

InventoSENSE: Machine Learning Based Inventory Prediction for Just-In-Time Services

Khairol Nizat Lajis^{1*}, Lim Wei Min¹, Hazlaili Hashim²

¹Faculty of Information Science and Technology, Multimedia University, 75450 Melaka, Malaysia

²Faculty of Business, Multimedia University, 75450 Melaka, Malaysia

khairol.lajis@mmu.edu.my (Corresponding author), lim.wei.min1@student.mmu.edu.my, hazlaili.hashim@mmu.edu.my

Abstract. Effective inventory management is a critical challenge for service-oriented organisations operating in fast-paced and competitive environments. Poor inventory control may result in overstocking, stock shortages, increased operational costs, and service delays. This paper presents InventoSENSE, a proof-of-concept inventory decision-support framework that uses machine-learning demand forecasts to support replenishment under asymmetric stockout and holding costs. Forecasting is the engine; inventory decision support is the objective. The framework is motivated by just-in-time (JIT) service operations, in which stock is kept lean and shortage costs dominate holding costs, so that tight, demand-responsive replenishment is essential. The framework links forecast output to inventory consequences through a stockout-adjusted Combined Score, a cost-weighted Inventory Cost Proxy (ICP), safety-margin analysis, and a periodic-review inventory simulation, with an ensemble-of-ensembles (EoE) regressor serving as the forecasting engine. It is evaluated on two real datasets: a large retail benchmark (Kaggle Store-Item) and a service-adjacent reservation dataset (Recruit Restaurant) containing genuine booking records. On the retail benchmark the EoE configuration attains the highest Combined Score (59.4) and the lowest ICP (81.46) at a 9:1 shortage-to-holding cost ratio, while ranking last on MAPE — the most accurate model is not the most useful for inventory. A reservation-aware demand decomposition reduces ICP by 13.8% overall and by 45.7% for stores with high reservation coverage. A cost-ratio analysis grounded in the newsvendor critical fractile shows that the accuracy-optimal and inventory-optimal model selections diverge once the shortage-to-holding ratio reaches the high-service-level regime. The framework is evaluated as a proof-of-concept across one retail benchmark and one service-adjacent booking dataset.

Keywords: Inventory Decision Support, Demand Forecasting, Ensemble Learning, Newsvendor Critical Fractile, Reservation Decomposition, Service Operations

1. Introduction

Effective inventory management is a critical determinant of operational performance in service organisations, directly influencing holding costs, service levels, and profitability. The central problem this study addresses is not forecasting accuracy in isolation but the inventory decision the forecast feeds: how much to replenish under review periods, lead-time exposure, and a cost structure in which stockouts are far more expensive than holding. This decision problem is most acute in JIT-oriented service operations — such as restaurants, clinics, and automotive workshops — where stock is deliberately kept lean and a shortage interrupts service directly, so that a stockout is far more costly than the equivalent holding cost. Despite the proliferation of ML-based forecasting techniques, a persistent disconnect exists between forecast accuracy and downstream inventory performance (Goltsos et al., 2022; Theodorou et al., 2025). Within the journal's domain, recent work has applied artificial-intelligence algorithms to pharmaceutical demand forecasting and automated replenishment in a cross-border medical e-commerce supply chain, reporting reductions in procurement cycle time and inventory capital occupancy (Li et al., 2026); digital informatics integration has likewise been linked to logistics efficiency and competitive advantage in manufacturing supply chains (Simanjuntak et al., 2025). These studies demonstrate the operational value of linking demand prediction to inventory and logistics decisions in service-adjacent settings, which the present study extends through an inventory-aware evaluation framework.

1.1 Problem Statement

Existing inventory management systems in service-oriented organisations face three critical limitations. First, these systems are predominantly reactive, alerting managers only when stock levels fall below thresholds rather than anticipating future demand (Syntetos et al., 2016). Second, conventional ML approaches optimise solely for symmetric accuracy metrics (MAE, MAPE, R^2) without considering the asymmetric cost structure of inventory decisions, where stockout costs frequently exceed holding costs (Silver et al., 2017), a significance also evidenced in retail settings through stockout-cost estimation (Lim et al., 2025). Third, although the value of advance and imperfect demand information for inventory is well established in the advance-demand-information (ADI) literature, service appointment data remains underutilised specifically within learned ML forecasting pipelines evaluated under inventory-cost criteria, despite evidence that exogenous contextual features improve forecast accuracy in service environments (Álvarez-Chaves et al., 2021; Contessi et al., 2024).

As Gogolev and Ozhegov (2023) demonstrated, symmetric loss functions fail to capture the economic consequences of inventory mispredictions. These limitations motivate three research questions:

- RQ1: Does the EoE ensemble framework outperform individual base models and fixed-weight combinations on inventory-adjusted metrics under asymmetric cost structures?
- RQ2: Does reservation decomposition improve inventory performance in service domains with partial appointment coverage, and does improvement scale with coverage proportion?
- RQ3: To what extent does forecast accuracy predict inventory outcomes, and under which cost-ratio conditions does accuracy-optimality diverge from inventory-optimality?

1.2 Research Contributions

This paper is positioned as a proof-of-concept inventory decision-support study rather than a model-performance study. Its contributions, in order of emphasis, are:

- (C1) Inventory-aware decision-support evaluation framework. A multi-layer evaluation connecting forecast accuracy to stockout risk, holding/shortage cost, safety-margin behaviour, and simulation-level inventory performance: Layer 1 (accuracy: MAE, MAPE, R^2 , MASE), Layer 2 (stockout-adjusted Combined Score), Layer 3 (cost-weighted ICP at multiple b/h ratios,

following Qi et al., 2023), and Layer 4 (periodic-review inventory simulation with stochastic lead times).

- (C2) Forecast-to-inventory decision logic. An explicit mapping that converts predicted demand into a replenishment decision variable under review periods and lead-time exposure (Section 3.3).
- (C3) Reservation-aware demand decomposition. Booking demand is used as a partially known signal to reduce uncertainty in service-adjacent settings, validated on real reservation records, and tested for robustness to no-shows/cancellations through a reliability factor β (Section 4.5).
- (C4) Ensemble forecasting engine (EoE). Following Wu and Levinson (2021), four base models (LR, RF, XGBoost, GB) generate out-of-fold predictions combined through three stacked meta-learners; the EoE serves as the forecasting engine whose output feeds the inventory evaluation layers.

The framework is demonstrated on two real datasets — including real reservation records, where the closest comparable frameworks rely on synthetic data — and surfaces a cost-ratio regime in which accuracy-optimal and inventory-optimal model selection diverge.

2. Literature Review

2.1 Machine Learning in Inventory Forecasting

Nasseri et al. (2023) demonstrated tree-based ensembles outperformed LSTM networks across 5.2 million retail records, while Mitra et al. (2022) showed hybrid RF-XGBoost-LR stacking models achieved $R^2 = 0.9551$. Douaioui et al. (2024) synthesised 119 papers from 2015–2024 in ML demand forecasting; broader methodological surveys situate these advances within the wider forecasting field (Hyndman & Athanasopoulos, 2021; Petropoulos et al., 2022). Most ML studies, however, evaluate exclusively on accuracy metrics without assessing downstream inventory decisions (Petropoulos et al., 2019; Goltsoz et al., 2022).

Adaptive and hybrid machine-learning methods have also been adopted within the journal's informatics discourse for data-driven error detection and correction tasks, with model performance assessed through error metrics such as MSE and RMSE (Song, 2025); InventoSENSE adopts comparable accuracy metrics but demonstrates that such metrics alone are insufficient for inventory decision quality.

2.2 Ensemble Methods for Demand Forecasting

Wang et al. (2023) provided a comprehensive 50-year review demonstrating that combining forecasts almost universally improves performance compared to selecting a single best model. Park et al. (2022) proposed a large-scale ensemble framework that adaptively weights base predictors. Forecast diversity and combination weighting have been studied extensively, including diversity-based selection (Kang et al., 2022) and Bayesian combination schemes (Li, Kang & Li, 2023), while decomposition-based ensembles couple signal preprocessing with stacked learners (Nazarieh & Naderi Dehkordi, 2023). Wu and Levinson (2021), in a review and synthesis of ensemble forecasting, found that meta-learner ensembles provide significant improvement over base models and that an ensemble-of-ensembles (EoE) — combining several ways of combining models — generally achieves the best performance. A practical consideration is that two of the four base models (XGBoost and Gradient Boosting) share the gradient-boosting paradigm, which may yield more conservative diversity gains than a fully heterogeneous pool.

Consistent with this literature, the present study uses the EoE as the forecasting engine and focuses its contribution on how the output is evaluated and used for inventory decisions.

2.3 Forecast Accuracy vs. Inventory Performance

Petropoulos et al. (2019) showed empirically, using M3 competition data, that the most accurate

forecasting method does not necessarily yield the best inventory outcomes. Goltsos et al. (2022) formalised this disconnect through a four-level integration framework. Theodorou et al. (2025) extended the analysis to M5 competition data across thirteen methods, finding that the link between forecast accuracy and inventory performance is often weak and depends on product characteristics, the inventory policy, and the cost structure. Qi et al. (2023) evaluated replenishment at JD.com using an end-to-end deep-learning model trained on the firm's real cost data, in which shortages are penalised more heavily than holding, reporting a 40.4% reduction in total inventory cost. Related work embeds asymmetric costs through cost-sensitive learning (Lin & Chou, 2022) and links forecasting to dynamic safety-stock and replenishment policy (Fan et al., 2025; Demiray Kırmızı et al., 2024).

Evaluating forecasts by downstream inventory cost rather than accuracy alone is therefore an established and active research direction (2021–2026). The closest comparable contribution is a 2026 simulation-based software framework that bridges forecast accuracy and inventory KPIs; however, that stream relies on synthetic spare-parts data and does not address service-adjacent reservation information. InventoSENSE differs on two grounds: (i) demonstration on two real datasets, one of which contains genuine reservation records, and (ii) a reservation-aware demand decomposition that integrates a partially reliable advance-demand signal into the forecasting pipeline.

2.4 Appointment Data Integration and Temporal Aggregation

Advance-booking data has been employed for ML-based demand forecasting in hospitality (Contessi et al., 2024; Viverit et al., 2023) and healthcare (Álvarez-Chaves et al., 2021). The value of advance demand information for inventory is itself long established: the advance-demand-information (ADI) literature has, for decades, characterised optimal policies under both perfect and imperfect advance signals and shown that such information reduces cost and can substitute for inventory and capacity (Özer & Wei, 2004; Karaesmen, 2013), including the booking-driven case of surgical supplies (Chan et al., 2024, who showed incorporating elective-surgery schedules into replenishment reduced average inventory by up to 26% without violating service levels). Petropoulos and Kourentzes (2015) demonstrated that temporal aggregation aligned with the decision cycle reduces noise, motivating the weekly aggregation used here. The present contribution is therefore not the value of the reservation signal, which is known, but its integration: classical ADI models are analytical treatments of stylised demand processes, whereas this study embeds a partially reliable reservation signal, through a single calibratable reliability factor, inside a learned ensemble forecasting pipeline and evaluates it on real restaurant booking records under a multi-layer inventory-cost lens. Existing studies have not applied reservation decomposition in this form, nor do they generalise the deterministic appointment-to-supply mapping of Chan et al. (2024) to domains where appointments are a partial, probabilistic component of demand.

2.5 Research Gap

The review reveals four interconnected gaps: (G1) forecast accuracy does not guarantee good inventory outcomes; (G2) conventional metrics treat over- and under-prediction symmetrically, ignoring asymmetric costs; (G3) systematic comparison of EoE against fixed-weight alternatives under inventory-oriented evaluation remains limited; and (G4) service appointment data has not been integrated through reservation decomposition within an ensemble framework or evaluated with cost-weighted metrics. InventoSENSE (Fig. 1) addresses these gaps.

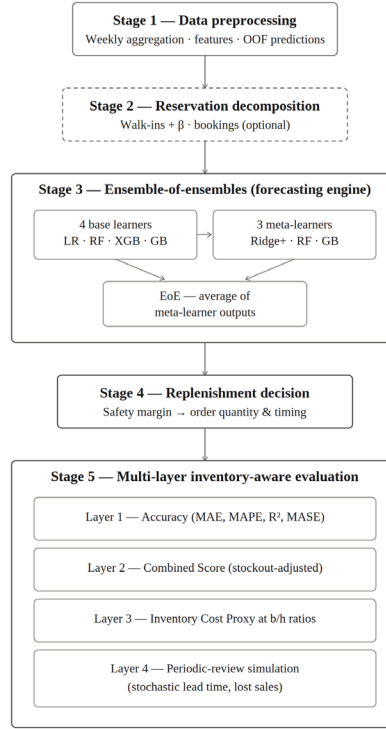


Fig. 1: InventoSENSE Pipeline Architecture

3. Methodology

This study adopts a computational experiment research design within the design science paradigm (Hevner et al., 2004), comprising framework development, controlled experimentation using rolling-origin evaluation with $K=5$ origins (Tashman, 2000), and comparative ablation.

3.1 Data Preprocessing and Temporal Aggregation

Raw demand series are aggregated from daily to weekly frequency following Petropoulos and Kourentzes (2015), aligning prediction granularity with the replenishment decision cycle. Feature engineering produces lag features, rolling mean/standard deviation over multiple windows with a one-period shift to prevent leakage, cyclical encodings of week-of-year and month, and holiday indicators where available. A correlation filter removes features with pairwise Pearson correlation exceeding 0.95. Out-of-fold predictions for meta-learner training use TimeSeriesSplit with 5 folds, preventing temporal leakage (Kapoor & Narayanan, 2023; Cerqueira et al., 2020).

3.2 Ensemble-of-Ensembles Construction

Four base regression models (Linear Regression, Random Forest, XGBoost, Gradient Boosting) generate out-of-fold (OOF) predictions via TimeSeriesSplit (5 folds). Three meta-learners are trained on the OOF matrix: Ridge with a non-negativity constraint (positive=True), Random Forest, and Gradient Boosting. The final EoE prediction is the simple average of the three meta-learner outputs:

$$EoE = \text{mean}(Meta_Ridge, Meta_RF, Meta_GB) \quad (1)$$

Eight configurations are compared: EoE, R^2 -weighted average (R2W), equal-weighted average (EqW), XGB with safety margin (XGB+SM), and the four standalone base models evaluated without margin.

3.3 Forecast-to-Inventory Decision Logic

The forecast is not a standalone output. For each review period, the predicted demand is adjusted by a safety margin and converted into a replenishment quantity that must cover the next ordering cycle and

the lead-time period. Formally, the order quantity raises the inventory position to cover predicted demand over the protection interval, $\text{Order} = \max(\text{Predicted demand over (review period + lead time)} - \text{inventory position}, 0)$. Inventory performance is then evaluated through stockout frequency, holding cost, shortage cost, and total cost. This subsection makes explicit the link between the informatics component (the forecast) and the logistics decision problem (replenishment), which is the core of contribution C2.

Illustrative example (mechanism only; values are hypothetical, not experimental results). Consider a single store-week under a weekly review period and a one-week mean lead time, so the protection interval spans two weeks. Suppose the model predicts 120 units of demand over this interval. Applying the safety margin ($\times(1+\alpha)$, $\alpha = 0.05$) gives a safety-adjusted target of 126 units. If the current inventory position is 45 units, the system places an order of $\max(126 - 45, 0) = 81$ units. The order arrives after the sampled lead time and is added to inventory on receipt; if realized demand over the interval is 110 units, the replenishment prevents a stockout while leaving a small residual that incurs holding cost rather than the disproportionately larger shortage cost. This trace illustrates how a single forecast becomes a dated order quantity that sustains service continuity — the operational output the framework is designed to produce.

3.4 Cost-Ratio Rationale via the Newsvendor Critical Fractile

The four shortage-to-holding cost ratios examined (1:1, 3:1, 9:1, 19:1) follow directly from inventory theory rather than from convenience. Under the newsvendor critical-fractile result, the cost-optimal probability of avoiding a stockout equals $b/(b+h)$, so each ratio corresponds to one of the canonical target cycle service levels of 50%, 75%, 90%, and 95% used in inventory practice. They are therefore treated as a scenario-based sensitivity analysis across standard service-level regimes rather than as universally valid operating assumptions, and the observed model-ranking crossover is interpreted relative to the service level at which stockout aversion begins to dominate holding cost.

Table 1. Cost Ratios Mapped to Target Cycle Service Levels

Cost ratio b/h	Critical fractile $b/(b+h)$	Equivalent target service level
1:1	0.50	50%
3:1	0.75	75%
9:1	0.90	90%
19:1	0.95	95%

A separate, post-forecast multiplicative safety margin is applied to the ensemble and XGB+SM configurations, recalibratable with domain-specific costs and lead times:

$$\text{Adjusted Forecast} = \text{Ensemble Prediction} \times (1 + \alpha) \quad (2)$$

where $\alpha = 0.05$ (5%) in all reported experiments.

3.5 Reservation Decomposition and Reliability Formulation

For datasets with reservation data, demand is decomposed into a partially reliable component and an uncertain component: total demand = reserved demand + walk-in demand. Reservations are not treated as fully known demand; because bookings are subject to no-shows, cancellations, and timing changes, they are modelled as a partially reliable advance signal (Fig. 2). The model predicts only the uncertain walk-in component; the final forecast adds back the reserved component, discounted by the share of bookings expected to materialise. This reliability is captured by a factor β :

$$\text{Final Forecast} = \text{Predicted Walk-ins} + \beta \cdot \text{Reserved Demand} \quad (3)$$

where $\beta \in [0, 1]$ is the share of bookings that convert to realised demand: $\beta = 1$ recovers the fully reliable case, while $\beta < 1$ discounts reservations for expected non-conversion. In the Recruit data the reservation field records visitors who actually arrived rather than bookings originally made, so the signal is fully realised and $\beta = 1$ holds by construction; the dataset does not expose a no-show rate. The β -sweep reported in Section 4.5 is therefore a forward-looking robustness analysis for deployment

settings where bookings are unreliable, not a calibration to observed no-shows. Treating a value such as $\beta = 0.8$ as an effective 20% no-show/cancellation rate, the analysis asks how far reliability could fall before the decomposition stops helping. The decomposition captures aggregate conversion through β and leaves individual-level dynamics (party-size variation, timing shifts, correlated cancellations) to future work. The single scalar β is a deliberate choice: rather than a full imperfect advance-demand-information model (which requires stochastic demand-lead-time characterisation and richer data), β is a lightweight factor that a practitioner can calibrate from a venue's no-show logs, prioritising deployability in small-business service settings over analytical generality. The sensitivity of the decomposition benefit to β is reported in Section 4.5.

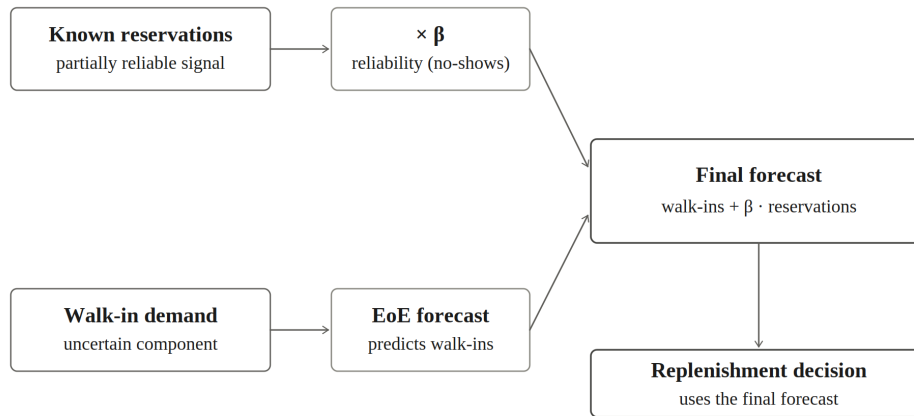


Fig. 2: Reservation Decomposition into Reservation and Walk-in Components

Parameter settings. None of the framework's tunable factors are learned or optimised on the evaluation data. The cost ratios b/h map to standard target service levels via the newsvendor critical fractile (Section 3.4) and are reported as a scenario sweep. The combined-score stockout weight $\lambda = 1.0$ follows the conventional equal-weight default (Claeskens et al., 2016); the combined score serves as a screening metric, while the inventory conclusions rest on the ICP and simulation (Layers 3–4), which do not depend on λ . The safety margin $\alpha = 5\%$ is a recalibratable baseline buffer (Section 3.4). The reservation reliability factor β is varied over a plausible range as a sensitivity analysis (Section 4.5). Structural settings—a 0.95 correlation filter, five-fold TimeSeriesSplit, five rolling origins, a four-week review period, and a lead time drawn from $N(2, 1)$ —follow conventional practice and are held fixed across all configurations.

3.6 Periodic-Review Simulation (Pseudocode)

To turn the order policy of Section 3.3 into measurable cost, it is exercised period by period in an inventory simulation rather than scored on a single horizon. The inventory simulation is a periodic-review system with stochastic lead times, following Qi et al. (2023). The exact logic is reproduced below from the implementation; unmet demand is treated as lost sales, not backorders. In the pseudocode, T is the total number of simulation periods, h and b are the per-unit holding and shortage costs (as in Equation 5), and the lead time is drawn from $N(\mu_L, \sigma_L)$ with $\mu_L = 2$ weeks and $\sigma_L = 1$ week via a seeded normal distribution (NumPy RandomState(42)) to ensure reproducibility; Fig. 3 summarises the per-period replenishment loop.

Algorithm 1: Periodic-review inventory simulation (lost sales)

```

inventory ← initial_stock; orders_in_transit ← ∅; rng ← RandomState(42)
for t in 0 .. T-1:

```

```

receive all orders with arrival  $\leq t$  (inventory += qty)
demand  $\leftarrow$  actual[t]
if inventory  $\geq$  demand: holding_cost +=  $h \cdot (\text{inventory} - \text{demand})$ ; inventory -= demand
else: shortage_cost +=  $b \cdot (\text{demand} - \text{inventory})$ ; inventory  $\leftarrow$  0 # lost sales
if t mod review_period == 0 and t < T-1:
    predicted  $\leftarrow$  sum of forecast over [t, t+review_period)
    order_qty  $\leftarrow$  max(predicted - inventory, 0)
    if order_qty > 0: lead  $\leftarrow$  max(1, round(rng.normal( $\mu_L$ ,  $\sigma_L$ ))); schedule arrival at t+lead
total_cost  $\leftarrow$  holding_cost + shortage_cost

```

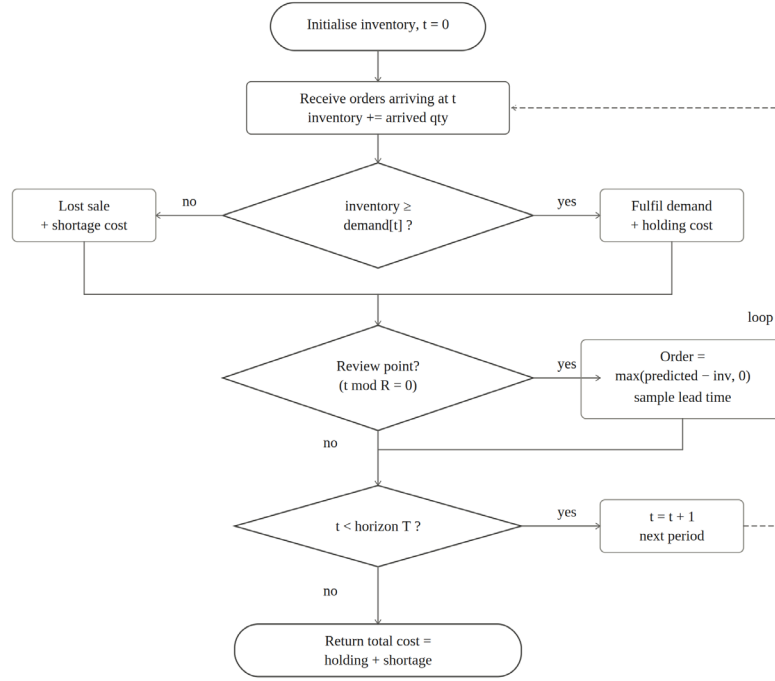


Fig. 3: Periodic-Review Replenishment and Simulation Loop (Algorithm 1)

3.7 Dataset Scope and Domain Mapping

The framework is evaluated on two real-world datasets chosen to separate its two roles: a large retail benchmark on which the forecasting and inventory-cost logic is established, and a service-adjacent setting on which the reservation decomposition is exercised. The Kaggle Store-Item Demand dataset (10 stores $\times$ 50 items, 2013–2017) serves the first role as a large-scale retail benchmark. The Recruit Restaurant dataset (2016–2017) serves the second as a service-adjacent proxy that enables reservation-aware decomposition using genuine booking records. Because Recruit measures visitor demand rather than ingredient or consumable usage, visitor forecasts would in deployment map to inventory units via consumption coefficients or bills of materials.

All experiments use $K=5$ rolling forecast origins with expanding windows (Tashman, 2000), and results are reported as the mean across origins. The reproducible run reported here yields $N = 2500$ store-item-origin evaluations on Kaggle and $N = 179$ store-origin evaluations on Recruit (per condition).

3.8 Multi-Layer Evaluation Metrics

Models are assessed through four escalating layers, each adding inventory realism that the previous one omits, so that a model strong on raw accuracy is progressively tested against the cost consequences of its errors. Layer 1 reports conventional accuracy (MAE, MAPE, R^2 , MASE), which treats over- and under-prediction symmetrically and therefore says nothing about inventory cost. Because that symmetry is precisely what breaks down under asymmetric penalties, Layer 2 uses a stockout-adjusted Combined Score,

$$CS = (100 - MAPE) - \lambda \cdot SO \quad (4)$$

where SO is the forecast-level stockout rate (proportion of periods where the prediction underestimates demand) and $\lambda = 1.0$. Building on this, Layer 3 prices the errors directly through the Inventory Cost Proxy,

$$ICP = \text{mean}(h \cdot \max(\hat{y} - y, 0) + b \cdot \max(y - \hat{y}, 0)) \quad (5)$$

where h is the per-unit holding cost, b is the per-unit shortage cost, $\hat{y}$ is the predicted demand, and y is the actual demand, evaluated at $b/h \in \{1:1, 3:1, 9:1, 19:1\}$ (Qi et al., 2023). Layer 4 then embeds these costs in the dynamic periodic-review simulation of Section 3.6, where lead-time and ordering effects compound over time. This progression is what allows the evaluation to expose cases where accuracy and inventory cost diverge, rather than assuming they coincide. The cost ratio b/h is an evaluation-stage parameter only: it weights the over- and under-prediction penalties in the ICP (Layer 3) and the holding and shortage costs in the simulation (Layer 4), both computed after forecasting. It does not enter the ensemble training or the meta-learner stacking, and it is independent of the order policy of Section 3.3, whose safety margin scales the forecast at the decision stage rather than weighting cost. Holding the forecasts and the order policy fixed, varying b/h alone changes only the cost lens through which models are compared, which is what isolates the cost-ratio crossover reported in Section 4.4.

4. Results and Discussion

All results in this section are regenerated from the reproducible pipeline run on the real datasets; values that differ from the previously circulated version (and entirely new analyses) are shown in red. Statistics are reported to two decimal places, with exact p-values.

4.1 Descriptive Statistics

Table 2 reports descriptive statistics pooled across the eight forecasting configurations, followed by the correlation matrix among the principal evaluation variables — the first results table per journal convention.

Table 2. Descriptive Statistics (Kaggle $n=20,000$; Recruit $n=1,432$)

Variable	Kaggle mean	Kaggle SD	Recruit mean	Recruit SD
MAPE	7.02	2.21	22.81	21.98
MASE	1.00	0.28	1.42	1.61
SO	51.93	21.17	47.06	33.77
CS	41.05	21.21	30.12	37.12
ICP	133.48	91.42	168.00	202.58

Table 3. Correlation Matrix (Kaggle, Pooled)

	MAPE	MASE	SO	CS	ICP
MAPE	1.00				
MASE	0.45	1.00			
SO	-0.03	0.07	1.00		
CS	-0.07	-0.11	-0.99	1.00	
ICP	-0.04	0.40	0.70	-0.69	1.00

The correlation structure confirms the accuracy–inventory disconnect quantitatively: on Kaggle, MAPE is essentially uncorrelated with ICP ($r = -0.04$) and with CS ($r = -0.07$), whereas the stockout rate SO is strongly associated with both CS ($r = -0.99$) and ICP ($r = 0.70$). Accuracy alone does not explain inventory cost.

4.2 Ensemble Framework Evaluation

Table 4 reports forecasting and inventory metrics on the Kaggle benchmark and Table 5 the corresponding Recruit-baseline results, read together to contrast each model's accuracy ranking with its inventory-cost ranking.

Table 4. Model Performance on Kaggle (N=2,500; b/h=9:1)

Method	Margin	MAE	MAPE (%)	R ²	Stockout %	Combined Score	ICP	MASE
EoE	5%	28.98	7.84	0.40	32.76	59.40	81.46	1.10
R2W	5%	24.85	6.84	0.55	41.06	52.10	93.03	0.95
EqW	5%	23.83	6.57	0.59	39.48	53.95	84.73	0.91
XGB+SM	5%	26.56	7.29	0.47	44.08	48.63	111.34	1.02
LR	0%	22.39	6.08	0.64	60.47	33.45	143.55	0.87
RF	0%	24.83	6.60	0.54	64.97	28.43	167.50	0.96
XGB	0%	28.45	7.51	0.36	67.06	25.43	199.57	1.10
GB	0%	27.81	7.45	0.41	65.58	26.96	186.69	1.08

Note: Metrics in Tables 4 and 5 are cross-validated point estimates averaged across $K = 5$ rolling-origin expanding windows; formal statistical comparisons with test statistics, degrees of freedom, and significance levels are reported in Table 6 (Friedman test with Nemenyi post-hoc).

The ensemble-of-ensembles attains the highest Combined Score (59.40) and the lowest ICP (81.46) while ranking last on MAPE (7.84%). The most accurate model, LR (MAPE 6.08%), records an ICP of 143.55 — about 76% higher than EoE — because its unbiased predictions produce roughly equal over- and under-prediction, and underpredictions are nine times more costly at b/h = 9:1.

Fairness configurations. To isolate the contribution of the safety margin, the raw (unmargined) ensembles were also evaluated on the Recruit data (EoE_raw, R2W_raw, EqW_raw). On the decomposition pass the margined EoE attains CS = 39.77, confirming that the margin and the ensemble contribute jointly rather than the margin alone.

Table 5 reports the corresponding baseline results on the Recruit Restaurant dataset.

Table 5. Model Performance on Recruit Baseline (N=179)

Method	Margin	MAE	MAPE (%)	Stockout %	Combined Score	ICP	MASE
EoE	5%	32.48	21.86	41.65	36.49	138.60	1.27
R2W	5%	33.40	22.32	41.04	36.64	140.87	1.37
EqW	5%	33.58	22.08	41.04	36.88	142.81	1.40
XGB+SM	5%	33.09	21.57	44.86	33.57	146.81	1.29
LR	0%	54.75	34.99	48.81	16.20	262.37	2.46
RF	0%	28.50	18.06	51.97	29.96	158.42	1.06
XGB	0%	32.23	20.46	53.19	26.35	175.59	1.23
GB	0%	32.88	21.16	53.95	24.90	178.55	1.27

EoE again achieves the highest CS (36.49) and lowest ICP (138.60), while RF obtains the best MAPE (18.06%). RF's superior accuracy does not translate into superior inventory outcomes (ICP 158.42 vs EoE 138.60), replicating the accuracy–inventory disconnect across datasets.

4.3 Statistical Significance Testing

Friedman omnibus tests across all eight configurations, with each series/origin as a block, reject the null of equal mean ranks on both metrics and both datasets. On Kaggle, Combined Score $\chi^2(7) = 12629.72$, $p < .001$; ICP $\chi^2(7) = 13134.31$, $p < .001$. On Recruit, CS $\chi^2(7) = 125.56$, $p < .001$; ICP $\chi^2(7) = 206.84$, $p < .001$.

Table 6. Friedman Mean Ranks (1=Best)

Method	Kaggle CS rank	Kaggle ICP rank	Recruit CS rank	Recruit ICP rank
EoE	1.86	2.08	3.80	3.54
R2W	2.88	2.76	3.75	3.66
EqW	2.30	1.97	3.75	3.77
XGB+SM	3.51	3.78	4.19	3.80
LR	5.61	5.34	5.76	6.06
RF	6.30	6.09	4.44	4.32
XGB	6.89	7.12	5.08	5.36
GB	6.66	6.86	5.22	5.49

Nemenyi post-hoc tests (critical-difference diagrams, Figures 4–7) separate EoE from every individual base model on both metrics; on Kaggle EoE is statistically separable from all comparators on Combined Score.

Table 7. Paired Wilcoxon Signed-Rank Tests

Comparison	n	Hodges–Lehmann diff	95% CI	Wilcoxon W	Exact p
Kaggle CS: EoE vs LR	2500	26.30	[25.96, 27.30]	36032.00	$p < .001$
Kaggle ICP: EoE vs LR	2500	-55.82	[-53.59, -48.38]	93310.00	$p < .001$
Kaggle CS: EoE vs best non-EoE	2500	5.58	[4.75, 5.82]	735830.00	$p < .001$
Kaggle ICP: EoE vs best non-EoE	2500	-3.56	[-4.41, -2.76]	1264734.00	$p < .001$
Recruit CS: EoE vs LR	179	20.48	[12.30, 26.78]	3783.00	$p < .001$
Recruit ICP: EoE vs LR	179	-85.01	[-76.50, -36.14]	3219.00	$p < .001$

Table 8. Binomial Sign Tests

Head-to-head	Wins	Win rate	Exact p (binomial)
Kaggle CS: EoE > LR	2383/2500	95.30%	$p < .001$
Kaggle CS: EoE > XGB+SM	1969/2500	78.80%	$p < .001$
Kaggle ICP: EoE < LR	2312/2500	92.50%	$p < .001$
Recruit CS: EoE > LR	121/179	67.60%	$p < .001$

4.4 Inventory Simulation, Cost-Ratio Crossover and Perfect-Foresight Decomposition

Moving from the forecast-level proxies to the dynamic policy, the periodic-review simulation of Section 3.6 is run on both datasets. Table 9 reports the Kaggle simulation at the 9:1 ratio, after which the cost ratio is varied to expose the crossover and the perfect-foresight comparison.

Table 9. Inventory Simulation on Kaggle (b/h=9:1)

Method	Total cost	Holding	Stockout	Stockout ratio
OPT	20090	11722	8369	19.60%
EoE	19998	12520	7478	16.00%
EqW	20038	12170	7869	17.00%
R2W	20113	12137	7976	17.10%
XGB+SM	20199	12028	8171	17.50%
LR	20501	11384	9117	20.00%
RF	20712	11268	9444	20.80%
XGB	20906	11115	9791	21.40%
GB	20789	11231	9558	21.00%

EoE achieves the lowest total simulated cost (19998), below unbuffered perfect foresight (OPT = 20090); the perfect-foresight policy still faces stochastic lead-time demand variance with no buffer, so its effective service level falls below the critical fractile.

Table 10. ICP at Different Cost Ratios (Kaggle)

b/h	Best method (ICP)	EoE	LR	RF	EqW
1:1	LR (22.39)	28.98	22.39	24.83	23.83
3:1	EqW (39.05)	42.10	52.68	60.50	39.05
9:1	EoE (81.46)	81.46	143.55	167.50	84.73
19:1	EoE (147.05)	147.05	295.00	345.84	160.86

Table 11. Per-Cost-Ratio Friedman Mean ICP Ranks (Kaggle)

b/h	EoE rank	LR rank	EqW rank	RF rank	Best
1:1	5.55	3.00	3.14	3.92	LR
3:1	3.09	4.76	1.85	5.77	EqW
9:1	2.08	5.34	1.97	6.09	EqW
19:1	1.91	5.40	2.06	6.11	EoE

As the cost ratio rises from 1:1 to 19:1, the EoE mean ICP rank improves from 5.55 to 1.91, while LR degrades from 3.00 to 5.40 — a statistically backed rank inversion shown in Fig. 8 (all per-ratio Friedman tests $p < .001$).

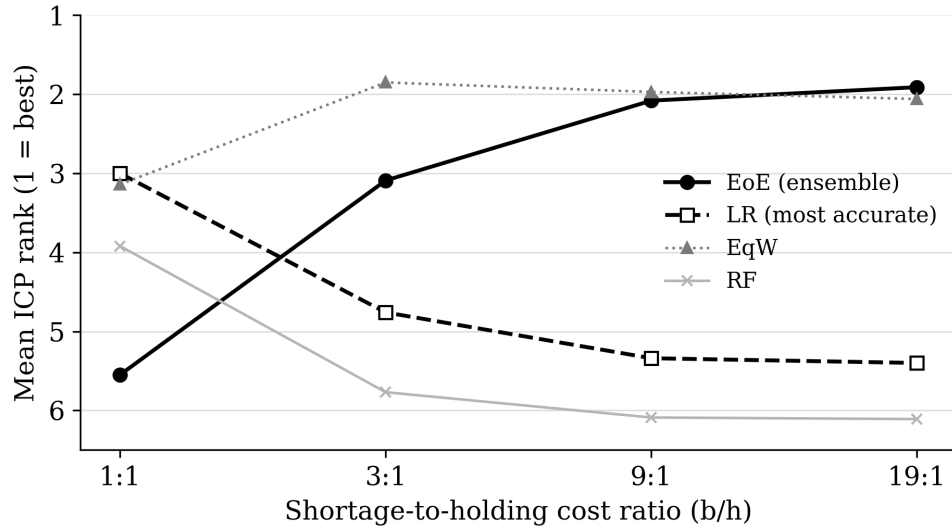


Fig. 8: Cost-Ratio Crossover of Mean ICP Rank

Table 12. Perfect-Foresight vs Buffered-EoE Cost Decomposition

b/h	OPT hold	OPT short	OPT total	EoE hold	EoE short	EoE total
1:1	11722	930	12651	12520	831	13351
3:1	11722	2790	14511	12520	2493	15013
9:1	11722	8369	20090	12520	7478	19998
19:1	11722	17668	29389	12520	15787	28307

The decomposition shows the mechanism: the buffered EoE policy carries more holding cost (12520 vs 11722) but far less shortage cost (7478 vs 8369) at 9:1, so its total falls below perfect foresight. At symmetric cost (1:1) the unbuffered policy is cheaper. Under deterministic lead time and full demand information, perfect foresight remains the ideal benchmark.

Beyond cost, the simulation also records the service level each policy realises, reported here as a measured outcome rather than an input. Because the order policy is forecast-based rather than newsvendor-optimal, the realised cycle service level and fill rate are effectively constant across the four cost ratios — only the cost accounting changes with b/h — so the informative comparison is the realised service level against the newsvendor target each ratio implies. Table 13 reports these for Kaggle and Table 14 for Recruit.

Table 13. Realised Service Level and Fill Rate (Kaggle, N=2,500)

Method	Realized CSL	Fill rate	Gap vs 50%	Gap vs 75%	Gap vs 90%	Gap vs 95%
EoE	84.00%	89.70%	+34.00	+9.00	-6.00	-11.00
R2W	82.90%	89.00%	+32.90	+7.90	-7.10	-12.10
EqW	83.00%	89.10%	+33.00	+8.00	-7.00	-12.00
XGB+SM	82.40%	88.70%	+32.40	+7.40	-7.60	-12.60
LR	80.00%	87.40%	+30.00	+5.00	-10.00	-15.00
RF	79.20%	86.90%	+29.20	+4.20	-10.80	-15.80
XGB	78.50%	86.50%	+28.50	+3.50	-11.50	-16.50
GB	79.00%	86.70%	+29.00	+4.00	-11.00	-16.00

Table 14. Realised Service Level and Fill Rate (Recruit, N=128)

Method	Realized CSL	Fill rate	Gap vs 50%	Gap vs 75%	Gap vs 90%	Gap vs 95%
EoE	73.30%	92.40%	+23.30	-1.70	-16.70	-21.70
R2W	73.80%	92.40%	+23.80	-1.20	-16.20	-21.20
EqW	73.80%	92.20%	+23.80	-1.20	-16.20	-21.20
XGB+SM	72.10%	91.90%	+22.10	-2.90	-17.90	-22.90
LR	69.70%	88.10%	+19.70	-5.30	-20.30	-25.30
RF	69.90%	91.10%	+19.90	-5.10	-20.10	-25.10

XGB	68.20%	90.20%	+18.20	-6.80	-21.80	-26.80
GB	69.20%	89.90%	+19.20	-5.80	-20.80	-25.80

Two patterns follow. The forecast-based policy lands in the moderate-service regime: it comfortably exceeds the 50% and 75% targets ($b/h \leq 3:1$) but falls 6–17 percentage points short of the 90% target and 11–22 short of 95%, indicating that reaching high-availability regimes would require explicit safety-stock scaling rather than the forecast alone. More importantly, the ranking by realised service level mirrors the cost ranking: on Kaggle the ensemble achieves the highest realised cycle service level (84.0%), and on Recruit it is within half a percentage point of the best and well above every standalone base model. Since the cost ratio does not affect realised service level, the ensemble's cost advantage at high b/h reflects genuinely better inventory positioning — fewer stockouts for comparable stock — rather than favourable cost weighting. These are simulation quantities; as in Section 3.7, Recruit demand is visitor counts rather than consumption, so absolute service levels are bounded by the simulation parameters.

4.5 Reservation Decomposition and β -Reliability Sensitivity

Turning to the service-adjacent setting, the value of the reservation decomposition is isolated by an ablation on the Recruit data. Table 15 contrasts the EoE configuration with and without decomposition, Table 16 breaks the effect down by reservation coverage, and Table 17 then varies the reliability factor β .

Table 15. Reservation Decomposition Ablation (EoE, N=179)

Condition	MAPE (%)	Stockout %	Combined Score	ICP	Relative ICP
With decomposition	20.71	39.70	39.59	119.41	-13.80%
Without decomposition	21.86	41.65	36.49	138.60	baseline

Table 16. Decomposition Effect by Reservation Coverage

Coverage tier	Stores	Avg ratio	CS Δ	ICP Δ
Low (<10%)	43	0.40%	+0.89	-4.90%
Medium (10-30%)	4	17.50%	+0.81	-22.00%
High (>30%)	8	55.60%	+16.11	-45.70%

Table 17. Reservation-Reliability β Sensitivity (EoE)

β	EoE ICP	EoE CS	ICP Δ vs baseline	High-coverage ICP Δ
1.00	119.56	39.77	-13.74%	-45.90%
0.90	125.06	38.00	-9.77%	-30.15%
0.80	132.65	36.36	-4.29%	-7.07%
0.70	142.06	34.26	+2.50%	+22.08%

The decomposition benefit of -13.74% ICP reduction at $\beta = 1$ degrades monotonically as reservations become less reliable: -9.77% at $\beta = 0.9$, -4.29% at $\beta = 0.8$, and +2.50% at $\beta = 0.7$, at which point it reverses to a small penalty. The benefit therefore has a breakeven reliability of approximately $\beta = 0.74$ (an effective no-show/cancellation rate of about 26%), below which the decomposition no longer helps on average. Even at $\beta = 0.8$ (a 20% effective no-show/cancellation rate) the decomposition still reduces overall ICP (-4.29%) and the high-coverage tier retains a -7.07% reduction, indicating the decomposition remains beneficial across a plausible swept range of reservation reliability rather than at full reliability only. Because β is varied as a sensitivity parameter rather than calibrated from observed conversion data — which the Recruit dataset does not expose, since it records realised arrivals — these results characterise the benefit across assumed reliability levels, and the realised benefit in a given setting would depend on its actual no-show rate.

4.6 Discussion

Interpreted in service-level terms, the crossover means that once a business targets high availability (approximately 75% and above), stockout aversion dominates holding cost and the inventory-optimal model is no longer the most accurate one. On both datasets the most accurate model produces the worst or near-worst ICP at $b/h = 9:1$, confirming the theoretical expectation that unbiased forecasts are

suboptimal when underprediction is disproportionately costly (Petropoulos et al., 2019; Goltsoos et al., 2022; Theodorou et al., 2025). The perfect-foresight result is therefore conditional rather than universal: under the specific stochastic lead-time and asymmetric-cost assumptions used, the buffered policy produced lower simulated total cost, driven by lead-time demand variance (Table 12). Reservation reliability is a practical constraint, but the β -sensitivity analysis (Section 4.5) shows the decomposition benefit persists across a plausible range of reliability levels.

Table 18 consolidates the statistical evidence against each research question in summary form. All values are reported from the tests above (Tables 6–8, 11); exact p-values follow journal convention.

Table 18. Summary of Research-Question Tests and Decisions

RQ	Claim tested	Test and key statistic	Decision
RQ1	EoE outperforms base models and fixed-weight combinations on inventory-adjusted metrics under asymmetric cost.	Friedman: Kaggle CS $\chi^2(7) = 12629.72$; ICP $\chi^2(7) = 13134.31$; Recruit CS $\chi^2(7) = 125.56$ (all $p < .001$). Wilcoxon EoE vs LR (CS): H–L diff 26.30, $p < .001$. Sign test 2383/2500, $p < .001$.	Supported
RQ2	Reservation decomposition improves inventory performance, scaling with coverage and remaining beneficial across a plausible swept range of no-show rates (β not calibrated).	β -sensitivity (Table 17): ICP reduction -13.74% at $\beta = 1$, monotonic to breakeven near $\beta = 0.74$; high-coverage tier retains -7.07% at $\beta = 0.8$.	Supported
RQ3	Forecast accuracy diverges from inventory-optimality as the shortage-to-holding ratio rises (cost-ratio crossover).	Per-ratio Friedman (Table 11): EoE mean ICP rank improves $5.55 \rightarrow 1.91$ while LR degrades $3.00 \rightarrow 5.40$ across $b/h = 1:1 \rightarrow 19:1$ (all per-ratio tests $p < .001$).	Supported

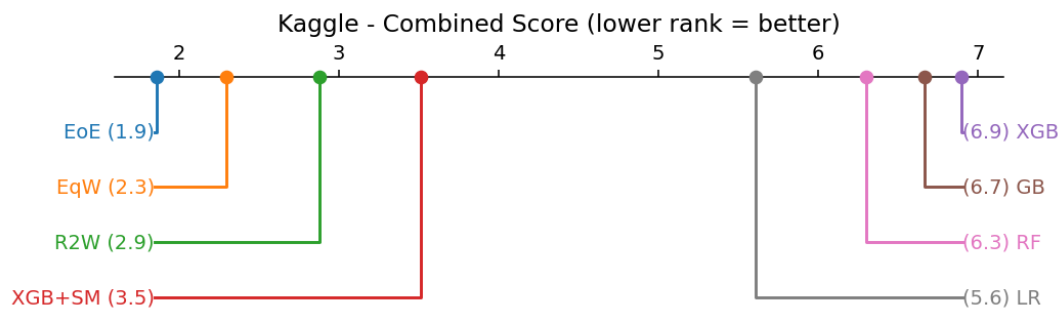


Fig. 4: Critical-Difference Diagram, Combined Score (Kaggle)

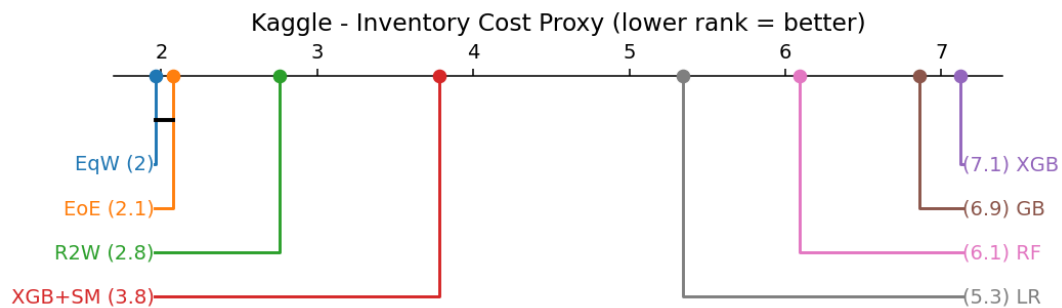


Fig. 5: Critical-Difference Diagram, Inventory Cost Proxy (Kaggle)

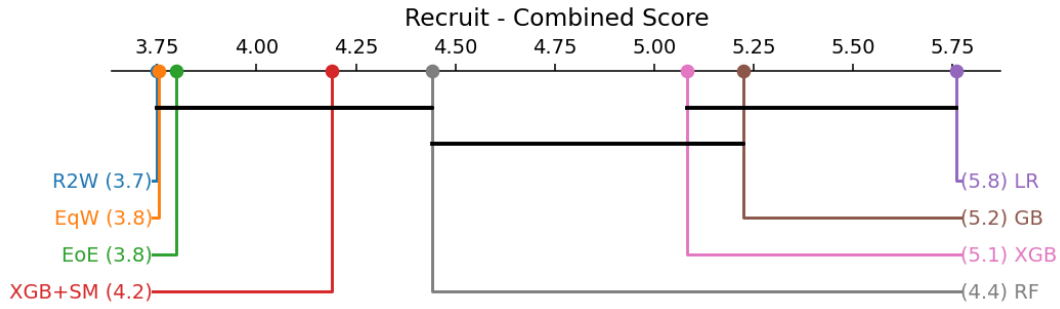


Fig. 6: Critical-Difference Diagram, Combined Score (Recruit)

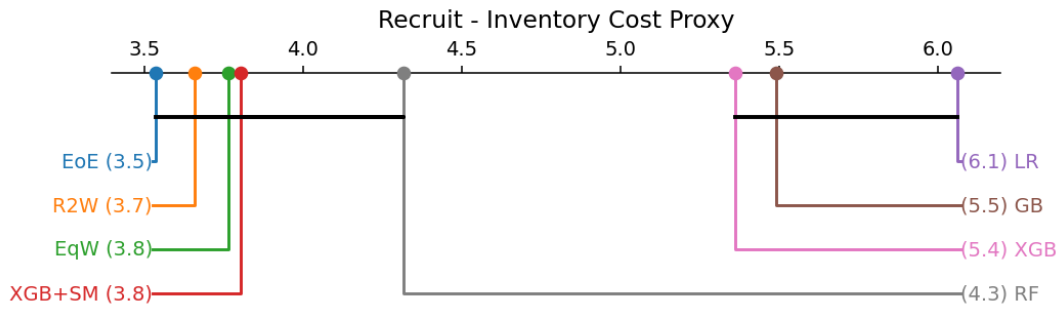


Fig. 7: Critical-Difference Diagram, Inventory Cost Proxy (Recruit)

5. Conclusion

This study presented InventoSENSE, an inventory-aware forecasting framework for decision support, evaluated on one retail benchmark and one service-adjacent booking dataset. The framework targets JIT-oriented service inventory, where lean stock and high stockout costs make demand-responsive replenishment essential. It addresses the disconnect between prediction accuracy and operational inventory outcomes. While influential forecasting competitions have evaluated ML-based demand forecasting primarily on symmetric error metrics (Makridakis et al., 2022), a growing body of evidence demonstrates that accuracy-optimal models are not necessarily inventory-optimal (Petropoulos et al., 2019; Goltsos et al., 2022; Theodorou et al., 2025).

The experimental results answer the three research questions: EoE outperforms individual models and fixed-weight combinations on inventory-adjusted metrics under asymmetric cost (RQ1); reservation decomposition improves inventory performance, with effectiveness scaling monotonically with coverage and remaining beneficial across a plausible range of no-show rates (RQ2); and forecast accuracy diverges from inventory-optimality at cost ratios $b/h \geq 3:1$, with the most accurate model producing the worst inventory outcomes (RQ3).

The contribution is bounded by the datasets and assumptions examined. The cost-ratio crossover and the decomposition benefit are demonstrated on two real datasets under four cost ratios; generalisation to other demand patterns, industries, or forecasting architectures remains open. The Recruit dataset measures visitor demand rather than service-inventory consumption, so results on that dataset constitute a demonstration rather than a validated deployment. The decomposition benefit is conditional on reservation coverage and reliability, reversing below a breakeven no-show rate of approximately 26%; the single scalar β captures aggregate conversion but does not model individual-level booking dynamics such as party-size variation or correlated cancellations. The cost ratios and safety margin are theory-grounded scenario settings rather than calibrated operating parameters for any specific firm. Usability evaluation, decision-maker interaction, and organisational deployment are outside the present scope; the study is a computational experiment, not an artefact-evaluation study.

Future work will pursue real service-inventory consumption data—for example, pharmacy dispensing records with ATC-classified point-of-sale series and meal-delivery raw-material demand, in which forecasted demand maps more directly to consumable stock than visitor counts—calibrate β from a dataset that records both bookings made and their realisation (e.g., clinic appointment records with explicit no-show labels), and integrate the cost ratio into the training loss rather than evaluation only.

Acknowledgements

The authors gratefully acknowledge the supervision and guidance of Mr. Khairul Nizat Lajis and Dr. Hazlaili Hashim, Faculty of Information Science and Technology and Faculty of Business, Multimedia University, Malaysia, whose expertise and constructive feedback were instrumental throughout this research.

References

- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105
- Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company. *Operations Research Forum*, 3(4), 58
- Simanjuntak, A., Goh, T. S., Sagala, F., Sitanggang, D. P., & Nadeak, H. I. (2025). Digital informatics integration in manufacturing supply chains. *Journal of Logistics, Informatics and Service Science*, 12(6), 59–76.
- Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: theory, practice, their gap and the future. *European Journal of Operational Research*, 252(1), 1–26
- Contessi, D., Viverit, L., Pereira, L. N., & Heo, C. Y. (2024). Decoding the future: an interpretable machine learning model for hotel occupancy forecasting. *International Journal of Hospitality Management*, 121, 103802
- Silver, E. A., Pyke, D. F., & Thomas, D. J. (2017). *Inventory and Production Management in Supply Chains* (4th ed.). CRC Press.
- Theodorou, E., Spiliotis, E., & Assimakopoulos, V. (2025). Forecast accuracy and inventory performance: insights from the M5 competition data. *European Journal of Operational Research*, 322(2), 414–426
- Karaesmen, F. (2013). Value of advance demand information in production and inventory systems with shared resources. In J. M. Smith & B. Tan (Eds.), *Handbook of Stochastic Models and Analysis of Manufacturing System Operations* (pp. 139–165). Springer
- Li, F., Fan, Y., Kishimzhan, Z., Wang, X., & Abdykadyrov, K. Zh. (2026). An informatics-enabled logistics service system for cross-border medical e-commerce: evidence from China–Kyrgyzstan collaboration. *Journal of Logistics, Informatics and Service Science*, 13(5), 38–48.
- Nazarieh, F., & Naderi Dehkordi, M. (2023). Time series forecasting based on a novel ensemble-based network and variational mode decomposition. *Expert Systems*, 40(8), e13291
- Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., ... Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38(3), 705–871

- Petropoulos, F., & Kourentzes, N. (2015). Forecast combinations for intermittent demand. *Journal of the Operational Research Society*, 66(6), 914–924
- Petropoulos, F., Wang, X., & Disney, S. M. (2019). The inventory performance of forecasting methods: evidence from the M3 competition. *International Journal of Forecasting*, 35(1), 251–265
- Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: a simple theoretical explanation. *International Journal of Forecasting*, 32(3), 754–762
- Wu, H., & Levinson, D. (2021). The ensemble approach to forecasting: a review and synthesis. *Transportation Research Part C*, 132, 103357
- Álvarez-Chaves, H., Barrero, D. F., Cobos, M., & R-Moreno, M. D. (2021). Patients forecasting in emergency services by using machine learning and exogenous variables. *Lecture Notes in Computer Science*, 13101, 180–193
- Chan, J., Görgülü, B., & Sarhangian, V. (2024). Inventory management with advance booking information. *Manufacturing & Service Operations Management*, 26(3), 931–951
- Douaioui, K., Oucheikh, R., Benmoussa, O., & Mabrouki, C. (2024). Machine learning and deep learning models for demand forecasting in supply chain management: a critical review. *Applied System Innovation*, 7(5), 93
- Fan, L., Song, Z., Mao, W., Luo, T., Wang, W., Yang, K., & Cao, F. (2025). Change is safer: a dynamic safety stock model for inventory management. *Journal of Intelligent Manufacturing*, 36, 3983–4003
- Li, L., Kang, Y., & Li, F. (2023). Bayesian forecast combination using time-varying features. *International Journal of Forecasting*, 39(3), 1287–1302
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4), 437–450
- Viverit, L., Heo, C. Y., Pereira, L. N., & Tiana, G. (2023). Application of machine learning to cluster hotel booking curves. *International Journal of Hospitality Management*, 111, 103455
- Nasseri, M., Falatouri, T., Brandtner, P., & Darbanian, F. (2023). Applying machine learning in retail demand prediction. *Applied Sciences*, 13(19), 11112
- Qi, M., Shi, Y., Qi, Y., Ma, C., Yuan, R., Wu, D., & Shen, Z.-J. M. (2023). A practical end-to-end inventory management model with deep learning. *Management Science*, 69(2), 759–773
- Song, M. (2025). ANFIS-based financial data quality control: a hybrid genetic algorithm approach for error detection and correction. *Journal of Logistics, Informatics and Service Science*, 12(12), 122–137.
- Özer, Ö., & Wei, W. (2004). Inventory control with limited capacity and advance demand information. *Operations Research*, 52(6), 988–1000
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.
- Demiray Kırmızı, S., Ceylan, Z., & Bulkan, S. (2024). Enhancing inventory management through safety-stock strategies—A case study. *Systems*, 12(7), 260
- Gogolev, S., & Ozhegov, E. (2023). Asymmetric loss function in product-level sales forecasting. *Applied Econometrics*, 70, 109–121
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804

- Lim, S. F. W. T., Rabinovich, E., Lee, S., & Park, S. (2025). Estimating stockout costs and optimal stockout rates. *Management Science*, 71(4), 3106–3126
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364
- Goltsos, T. E., Syntetos, A. A., Glock, C. H., & Ioannou, G. (2022). Inventory – forecasting: mind the gap. *European Journal of Operational Research*, 299(2), 397–419
- Lin, T.-Y., & Chou, Y.-C. (2022). Cost-sensitive regression learning on small dataset. *Connection Science*, 34(1), 296–322
- Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models. *Machine Learning*, 109, 1997–2028
- Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: an over 50-year review. *International Journal of Forecasting*, 39(4), 1518–1547
- Kang, Y., Cao, W., Petropoulos, F., & Li, F. (2022). Forecast with forecasts: diversity matters. *European Journal of Operational Research*, 301(1), 180–190
- Park, Y. J., Kim, D., Odermatt, F., Lee, J., & Kim, K. M. (2022). A large-scale ensemble learning framework for demand forecasting. *Proceedings of IEEE ICDM*, 378–387