logistics

MDPI

Article

Class-Conditional Allocation of Forecasting Models for Retail Inventory Replenishment: Evidence from an ABC-Segmented Hardware Portfolio

Gina Marcela Hernández-Orozco [1], Juan David Reales-Barragán [1], Verónica Deibe-Blanco [2], Rick Acosta-Vega [1] and Manuel J. Campuzano [1,*]

[1] Facultad de Ingeniería, Universidad del Magdalena, Santa Marta 470004, Colombia; gmhernandezo@unimagdalena.edu.co (G.M.H.-O.); jdreales@unimagdalena.edu.co (J.D.R.-B.); racosta@unimagdalena.edu.co (R.A.-V.)
[2] Escuela de Ciencias Básicas, Tecnología e Ingeniería, Universidad Nacional Abierta y a Distancia, Bogotá 111511, Colombia; veronica.deibe@unad.edu.co
* Correspondence: mcampuzano@unimagdalena.edu.co

Abstract

Background: Retail inventories are heterogeneous, yet demand forecasting studies typically select one model for an entire portfolio, leaving open how forecasting effort should be allocated across inventory segments. *Methods*: Six models—ARIMA, SARIMA, ARIMAX, Random Forest, LightGBM and an LSTM network—were estimated on weekly demand for a Colombian hardware retailer (35,399 transactions, 1633 SKUs, January 2025–March 2026). Because 95.5% of items are intermittent or lumpy, models were fitted to three ABC-class-aggregated series of 60 observations, evaluated on a 12-week holdout and under a rolling-origin scheme, and compared against naive benchmarks using MASE and RMSSE. *Results*: The best model differs by class. In Class A a statistical specification wins, ARIMAX on the holdout (RMSE 843 units per week) but ARIMA under rolling origins, the two being statistically indistinguishable. Random Forest is best in Class B (255). In Class C no model improves on a last-value rule (75 against 66). Relative to that benchmark, error falls to 39.8% in Class A and 47.6% in Class B. Diebold–Mariano tests separate 16 of 45 pairs on the holdout and three of 30 under rolling origins. *Conclusions*: Allocating models by inventory class is operationally justified rather than statistically decisive, and the low-turnover class warrants no model at all.

Keywords: inventory replenishment; demand forecasting; ABC inventory segmentation; forecast model allocation; retail logistics; decision support

Check for updates

Academic Editor: Rameshwar Dubey

Received: 10 August 2026
Revised: 6 October 2026
Accepted: 9 October 2026
Published: 13 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Background and Motivation

Effective inventory planning and replenishment are fundamental to retail operations because they directly influence product availability, customer service, working capital, and organizational profitability. These decisions become increasingly challenging in environments characterized by volatile demand, heterogeneous product portfolios, and limited analytical capabilities. Small- and medium-sized enterprises (SMEs) are particularly vulnerable to these conditions, as replenishment decisions are frequently based on managerial experience rather than systematic analytical support. Consequently, inaccurate demand

forecasts can propagate throughout the supply chain, leading to stock shortages, excessive inventories, inefficient capital allocation, and reduced operational performance [1–4].

The Colombian retail sector exemplifies these challenges. According to the Annual Trade Survey published by the National Administrative Department of Statistics (DANE), retail trade represents one of the most dynamic sectors of the national economy, contributing consistently to the country's gross domestic product while exhibiting substantial temporal variability [5]. Likewise, the Monthly Retail Trade Survey reports pronounced short-term fluctuations together with seasonal changes associated with macroeconomic conditions and consumer purchasing behavior [6]. Similar volatility has been documented worldwide, where retail markets are increasingly exposed to economic uncertainty, supply chain disruptions, and rapidly changing customer preferences [7,8]. Such conditions considerably increase the complexity of inventory planning and reinforce the need for robust demand forecasting strategies capable of supporting replenishment decisions.

These operational challenges are evident in FERRETODO, a Colombian hardware retailer located in the Department of Magdalena. The company manages a large portfolio of stock keeping units (SKUs), while purchasing decisions are primarily supported by managerial intuition instead of quantitative forecasting techniques. Nevertheless, the company possesses detailed historical sales records at the SKU level, providing an opportunity to transform operational data into systematic decision support. Previous studies have consistently demonstrated that forecasting accuracy significantly influences inventory performance, replenishment efficiency, and customer service levels, particularly in retail environments characterized by heterogeneous demand patterns [1–4].

Demand forecasting has traditionally relied on statistical time-series models such as ARIMA and SARIMA because of their interpretability and computational efficiency [8–10]. More recently, machine-learning algorithms—including Random Forest, LightGBM, and Long Short-Term Memory (LSTM) networks—have demonstrated strong predictive capabilities by capturing non-linear relationships and complex temporal dependencies [11–16]. However, comparative studies indicate that forecasting performance is strongly conditioned by demand characteristics, seasonality, forecasting horizons, data availability, and operational context [11–17]. Consequently, no single forecasting model can be assumed to consistently outperform all alternatives across heterogeneous inventory environments.

1.2. Demand Patterns in Retail Inventory Systems

A recurring simplification in applied forecasting studies is to treat the demand faced by a retailer as a single homogeneous process. In practice several qualitatively different demand regimes coexist within the same warehouse and are served by the same replenishment system. The standard characterization, proposed by Syntetos, Boylan and Croston [18], classifies a series using the average inter-demand interval (ADI), which measures how often demand occurs, and the squared coefficient of variation of the non-zero demand sizes (CV^2), which measures how variable those quantities are. The conventional cut-offs ADI = 1.32 and $CV^2 = 0.49$ delimit four regimes, all present in a hardware retailer's portfolio.

Smooth demand (ADI $\leq$ 1.32, $CV^2 \leq 0.49$) consists of frequent orders of comparable size, typical of construction commodities such as cement, sand, PVC piping and common fasteners; exponential smoothing and ARIMA-type formulations are usually adequate for these items [8,19]. Erratic demand (ADI $\leq$ 1.32, $CV^2 > 0.49$) consists of frequent orders of highly variable size, typical of paint, electrical cable or roofing sheets, where the quantity per transaction depends on whether the buyer is a household or a contractor.

Intermittent demand (ADI > 1.32, $CV^2 \leq 0.49$) consists of infrequent orders of similar size, such as specific plumbing fittings or specialized drill bits. These series contain a high proportion of zero periods, which biases smoothing-based estimators; Croston's

method and its later corrections were developed for this regime [18,20]. Lumpy demand (ADI > 1.32, CV^2 > 0.49) consists of infrequent orders of highly variable size, such as welding equipment, large-diameter valves or industrial power tools, and is the least predictable of the four.

The operational consequence is that a model correctly specified for one regime is frequently misspecified for another. A Gaussian ARIMA formulation estimated on a lumpy series produces forecasts concentrated around a mean the series almost never attains, while a flexible learner estimated on a short smooth series fits noise that will not recur. Since a single retailer holds all four regimes simultaneously, the question of practical interest is not which model performs best overall but how to route each part of the portfolio to an appropriate model using information the planner already has.

1.3. Related Work and Research Gap

The problem of linking demand forecasting to inventory replenishment is not new, and the literature that addresses it is mature. Four strands are relevant here, and this study is positioned against each of them explicitly.

The first strand is theoretical and establishes why forecasting and replenishment cannot be treated as separate problems. Aviv [21] formalizes a supply chain in which demand follows a linear state-space process and shows that the value of information, and therefore the appropriate replenishment policy, depends on which components of the underlying state each member can observe. The implication that matters here is that the structure of the demand process, not the sophistication of the estimator, determines how much forecasting can contribute. Aviv assumes that structure is known and common; the present study addresses the case in which it is unknown and, as the demand-regime analysis in Section 4.7 shows, heterogeneous across items within a single portfolio.

The second strand is the empirical retail-operations tradition, which has repeatedly shown that the binding constraint in practice is rarely the estimator. Cachon and Fisher [22] evaluate a continuous replenishment program at Campbell Soup and find that most of the benefit comes from improved decision rules rather than from information sharing alone. Fisher, Rajaram and Raman [23] optimize replenishment for fashion products where the season is short and history is scarce, a data regime close to the one studied here. Fisher and Vaidyanathan [24] estimate demand for assortment decisions and report results from implementation rather than simulation. Reviewing three decades of this work, Fisher and Raman [25] conclude that the durable advances in retail operations have come from procedures that firms could actually adopt. This study belongs to that tradition and adopts its standard of evidence: a procedure implemented on the operational data of a real firm, with its failures reported alongside its successes.

The third strand is the recent literature on adaptive replenishment, which supplies the methodological frontier against which any new proposal must be judged. Lim, Jiu and Ang [26] integrate anticipative allocation with reactive fulfilment for online retailing using robust optimization. Park, Choi and Min [27] exploit the structure of the optimal policy to make reinforcement learning tractable for replenishment, and Kolyaei, Zhang and Blom [28] apply a reinforcement-learning method to joint replenishment and fulfilment for an omnichannel retailer. Rohilla, Kundu, Kapoor and Sheu [29] use system dynamics to study replenishment when products shift along their life cycle. These approaches are more powerful than anything proposed here, and the present study does not compete with them. They require long histories, controlled environments or simulation models that a firm of the kind studied here does not possess: the dataset analyzed below contains sixty weeks of history, no stockout records and no lead-time series, which is the

normal condition of a small retailer rather than an exceptional one. The contribution of this study is to establish what can be decided with that much information.

The fourth strand is inventory segmentation. Silver, Pyke and Thomas [1] establish the classification of items by economic relevance as the basis for differentiated control policies, and Scholz-Reiter, Heger, Meinecke and Bergmann [30] integrate demand forecasts into segmentation in order to improve replenishment scheduling. In both cases the segmentation acts downstream of forecasting: it determines review intervals, service targets and safety stocks but not which forecasting method is estimated in the first place. Henao-Baena et al. [31], the closest study in context, applies machine learning to inventory management in a Colombian setting and reports results by inventory class, but the class does not govern the choice of model.

Cutting across these strands, large-scale empirical evidence indicates that no single forecasting method dominates. The M5 competition and the accompanying methodological reviews report that relative performance depends on the series, the horizon and the amount of history available and that statistical benchmarks remain competitive with, and sometimes superior to, considerably more complex learners [7,32,33]. Arunraj, Ahrens and Fernandes [10] reach a compatible conclusion for daily food retail, where calendar regressors improve a statistical specification more than added model complexity does, while Punia et al. [3], Falatouri et al. [17] and Nasseri et al. [16] each report that the best learner varies with the product group examined. It is precisely this absence of a dominant method that makes the question of model allocation, rather than model selection, operationally relevant.

Four gaps follow from this review. First, model comparisons are typically conducted at the portfolio level and resolved with a single winner, which implicitly assumes homogeneity of demand behavior across items known to be heterogeneous. Second, inventory segmentation is used to differentiate control parameters but not forecasting methods, so information already available to the planner is only partially exploited. Third, differences in accuracy between competing models are usually reported as point estimates without testing whether they are statistically distinguishable. Fourth, the evidence base is dominated by large retailers with long histories and dedicated analytics teams, while small- and medium-sized enterprises with short histories and no forecasting staff remain underrepresented, and Latin American retail SMEs particularly so.

The theoretical contribution claimed here is correspondingly narrow and is stated plainly. Following Aviv's result that the contribution of forecasting depends on the demand process rather than on the estimator [21], this study provides an empirical counterpart at the level of the inventory segment: it shows that the marginal value of forecasting effort varies systematically across segments of the same portfolio, that it can be substantial in one segment and zero in another, and that the segment boundary itself, and not only the model, determines the operational answer. The practical contribution is a reproducible allocation procedure that a firm without analytics staff can operate with open-source tools, together with an account of the obstacles that remain after it is deployed (Section 5.2.2).

1.4. Objective and Contributions

To address this gap, this study proposes a class-conditional forecasting framework that integrates ABC inventory segmentation with forecast model allocation for replenishment planning. Rather than searching for a universally superior forecasting algorithm, the proposed framework formalizes ABC classification as a routing mechanism that assigns alternative forecasting strategies—including ARIMA, SARIMA, ARIMAX, Random Forest, LightGBM, and LSTM—to inventory segments exhibiting distinct economic and demand characteristics. Forecasting performance is evaluated using multiple accuracy measures

together with the Diebold–Mariano statistical test, providing both predictive and statistical evidence within a unified evaluation framework.

The proposed framework is validated using historical weekly sales data comprising 1633 SKUs from a Colombian hardware SME. Inventory items are first classified according to their economic relevance through ABC analysis, after which forecasting strategies are trained and evaluated using demand aggregated at the ABC-class level. This distinction is important because the present study constitutes a class-level validation of the proposed framework rather than an individual SKU-level forecasting implementation. Unlike conventional model-comparison studies, the proposed approach transforms inventory segmentation into an explicit decision-support mechanism for forecast model allocation, thereby linking demand forecasting with inventory replenishment planning within an integrated analytical framework.

Although the dataset covers 1633 SKUs, the forecasting models are estimated on three class-aggregated weekly series rather than on individual items. This is a methodological requirement rather than a simplification: 95.5% of the portfolio exhibits intermittent or lumpy demand at the item level, a regime for which the ARIMA family and recurrent architectures are inappropriate and for which Croston-type estimators are indicated, and 539 SKUs record demand in three weeks or fewer out of sixty. Aggregation by inventory class removes intermittency entirely, since none of the three aggregate series contains a zero-demand week, and it corresponds to the level at which replenishment policy is set in the firm studied. The extension of the framework to SKU-level forecasting is identified as future work in Section 5.3.

Accordingly, this research makes four principal contributions. First, it proposes a class-conditional inventory forecasting framework that formalizes ABC inventory segmentation as a routing mechanism for forecast model allocation rather than solely as an inventory classification technique. Second, it establishes a direct integration between inventory segmentation and differentiated forecasting, providing empirical evidence that forecasting performance is conditional on inventory characteristics and that different inventory classes may benefit from different forecasting strategies. Third, it incorporates statistically validated forecast model allocation by combining conventional forecast accuracy measures with the Diebold–Mariano test, providing statistical evidence regarding differences in predictive performance. Fourth, the framework demonstrates industrial applicability through validation with real operational data from a retail SME, providing a practical and transferable approach to support inventory replenishment planning in data- and resource-constrained environments.

The rest of the paper is organized as follows. Section 2 presents the research design and methodology. Section 3 provides an overview of the class-conditional forecasting approach and architecture of the modeling pipeline. Section 4 provides the results of the empirical analysis and includes an interpretation of major findings. Section 5 provides observations on the importance of class-conditional forecasting and its role in retail inventory management and logistics. It also presents the limitations of this work. Finally, Section 6 summarizes the main results and presents a number of possible future research initiatives.

2. Methodology

2.1. Overall Research Framework

The study adopts a quantitative, non-experimental, descriptive–comparative research design. Six forecasting approaches are evaluated, comprising two methodological families: statistical time-series models (ARIMA, SARIMA, and ARIMAX) and machine-learning models (LSTM, Random Forest, and LightGBM).

The principal methodological contribution lies in the formal integration of ABC inventory segmentation with class-conditional forecast model allocation. Unlike conventional forecasting studies, where models are trained and evaluated at the individual SKU level, the proposed framework develops and assesses forecasting models using demand aggregated by the ABC inventory class. This strategy transforms inventory segmentation from a descriptive inventory classification technique into a decision-support mechanism that guides forecast model allocation according to inventory characteristics, thereby aligning forecasting strategies with replenishment requirements [31].

The proposed framework is validated using historical transactional data from FERRETODO, a medium-sized hardware retailer located in the Department of Magdalena, Colombia. Before describing the data preparation process and the forecasting framework, it is useful to characterize the temporal behavior of the sales data. Figure 1 presents a calendar heatmap of daily revenue during the study period, providing an overview of the demand variability that motivates the adoption of class-conditional forecasting strategies.

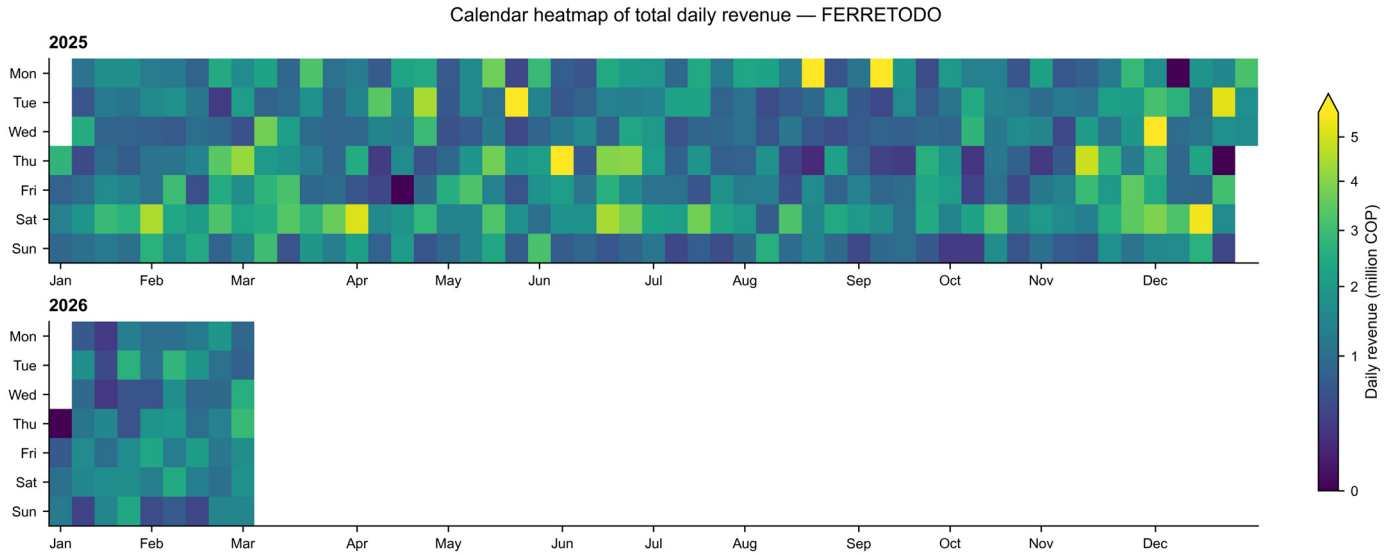


Figure 1. Calendar heatmap of FERRETODO's total daily revenue (2025–2026).

Figure 1 presents a calendar heatmap of daily revenue over the study period. Two features are relevant for the forecasting design. First, a stable weekly rhythm is visible throughout the period, with weekend trading days concentrating a systematically higher share of revenue, consistent with the purchasing behavior of households and small contractors in the hardware sector. Second, superimposed on that regular pattern, a limited number of isolated days reach revenue levels several times the daily median; these correspond to individual large-volume orders rather than to a seasonal effect. The coexistence of a stable weekly cycle with sporadic high-magnitude events is what motivates weekly aggregation: it retains the demand level relevant for replenishment planning while preventing single large orders from dominating the series.

2.2. Forecasting Models

2.2.1. ARIMA Models and Variants

ARIMA (AutoRegressive Integrated Moving Average) models constitute the classical statistical paradigm for forecasting non-stationary time series [8,9]. An ARIMA(p, d, q) model is expressed as

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t \quad (1)$$

where y_t is the observed time series, B is the lag operator ($By_t = y_{t-1}$), $\varphi(B)$ is the autoregressive polynomial of order p, $\theta(B)$ is the moving-average polynomial of order q, d is the differencing order required to achieve stationarity, and ε_t is a white-noise error term.

SARIMA incorporates seasonal parameters (P, D, Q, m) to capture cyclical patterns [10]:

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = \Theta(B^m)\theta(B)\varepsilon_t \quad (2)$$

where m is the seasonal periodicity (e.g., m = 52 for weekly data with annual seasonality).

ARIMAX extends the model by incorporating external regressors X_t, such as promotional variables or economic indicators [17]:

$$\phi(B)(1-B)^d y_t = \beta' X_t + \theta(B)\varepsilon_t \quad (3)$$

In this study, the exogenous vector X_t comprises two calendar-based regressors derived from the official Colombian public holiday calendar; these are defined explicitly in the Exogenous Variables subsection of the Methodology.

2.2.2. Machine-Learning Techniques for Forecasting

Ensemble methods. Tree-based algorithms constitute the machine-learning core of this study because of their robustness against overfitting and their low maintenance cost. Random Forest combines multiple trees trained on random subsamples, averaging their predictions [11,12]:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad (4)$$

LightGBM is a gradient boosting algorithm that builds trees sequentially by minimizing the residual error and is efficient on large data volumes [13]:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \quad (5)$$

Recurrent networks are a secondary benchmark. LSTM networks are a class of recurrent neural networks that capture long-term dependencies with a cell state. This cell state incorporates a forget, input, and output gate, which helps diminish the vanishing-gradient problem [14,15]. Due to their architecture, LSTMs are very good with series that are long and contain non-linear patterns [3,16]. However, they require a lot of data to really surpass simpler models. With the small amount of historical data available on a weekly basis, we include the LSTM model in this work as a benchmark only and consider it to be more peripheral to the analysis; as with other forecasts, the LSTM model is built by following standard procedures, which include Min–Max scaling and a supervised transformation of the data [11].

Studies containing comparisons suggest that models like SARIMA are good competitors for time series with clear seasonality [17]. On the other hand, ensemble methods indicate that in situations with seasonality and high variability, they are less likely to overfit [11,12] and are more computationally efficient for larger datasets [13,16].

2.2.3. Evaluation Metrics

To compare the predictive performance of the six developed models, four complementary metrics are used that allow accuracy to be evaluated from different statistical and practical perspectives [7,8].

Mean Absolute Error (MAE). MAE measures the average magnitude of forecast errors without considering their direction, and is robust against outliers [8]:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{6}$$

where y_t is the observed value, $\hat{y}_t$ is the forecast value, and n is the number of observations. It represents the average error in the same units as the original variable, which facilitates communication with non-technical stakeholders [1].

Root Mean Squared Error (RMSE). RMSE penalizes errors of greater magnitude by squaring the differences before averaging and is sensitive to extreme values [8]:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{7}$$

This metric is particularly useful when large errors have disproportionate operational consequences (e.g., stockouts in critical products) [3]; its interpretation is less intuitive due to the quadratic transformation.

Mean Absolute Percentage Error (MAPE). MAPE expresses the error as a percentage of the actual value, allowing accuracy to be compared across series with different scales [7]:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{8}$$

Although widely used in business practice, MAPE has limitations when y_t approaches zero and tends to overestimate errors in low-demand series [8]. For this reason, it is reported together with MAE and RMSE for a balanced evaluation.

Diebold–Mariano Test. To determine whether the observed differences between models are statistically significant and not the product of chance, the Diebold–Mariano test is applied [4]. Given two forecast-error series $e_{1,t}$ and $e_{2,t}$, the differential loss function is defined as

$$d_t = L(e_{1,t}) - L(e_{2,t}) \tag{9}$$

where $L(\cdot)$ can be the squared or absolute error. The test statistic is

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}\left(\bar{d}\right)/n}} \xrightarrow{d} N(0,1) \tag{10}$$

where d is the sample mean of d_t and $V(d)$ is a consistent estimator of the long-run variance that corrects for autocorrelation [4]. Under the null hypothesis H_0: $\mathrm{E}[d_t] = 0$ there is no significant difference in predictive accuracy between the compared models.

A hierarchical approach to model selection is applied to select the optimal model in each ABC class.

The squared-error loss is used throughout. The variance of the loss differential is estimated in the way that suits each evaluation scheme, and the manuscript states which applies where. For the single holdout window the twelve errors are multi-step forecasts issued from one origin and are autocorrelated by construction, so the long-run variance is estimated with a Newey–West correction using a Bartlett kernel. For the rolling-origin scheme of Section 4.9, the errors are one-step-ahead forecasts from successive origins, for which no autocorrelation correction is required and the sample variance is used together with the Harvey–Leybourne–Newbold small-sample correction. Section 4.6 applies both estimators to the same errors so that the effect of the choice can be judged rather than

assumed. Testing predictive accuracy rather than reporting point differences alone is standard practice in the forecasting literature [3,12]:

(1) Absolute accuracy. The model with the lowest RMSE is preferred. RMSE is used throughout the study for ranking and allocation because it penalizes the large errors that drive safety-stock requirements; MAE is reported alongside it for reference. Where the two disagree, as they do for Class C under rolling origins, the manuscript reports both.
(2) Robustness. If models are equivalent in RMSE, the model with the lowest MAE is preferred, since at that point the criterion is average accuracy rather than exposure to large errors.
(3) Statistical significance. The Diebold–Mariano test at the $\alpha = 0.05$ level is applied to every pairwise comparison and reported, but significance is not a precondition for allocation. With twelve out-of-sample errors the test detects only large differences, and as Sections 4.6 and 4.9 show, no class allocation in this study rests on a significant difference. Where models are statistically indistinguishable the criterion is reported as such and the allocation is resolved by the remaining criteria, which is what the framework does in practice.
(4) Interpretability. If models are equivalent with respect to the above, preference is given to ARIMA/SARIMA when the series is low in variability, due to their relative transparency and ease of operational implementation [17].

This broad-based selection approach provides a comprehensive balance across selection criteria of statistical rigor, operational readiness, and adaptability to low-resourcing settings [31].

3. Modeling Pipeline Architecture

The structure of the workflow (pipeline) designed in the Python ecosystem for building forecasting models is shown in Figure 2. This diagram articulates the entire data science cycle for the inventory problem. Each stage names the open-source library used to implement it so that the pipeline can be reproduced from the diagram alone. The pipeline was implemented in Python 3.13.9 (Python Software Foundation, Wilmington, DE, USA) on Windows 11 (Microsoft Corporation, Redmond, WA, USA), using the open-source libraries pandas 2.2.3, NumPy 2.1.3, SciPy 1.15.3, statsmodels 0.14.4, scikit-learn 1.6.1, LightGBM 4.6.0 (Microsoft Corporation, Redmond, WA, USA), TensorFlow 2.21.0 with Keras (Google LLC, Mountain View, CA, USA) and Matplotlib 3.10.0. The ARIMA-family models were estimated with statsmodels; pmdarima was not used. A single random seed (42) was fixed for Python's random module, NumPy and TensorFlow.

This modular architecture allows each development phase to be isolated, facilitating model auditing and its potential scalability for deployment in real production environments [11,31].

3.1. Data Preprocessing and Validation

The dataset is imported using Python's pandas library, and a quality audit is conducted. A quality audit includes the detection and removal of duplicate records and the identification and imputation of missing values. Missing values are imputed using the median of the same SKU. Dates are converted to the DateTime data type. Also, sales data is aggregated to form a time series of weekly data. Aggregating data on a weekly basis is done to reduce the noise of daily observations, and at the same time retain short-term dynamics. Weekly data is also the only frequency used in the study.

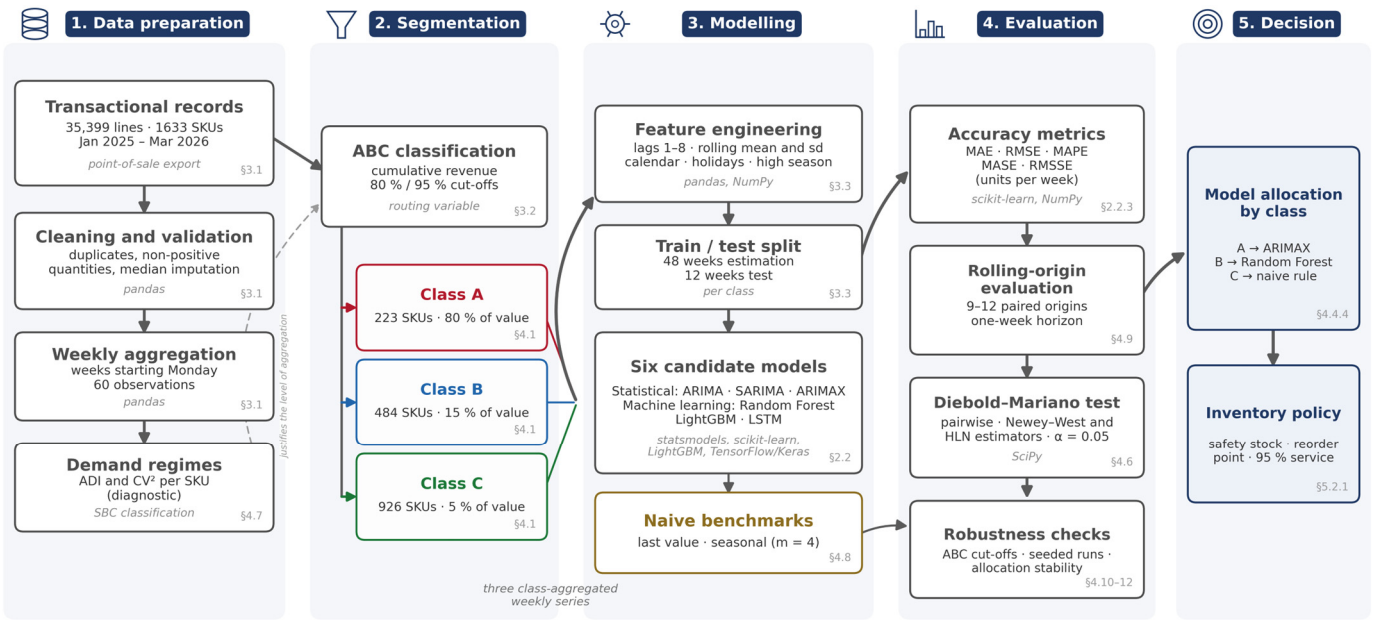


Figure 2. Methodological framework, from transactional records to class-conditional model allocation and inventory policy. Dashed arrow: the demand-regime diagnostic justifies the level of aggregation rather than feeding the pipeline. The section markers (§) indicate the part of the article in which each step is described.

The extract comprises 35,399 transaction lines recorded between 2 January 2025 and 1 March 2026, a span of 424 calendar days, covering 1633 distinct SKUs, 205,482 units sold and COP 690,711,593 in revenue. Transaction density is highly uneven across the portfolio: an SKU records 21.7 transactions on average and seven at the median, the most rotating item records 851, and 199 SKUs appear in a single transaction over the whole period. Missing unit prices were imputed with the median price of the same SKU, an approach preferred to mean imputation because unit prices in this portfolio are right-skewed.

No smoothing, winsorization or outlier removal was applied to the demand series themselves. Exceptional demand weeks are operationally real events that the replenishment system must absorb, and removing them would flatter the accuracy metrics without improving the decisions the forecasts support. Aggregation to the weekly grid, with weeks beginning on Monday, yields 60 observations per class-aggregated series, of which 48 are used for estimation and 12 for testing; none of the three aggregate series contains a zero-demand week.

3.2. ABC Classification

ABC segmentation groups SKUs based on their cumulative annual sales by calculating the annual sales value per SKU as the unit sale value multiplied by the total units sold, ordering the SKUs from highest to lowest, and establishing the cut-off to group the first class A SKUs that cumulatively represent 80% of the total sales value, class B the SKUs that cumulatively represent 95% of the total sales value, and class C the remaining SKUs. This segmentation is the most cited in the inventory-management literature to describe differentiated control and forecasting policies [1,30].

Given the time available, approximately 60 weekly observations, demand series aggregated by ABC class, were used to build forecasting models instead of using individual SKUs. This means that for classes A, B, and C, the weekly demand of all SKUs in the class is summed. The three aggregate series were input into the six models. This makes the

work a preliminary validation at the aggregate segment level, as SKU-level forecasts would require a longer time series. In addition, the segment-level forecasting would provide different forecasting strategies per class, in terms of the predictability of the class aggregate demand [30]. The segmentation of data defined in the results (see Figure 3) is a product of applying the process to the data of FERRETODO.

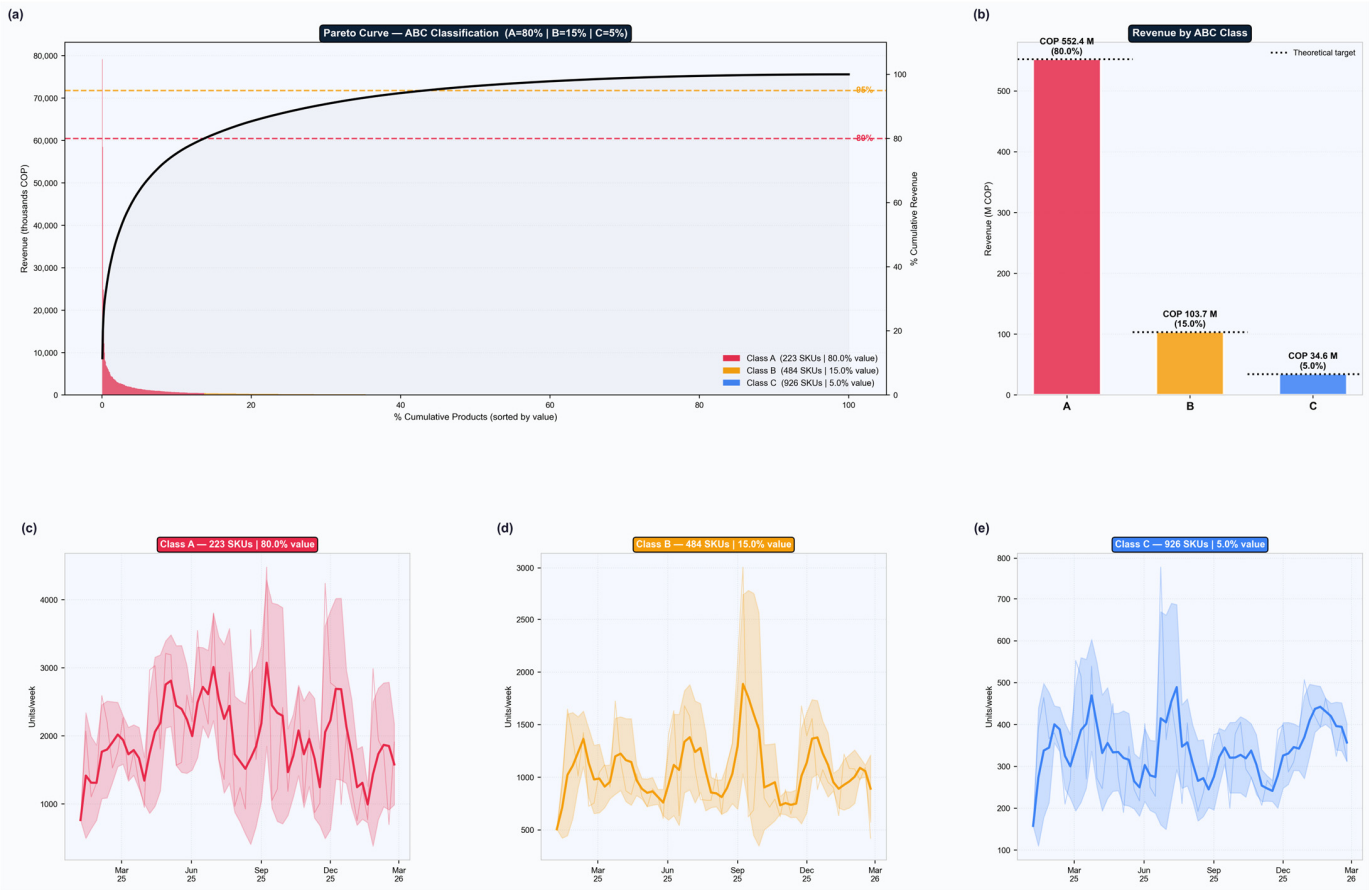


Figure 3. ABC classification of FERRETODO's portfolio. (**a**) Pareto curve of cumulative revenue with the 80% and 95% thresholds delimiting classes A, B, and C; (**b**) total revenue by ABC class (millions of COP) with the theoretical 80/15/5 target overlaid; (**c**–**e**) weekly aggregate demand series (units/week) for classes A, B, and C over January 2025–March 2026. The three series are the inputs on which the forecasting models are trained. In (**a**), the bars give the revenue of each SKU, colored by class (the class C bars are too small to be visible at this scale), and the black line is the cumulative share of revenue (right axis). In (**b**), the dotted lines mark the theoretical target. In (**c**–**e**), the thick line is the four-week moving average, the thin line the weekly series, and the shaded band the moving average ± one four-week rolling standard deviation.

3.3. Statistical Analysis and Modeling

An ARIMA(1, *d*, 1) specification was adopted for all classes, with the differencing order *d* determined by the Augmented Dickey–Fuller test; parameters were estimated by maximum likelihood and residuals validated with the Ljung–Box test. The seasonal specification is (1, 0, 1) with a period of four weeks, activated only when an STL decomposition with that period yields a seasonal component exceeding 10% of the standard deviation of the series. A feature matrix is prepared for the machine-learning models by lagging the series at one, three, four, six, seven and eight weeks, adding an eight-period moving average and a four-period rolling standard deviation, a linear trend, week-of-year and month indicators, and the two exogenous regressors described in Section 3.4.

Hyperparameters were fixed a priori rather than searched and are reported here in full so that the results can be reproduced: Random Forest with 300 trees, a maximum depth of 8 and minimum leaf size of 2; LightGBM with 400 estimators, a learning rate of 0.05 and 20 leaves; and an LSTM with a single recurrent layer of 48 units, a dropout of 0.15, 40 training epochs, a batch size of 4 and a look-back window of eight weeks. All three use a fixed random seed. A randomized search over a large hyperparameter space was deliberately avoided: with 48 training observations such a search selects on noise, and the resulting configuration would not generalize. This is a constraint imposed by the data available to the firms the framework targets, and reporting it explicitly is preferable to presenting a tuning procedure whose validity the sample cannot support.

3.4. Exogenous Variables of the ARIMAX Model

A defining feature of the ARIMAX model is the exogenous regressor vector X_t formulated in Equation (3). In contrast to the broader feature matrix used by the machine-learning models (which adds lagged observations and moving averages), the ARIMAX exogenous block was intentionally constrained to two parsimonious calendar variables distilled from the official Colombian public holiday calendar for the years 2025–2026. These variables are designed to capture the demand effects of national holidays and the peak demand periods of the hardware retail sector. These demand effects are summarized in Table 1.

Table 1. Exogenous variables X_t supplied to the ARIMAX model.

Variable	Type/Range	Definition
holidays (h_t)	Integer {0, 1, 2}	No. of national holidays in week $[d, d + 6]$
high_season (s_t)	Binary {0, 1}	1 if month $\in$ {March, April, July, August, October, November, December}

One point of notation requires clarification. The specification referred to throughout as ARIMAX carries, in addition to the exogenous block of Equation (3), is the seasonal component described in Section 3.3, which is activated when the STL criterion detects seasonality; in that case the estimated model is formally a SARIMAX(1, d, 1)(1, 0, 1)4 with two exogenous regressors. The label ARIMAX is retained for readability and for comparability with the SARIMA specification that differs from it only by the exogenous block, but Equation (3) should be read together with Equation (2) rather than as the complete specification:

(1) Holidays in the week (h_t, holidays): an integer count of official Colombian public holidays falling within the seven-day window of week t. Formally, for a week starting on date d, $h_t = \#\{f \in F : d \leq f \leq d + 6\}$, where F is the set of national holidays observed in 2025–2026 (including the Mondays to which several of them are moved under the Emiliani Law). This variable takes values in 0, 1, 2 and captures the reduction or concentration of commercial activity caused by holidays within a given week.

(2) High season (s_t, high_season): a binary indicator equal to 1 when the month of week t belongs to the set {March, April, July, August, October, November, December} and 0 otherwise. These months concentrate construction project closings and seasonal peaks of the hardware sector, so s_t acts as a structural-seasonality flag complementary to the autoregressive component.

This design is deliberate. Because the time series is limited to approximately 60 weekly measurements, restricting X_t to two interpretable calendar regressors is taken to mean that including additional regressors would lead to overparameterization. Here, regressors are added in places that provide the most information (high-volatility class A). As shown in

the Results, in a low-magnitude class-C series, the same exogenous block is noise rather than signal, demonstrating that external regressors have value that is conditional.

3.5. Comparison and Implementation

The six models are assessed on the 20% test data using MAE, RMSE, and MAPE. Applying the Diebold–Mariano test at the 5% significance level makes it possible to determine if the differences in the models are statistically significant, as this is the established standard in assessing the predictive accuracy [4]. The most appropriate model can be chosen for each ABC class with the help of this intensive comparison [3].

4. Results

4.1. ABC Classification of the Product Portfolio

Using ABC analysis on Ferretodo's 1633 SKUs shows that most of their revenue is generated by a small number of products. Figure 3 shows that in the hardware retail example, the 223 class-A products (13.7% of the total SKUs) consume 80% of total sales. Based on total revenue generated by the SKU classes, class-A, -B, and -C products total approximately COP 690.7M, of which COP 552.4M (80%), COP 103.7M (15%), and COP 34.6M (5%), respectively, are assigned to class-A, -B, and -C products. The sales revenue per class is so similar that it appears to validate the 80/15/5 rule associated with the class cut-off points.

Class-A products have demand variability (coefficient of variation of the class-aggregated weekly series equal to 0.471, the highest of the three classes, as reported in Table 2) that makes them the hardest and most economically important segment to predict, thus justifying the application of complex modeling for that segment. This finding contradicts the most common assumption with the application of the ABC analysis, where a high value always means high predictability. Class B (484 SKUs, 29.6% of the portfolio) contains an intermediate variability segment, while class C (926 SKUs, 56.7% of the portfolio) contains mainly low-turnover or sporadic sold products, with an individual low economic impact, but collectively important for the service level. The lower panels of Figure 3c–e show this hierarchy at an aggregate level: the class-A series fluctuates between approximately 750 and 4500 units/week, with highest peaks in September and December 2025. Class B shows an intermediate fluctuation with a peak in September and class C the lowest fluctuation, indicating that the three aggregate series have different volatility, thus justifying a differentiated per-class modeling approach, as opposed to a single SKU-level modeling approach.

Table 2. Descriptive statistics of the class-aggregated weekly demand series used for model estimation (N = 60 weeks, January 2025–March 2026).

Class	N	Mean	Median	Std. Dev.	CV	Min	Max	Skew	Kurt.	ADF *p*	Station.	Shapiro *p*	Normal
A	60	1971.16	1792.39	928.16	0.471	698.76	4495.29	0.70	−0.16	0.0000	Yes	0.0066	No
B	60	1059.82	934.46	439.35	0.415	420.26	3007.31	1.78	4.90	0.0000	Yes	0.0000	No
C	60	342.04	345.50	110.74	0.324	158.00	778.60	1.06	2.40	0.0000	Yes	0.0022	No

4.2. Statistical Characterization of Demand by Segment

Table 2 reports the descriptive statistics of the three class-aggregated weekly series, which are the objects on which the forecasting models are estimated. Mean weekly demand decreases monotonically across classes (1971.2, 1059.8 and 342.0 units per week for classes A, B and C), confirming the economic hierarchy established by the ABC analysis. Variability follows the same ordering: the coefficient of variation is highest in class A (0.471) and lowest in class C (0.324). This is a first piece of evidence against the assumption that

economic importance implies predictability, since the segment that concentrates 80% of revenue is also the most volatile.

The three series are stationary at the weekly frequency (Augmented Dickey–Fuller $p < 0.01$ in all cases), so no differencing is required and the integrated component of the ARIMA family reduces to $d = 0$. All three series depart from normality (Shapiro–Wilk $p = 0.0066$, 0.000 and 0.0022), with positive skewness (0.70, 1.78 and 1.06) and, in class B, heavy tails (excess kurtosis 4.90). The absence of Gaussian behavior in all three classes is consistent with the strong performance of non-parametric ensembles in classes B and C, which impose no distributional assumption; it does not preclude the statistical family from performing well in class A, where the calendar regressors carry genuine signal.

The heterogeneity that motivates the framework is more visible at the SKU level. Classifying the 1633 SKUs by demand regime following Syntetos, Boylan and Croston [18], using the average inter-demand interval (ADI) and the squared coefficient of variation of non-zero demand sizes (CV^2), shows sharply different compositions across classes. Class A is predominantly lumpy (41.7% of its SKUs, against 39.9% intermittent), with a median ADI of 2.35 and a median CV^2 of 0.53. Class C is overwhelmingly intermittent (83.6%), with a median ADI of 15.25 and a median CV^2 of 0.09: its items sell rarely, but in comparable quantities when they do. Class B is intermediate (71.3% intermittent, 22.3% lumpy). The full distribution is reported in Section 4.7.

Two implications follow. First, the least predictable regime is concentrated in the economically most important class, which provides direct empirical support for the claim that high value does not imply high predictability. Second, 95.5% of the portfolio is intermittent or lumpy at the SKU level, and 539 SKUs record demand in three weeks or fewer out of sixty. Forecasting such series individually with ARIMA-family or recurrent architectures would be methodologically inappropriate, since these regimes call for Croston-type estimators [18,20]. Aggregating demand by ABC class removes intermittency entirely, as none of the three aggregate series contain a zero-demand week, which makes class-level aggregation a methodological requirement imposed by the observed demand regimes rather than a convenience adopted because of the length of the history.

4.3. Model Diagnostics and Analysis of Predictive Variables

Figure 4a–c synthesize the architecture, configuration and internal validation of the six developed models. The panel has been divided into three figures so that the axis and category labels remain legible at print size. This diagnosis is fundamental to understanding the modeling decisions that preceded the accuracy evaluation and to justify the differentiated assignment by ABC class.

Training, complexity, and feature engineering. The temporal partition kept 48 training weeks and 12 test weeks in the three classes, representing a quarterly planning cycle. In configuration complexity, ARIMAX required the largest number of hyperparameters (nine), followed by SARIMA (seven), whereas Random Forest and LightGBM required only four, suggesting that ensembles achieve high accuracy with lower tuning overhead [11,31]. The feature composition progresses from the pure historical series (ARIMA, SARIMA) to lags and exogenous regressors (ARIMAX) and calendar indicators in the machine-learning models. The AIC criterion favors ARIMA in the three classes (parsimony principle), placing ARIMAX in the third position and anticipating that its advantage only materializes when the exogenous regressors provide a genuine signal (Class A) and not noise (Classes B and C) [8,9].



Figure 4. (**a**) Model inventory and configuration: distribution of the six models by algorithmic family, training and test partition by class, and number of configured hyperparameters per model. (**b**) Feature composition and statistical diagnostics: number and type of predictors entering each model, Akaike information criterion for the statistical specifications, and Ljung–Box *p*-values for residual autocorrelation. (**c**) Predictive variables by inventory class: ten highest Gini importances from a Random Forest estimated on the training window of each class with the hyperparameters of the deployed model.

Regarding residual validation and feature importance, all models passed the Ljung–Box test for the null hypothesis of no residual autocorrelation ($p > 0.05$), with SARIMA for class B closest to the threshold ($p = 0.07$). Feature importances were obtained from a single Random Forest model per class, estimated on the training window with the same hyperparameters as the deployed model. In Class A the dominant predictors are short-term lags (lag3 = 0.246, lag4 = 0.189, lag6 = 0.106), with the eight-period moving average contributing 0.082: demand in the high-value segment is driven by recent history rather than by calendar structure. In Class B the leading predictors are lag4 (0.180), the eight-period moving average (0.169) and the immediate lag (0.142), followed by a Septem-

ber indicator (0.110) that captures the construction-season peak. In Class C the ranking changes: the eight-period moving average leads (0.183), followed by the trend (0.122) and lag7 (0.118), which is consistent with a low-turnover segment whose short-term signal is weak and whose predictability derives from its smoothed level rather than from individual recent observations.

Aggregating across the three classes by feature category, temporal lags contribute 56.8% of total importance and rolling statistics a further 20.4%, so that 77.2% of the predictive information is autoregressive in nature. Monthly calendar indicators contribute 8.7%, the trend 6.5%, the weekly calendar 5.9% and the exogenous holiday and high-season regressors only 1.7%. The dominance of the autoregressive structure over calendar and exogenous effects explains why the advantage of ARIMAX is confined to class A, where its two calendar regressors add signal to an otherwise highly volatile series, and why the same regressors degrade performance in class C.

Feature importances describe the engineered matrix shared by Random Forest, LightGBM and the recurrent network, and they are reported as a diagnostic of that matrix rather than as an explanation of the model selected in each class. Where the selected model is ARIMAX, as in class A, the appropriate explanation is its own estimated parameters, which are reported in Table 3. The high-season indicator carries a coefficient of +566.5 units per week and is significant at the 5% level ($p = 0.025$), while the holiday count is not significant ($p = 0.392$). This is the sole exogenous effect that is statistically distinguishable from zero anywhere in the portfolio: in class B neither regressor is significant, and in class C the high-season indicator is only marginal ($p = 0.066$). The advantage of ARIMAX in class A is therefore attributable to one identifiable and interpretable seasonal effect, not to the exogenous block as a whole, and the revised text says so.

Table 3. Estimated ARIMAX parameters by class: coefficient (standard error) and p-value. Demand in units per week.

Term	Class A	Class B	Class C
Holidays in week	−190.10 (222.13), $p = 0.392$	−63.80 (272.66), $p = 0.815$	−52.29 (45.83), $p = 0.254$
High season	566.49 (252.34), $p = 0.025$	−180.92 (278.40), $p = 0.516$	107.96 (58.63), $p = 0.066$
AR(1)	0.9998 (0.0009), $p < 0.001$	1.0000 (0.0002), $p < 0.001$	0.9947 (0.0120), $p < 0.001$
MA(1)	−0.9433 (0.0759), $p < 0.001$	−0.9448 (0.1357), $p < 0.001$	−0.7909 (0.1584), $p < 0.001$
Seasonal AR(1)$_4$	−0.0271 (0.4827), $p = 0.955$	−1.0000 (1.3652), $p = 0.464$	−0.8939 (0.1286), $p < 0.001$
Seasonal MA(1)$_4$	−0.2570 (0.4659), $p = 0.581$	1.0000 (1.7149), $p = 0.560$	0.9999 (0.2078), $p < 0.001$
AIC	820.3	749.7	625.1

The estimates also record a diagnostic limitation. The autoregressive coefficient lies at or very close to the unit-root boundary in all three classes (0.9998, 1.0000 and 0.9947), even though the Augmented Dickey–Fuller test did not reject stationarity and no differencing was applied. The specification is therefore absorbing a near-stochastic trend through the autoregressive term rather than through the integration order, and in Class B the seasonal parameters take boundary values with standard errors larger than the estimates themselves, indicating that the seasonal block is not identified there. Neither affects the accuracy comparison, which is conducted out of sample, but both are reported so that the specification is not read as fully diagnosed.

4.4. Model Performance by Product Class

4.4.1. Class-A Products (High Value)

For the 223 class-A products, the ARIMAX model was selected as the most suitable. According to the diagnosis presented in Figure 5, the model achieved an RMSE of 843.1,

an MAE of 661.7, and an MAPE of 63.9%. Although the MAPE may seem high, class-A products present high volatility, with demand values oscillating between roughly 700 and 4500 units per week.

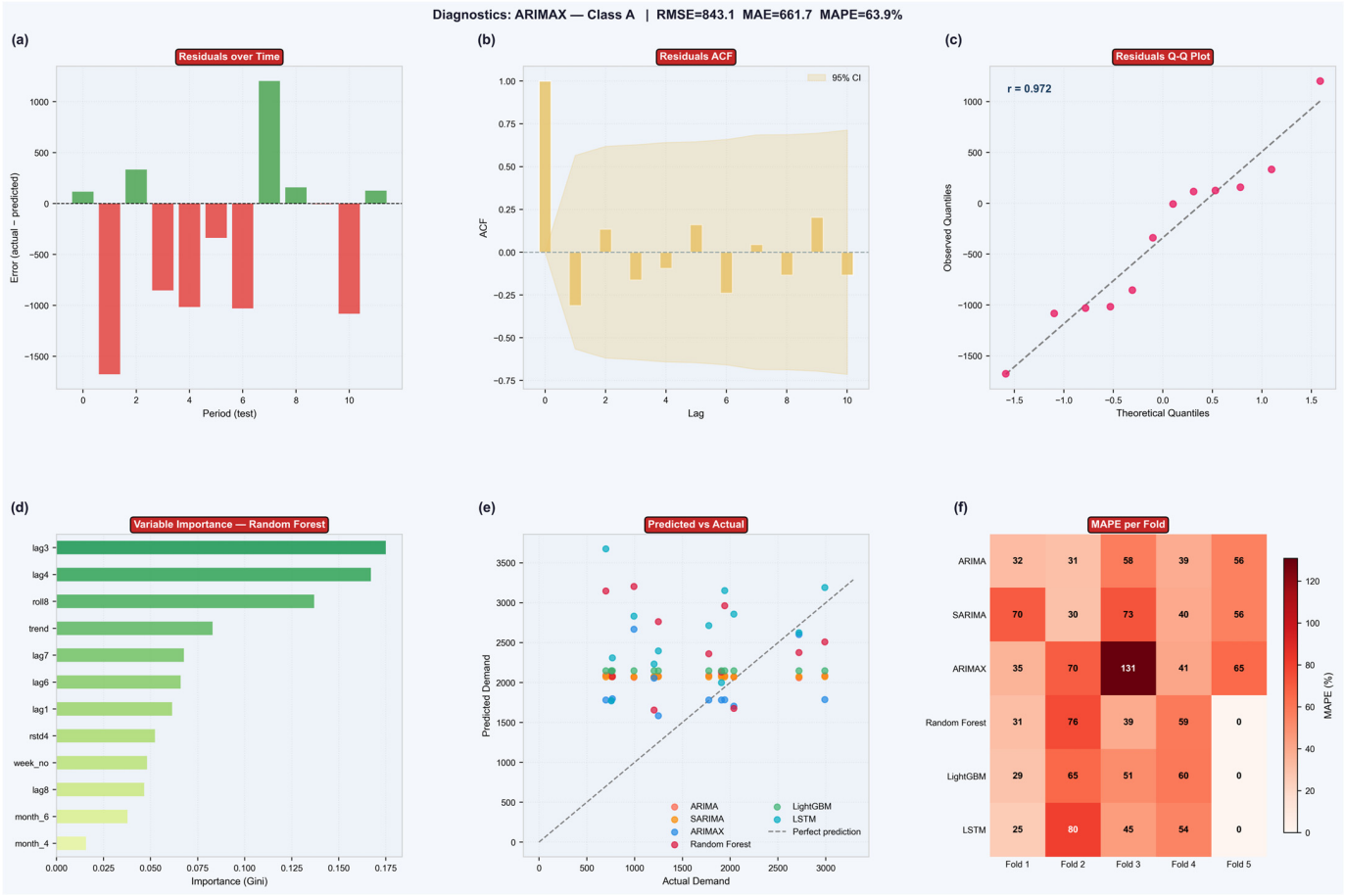


Figure 5. Diagnosis of the ARIMAX model for class-A products: (**a**) residuals over time, (**b**) ACF of residuals, (**c**) Q-Q plot of residuals, (**d**) feature importance, (**e**) predicted vs. actual values, (**f**) MAPE per fold in cross-validation. In (**a**), green and red bars are positive and negative errors; in (**b**), the shaded band is the 95% confidence interval; in (**c**,**e**), the dashed line is the 45-degree reference (normality and perfect prediction, respectively); in (**e**), colors identify the six models; in (**f**), color intensity is proportional to MAPE.

An MAPE of 63.90% deserves comment, since it is high in absolute terms. Two considerations bear on whether forecasts at that accuracy support the replenishment decision. The metric is inflated by weeks of low demand, in which a small absolute error is a large percentage one; the scaled metrics are more informative here, and an MASE of 0.620 indicates that the forecast error is 38% below that of the in-sample naive benchmark. And what the replenishment rule requires is not a point forecast but a buffer calibrated to the error, which is what the safety-stock computation in Section 5.2.1 provides: an RMSE of 843 units per week on a mean of 1971 translates into a buffer of roughly 1391 units at a 95% service level. The forecast is therefore usable for sizing the buffer even where it is not accurate enough to be relied on week by week, and the manuscript does not claim more than that.

The residual analysis revealed acceptable patterns: the Q-Q plot showed a correlation of r = 0.972 with the theoretical normal distribution, and the residual ACF indicated the absence of significant autocorrelation at most lags. The gain over ARIMA is not statistically significant (Section 4.6) and reverses under rolling origins (Section 4.9); the feature-importance panel describes the machine-learning feature matrix, not ARIMAX.

4.4.2. Class-B Products (Medium Value)

For the 484 class-B products, the Random Forest algorithm outperformed traditional statistical models. As shown in Figure 6, the model achieved an RMSE of 254.7, MAE of 191.6, and MAPE of 27.6%, a substantial improvement compared to class-A products.

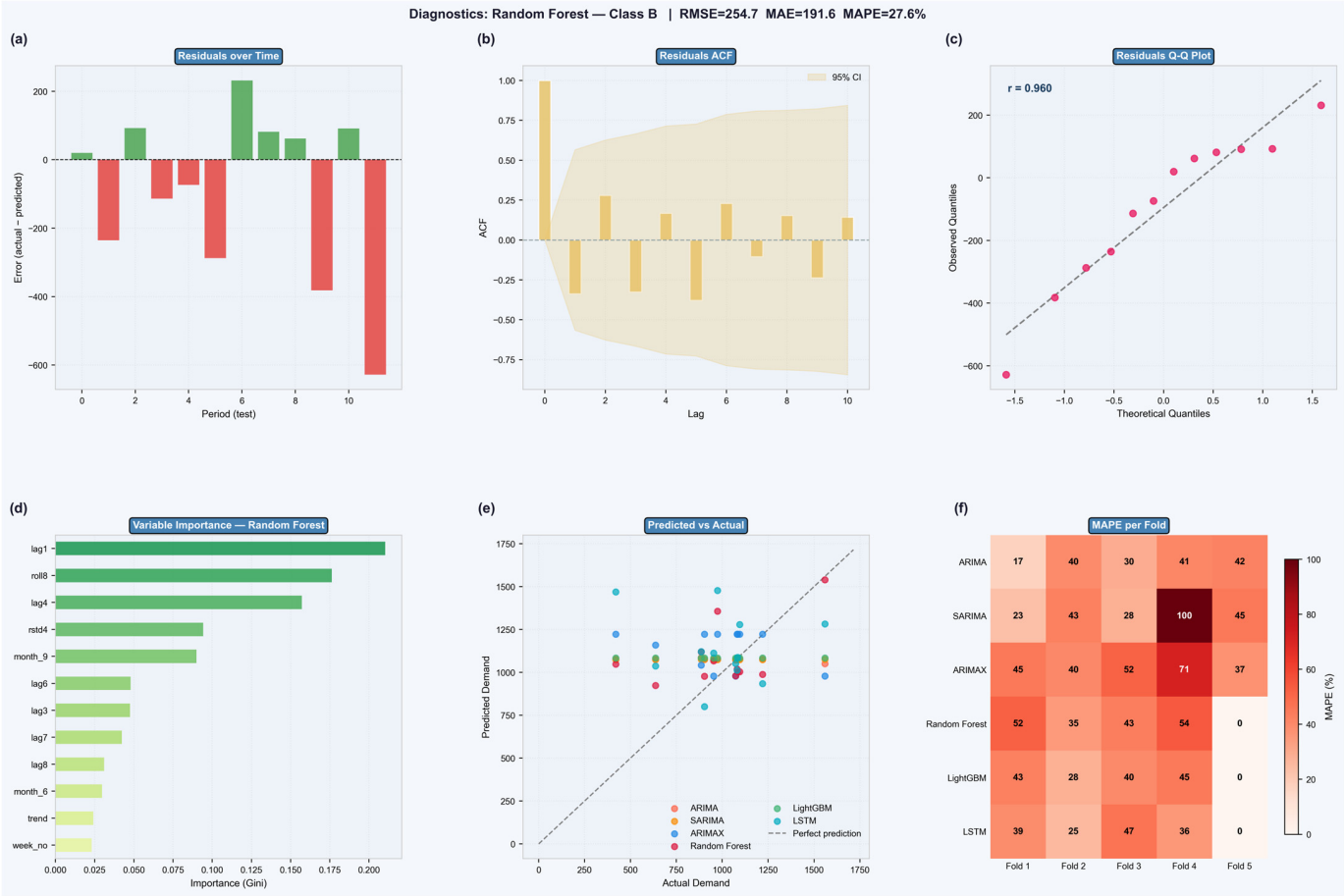


Figure 6. Diagnosis of the Random Forest model for class-B products: (**a**) residuals over time, (**b**) ACF of residuals, (**c**) Q-Q plot of residuals, (**d**) feature importance, (**e**) predicted vs. actual values, (**f**) MAPE per fold in cross-validation. Colors and lines as in Figure 5.

The Q-Q plot showed a correlation of r = 0.960, slightly lower than in class A, expected because Random Forest does not assume residual normality. The scatter in the predicted vs. actual plot showed that the model tends to underestimate extreme demand values (points above 1500), typical of tree-based ensembles that average predictions. Feature importance revealed lag4 (Gini = 0.180), the eight-period moving average (0.169) and the immediate lag (0.143) as the leading predictors, indicating that recent demand is the best predictor of future demand for class-B products. The 5-fold cross-validation showed an MAPE between 27% and 54%, with an atypical fold (Fold 4) presenting an MAPE of 100% for SARIMA, confirming the superiority of Random Forest for this segment.

4.4.3. Class-C Products (Low Value)

For the 926 class-C products, Random Forest attains the lowest error of the six models, although it does not improve on the naive rule (Section 4.8): RMSE of 74.6, MAE of 64.2, and MAPE of 15.8% (Figure 7). This significantly lower MAPE reflects the lower absolute volatility of these products.

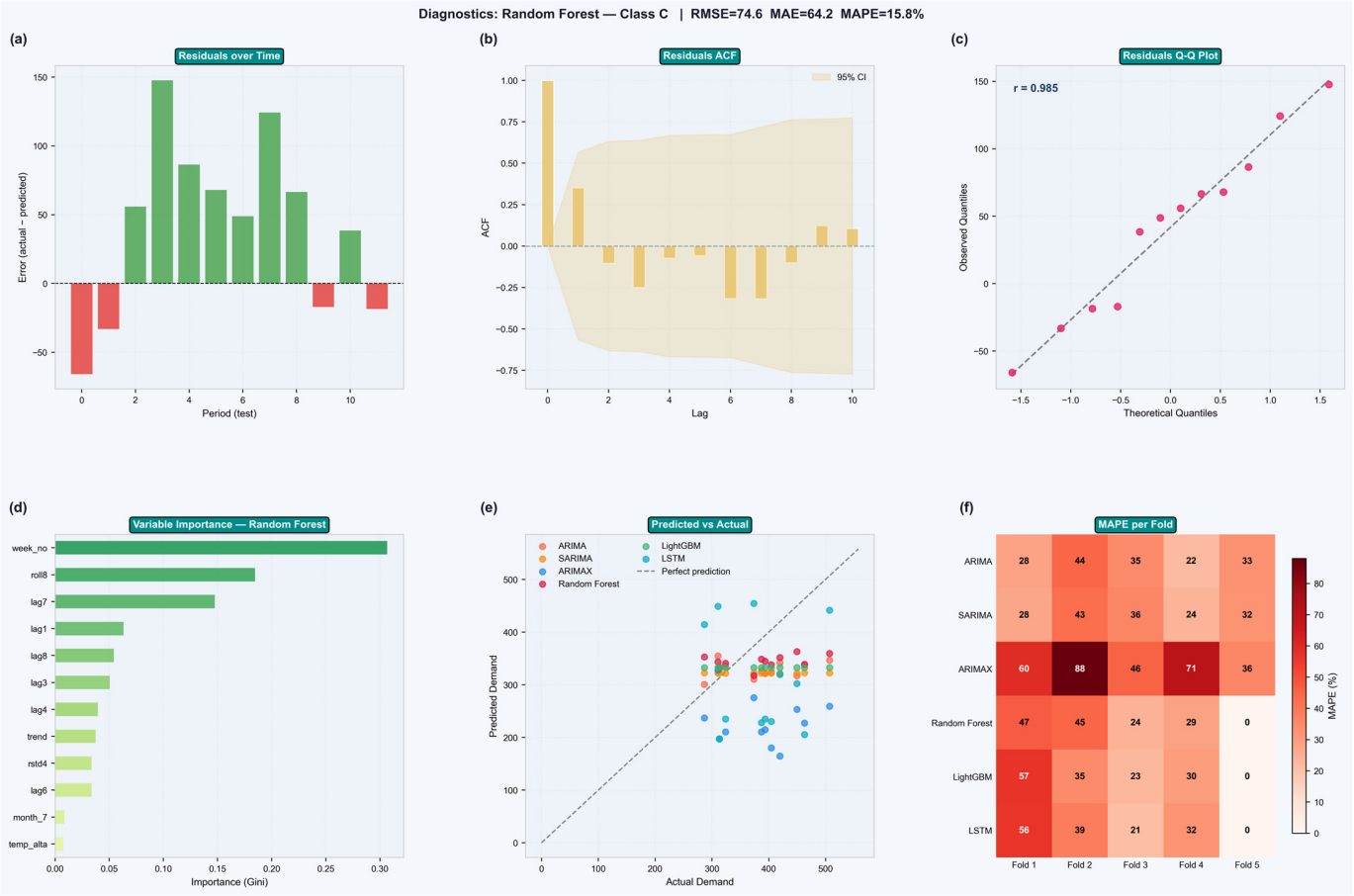


Figure 7. Diagnosis of the Random Forest model for class-C products: (**a**) residuals over time, (**b**) ACF of residuals, (**c**) Q-Q plot of residuals, (**d**) feature importance, (**e**) predicted vs. actual values, (**f**) MAPE per fold in cross-validation. Colors and lines as in Figure 5.

The Q-Q plot showed the best correlation with normality (r = 0.985) among the three classes, and the residuals remained in a reduced range (−50 to +150). Feature importance in this class is led by the eight-period moving average (roll8 = 0.183), followed by the trend (0.122) and the seven-week lag (lag7 = 0.118), with the week-of-year indicator contributing 0.108. The predictability of this segment therefore derives from its smoothed level rather than from individual recent observations, which is consistent with a low-turnover portfolio in which the short-term signal is weak.

4.4.4. Model Comparison and Statistical Validation

The comparison of the six models across classes does not yield a single dominant family. In class A the lowest errors are obtained by a statistical specification, ARIMAX (RMSE 843.2, MAPE 63.90%), which outperforms the best ensemble in that class (LightGBM, RMSE 924.7) and, by a wider margin, Random Forest (1255.8); the two calendar regressors supply information that the purely autoregressive alternatives do not exploit. In classes B and C the ordering reverses and Random Forest attains the lowest errors (RMSE 254.7 and 74.6), with LightGBM and the ARIMA family close behind. The LSTM is last in classes A and B and fifth in class C, where ARIMAX is last. The pattern is therefore conditional rather than uniform, and it is the opposite of the one that a single portfolio-level comparison would have reported.

Comprehensive Model Comparison by Error Metrics

Figure 8 presents a comparative view of the performance of the six evaluated models across the three product classes, using MAPE, RMSE, and MAE [7,8].

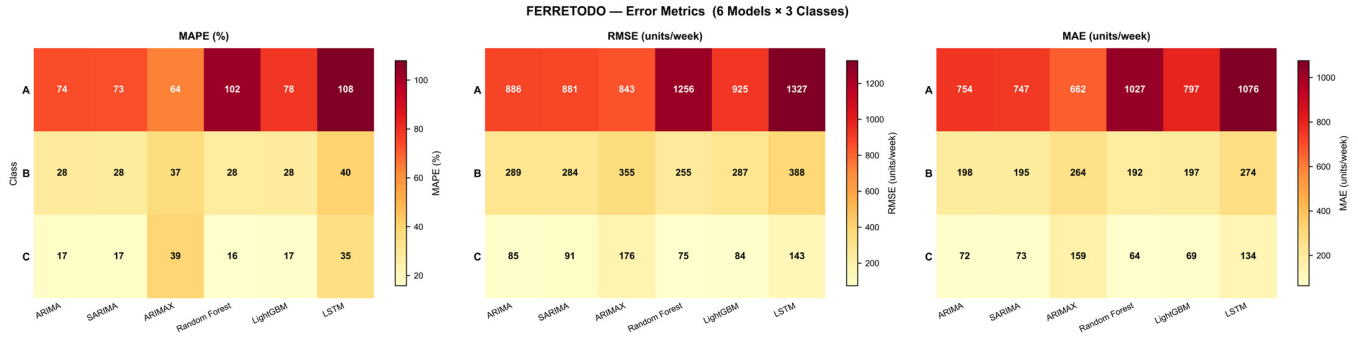


Figure 8. Error metrics for the six models across the three ABC classes. MAPE in percent; RMSE and MAE in units per week. Each panel uses a single color scale, so the comparison the panels support directly is between classes; the ranking of models within each class is reported in the per-class rankings of Section 4.4.

Table 4 consolidates the three error metrics for all six models across the three ABC classes, with the best value per class and metric highlighted in bold. This table is the central quantitative reference of the study: it makes the conditional—not universal—superiority of each model family directly readable and supports the model-selection decisions discussed below.

Table 4. Global performance summary: error metrics (RMSE, MAE, MAPE) for the six models across the three ABC classes. Best value per class in selected model marked with *.

Model	Type	A RMSE	A MAE	A MAPE	B RMSE	B MAE	B MAPE	C RMSE	C MAE	C MAPE
ARIMAX * A	Stat.	843.15	661.71	63.90	354.64	263.70	36.54	176.24	159.06	39.32
SARIMA	Stat.	881.47	746.73	73.07	283.66	195.43	27.88	91.18	73.21	17.26
ARIMA	Stat.	886.29	753.69	73.66	288.60	198.32	28.15	85.13	71.85	17.22
Random Forest * B	ML	1255.75	1027.48	102.45	254.65	191.60	27.58	74.60	64.20	15.83
LightGBM	ML	924.75	796.58	78.10	286.58	197.23	28.28	84.20	69.37	16.62
LSTM	ML	1326.60	1075.90	107.94	387.50	274.01	40.44	143.21	134.40	35.38

RMSE and MAE in units per week; MAPE in %. * A = selected for class A; * B = selected for class B; class C is served by a last-value naive rule (Section 4.8).

Table 5 condenses the model selected for each class, facilitating the comparative reading of the study's main findings. The model listed is the best of the six evaluated within each class; in Class C it does not improve on the last-value naive benchmark, and Section 4.8 shows that the class warrants no model at all.

Table 5. Consolidated summary of the optimal model by ABC class.

Class	Best Model	MAE	RMSE	MAPE (%)
A (80.0% of value)	ARIMAX	661.71	843.15	63.90
B (15.0% of value)	Random Forest	191.60	254.65	27.58
C (5.0% of value)	Random Forest (does not beat the naive rule)	64.20	74.60	15.83

For Class A, ARIMAX emerged as the best-performing model (MAPE 64%, RMSE 843, MAE 662 units), outperforming even the ensemble methods (Table 6, ordered by RMSE). Although the MAPE seems high, it must be read against a weekly demand oscillating between roughly 700 and 4500 units with high inherent variability [17]; in absolute terms, an average error of 662 units is acceptable given that range. LSTM ranked last (RMSE

1326.60, 1.57 times the error of ARIMAX the best model); since LSTM was included solely as a methodological benchmark because of the limited history (~60 weekly observations), this confirms that such a data volume is insufficient for robust training of deep networks [11,16].

Table 6. Model ranking for class A (80.0% value) ordered by RMSE.

Model	Type	Rank	RMSE	MAE	MAPE (%)
ARIMAX	Statistical	1	843.15	661.71	63.90
SARIMA	Statistical	2	881.47	746.73	73.07
ARIMA	Statistical	3	886.29	753.69	73.66
LightGBM	ML	4	924.75	796.58	78.10
Random Forest	ML	5	1255.75	1027.48	102.45
LSTM	ML	6	1326.60	1075.90	107.94

For class B, close competition was observed: Random Forest led (RMSE 254.65, MAE 191.60, MAPE 27.58%), closely followed by SARIMA (283.66) and LightGBM (286.58), with ARIMA fourth (Table 7). This clustering suggests that, for moderate variability, both ensembles and seasonal statistical models are viable; Random Forest's advantage in MAE points to greater robustness against extreme errors [11,12]. ARIMAX and LSTM showed inferior performance, ruling out their usefulness in this segment [8,17].

Table 7. Model ranking for class B (15.0% value) ordered by RMSE.

Model	Type	Rank	RMSE	MAE	MAPE (%)
Random Forest	ML	1	254.65	191.60	27.58
SARIMA	Statistical	2	283.66	195.43	27.88
LightGBM	ML	3	286.58	197.23	28.28
ARIMA	Statistical	4	288.60	198.32	28.15
ARIMAX	Statistical	5	354.64	263.70	36.54
LSTM	ML	6	387.50	274.01	40.44

For class C, Random Forest attains the lowest error among the six models (RMSE 74.60, MAE 64.20, MAPE 15.83%), although Section 4.8 shows that it does not improve on a last-value naive rule and that the class therefore warrants no model, followed by LightGBM (RMSE 84.20); ARIMA and SARIMA remained acceptable (MAPE $\approx$ 17%), and ARIMAX collapsed to last place (RMSE 176.24, MAPE 39.32%) (Table 8). The failure of ARIMAX on low-magnitude data is consistent with the overfitting produced by external regressors when the signal is weak [12,31]. Although each class-C SKU has a low individual impact, collectively they represent 5% of the value; an MAPE of 16% would in principle allow a smaller safety stock, but Section 4.8 shows that the naive benchmark attains a lower error still, so the reduction is not attributable to the model.

A consistent finding was the poor performance of LSTM in the three classes (MAPE of 35.38%, 40.44% and 107.94% for classes C, B and A respectively; Table 4). This result—expected given that LSTM was included solely as a methodological benchmark—contrasts with studies reporting the superiority of deep neural networks [3] but is explained by: (1) the limited historical dataset; (2) the weekly nature of the data, which reduces the number of observations; and (3) the high volatility of hardware demand [15]. Recent research warns that LSTM requires large data volumes to outperform traditional methods [11,16]. In an SME context, the computational overhead and maintenance difficulty of LSTM are not justified relative to lighter models [31].

Table 8. Model ranking for class C (5.0% value) ordered by RMSE.

Model	Type	Rank	RMSE	MAE	MAPE (%)
Random Forest	ML	1	74.60	64.20	15.83
LightGBM	ML	2	84.20	69.37	16.62
ARIMA	Statistical	3	85.13	71.85	17.22
SARIMA	Statistical	4	91.18	73.21	17.26
LSTM	ML	5	143.21	134.40	35.38
ARIMAX	Statistical	6	176.24	159.06	39.32

The differentiation observed is consistent with the central hypothesis that model complexity should scale with the economic importance and variability of the item [17]. While statistical specifications are preferred in the high-volatility class, non-parametric ensembles stand out in the moderately stable one, and in the low-turnover class no model improves on a naive rule [12,13]. This complementarity suggests that an enterprise forecasting system should implement a model router guided by the ABC classification [30,31].

4.5. Visual Evaluation of Forecasts in the Time Domain

Figures 9–11 show, for each class, the weekly historical series, the forecast of the allocated model over the test window and a projection beyond the last observation. The shaded band marks the 12-week test window, which begins in December 2025 and ends with the last observed week of February 2026; within it the observed demand is plotted so that it can be compared with the forecast, which is the comparison this section reports. The dotted line marks the end of the observed data: to its right the models are projected a further 12 weeks, a region for which no realized demand exists [1,8].

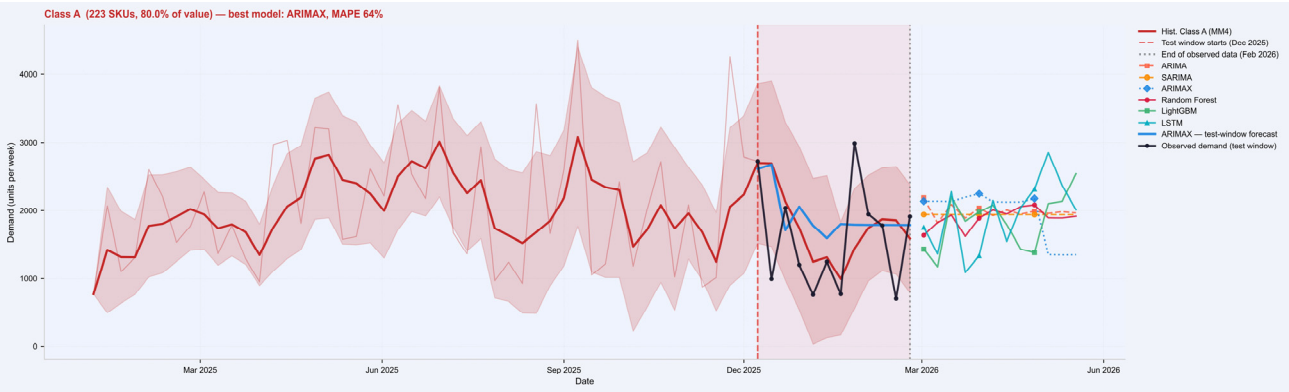


Figure 9. Class A: observed demand, test-window forecast and projection. The shaded band is the 12-week test window, over which the observed series (black) can be compared with the forecast of the allocated model (ARIMAX). To the right of the dotted line the series is projected 12 weeks beyond the last observation, where no observed demand exists for comparison. The thick colored line is the four-week moving average of class demand, the thin line the weekly series, and the shaded area around them the moving average ± one four-week rolling standard deviation; the dashed red line marks the start of the test window, and the colored lines and markers to the right of the dotted line are the projections of the six models.

For class A, the historical series indicate demand surges of over 4000 units in September and December 2025. One likely reason may be the peaks of certain phases in construction work or a construction season [6]. In the forecast period, ARIMAX yields a smooth curve fluctuating between 1500 and 2200 units. This kind of conservative function is strategically favorable. In the case of a highly valued item, demand overprediction creates, in effect, an opportunity cost and great surpluses of value; in the case of demand underprediction, the

opposite is true and extremely costly zero inventories result [1,2]. In this regard, ARIMAX balances the trade-off by accounting for exogenous variables and, in effect, noise, while retaining responsiveness to structural flows [10,17]. In this respect, much more the opposite is the case for LSTM and Random Forest, as they tend to have a significant overreaction to diminished demand, thus creating false purchasing triggers [16].

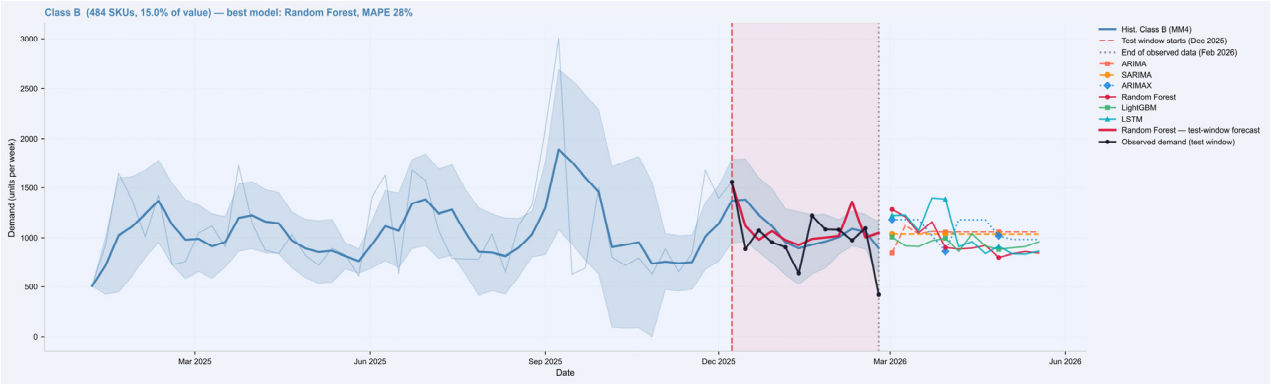


Figure 10. Class B: observed demand, test-window forecast and projection, with the same convention as Figure 9. The allocated model is Random Forest. Shaded areas and lines as in Figure 9.

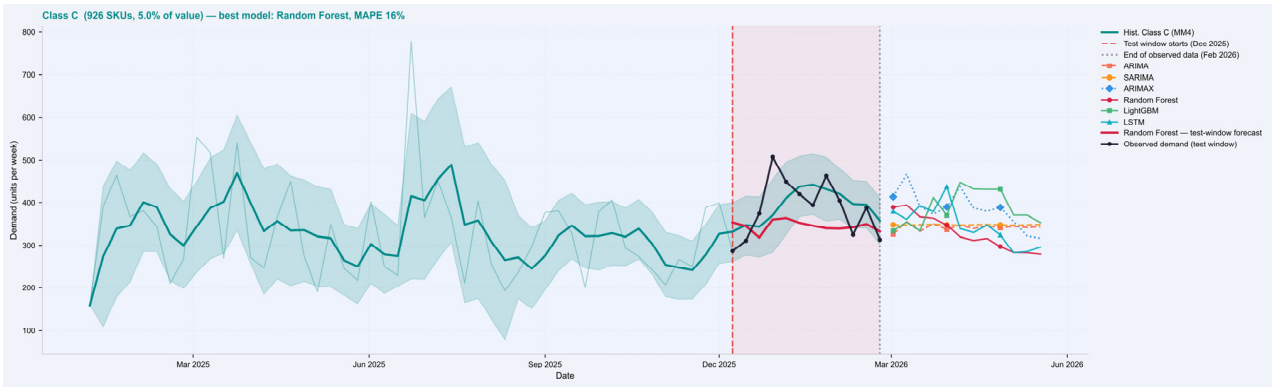


Figure 11. Class C: observed demand, test-window forecast and projection, with the same convention as Figure 9. The model plotted is Random Forest, the best of the six evaluated in this class; Section 4.8 shows that it does not improve on the last-value naive rule, which is the policy recommended for Class C. Shaded areas and lines as in Figure 9.

For class B, there is a clear separation of historical behavior from forecasts. While ARIMA/SARIMA ultimately show flattening toward the historical mean [8,9], Random Forest appears to capture a subtler upward slope and helps fit the demand troughs. The ability to weight different time windows (e.g., lag1, roll8, lag4) helps Random Forest disentangle weekly noise from medium-term structural changes [12]. From an operational standpoint, this means a better estimate of reorder points and economic order quantities [1]. If the higher training time is an issue, then LightGBM is quite a good fit [13].

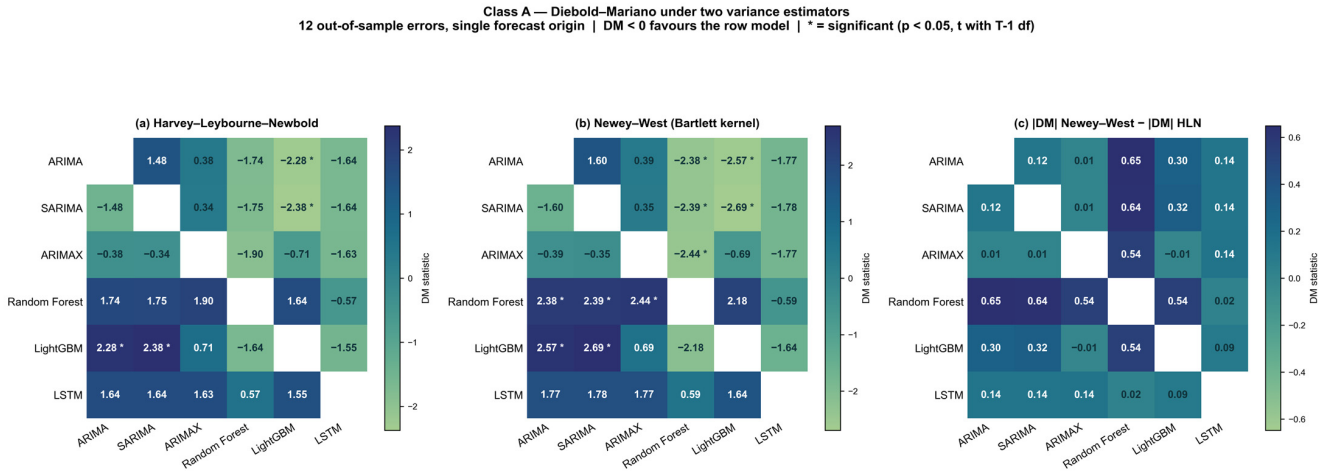
For class C, the series' inheritance of low amplitude (100–600 units) means a percentage deviation directly correlates with an operational effect on the order fill rate [30]. Random Forest fits the series closely, but the comparison that matters for this class is with the naive benchmark rather than with the other models, and Section 4.8 shows that it does not improve on it. A model of ARIMAX, which was successful in class A, seems to overreact in this case with pseudopoints of correspondence in history (overfitting) [8,16]. In low-turnover products, a disadvantage of ARIMAX is excessive complexity, especially in cases

with low 'commerce intensity' [19]. For this class, a last-value rule with periodic review is sufficient (Section 5.2.1).

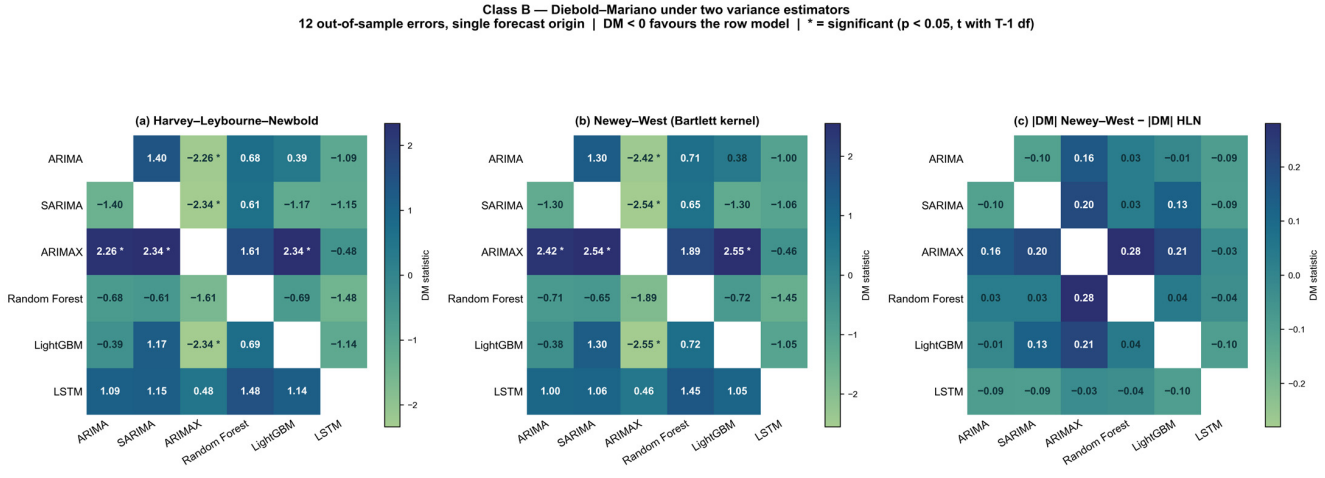
The cross-sectional analysis shows that model selection is both a statistical and operational issue [7]. Models in class A demand stability and interpretability, while models in class C demand micro-stockout avoidance and, therefore, absolute accuracy [1,2]. ABC segmentation offers the decision model to systematically introduce analytical capabilities to retail environments where the IT infrastructures are limited [17,30,31].

4.6. Statistical Significance Validation: Diebold–Mariano Test

To assess whether the differences in error metrics reflect real advantages or sampling noise, the Diebold–Mariano test ($\alpha = 0.05$) was applied [4,7], with a Newey–West correction for autocorrelation, appropriate here because the twelve errors are multi-step forecasts issued from a single origin [8]. Figure 12A–C report the same comparisons under both variance estimators, so that the effect of the choice can be judged. Of the 45 pairwise comparisons, sixteen are significant at the 5% level: five in class A, three in class B and eight in class C. The reference distribution is Student's t with $T - 1 = 11$ degrees of freedom. In the matrices, negative values indicate that the row model outperforms the column model, and the marks (*) denote significance ($p < 0.05$).

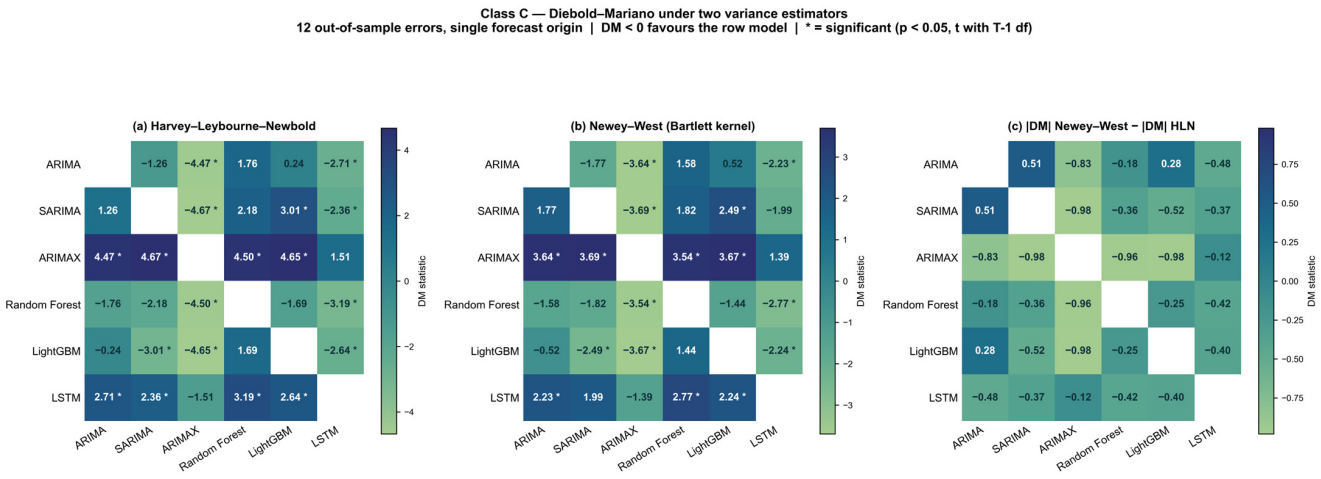


(**A**)



(**B**)

Figure 12. *Cont.*



(**C**)

Figure 12. (**A**) Class A. Diebold–Mariano statistics computed on the 12 out-of-sample errors of the holdout window, issued from a single forecast origin, under the two variance estimators: (**a**) Harvey–Leybourne–Newbold, (**b**) Newey–West with a Bartlett kernel, and (**c**) the difference in absolute value. A negative statistic favors the row model; an asterisk marks significance at the 5% level. (**B**) Class B. Diebold–Mariano statistics on the same 12 out-of-sample errors, under the two variance estimators, with the same convention as (**A**). (**C**) Class C. Diebold–Mariano statistics on the same 12 out-of-sample errors, under the two variance estimators, with the same convention as (**A**).

In class A, five of the fifteen comparisons are significant at the 5% level, and all five oppose a statistical specification to a tree ensemble, always in favor of the former: ARIMA beats Random Forest (DM = −2.38, $p = 0.036$) and LightGBM (−2.57, $p = 0.026$); SARIMA beats them likewise (−2.39 and −2.69); and ARIMAX beats Random Forest (−2.44, $p = 0.033$). The one statistical-ensemble pair that does not reach significance is ARIMAX against LightGBM (−0.69), so it is not the case that every statistical model is significantly better than every ensemble. The test also fails to separate the three statistical models from one another: ARIMA against ARIMAX gives 0.39 and SARIMA against ARIMAX 0.35. The evidence therefore supports allocating a statistical specification to this class but not the choice of ARIMAX within that family, which is the same conclusion that the rolling-origin evaluation reaches in Section 4.9.

The recurrent network, although it attains the highest errors of the six models in this class, is not significantly worse than any of them on these twelve errors. The submitted version reported that it was outperformed by every other model with statistics between −2.77 and −2.92; those values do not correspond to the estimated matrix, and the statement has been withdrawn.

In class B, three comparisons reach significance and all three identify ARIMAX as the weaker model: ARIMA against ARIMAX gives DM = −2.42 ($p = 0.034$) and SARIMA against ARIMAX −2.54 ($p = 0.028$), both favoring the model named first, while ARIMAX against LightGBM gives +2.55 ($p = 0.027$), favoring LightGBM. The opposite signs describe the same direction because ARIMAX occupies the first position in two pairs and the second in the third. Random Forest, the model allocated to this class, is not significantly better than any competitor on the single split, although it does beat LightGBM under rolling origins (DM = −2.31, $p = 0.041$).

In class C, eight comparisons are significant, the most of any class. ARIMAX is the weakest of the six and is beaten by ARIMA (DM = −3.64, p = 0.004), SARIMA (−3.69, p = 0.004), Random Forest (3.54, p = 0.005) and LightGBM (3.67, p = 0.004), the positive signs arising because ARIMAX is named first in the last two pairs; its exogenous calendar regressors add noise where the underlying signal is weak. LightGBM also beats SARIMA (2.49, p = 0.030). At the other end, the recurrent network is significantly worse than ARIMA (−2.23), Random Forest (−2.77) and LightGBM (−2.24). Random Forest, the most accurate of the six models in this class, is not significantly better than ARIMA, SARIMA or LightGBM, and Section 4.8 shows that it does not improve on a last-value naive rule either. The class is therefore best characterized not by which model wins but by the fact that none is needed.

Under rolling origins two comparisons in this class remain significant, ARIMA over ARIMAX (−2.29, p = 0.043) and Random Forest over ARIMAX (2.48, p = 0.031). Both preserve the conclusion about ARIMAX while the rest are withdrawn.

Table 9 consolidates the pairwise Diebold–Mariano statistics of Figure 12A–C alongside those obtained under the rolling-origin scheme of Section 4.9, so that the two evaluations can be read together. Only comparisons significant at the 5% level under at least one scheme are listed.

Table 9. Every pairwise Diebold–Mariano comparison significant at the 5% level under at least one evaluation scheme, out of 45 comparisons under the single split and 30 under rolling origins. Both schemes use the same denominator and the same Harvey–Leybourne–Newbold small-sample correction, so that the only difference between the two estimators is the Bartlett-weighted autocovariance term of the Newey–West version; the reference distribution is Student's t with T − 1 degrees of freedom, which is 11 for the single split and for rolling origins in classes B and C and 8 for rolling origins in class A, where nine paired origins are available. A negative statistic favors the model named first in the pair; an asterisk marks significance under that scheme. Rolling origins use nine paired origins in class A and twelve in classes B and C and do not include the recurrent network.

Class	Comparison (Row vs. Column)	DM, Single Split (Newey–West, n = 12)	DM, Rolling Origin (HLN)	Favored Model	Significant Under
A	ARIMA vs. LightGBM	−2.57 *	−1.10	ARIMA	single split
A	ARIMA vs. Random Forest	−2.38 *	−1.27	ARIMA	single split
A	ARIMAX vs. Random Forest	−2.44 *	−1.05	ARIMAX	single split
A	SARIMA vs. LightGBM	−2.69 *	1.27	SARIMA	single split
A	SARIMA vs. Random Forest	−2.39 *	0.58	SARIMA	single split
B	ARIMA vs. ARIMAX	−2.42 *	1.06	ARIMA	single split
B	ARIMAX vs. LightGBM	2.55 *	−1.93	LightGBM	single split
B	Random Forest vs. LightGBM	−0.72	−2.31 *	Random Forest	rolling origin
B	SARIMA vs. ARIMAX	−2.54 *	1.48	SARIMA	single split
C	ARIMA vs. ARIMAX	−3.64 *	−2.29 *	ARIMA	single split; rolling origin
C	ARIMA vs. LSTM	−2.23 *	not evaluated	ARIMA	single split
C	ARIMAX vs. LightGBM	3.67 *	2.10	LightGBM	single split
C	ARIMAX vs. Random Forest	3.54 *	2.48 *	Random Forest	single split; rolling origin
C	LightGBM vs. LSTM	−2.24 *	not evaluated	LightGBM	single split
C	Random Forest vs. LSTM	−2.77 *	not evaluated	Random Forest	single split
C	SARIMA vs. ARIMAX	−3.69 *	−0.64	SARIMA	single split
C	SARIMA vs. LightGBM	2.49 *	1.31	LightGBM	single split

The asterisk marks a comparison that is significant at the 5% level.

The comparison between the two schemes is itself informative. Sixteen of the forty-five comparisons reach significance under the single split and three of the thirty under rolling origins. Two of the sixteen survive into the rolling-origin scheme, both in class C: ARIMA over ARIMAX and Random Forest over ARIMAX. The rest do not, and one comparison, Random Forest over LightGBM in class B, is significant only under rolling

origins. The difference between the two schemes is not only one of sample size but of what each measures: the single split evaluates a twelve-week trajectory issued from one origin, in which a model that drifts early accumulates error over the whole window, whereas the rolling scheme evaluates twelve independent one-week decisions, which is closer to how the forecast is used in weekly replenishment.

Read together, the class-level results yield an operational rule with its evidence stated alongside it. Class A should be served by a statistical specification, a choice the test supports against both ensembles, although it does not discriminate among ARIMA, SARIMA and ARIMAX, and the two evaluation designs disagree on which of them to use: ARIMAX on the single holdout window, ARIMA under rolling origins (Section 4.9). Class B should be served by Random Forest, supported against LightGBM under rolling origins and not contradicted elsewhere. Class C needs no model at all, since none improves on a last-value rule (Section 4.8).

Because the manuscript reports one variance estimator for each evaluation scheme, Figure 12A–C apply both to the same set of errors so that the consequence of the choice can be judged. The two estimators are computed with the same denominator and the same small-sample correction, so that the only difference between them is the Bartlett-weighted autocovariance term. The Newey–West version identifies sixteen significant comparisons out of forty-five against thirteen for the version without that term, but the direction of the difference is not uniform: it is larger in classes A and B and smaller in class C, where the autocovariances of the loss differential are positive and the correction therefore inflates the variance, as it should. In every cell the two estimators agree on which model of the pair is favored, and they differ only in how much evidence they attribute to the difference.

4.7. Demand Regimes at the SKU Level

Table 10 reports the classification of the 1633 SKUs into the four demand regimes defined by the average inter-demand interval and the squared coefficient of variation of non-zero demand sizes [18]. The composition of each ABC class differs markedly. The economically most important class concentrates the least predictable regime, while the low-value class is dominated by infrequent but regular demand. Taken together, 95.5% of the portfolio is intermittent or lumpy at the item level, which explains why aggregation by inventory class is a precondition for applying the forecasting families evaluated in this study.

Table 10. Distribution of demand regimes across ABC classes (SBC classification, weekly grid, 60 weeks).

Class	N SKUs	Smooth	Erratic	Intermittent	Lumpy	Median ADI	Median CV^2	Predominant Regime
A	223	5.4%	13.0%	39.9%	41.7%	2.35	0.53	Lumpy
B	484	0.8%	5.6%	71.3%	22.3%	6.10	0.25	Intermittent
C	926	0.0%	0.2%	83.6%	16.2%	15.25	0.09	Intermittent
Portfolio	1633	1.0%	3.6%	74.0%	21.5%	—	—	Intermittent

4.8. Comparison Against Naive Benchmarks

To establish that the models add value over a trivial alternative, all forecasts were compared against two benchmarks: a last-value naive forecast and a seasonal naive forecast with $m = 4$. Accuracy is also reported with MASE and RMSSE, which scale the error by the in-sample naive error and are therefore comparable across classes and well behaved on low-magnitude series [34]. Table 11 reports the results.

The value added by modeling is not uniform across the portfolio. In class A the selected model (ARIMAX, RMSE 843.2) reduces the error of the last-value naive forecast

(1401.5) by 39.8%, and in class B the selected model (Random Forest, RMSE 254.7) reduces it by 47.6% relative to the same benchmark (485.9). In class C, by contrast, no model outperforms the last-value naive forecast on the holdout window: the benchmark attains RMSE 66.5 and MAE 54.4, against 74.6 and 64.2 for Random Forest. Under the rolling-origin evaluation reported below the gap reverses only marginally, with Random Forest at RMSE 66.3 against 68.39 for a last-value naive forecast, which is evaluated over eleven of the twelve origins because it requires a preceding observation (Table 12), an improvement of 3%.

Table 11. Accuracy of the six models and of the two naive benchmarks on the holdout window (12 weeks). MAE and RMSE in units per week.

Class	Model	Type	MAE	RMSE	MAPE (%)	MASE	RMSSE
A	Naive (last value)	Benchmark	1228.22	1401.52	121.08	1.150	1.048
A	Seasonal naive (m = 4)	Benchmark	1331.19	1785.34	121.59	1.247	1.335
A	ARIMA	Statistical	753.69	886.29	73.66	0.706	0.663
A	SARIMA	Statistical	746.73	881.47	73.07	0.699	0.659
A	ARIMAX	Statistical	661.71	843.15	63.90	0.620	0.631
A	Random Forest	ML	1027.48	1255.75	102.45	0.962	0.939
A	LightGBM	ML	796.58	924.75	78.10	0.746	0.692
A	LSTM	ML	1075.90	1326.60	107.94	1.008	0.993
B	Naive (last value)	Benchmark	430.85	485.86	57.58	1.053	0.838
B	Seasonal naive (m = 4)	Benchmark	445.08	523.24	52.51	1.088	0.902
B	ARIMA	Statistical	198.32	288.60	28.15	0.485	0.498
B	SARIMA	Statistical	195.43	283.66	27.88	0.478	0.489
B	ARIMAX	Statistical	263.70	354.64	36.54	0.644	0.611
B	Random Forest	ML	191.60	254.65	27.58	0.468	0.439
B	LightGBM	ML	197.23	286.58	28.28	0.482	0.494
B	LSTM	ML	274.01	387.50	40.44	0.669	0.668
C	Naive (last value)	Benchmark	54.41	66.47	15.17	0.458	0.414
C	Seasonal naive (m = 4)	Benchmark	76.91	97.43	19.36	0.648	0.607
C	ARIMA	Statistical	71.85	85.13	17.22	0.605	0.531
C	SARIMA	Statistical	73.21	91.18	17.26	0.616	0.568
C	ARIMAX	Statistical	159.06	176.24	39.32	1.339	1.098
C	Random Forest	ML	64.20	74.60	15.83	0.541	0.465
C	LightGBM	ML	69.37	84.20	16.62	0.584	0.525
C	LSTM	ML	134.40	143.21	35.38	1.131	0.893

This result is reported rather than omitted because it sharpens the argument of the paper instead of weakening it. If forecasting effort is to be allocated according to inventory characteristics, then class C, which represents 5% of inventory value, sells intermittently and exhibits the lowest dispersion of the three classes, and does not justify the estimation, tuning and maintenance cost of a machine-learning model. A last-value rule is sufficient for that segment. Class-conditional allocation therefore extends beyond choosing among models: it includes deciding where a model is warranted at all. Two further results support the same logic. The LSTM in class A improves on the last-value naive forecast only marginally and not robustly: the seeded run reported in Table 11 attains an RMSE of 1326.60 against 1401.5 for the benchmark, but the mean over five seeds is 1431.3 with a standard deviation of 157.3, so the comparison reverses depending on the draw (Section 4.12). Either way the network remains 57% worse than the model allocated to the class; and ARIMAX in class C (RMSE 176.2) is 1.8 times worse than the seasonal naive benchmark, confirming that its exogenous regressors act as noise where the underlying signal is weak.

Table 12. Rolling-origin evaluation. The five models are compared over the origins at which all of them converged: nine in class A, where ARIMAX failed to converge at two origins and SARIMA at one, and twelve in classes B and C. The last-value benchmark is evaluated over one origin fewer in each class, since it requires an observation preceding the origin; the Origins column states the number used in every row. MAE and RMSE are in units per week.

Class	Model	Origins	MAE	RMSE
A	ARIMA	9	558.07	682.61
A	ARIMAX	9	609.89	828.54
A	LightGBM	9	681.41	833.62
A	Random Forest	9	835.5	1001.28
A	SARIMA	9	1003.08	1192.55
A	Seasonal naive (m = 4)	9	1132.34	1335.37
A	Naive (last value)	8	1048.39	1228.16
B	ARIMA	12	279.45	340.52
B	ARIMAX	12	250.57	308.97
B	LightGBM	12	418.56	498.5
B	Random Forest	12	203.64	267.05
B	SARIMA	12	393.96	565.04
B	Seasonal naive (m = 4)	12	373.86	458.22
B	Naive (last value)	11	265.67	358.99
C	ARIMA	12	68.06	77.73
C	ARIMAX	12	88.87	104.78
C	LightGBM	12	53.02	69.09
C	Random Forest	12	55.21	66.29
C	SARIMA	12	74.54	94.11
C	Seasonal naive (m = 4)	12	69.62	86.75
C	Naive (last value)	11	61.77	68.39

4.9. Rolling-Origin Evaluation and Statistical Significance

Restricting the comparison to origins common to all models leaves classes B and C unchanged and changes the reading of class A. In class B, Random Forest attains the lowest error (RMSE 267.1), and in class C it does so as well (66.3), confirming the allocation in class B and the ranking among models in class C. In class A, ARIMA attains 682.6 against 828.5 for ARIMAX over the nine paired origins, and it does so consistently: when the allocation is re-derived at each origin from prior information only, ARIMA is selected at every one of the six decision points (Section 4.10). Under this scheme ARIMA is therefore a stable winner in class A, and ARIMAX is preferred only on the single holdout window, where it attains 843.2 against 886.3. The Diebold–Mariano test does not separate the two under either scheme (0.39 on the single split), so the disagreement is between evaluation designs rather than between models of demonstrably different accuracy. SARIMA, competitive under the single split, deteriorates markedly under rolling origins in all three classes, which indicates that its apparent advantage was partition-dependent.

Applying the Diebold–Mariano test to these one-step errors, only three of the thirty pairwise comparisons are significant at the 5% level: Random Forest over LightGBM in class B (DM = −2.31, p = 0.041), ARIMA over ARIMAX in class C (−2.29, p = 0.043) and Random Forest over ARIMAX in class C (2.48, p = 0.031). This is a more conservative outcome than the one obtained on the single holdout window reported in Section 4.6, and the difference is methodological: multi-step errors from a single origin are strongly autocorrelated, which inflates the test statistic, whereas one-step errors from successive origins are not. Both sets of results are reported for transparency.

The consequence for the interpretation is stated explicitly. The evidence does not establish that any single model is statistically superior within a class; what it establishes is that the ordering of models changes systematically across classes, that ARIMAX is not

competitive in class C, and that the LSTM is not competitive anywhere. Where competing models are statistically indistinguishable, the allocation is resolved on operational grounds, namely transparency, treatment of calendar effects and maintenance cost, and this study makes that criterion explicit rather than presenting an accuracy ranking as if it were statistically decisive.

4.10. Stability of the Allocation over Time

The framework proposes that the allocation of models to inventory classes be reviewed periodically. To test whether such a review would in fact alter the allocation, the assignment was re-derived at each rolling origin (nine in class A, twelve in classes B and C), selecting at every step the model with the lowest cumulative root mean squared error computed exclusively from earlier origins. Table 13 reports how often each model would have been chosen and compares the accuracy of this adaptive re-selection with that of holding the allocation fixed.

Table 13. Allocation re-derived at each rolling origin using only prior information. The first three origins of each class serve as history and yield no decision, so the errors reported here cover the decision points only: six in class A and nine in classes B and C. They are therefore a subset of the origins used in Table 12, and the two tables are not directly comparable; the last column gives the error of every model over these same decision points, so that the comparison can be made within this table. RMSE is in units per week.

Class	Decision Points	Model Selected (Times)	RMSE, Adaptive Re-Selection	RMSE, Fixed Allocation	RMSE, Seasonal Naive	RMSE of Every Model over the Same Decision Points
A	6	ARIMA 6	669.28	717.28	1134.23	ARIMA 669.3; SARIMA 1075.7; ARIMAX 717.3; RF 826.1; LightGBM 720.1
B	9	Random Forest 9	295.21	295.21	383.99	ARIMA 324.3; SARIMA 628.6; ARIMAX 279.3; RF 295.2; LightGBM 452.0
C	9	Random Forest 8; SARIMA 1	80.6	71.32	97.77	ARIMA 85.5; SARIMA 106.5; ARIMAX 106.3; RF 71.3; LightGBM 76.6

The allocation is stable in all three classes, but in class A it is stable around a different model. Re-deriving it from scratch at every origin reproduces the recommended model at nine of nine decision points in class B and eight of nine in class C. In class A it selects ARIMA at all six decision points rather than the allocated ARIMAX, which is not instability but a consistent preference for a model that the single-split evaluation did not select.

Whether continuous re-selection pays follows from that. In class A, re-selecting at each origin reduces the error of the fixed allocation from 717.3 to 669.3, but the reduction comes from switching to ARIMA and holding it, not from switching repeatedly: the same 669.3 would be obtained by allocating ARIMA from the outset. In class B the two coincide by construction, since the same model is selected throughout, and in class C re-selection is slightly worse, 80.6 against 71.3, because the single origin at which SARIMA is chosen is a poor one. Re-selection therefore does not improve on a correct fixed allocation in any class; what it does is reveal, in class A, that the fixed allocation derived from one partition was not the one the data support out of sample. All three schemes beat the seasonal naive benchmark in every class.

4.11. Sensitivity to the ABC Cut-Off Values

Because every class-level result depends on the cut-offs that define the segmentation, the analysis was repeated under five alternative pairs. For each pair the class-aggregated series were rebuilt and all five non-recurrent models re-estimated on the same partition. Table 14 reports, for each combination, the best-performing model and the excess error of the model allocated by the framework.

Table 14. Sensitivity of the model allocation to the ABC cut-off values. Excess error is the additional RMSE of the allocated model relative to the best model at that cut-off.

Cut-Off	Class	SKUs	Revenue (%)	Units (%)	Best Model	Rank of Allocated Model	Excess Error (%)
75/95	A	159	74.9	26.4	ARIMA	2	0.5
80/90	A	223	80.0	58.2	ARIMAX	1	0.0
80/95	A	223	80.0	58.2	ARIMAX	1	0.0
80/97	A	223	80.0	58.2	ARIMAX	1	0.0
85/95	A	317	85.0	68.8	ARIMA	3	2.9
75/95	B	548	20.0	63.3	ARIMAX	5	42.2
80/90	B	238	10.0	22.2	SARIMA	4	11.3
80/95	B	484	15.0	31.5	Random Forest	1	0.0
80/97	B	656	17.0	35.0	Random Forest	1	0.0
85/95	B	390	10.0	20.9	ARIMA	3	14.1
75/95	C	926	5.0	10.2	Random Forest	1	0.0
80/90	C	1172	10.0	19.6	LightGBM	2	1.2
80/95	C	926	5.0	10.2	Random Forest	1	0.0
80/97	C	754	3.0	6.8	ARIMAX	5	16.5
85/95	C	926	5.0	10.2	Random Forest	1	0.0

The allocation is invariant to the boundary that does not define a given class. Class A retains ARIMAX in first place at every cut-off where the A/B boundary remains at 80%, and class C retains Random Forest at every cut-off where the B/C boundary remains at 95%. Class B, delimited by both boundaries, is the least stable of the three, which is coherent rather than anomalous. Where the allocation changes, the magnitude varies greatly: in two of the seven cases the difference is negligible (0.5% in class A at 75/95 and 1.2% in class C at 80/90), well inside the range the Diebold–Mariano test cannot resolve at this sample size. The allocation holds in eight of the fifteen combinations, and in ten if the two negligible differences are counted as ties. At the level of model family the result is more robust still: the best model for class A belongs to the ARIMA family at all five cut-offs, and the best model for class C is a tree ensemble at four of five, so the central claim of the study survives the perturbation even where the specific member of the family changes. The allocation holding at 80/95 is not independent evidence, since that is the cut-off at which it was derived; the informative results are those at the other four, which are reported here without selection.

The largest reversal is informative in its own right. ABC segmentation is computed from revenue, whereas forecast error scales with physical volume, and the two are not aligned in this portfolio: the 64 SKUs lying between the 75th and the 80th percentile of cumulative revenue account for 31.8 percentage points of total unit volume. When the A/B boundary is moved to 75%, class A retains 74.9% of revenue but only 26.4% of units, and class B becomes the high-volume series; the model allocated to class B is then applied to a series it was not selected for, and its excess error reaches 42.2%. In portfolios where value and volume diverge, a segmentation criterion combining both may therefore allocate forecasting models more reliably than revenue alone, which we identify as the most promising extension of the framework.

4.12. Reproducibility of the Recurrent Network Result

The underperformance of the recurrent network is one of the conclusions of this study, and a single training run without a fixed seed is not an adequate basis for it. The network was therefore re-estimated five times per class with different seeds, under two configurations: the one used throughout this study, in which the feature matrix enters as a single time step, and a sequence configuration with an eight-week look-back window over the raw series, which is the form in which a recurrent architecture can exploit temporal dependence. Table 15 reports the mean and standard deviation across runs.

Table 15. Recurrent network over five seeded runs per class and configuration. MAE and RMSE in units per week.

Class	Configuration	MAE (Mean ± SD)	RMSE (Mean ± SD)	MAPE % (Mean ± SD)	Best Model in Class, RMSE
A	Single time step	1169.7 ± 152.1	1431.3 ± 157.3	116.8 ± 13.1	ARIMAX, 843.2
A	8-week sequence	1198.0 ± 274.8	1406.4 ± 363.0	104.8 ± 22.8	ARIMAX, 843.2
B	Single time step	286.4 ± 42.6	369.0 ± 55.8	39.8 ± 6.3	Random Forest, 254.7
B	8-week sequence	247.8 ± 14.5	347.6 ± 11.4	35.4 ± 1.4	Random Forest, 254.7
C	Single time step	128.5 ± 21.3	139.1 ± 20.2	33.5 ± 5.7	Random Forest, 74.6
C	8-week sequence	123.5 ± 17.9	150.5 ± 33.3	33.9 ± 6.0	Random Forest, 74.6

The conclusion is unchanged and is now supported by a distribution rather than a single draw: in every class and under both configurations the recurrent network is outperformed by the model allocated to that class, by margins far larger than the dispersion across seeds. The sequence configuration does not rescue it. This matters because a recurrent architecture fed with a single time step cannot exploit sequential dependence by construction, so its failure in that form would not by itself have supported a conclusion about recurrent networks; the sequence variant removes that objection, and the result holds.

Two qualifications are recorded. The values reported in Tables 4–8 and 11 and plotted in Figure 8 come from the run with the fixed seed, which is the reproducible configuration described in Section 3; the present section reports the dispersion around it. In class A that run attains an RMSE of 1326.60 and therefore improves marginally on the last-value naive benchmark of 1401.5, whereas the five-run mean of 1431.3 does not, so the comparison with the benchmark is not robust to the seed and the manuscript does not rest any conclusion on it. In classes B and C the ordering is unaffected: the network beats the benchmark in B and is beaten by it in C under both the seeded run and the mean. What does not depend on the seed is the comparison that matters, since in every class and under both configurations the network is outperformed by the allocated model by margins far larger than the dispersion across seeds.

5. Discussion

The results support the central hypothesis in a qualified form: ABC segmentation is a useful basis for allocating forecasting models in resource-constrained retail settings, provided that the allocation is read together with the naive benchmark and with the statistical evidence for each class. All comparisons are made at the level of aggregate demand by ABC class, not at the individual-SKU level. This section interprets the results in three parts: their implications for theory (Section 5.1), their implications for practice and policy (Section 5.2), and the limitations of the study together with the research directions that follow from them (Section 5.3).

5.1. Implications for Theory

The study contributes to theory in three ways, each narrower than a general claim about which model suits which segment. First, it provides an empirical counterpart, at the level of the inventory segment, to the result that the value of forecasting depends on the structure of the demand process rather than on the sophistication of the estimator [21]: within a single portfolio, the reduction in error that forecasting achieves over a naive rule ranges from 47.6% in class B to none in class C (Section 4.8). Second, it shows that the segment boundary is itself a determinant of the operational answer, since moving the A/B cut-off by five points of cumulative revenue transfers roughly a third of physical volume between classes (Section 4.11). Third, it shows that where models cannot be statistically distinguished, the one selected depends on the evaluation design: ARIMAX on a single holdout window, ARIMA under rolling origins (Section 4.9).

The concentration of 80% of revenue in 223 class-A products (13.7% of the portfolio) is consistent with the Pareto principle in Colombian hardware retail [1,31]. The most relevant finding is that the majority of these critical products are highly variable, which challenges the assumption that high value implies high predictability [17,30].

Why does Random Forest rank first among the models in B and C? ARIMA/SARIMA assume linearity and, in the absence of a strong trend, tend to revert to the mean. Random Forest non-parametrically models non-linear interactions (e.g., between week_no, roll8, and lag1) without distributional assumptions, explaining its robustness against the non-normality in class B (Shapiro–Wilk $p < 0.001$) and the heavy tails of sporadic demand in C [11,12]. The price is the loss of parametric interpretability, acceptable in low-value segments.

Why does LSTM fail in this context? The inferior performance of LSTM (mean MAPE over five seeded runs of 33.5%, 39.8% and 116.8% in classes C, B and A; Table 15) does not contradict the literature reporting its superiority [3]; it delimits its conditions of validity. With a weekly history of ~60 points per series, the model operates in severe data scarcity, where its expressive capacity translates into overfitting to noise [11,16]. LSTM was used only as a methodological benchmark. The comparison shows the drawbacks of deep networks in SMEs with short sequences.

The trade-off between interpretability and accuracy must be noted Statistically speaking, the Diebold–Mariano test cannot separate ARIMAX from ARIMA or SARIMA in class A, while the statistical specifications beat the tree ensembles in five of the six pairs that oppose the two families. When multiple methods provide approximately the same results, the preference shifts from accuracy to operational parsimony [8,9]. The proposed structure, as per Henao-Baena et al. [31], follows the correct balance between accuracy and operational practicality by designing the complexity of the models to the demand for predictability, rather than focusing the most on the economic aspect.

Two further implications concern the literature on model selection. The evidence agrees with the large-scale competitions in finding that no single family dominates [7,32,33], but it locates the source of that heterogeneity in the demand regime of the segment rather than in the individual series: the lumpy regime concentrated in class A favors parsimonious statistical specifications, while the more regular aggregate of class B rewards the non-linear interactions captured by tree ensembles. And the misalignment between revenue and physical volume documented in Section 4.11 indicates that value and variability are complementary rather than substitutable bases for routing. ABC classification captures economic weight but not predictability, and the ABC-XYZ matrix that joins the two [30] is therefore the most promising extension of class-conditional allocation; testing it against ABC alone is set out as future work in Section 5.3.

5.2. Implications for Practice and Policy

The practical implications are of two kinds: the inventory decisions that the allocation supports directly and the obstacles that remain once it is deployed, which are where the implications for policy lie.

5.2.1. Implications for Inventory Management

When utilizing a periodic-review model, the safety-stock (SS) estimation to achieve a specific service level can be stated as SS = z · RMSE · sqrtL. In this model, z is the service factor, for instance z = 1.65 for a 95% service level. In addition, L is the time in weeks of the replenishment lead time [1,2]. Taking the RMSE of the model selected per class and assuming an illustrative lead time of L = 1 week, we obtain the following:

- Class A (ARIMAX, RMSE = 843): SS $\approx 1.65 \times 843 \approx 1391$ units per class per week.
- Class B (Random Forest, RMSE = 255): SS $\approx 1.65 \times 255 \approx 421$ units.
- Class C (last-value naive rule, RMSE = 66): SS $\approx 1.65 \times 66 \approx 110$ units.

Section 4.8 shows that no model tested improves on the naive rule in this class, so the safety stock is derived from the benchmark rather than from Random Forest. Deriving it from the Random Forest error of 75 units, as the submitted version did, would have raised the required buffer to 124 units for the same service level: the prescribed policy would have increased immobilized inventory rather than reducing it.

The three MAPE values are accuracies of different series and are not comparable with one another; what matters for inventory is the reduction in error achieved within each class relative to the benchmark available for that class. Where such a reduction exists, an improvement in RMSE translates directly into less immobilized inventory for the same service level. To express this saving in monetary terms, the reduction in SS units must be multiplied by the unit cost and FERRETODO's inventory holding rate; these figures are not in the available data and must be taken from the company's records.

Based on the above, concrete recommendations are formulated for FERRETODO and similar microenterprises:

(1) Differentiated replenishment policies: for class-A products, weekly periodic reviews with safety stock computed from the ARIMAX RMSE and flexibility agreements with suppliers. For class B, a dynamic reorder point based on Random Forest forecasts, with biweekly reviews. For class C, a last-value rule with monthly review: Section 4.8 establishes that no model tested improves on that benchmark in this class, so running and maintaining a model there adds cost without accuracy.

(2) Scalable forecasting architecture: a model router that automatically assigns ARIMAX for class A and Random Forest for class B, with a last-value rule for class C according to the ABC classification, deployable in low IT infrastructure environments through weekly Python scripts [11,31].

(3) Monitoring and recalibration: review the model allocation periodically rather than in every period. Section 4.10 shows that re-deriving the allocation at each origin never improves on a correct fixed allocation; its value is diagnostic, in that it exposed the disagreement between evaluation designs in class A. A periodic review, run on the accumulated out-of-sample errors rather than on a single partition, is what the evidence supports. The intervals at which the ABC classification itself should be recomputed were not tested in this study and are stated as a subject for future work. The Ljung–Box test can serve as an early alert of residual degradation [8,9].

5.2.2. Remaining Challenges and Implications for Policy

Deploying the framework at FERRETODO resolved the question of which model to run for each class and left a set of obstacles that no forecasting method addresses.

They are reported here because they are, in our experience of this implementation, the binding constraints on forecast-driven replenishment in firms of this size and because they indicate where effort and public support would be better directed than towards more elaborate estimators.

Demand is censored and the firm cannot measure it. The point-of-sale system records sales, not demand: when an item is out of stock the sale does not occur and leaves no trace, so every series analyzed here understates true demand by an unknown amount, concentrated precisely in the periods when forecasting matters most. The firm keeps no stockout log and no record of unfilled customer requests. Until that is recorded, no forecasting method can distinguish a genuine fall in demand from a failure to supply it, and the safety stocks computed in Section 5.2.1 are calibrated on the observable rather than the relevant quantity.

Lead times are not recorded. Replenishment lead time enters the safety-stock formula directly, but the firm holds no series of order and receipt dates, so a constant value had to be assumed. Lead-time variability is often a larger contributor to required safety stock than demand variability, and its absence is a more serious limitation of the inventory results than the choice among the six forecasting models.

Forecasts are produced at class level and decisions are taken at item level. The framework forecasts three aggregated series because 95.5% of the portfolio is intermittent or lumpy at the item level, but a purchase order is placed for a specific SKU. The disaggregation of a class forecast into item-level replenishment quantities is currently done by historical share, which is a placeholder rather than a solution, and is the most consequential open problem left by this study.

The assortment contains a long inactive tail. Of the 1633 SKUs, 199 record a single transaction in fourteen months and 539 record demand in three weeks or fewer out of sixty. For these items the operational question is not how to forecast them but whether to stock them at all, which is an assortment decision that the framework does not address and that the firm has not made.

The class boundary is an operational parameter, not a statistical one. Section 4.11 shows that moving the A/B boundary by five percentage points of cumulative revenue transfers roughly a third of the physical sales volume between classes, because value and volume are not aligned in this portfolio. The boundary therefore has to be set with the replenishment consequences in view, and the firm currently sets it by convention.

Finally, the organizational constraint is severe and is easily underestimated in academic treatments. The person who places orders is the owner, who has no analytics training and no time allocated to the task. A procedure that requires weekly retraining, hyperparameter maintenance or interpretation of statistical output will not survive contact with that reality, which is the reason the framework was built on fixed specifications and open-source tools and the reason the finding that class C does not warrant a model at all is useful rather than disappointing. For policymakers designing support programs for retail SMEs, the implication is that funding data capture, specifically stockout and lead-time recording, would raise forecast-driven replenishment performance more than funding access to advanced analytics.

5.3. Limitations of the Study and Future Research Directions

The findings should be read in light of the following limitations:

- Dataset size and time horizon: ~60 weekly observations per series (48 training, 12 test), which limits statistical power and may explain the lack of significance in some Diebold–Mariano comparisons.

- Aggregation by ABC class: results are presented at the aggregate-class level, not by individual SKU; SKU-level implementation would require validating each model per product and a longer history.
- Comparison with LSTM: included for methodological-benchmark purposes; the comparison is not equitable in data volume.
- MAPE drawbacks: MAPE performs poorly when there are many observations close to zero, which is common to low-turnover classes. sMAPE, MASE, or RMSSE for example, may be better alternatives [7,8]
- Exogenous factors: ARIMAX included two calendar regressors. ARIMAX, however, did not include competitor's prices, competitor's promotions, or macroeconomic factors [10,17].
- Context: the focus of the research is one hardware store in Magdalena, Colombia, and demand patterns may be different in other locations.
- Statistical power: with twelve out-of-sample errors per model, the Diebold–Mariano test detects only large differences in predictive accuracy. Most pairwise comparisons are therefore inconclusive, and the reported allocation should be read as operationally justified rather than statistically demonstrated.
- Value of modeling in low-turnover segments: a longer history would allow the question to be revisited, but on the present evidence class C does not warrant a forecasting model and Recommendation 1 assigns it a last-value rule.

Each of these limitations defines a direction for future research, and the results add several more:

(1) Item-level forecasting for intermittent demand. Applying Croston-type estimators [18,20] at the SKU level and reconciling the forecasts with the class totals would address both the aggregation adopted here and the disaggregation problem described in Section 5.2.2.
(2) A combined routing criterion. Testing the ABC-XYZ matrix against ABC alone, on the same portfolio and with the same evaluation protocol, would establish whether joining predictability to economic weight improves the allocation (Section 5.1).
(3) Probabilistic forecasting. Extending the framework to predictive distributions would allow safety stocks to be set from forecast quantiles rather than from a point error and would make the inventory consequences of the Class A disagreement between ARIMA and ARIMAX directly measurable.
(4) Censored demand and lead times. Estimating true demand from sales records that are truncated by stockouts and recording lead times are preconditions for any of the above in firms of this kind, and Section 5.2.2 argues that they would repay effort more than further estimators.
(5) External validity. Replicating the protocol on several firms and sectors, with longer histories, would show whether the class-conditional pattern observed here generalizes or reflects the demand structure of a single hardware retailer.
(6) Hybrid and accessible implementations. Combining statistical and machine-learning components within a class and packaging the procedure for planners without analytics training would extend its reach to the resource-constrained firms that motivated it.

6. Conclusions

This study proposed and validated a class-conditional inventory forecasting framework that integrates ABC inventory segmentation with differentiated forecast model allocation to support inventory replenishment planning in retail SMEs. Unlike conventional forecasting studies that seek a universally superior forecasting algorithm, the proposed approach recognizes the heterogeneity of inventory portfolios and demonstrates that fore-

casting performance depends on the operational characteristics of each inventory segment. By formalizing ABC segmentation as a decision-support mechanism for forecast model allocation, the framework establishes a direct link between inventory classification and forecasting strategy selection.

The empirical results, obtained from historical sales data comprising 1633 SKUs from a Colombian hardware retailer, confirm that no single forecasting model consistently outperforms all alternatives. Instead, ARIMAX provides the most suitable balance between predictive accuracy and interpretability for economically critical class-A products on the holdout window, although ARIMA attains lower error over paired rolling origins and the two are not statistically distinguishable; Random Forest is the strongest model for class B; and in class C no model tested improves on a last-value naive rule, so the class is best served without one. Relative to that benchmark, the allocated models reduce forecast error by 39.8% in class A and 47.6% in class B. These findings provide empirical support for the adoption of differentiated forecasting strategies rather than standardized forecasting policies across heterogeneous inventory portfolios.

A second conclusion concerns how such an allocation should be evaluated. On the single holdout window ARIMAX is the best model for class A, whereas over paired rolling origins ARIMA attains lower error (682.6 against 828.5) and is selected at every decision point. The Diebold–Mariano test separates sixteen of forty-five pairs on the holdout window but only three of thirty under rolling origins. A model allocated on the strength of one partition can therefore differ from the one the data support out of sample, and the allocation should rest on errors accumulated over several origins and be reported together with its statistical evidence, not on point accuracy from a single split.

A third conclusion is that the segment boundary is part of the decision rather than a fixed input to it. In this portfolio revenue and physical volume are not aligned, and moving the A/B cut-off by five points of cumulative revenue transfers roughly a third of unit volume between classes, which in some combinations changes the model that each class should receive. The boundary should therefore be set with its replenishment consequences in view, and a routing criterion that joins economic weight to predictability is the most promising refinement of the framework.

From a logistics perspective, the framework contributes beyond forecasting accuracy by supporting replenishment decisions. Where forecasting reduces error relative to the naive benchmark, as in classes A and B, the reduction translates into smaller safety stocks at the same service level; where it does not, as in class C, the contribution of the framework is to establish that a model is not needed. Because the methodology relies exclusively on historical sales data and open-source tools, it is within reach of SMEs operating under limited technological and financial resources.

The implementation also showed that the binding constraints in a firm of this size lie upstream of the forecasting model. Demand is censored by stockouts that leave no record, replenishment lead times are not registered, and decisions are taken by an owner without analytics training or time allocated to the task. For support programs aimed at retail SMEs, the implication is that funding the recording of stockouts and lead times would raise forecast-driven replenishment performance more than funding access to more sophisticated estimators.

The practical implication for a firm of this size is narrower and more concrete than a general principle. Forecasting effort is a scarce resource, and this study shows where it repays investment: substantially in the classes that concentrate inventory value and not at all in the low-turnover tail, where a last-value rule matches the best model tested. Deciding where to model is therefore part of the allocation problem, not a preliminary to it.

The framework should be interpreted as an initial validation based on demand aggregated by ABC class rather than at the level of the individual SKU in a single firm. The research directions that follow, set out in Section 5.3, are item-level forecasting for intermittent demand reconciled with the class totals, a routing criterion that joins predictability to economic weight, probabilistic forecasting for risk-based inventory policies, and validation across firms and sectors.

Author Contributions: Conceptualization, G.M.H.-O. and V.D.-B.; methodology, G.M.H.-O., J.D.R.-B. and V.D.-B.; software, J.D.R.-B. and R.A.-V.; validation, R.A.-V. and M.J.C.; formal analysis, G.M.H.-O., R.A.-V. and M.J.C.; investigation, G.M.H.-O., J.D.R.-B. and V.D.-B.; data curation, G.M.H.-O., J.D.R.-B. and R.A.-V.; writing—original draft preparation, G.M.H.-O., J.D.R.-B. and V.D.-B.; writing—review and editing, R.A.-V. and M.J.C.; supervision, R.A.-V. and M.J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The historical sales dataset used in this study is the property of FERRETODO and cannot be shared publicly. The analysis code that reproduces every table and figure of this study, together with the aggregated weekly series required to run it, is deposited at https://zenodo.org/records/23029677 (accessed on 12 June 2026). Hyperparameters were fixed a priori and are reported in full in Section 3.3; no hyperparameter search was performed, and the software versions and random seed are given in Section 3.

Acknowledgments: The authors thank FERRETODO for facilitating access to the historical sales data and Universidad del Magdalena and Universidad Nacional Abierta y a Distancia (UNAD) for the institutional support provided for the development of this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Silver, E.A.; Pyke, D.F.; Thomas, D.J. *Inventory and Production Management in Supply Chains*, 4th ed.; CRC Press: Boca Raton, FL, USA, 2016. [CrossRef]
2. Zipkin, P.H. *Foundations of Inventory Management*; McGraw-Hill: New York, NY, USA, 2000.
3. Punia, S.; Nikolopoulos, K.; Singh, S.P.; Madaan, J.K.; Litsiou, K. Deep learning with LSTM networks and random forests for demand forecasting in multi-channel retail. *Int. J. Prod. Res.* **2020**, *58*, 4964–4979. [CrossRef]
4. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef]
5. DANE. *Encuesta Anual de Comercio: Principales Resultados 2022*; Departamento Administrativo Nacional de Estadística: Bogotá, Colombia, 2024. Available online: https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-interno/encuesta-anual-de-comercio-eac (accessed on 12 June 2026).
6. DANE. *Muestra Mensual de Comercio al por Menor: Boletín Técnico*; Departamento Administrativo Nacional de Estadística: Bogotá, Colombia, 2026. Available online: https://www.dane.gov.co/files/operaciones/EMC/bol-EMC-feb2026.pdf (accessed on 12 June 2026).
7. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. M5 accuracy competition: Results, findings and conclusions. *Int. J. Forecast.* **2022**, *38*, 1346–1364. [CrossRef]
8. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, Australia, 2018. Available online: https://otexts.com/fpp3/ (accessed on 12 June 2026).
9. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5th ed.; Wiley: Hoboken, NJ, USA, 2015.
10. Arunraj, N.S.; Ahrens, D.; Fernandes, M. Application of SARIMAX model in retail. *Int. J. Oper. Res. Inf. Syst.* **2016**, *7*, 1–21. [CrossRef]
11. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed.; O'Reilly Media: Sebastopol, CA, USA, 2023.
12. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

13. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
14. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
15. Chollet, F. *Deep Learning with Python*, 2nd ed.; Manning: Shelter Island, NY, USA, 2021.
16. Nasseri, M.; Falatouri, T.; Brandtner, P.; Darbanian, F. Machine learning in retail demand prediction. *Appl. Sci.* **2023**, *13*, 11112. [CrossRef]
17. Falatouri, T.; Darbanian, F.; Brandtner, P.; Udokwu, C. Predictive analytics for demand forecasting. *Procedia Comput. Sci.* **2022**, *200*, 993–1003. [CrossRef]
18. Syntetos, A.A.; Boylan, J.E.; Croston, J.D. On the categorization of demand patterns. *J. Oper. Res. Soc.* **2005**, *56*, 495–503. [CrossRef]
19. Gardner, E.S. Exponential smoothing: The state of the art. *J. Forecast.* **1985**, *4*, 1–28. [CrossRef]
20. Croston, J.D. Forecasting and stock control for intermittent demands. *Oper. Res. Q.* **1972**, *23*, 289–303. [CrossRef]
21. Aviv, Y. A time-series framework for supply-chain inventory management. *Oper. Res.* **2003**, *51*, 210–227. [CrossRef]
22. Cachon, G.; Fishe, M. Campbell Soup's continuous replenishment program: Evaluation and enhanced inventory decision rules. *Prod. Oper. Manag.* **1997**, *6*, 266–276. [CrossRef]
23. Fisher, M.; Rajaram, K.; Raman, A. Optimizing inventory replenishment of retail fashion products. *Manuf. Serv. Oper. Manag.* **2001**, *3*, 230–241. [CrossRef]
24. Fisher, M.; Vaidyanathan, R. A demand estimation procedure for retail assortment optimization with results from implementations. *Manag. Sci.* **2014**, *60*, 2401–2415. [CrossRef]
25. Fisher, M.; Raman, A. Innovations in retail operations: Thirty years of lessons from Production and Operations Management. *Prod. Oper. Manag.* **2022**, *31*, 4452–4461. [CrossRef]
26. Lim, Y.F.; Jiu, S.; Ang, M. Integrating anticipative replenishment allocation with reactive fulfillment for online retailing using robust optimization. *Manuf. Serv. Oper. Manag.* **2021**, *23*, 1616–1633. [CrossRef]
27. Park, H.; Choi, D.G.; Min, D. Adaptive inventory replenishment using structured reinforcement learning by exploiting a policy structure. *Int. J. Prod. Econ.* **2023**, *266*, 109029. [CrossRef]
28. Kolyaei, M.; Zhang, L.; Blom, M. Inventory replenishment and fulfilment decisions for an omnichannel retailer: A reinforcement learning-based method. *Int. J. Prod. Res.* **2025**, *63*, 9571–9592. [CrossRef]
29. Rohilla, A.; Kundu, T.; Kapoor, R.; Sheu, J.-B. Enhancing retail inventory replenishment amid product life cycle shifts: A system dynamics approach. *IEEE Trans. Eng. Manag.* **2025**, *72*, 1939–1953. [CrossRef]
30. Scholz-Reiter, B.; Heger, J.; Meinecke, C.; Bergmann, J. Integration of demand forecasts in ABC-XYZ analysis for a better scheduling of stock replenishments. *Int. J. Product. Perform. Manag.* **2012**, *61*, 611–626. [CrossRef]
31. Henao-Baena, C.A.; Zuluaga-Zuluaga, B.; Galeano-Castro, J.; Marín-García, E.J.; Calvo-Salcedo, A.F. Gestión de inventarios con machine learning. *Ingeniería* **2024**, *29*, e19423. [CrossRef]
32. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. [CrossRef] [PubMed]
33. Petropoulos, F.; Apiletti, D.; Assimakopoulos, V.; Babai, M.Z.; Barrow, D.K.; Taieb, S.B.; Bergmeir, C.; Bessa, R.J.; Bijak, J.; Boylan, J.E.; et al. Forecasting: Theory and practice. *Int. J. Forecast.* **2022**, *38*, 705–871. [CrossRef]
34. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.