment introduces excess stock. Ranking by disruption probability is also different from ranking by expected preventable loss. This study contributes an explicit simulation and monitoring design for examining those tradeoffs. The simulation holds the forecasting rule and available operational information consistent across policies, separates exposure detection from inventory outcomes, permits reviewers to approve unchanged orders, and includes a direct review charge. It reports service, lost units, holding, review load and terminal inventory alongside cost. The purpose is to assess an implementable review heuristic under transparent conditions, not to introduce a new inventory-optimal control algorithm. The empirical basis is M5 recorded retail sales [5], used as an exogenous request proxy before injecting controlled disruptions. The study does not recover true customer demand from historical stockout-censored sales, observe actual planners, or evaluate a retail deployment. These boundaries are central to interpretation.

2. Related Work

Human-automation design distinguishes information acquisition, analysis, decision selection and action implementation [6]. Replenishment systems can distribute these functions differently: an automated forecast may coexist with a human inventory count or a manual order-release decision. The present framework places the human at a review stage between candidate-order calculation and execution. Learning-to-defer research studies when a model should pass a decision to an expert [7]. Here the score is trained to identify event exposure using simulation labels, rather than learning an expert's comparative performance. It is therefore a budgeted triage heuristic, not an implementation of a consistency result for optimal deferral. Learning the incremental value of review would require additional data about reviewer actions and their counterfactual consequences.

Forecast monitoring predates modern machine learning. Tracking signals summarize discrepancies between forecasts and observations [8], while concept-drift research addresses changes in the processes generating data [9]. In inventory systems, apparent forecast errors can also change because the ordering policy changes stock availability. A sequence of low observed sales can indicate weak demand, a stockout or a record error. Monitoring should therefore retain operational state as well as forecast errors. Work on technical debt and production readiness emphasizes that a model forms part of a larger system with data, feedback and monitoring dependencies [10, 11]. These ideas support persistent decision records and controlled evaluation. They do not establish that logging alone improves inventory outcomes. The experiment below tests a frozen routing scorer; adaptive retraining and the causal value of a continuing audit program remain future work.

3. Decision and Review Framework

3.1. Information available before ordering

At a weekly decision epoch, the system has recorded on-hand inventory, open orders and their promised dates, recent receipts, prior order quantities and completed sales observations. Its forecast uses only observations available before the decision. In simulation, each policy's realized sales can differ because its inventory position differs. The same forecasting rule is applied separately to those policy-specific observations. No automated policy receives the simulator's unobserved lost requests. The reviewer can acquire side information about an injected event. Information availability, announcement time, probability of discovery and adjustment noise are explicit simulation assumptions. The reviewer never receives the realized future request sequence. A review may approve the existing order, change its quantity or correct an inventory record. A delayed approval still has an operational effect, even if the quantity stays unchanged.

3.2. Count forecast and candidate order

The forecast uses the latest 56 observed daily sales, combining pre-simulation history with sales generated under the policy. The daily mean μ and sample variance v define a count distribution. If v exceeds μ , a negative binomial distribution is used; otherwise the model uses a Poisson distribution with mean μ . Zero recent sales imply a degenerate zero forecast. There is no substitution of a quantile average for a mean. For a negative binomial variable parameterized by success probability p and shape r , the mean is $r(1 - p)/p$ and the variance is $r(1 - p)/p^2$. Matching μ and v gives $p = \mu/v$ and $r = \mu^2/(v - \mu)$. Conditional independence and a common daily

distribution imply that cumulative requests over T days have shape Tr and the same p . Under the Poisson model, the corresponding mean is $T\mu$. These are modeling assumptions, not established properties of $M5$ demand.

Let $Q_{0.95}(T)$ denote the 0.95 quantile of cumulative requests under this count model. The review period R is seven days, and nominal lead time L is three days. The candidate order is

$$a_{i,t} = \max\{0, Q_{0.95}(R + L) - IP_{i,t}\}$$

Where recorded inventory position IP is recorded on-hand inventory plus outstanding ordered units. Orders and inventory are in whole units. The 0.95 level is a fixed stocking heuristic. It is neither a guarantee of realized cycle service nor a cost-optimal fractile derived from the holding and shortage charges used below. In particular, a per-unit-day holding charge and a per-unit lost-sales penalty cannot be substituted into a single-period critical-fractile formula without an appropriate cost-horizon derivation.

3.3. Exposure scoring and capacity allocation

The score uses nine operational features: relative forecast dispersion, recent forecast error, a zero-sales signal conditional on positive recorded inventory, recent receipt lateness, an order-size deviation, forecast volume, overdue open orders, a recent sales-to-forecast ratio and its absolute value. Definitions are retained in the executable feature function. Only completed observations and existing records are available at scoring time. The error signal is motivated by monitoring work [8] but is not claimed to reproduce Trigg's original statistic.

Features are standardized and fitted with a class-balanced logistic regression on the separate calibration series. Logistic regression provides an interpretable linear ranking in feature space [12]. Because class weighting changes the fitted probability scale, the resulting value is used as a score, not as a calibrated probability of disruption or harm.

$$s_{i,t} = \frac{1}{1 + \exp[-(\beta_0 + \beta^T x_{i,t})]}$$

For N candidates and review fraction B , the risk-based policy sends the highest-scoring round(BN) decisions to review. Random review sends the same number to review using a random permutation. Full review inspects all candidates, and automation inspects none. Candidates include decisions whose proposed order is zero: a missed order or inaccurate record can also merit inspection. The training label marks an injected event that is active during or begins within the nominal ten-day protection interval. The label is defined from the common exogenous event schedule, so its denominator is identical across review policies. For an inventory-record event, exposure continues after a policy corrects the record. This deliberate definition supports comparable detection summaries but does not identify unresolved harm. It may reward repeated detection of an already addressed event. Precision and recall must therefore be interpreted as event-exposure measures, not as prevented-error measures.

3.4. Decision logging and continuing evaluation

A decision record should retain the forecast version, historical data cutoff, candidate quantity, recorded inventory position, score, routing reason, reviewer action, release time, receipts and resulting sales. Reason codes and later audits can distinguish supplier problems, inventory discrepancies and demand changes. Forecast monitoring, reviewer behavior and service outcomes should be examined together, since policy-induced stockouts affect subsequent sales observations. The simulation implements weekly outcome records and capacity checks. The broader framework would use these records for periodic review of the forecaster, ranking rule and reviewer process. That continuing evaluation is a proposed operating procedure: the experiment does not retrain the scorer during evaluation or measure gains from adaptation.

Logged reviewed cases alone are not an unbiased sample of all decisions. Future deployment evaluation would require planned audits of some unreviewed decisions and, where operationally appropriate, randomized allocation of limited review capacity. A

decision log also does not reveal the outcome of an alternative unexecuted order. Causal estimates of override value need a valid comparison design or an explicitly validated simulation.

4. Simulation Design

4.1. Historical basis and separation of samples

The M5 dataset contains daily sales for 30,490 item–store combinations [5]. From the fixed random sample of 3,000 series, the simulation retains series with at least one year of history before the final 112 days and at least ten positive-sales days in that year. A seeded permutation selects 1,000 evaluation series and a disjoint set of 500 calibration series. The horizon spans 29 February–19 June 2016 and contains 16 weekly decisions per series. The final 112 recorded observations are treated as baseline requests. Controlled demand and operational events generate the simulator's request and supply paths. Fulfilled requests become observed sales; unfulfilled requests are lost and are unavailable to the forecasting rule. This construction uses real sales patterns but remains a synthetic inventory environment. It does not imply that those historical requests were uncensored or that the injected events occurred at Walmart.

4.2. Injected events

Each series receives at most one event type in a replication. Ten percent are assigned a promotion, five percent a demand collapse, ten percent a supplier disruption and five percent a phantom-stock event; the remaining 70% have none. Realized counts fluctuate with the draws. Event start times are sampled within the horizon, with the mechanisms specified in the accompanying code.

Table 1. Event Mechanisms used in the Simulation

Event	Assigned probability	Mechanism
Promotion	0.10	Fourteen days of Poisson requests with mean 2.5 times baseline requests plus 0.75 times the pre-horizon 28-day mean
Demand collapse	0.05	Each baseline request survives independently with probability 0.20 until the horizon ends
Supplier delay	0.10	Orders released during a 21-day disruption take ten rather than three days to arrive
Phantom stock	0.05	Physical stock is removed while the record remains unchanged; the removed quantity cannot exceed physical stock

The intended phantom removal is the larger of two units and 1.5 weeks of pre-horizon mean sales, rounded to an integer. Actual removal is capped at available physical stock, and the record offset equals that actual removal. A successful count corrects the offset without restoring lost physical units. This distinction preserves stock conservation.

4.3. Reviewer behavior

The main knowledge probability is 0.80 when event information is available. Promotion and collapse information can be known seven days before the event, supplier information three days before it, and a phantom-stock discrepancy only once it exists. A reviewer without information approves the candidate quantity unchanged. The main review process delays order release by one day. An informed demand adjustment starts from the pre-simulation 56-day mean, uses the event multiplier on affected future days, and uses the baseline mean outside the event. This avoids multiplying a forecast that has already adapted to the event. The pre-simulation Fano factor is held constant, so variance scales linearly with the adjusted mean.

This is a simplifying reviewer model, not an estimate of actual human forecasting behavior. The promotion reviewer uses the stated multiplier and baseline but does not recover the simulator's additional promotion component or realized future baseline counts. Since the negative binomial success probability remains common across days, the assumed daily shapes can be summed for a cumulative distribution.

Informed supplier review changes the protection period to reflect the longer lead time. An informed inventory count corrects recorded stock before recalculating the order. Revised protection periods include the review delay. Multiplicative quantity noise has log standard deviation 0.125 and a log mean chosen so its multiplier has expectation one; it is applied only to informed quantity adjustments. There is no assumption that uninformed reviewers always make a noisy adjustment. Random numbers for information discovery and adjustment noise are indexed by series, cycle, disruption replication and reviewer seed. The same indexed draws are used across policies. Random routing uses a separate stream. This pairing avoids changing all later reviewer draws merely because one policy reviews a different earlier decision.

4.4 Costs service and horizon accounting

The main cost is the sum of end-of-day holding, lost-request penalties and review charges. Holding costs one illustrative unit per unit of inventory per day; lost requests cost 19 per unit; every reviewed decision costs one. These are scenario values, not estimated monetary costs.

$$C = \sum_i \sum_t I_{i,t} + 19 \sum_i \sum_t U_{i,t} + \sum_i \sum_k V_{i,k}$$

Here I is physical end-of-day inventory, U is unfulfilled requests, and V indicates review. Fill rate is one minus total unfulfilled requests divided by total requests. Cycle service is the share of series-week cycles with no lost requests, including cycles with zero requests. These measures have different denominators and are reported separately. Initial inventory is the same model-based ten-day quantile for every policy. There are no purchasing, fixed-ordering, disposal or terminal salvage charges. Pending orders and residual stock at the end are reported, because a finite horizon can shift costs beyond the observation window. The simulation checks the physical balance between initial stock, receipts, fulfilled requests, phantom removals and terminal stock. These checks establish accounting consistency, not realism of the inventory assumptions.

4.5. Replications and prespecified sensitivities

The main comparison uses five independent disruption draws, with three reviewer seeds in each draw. Each draw fits the same nine-feature scorer on its 500 calibration trajectories and holds that scorer fixed during evaluation. Calibration uses complete simulated trajectories of separate series over the same calendar dates as evaluation. Separation is therefore by series, not by calendar time; the design tests generalization across series within an environment rather than prospective deployment. The main constrained review fraction is 10%. Outcomes are paired within draw and reviewer seed. Results are averaged over reviewer seeds within each draw before summarizing variation across environments. Five draws provide a limited description of Monte Carlo variation; they are not treated as a large sample for precise confidence claims.

Additional comparisons vary the constrained budget to 5% and 20%, knowledge probability to 0.50 and 1.00, and review delay to zero and two days. These sensitivities use the first disruption draw and one reviewer seed and are explicitly descriptive. Review-cost sensitivities at zero and five units can be calculated from saved review counts without rerunning the physical trajectories, since these policies do not optimize their decisions against the charge. The supplementary replication package includes source checksums, sampled series, event schedules, model settings, per-series outcomes, weekly audits and numerical checks.

5. Results

5.1. Service and cost

Table 2 summarizes the main policies, averaging reviewer seeds within each disruption draw and then averaging the five draws. Each constrained policy reviews exactly 100 of 1,000 decisions per cycle, or 1,600 decisions per run. Automation has a fill rate of 90.289%; random and risk-based review have fill rates of 90.614% and 90.681%. Full review has the

highest average fill rate, 90.958%, but a lower share of cycles without lost requests than the other policies. The difference illustrates why unit fill and cycle service must be reported separately.

Table 2. Main Outcomes over Five Disruption Draws. Cost Covers the Full 112-Day Horizon in Illustrative Units. Mean Stock is Physical End-Of-Day Inventory per Series per Day

Policy	Fill %	No-loss cycles %	Cost per series	Mean stock
Automation	90.289	87.142	1618.92	11.894
Random review 10%	90.614	87.363	1613.16	11.914
Risk review 10%	90.681	87.536	1621.86	12.010
Full review	90.958	86.969	1563.18	11.430

Risk-based review does not have the lowest cost. Its mean cost per series is 1,621.86, compared with 1,613.16 for random review and 1,618.92 for automation. Full review costs 1,563.18 under the assumed review charge. The outcome does not support a claim that selecting the highest exposure scores is cost-superior to random allocation or that full review is necessarily harmful.

Table 3 decomposes cost. Relative to random review, risk-based review has lower lost-sales penalties but greater holding cost. The two constrained policies incur the same direct review charge. Full review reduces both holding and lost-sales penalties sufficiently to offset its larger review charge in this main scenario.

Table 3. Mean Cost Components per Series over the Horizon. Rounding May Prevent Displayed Components from Summing Exactly to Table 2

Policy	Holding	Lost-sales penalty	Review
Automation	1332.15	286.77	0.00
Random review 10%	1334.39	277.17	1.60
Risk review 10%	1345.07	275.19	1.60
Full review	1280.19	266.99	16.00

5.2. Exposure ranking and variation across environments

Risk review's mean exposure precision is 23.37%, against 11.03% for random review; recall is 20.79%, against 9.81%. These values describe the fixed event-exposure label. On automated evaluation trajectories, the fitted score's area under the ROC curve ranges from 0.636 to 0.664 across draws, with average precision from 0.234 to 0.272. These diagnostics indicate useful discrimination of the specified exposure label, but do not measure the value of a review action. The paired risk-versus-random comparison has an average cost increase of 0.5405% across the five environment means. The environment-specific cost change ranges from a 0.0907% reduction to a 0.9580% increase. Average fill improves by 0.0671 percentage points, ranging from -0.1662 to $+0.3403$ points across environments. Risk routing has lower cost in only one of five draws and higher fill in three. Its stronger exposure detection therefore does not translate into a consistent cost or fill advantage at the fixed budget.

Compared with automation, full review has mean paired cost savings of 3.44% across draws, with a range of 2.18%–4.09%, and a fill increase of 0.6695 percentage points. It also has a 0.1737-point reduction in no-loss-cycle share. These comparisons concern the main information, delay and cost assumptions; they do not establish full review as a generally optimal policy.

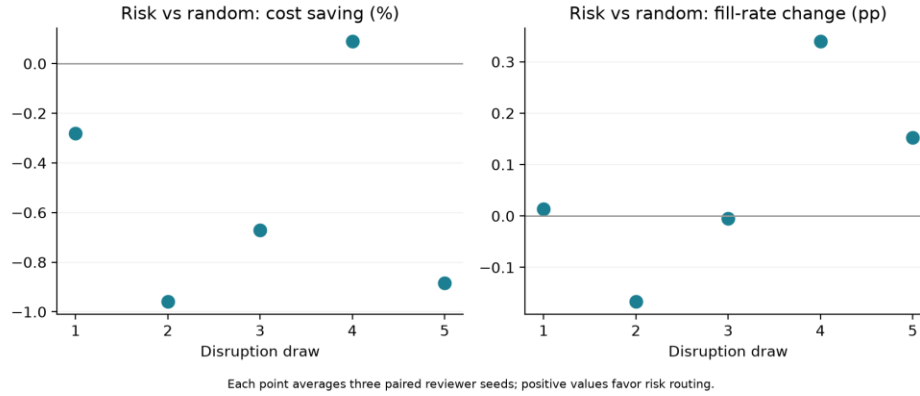


Figure 1. Paired Risk-Versus-Random Outcomes at a 10% Review Budget. Each Point Averages Three Reviewer Seeds within One Disruption Draw. Positive Values Favor Risk Routing. The Spread is Descriptive and is not a Confidence Interval

5.3 Sensitivity to capacity information delay and review price

Sensitivity results use the first disruption draw and reviewer seed zero only. Raising the risk-review budget from 5% to 20% changes fill from 90.388% to 90.313% and cost per series from 1,633.50 to 1,644.41. For random review, the same budget change raises fill from 90.490% to 90.647% and reduces cost from 1,633.79 to 1,627.24. These single-environment observations give no basis for a monotonic benefit from reviewing more high-score decisions.

When reviewer knowledge increases from 0.50 to 1.00, full-review fill rises from 90.172% to 91.248%, and cost per series falls from 1,583.83 to 1,562.36. At a 10% budget, risk-review fill changes from 90.229% to 90.334%, with cost declining from 1,636.28 to 1,630.55. The magnitude is specific to the event and reviewer model; complete knowledge of event attributes still does not reveal future realized requests.

Table 4 shows a substantial delay tradeoff for full review. Faster release improves fill but can increase holding cost; the lowest accounting cost here is associated with the worst fill. A service constraint would therefore change the interpretation of a cost-only comparison.

Table 4. Full-Review Delay Sensitivity in the First Environment and Reviewer Seed. Other Main Assumptions are Held Fixed

Review delay in days	Full-review fill %	Full-review cost per series
0	92.350	1648.19
1	90.885	1576.55
2	88.356	1534.99

Changing the review charge affects cost rankings without changing these policies' physical trajectories. In the first environment and seed, full review costs 1,576.55 per series at a charge of one unit per review, against automation at 1,630.52. At five units per review, full review costs 1,640.55 and is more expensive than automation. This sensitivity prevents treating the main full-review cost advantage as independent of the assumed value of planner time.

5.4. Accounting and terminal state

The saved artifact audit reconstructs 66 evaluation policy trajectories and 1,056,000 weekly decision records. Inventory balances are exact, review counts match every prescribed weekly budget, and no uninformed review changes the proposed quantity. Reconstructed rolling means and variances match the saved causal features. A future-request perturbation check leaves earlier

decisions unchanged, and a discrete convolution check validates the assumed negative binomial aggregation. Mean terminal on-hand inventory per series is 11.090 for automation, 11.325 for random review, 11.331 for risk review and 12.090 for full review. Terminal pipeline is zero in these runs because the event timing and final order schedule allow all orders to arrive within the horizon. Residual stock nevertheless has no terminal credit or penalty in the objective, so extending the horizon or valuing remaining stock could alter cost comparisons.

6. Discussion

6.1. Detecting exposure is different from improving decisions

The scoring target identifies exposure to an event, not the financial loss from the existing order and not the expected gain from review. A high-volume decision with a modest exposure probability can have greater expected review value than a low-volume decision with a high probability. The logistic ranker includes a volume feature but is not trained to maximize saved cost. Its service and cost effects must be measured directly. The distinction is especially important for phantom-stock events. A corrected record can remain within the fixed exposure label's active period. Routing precision can therefore count an exposed decision whose specific inventory discrepancy has already been corrected. This limits precision and recall as operational performance indicators. Future work should distinguish initial exposure, unresolved state and expected avoidable loss while preserving a valid comparison across policies with different state histories.

6.2. Information delay and reviewer cost

Review creates an opportunity to use information that the forecaster lacks. It also introduces latency and a direct charge. The simulation allows unchanged approval because automatic modification would confound the value of inspection with a forced tendency to alter orders. Under this model, even unchanged approval may be costly when it delays replenishment. Conversely, accurate side information may justify a larger order or an earlier stock correction that increases holding while reducing lost sales. These mechanisms explain why there is no general dominance result for human review. The effect depends on available information, where review is allocated, how quantity adjustments are made and how service is valued relative to stock and labor. An organization should estimate these quantities from its own process before adopting the numerical thresholds or costs in this study.

6.3. What continuing evaluation would need to establish

Weekly outcome tracking can reveal whether a fixed ranker continues to identify useful cases and whether review behavior changes. It does not, on its own, correct selection bias, uncover unobserved demand or estimate the counterfactual outcome of an overridden decision. The proposed operational procedure therefore requires independent audits and explicit evaluation periods alongside score monitoring. A next experiment could compare frozen and periodically refitted scorers under a prespecified change in disruption frequencies. That design should delay labels until the relevant outcome is observable and retain a separate evaluation period. Another extension could train on the difference in simulated cost between review and automation. Such a value-of-review target would better match the decision objective, but its validity would remain limited by the simulator and reviewer assumptions.

7. Limitations

The study uses one retailer's observed sales as a baseline request proxy. Historical inventory availability is not observed, and the imposed events, reviewer knowledge and review delay are synthetic. There are no human participants. The results cannot substantiate claims about actual planner competence, production feasibility or realized savings in retail operations. The count forecast treats recent daily observations as identically distributed and conditionally independent. It omits day-of-week effects, promotion covariates and inventory-censoring correction. Stockouts can suppress subsequent forecasts because the forecast sees fulfilled sales rather than lost requests. This feedback is shared by all policies but remains a substantive modeling limitation. The observed service levels should not be interpreted as validation of a nominal 95% protection quantile.

The reviewer uses a fixed pre-simulation baseline and known event multipliers when informed. Actual planners may have different knowledge, learn over time, disagree, or face correlated workload constraints. The simulation has unlimited supplier capacity,

no minimum order quantities, no pack rounding beyond units and no shared warehouse constraint. A nominal review percentage counts decisions equally even though real decisions require different amounts of labor. Five disruption draws and three reviewer seeds do not cover all relevant environments. Sensitivity runs vary selected parameters separately in one environment and do not estimate their interactions. The finite horizon omits terminal valuation and some delayed costs. The scorer is refitted for each independently generated environment, which assumes representative calibration data; its transfer to an unseen disruption mixture is not tested. Finally, continuous monitoring is implemented as logging and evaluation, while online learning or recalibration benefits are not experimentally demonstrated.

8. Conclusion

Human override should be evaluated as a decision process with information, quantity changes, delay and cost. The proposed framework makes those components explicit and compares constrained review policies under a common forecasting rule and auditable inventory accounting. Risk-based routing identifies more event-exposed decisions than random review, yet that detection advantage does not produce lower average cost at the same budget. Full review performs well under the main cost assumptions but loses its cost advantage when review becomes sufficiently expensive in the examined sensitivity scenario. These findings support measured, context-specific evaluation of review allocation rather than an assumption that more review necessarily improves performance. Real-world validation and an objective based on the expected value of review remain necessary extensions.

Declarations

- Funding: This research received no external funding.
- Competing interests: The author declares no competing interests.
- Data and code availability: M5 is publicly available through the competition associated with [5]. The accompanying replication package provides data retrieval instructions and checksums, sample identifiers, simulation code, event schedules, scorer settings, per-series results and weekly audit summaries. The public data remain subject to their original terms.

References

- [1] N. DeHoratius and A. Raman, "Inventory record inaccuracy: An empirical analysis," *Management Science*, 54(4), 627–641, 2008. doi:10.1287/mnsc.1070.0789.
- [2] A. G. Kök and K. H. Shang, "Inspection and replenishment policies for systems with inventory record inaccuracy," *Manufacturing & Service Operations Management*, 9(2), 185–205, 2007. doi:10.1287/msom.1060.0136.
- [3] K. H. van Donselaar, V. Gaur, T. van Woensel, R. A. C. M. Broekmeulen and J. C. Fransoo, "Ordering behavior in retail stores and implications for automated replenishment," *Management Science*, 56(5), 766–784, 2010. doi:10.1287/mnsc.1090.1141.
- [4] R. Fildes, P. Goodwin, M. Lawrence and K. Nikolopoulos, "Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning," *International Journal of Forecasting*, 25(1), 3–23, 2009. doi:10.1016/j.ijforecast.2008.11.010.
- [5] S. Makridakis, E. Spiliotis and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *International Journal of Forecasting*, 38(4), 1346–1364, 2022. doi:10.1016/j.ijforecast.2021.11.013.
- [6] R. Parasuraman, T. B. Sheridan and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 30(3), 286–297, 2000. doi:10.1109/3468.844354.
- [7] H. Mozannar and D. Sontag, "Consistent estimators for learning to defer to an expert," *Proceedings of Machine Learning Research*, 119, 7076–7087, 2020. <https://proceedings.mlr.press/v119/mozannar20b.html>.
- [8] D. W. Trigg, "Monitoring a forecasting system," *Operational Research Quarterly*, 15(3), 271–274, 1964. doi:10.1057/jors.1964.48.
- [9] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, 46(4), Article 44, 1–37, 2014. doi:10.1145/2523813.
- [10] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo and D. Dennison, "Hidden technical debt in machine learning systems," *Advances in Neural Information Processing Systems*, 28, 2503–2511, 2015.
- [11] E. Breck, S. Cai, E. Nielsen, M. Salib and D. Sculley, "The ML test score: A rubric for ML production readiness and technical debt reduction," *2017 IEEE International Conference on Big Data*, 1123–1132, 2017. doi:10.1109/BigData.2017.8258038.
- [12] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009. doi:10.1007/978-0-387-84858-7.