
| RESEARCH ARTICLE

Addressing Demand Forecasting Challenges in the U.S. Retail Supply Chains: An AI-Driven Machine Learning Evaluation of Historical, Operational, and Macroeconomic Information

Sujoy Saha— Master of Science in Business Analytics, (University of New Haven, CT, USA), Master of Science in Statistics, (National University, Bangladesh), Bachelor of Science in Statistics, (National University, Bangladesh)

Email : ssujoy26@gmail.com, ORCID: 0009-0000-9358-7813

Md Firoze Mehedi— Master's in Logistics Systems Engineering, School Of Engineering, Bowling Green State University

Email : mmehedi@bgsu.edu, ORCID: 0009-0006-9717-8543

Tariqul Islam— MSIT- concentration in Artificial Intelligence , (University of the Potomac, Arlington, Virginia)

Email : info.tariqulislam@gmail.com, ORCID: 0009-0008-0026-0423

Md Kamrul Islam — Master of Science in Information Technology, (Washington University of Science and Technology)

Email : mdkamrulislam011994@gmail.com, ORCID: 0009-0001-3765-3107

Corresponding Author: Sujoy Saha, **E-mail:** ssujoy26@gmail.com

| ABSTRACT

U.S. retail supply chains face persistent demand-forecasting challenges because product demand varies across items, stores, and time, complicating inventory and replenishment planning. Improving forecast accuracy while determining which available data sources provide meaningful predictive value is therefore important for data-driven retail supply chain management. This study addresses this challenge by applying machine learning to evaluate the incremental predictive value of historical demand, retail and operational information, and macroeconomic indicators for large-scale U.S. retail demand forecasting. Using the M5 dataset, three LightGBM specifications were evaluated under a common modeling framework across 25,940 item-store series and five rolling 28-day forecast windows. Model A incorporated lagged and rolling historical-demand features; Model B additionally incorporated calendar, SNAP, event, and selling-price information; and Model C further incorporated two-month-lagged consumer price index, unemployment, and retail sales indicators. Forecast performance was compared with a 28-day seasonal-naïve benchmark using MAE, RMSE, and WAPE, supplemented by paired statistical and bootstrap robustness analyses. Mean WAPE decreased from 0.9006 for the seasonal-naïve benchmark to 0.7022 for Model A, representing an improvement of approximately 22.03%. Model B further reduced mean WAPE to 0.6979, an additional improvement of approximately 0.61%, while Model C produced only a negligible further reduction to 0.6978, approximately 0.02%. Paired and series-average analyses showed a large and robust improvement for Model A, a small positive but heterogeneous average improvement for Model B, and negligible aggregate incremental forecasting value for Model C, whose bootstrap confidence interval for the mean MAE improvement included zero. These findings demonstrate that the history-based LightGBM specification generated the dominant forecasting improvement, while retail and operational information provided a modest incremental benefit and the examined lagged macroeconomic indicators contributed little additional aggregate predictive value. For U.S. retail supply chain forecasting, the findings demonstrate how machine-learning-based evaluation of progressively richer information sets can identify which data sources provide meaningful forecasting gains, supporting evidence-based model development for inventory and replenishment planning.

| KEYWORDS

Machine Learning (ML), Demand Forecasting, Supply Chain Analytics, Retail Demand Forecasting, LightGBM, Macroeconomic Indicators, Retail Analytics

| ARTICLE INFORMATION

ACCEPTED: 02 August 2026

PUBLISHED: 01 September 2026

DOI: 10.32996/fcsai.2026.5.9.24

1. Introduction

Accurate demand forecasting is a fundamental component of retail supply chain planning because decisions concerning inventory availability, replenishment, resource allocation, and customer service depend on reliable estimates of future demand. Forecasting becomes particularly challenging in large-scale retail environments, where organizations must manage thousands of product-location combinations characterized by seasonality, intermittent demand, promotional effects, price variation, and changing market conditions. Inaccurate forecasts can contribute to excess inventory and associated holding costs when demand is overestimated, while underestimation can increase the risk of stockouts, lost sales, and service-level deterioration. Consequently,

improving forecasting accuracy remains an important analytical objective for retail and supply chain decision-making (Babai et al., 2022; Fildes et al., 2022; Makridakis et al., 2022). Advances in machine learning have expanded the range of methods available for large-scale retail demand forecasting. Unlike conventional univariate forecasting approaches that primarily rely on the historical behavior of individual time series, machine-learning models can learn nonlinear relationships across large collections of related sales series while incorporating multiple explanatory variables. This capability is particularly relevant in retail environments, where demand may be associated with calendar patterns, special events, pricing, promotions, and other operational factors. Evidence from the M5 forecasting competition further demonstrated the potential of global machine-learning approaches for large-scale retail forecasting, with gradient-boosting methods, particularly LightGBM, playing a prominent role among high-performing solutions (Fildes et al., 2022; Makridakis et al., 2022).

Despite these advances, an important question concerns the incremental value of different information sources in large-scale retail demand forecasting. Historical demand patterns may contain substantial predictive information, while calendar effects, events, pricing, and other retail-specific variables can provide additional operational context. Macroeconomic conditions may also influence consumer demand, motivating the inclusion of indicators such as inflation and labor-market conditions in forecasting models. However, the practical forecasting benefit of adding such broader economic information should be evaluated relative to models that already incorporate historical demand and retail-level covariates. A structured comparison of progressively richer predictor sets can therefore help distinguish whether improvements arise primarily from demand history, retail and operational information, or additional macroeconomic context (Haque et al., 2023). Accordingly, this study investigates the incremental forecasting value of progressively richer information sets for large-scale retail demand prediction. Using the M5 retail dataset, three LightGBM forecasting models are evaluated under a common modeling framework. Model A uses only lagged and rolling historical demand features; Model B extends this specification with calendar, SNAP, event, and selling-price information; and Model C further incorporates lagged macroeconomic indicators representing consumer prices, unemployment, and retail sales. The models are evaluated against a 28-day seasonal-naïve benchmark using five rolling 28-day forecast windows. This design enables the study to assess whether retail and macroeconomic information provides incremental predictive value beyond that contained in historical demand patterns while maintaining a consistent forecasting algorithm across model specifications.

This study makes three main contributions to the retail demand-forecasting literature. First, it provides a controlled comparison of progressively richer predictor sets within the same machine-learning algorithm, allowing the incremental predictive value of historical demand, retail and operational variables, and macroeconomic indicators to be examined separately. Second, it evaluates forecasting performance at scale across 25,940 item-store series and five rolling 28-day forecast windows, providing evidence across multiple forecast origins rather than relying on a single holdout period. Third, the study complements aggregate forecast-accuracy measures with paired statistical and robustness analyses at the item-store level, enabling small average improvements to be distinguished from effects that are consistent across individual series and forecasting windows. The remainder of this paper is organized as follows. Section 2 reviews the literature on retail demand forecasting, machine-learning approaches, and the use of operational and macroeconomic information in forecasting. Section 3 describes the M5 dataset, feature construction, model specifications, rolling backtest design, and evaluation methodology. Section 4 presents the forecasting and statistical comparison results. Section 5 discusses the findings and their implications for retail and supply chain forecasting. Section 6 outlines the limitations of the study and directions for future research, and Section 7 concludes the paper.

2. Literature Review

Retail demand forecasting has traditionally relied on statistical time-series methods that model historical patterns such as level, trend, and seasonality. These approaches remain useful because of their interpretability and relatively straightforward implementation, particularly when individual series exhibit stable temporal structures. However, large-scale retail forecasting presents additional challenges because organizations may need to generate forecasts simultaneously for thousands of related product-location series while accounting for heterogeneous demand patterns and potentially informative external variables. The

growth of machine-learning methods has therefore expanded the forecasting toolkit by enabling models to learn relationships across large collections of series and incorporate multiple explanatory features within a common predictive framework (Fildes et al., 2022, Makridakis et al., 2022).

Among machine-learning approaches, gradient-boosted decision trees have become particularly relevant for large-scale demand forecasting because they can capture nonlinear relationships and interactions among numerous predictors while remaining computationally efficient on large tabular datasets. LightGBM is a gradient-boosting framework designed for efficient tree-based learning and has been widely applied to prediction problems involving large numbers of observations and explanatory variables (Ke et al., 2017). Its relevance to retail forecasting was further demonstrated in the M5 competition, where machine-learning approaches based on gradient boosting formed an important component of several high-performing forecasting solutions. These results illustrate the suitability of global tree-based models for exploiting information shared across large collections of related retail demand series (Makridakis et al., 2022).

Beyond historical sales patterns, retail demand forecasting can benefit from explanatory variables that capture the operational and temporal context in which demand occurs (Huber & Stuckenschmidt, 2020). Calendar characteristics can represent recurring weekly and monthly patterns, while holidays and special events may generate temporary changes in purchasing behavior. Pricing information can also provide useful predictive signals because changes in selling prices may coincide with changes in product-level demand. Large-scale retail datasets such as the M5 data facilitate the integration of these explanatory variables with historical demand, including calendar information, event indicators, state-level Supplemental Nutrition Assistance Program (SNAP) indicators, and weekly selling prices. Incorporating such covariates into machine-learning models provides a means of evaluating whether retail and operational information contributes additional forecasting value beyond lagged demand history alone (Makridakis et al., 2022).

Retail demand may also be associated with broader economic conditions that influence household purchasing behavior and aggregate consumption patterns. Indicators such as consumer prices, unemployment, and retail sales provide information about changes in the economic environment that may not be represented directly by product-level sales histories or retail operational variables. From a forecasting perspective, incorporating such indicators may therefore provide additional contextual information when consumer demand responds to changing economic conditions. However, the usefulness of macroeconomic variables for product-level forecasting is ultimately an empirical question, particularly when models already contain detailed historical demand and retail-specific predictors. Evaluating their incremental predictive contribution is consequently important for determining whether the additional information produces meaningful improvements in forecast accuracy (Haque et al., 2023). Reliable comparison of forecasting models also depends on the design of the evaluation procedure. Performance measured over a single holdout period may reflect conditions specific to that particular forecast origin and may therefore provide an incomplete assessment of relative model accuracy. Rolling-origin evaluation addresses this concern by repeatedly estimating or updating forecasting models using information available before each forecast origin and assessing their predictions over subsequent periods (Tashman, 2000). Evaluating competing specifications across multiple forecast windows can consequently provide stronger evidence about whether observed improvements are persistent or dependent on particular time periods. This consideration is especially relevant when assessing the incremental value of additional predictor groups, because small average gains may mask heterogeneous performance across forecast origins or individual retail series. Taken together, the literature demonstrates the potential of machine-learning methods for large-scale retail forecasting and highlights the availability of multiple sources of predictive information, ranging from historical demand to retail-specific and broader economic variables. However, the availability of additional predictors does not necessarily imply that each information source provides meaningful incremental forecasting value. A controlled comparison in which the forecasting algorithm is held constant while predictor groups are introduced progressively can help isolate these contributions. Accordingly, the present study compares a demand-history specification with successive extensions incorporating retail and operational variables and lagged macroeconomic indicators. The use of multiple rolling forecast windows further allows the consistency of any incremental forecasting gains to be examined across different forecast origins.

3. Methodology

3.1 Dataset Description

The empirical analysis used the M5 Forecasting – Accuracy dataset, which contains large-scale hierarchical retail sales information from Walmart stores in the United States. The original sales data comprise 30,490 item-store time series representing 3,049 products sold across 10 stores located in California, Texas, and Wisconsin. Each original series contains 1,941 daily sales observations covering January 29, 2011 through June 19, 2016. In addition to daily unit sales, the M5 data provide calendar information describing dates, events, and state-level Supplemental Nutrition Assistance Program (SNAP) indicators, together with weekly product-level selling-price information. These data provide the retail demand history and operational variables used to construct the forecasting models in this study (Makridakis et al., 2022). Following preparation and integration of the sales, calendar, and selling-price data, the usable modeling population consisted of 25,940 item-store time series. The

associated candidate price data contained 6,385,677 records, all of which were successfully matched during validation, with no unmatched price records. The forecasting target was daily unit demand for each item-store series. The final modeling population was retained consistently across the rolling evaluation windows, providing a common set of retail series for comparing the alternative forecasting specifications.

3.2 Feature Engineering and Predictor Construction

Historical demand features were constructed to capture recent and recurring sales patterns while preventing information from the forecast horizon from entering the predictor set. Four lag variables were generated at 28, 35, 42, and 56 days. In addition, two shifted rolling-mean features were constructed: a 7-day rolling mean and a 28-day rolling mean, with both calculations shifted by 28 days. The resulting six historical-demand predictors were therefore `lag_28`, `lag_35`, `lag_42`, `lag_56`, `rolling_mean_7_shift28`, and `rolling_mean_28_shift28`. Beginning the demand-history features at a 28-day lag ensured that actual sales occurring within the corresponding 28-day forecast horizon were not used as predictors.

Retail and operational covariates were incorporated to capture additional information potentially associated with short-term demand variation. Five variables were included: day of the week (`wday`), month, a state-specific SNAP indicator (`snap`), an event indicator (`event_flag`), and selling price (`sell_price`). The calendar variables represented recurring temporal patterns, while the SNAP and event indicators captured information associated with program participation dates and special events recorded in the M5 calendar. Selling price provided product-level pricing information for the corresponding forecast horizon. Because horizon selling prices were used as predictors, the analysis assumes that planned or future selling-price information is available at the forecast origin.

Three macroeconomic indicators were incorporated to represent broader economic conditions: the Consumer Price Index (CPI), unemployment rate, and retail sales. The economic feature dataset contained monthly observations that had already been lagged by two months to approximate information availability at the time of forecasting. These lagged variables were aligned with the corresponding retail observations and entered the modeling data as `cpi`, `unemployment`, and `retail_sales`. For each 28-day backtest window, the applicable macroeconomic values were frozen at the forecast origin and held constant throughout the forecast horizon. This procedure prevented subsequently observed macroeconomic information from entering forecasts retrospectively and therefore reduced the risk of look-ahead leakage. The two-month lag should be interpreted as an operational information-availability assumption rather than as a reconstruction of historical real-time macroeconomic data vintages.

3.3 Forecasting Model Development

Three forecasting specifications were developed using the same LightGBM gradient-boosted decision-tree regression algorithm, with the predictor set progressively expanded across models (Ke et al., 2017). Model A served as the demand-history specification and contained the six lagged and shifted rolling-demand features. Model B retained these six predictors and added the five retail and operational variables—day of the week, month, SNAP, event indicator, and selling price—for a total of 11 predictors. Model C retained all Model B predictors and further incorporated CPI, unemployment, and retail sales, resulting in 14 predictors. Holding the forecasting algorithm constant across the three specifications allowed differences in predictive performance to be attributed to the incremental information contained in the additional predictor groups rather than to differences in modeling algorithms. All three model specifications were estimated using a fixed LightGBM configuration to maintain comparability across predictor sets. The objective function was `regression_l1`, with 300 boosting estimators and a learning rate of 0.05. The number of leaves was set to 31 and the minimum number of observations per child node (`min_child_samples`) to 50. The configured subsample and feature-subsampling parameters were 0.8, while L1 and L2 regularization parameters (`reg_alpha` and `reg_lambda`) were each set to 0.1. A random seed of 42 was used to support reproducibility, and parallel processing was enabled using all available CPU cores. The same configuration was applied to Models A, B, and C, with no documented early stopping or validation-set-based hyperparameter optimization.

3.4 Rolling Backtest and Training Design

Forecast performance was evaluated using a rolling backtest design comprising five non-overlapping 28-day forecast windows. For each forecast origin, model training was restricted to the preceding 365 days, with the training period ending one day before the corresponding evaluation window. The five evaluation periods were January 4–31, February 1–28, February 29–March 27, March 28–April 24, and April 25–May 22, 2016. Each evaluation window contained forecasts for all 25,940 item-store series over 28 days, yielding 726,320 observations per window and 3,631,600 evaluation observations across the complete backtest. This rolling design enabled model performance to be assessed across multiple forecast origins while preserving the temporal ordering between model training and evaluation (Tashman, 2000).

Because of computational memory constraints, model estimation used a deterministic random sampling procedure within the physical training-data batches rather than the complete set of eligible training observations. For each rolling forecast window, the 365-day training period contained 9,468,100 eligible observations. A maximum of 20,000 observations was randomly

sampled from each of 130 physical data batches, resulting in 2,600,000 training observations per window, equivalent to approximately 27.46% of the eligible training data. The sampling procedure was applied consistently across the model specifications, while the evaluation data were not sampled; all 726,320 observations in each 28-day evaluation window were retained for forecast assessment.

3.5 Benchmark and Forecast Evaluation Metrics

A 28-day seasonal-naive forecast was used as the benchmark for evaluating the machine-learning models. Under this benchmark, demand observed 28 days earlier served as the forecast for the corresponding horizon day. The benchmark therefore provided a simple historical reference that preserved the 28-day seasonal structure of the forecasting task without incorporating additional retail, operational, or macroeconomic predictors. Comparing Models A, B, and C with this common benchmark enabled the analysis to distinguish the forecasting gains associated with machine-learning-based demand modeling from the smaller incremental effects associated with progressively richer predictor sets.

Forecast accuracy was evaluated using three complementary error measures: mean absolute error (MAE), root mean squared error (RMSE), and weighted absolute percentage error (WAPE) (Hyndman & Koehler, 2006). MAE measures the average absolute deviation between predicted and observed demand, providing an interpretable measure of typical forecast error in the units of the target variable. RMSE places greater weight on larger forecast errors through squared deviations and therefore provides additional information about sensitivity to relatively large prediction errors. WAPE expresses total absolute forecast error relative to total observed demand, facilitating comparison of aggregate forecasting performance across evaluation windows with different demand volumes. Lower values of all three metrics indicate better forecasting performance. Model performance was calculated separately for each of the five rolling backtest windows and subsequently summarized across windows using the mean and standard deviation of each metric.

3.6 Statistical and Robustness Analysis

To complement the aggregate forecast-accuracy measures, paired statistical comparisons were conducted using item-store-level MAE differences. The primary paired analysis was performed across 129,700 series-window observations, corresponding to 25,940 item-store series evaluated across five forecast windows. Three comparisons were examined: Model A versus the seasonal-naive benchmark, Model B versus Model A, and Model C versus Model B. For each comparison, the mean paired MAE difference and relative improvement were calculated, together with the proportions of observations for which the comparison model improved or worsened forecast accuracy. Statistical differences in the paired error distributions were assessed using the Wilcoxon signed-rank test (Wilcoxon, 1945), and 95% confidence intervals for the mean paired MAE differences were estimated using 2,000 bootstrap iterations (Efron, 1979). As an additional robustness check, errors were averaged across the five windows for each of the 25,940 item-store series and the paired comparisons were repeated at the series level, reducing the dependence associated with repeated observations of the same series across forecast windows.

3.7 Temporal Integrity and Experimental Controls

Several design controls were applied to preserve temporal integrity and comparability across the forecasting experiments. Historical demand predictors were restricted to lagged or shifted values beginning 28 days before the forecast horizon, while the 365-day training period for each backtest ended before its corresponding evaluation window. Calendar, event, and SNAP variables were treated as information known for the forecast horizon, and planned selling prices were assumed to be available at the forecast origin. Macroeconomic indicators incorporated the predefined two-month availability lag and were frozen throughout each 28-day forecast window. The same LightGBM configuration, training-sampling procedure, evaluation population, and accuracy metrics were applied consistently across Models A, B, and C. These controls were intended to minimize look-ahead leakage and ensure that differences in forecast performance primarily reflected changes in the predictor information available to each model.

4. Results

The forecasting results demonstrate a substantial improvement from the seasonal-naive benchmark to the machine-learning specifications, followed by progressively smaller gains as additional predictor groups were introduced. Across the five rolling Page | 4 IJLLT 5(1): 00-00 backtest windows, the 28-day seasonal-naive benchmark produced a mean MAE of 1.3057, mean RMSE of 3.0875, and mean WAPE of 0.9006. Model A, which used only lagged and rolling historical-demand features, reduced these values to 1.0182, 2.5648, and 0.7022, respectively. Model B, incorporating additional retail and operational variables, achieved mean MAE, RMSE, and WAPE values of 1.0120, 2.5625, and 0.6979. Model C, which further incorporated the lagged macroeconomic indicators, produced corresponding values of 1.0119, 2.5575, and 0.6978. Overall, the largest improvement occurred between the seasonal-naive benchmark and Model A, whereas the subsequent extensions from Model A to Model B and from Model B to Model C resulted in considerably smaller changes in aggregate forecast accuracy, as summarized in Table 1 and illustrated by the mean WAPE comparison in Figure 1.

Table 1. Overall forecasting performance across five rolling backtest windows

Model	Mean MAE	SD MAE	Mean RMSE	SD RMSE	Mean WAPE	SD WAPE
Seasonal Naive	1.3057	0.023	3.0875	0.1713	0.9006	0.0326
Model A	1.0182	0.0205	2.5648	0.134	0.7022	0.0215
Model B	1.012	0.0208	2.5625	0.1262	0.6979	0.0208
Model C	1.0119	0.0215	2.5575	0.1313	0.6978	0.0214

Note. MAE = mean absolute error, RMSE = root mean squared error, WAPE = weighted absolute percentage error, SD = standard deviation across the five rolling backtest windows. Model A uses historical-demand predictors, Model B adds retail and operational predictors, Model C further adds lagged macroeconomic indicators.

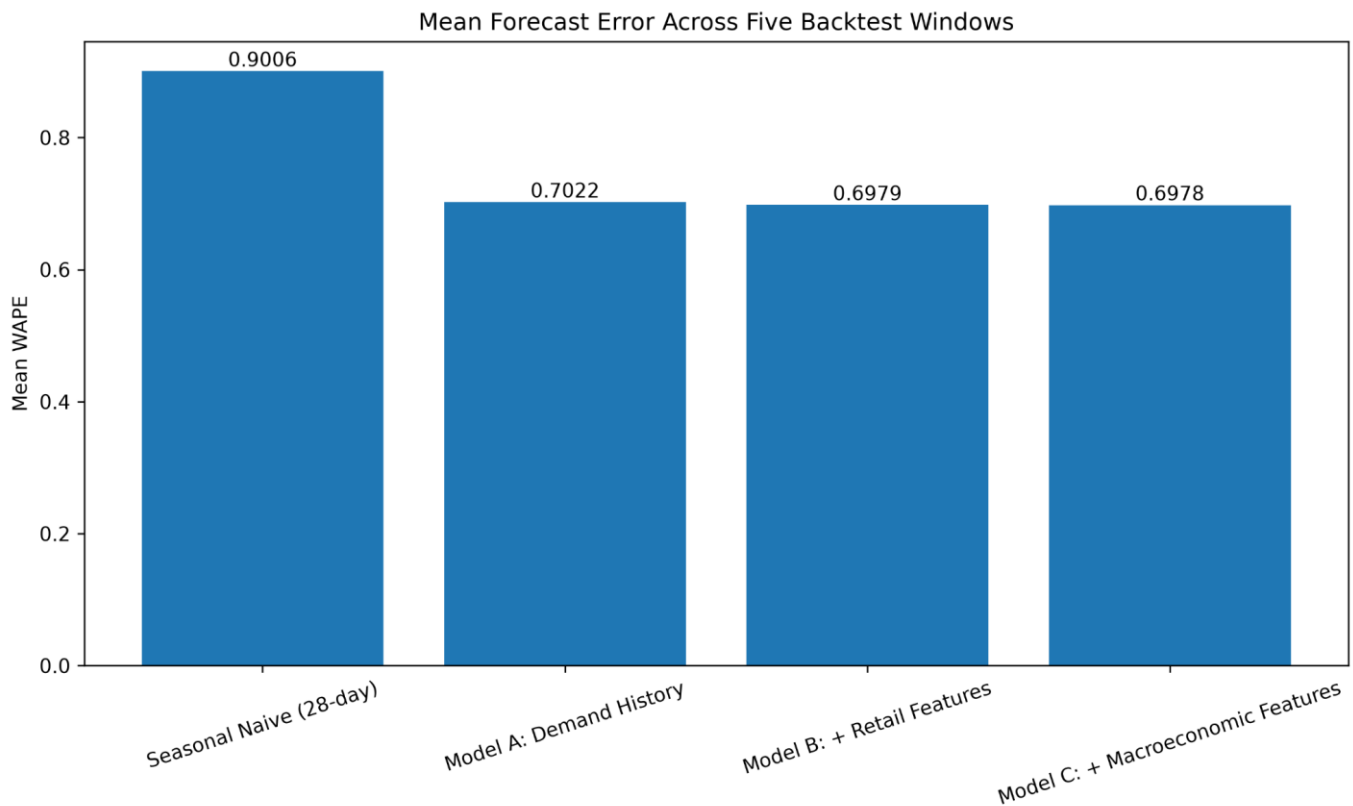


Figure 1. Mean WAPE across the five rolling backtest windows for the seasonal-naive benchmark and Models A–C. Lower WAPE indicates better forecasting performance.

The window-level results showed that all three LightGBM specifications achieved lower WAPE than the seasonal-naive benchmark in each of the five evaluation periods. Model A recorded WAPE values ranging from 0.6798 to 0.7341, compared with 0.8727 to 0.9542 for the seasonal-naive benchmark. The addition of retail and operational predictors in Model B produced lower WAPE than Model A in all five windows, although the magnitude of improvement was relatively small. Model C did not consistently improve upon Model B: it achieved lower WAPE in Windows 3 and 4 but slightly higher WAPE in Windows 1, 2, and 5. Thus, while the machine-learning specifications consistently outperformed the seasonal-naive benchmark across forecast origins, the incremental contribution of the macroeconomic predictors varied across evaluation periods, as summarized in Table 2 and illustrated in Figure 2.

Table 2. WAPE performance across the five rolling forecast windows

Forecast Window	Seasonal Naive	Model A	Model B	Model C
Window 1	0.954199	0.734113	0.729028	0.729409
Window 2	0.879081	0.712639	0.707142	0.708615
Window 3	0.906275	0.695861	0.692944	0.692344
Window 4	0.890716	0.688557	0.683569	0.681629
Window 5	0.872713	0.679780	0.676885	0.677025

Note. Lower WAPE indicates better forecasting performance. Windows 1–5 correspond respectively to January 4–31, February 1–28, February 29–March 27, March 28–April 24, and April 25–May 22, 2016.

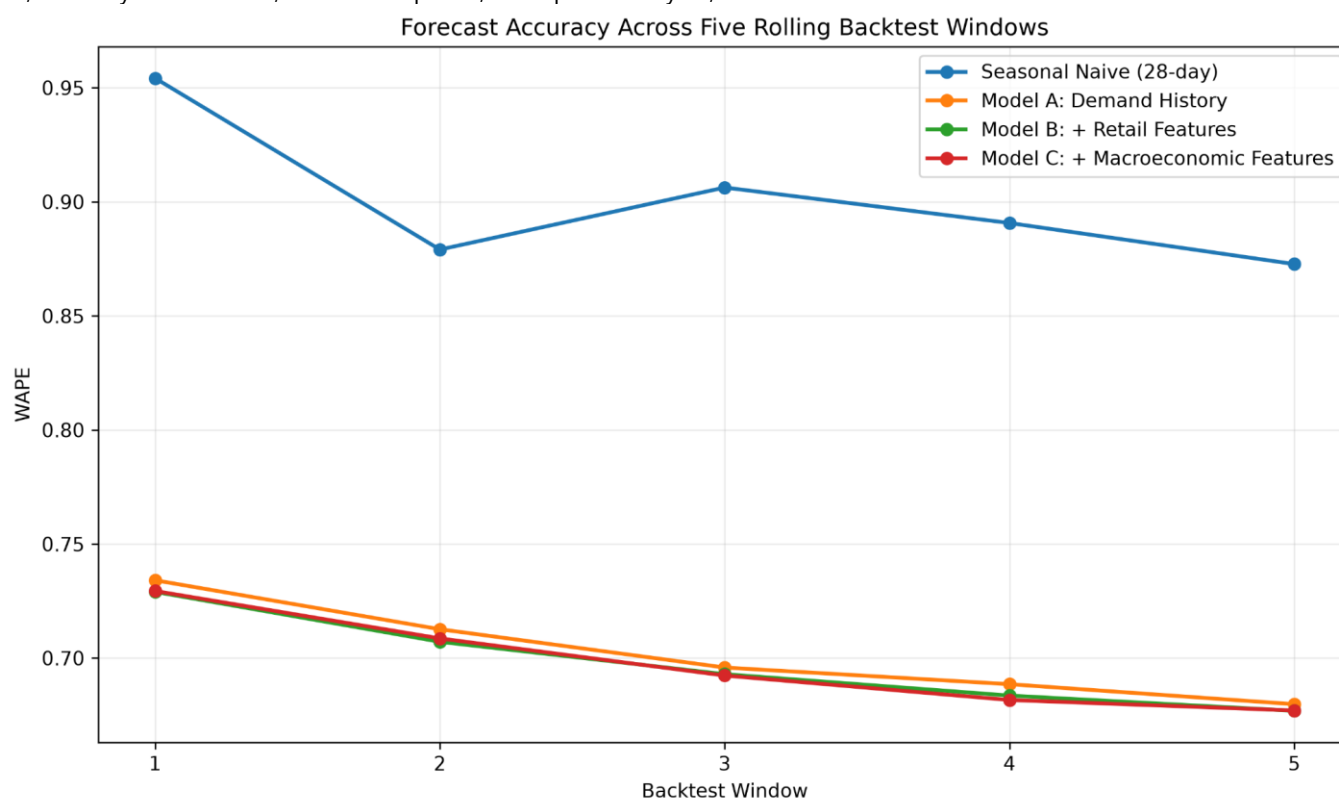


Figure 2. WAPE performance across the five rolling forecast windows for the seasonal-naive benchmark and Models A–C. Lower WAPE indicates better forecasting performance.

The magnitude of improvement decreased substantially as the predictor set was progressively expanded. Relative to the seasonal-naive benchmark, Model A reduced mean WAPE from 0.9006 to 0.7022, corresponding to a relative improvement of approximately 22.031%. Adding the retail and operational predictors in Model B further reduced mean WAPE from 0.7022 to 0.6979, representing an additional improvement of approximately 0.609%. By comparison, incorporating the lagged macroeconomic indicators in Model C changed mean WAPE only from 0.6979 to 0.6978, an incremental improvement of approximately 0.016%. These results indicate that the dominant forecasting gain was obtained from the machine-learning model based on historical-demand information, while retail and operational variables provided a smaller additional gain and the three macroeconomic indicators contributed little additional improvement in aggregate WAPE.

The superiority of Model A over the seasonal-naive benchmark was also evident in the paired error analysis. Across 129,700 item-store-window observations, the mean paired MAE reduction was 0.2875, corresponding to a relative improvement of approximately 22.017%. Model A produced lower MAE than the seasonal-naive benchmark in 82.28% of the series-window comparisons, whereas higher MAE was observed in 8.28%. The Wilcoxon signed-rank test indicated a detectable difference in the paired error distributions ($p < 0.001$), and the 95% bootstrap confidence interval for the mean MAE reduction was [0.2846,

0.2904], remaining clearly above zero. When forecast errors were first averaged across the five windows for each of the 25,940 item-store series, the mean MAE reduction remained 0.2875, with a 95% bootstrap confidence interval of [0.2828, 0.2919]. At this series-average level, Model A improved upon the seasonal-naïve benchmark for approximately 98.30% of the item-store series. Together, these results show that the improvement of Model A over the seasonal-naïve benchmark was large and robust across both series-window and series-average comparisons.

The incremental improvement from Model A to Model B was substantially smaller but remained positive on average. Across the 129,700 series-window observations, Model B reduced MAE by an average of 0.0062 relative to Model A, corresponding to a relative improvement of approximately 0.606%. Model B achieved lower MAE in 38.89% of the series-window comparisons and higher MAE in 46.11%. The Wilcoxon signed-rank test indicated a detectable difference in the paired error distributions ($p < 0.001$), while the 95% bootstrap confidence interval for the mean MAE reduction was [0.0055, 0.0068], remaining above zero. At the series-average level, based on 25,940 item-store series, the mean MAE reduction remained 0.0062, with a 95% bootstrap confidence interval of [0.0052, 0.0071]. Model B improved upon Model A for approximately 45.74% of the series and worsened performance for approximately 52.54%. These findings indicate a small positive average contribution from the retail and operational predictors, but also substantial heterogeneity across individual item-store series.

The addition of the lagged macroeconomic indicators in Model C produced only a negligible change in forecast accuracy relative to Model B. Across the 129,700 series-window observations, the mean paired MAE reduction was approximately 0.0002, corresponding to a relative improvement of approximately 0.017%. Model C achieved lower MAE in 42.08% of the series-window comparisons and higher MAE in 45.95%. Although the Wilcoxon signed-rank test detected a difference in the paired error distributions ($p < 0.001$), the 95% bootstrap confidence interval for the mean MAE reduction was [−0.0001, 0.0005], which included zero. The series-average analysis yielded a similarly small mean MAE reduction of approximately 0.0002, with a 95% bootstrap confidence interval of [−0.0001, 0.0005]. At this level, Model C improved upon Model B for approximately 44.80% of the 25,940 item-store series and worsened performance for approximately 54.02%. Taken together with the window-level results, these findings indicate that the three lagged macroeconomic indicators provided negligible aggregate incremental forecasting value over Model B, with heterogeneous and inconsistent effects across forecast windows and item-store series. The paired MAE reductions and their 95% bootstrap confidence intervals for all three model comparisons are summarized in Table 3 and illustrated in Figure 3.

Table 3. Paired MAE comparisons across item-store–window observations

Comparison	N	Mean MAE Reduction	Relative Improvement (%)	Improved (%)	Worsened (%)	Wilcoxon p-value	95% Bootstrap CI
Model A vs. Seasonal Naïve	129,700	0.2875	22.0167	82.2783	8.276	<0.001	[0.2846, 0.2904]
Model B vs. Model A	129,700	0.0062	0.606	38.8921	46.1149	<0.001	[0.0055, 0.0068]
Model C vs. Model B	129,700	0.0002	0.0169	42.0786	45.953	<0.001	[−0.0001, 0.0005]

Note. Positive mean MAE reduction indicates lower average MAE for the first model named in each comparison. Confidence intervals are 95% bootstrap intervals based on 2,000 iterations. The Wilcoxon signed-rank test evaluates paired error distributions and should be interpreted together with the magnitude of the mean difference and its bootstrap confidence interval.

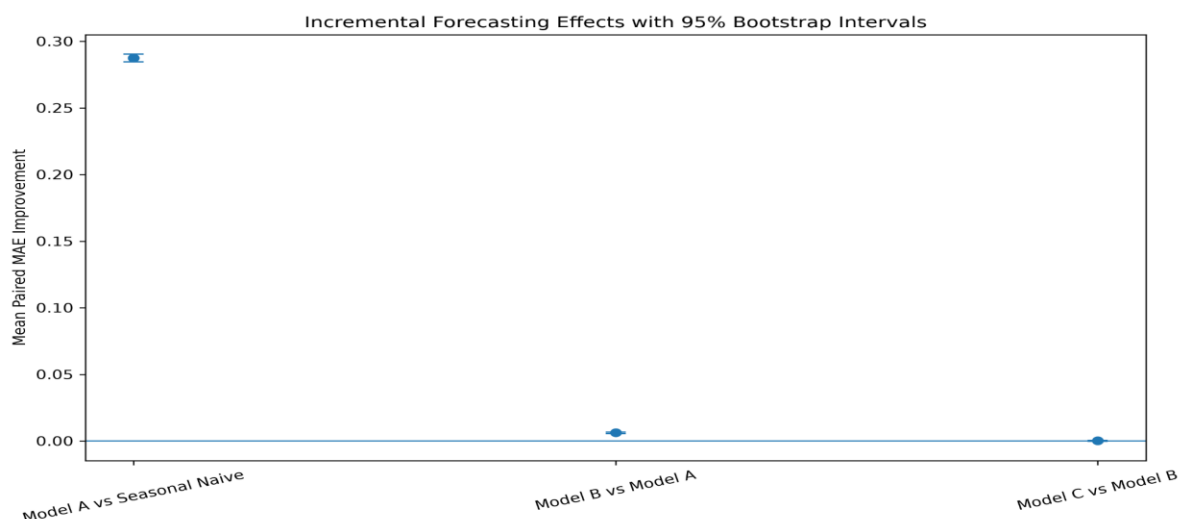


Figure 3. Mean paired MAE reductions and 95% bootstrap confidence intervals for the incremental forecasting comparisons. Positive values indicate lower mean MAE for the first model in each comparison; the horizontal zero line indicates no mean MAE improvement.

The series-average robustness analysis produced conclusions consistent with the primary series-window comparisons. After averaging forecast errors across the five evaluation windows for each of the 25,940 item-store series, Model A reduced mean MAE by 0.2875 relative to the seasonal-naive benchmark, corresponding to a relative improvement of approximately 22.017%, and achieved lower average MAE for 98.30% of the series. Model B reduced mean MAE by 0.0062 relative to Model A, an improvement of approximately 0.606%, although lower average MAE was observed for 45.74% of the series and higher average MAE for 52.54%. Model C produced a mean MAE reduction of only approximately 0.0002 relative to Model B, corresponding to approximately 0.017%, with improvement observed for 44.80% of the series and deterioration for 54.02%. The bootstrap confidence intervals remained clearly above zero for Model A versus the seasonal-naive benchmark [0.2828, 0.2919] and Model B versus Model A [0.0052, 0.0071], whereas the interval for Model C versus Model B [−0.0001, 0.0005] included zero. This robustness analysis therefore reinforces the large improvement associated with Model A, the small positive average contribution of the retail and operational predictors, and the absence of a robust positive mean-error improvement from the additional macroeconomic indicators.

A complementary model-win analysis further illustrated the heterogeneity in forecasting performance across individual series-window observations. Among the 129,700 item-store–window cases, Model A achieved the lowest MAE in 53,436 cases (41.20%), representing the largest share among the four forecasting approaches. Model B achieved the lowest MAE in 28,388 cases (21.89%), while Model C was the best-performing specification in 27,103 cases (20.90%). The seasonal-naive benchmark achieved the lowest MAE in the remaining 20,773 cases (16.02%). These results show that no single forecasting specification performed best for every item-store series and forecast window.

In particular, although Models B and C produced slightly lower aggregate error measures than Model A, Model A remained the most frequent individual series-window winner, highlighting substantial cross-series and temporal heterogeneity in the benefits of the additional predictor groups, as summarized in Table 4 and illustrated in Figure 4.

Table 4. Frequency of lowest-MAE performance across series-window observations

Forecasting Approach	Number of Wins	Win Rate (%)
Seasonal Naive	20,773	16.02
Model A	53,436	41.20
Model B	28,388	21.89
Model C	27,103	20.90
Total	129,700	100

Note. A win indicates that the forecasting approach achieved the lowest MAE among the four approaches for a given item-store-window observation. Percentages are rounded to two decimal places.

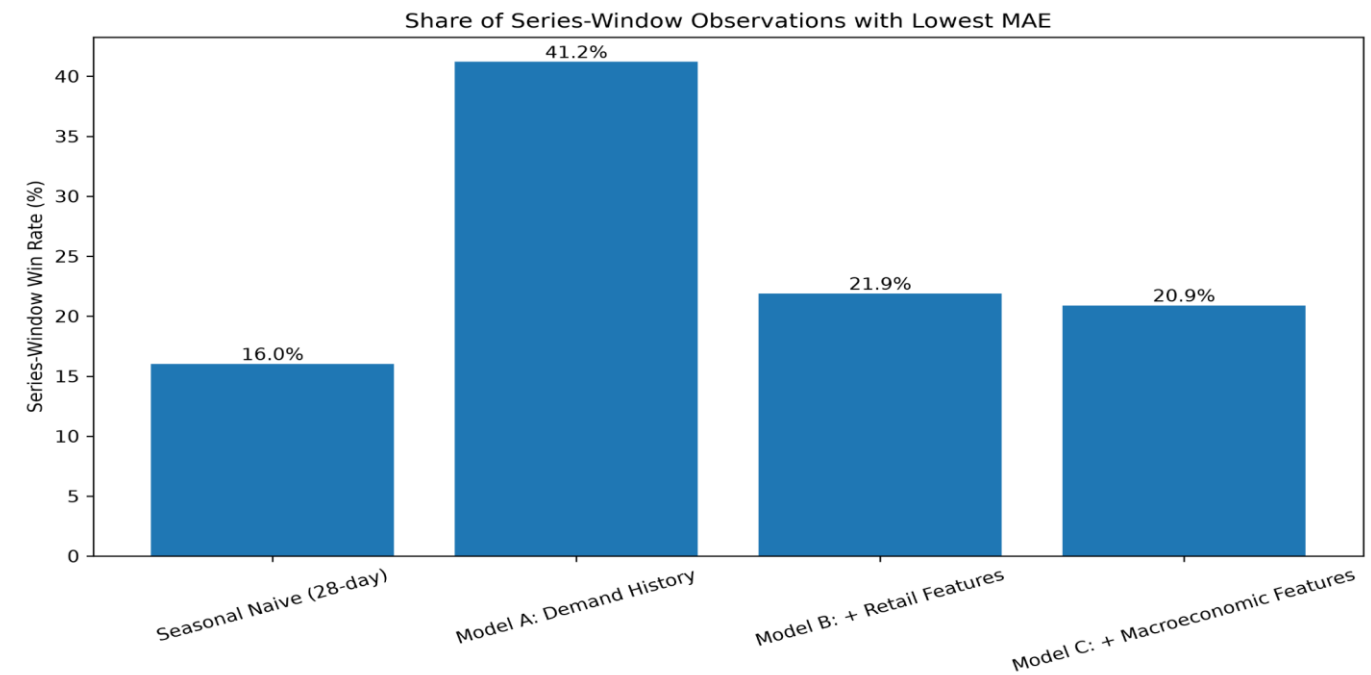


Figure 4. Share of series-window observations for which each forecasting approach achieved the lowest MAE across 129,700 item-store-window comparisons.

Overall, the results reveal a clear hierarchy in the magnitude of forecasting gains associated with the progressively expanded predictor sets. The largest improvement was achieved by Model A relative to the seasonal-naive benchmark, demonstrating the substantial predictive value captured by the historical-demand features within the LightGBM framework. The addition of retail and operational information in Model B yielded a smaller but positive average improvement, although its benefits varied considerably across individual item-store series. In contrast, the addition of the three lagged macroeconomic indicators in Model C produced negligible aggregate improvement and inconsistent effects across forecast windows and series. The combined evidence from aggregate accuracy measures, window-level performance, paired statistical comparisons, bootstrap confidence intervals, and robustness analyses therefore indicates that most of the forecasting improvement was associated with historical demand information, with more limited incremental gains from the additional predictor groups.

5. Discussion

The findings demonstrate that the predictive value of the examined information sources was highly uneven. Historical demand information accounted for the dominant improvement in forecast accuracy, as the history-based LightGBM specification substantially outperformed the seasonal-naive benchmark across all five evaluation windows. Retail and operational variables provided additional predictive information, but their average contribution was considerably smaller and heterogeneous across item-store series. In contrast, incorporating the three lagged macroeconomic indicators—CPI, unemployment, and retail sales—provided little additional aggregate forecasting value beyond the demand-history and retail/operational predictors. The results therefore suggest that, for the forecasting setting examined in this study, progressively increasing the breadth of available information did not translate proportionally into improved predictive accuracy. Rather, the usefulness of additional information depended on whether it contributed sufficiently relevant predictive signals beyond that already captured by historical demand patterns and retail-level covariates.

The substantial improvement of Model A over the seasonal-naive benchmark highlights the importance of representing demand history through multiple temporal features rather than relying on a single seasonal reference point. Whereas the benchmark generated forecasts using demand observed 28 days earlier, Model A incorporated multiple lagged observations at 28, 35, 42, and 56 days together with 7-day and 28-day rolling means shifted by 28 days. This broader representation allowed the model to

use information from several historical horizons and smoothed demand patterns simultaneously, while the nonlinear structure of LightGBM enabled these predictors to be combined flexibly across the large collection of item-store series. The magnitude and consistency of the observed improvement suggest that these historical-demand features captured substantial predictive information that was unavailable to the single-lag seasonal-naïve benchmark. Importantly, however, the comparison reflects both the richer historical feature representation and the use of the LightGBM modeling framework; it should therefore not be interpreted as isolating the effect of any individual lag or rolling feature.

The transition from Model A to Model B indicates that retail and operational information provided additional predictive value beyond historical demand, but that this contribution was modest and heterogeneous. Model B augmented the historical-demand features with weekday, month, SNAP participation, event occurrence, and selling price information, resulting in a small average improvement in MAE and WAPE. However, the paired analyses showed that the benefit was not uniform across item-store series, with some series improving and others experiencing higher forecast errors. This pattern suggests that the usefulness of retail and operational covariates may depend on the demand characteristics and commercial context of individual products and stores. Calendar effects, events, SNAP-related purchasing patterns, and price information can provide context that is not represented directly by lagged demand, but their incremental predictive signal may overlap with patterns already embedded in recent demand history. Consequently, the results support the inclusion of retail and operational information as a potentially useful extension to historical-demand features while also demonstrating that a richer predictor set does not necessarily improve forecasts for every individual series.

The transition from Model B to Model C provides a more cautious assessment of the incremental predictive value of macroeconomic information. Adding the two-month-lagged CPI, unemployment, and retail sales indicators produced only a negligible reduction in aggregate forecast error, and the direction of the change was not consistent across the five evaluation windows. Although the Wilcoxon signed-rank test detected differences in the paired error distributions, the bootstrap confidence intervals for the mean MAE reduction included zero in both the series-window and series-average analyses. Given the very large number of paired observations, the small Wilcoxon p-values should therefore be considered together with the magnitude of the estimated improvement and its bootstrap uncertainty rather than interpreted alone as evidence of a meaningful forecasting gain. In this setting, the three macroeconomic indicators did not provide a robust positive improvement in mean forecast error beyond the information already contained in historical demand and retail and operational predictors. This finding does not imply that macroeconomic conditions are generally irrelevant to retail demand; rather, it indicates that the specific indicators, lag structure, forecast horizon, and modeling design examined here yielded little additional aggregate predictive value.

The results also reveal substantial heterogeneity in the relative performance of the forecasting specifications across item-store series and forecast windows. Although Models B and C achieved slightly lower aggregate error measures than Model A, Model A recorded the lowest MAE in the largest proportion of individual series-window cases, accounting for 41.20% of the 129,700 comparisons. Similarly, the incremental extensions did not improve every series: Model B produced lower series-window MAE than Model A in 38.89% of comparisons, while Model C improved upon Model B in 42.08%. Temporal variation was also evident, particularly for the macroeconomic extension, as Model C achieved lower WAPE than Model B in only two of the five forecast windows. These patterns demonstrate that small improvements in aggregate forecasting metrics can coexist with substantial variation in model performance across individual forecasting units and time periods. Accordingly, model evaluation in large-scale retail forecasting should consider not only overall average accuracy but also the distribution and stability of forecasting gains across series and forecast origins.

The findings are broadly consistent with prior evidence demonstrating the value of machine-learning approaches and enriched predictor information in large-scale retail forecasting, while also providing a more nuanced view of the incremental contribution of external variables. Evidence from the M5 Accuracy Competition showed that machine-learning approaches can perform strongly in large-scale retail forecasting settings and that explanatory variables can contribute useful forecasting information (Makridakis et al., 2022). The present results are consistent with this broader evidence, as the LightGBM specifications substantially outperformed the seasonal-naïve benchmark and the addition of retail and operational predictors produced a further, although modest, average improvement. At the same time, previous research has emphasized the potential value of incorporating macroeconomic conditions, including CPI and unemployment, into retail-demand forecasting models (Haque et al., 2023). The present study qualifies this expectation by showing that, once historical-demand and retail and operational information were already represented, adding CPI, unemployment, and retail sales with a two-month operational lag yielded negligible aggregate incremental improvement. This difference should not be interpreted as a contradiction, because the predictive contribution of macroeconomic information may depend on the dataset, forecast horizon, variable construction, lag structure, modeling framework, and information already contained in the baseline predictor set. Rather, the results emphasize

the importance of evaluating the incremental predictive value of external information under realistic forecasting conditions instead of assuming that additional variables will necessarily improve forecast accuracy.

From a practical supply-chain forecasting perspective, the findings suggest that expanding model complexity through additional data sources should be guided by demonstrated incremental predictive value rather than by data availability alone. The substantial improvement achieved using historical-demand features indicates that organizations can obtain considerable forecasting gains from systematically structuring and modeling their internal demand histories. Retail and operational information may provide further value, particularly when calendar, event, SNAP, and planned pricing information is available at the forecast origin, although the observed gains in this study were comparatively modest and heterogeneous. By contrast, the negligible aggregate improvement associated with the three lagged macroeconomic indicators suggests that integrating external economic data can introduce additional data-engineering and maintenance requirements without necessarily producing commensurate improvements in short-horizon item-store forecasting accuracy. For practitioners, these results support an incremental model-development strategy in which additional predictor groups are evaluated against a clearly defined baseline and their forecasting contribution is assessed in terms of its magnitude, consistency, and practical relevance.

Taken together, the findings demonstrate that the value of additional information in large-scale retail demand forecasting is incremental rather than uniform. Historical demand features provided the strongest predictive foundation in the present modeling framework, retail and operational variables generated a smaller additional average benefit, and the three lagged macroeconomic indicators contributed little further improvement at the aggregate level. At the same time, variation across item-store series and forecast windows shows that aggregate performance measures alone do not fully characterize the usefulness of progressively richer predictor sets. These findings underscore the importance of evaluating forecasting models not only according to whether additional variables improve average accuracy, but also according to the magnitude, consistency, and robustness of those improvements across forecasting units and time periods. This perspective provides the basis for considering the methodological and empirical limitations of the present study and the opportunities for extending the analysis in future research.

6. Limitations and Future Research

6.1 Limitations

Despite the forecasting improvements observed in this study, several limitations should be considered when interpreting the findings. First, computational constraints required the LightGBM models to be trained on 2,600,000 observations from 9,468,100 eligible training observations in each forecast window, corresponding to approximately 27.46% of the available training data. Although the same deterministic sampling procedure was applied across Models A, B, and C and the complete evaluation samples were retained, training on a subset of the available historical observations may have influenced model estimation. In addition, the analysis was restricted to LightGBM with a fixed model configuration. This design facilitated comparison of the incremental predictive value of different predictor sets within a common modeling framework, but the findings do not establish that LightGBM is superior to alternative forecasting methods or that the observed contribution of the predictor groups would remain unchanged under different model architectures.

The temporal and external-data design also imposes limitations. The evaluation covered five non-overlapping 28-day forecast windows from January through May 2016 and therefore may not capture model performance across a broader range of seasonal patterns or economic conditions. Moreover, the macroeconomic extension was limited to CPI, unemployment, and retail sales using a two-month operational availability lag, and exact historical data-release vintages were not reconstructed. The negligible incremental improvement observed for Model C should therefore be interpreted specifically within this indicator set and lag structure rather than generalized to all macroeconomic information. Models B and C also assumed that planned or future selling-price information was available at the forecast origin, an assumption that may not hold in all operational retail environments.

Additional limitations concern interpretation and generalizability. The study evaluated predictor groups collectively and did not estimate the contribution of individual variables or examine performance across subgroups such as states, stores, product categories, or demand patterns. The findings therefore cannot identify which individual retail, operational, or macroeconomic variables were responsible for the observed changes in forecasting performance. Furthermore, the study was predictive rather than causal, and associations between the included predictors and forecast accuracy should not be interpreted as causal effects on retail demand. The primary paired analysis also contained repeated observations of the same 25,940 item-store series across five forecast windows; although the series-average robustness analysis provided a more conservative assessment and produced consistent overall conclusions, it does not eliminate all potential dependence within the hierarchical retail data. Finally, because the analysis was conducted using the M5 dataset, the findings should not be assumed to generalize automatically to other retailers, industries, geographic markets, or forecasting horizons.

6.2 Future Research

Future research should extend the present analysis across broader datasets, forecasting periods, and modeling frameworks to assess the robustness and generalizability of the findings. The progressive predictor framework could be evaluated using the complete eligible training data, additional rolling forecast windows, alternative forecast horizons, and other forecasting approaches, including statistical, machine-learning, neural, and ensemble models. Further investigation of external information could incorporate additional macroeconomic indicators, alternative lag structures, higher-frequency variables, and historical data-release vintages, while alternative representations of future price information could be examined under different operational availability assumptions. Feature-ablation and model-interpretability methods such as SHAP could also be used to quantify the contribution of individual predictors, while subgroup analyses could determine whether predictive performance differs across stores, product categories, demand patterns, or geographic markets. Finally, replication using additional retail and supply-chain datasets would help determine whether the substantial forecasting gains associated with the history-based LightGBM specification and the limited incremental contribution of the examined macroeconomic indicators persist beyond the M5 forecasting environment.

7. Conclusion

This study addressed the challenge of improving demand forecasting in large-scale U.S. retail supply chains by evaluating how progressively richer information sets contribute to forecasting accuracy within an AI-driven machine learning framework. Using the M5 dataset, three LightGBM specifications incorporating historical demand, retail and operational information, and lagged macroeconomic indicators were evaluated against a 28-day seasonal-naïve benchmark across 25,940 item-store series and five rolling forecast windows. The history-based LightGBM specification produced the dominant forecasting improvement, reducing mean WAPE from 0.9006 for the seasonal-naïve benchmark to 0.7022, an improvement of approximately 22.03%. Adding retail and operational information further reduced mean WAPE to 0.6979, although the incremental improvement was comparatively modest and heterogeneous across item-store series. Incorporating lagged CPI, unemployment, and retail sales indicators produced only a negligible further reduction to 0.6978, with no robust positive improvement in mean MAE relative to Model B. These findings demonstrate that addressing large-scale retail demand-forecasting challenges does not necessarily require continuously expanding the number or breadth of data sources. Rather, substantial forecasting gains can be achieved by effectively modeling historical demand information, while additional retail, operational, and external economic variables should be evaluated according to their demonstrated incremental predictive contribution. For U.S. retail supply chains, improved forecasting accuracy can support more informed inventory and replenishment planning, strengthen demand visibility, and potentially help retailers mitigate stockout and excess-inventory risks associated with inaccurate demand forecasts. The study therefore contributes an empirical framework for evaluating how different information sources can be incorporated into machine-learning-based demand forecasting while avoiding the assumption that greater data complexity will necessarily yield meaningful forecasting improvements. Future research should evaluate the framework across broader retail and supply-chain datasets, additional forecast periods and horizons, alternative machine-learning architectures, and richer representations of external economic information to determine whether these forecasting patterns persist across different operational environments.

Funding: No Funding is available for this research

Conflicts of Interest: The author(s) have no conflict of interest to declare

Ethical Statement : This study did not require ethical approval because it used the publicly available M5 Forecasting – Accuracy dataset and did not involve human participants, personally identifiable information, or the collection of primary human-subject data.

References

- [1]. Babai, M. Z., Boylan, J. E., & Rostami-Tabar, B. (2022). *Demand forecasting in supply chains: A review of aggregation and hierarchical approaches*. *International Journal of Production Research*, 60(1), 324–348.
<https://doi.org/10.1080/00207543.2021.2005268>
- [2]. Efron, B. (1979). *Bootstrap methods: Another look at the jackknife*. *The Annals of Statistics*, 7(1), 1–26.
<https://doi.org/10.1214/aos/1176344552>
- [3]. Fildes, R., Ma, S., & Kolassa, S. (2022). *Retail forecasting: Research and practice*. *International Journal of Forecasting*, 38(4), 1283–1318. <https://doi.org/10.1016/j.ijforecast.2019.06.004>
- [4]. Haque, M. S., Amin, M. S., Miah, J., Cao, D. M., Sayed, M. A., & Ahmed, S. (2023). *Retail demand forecasting: A comparative study for multivariate time series*. *Journal of Mathematics and Statistics Studies*, 4(4), 40–46.
<https://doi.org/10.32996/jmss.2023.4.4.4>

- [5]. Huber, J., & Stuckenschmidt, H. (2020). *Daily retail demand forecasting using machine learning with emphasis on calendric special days*. *International Journal of Forecasting*, 36(4), 1420–1438. <https://doi.org/10.1016/j.ijforecast.2020.02.005>
- [6]. Hyndman, R. J., & Koehler, A. B. (2006). *Another look at measures of forecast accuracy*. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- [7]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. *Advances in Neural Information Processing Systems*, 30, 3146–3154. <https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
- [8]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). *M5 accuracy competition: Results, findings, and conclusions*. *International Journal of Forecasting*, 38(4), 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>
- [9]. Tashman, L. J. (2000). *Out-of-sample tests of forecasting accuracy: An analysis and review*. *International Journal of Forecasting*, 16(4), 437–450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
- [10]. Wilcoxon, F. (1945). *Individual comparisons by ranking methods*. *Biometrics Bulletin*, 1(6), 80–83. <https://doi.org/10.2307/3001968>