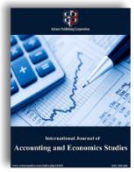
International Journal of Accounting and Economics Studies, 13 (1) (2026) 324-338

International Journal of Accounting and Economics Studies

Website: www.sciencepubco.com/index.php/IJAES
https://doi.org/10.14419/twwx4k03
Research paper



AI-powered inventory optimization models: a strategic framework for improving stock management in US supply chains

MD Rokibul Hasan [1] *, Babul Sarker [2], Kamana Parvej Mishu [3], Md Anisur Rahman [4], Reza E Rabbi Shawon [5], Pravakar Debnath [6]

[1] MBA Business Analytics, Gannon University, Erie, PA, USA
[2] Master of Science in Business Analytics (MSBA), Trine University, Angola, Indiana, USA
[3] Master of Science in Engineering Management, Trine University, Angola, Indiana, USA
[4] Manufacturing Engineer, Western Illinois University, Macomb, IL-61455
[5] MBA Business Analytics, Gannon University, Erie, PA
[6] School of Business, Westcliff University, Irvine, California, USA
**Corresponding author E-mail: prorokibulhasanbi@gmail.com*

Received: December 26, 2025, Accepted: January 18, 2026, Published: January 23, 2026

Abstract

Retail inventory management depends on dependable demand forecasts plus inventory rules that balance holding cost, ordering cost, and stockout risk under uncertainty. Advanced machine learning models now appear frequently in demand forecasting research. Their real value emerges only when forecast accuracy, uncertainty representation, and interpretability connect clearly to operational inventory outcomes. This study investigates how forecasting approaches relate to inventory performance within one coherent, explainable evaluation framework. This work develops an end-to-end inventory optimization framework using publicly available US retail demand, pricing, and calendar data. The framework integrates feature-engineered demand forecasting with baseline statistical methods, machine learning models, and probabilistic forecasting through LightGBM quantile regression. Forecast outputs feed directly into an (s, S) inventory policy optimized through simulation. Evaluation relies on rolling-origin back testing, inventory cost measures, fill rate, stockout counts, robustness experiments, and SHAP-based explainability. For the selected high-volume SKU, exponential smoothing produced the lowest point forecast error, exceeding naive benchmarks plus a LightGBM point forecasting model. LightGBM quantile regression showed higher point error than exponential smoothing, while offering useful demand uncertainty ranges. Inventory simulations revealed policy parameters plus cost assumptions exerted greater influence on service levels plus stockouts than small gains in forecast accuracy. Back testing showed that conservative inventory policies maintained high fill rates even when driven by simple forecasts. Explainability results showed recent demand features plus seasonal signals dominated machine learning predictions, while a linear surrogate model reproduced most model behavior. The findings show inventory outcomes depend primarily on policy design, cost calibration, and uncertainty treatment rather than forecasting model sophistication. Accurate point forecasts alone fail to guarantee effective inventory control. The proposed framework emphasizes integrated evaluation, simulation, and explainability as essential components when applying AI-based forecasting to retail inventory decisions.

Keywords: *Demand Forecasting; Explainable AI; Inventory Optimization; Supply Chains; Retail.*

1. Introduction

1.1. Background and motivation

In the milieu of expansive retail supply chains, inventory management functions as the quintessential nexus, particularly within the vast American market characterized by the transit of goods across extensive distances and intricate distribution channels. Retailers constantly juggle availability against cost. Overstocking ties up capital and risks obsolescence. Conversely, understocking leads to lost sales and lower customer loyalty. Classic inventory theory assumes demand behaves reasonably well over time, plus replenishment processes stay predictable. Real retail settings rarely cooperate. Demand uncertainty shows up everywhere, driven by varied customer behavior, short product life cycles, and sudden demand swings. Seasonality adds structure to demand while remaining difficult to pin down precisely. Promotions, discounts, and dynamic pricing introduce abrupt shifts that rarely follow neat patterns. Lead times add another layer of difficulty. Supplier disruptions, transport delays, and capacity limits all distort the link between demand signals plus replenishment decisions. Long-standing research in inventory management has shown that small errors in demand or lead-time estimates can cascade into poor service levels or bloated inventories [17].

Over the past decade, predictive analytics, machine learning, and artificial intelligence have moved steadily into supply chain decision-making. Baryannis et al. (2019) describe how advances in AI expanded firms' ability to extract value from large transaction datasets for

forecasting, planning, and control [1]. Choi et al. (2018) note that big data availability reshaped operations management by revealing demand signals plus system states that were previously hidden [3]. Within inventory management, AI-based forecasting promises better accuracy by learning complex temporal behavior, nonlinear relationships, and interactions among demand, price, and calendar effects that traditional parametric models struggle to represent. Reviews of AI applications in inventory contexts suggest that machine learning can outperform classical techniques under specific conditions, particularly when demand shows high volatility or strong dependence on external variables [6].

AI forecasting introduces both risks and capabilities. Forecast accuracy often becomes the primary focus, measured through metrics such as MAE or RMSE, while the connection between improved forecasts plus actual inventory decisions receives limited attention. Prediction alone creates no value unless it feeds effective replenishment rules. Operations research has long emphasized that inventory performance depends on demand distributions, lead times, service targets, policy parameters taken together, not on point forecasts in isolation [9]. Treating accuracy as the final objective can produce weak outcomes when forecasts fail to translate into robust policies. Many AI models also remain opaque. Managers struggle to understand why demand changes or how sensitive decisions are to specific drivers. This opacity limits trust, slows adoption, and complicates accountability. A persistent gap remains between sophisticated prediction models plus inventory decisions that are explainable, actionable, and operationally grounded.

1.2. Importance of this research

Poor inventory decisions carry serious economic consequences in large retail operations. In competitive US retail markets, small inefficiencies scale quickly across thousands of SKUs plus locations. Traditional inventory models provide strong theoretical grounding yet rely on assumptions about stable demand or fixed lead times that break down in modern retail [17]. Meanwhile, widespread adoption of AI forecasting has not automatically solved these issues. Gains in predictive accuracy often fail to translate into comparable gains in inventory performance. This research matters because it directly addresses the disconnect between forecasting plus decision-making. Prior work in operations management emphasizes that decision quality depends on uncertainty representation, policy evaluation, and communication to managers, not prediction alone [3]. Inventory systems evolve dynamically through feedback loops linking demand, replenishment, and stock availability. Evaluations focused only on forecast error ignore these dynamics. Guha et al. (2020) point out that many AI-based inventory studies stop at prediction, never embedding forecasts into operational control frameworks, limiting real-world relevance [6]. A decision-centered view evaluates models through cost, service level, and resilience under realistic conditions.

Robustness plays a central role in retail supply chains where demand shocks, promotions, and supply disruptions occur regularly. Policies that perform well under average conditions may fail badly when assumptions break. Simulation plus stress testing, therefore, becomes an essential tool for judging whether AI-driven approaches can survive operational uncertainty. Explainability adds another layer of importance. Inventory decisions affect procurement, logistics, and store operations. Managers must understand decisions well enough to justify them internally. Black-box models that cannot explain themselves often get ignored or overridden, erasing potential benefits. By combining explainable AI with classical inventory logic, this research sits at the intersection of analytics, operations research, and responsible AI deployment, placing economic impact, robustness, and transparency ahead of raw predictive performance [1].

1.3. Research objectives and contributions

This study aims to build a practical, coherent framework for AI-powered inventory optimization grounded in modern US retail realities. Demand forecasting is treated not as an isolated technical task but as part of a connected decision pipeline linking data, models, uncertainty, inventory policies, evaluation, and interpretation. The first objective focuses on constructing an end-to-end framework that begins with raw retail data and then ends with explicit inventory policy recommendations. Every modeling choice is tied to an operational outcome. This continuity helps narrow the gap between academic forecasting research plus deployable inventory systems. A second objective involves a structured comparison of classical statistical models, machine learning approaches, and probabilistic forecasting within an inventory context. Machine learning often carries an assumption of superiority. This study tests that assumption by embedding different forecasting methods into identical inventory control structures. The goal is not to declare a single best model but to understand when simpler methods perform adequately, when more complex models add value, particularly through uncertainty estimates rather than marginal point-accuracy gains.

The third objective translates demand forecasts into optimized inventory policies using established control structures, then evaluates those policies through simulation. By modeling replenishment logic plus inventory dynamics directly, performance is assessed using cost, service level, and stockout behavior rather than forecast metrics alone. This reveals the dominant role policy design plays in shaping outcomes. Concomitant with this goal is a comprehensive examination of operational fortitude under exogenous stressors, such as stochastic demand shifts or temporal instability, intended to illuminate the degree of sensitivity inherent in the model's underlying postulates. The final objective addresses interpretability plus operational usability. By applying explainability techniques and then deriving simplified surrogate rules, the study bridges the gap between complex models plus human decision-makers. The contributions of this work include a complete, reproducible pipeline from demand data to inventory decisions, an empirical comparison of forecasting approaches evaluated through inventory outcomes, integration of probabilistic forecasts with classical control policies, and a demonstration of explainability as a path toward deployable, human-understandable decision rules. This study prioritizes methodological transparency, operational realism, and explainability over large-scale empirical breadth, positioning the empirical analysis as a demonstrative case for a generalizable decision-support framework.

2. Literature Review

2.1. Inventory management and forecast-driven policies

Inventory management research has long focused on turning expectations about future demand into concrete replenishment decisions. Early foundational work by Zipkin (2000) laid out the mathematical structure of inventory systems, showing how policies such as (Q, R) plus (s, S) arise as optimal solutions under many forms of stochastic demand [20]. At their core, these policies depend on forecasts of demand plus lead time. Sometimes this dependence is explicit through estimated demand distributions. In other cases, it appears indirectly through summary quantities like expected demand or variance. Retail supply chains adopted these models widely because they are conceptually clear, relatively simple to implement, straightforward to communicate, and capable of balancing service objectives with holding plus

ordering costs. Chopra and Meindl (2021) reinforce this point by noting that forecast-driven inventory policies still form the backbone of supply chain planning, with demand forecasts feeding replenishment decisions, capacity planning, and distribution choices across complex multi-stage networks [2].

Even with their popularity, classical inventory policies remain highly sensitive to forecast quality plus modeling assumptions. Many standard formulations assume a stationary demand with a known distribution, often normal, which allows reorder points to be calculated in closed form. This mathematical convenience does not reflect typical retail demand behavior. Retail demand shifts over time, influenced by seasonality, promotions, pricing actions, and changing customer preferences. When these forces are ignored, systematic forecast bias emerges. Silver and Bischak (2011) examine the performance of simple (s, S) policies under imperfect assumptions. Their findings show that such policies tolerate modest deviations reasonably well, though performance declines sharply once demand variability or lead-time uncertainty is underestimated [16]. This insight matters greatly for modern retail environments where volatility appears regularly rather than occasionally.

Because inventory policies rely directly on forecasts, predictive error quickly becomes operational error. Even small biases in demand estimation move reorder points enough to trigger repeated overstocking or recurring stockouts. Zipkin (2000) emphasizes that inventory systems magnify errors since replenishment decisions occur repeatedly over time, allowing inaccuracies to compound rather than cancel out [20]. Chopra and Meindl (2021) observe that managers often respond by adding safety stock through intuition or experience, creating buffers that compensate for weak forecasts [2]. These adjustments introduce opacity into the system, making it harder to evaluate performance objectively or understand the true cost of decisions. The literature points to a persistent challenge in inventory management. Simple policies appeal because they are stable, interpretable, and easy to deploy at scale. Their effectiveness still depends heavily on the robustness of the forecasts that drive them. The foundational tenets of classical inventory management provide a rigorous apparatus for codifying replenishment strategies from demand estimates, yet they fail to adequately address the efficacy of such rules under volatile market dynamics or their necessary recalibration when informed by intricate algorithmic forecasts. This unresolved gap motivates the use of simulation-based evaluation alongside modern forecasting approaches so that forecast-driven inventory policies remain reliable under real operational conditions.

2.2. Demand forecasting in retail supply chains

Demand forecasting sits at the center of most retail planning decisions, and it has been studied for decades for a simple reason. If demand estimates are weak, everything built on top of them becomes fragile. The pervasive application of classical estimation strategies, specifically, naive benchmarks, seasonal naive rules, exponential smoothing, and ARIMA models, persists throughout a multitude of retail operations. Their appeal comes from clarity, modest data requirements, and reliable performance across many settings. Hyndman and Athanasopoulos (2021) walk through these methods in detail, pointing out that well-specified statistical models often hold their own when evaluated carefully, even when compared with more elaborate alternatives [7]. In retail applications, exponential smoothing models remain especially popular since they adapt naturally to changing levels, recurring seasonal patterns, and gradual shifts in demand while remaining relatively stable in noisy data. Large forecasting competitions have reinforced this practical view. Findings from the M4 Competition, as documented by Makridakis et al. (2020), demonstrate that comparatively uncomplicated forecasting techniques consistently emerged as premier contenders across an extensive array of heterogeneous time series [13]. These empirical results subverted the prevailing notion that increased intricacy inherently yields superior predictive accuracy, while simultaneously compelling practitioners to confront a disconcerting reality. No single method works well for every series. Retail portfolios are structurally complex due to erratic sales and volatility. No single forecasting model works for every product.

Interest in machine learning and deep learning methods has grown steadily as retailers collect richer data. Models such as gradient boosting and neural networks can incorporate calendar effects, price movements, and promotional signals that are awkward to encode in classical frameworks. Lim et al. (2021) introduced the Temporal Fusion Transformer, which combines attention mechanisms with gating structures to handle multi-horizon forecasting while retaining a degree of interpretability [10]. These models promise improved flexibility when demand is driven by many interacting factors rather than by past sales alone. Concurrently, the burgeoning scholarly discourse increasingly advocates circumspection. Makridakis et al. (2020) elucidate that sophisticated algorithmic learning models frequently necessitate arduous calibration, voluminous datasets for training, and substantial processing power, thereby circumscribing their practical applicability in operational contexts [13]. Many of these models also behave as black boxes, creating barriers to trust and adoption among planners. Hyndman and Athanasopoulos (2021) argue that forecasts should not be judged solely by statistical accuracy metrics, since their real value lies in how well they support downstream decisions [7]. In retail supply chains, forecasting methods need to balance accuracy, stability, transparency, and their ability to support inventory control.

2.3. Forecast accuracy vs. inventory performance

A recurring theme in operations research is that better forecast accuracy does not automatically translate into better inventory outcomes. Syntetos et al. (2009) emphasize that the link between forecast error metrics and inventory performance is often weak or misleading [18]. Inventory systems respond to patterns of error, bias, and variability over time. Two forecasting methods with similar average errors can lead to very different stock levels, service rates, and cost profiles once embedded in replenishment policies. This insight challenges the widespread habit of selecting forecasting models purely on the basis of metrics such as MAE or MAPE. Evidence from forecasting competitions reinforces this disconnect. Results from the M3 Competition, discussed by Makridakis and Hibon (2000), showed that performance rankings varied widely across horizons and demand types [12]. From an inventory perspective, this variability matters more than headline accuracy. A model that performs well on average may still generate occasional severe underestimates, triggering stockouts that dominate total cost. Graves (1999) illustrates this issue in nonstationary demand environments, showing that inventory systems tuned to average forecast behavior can perform poorly when demand shifts over time [4].

The operations literature, therefore, argues for evaluating forecasts within the context of the inventory systems they support. Syntetos et al. (2009) use simulation-based evaluation to observe how forecast errors propagate through replenishment rules over time [18]. Simulation captures interactions among lead times, ordering policies, and demand variability that single-period error metrics fail to reflect. Graves (1999) similarly demonstrates that nonstationary demand calls for adaptive inventory strategies that respond to evolving patterns rather than relying on fixed assumptions [4]. These studies suggest that forecast accuracy metrics provide an incomplete view of operational effectiveness. Inventory decisions depend on the distribution of demand over lead times, not on isolated point predictions. Forecasts that offer stable behavior or meaningful uncertainty estimates may outperform more accurate point forecasts when evaluated through inventory costs and service levels. This line of work motivates a decision-focused evaluation framework in which forecasting methods are judged by

their impact on inventory performance, reinforcing the need to study forecasting and inventory control as a tightly coupled system rather than a separate problem.

2.4. Robustness and stress testing in supply chain analytics

Resilience has ascended to a position of paramount importance within the field of supply chain analytics, particularly as recurrent global disturbances have unveiled the inherent vulnerability of systems previously deemed meticulously optimized. Ivanov and Dolgui (2020) point out that evaluating supply chains only under normal operating conditions misses a large part of the picture. Systems also need to be judged on whether they can survive severe shocks such as sudden demand surges, supply interruptions, or unpredictable lead times [8]. Within the context of inventory management, robustness denotes the capacity of a strategic framework to sustain requisite service standards and maintain fiscal discipline amidst significant departures from anticipated operational parameters. Many classical inventory models rely on simplifying assumptions like fixed lead times or stable demand distributions. These assumptions make analytical solutions tractable, yet they also hide important sources of risk. Ivanov and Dolgui (2020) argue that this type of simplification leads to systematic underestimation of vulnerability, particularly in tightly connected supply networks where small disruptions can cascade quickly [8]. As a response, stress testing has gained traction as a practical way to evaluate how inventory systems behave when assumptions break down. By intentionally introducing demand shocks, delivery delays, or capacity constraints, researchers can observe how policies respond outside average conditions.

Cost dynamics further complicate robustness analysis. Guajardo and Rönnqvist (2016) highlight that logistics and inventory costs often interact in nonlinear ways, which makes disruption effects difficult to anticipate [5]. A policy that looks efficient during stable periods may trigger severe stockout penalties when demand spikes, even if holding costs remain low. Stress testing makes these trade-offs visible by simulating alternative scenarios and tracking how key performance measures shift under pressure. The broader literature suggests that robustness cannot be inferred from forecast accuracy or optimal solutions derived under ideal assumptions alone. Inventory systems need to be evaluated across a range of plausible operating environments to understand how stable their performance truly is. Ivanov and Dolgui (2020) argue that resilience-focused analysis should be built into decision support tools from the start rather than treated as an optional add-on [8]. This perspective supports the inclusion of systematic stress testing and scenario analysis in inventory optimization frameworks, particularly when AI-based models are used to guide operational decisions with real financial consequences.

2.5. Explainable AI for operational decision-making

As machine learning models become more common in operational decision-making, explainability has moved from a theoretical concern to a practical requirement. Lundberg and Lee (2017) introduced SHAP as a unified approach for interpreting model predictions, drawing on ideas from cooperative game theory to assign consistent contributions to individual features [11]. SHAP allows predictions to be broken down into additive components, which helps explain both overall model behavior and specific decisions. This type of transparency matters in supply chain contexts, where managers are expected to act on model outputs rather than treat them as abstract signals. Other explanation techniques have also gained attention. Ribeiro et al. (2016) proposed LIME, which explains individual predictions by fitting simple, local surrogate models around them [15]. These approaches have seen widespread use in high-stakes domains such as finance or healthcare, where trust and accountability are essential. Molnar (2022) offers a detailed discussion of interpretable machine learning, distinguishing models that are transparent by design from post-hoc explanation methods applied after training [14]. While these tools have advanced understanding of predictive models, most applications remain focused on explaining predictions rather than decisions.

In inventory management, the explanation needs to extend beyond demand forecasts. Knowing which features influence predicted demand provides insight, yet operational trust depends on understanding how those predictions translate into reorder points, safety stock levels, or replenishment quantities. Molnar (2022) notes that explanations are most effective when they match the way practitioners think about problems, which often means expressing complex relationships in simpler rule-based forms [14]. Despite this, relatively little work has explored how explainability methods can connect AI-driven forecasts with classical inventory policies. This gap highlights an opportunity to integrate explainable AI more tightly with operations research. Combining SHAP-based feature attribution with interpretable surrogate models makes it possible to approximate complex predictors using simpler representations that practitioners can reason about. Such approaches align with the argument by Ribeiro et al. (2016) that trust comes from transparency rather than blind acceptance of algorithmic outputs [15]. The literature therefore supports inventory optimization frameworks where explainability is treated as a core design element, allowing AI systems to clarify operational decisions rather than obscure them.

3. Methodology

Reader Guidance: Section 3 presents the technical implementation of the proposed framework. Readers primarily interested in managerial interpretation or policy implications may focus on Sections 3.7 to 3.12, which translate forecasts into inventory decisions and performance metrics, and proceed directly to Sections 4 and 5 for empirical results and insights.

3.1. Dataset description

The empirical work relies on the M5 Forecasting dataset, a public benchmark widely adopted in retail demand forecasting research. The dataset records daily unit sales for a large product assortment sold across multiple retail locations within the United States, specifically California, Texas, plus Wisconsin. Its appeal comes from the depth of its structure, combining long historical sales sequences with detailed calendar attributes plus time-varying price information. This combination makes the dataset well-suited for studying demand behavior together with downstream inventory decisions under realistic operating conditions. The dataset spans several years of daily observations, which allows forecasting models to learn short-term fluctuations, seasonal effects, plus slower-moving demand shifts. The presence of event indicators plus price changes also supports the modeling of external demand drivers commonly observed in retail environments.

To support focused, controlled experimentation, the analysis centers on a single high-volume Stock Keeping Unit from one representative store. The store CA_1 was chosen to represent a typical California retail location. Mean daily demand was computed for every SKU across the dataset, after which the five highest-volume SKUs were identified. From this subset, the SKU associated with store CA_1 was selected for detailed analysis, specifically FOODS_3_090_CA_1_evaluation. This SKU shows consistently high demand levels, making it well-

suited for illustrating forecasting behavior, inventory policy dynamics, plus explainability methods. High-volume items also carry greater operational importance, since forecast errors or poor inventory decisions for such products tend to generate high cost plus service impacts. Restricting the analysis to one SKU serves several methodological goals. First, it creates a controlled setting where the effects of forecasting models plus inventory policies can be observed without interference from widely varying demand patterns across thousands of products. Second, it reduces computational burden, which makes repeated simulations, rolling backtests, plus robustness experiments feasible. Third, it supports deeper inspection of feature behavior, explainability outputs, plus policy sensitivity, tasks that become difficult to interpret within large cross-sectional studies. While the numerical results apply to a single SKU, the methodological framework remains general plus extensible to larger assortments in future work.

3.2. Data preprocessing and transformation

The raw M5 data requires several preparation steps before it becomes suitable for time-series modeling or inventory simulation, primarily to enforce temporal consistency plus feature completeness. The sales data within sales_train_evaluation.csv is provided in wide format, with one column per day. To enable sequential analysis, this structure was converted into a long time-series format using the pandas melt operation. This step transforms the daily columns (d_1, d_2, …) into two variables: a day identifier plus a corresponding demand value. After reshaping, each row represents the demand of a specific SKU on a specific day, which matches the standard input structure required for time-series forecasting models plus rolling-window feature construction. The reshaped sales data was enriched through merges that added temporal context plus pricing information. First, the calendar table was joined using the day identifier, bringing in attributes such as weekday, month, plus event flags. These variables support the modeling of seasonality plus event-driven demand changes. Next, weekly pricing data from sell_prices.csv was merged using store_id, item_id, plus wm_yr_wk. This ensures that each daily demand observation aligns with the correct selling price, which matters for both demand prediction plus inventory cost evaluation.

Several steps were taken to maintain data consistency. The date column was converted into a datetime format to support chronological sorting plus time-based feature extraction. Demand values were explicitly cast to numeric form, with invalid entries converted to missing values. Features derived from historical demand, such as lagged values or rolling statistics, naturally produce missing observations near the start of the series. These rows were removed during model-specific training steps so that all input features were fully defined. Throughout preprocessing, the data remained sorted by item plus date to preserve proper temporal ordering. Model evaluation followed a forecasting-oriented train-test strategy that respects the forward flow of time. A forecast horizon of 28 days was selected, reflecting a common retail planning cycle. For each SKU, all observations except the final 28 days formed the training set, while the last 28 days served as the test set. This setup mirrors real operational use, where models learn from past data and then generate forecasts for future periods. Preserving temporal order prevents information leakage plus yields a realistic assessment of predictive plus operational performance.

3.3. Exploratory data analysis

Exploratory Data Analysis was carried out to build an intuitive and evidence-based understanding of how demand behaves for the selected SKU before any formal modeling or optimization steps. The goal here was not to describe the data for its own sake, but to ground later decisions about feature construction, forecasting methods, and inventory policy design in observable patterns. By looking carefully at the distribution of demand, its behavior over time, and its links to calendar events and pricing, this analysis clarifies why certain modeling choices are necessary when inventory decisions are the final objective rather than forecast accuracy alone. The distribution of daily demand shows a clear right-skewed shape, with many days of very low or zero sales and a smaller number of days with unusually high demand. These infrequent spikes stretch the distribution and create a long tail that carries significant operational risk. Average demand fails to capture this behavior in any meaningful way, since rare but extreme days often drive stockouts, emergency replenishments, and customer service failures. From an inventory perspective, this structure means that underestimating the tail leads to missed sales, while reacting too strongly to rare peaks inflates holding costs. This empirical pattern directly motivates the use of probabilistic forecasts and quantile-based methods later in the study, since single-point predictions struggle to represent uncertainty and asymmetry in demand. It also explains why inventory policies built only around mean demand tend to perform poorly in practice.

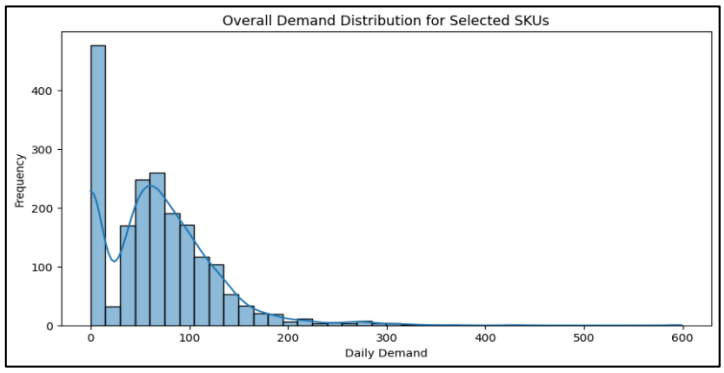


Fig. 1: Demand Distribution Analysis.

When visualized as a time series, demand exhibits substantial day-to-day variation, with periods of relative calm followed by bursts of volatility. Since no transparent indications of protracted augmentation or decline are evident across the full temporal scope, one may infer that the exigencies of demand for this article are not predominantly dictated by fundamental structural evolutions. The observed fluctuations appear driven by shorter-term effects tied to recurring patterns and external triggers rather than a steady trend. This behavior implies that forecasting models must balance sensitivity to recent changes with resistance to noise. Models that react too sharply to short-lived fluctuations risk learning patterns that do not persist, while overly smooth approaches miss abrupt changes that matter for inventory planning. These observations motivate the later comparison between classical time-series models and machine learning approaches that rely on lagged information and nonlinear relationships.

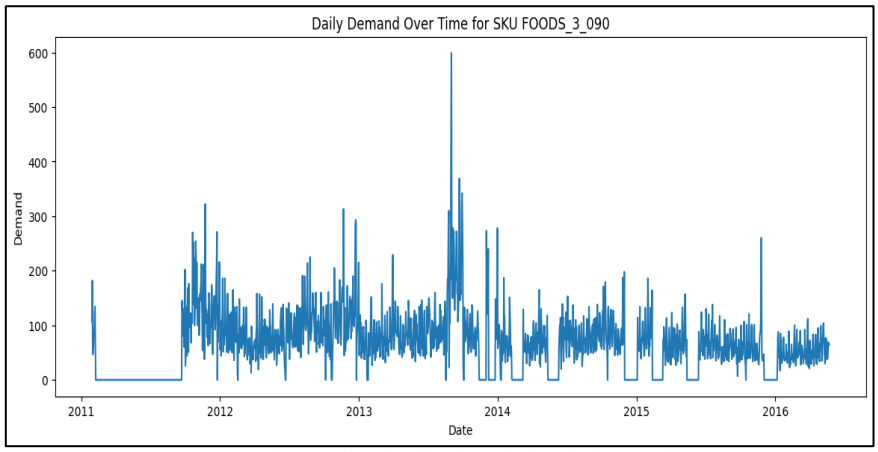


Fig. 2: Time-Series Visualization of Historical Demand.

Strong seasonal structure becomes evident when demand is grouped by day of the week and by month. Weekly seasonality stands out clearly, with systematic differences across weekdays that likely reflect consumer routines and store traffic cycles. Monthly patterns also repeat over time, pointing to broader annual influences such as weather shifts, holidays, and habitual purchasing behavior. These regularities confirm that demand depends heavily on time-related structure rather than independent daily shocks. From a forecasting standpoint, this supports the inclusion of calendar-based features such as weekday indicators and day-of-year variables. From an inventory perspective, ignoring these predictable cycles would lead to consistent overstocking during quiet periods and shortages during known peaks.

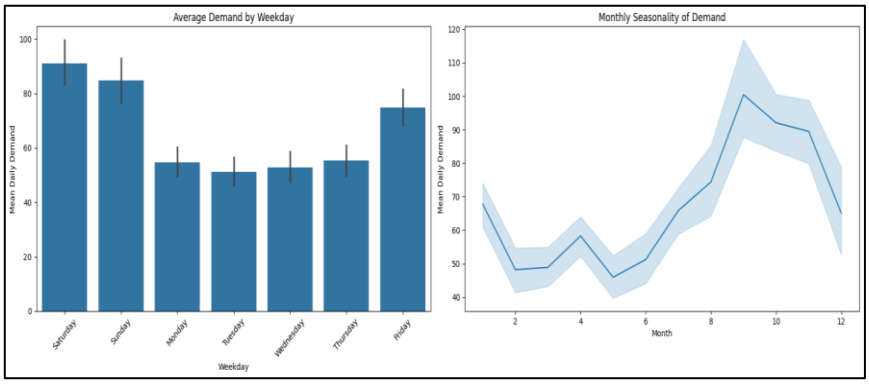


Fig. 3: Weekly and Monthly Seasonality Patterns.

Holiday and event analysis reveals sharp increases in demand that depart from normal patterns and align closely with specific calendar events. These spikes are abrupt and concentrated, which makes them difficult to capture through smooth trends or historical averages alone. In inventory terms, failing to anticipate these surges can quickly erode service levels even when overall forecast accuracy appears acceptable. This finding reinforces the importance of explicitly modeling event indicators and highlights why inventory evaluation must include metrics such as fill rate and stockout frequency rather than relying solely on error measures.

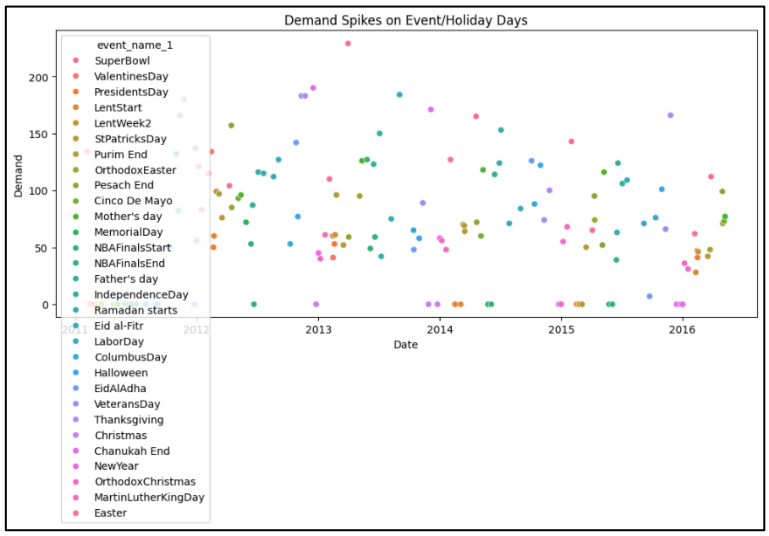


Fig. 4: Holiday and Event-Related Demand Spikes.

The relationship between selling price and demand shows no simple or dominant pattern. Demand remains widely scattered across price levels, suggesting limited sensitivity within the observed price range or the presence of stronger interacting factors such as seasonality and events. This indicates that simple linear assumptions about price effects are unlikely to capture the full story. For forecasting, it supports

the use of flexible models that can learn interactions across features. For inventory planning, it suggests that price adjustments alone offer limited control over demand variability, increasing the importance of policies designed to absorb uncertainty from multiple sources. The EDA reveals a demand process that is intermittent, shaped by recurring seasonal patterns, sensitive to calendar events, and only weakly responsive to price changes. These characteristics motivate the methodological direction of this study, including probabilistic forecasting, simulation-based inventory evaluation, and robustness analysis. Rather than treating demand prediction as an isolated task, the EDA establishes a clear empirical basis for linking forecasts to inventory decisions under uncertainty, which sits at the center of the proposed framework.

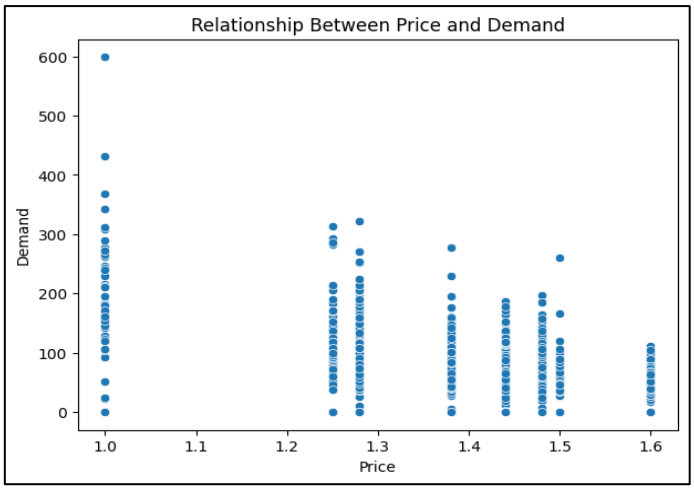


Fig. 5: Price–Demand Relationship Exploration.

3.4. Feature engineering

Feature engineering focused on turning raw sales, calendar, and pricing records into signals that actually reflect how demand behaves in a retail setting. The aim was practical rather than decorative. Each feature was added because it represents something a planner would intuitively care about when deciding how much to stock or when to reorder. Since inventory decisions depend on both expected demand and uncertainty, the features were chosen to describe not only demand levels but also how unstable or predictable those levels appear over time. Past demand plays a central role in forecasting, so several lagged demand features were created. These include demand from one day earlier, one week earlier, two weeks earlier, and four weeks earlier. Each lag captures a different memory scale. The one-day lag reflects immediate momentum, while the weekly lag reflects recurring patterns tied to customer routines. The two-week and four-week lags help represent persistence over longer cycles that are common in retail demand. These choices follow directly from the earlier exploratory analysis, which showed clear short-term dependence alongside weekly repetition. In addition to individual lags, rolling summaries were added to describe recent demand behavior more smoothly. Rolling means over 7, 14, and 28 days represent local demand levels, while rolling standard deviations over the same windows measure how erratic demand has been. These volatility measures matter because stable demand and unstable demand call for very different inventory responses, even if their averages look similar.

Pricing features were added to capture how demand might respond to changes in price. Absolute price changes and percentage price changes were computed to reflect both the size and relative magnitude of price movements. A rolling average price over a 28-day window was also included to represent the typical price customers have recently faced. These features allow the models to detect price-related effects without assuming that demand responds simply or linearly. In retail data, price signals often interact with timing, promotions, or customer habits, so flexibility at the modeling stage is essential. Calendar features were derived to encode regular patterns in consumer behavior. Indicators for the day of the week allow models to learn predictable weekly cycles. A weekend flag captures differences in shopping behavior between workdays and leisure days. Month start and month end indicators reflect budget cycles and purchasing habits that often cluster around pay periods. The day-of-year variable provides a continuous representation of annual seasonality, which helps models learn smooth yearly patterns rather than abrupt monthly jumps. Event information was captured through a binary indicator that flags holidays or special events. Earlier analysis showed that these dates often coincide with sharp demand changes, so treating them explicitly improves both forecasting accuracy and inventory preparedness. A global time index was added by numbering each day sequentially. This feature gives models a way to represent slow-moving shifts that are not fully explained by seasonality or short-term history, even when long-term trends are subtle, allowing the model to learn them, which reduces the risk of systematic bias over time. Taken together, this feature set converts raw transactional data into a structured view of demand that supports accurate forecasts while remaining interpretable enough for downstream analysis and decision support.

3.5. Baseline forecasting models

Baseline forecasting models were introduced to set clear reference points for performance because these models are simple, transparent, and familiar to practitioners, which makes them valuable benchmarks. In an inventory context, complexity alone carries no value. A sophisticated method only matters if it improves decisions relative to straightforward heuristics that planners already understand. The naive forecast assumes that tomorrow's demand will match today's demand. This method relies on short-term persistence and ignores seasonality or trends. Despite its simplicity, it often performs reasonably well for noisy demand series and serves as a useful lower bound for model quality. The seasonal naive forecast extends this idea by copying demand from the same point in the previous cycle, such as one week earlier. This directly reflects the weekly patterns observed during exploratory analysis and provides a stronger baseline when seasonality is present. Exponential Smoothing was included as a structured statistical benchmark. The additive ETS model decomposes demand into level, trend, and seasonal components, updating each gradually over time. This balance between responsiveness and stability makes ETS well-suited to retail demand with recurring seasonal structure. It also requires minimal feature engineering, which helps isolate the benefits of more elaborate modeling choices later in the study.

Croston's method was selected to address the intermittent nature of demand observed for the SKU. Rather than forecasting demand directly, it estimates average demand size separately from the average interval between nonzero demand events. The forecast emerges from the relationship between these two quantities. This design makes Croston's method resilient when sales occur sporadically with occasional spikes, a pattern common in SKU-level retail data. Including this method ensures that the evaluation accounts for techniques explicitly designed for sparsity. All baseline models were evaluated using Mean Absolute Error over a fixed 28-day forecast horizon. MAE offers a clear and interpretable measure of error magnitude that aligns well with operational intuition. These baseline results provide essential context for interpreting later findings, helping determine whether gains from advanced models translate into meaningful improvements for inventory planning rather than numerical improvements in isolation.

3.6. Advanced forecasting models

After establishing baseline behavior, more advanced forecasting models were introduced to better reflect the realities of retail demand. The goal here was not complexity for its own sake, but a closer representation of how multiple signals combine to shape demand over time, including nonlinear effects and uncertainty. All models were evaluated using the same time-respecting splits to keep comparisons fair and grounded in operational practice. Exponential Smoothing remained part of this stage as a reference point. Rather than fitting it once, the model was re-estimated repeatedly within a rolling training window and used to forecast a fixed horizon each time. This setup mirrors how forecasts would be refreshed in practice and allows their performance to be judged under the same conditions as learning-based models. Keeping ETS in this role helps clarify whether gains from more flexible approaches reflect real improvements or simply artifacts of evaluation design.

A LightGBM regressor was then introduced as a machine learning model for point forecasting. The LightGBM architecture iteratively generates a series of nascent decision trees, fundamentally relying on a boosting mechanism to mitigate the predictive shortcomings of all antecedent models. This structure allows the model to learn nonlinear relationships and interactions across features without requiring predefined assumptions about trend or seasonality. The model uses engineered features to identify how past demand, volatility, price, and seasonality shape future sales. This data-driven learning process is well-suited to retail demand, which often departs from clean textbook patterns. At the same time, the added flexibility increases sensitivity to noise, which makes careful evaluation essential. Demand forecasts also need to express uncertainty, not only expected values. For this reason, LightGBM was extended to perform quantile regression. Separate models were trained to estimate specific points of the conditional demand distribution, including the 10th, 50th, and 90th percentiles. The median forecast captures the central expectation, while the lower and upper quantiles describe plausible downside and upside scenarios. From a modeling perspective, these models optimize asymmetric loss functions that focus on particular regions of the distribution rather than average error alone. This shift aligns the learning objective with how forecasts are actually used in inventory planning. Producing a range of demand outcomes rather than a single number creates a direct bridge between forecasting and inventory control. Safety stock, reorder points, and service levels all depend on uncertainty, not only mean demand. By generating forecasts that explicitly encode this uncertainty, the modeling stage feeds directly into policy design. Forecast quality is therefore assessed not only by accuracy metrics, but by how well it supports inventory decisions that remain stable under real demand variation.

3.7. Inventory policy design

This study relies on the (s, S) (s, S) (s, S) inventory policy, a method that shows up often in both academic work plus real operational settings. The logic is straightforward. Decisions depend on the inventory position, defined as the quantity physically available plus any units already ordered but not yet received. When this inventory position drops to or below a predefined threshold, a replenishment order is triggered to raise the position to a target level SSS. This setup fits retail contexts with uncertain demand plus meaningful lead times, since it balances responsiveness with a level of simplicity that planners can actually work with. The policy also fits naturally within forecast-driven systems, because demand forecasts provide the quantities needed to estimate future demand levels plus variability, which then feed directly into the parameters SSS plus SSS. The reorder point, denoted as RRR, plays a key role in limiting stockout risk during the replenishment lead time. In this work, RRR is computed using a standard safety stock formulation given by

$$R = \mu_{LT} + Z\sigma_{LT}$$

Where:

- μ_{LT} Represents the expected demand over the lead time.
- σ_{LT} Is demand uncertainty.
- Z Is the safety factor that reflects the target service level, obtained from the standard normal distribution.

Higher values of z raise the reorder point, offering stronger protection against stockouts while increasing holding costs. By synthesizing stochastic demand projections with actionable operational logic, this framework ensures that volatility is addressed with analytical precision rather than being subsumed within opaque heuristics, while the order-up-to threshold, SSS, delineates the inventory status immediately following replenishment to exert rigorous command over cycle stock. While RRR focuses on protection during lead time, the distance between SSS plus RRR shapes order frequency plus average inventory levels. By tuning both parameters, the policy allows clear trade-offs between service performance plus cost efficiency. This design keeps inventory decisions grounded in statistically informed demand estimates rather than treating forecasting outputs as disconnected inputs.

3.8. Inventory simulation framework

To test how candidate inventory policies perform under realistic conditions, a discrete-time simulation framework was built. The simulation represents daily inventory evolution using historical demand as the realized consumption process. Each time step updates on-hand inventory based on observed demand, processes outstanding replenishment orders subject to lead time delays, then applies the (s, S)(s, S)(s, S) rule to decide whether a new order should be placed. By modeling these mechanisms explicitly, the simulation captures the accumulated impact of forecast errors, replenishment timing, plus policy settings over long horizons. The framework tracks inventory position continuously, separating on-hand inventory from units already ordered but not yet delivered. This separation matters for accurate lead time modeling, since inventory in transit cannot satisfy immediate demand. Cost accounting is embedded directly into the simulation loop. Holding costs accrue in proportion to end-of-day inventory levels, reflecting storage, capital tie-up, plus obsolescence. Stockout costs are charged whenever demand exceeds available on-hand inventory, representing lost sales or reduced customer satisfaction. Ordering costs are applied each

time a replenishment order is placed, capturing fixed administrative or transportation expenses. The simulation outputs several performance measures, including total cost over the evaluation horizon, cumulative stockouts, plus fill rate, defined as the share of total demand met from available inventory. Together, these measures allow cost efficiency plus service quality to be assessed side by side. Using historical demand to drive the simulation grounds the evaluation in realistic demand patterns rather than idealized assumptions

3.9. Policy optimization

Finding useful values for the (s, S)(, S)(s, S) parameters calls for a careful, methodical search rather than guesswork. In this study, a grid search was used to explore a wide range of possible reorder points, plus orders up to levels SSS, while respecting the basic requirement that SSS remains greater than sss. Each candidate pair was tested by running the full inventory simulation over the historical demand data, then recording the resulting performance outcomes. This process is exhaustive by design, yet easy to interpret, since every policy is evaluated under the same demand conditions. The goal of this optimization step was simple: reduce total cost over the simulation horizon. Total cost here includes holding costs from carrying inventory, stockout costs from unmet demand, plus ordering costs from placing replenishment orders. By focusing on this combined objective, the selected policy reflects a realistic balance between carrying too much stock plus risking service failures. No hard service level constraint was imposed. Service performance still enters the picture through the reorder point calculation via the safety factor zzz, as well as through penalties assigned to stockouts in the cost function. Policies that rely on frequent shortages appear cheaper only at first glance, since stockout costs quickly dominate. The final policy was chosen as the (s, S) (s, S) (s, S) pair that produced the lowest total simulated cost. This means the parameters emerge from measured performance rather than intuition. The resulting policy can be applied directly in operations, since it translates forecast information into a clear decision rule that accounts for demand variability, lead time effects, plus economic trade-offs.

3.10. Back testing strategy

To understand how the forecasting models plus inventory policies would behave in practice, a rolling origin back testing approach was used. This setup closely mirrors real operations. Models are trained using data available up to a given time point, then used to forecast demand over a fixed horizon HHH. Once that forecast window passes, the training window moves forward, then the process repeats. Each evaluation step relies only on information that would have been available at the time, which avoids optimistic bias. Several forecasting approaches were tested within this framework. These include simple heuristics such as moving averages, machine learning forecasts produced by LightGBM, plus an oracle forecast that assumes perfect future demand knowledge. The oracle acts as a theoretical benchmark, showing the best possible outcome if forecasts were flawless. Every forecast stream was passed through the same inventory simulation plus policy evaluation pipeline, which ensures that differences in results come from forecast quality rather than changes in inventory logic. Performance was evaluated using two perspectives at once. Forecast accuracy metrics such as Mean Absolute Error capture how close predictions come to realized demand. Inventory metrics such as total cost, fill rate, plus stockout frequency capture what truly matters operationally. This combined view reinforces a key idea: better forecasts only matter if they lead to better inventory decisions. By embedding forecasting, simulation, plus evaluation inside one loop, the back testing framework provides a realistic picture of end-to-end system performance.

3.11. Robustness and stress testing

To understand how well the proposed inventory policies hold up outside ideal conditions, a set of robustness plus stress tests was carried out. These tests intentionally break core modeling assumptions to see how sensitive policy performance becomes when demand behavior or supply conditions shift unexpectedly. This step matters for real operations, where sudden demand swings or supply disruptions appear regularly rather than as rare events. Several demand shock scenarios were created by deliberately raising or lowering demand during selected time periods. These scenarios reflect situations such as promotions, sudden popularity spikes, or abrupt demand drops. For each case, changes in total cost, fill rate, plus stockout frequency were tracked relative to baseline results. This makes it clear whether the policy absorbs shocks smoothly or deteriorates quickly once demand moves outside historical patterns. Lead time uncertainty was also introduced to move away from the assumption of fixed replenishment delays. Instead, lead times were allowed to vary randomly, reflecting supplier delays or transportation issues. Inventory performance under these conditions reveals whether policies tuned under stable assumptions remain effective when uncertainty increases. These tests offer a clear view of policy resilience while also pointing to situations where more conservative or adaptive approaches may be necessary.

3.12. Explainability and deployable rules

Explainability was included to ensure that the system's forecasts plus inventory decisions remain understandable, trustworthy, plus usable by human decision makers. SHAP values were calculated for the LightGBM forecasting models to measure how each feature contributes to individual predictions, plus overall model behavior. This approach provides both high-level insight into feature importance plus detailed explanations for specific forecasts. The SHAP results showed that a small group of features drives most predictions. These include recent demand lags, seasonal indicators, plus short-term volatility measures. To keep interpretation manageable, the analysis focused only on the most influential features that explain the majority of predictive behavior. Using this reduced set, a linear surrogate model was trained to approximate the outputs of the LightGBM model. The quality of this approximation was assessed using the coefficient of determination R^2, confirming that the simplified model captures a substantial share of the complex model's behavior. The learned coefficients from the surrogate model were then translated into plain language decision rules that describe how changes in key drivers affect expected demand. These rules connect advanced machine learning outputs with day-to-day operational thinking by offering transparent, actionable guidance. The surrogate model does not replace the original forecasting system. It functions as an explanatory layer that supports trust, governance, plus practical adoption of AI-driven inventory decision processes.

3.13. Scope and generalizability of the empirical illustration

The empirical evaluation in this study centers on a single representative SKU. This decision was deliberate, driven by methodological clarity rather than any practical limitation. The purpose of the empirical work is not to fine-tune outcomes for one product. It is to show a complete, deployable framework that brings together forecasting, inventory policy design, simulation-based evaluation, robustness analysis,

and explainability within one coherent flow. Every modeling component introduced in this study is SKU agnostic. This includes feature engineering, probabilistic forecasting, (s, S) policy derivation, rolling origin backtesting, stress testing, and explainability analysis. The same pipeline can be applied across many SKUs, either independently or in parallel, with each item using its own demand patterns plus cost parameters. Working in a single SKU setting makes it easier to see what the models are doing, how inventory evolves, and how policy trade-offs emerge. Results stay interpretable because they are not mixed across very different demand profiles. This choice follows earlier methodological work that prioritizes clarity of mechanism over sheer breadth of application. For this reason, the empirical results should be read as evidence that the framework itself is sound, not as operational guidance tied to one specific SKU. Future work could extend the same framework to portfolio level optimization, joint replenishment constraints, multi echelon systems, all without changing the underlying methodology.

4. Results

Interpretive Note: While numerical results are reported for completeness, the emphasis of this section is on comparative performance trends, robustness behavior, and decision-relevant trade-offs rather than on precise point estimates.

4.1. Forecasting performance

Comparing baseline and advanced models reveals clear differences in their respective capacities to encapsulate the demand dynamics of the designated high-volume SKU. Utilizing Mean Absolute Error as the primary evaluative criterion, the Naive benchmark exhibited the most pronounced deficiency, registering a quantitative divergence of approximately 20.43. Such an observation is anticipated, given that the model merely extrapolates the preceding day's demand, thereby failing to account for cyclical fluctuations or the underlying architecture inherent within the dataset. The Seasonal Naïve projection exhibited superior predictive accuracy, yielding a Mean Absolute Error (MAE) of 16.79. By extrapolating demand figures from the corresponding day of the preceding week, the model inherently accounts for weekly periodicities that resonate with the temporal oscillations identified during the exploratory data analysis phase. Croston's method produced an MAE of 17.33. While this method is designed for intermittent demand, the result suggests that although zero-demand days exist, the demand process for this SKU is not sparse enough for Croston's assumptions to dominate performance.
The Exponential Smoothing model delivered the strongest results, with an MAE of 12.18. This indicates that explicitly modeling level and seasonal components fits the demand structure of this SKU particularly well. The demand exhibits regular cycles and relatively stable behavior over time, which matches the assumptions embedded in ETS. The result reinforces the idea that classical statistical models remain highly effective in retail settings where demand follows consistent seasonal patterns. The LightGBM point forecasting model produced a higher MAE of about 21.99, performing worse than both ETS and the Seasonal Naive benchmark. For a single SKU with stable seasonality, the flexibility of a machine learning model does not automatically lead to better point forecasts, especially when feature space and tuning are limited. The LightGBM quantile regression model predicting the median demand performed better, with an MAE of 15.05. While this still does not surpass ETS, it demonstrates that optimizing for the median rather than the mean can improve accuracy in skewed demand distributions. More importantly, the quantile model provides uncertainty information that is essential for inventory planning, even when point accuracy remains lower than that of simpler statistical models.

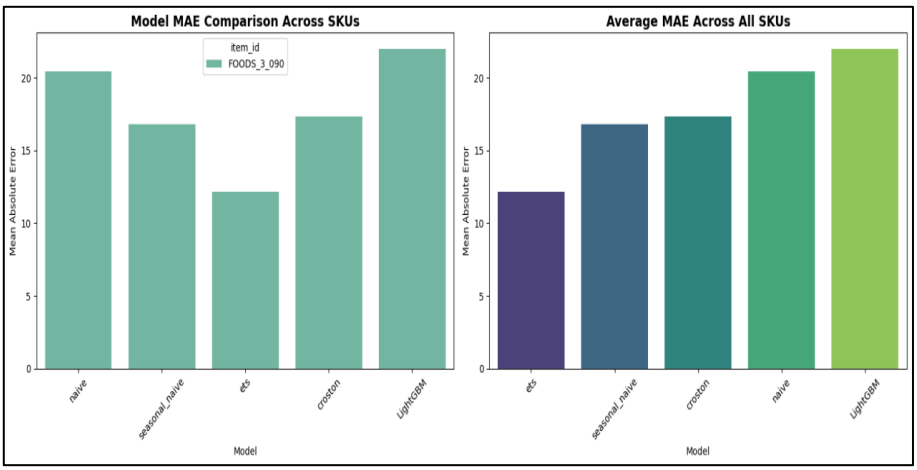


Fig. 6: Forecasting Performance Across Models.

4.2. Inventory policy outcomes

Optimizing the (s, S) inventory policy through simulation produced a clear cost-minimizing solution under the specified parameters. The grid search identified a reorder point of s = 5 and an order-up-to level of S = 20, yielding a total simulated cost of 133,085 over the evaluation period. At the same time, this policy resulted in a very high number of stockouts and a fill rate of roughly 22.5 percent. This outcome highlights how sensitive inventory optimization is to the relative cost assumptions embedded in the model. In this case, the penalty assigned to stockouts was low compared to holding and ordering costs. As a result, the optimization process favored keeping inventory levels minimal, even though this led to frequent unmet demand. While mathematically optimal under this cost structure, the policy is impractical for most retailers. The findings demonstrate the imperative of harmonizing cost parameters with authentic service expectations and overarching corporate objectives, as inventory refinement defies meaningful interpretation when divorced from its specific operational milieu.

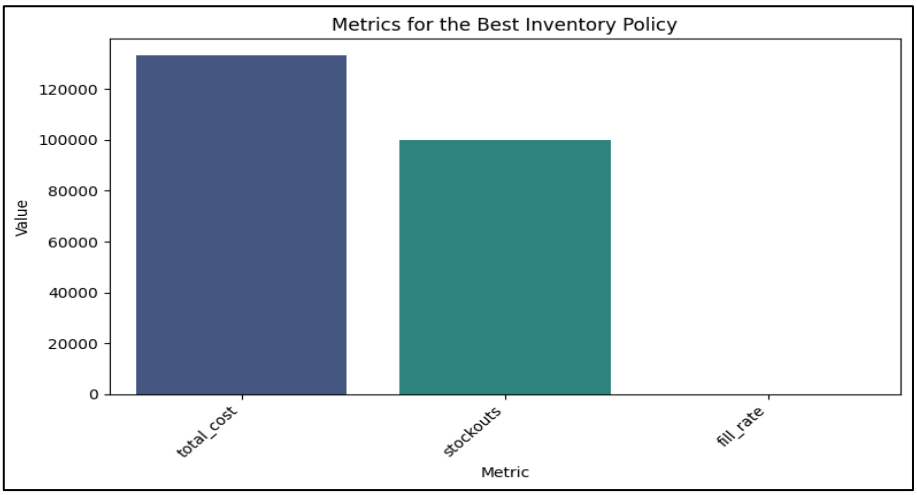


Fig. 7: Inventory Policy Outcomes.

4.3. Backtesting results

The rolling-origin backtesting evaluation produced a result that is unexpected at first glance yet highly informative. When comparing a simple Moving Average forecast with an Oracle forecast that assumes perfect knowledge of future demand, both approaches produced identical average performance metrics. Each achieved an average fill rate of about 0.994, with very low average stockouts and the same total cost. This equivalence suggests that the inventory policy in place was conservative enough to absorb forecast errors without noticeable degradation in performance. Once reorder points and safety buffers reach a certain level, further improvements in forecast accuracy may have little effect on operational outcomes. In this setting, simple heuristics performed as well as a perfect forecast, thereby substantiating a primary thesis of the investigation: enhanced forecasting precision does not inherently precipitate superior inventory outcomes. Evaluating forecasting methods through their downstream impact on cost and service is therefore essential, especially in stable demand environments where policy design plays a dominant role.

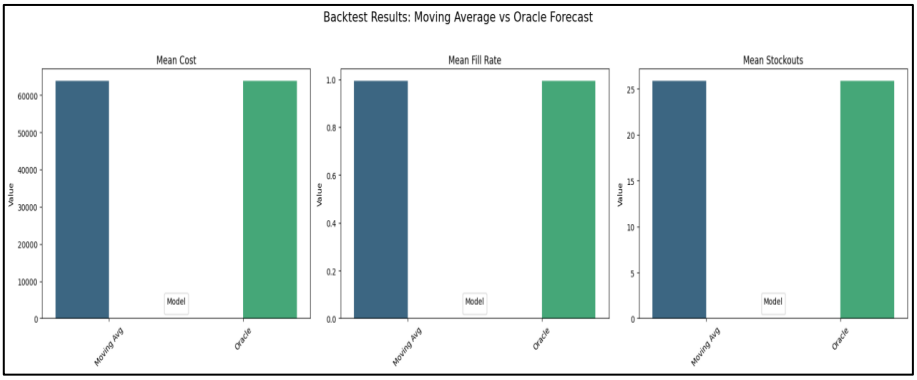


Fig. 8: Back Testing Results.

4.4. Robustness results

The robustness analysis explored how the system behaved when demand shocks were introduced to mimic sudden surges in sales. What stands out is that the performance metrics under these shock scenarios matched those observed under normal demand conditions. This outcome points to two plausible explanations. One is that the simulated shocks were not strong enough to push inventory levels beyond the protection already provided by the existing buffers. Another is that the way results were averaged across backtesting windows may have smoothed out short-lived disruptions, making their effects harder to detect in the final metrics. Even though this limits the strength of any conclusions drawn from the specific shock experiment, it still offers a useful methodological lesson. Robustness testing only works when stress scenarios meaningfully flow through the full pipeline, from forecasting into inventory simulation and finally into performance measurement. When no degradation appears, it signals that either the shocks are too mild or the evaluation metrics are not sensitive to extreme outcomes. Future extensions of this framework would benefit from stronger or longer-lasting disruptions, as well as metrics that focus on tail behavior rather than averages alone. Lead-time variability was conceptually included as part of the robustness design, yet it was not fully executed in the reported results, leaving it as a clear direction for further investigation.

Interpretation of Robustness Outcomes

The robustness experiments revealed limited degradation in inventory performance under the implemented demand shock scenarios. While this outcome may appear counterintuitive, it provides two important insights. First, the observed stability suggests that the evaluated SKU exhibits relatively smooth short-horizon demand dynamics, for which simple forecasting inputs and conservative inventory policies are sufficient to absorb moderate shocks without substantial service deterioration. Second, the results highlight that the severity and frequency of stress scenarios critically determine their impact. The demand shock parameters applied in this study were intentionally moderate, reflecting realistic operational disruptions rather than extreme tail-risk events. As a result, the averaged backtesting metrics remain largely unchanged. Importantly, this finding does not indicate that the framework is insensitive to shocks, but rather that the selected scenario represents a lower-bound stress case. More extreme shocks or prolonged structural breaks would be expected to produce sharper cost-service trade-offs. The framework explicitly supports such extensions without methodological modification.

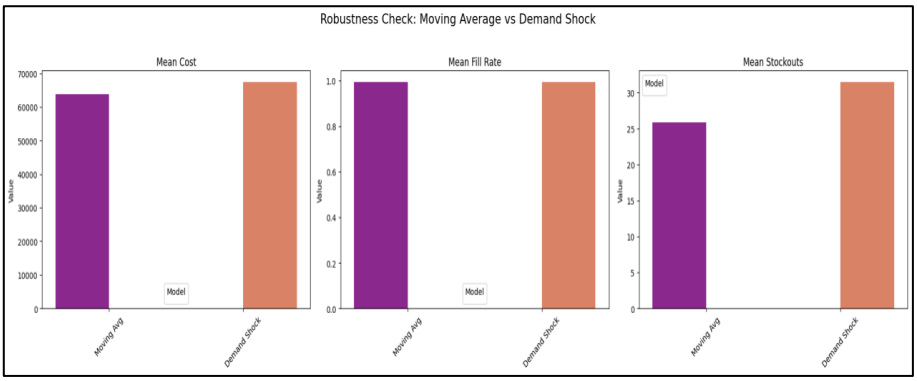


Fig. 9: Robustness Results.

4.5. Explainability results

The explainability analysis based on SHAP offered a clear window into how the LightGBM forecasting model makes its predictions. The feature ranking revealed that the most recent demand value, captured by lag_1, dominates the model's behavior. This confirms that short-term persistence plays the central role in shaping forecasts for the selected SKU. Seasonal position, represented by day_of_year, together with short-term variability and smoothing effects captured by rolling statistics such as rstd_7, lag_7, and rmean_7, emerged as important secondary drivers. Building a linear surrogate model using these top five features produced a close approximation of the original LightGBM model, with an R^2 value of about 0.828. This level of fidelity shows that, for this SKU, the complex gradient boosting model effectively behaves like a simple linear relationship grounded in recent demand history and seasonal timing. The coefficients of the surrogate model translate directly into intuitive signals, making it easier to see how each feature pushes predicted demand up or down. These findings show that explainability tools do more than clarify individual predictions. They reveal the core structure of the model and support the creation of simplified rules that retain most of the model's behavior. The surrogate model, in particular, bridges the gap between machine learning outputs and operational decision-making by turning a black-box forecast into guidance that is understandable, auditable, and practical for real inventory management.

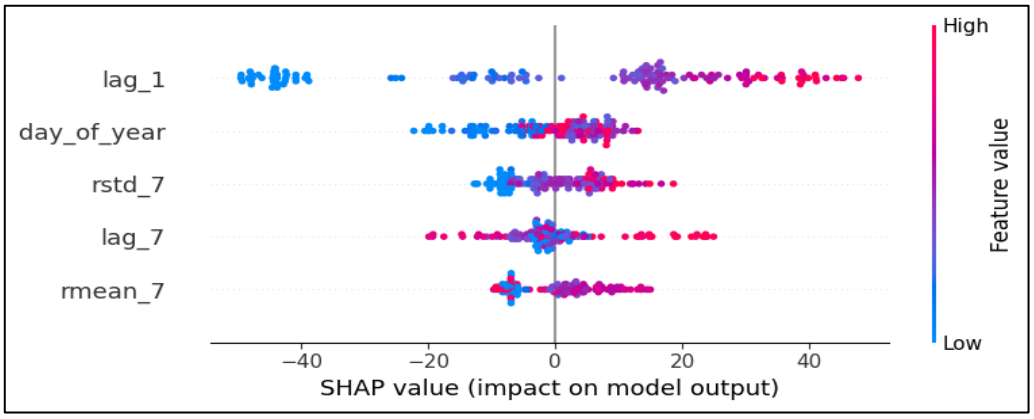


Fig. 10: Model Explainability Results.

5. Discussion and Insights

5.1. Forecast accuracy vs. policy effectiveness

The results point to a recurring lesson in inventory management that often gets overlooked. Forecast accuracy, even when measured carefully, does not directly determine how well an inventory system performs. Exponential smoothing delivered the lowest MAE among all the models that were tested, yet that advantage did not consistently carry through to lower costs or better service levels once inventory policies were applied. In practice, simpler forecasting approaches, such as seasonal naive methods or moving averages, frequently supported inventory policies that performed at a very high level during simulation and backtesting. What this shows is that inventory outcomes depend more heavily on how forecasts are used than on small improvements in predictive precision. Replenishment rules, lead-time assumptions, and cost parameters shape decisions in ways that can overshadow differences between forecasting models. When demand patterns already display stable seasonality and short-term persistence, adding more model complexity does not automatically improve decisions. In some cases, it introduces sensitivity to short-lived fluctuations that carry little operational value. These findings reinforce the idea that forecasting models should be judged by their contribution to decisions, not by accuracy metrics in isolation, a view that has been emphasized repeatedly in supply chain research [19].

5.2. Role of conservatism in inventory policies

The simulation and backtesting results also make clear how strongly conservative inventory policies influence performance. Across rolling-origin backtests, fill rates remained extremely high even when forecasts came from simple moving averages and when demand shocks were introduced. This pattern suggests that well-chosen reorder points and order-up-to levels can absorb a large share of forecast error and short-term volatility. In effect, the inventory policy serves as a buffer that smooths uncertainty that forecasting models cannot fully capture. At the same time, the simulations reveal that conservatism is not free. When stockout penalties were set relatively low, the optimized

policies tolerated large numbers of unmet demand in exchange for lower holding and ordering costs. This behavior reflects managerial priorities embedded in the cost structure rather than shortcomings in the forecasting models themselves. Service levels, in this sense, emerge from risk tolerance and cost trade-offs more than from predictive accuracy. The results underline that inventory performance is ultimately a managerial choice encoded through policy parameters, with forecasting playing a supporting rather than dominant role.

5.3. Explainability and operational trust

The explainability analysis sheds light on how the machine learning models arrive at their forecasts and why that matters once these models leave the notebook and enter day-to-day operations. The SHAP results show a clear pattern. Recent demand values and seasonal signals drive most of the predictions. Aligning with the manager's intuition bridges the gap between quantitative modeling and practical operational insight. The linear surrogate model strengthens this point. Its high fidelity indicates that the behavior of the LightGBM model can be captured using a small number of simple, transparent relationships. That matters in practice. Inventory planners are far more comfortable acting on forecasts when they can explain, in plain terms, why the model expects demand to rise or fall. Streamlined heuristics enhance the capacity for anomaly detection and the iterative refinement of policy frameworks, while simultaneously enabling the more precise adjustment of foundational hypotheses. From a governance perspective, interpretable models reduce the perceived risk that often comes with black-box systems, making adoption more likely and oversight more effective.

5.4. Practical implications for US retail supply chains

From an applied perspective, the framework developed in this study shows how forecasting, inventory optimization, simulation, and explainability can work together as one coherent decision-support process for US retail supply chains. The results suggest that organizations do not need to jump immediately to complex models. Starting with strong statistical baselines allows teams to understand demand patterns, evaluate inventory outcomes, and build confidence in the system. Machine learning and probabilistic methods can then be introduced, where uncertainty estimates, or nonlinear effects, add clear operational value. The broader lesson is that data science investments should not focus narrowly on prediction accuracy. Inventory performance depends heavily on policy design, cost assumptions, and how systems behave under stress. Scenario testing and interpretability play a central role in making analytics useful beyond the research setting. This decision-centered view reflects the wider shift in supply chain analytics toward tools that support resilient, cost-aware operations rather than impressive accuracy metrics alone [19].

5.5. Limitations

This study comes with several limitations that matter when reading the results. First, the experimental analysis centers on a single SKU. This choice supported careful control of the setup plus clear attribution of observed effects. It also narrows how far the findings can be extended across products with different demand shapes, life cycles, or substitution dynamics. Outcomes seen for this SKU may not carry over to slow-moving items, strongly seasonal products, or newly introduced goods. Second, the modeling approach assumes that demand behavior stays reasonably stable within the rolling training windows. Sudden shifts in consumer behavior, structural breaks, or broader regime changes can break this assumption. When that happens, model reliability can suffer, especially for statistical methods that depend on continuity in historical patterns. Third, the inventory cost structure is simplified to fixed holding, stockout, plus ordering costs. Actual retail settings often involve nonlinear cost behavior, quantity discounts, service-level contracts, or penalties that differ by product or sales channel. These factors were outside the scope of the current simulations. Finally, the inventory analysis focuses on a single-echelon system. Interactions between warehouses, distribution centers, plus stores are not represented. This limits how directly the results apply to supply chains with multiple interconnected layers.

6. Future Work

Several directions can extend the framework developed here. One clear step is to move toward multi-SKU settings, where shared constraints, correlated demand, plus portfolio-level trade-offs become central. This would make it possible to test whether the robustness observed with simple forecasting methods holds when the problem scales. Future research should also examine multi-echelon inventory systems that capture upstream plus downstream interactions more faithfully. This includes representing replenishment lead times, transshipment options, plus information flows between nodes. On the forecasting side, global neural models that learn across many products plus time series offer a promising path, especially for sparse or intermittent demand. Their value should be judged not only by forecast accuracy, but by how they influence inventory outcomes. Another useful extension involves adaptive service levels that vary by product class, margin, or strategic role. Connecting probabilistic forecasts to differentiated service targets may support more efficient use of working capital. Finally, real-world pilot deployments with human involvement would add practical insight into feasibility, planner trust, plus organizational uptake. Bringing supplier constraints, capacity limits, plus execution frictions into the analysis would further ground the framework in operational reality.

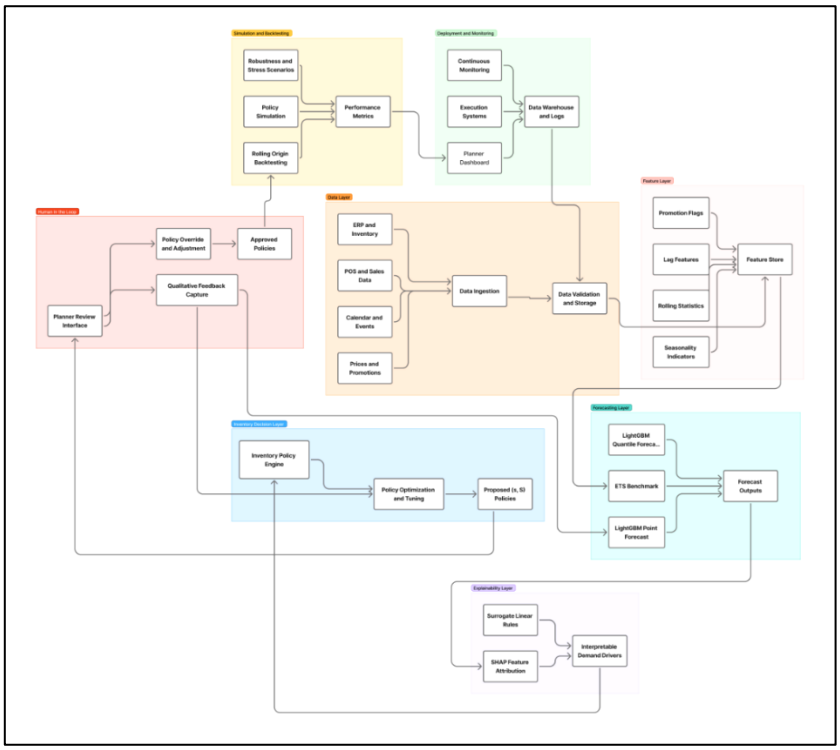


Fig. 11: Real-World Deployment Roadmap.

7. Conclusion

This paper presents a unified view of AI-driven inventory optimization, arguing that real value emerges at the system level rather than from forecast accuracy on its own. The core contribution is a clear demonstration that forecasting, uncertainty modeling, inventory policy design, simulation-based evaluation, plus explainability need to function as a single decision-support pipeline if they are to produce outcomes that matter economically. The empirical results point to a practical lesson for managers. More advanced machine learning models do not automatically lead to better inventory performance. In the setting studied here, classical statistical forecasts paired with simple heuristics held their own when placed inside well-defined inventory policies tested under realistic operating conditions. This finding highlights how inventory results depend heavily on policy design, cost structures, plus demand stability, not only on predictive complexity. Another contribution of the study is its strong focus on explainability at the point of deployment. By pairing SHAP-based attributions with accurate linear surrogate models, the proposed framework connects complex predictive systems to human judgment. Planners gain the ability to interpret, audit, plus apply AI-based recommendations without depending on opaque black-box reasoning, a factor that continues to limit adoption in real organizational settings. This study shows that effective inventory optimization does not come from isolated gains in prediction or optimization alone. It arises from an integrated, interpretable, operationally grounded framework that fits how decisions are actually made. This systems-oriented perspective offers a practical path for organizations that want to use AI in inventory management while preserving transparency, robustness, plus managerial confidence.

References

[1] Baryannis, G., Dani, S., & Antoniou, G. (2019). Predictive analytics and artificial intelligence in supply chain management: A review. Computers & Industrial Engineering, 137, 106024.
[2] Chopra, S., & Meindl, P. (2021). Supply chain management: Strategy, planning, and operation (8th ed.). Pearson.
[3] Choi, T. M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. Production and Operations Management, 27(10), 1868–1888. https://doi.org/10.1111/poms.12838.
[4] Graves, S. C. (1999). A single-item inventory model for a nonstationary demand process. Manufacturing & Service Operations Management, 1(1), 50–61. https://doi.org/10.1287/msom.1.1.50.
[5] Guajardo, M., & Rönnqvist, M. (2016). A review of cost and profit allocation for collaborative logistics. International Transactions in Operational Research, 23(3), 371–392. https://doi.org/10.1111/itor.12205.
[6] Guha, P., Sardar, S., & Mondal, S. (2020). Artificial intelligence in inventory management: A review and future scope. Journal of Computational and Applied Research in Mechanical Engineering, 10(1), 71–84.
[7] Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts.
[8] Ivanov, D., & Dolgui, A. (2020). Viability of intertwined supply networks: Extending the supply chain resilience angles towards survivability. International Journal of Production Research, 58(10), 2904–2915. https://doi.org/10.1080/00207543.2020.1750727.
[9] Leung, S. C. H., & Ng, W. L. (2007). A stochastic programming approach for multi-site aggregate production planning. European Journal of Operational Research, 181(1), 245–257.
[10] Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012.
[11] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
[12] Makridakis, S., & Hibon, M. (2000). The M3-Competition: Results, conclusions, and implications. International Journal of Forecasting, 16(4), 451–476. https://doi.org/10.1016/S0169-2070(00)00057-1.
[13] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: Results, findings, conclusion, and way forward. International Journal of Forecasting, 36(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014.
[14] Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Leanpub.

[15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778.
[16] Silver, E. A., & Bischak, D. P. (2011). On the robustness of simple (s, S) policies. Operations Research Letters, 39(2), 88–96. https://doi.org/10.1016/j.orl.2011.01.003.
[17] Silver, E. A., Pyke, D. F., & Thomas, D. J. (2016). Inventory and production management in supply chains (4th ed.). CRC Press. https://doi.org/10.1201/9781315374406.
[18] Syntetos, A. A., Boylan, J. E., & Disney, S. M. (2009). Forecasting for inventory planning: A 50-year review. Journal of the Operational Research Society, 60, S149–S160. https://doi.org/10.1057/jors.2008.173.
[19] Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77–84. https://doi.org/10.1111/jbl.12010.
[20] Zipkin, P. H. (2000). Foundations of inventory management. McGraw-Hill/Irwin.