The Eastasouth Journal of Information System and Computer Science
Vol. 2, No. 01, August 2024, pp. 203-215
ISSN: 3025-566X, DOI: 10.58812/esiscs.v2i01

Artificial Intelligence and Machine Learning in U.S. Business: Applications, Adoption Challenges and Implications for Strategic Decision-Making

Udoy Sankar Saha[1], Amena Akther[2], Preety Arifa Momotaj[3], Mohammad Didar Hossain[4], Md Tanvirul Islam[5], Semanto Saha Utsab[6]

[1] St. Francis College, Brooklyn, New York, USA, udoy.saha.sfc2026@gmail.com
[2] International American University, Los Angeles, California, USA, amenaakther017138@gmail.com
[3] International American University, Los Angeles, California, USA, arifarahmanpreety@gmail.com
[4] St. Francis College, Brooklyn, New York, USA, ecodidar@gmail.com
[5] Atlantis University, Miami, Florida, USA, imdtanvirulislam11@gmail.com
[6] Montclair State University, Montclair, New Jersey, USA, semanto.utsab13@gmail.com

Article Info

Article history:

Received August, 2024
Revised August, 2024
Accepted August, 2024

Keywords:

Artificial Intelligence;
Business Analytics;
Demand Forecasting;
Inventory Management;
Machine Learning;
Operational Resilience;
Quantile Regression;
United States Business

ABSTRACT

Artificial intelligence and machine learning offer opportunities to improve business planning, yet greater predictive accuracy does not necessarily translate into better operational decisions. This study examines that relationship through a reproducible simulation of retail demand forecasting and inventory allocation, contextualized within the U.S. business environment. Five independent synthetic panels, each comprising ten stores, eight products, and 156 weeks, support comparisons across seven forecasting and ordering policies. Histogram gradient boosting models estimate conditional mean demand and its 80th percentile, while benchmark policies include historical forecasts, moving averages, and validation-derived safety buffers. Chronological partitions prevent future information from entering model development, and an unannounced demand increase tests performance under distribution shift. With shortage costs set at four times surplus costs, mean boosting achieved the lowest mean absolute error of 10.17 units. Quantile boosting achieved the lowest average inventory loss of 41.05 synthetic cost units, compared with 57.22 for unbuffered mean forecasts and 42.85 for buffered mean forecasts. Its incremental loss reduction over the stronger buffered baseline was 4.22%. However, quantile coverage declined from 83.1% in unshifted stores to 64.5% in shifted stores, revealing vulnerability despite favorable average performance. These findings demonstrate the importance of aligning machine learning objectives with business costs, comparing credible alternatives, and monitoring reliability across operating conditions. The results provide methodological guidance for U.S. business applications; validation using real organizational data remains necessary before claiming financial or operational benefits.

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

Corresponding Author:

Name: Udoy Sankar Saha
Institution: St. Francis College, Brooklyn, New York, USA
Email: udoy.saha.sfc2026@gmail.com

1. INTRODUCTION

Business managers rarely purchase a prediction for its own sake. They need to decide how much inventory to hold, which supplier deserves attention, how to allocate staff, or whether a customer interaction requires escalation. Artificial intelligence becomes useful when its output improves one of these decisions at an acceptable cost. A model with a smaller statistical error can still encourage the wrong action if the evaluation metric ignores the consequences of underestimating and overestimating demand.

The United States provides a useful setting for this question because business AI adoption has varied across activities and industries. [1] described early Business Trends and Outlook Survey evidence, reporting that approximately four percent of nonfarm employer businesses used AI to produce goods or services in a late 2023 snapshot, with substantial sector differences. That is historical survey evidence tied to a particular definition and population. It is neither a current adoption estimate nor a measure of all employee experimentation with AI tools.

Research on work assistance offers a different kind of evidence. [2] in the original 2023 working paper version of their customer support study, reported heterogeneous productivity effects following introduction of a conversational assistant. The findings concern a specific workplace and task, rather than all American businesses. They nevertheless show why aggregate claims about AI value should be examined at the level of the workflow and the worker.

[3] studied AI assistance in professional writing tasks. Their experiment supports a task specific relationship between assistance and productivity, but writing performance does not establish inventory savings, procurement reliability, or firm profitability. Treating every application as the same intervention would conceal differences in information quality, oversight, and business consequences.

This paper focuses its computational analysis on demand forecasting and inventory decisions. The setting is relevant to retailers, distributors, and small or midsized businesses with repeated sales records, but the experiment uses no real firm data. A synthetic panel makes the information boundary, demand process, and decision costs visible. It also allows a controlled demand shift without attributing the resulting pattern to a real organization.

The research asks three questions. First, do policies with better point forecast accuracy also produce lower inventory loss under asymmetric costs? Second, does direct conditional quantile estimation improve on a conventional mean forecast with a validation derived buffer? Third, how does an unannounced demand shift affect the reliability of the resulting order quantities? The contribution is an executable benchmark and an interpretation of the accuracy cost relationship, rather than a new boosting algorithm or a national estimate of AI productivity.

2. LITERATURE AND CONCEPTUAL FOUNDATIONS

2.1 Business value depends on the decision

[4] distinguished predictive and explanatory modeling goals. In a business setting, this distinction matters whenever a manager moves from forecasting outcomes to claiming that a policy causes those outcomes. A sales forecast can identify a pattern without establishing that changing price, advertising, or staffing will produce the predicted change.

[5] reviewed machine learning methods and their relationship to economic analysis, including causal inference. Their discussion motivates keeping operational prediction separate from evaluation of a managerial intervention. The experiment in this paper predicts demand under a fixed simulated environment; it does not estimate the causal return on AI adoption or optimize prices.

[6] connected predictive information to prescriptive decisions. That perspective is central here: the appropriate ordering target depends on the cost of shortages relative to excess stock. Forecasting accuracy and operational performance are related, but their objectives are not interchangeable.

2.2 *Forecasting methods and uncertainty*

[7] developed gradient boosting as a framework for sequential function approximation. The present implementation uses histogram gradient boosting because it can fit nonlinear relationships in structured features and supports both squared error and quantile losses. No claim is made that this model class is always preferable to statistical time series models or neural networks.

[8] demonstrated the scalability of a particular tree boosting system, XGBoost. Their work illustrates the broader relevance of boosting for structured prediction. XGBoost itself is not used in this experiment, so no results are attributed to that implementation.

[9] reported results from the M5 accuracy competition, an important empirical benchmark for retail forecasting. The current study does not reproduce M5 or use its data. It draws the narrower lesson that repeated retail series support meaningful benchmarking and that model comparisons should be explicit about information and aggregation levels.

[10] established regression quantiles. Estimating a conditional upper quantile is useful when the cost of underprediction exceeds the cost of overprediction. The quantile target in this study follows a stated inventory loss function rather than an arbitrary preference for larger forecasts.

[11] developed principles for evaluating probabilistic forecasts using appropriate scoring rules. Here, pinball loss assesses the chosen quantile, while observed coverage checks whether demand falls below the predicted ordering level at the expected frequency. A favorable average score does not eliminate the need to examine performance under shift.

2.3 *Evaluation integrity and operational systems*

[12] examined weaknesses in common forecast accuracy measures. This study avoids percentage errors that can become unstable near zero demand and reports absolute error, squared error, and decision loss. Scaled errors would be useful when comparing series with very different units but are not included in the present benchmark.

[13] analyzed conditions under which cross validation can be valid for autoregressive prediction. Rather than treating random splitting as universally valid or invalid, the present design uses explicit chronological partitions, and an information set that matches one week ahead planning. This choice makes the simulated distribution shift observable and easy to audit.

[14] documented leakage as a source of unreliable machine learning conclusions. Feature construction here excludes the target week demand, and neither hyperparameter selection nor safety buffer estimation uses test outcomes. Earlier realized test demand is available only when forecasting a later week, as it would be in sequential operation.

[15] emphasized that machine learning systems incur maintenance costs and dependencies beyond model fitting. The business implications therefore extend to data ownership, feature freshness, monitoring, and fallback policies. These organizational issues are discussed as implementation requirements; their costs are not measured by the simulation.

2.4 *Human oversight and accountable deployment*

[16] examined the relationship between automation and augmentation in management. In inventory planning, a useful distinction concerns which actions the system may recommend, which it

may execute, and who can override a recommendation. Human review changes the operational policy and should be evaluated rather than assumed to improve every decision.

[17] developed a framework for attributing model predictions to input features. Such explanations can help examine whether a forecast depends on recent demand or promotional information. They do not establish causal effects, certify fairness, or replace testing. Attribution analysis is a possible deployment extension, not a procedure executed in this benchmark.

[18] proposed model cards for documenting intended use and evaluation conditions. This study follows the same reporting principle by specifying the synthetic population, features, decision rule, and limitations. A model developed for these conditions should not be silently repurposed for credit decisions or employment screening.

[19] presented the NIST AI Risk Management Framework as voluntary, context sensitive guidance. Its governance functions inform the discussion of ownership, measurement, and monitoring. Referencing that framework does not certify a system or establish compliance with any particular legal requirement.

Table 1. Business applications require different evaluation targets

Application	Prediction or output	Decision measure
Retail planning	Demand distribution	Shortage and surplus cost
Procurement	Supplier delay or failure risk	Disruption cost and review workload
Customer intelligence	Response or retention likelihood	Incremental value from an intervention
Fraud monitoring	Suspicious transaction score	Loss prevented at a fixed review capacity
Generative assistance	Suggested text or response	Task quality, completion time, and correction burden

3. RESEARCH DESIGN

3.1 Study type and evidentiary boundaries

The study combines a targeted literature synthesis with a reproducible computational experiment. The literature selection covers methods and business context published through 2023. It is not an exhaustive systematic review, and no formal screening counts or meta-analytic effect are claimed. The experiment was implemented and executed for this paper; its outputs are generated measurements rather than illustrative numbers inserted into a results section.

Every store, product, demand observation, and cost in the benchmark is synthetic. There are no participating companies, survey respondents, customer identifiers, or observed transactions. The setting is relevant to U.S. business operations but is not calibrated to the distribution of U.S. firms. The national context comes from separately identified published sources, not from relabeling simulated observations as American company data.

The benchmark tests a familiar decision principle under a transparent environment. It cannot establish a novel universal learning method. Its value lies in exposing how the apparent magnitude of an AI benefit changes when a credible, inexpensive alternative is included.

3.2 Simulated panel

Five independent replications use random seeds 11, 29, 47, 71, and 101. Each replication contains ten stores, eight products, and 156 sequential weeks, yielding 80 store product series and 12,480 raw observations. Across replications, the simulation generates 62,400 observations. Requiring a 52 week lag leaves 8,320 usable rows per replication.

Let s identify a store, p a product, and t a week beginning at zero. Store effects as are drawn from a normal

distribution with standard deviation 0.18; product effects bp use standard deviation 0.35. A promotion indicator P has probability 0.18. The relative price R is the clipped value of 1 plus normal noise with standard deviation 0.07 minus 0.12P, bounded between 0.65 and 1.25.

Demand follows a negative binomial distribution with size parameter 18 and conditional mean μ. The exact generator is log μspt = 3.25 + as + bp + 0.25 sin(2πt/52) + 0.12 cos(2πt/52) + 0.0015t + 0.45P − 1.6 log R + 0.10 log[(Ds,p,t−1 + 5)/30] + δst. The lag term equals zero at the first observation. In NumPy, the negative binomial probability is 18/(18 + μ), giving variance $\mu + \mu^2/18$.

The shift term δ equals 0.30 for stores zero through four from week 136 onward and zero otherwise. Holding other inputs fixed, it initially multiplies mean demand by exp (0.30), approximately 1.35. Its longer path also interacts with lagged demand. Store effects, product effects, and the shift indicator are not supplied directly to the learning algorithm.

These parameters were chosen to create interpretable seasonality, promotion effects, overdispersion, and a controlled shift. They were not estimated from U.S. retail microdata. Promotion and price have a programmed relationship, which should not be interpreted as an estimated demand elasticity.

3.3 *Information available at the decision time*

The fourteen input features are store identifier, product identifier, relative price, promotion, sine and cosine annual seasonality, normalized time, five demand lags at one, two, four, thirteen, and fifty two weeks, and trailing means over four and thirteen weeks. Price and promotion are treated as planned values known before the target week begins. Trailing means use shifted demand, so they exclude the week being predicted.

Store and product identifiers enter as numeric variables in this implementation. Tree splits can represent some group differences, but numeric coding imposes an arbitrary ordering. Alternative encodings and unseen store generalization are not evaluated. All test stores and products have training history.

Testing is sequential one step ahead. Once a test week has elapsed, its realized demand can enter features for a later test week. The model itself is not refitted. This differs from predicting an entire twenty-week horizon at one origin; the latter would require recursively generated lags or a separate multi horizon design.

3.4 *Chronological partitions*

Training covers weeks 52 through 115, giving 5,120 rows. Validation covers weeks 116 through 135, giving 1,600 rows. Testing covers weeks 136 through 155, also giving 1,600 rows. The five replications therefore provide 8,000 test observations for each policy, while the independent simulation replication remains the unit used to summarize variability.

Models are fitted on the training partition only. Validation selects between two tree sizes and estimates residual buffers. The selected estimator is not refitted on combined training and validation data because the buffer should correspond to the model that produced the validation residuals. Test rows are never used for these choices.

Table 2. Executed simulation configuration

Element	Specification
Independent panels	5 random seeds: 11, 29, 47, 71, 101
Panel structure	10 stores × 8 products × 156 weeks
Training	Weeks 52–115; 5,120 rows per panel
Validation	Weeks 116–135; 1,600 rows per panel
Testing	Weeks 136–155; 1,600 rows per panel

Element	Specification
Shift	Log demand increments 0.30 in stores 0–4 during testing
Learning models	Histogram boosting for mean and 0.80 quantile
Search	Maximum leaves 7 or 15; training only fitting
Decision loss	8 shortage units + 2 surplus units
Information	One week ahead; realized past demand available

4. FORECASTING AND INVENTORY POLICIES

4.1 Baseline forecasts

The seasonal naive policy predicts demand observed fifty-two weeks earlier. The last week policy uses the most recent demand, and the moving average policy uses the mean of the preceding four observations. These baselines have no learned model parameters, but they still depend on properly timestamped historical data.

Two additional baselines make the operational comparison more demanding. For each replication, a moving average plus buffer policy adds the 80th percentile of validation residuals to the moving average forecast. A boosted mean plus buffer policy applies the same construction to validation residuals of the mean model. Negative final forecasts are clipped to zero.

A single global residual offset is inexpensive and transparent. It assumes a stable additive correction across stores, products, and time. Conditional quantile estimation can potentially capture differences that the global buffer misses. Conversely, the simpler offset can be adequate when such differences are weak, so it is an important comparator.

4.2 Mean and quantile gradient boosting

The mean model minimizes squared error. The quantile model minimizes pinball loss at $\tau = 0.80$. Both use scikit learn histogram gradient boosting with 180 iterations, learning rate 0.06, minimum leaf size 25, L2 regularization 1, and no early stopping. The only searched structural parameter is maximum leaf count, with candidates 7 and 15.

For the mean model, validation mean squared error selects the candidate. For the quantile model, validation pinball loss selects it. Random state follows the replication seed. The identical feature set and small search space improve interpretability of the comparison but do not establish the best achievable result for either model class.

[20] documented the scikit learn software ecosystem. The actual computational environment for this experiment uses scikit learn 1.8.0, with exact Python and dependency versions recorded in the supplementary environment file. The 2011 citation credits the software publication; it does not imply that the current estimator implementation existed unchanged in 2011.

4.3 Decision loss and quantile target

Every policy converts its nonnegative forecast f into an integer quantity q = ceil(f). The stylized weekly loss is $C(q,D) = 8 \max(D - q,0) + 2 \max(q - D,0)$. Shortage is four times as costly per unit as surplus. These coefficients are synthetic cost units, not measured dollars, margins, or accounting expenses.

For continuous quantities and known conditional demand distribution F, the expected loss is minimized at F inverse(8/(8 + 2)) = F inverse(0.80). This standard relationship motivates the quantile model. With discrete demand, integer ordering is appropriate, but finite sample estimation, rounding, and distribution shift prevent perfect attainment of the theoretical target.

Pinball loss is $\varrho\tau(u) = u[\tau - I(u < 0)]$, where $u = D - f$. Without integer rounding, the selected inventory loss equals ten times the pinball loss at 0.80. With rounding, the reported values need not satisfy that equality exactly. This relationship is why forecast error and inventory loss should not be treated as

two independent confirmations of the same benefit.

The loss is a single period mismatch measure. There is no inventory carryover, replenishment delay, purchase budget, substitution, shelf capacity, or stock dependent censoring of sales. The quantities therefore represent weekly availability decisions under a simplified newsvendor setting, rather than a complete inventory control system.

4.4 Evaluation measures

Mean absolute error and root mean squared error compare point predictions to demand before rounding. Average decision loss uses rounded quantities. Coverage is the fraction of observations with demand no greater than the quantity. Unit fill is total min(D,q) divided by total demand within each replication and subgroup.

Coverage and fill answer different questions. Coverage describes how often the quantity meets or exceeds total demand. Fill describes what fraction of requested units can be supplied in this simplified setting. Neither is a calibrated probability of service quality across all real business settings.

Metrics are first computed separately for each replication. Tables report the arithmetic mean and sample standard deviation across five replications. These standard deviations are Monte Carlo dispersion measures, not confidence intervals for U.S. firms. No significance test or national effect estimate is reported.

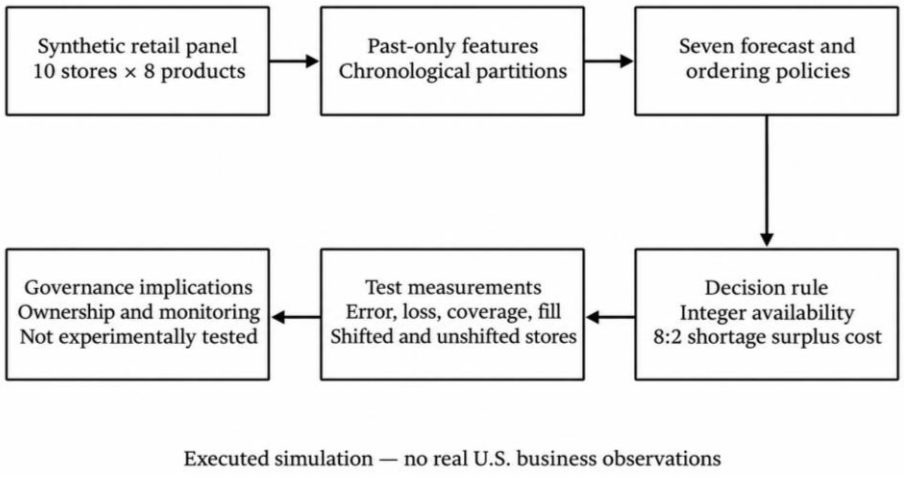


Figure 1. Executed workflow and its evidence boundaries. The governance stage describes recommendations rather than measured organizational outcomes.

5. RESULTS

5.1 Forecast accuracy and decision loss

Table 3 reports the executed benchmark. The boosted mean model achieved the lowest average absolute error, 10.17 units, and the lowest average root mean squared error, 14.56 units. The boosted quantile model had higher absolute error, 11.65 units, consistent with its deliberately higher ordering target. Comparing the latter only on point error would overlook its intended purpose.

Average decision loss was 41.05 for boosted quantiles, 42.85 for the buffered boosted mean, and 57.22 for the unbuffered boosted mean. The four-week moving average produced loss of 80.88, while adding its validation derived buffer reduced loss to 65.44. Last week and seasonal naive policies produced losses of 93.88 and 124.73, respectively.

Using ratios of replication averaged losses, direct quantile estimation reduced loss by 28.26 percent relative to unbuffered boosted means, 49.25 percent relative to unbuffered moving averages, and 4.22 percent relative to buffered boosted means. The

last comparison is the most informative measure of incremental value over a cost aware machine learning baseline. These percentages describe the synthetic environment only.

Table 3. Test performance across five synthetic panels

Policy	MAE	RMSE	Loss	Coverage %
Seasonal naive	19.13 ± 3.48	27.58 ± 5.34	124.73 ± 24.31	35.5
Last week	18.01 ± 3.02	26.20 ± 4.79	93.88 ± 15.75	49.5
Moving average	14.73 ± 2.42	21.65 ± 3.91	80.88 ± 13.66	54.0
Moving average + buffer	17.43 ± 2.85	22.64 ± 3.95	65.44 ± 11.49	74.9
Boosted mean	10.17 ± 1.59	14.56 ± 2.53	57.22 ± 9.99	50.3
Boosted mean + buffer	11.80 ± 1.81	15.25 ± 2.54	42.85 ± 7.25	74.4
Boosted quantile	11.65 ± 1.89	15.42 ± 2.73	41.05 ± 6.73	73.8

Note. MAE, RMSE, and loss are mean ± sample standard deviation across five panels. Coverage is the mean percentage. All quantities come from synthetic data.

The unbuffered mean versus quantile comparison combines different forecasting targets and decision policies. It should not be described as proof that one algorithm is categorically more accurate. Adding a simple validation buffer captures a large share of the apparent operational gain. The remaining difference favors the conditional quantile model on average, but it is smaller than the headline comparison with an unadjusted mean forecast.

The quantile policy supplied an average of 92.9 percent of requested units, compared with 84.8 percent for unbuffered boosted means and 92.3 percent for buffered boosted means. Its coverage was 73.8 percent overall, below the nominal 80 percent quantile target. Thus, the policy with the lowest loss still exhibited a reliability problem under the pooled test conditions.

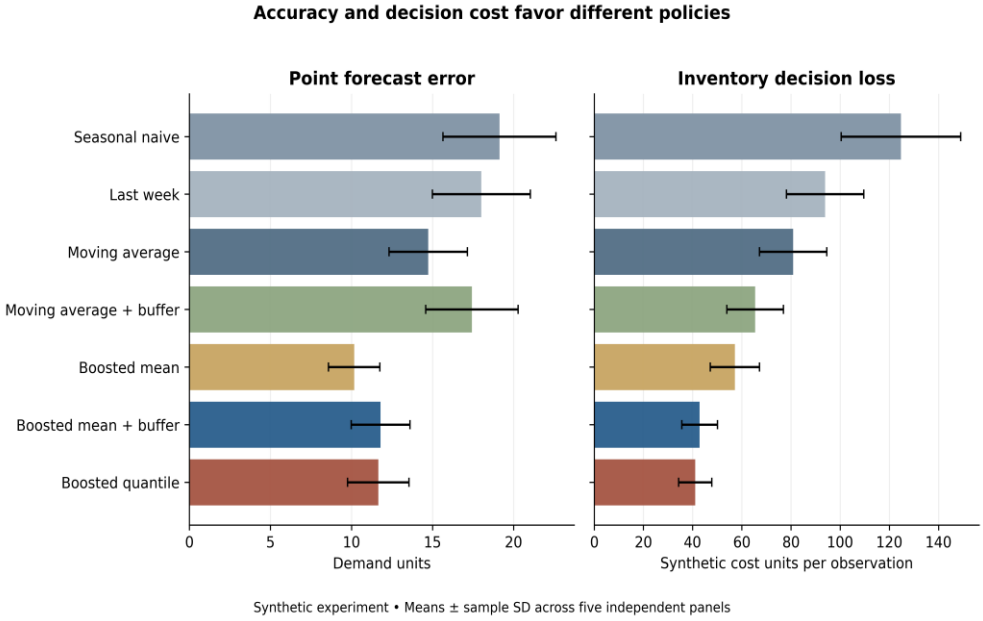


Figure 2. Point forecast error and decision loss across seven policies. Bars are means and error bars are sample standard deviations across five synthetic panels. Costs are not observed dollars.

5.2 Distribution shift

Table 4 separates the five shifted stores from the five unshifted stores. Quantile coverage averaged 83.1 percent in unshifted stores and 64.5 percent in shifted stores. Their corresponding average losses were 31.77 and 50.32 cost units. The conditional quantile model therefore adapted imperfectly through observed lag features after the demand increase, despite receiving realized prior week demand during sequential testing.

The boosted mean policy also suffered under shift: coverage was 61.5 percent in unshifted stores and 39.0 percent in shifted stores. Its decision loss increased from 38.25 to 76.18. These subgroup differences reflect both the imposed shift and the particular simulated store effects; store identities were not randomized to shift within each replication. They are descriptive stress comparisons, not randomized causal estimates of shift effects.

Table 4. Test coverage and loss by demand shift

Policy	Store group	Coverage %	Mean loss
Boosted mean	Unshifted	61.5	38.25
Boosted mean	Shifted	39.1	76.18
Boosted mean + buffer	Unshifted	84.9	32.72
Boosted mean + buffer	Shifted	63.8	52.99
Boosted quantile	Unshifted	83.1	31.77
Boosted quantile	Shifted	64.5	50.32

The subgroup pattern shows why an aggregate loss metric is insufficient for operational monitoring. An acceptable overall result can combine a relatively cautious policy in one group with frequent shortages in another. Coverage above 80 percent in unshifted stores also demonstrates that a pooled quantile objective does not force equal calibration across subgroups.

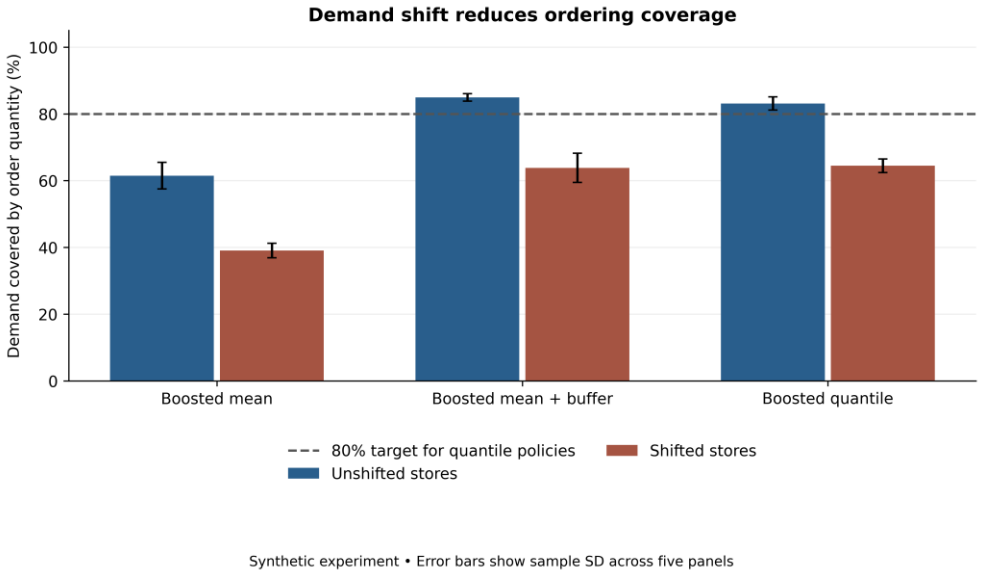


Figure 3. Coverage in shifted and unshifted stores. The dashed line marks the quantile target; it is not a target for the unbuffered mean model. Error bars show replication standard deviations.

5.3 Replication and uncertainty

All results come from five independently generated panels. The sample standard deviation of loss was 6.73 for boosted quantiles and 7.24 for buffered boosted means. These

quantities reflect variation in simulated stores, products, promotions, demand noise, and estimator fitting. They should not be used as evidence of regional variation in the United States.

The study does not claim a formal statistical advantage of the quantile policy based solely on its lower averaged loss. A larger Monte Carlo study, paired difference analysis, sensitivity to cost ratios, and evaluation across other demand generators would be necessary for a stronger computational conclusion. The available results establish the observed ranking in this benchmark and its material failure under the included shift.

6. DISCUSSION

6.1 Why the operational target matters

The main finding is a mismatch between the best point forecast and the best ordering policy. The boosted mean model predicts closer to realized demand on average, but its quantities do not reflect the larger shortage penalty. The quantile model willingly increases some overestimation to reduce more expensive underestimation. Its lower decision loss therefore follows an explicit business objective.

This is not a newly discovered optimization principle. The contribution is a transparent demonstration that includes a stronger buffered baseline and a shift analysis. Without the buffered comparator, the apparent benefit of the more specialized model would look much larger. Business managers need precisely this distinction when comparing a sophisticated AI investment with a simpler improvement to an existing planning process.

The 4.22 percent loss difference relative to buffered means might justify further investigation, but it is not a return on investment. The analysis excludes integration work, model maintenance, procurement constraints, labor, and deployment costs. A firm could rationally choose the buffered model if its operational simplicity outweighed a small difference in expected mismatch loss.

6.2 Interpretation for United States business

The experiment is relevant to U.S. retail and distribution decisions because its features represent recurring planning information: historical demand, prices, promotions, and seasonal patterns. Its applicability remains conditional. A company with sparse sales, long supplier lead times, or strong substitution across products faces a different problem from the single period benchmark.

Small firms may benefit first from consistent product identifiers, reconciled sales records, and a reliable demand baseline. Larger firms may need coordinated forecasts across stores and distribution centers. The present study does not compare firms by size, quantify a national productivity effect, or establish that advanced methods are necessary for every organization.

Customer intelligence, fraud detection, procurement scoring, and generative assistance involve different objectives. A customer retention model needs an intervention policy and a way to measure incremental retention. A fraud detector must be evaluated at a realistic review budget. A supplier risk score needs a definition of failure and careful handling of missing outcomes. The forecasting result does not validate these applications, but the requirement to connect predictions to decisions carries over as an evaluation principle.

6.3 Human oversight and organizational responsibility

A practical planning interface should present the recommendation, the information date, and the uncertainty or known limitation relevant to the decision. An explanation that recent demand and promotions drove a forecast may help a planner review it. It does not show that altering promotions would have the predicted causal effect.

Override behavior also requires measurement. If planners override most high-risk recommendations, the realized policy differs from the model evaluated offline. An audit record should distinguish the original prediction, executed quantity, reason for override, and later outcome. These records can reveal whether the system supports decision making or merely adds another unexamined number to a dashboard.

Monitoring must match the decision loss. A rise in shortage frequency among a store group deserves attention even when overall mean absolute error is stable. The simulation shows this issue directly: pooled performance did not preserve 80 percent coverage after shift. A real deployment would need a registered response, such as investigation, temporary buffer adjustment, or model retraining, followed by evaluation of that response.

6.4 Governance as part of the operating model

Data access should follow the minimum information needed for the forecast. This experiment uses store product aggregates and has no need for customer names, payment credentials, or employee attributes. A real implementation should document any added fields, their purpose, retention, and access permissions.

Versioning connects governance to reproducibility. The model checkpoint, feature definitions, cost assumptions, and forecast timestamp together define the decision system. Changing the shortage penalty changes the optimal quantile even if the underlying demand data are unchanged. Such a change should be reviewed as a policy modification, rather than hidden inside a technical update.

Table 5. Recommended controls for a business pilot

Control	Evidence to retain	Purpose
Decision ownership	Named owner and approved action scope	Connect recommendations to accountability
Feature timing	Timestamp and availability of every input	Prevent leakage and stale forecasts
Cost assumptions	Shortage, surplus, and integration costs	Make policy objectives reviewable
Shift monitoring	Coverage and loss by business unit	Detect uneven deterioration
Override tracking	Prediction, executed quantity, reason	Evaluate the actual operating policy
Release documentation	Version, validation period, fallback	Support reproducibility and rollback

7. LIMITATIONS AND EXTERNAL VALIDATION AGENDA

The most important limitation is that the business data are entirely synthetic. The experiment supports conclusions about an implemented simulator, not the prevalence or profitability of AI use among U.S. companies. Its parameters create conditions favorable to learning from structured features; different demand processes could change the ranking.

The comparison set is deliberately limited. It omits ARIMA, exponential smoothing, linear quantile regression, random forests, neural networks, hierarchical reconciliation, and optimized inventory control. Adding these methods would help distinguish the benefit of flexible conditional estimation from the general value of a cost aware target. The small hyperparameter search is transparent but cannot establish best in class forecasting performance.

The environment includes only one upward shift pattern, one cost ratio, and one twenty-week test window. The shifted stores are fixed rather than randomized, and the five replications provide limited Monte Carlo

precision. No ablation isolates each feature. No claims are made about adversarial robustness, cybersecurity performance, fairness across demographic groups, or formal privacy protection.

Demand is fully observed regardless of the order quantity. Real sales can be censored by stockouts, making demand estimation harder and creating a feedback loop between policy and training data. Inventory does not carry forward, and customers do not substitute between products. The simplified loss therefore cannot be interpreted as measured profit or working capital savings.

External validation should begin with a legally accessible, documented transaction panel and a clear prediction timestamp. A replication should test multiple origins, regions, business units, and product groups, with a held out period that reflects actual operating conditions. The historical comparator must include the firm's existing buffers and planner adjustments, not an artificially weak baseline.

A subsequent pilot should compare executed policies under a design that accounts for shared inventory and spillovers. Outcomes should include realized service, waste, review workload, and total operating cost. Claims about causal business benefit require such an intervention study or another credible identification strategy. Model accuracy alone does not supply that evidence.

8. CONCLUSION

AI and machine learning in U.S. business should be evaluated through the decisions they support. In this reproducible synthetic retail experiment, mean boosting produced the smallest point forecast errors, while quantile boosting produced the lowest loss under asymmetric shortage and surplus costs. A simple validation derived buffer substantially narrowed the advantage of direct quantile modeling. The best average policy also lost coverage after demand shift. These findings support an evaluation approach that combines credible baselines, business specific costs, chronological testing, and subgroup monitoring. They do not establish performance in real American firms; that requires external data and operational validation.

REFERENCES

[1] C. Breaux and E. Dinlersoz, "Only 3.8% of Businesses Use AI to Produce Goods and Services, Highest Use in Information Sector," 2023, *U.S. Census Bureau*. [Online]. Available: https://www.census.gov/library/stories/2023/11/businesses-use-ai.html

[2] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative AI at Work," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2304.11771v1.

[3] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, no. 6654, pp. 187–192, Jul. 2023, doi: https://doi.org/10.1126/science.adh2586.

[4] G. Shmueli, "To Explain or to Predict?," *Stat. Sci.*, vol. 25, no. 3, pp. 289–310, Oct. 2010, doi: https://doi.org/10.1214/10-STS330.

[5] S. Athey and G. W. Imbens, "Machine learning methods that economists should know about," *Annu. Rev. Econom.*, vol. 11, pp. 685–725, 2019, doi: https://doi.org/10.1146/annurev-economics-080217-053433.

[6] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics," *Manage. Sci.*, vol. 66, no. 3, pp. 1025–1044, 2020, doi: https://doi.org/10.1287/mnsc.2018.3253.

[7] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, Nov. 2001, doi: https://doi.org/10.1214/aos/1013203451.

[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. doi: https://doi.org/10.1145/2939672.2939785.

[9] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *Int. J. Forecast.*, vol. 38, no. 4, pp. 1346–1364, 2022, doi: https://doi.org/10.1016/j.ijforecast.2021.11.013.

[10] R. Koenker and G. Bassett Jr., "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978, doi: https://doi.org/10.2307/1913643.

[11] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007, doi: https://doi.org/10.1198/016214506000001437.

[12] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp.

679–688, 2006, doi: https://doi.org/10.1016/j.ijforecast.2006.03.001.

[13] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Comput. Stat. Data Anal.*, vol. 120, pp. 70–83, 2018, doi: https://doi.org/10.1016/j.csda.2017.11.003.

[14] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: https://doi.org/10.1016/j.patter.2023.100804.

[15] D. Sculley *et al.*, "Hidden Technical Debt in Machine Learning Systems," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

[16] S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation–augmentation paradox," *Acad. Manag. Rev.*, vol. 46, no. 1, pp. 192–210, 2021, doi: https://doi.org/10.5465/amr.2018.0072.

[17] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[18] M. Mitchell *et al.*, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, in FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 220–229. doi: https://doi.org/10.1145/3287560.3287596.

[19] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, NIST AI 100-1, 2023. doi: https://doi.org/10.6028/NIST.AI.100-1.

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011, [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html