forecasting

MDPI

Article

Forecasting Intermittent Sales in Fashion Retail: A Two-Stage Machine Learning Approach

Betül Yılmaz Sucuoğlu [1,*], Ömer Faruk Beyca [1] and Fuat Kosanoğlu [2]

[1] Department of Industrial Engineering, Istanbul Technical University, Istanbul 34467, Türkiye; beyca@itu.edu.tr
[2] Department of Management Engineering, Istanbul Technical University, Istanbul 34467, Türkiye; kosanoglu@itu.edu.tr
* Correspondence: yilmazb18@itu.edu.tr

Highlights

What are the main findings?

- Two-stage machine learning models, particularly Two-Stage XGBoost, achieve superior forecasting accuracy (lowest WRMSSE) for intermittent fashion retail sales, outperforming most one-stage ML configurations, classical benchmarks (Croston, SBA), and deep learning baselines (LSTM, TFT).
- Decomposing the forecasting task into demand occurrence classification and conditional demand magnitude estimation effectively reduces overestimation during zero-demand periods in sparse data.

What are the implications of the main findings?

- Retailers can achieve more reliable inventory planning and stock risk mitigation in volatile fashion markets by adopting structured two-stage machine learning frameworks.
- The study highlights the critical role of rich exogenous feature engineering and contextual factors (e.g., local climate, cultural events) for effective demand forecasting in unique, under-researched markets like Iraq.

Abstract

Intermittent sales patterns, prevalent in fast-fashion retail, pose a critical challenge for conventional forecasting methods. This study empirically compares one-stage and two-stage machine learning (ML) frameworks with classical benchmarks (Croston, SBA). The two-stage approach uses a Random Forest classifier for demand occurrence, followed by regression models (RF, GBM, XGBoost, LightGBM) for magnitude. Models are evaluated using weekly sales data from an Iraqi fashion retailer, incorporating rich exogenous features like product attributes, pricing, weather, and special events across 64 unique attribute-defined product group time series. Performance is assessed via a fixed 13-week holdout and rolling-origin cross-validation, with LSTM and Temporal Fusion Transformer (TFT) serving as deep learning benchmarks. Empirical findings show that machine learning configurations achieve superior WRMSSE accuracy, with two-stage models often outperforming one-stage counterparts, and both significantly surpassing classical and deep learning baselines. The Two-Stage XGBoost yielded the lowest WRMSSE, establishing the feature-engineered two-stage framework as the strongest overall for this intermittent retail setting. Furthermore, a detailed SHAP analysis elucidated the distinct feature contributions to demand occurrence versus demand magnitude, providing actionable insights for inventory management. This rigorous benchmarking analysis offers practical implications

Academic Editor: Erol Egrioglu

Received: 10 April 2026
Revised: 6 June 2026
Accepted: 24 June 2026
Published: 30 June 2026

for inventory planning and demand management in volatile markets, highlighting the effectiveness of explicit demand occurrence modeling.

Keywords: intermittent demand forecasting; two-stage machine learning; retail forecasting; fast fashion; random forest; gradient boosting models

1. Introduction

Sales forecasting remains a fundamental task in retail operations, with significant implications for inventory management, customer satisfaction, and overall profitability. As retail environments become increasingly dynamic, forecasting accuracy is critical to aligning supply with volatile demand patterns [1]. Retailers must continuously balance inventory availability with logistics and storage costs, a challenge compounded by frequent changes in consumer behavior and market conditions [2–4].

Inaccurate demand forecasting often leads to substantial financial consequences, including stockouts and overstock situations, which in turn erode profit margins [5,6]. These challenges are particularly acute in the fashion retail sector, where demand is characterized by high volatility, short product life cycles, and rapid shifts in consumer preferences [7,8]. The complexity of forecasting is further intensified by a range of external factors, including seasonal sales fluctuations, weather sensitivity, the breadth of product assortments, and the obsolescence of designs [9]. Moreover, recent work by Turatti et al. [10] emphasizes the importance of accurately modeling volatility in forecasting systems—an aspect that is especially pertinent in fashion retail, where demand is inherently uncertain and prone to frequent fluctuations.

Beyond product- and market-specific complexities, demand in the retail sector is further influenced by temporal and behavioral factors, including calendar effects (e.g., holidays, promotional campaigns), localized consumer behavior, and dynamic pricing strategies [11,12]. These elements introduce additional variability, necessitating the identification and integration of a diverse set of demand drivers to improve forecast accuracy [13,14].

Intermittent demand represents a particularly challenging class of forecasting problems, characterized by long sequences of zero demand interspersed with irregular and often volatile positive observations. Traditional forecasting models, which implicitly assume continuous demand patterns, tend to perform poorly under such conditions, leading to biased estimates and systematic over- or under-stocking decisions. The importance of explicitly addressing intermittency has been well established in the forecasting literature, particularly in inventory-driven contexts where demand occurrence and demand magnitude follow distinct stochastic processes. Consequently, intermittent demand forecasting has emerged as a specialized research stream with dedicated methodological treatments and evaluation practices. Established methods such as Croston-type estimators and aggregation–disaggregation frameworks have been widely adopted to address such settings; however, their comparative performance relative to modern machine learning approaches remains context-dependent, particularly in fashion retail environments characterized by rich exogenous information.

In response to these challenges, granular data collection at the individual retail location level gains prominence. By incorporating localized market dynamics, historical sales patterns, product characteristics, and temporal indicators from individual stores, retailers can gather rich information. While such detailed data often enables precise sales predictions at the individual store–product level, enhancing forecasting granularity and relevance [15,16], our study focuses on forecasting at the product group level, aggregating

sales across all stores to leverage collective sales patterns and mitigate extreme intermittency. This approach facilitates more informed inventory and replenishment decisions, reduces the risks of stock imbalances, and improves service levels—contributing to operational efficiency and long-term business sustainability [3,17,18]. This study investigates product group-level demand forecasting in a multinational fashion retail setting, focusing on the empirical evaluation of alternative modeling strategies under highly intermittent and volatile demand conditions. Specifically, we forecast weekly sales for 64 unique time series, each representing a distinct product group defined by core product attributes (e.g., Color Group, NOS, Fit, CollarType, Thickness) with sales aggregated across all stores, rather than individual SKUs or individual store–SKU combinations. Rather than proposing a new forecasting framework, the study builds on established modeling paradigms and examines their empirical performance in a realistic operational context. Using historical sales data from previous seasons, the objective is to forecast future sales quantities for specific product categories by accounting for product-level characteristics, special event effects, and meteorological data—variables that are often difficult to operationalize jointly at product level in real-world retail settings. The study evaluates a two-stage modeling strategy—where demand occurrence is first classified and demand magnitude is subsequently estimated—as a pragmatic approach commonly discussed in the intermittent-demand literature. Its performance is empirically compared against one-stage machine learning models under identical data splits and evaluation protocols. To the best of our knowledge, this study represents one of the first empirical comparisons of one-stage and two-stage machine learning approaches specifically for intermittent demand forecasting in fashion retail environments.

While recent studies (e.g., [10]) have explored two-stage modeling strategies in multichannel retail settings, this study focuses on empirically assessing the performance of such approaches in a highly intermittent and volatile fashion retail environment. Our work, while application-driven, aims to empirically test a key theoretical proposition in intermittent demand forecasting: that explicitly modeling the distinct stochastic processes of demand occurrence and demand magnitude, particularly when leveraging rich exogenous features via machine learning, offers a superior approach compared to methods that implicitly or less flexibly handle these components. Building on existing decoupled forecasting frameworks in the literature, the study examines whether a pragmatic classification–regression decomposition yields measurable performance differences relative to single-stage models when applied at the product level with rich exogenous information, such as granular price categories and weather anomaly indicators.

The benchmark is further extended by incorporating two modern deep learning baselines; namely, LSTM and Temporal Fusion Transformer (TFT). These architectures are included not as proposed methods, but as contemporary benchmark models that allow us to assess whether a feature-engineered two-stage machine learning framework remains competitive relative to more architecture-driven sequence models. This additional comparison is particularly relevant in intermittent retail demand settings, where sparse targets, rich exogenous drivers, and operational interpretability requirements coexist. Understanding the specific features driving demand at each stage of a two-stage model is crucial for transforming predictive accuracy into actionable business strategies.

This study is particularly distinctive due to its focus on the Iraqi fashion retail market, a region often underrepresented in the global forecasting literature. Unlike the more stable and saturated markets typically studied in Western or East Asian contexts, the Iraqi market presents a unique blend of rapidly evolving consumer preferences, specific cultural and religious calendars, and a climate characterized by extreme weather conditions. These localized dynamics, coupled with a developing retail infrastructure, introduce distinct challenges and opportunities for demand forecasting that are not fully captured by studies

primarily focused on established markets. By investigating this specific environment, our research provides valuable insights into the transferability and robustness of advanced forecasting methodologies under conditions of heightened volatility and unique contextual drivers.

The contribution of this work is therefore contextual and empirical, offering systematic model comparisons using a transparent rolling-origin evaluation protocol and deriving managerial insights that may support retail inventory and demand planning decisions in volatile fashion contexts. In addition to classical intermittent-demand baselines and one-stage/two-stage machine learning models, the benchmark includes modern deep learning comparators, thereby strengthening the study's empirical positioning relative to current forecasting practice. Model performance is evaluated using WRMSSE (Weighted Root Mean Squared Scaled Error), the standard metric for intermittent demand forecasting established in forecasting literature], alongside supporting metrics to ensure comprehensive assessment.

The remainder of this article is structured as follows. Section 2 reviews the relevant literature on demand forecasting, intermittent demand, and machine learning applications in retail. Section 3 presents the materials and methods, including the data structure, feature construction, forecasting framework, and benchmark models. Section 4 reports the empirical results. Section 5 discusses the findings and their implications. Finally, Section 6 concludes the study and outlines directions for future research.

2. Literature Review

Intermittent and lumpy demand forecasting constitutes a distinct forecasting setting characterized by long sequences of zero demand interspersed with irregular positive observations, where demand occurrence and demand size may follow different stochastic dynamics. In such cases, conventional continuous-demand forecasting approaches may yield biased estimates and suboptimal inventory decisions. Consequently, the forecasting literature has developed dedicated methods and benchmarking practices specifically tailored to intermittent demand, including Croston-type estimators, aggregation–disaggregation frameworks, and empirical studies comparing statistical and machine learning approaches under retail-like data conditions [19,20]. Building on these established foundations, the present study positions its contribution as an application-driven empirical benchmark in a product-level fashion retail environment with rich exogenous variables, rather than as a methodological innovation.

A comprehensive review by Giannopoulosa et al. [21] thoroughly examines the evolution, taxonomy, and implementation details of ML algorithms in this challenging domain. Their work highlights that various ML techniques, including artificial neural networks (ANNs), tree-based methods, and ensemble approaches, offer promising results by learning nonlinear relationships and capturing complex patterns within sparse datasets, thereby overcoming the limitations of conventional methods.

Classical intermittent-demand methods explicitly model the stochastic structure of zero-inflated series. The seminal work of Croston [22] introduced a decomposition-based estimator separating demand size and inter-arrival intervals, later refined through bias-adjusted variants such as the Syntetos–Boylan Approximation (SBA) and the Teunter–Syntetos–Babai (TSB) method. In parallel, aggregation–disaggregation approaches such as ADIDA (Aggregate–Disaggregate Intermittent Demand Approach) were proposed to stabilize sparse demand patterns before re-disaggregation to the original frequency [23]. These methods have become standard baselines in intermittent-demand benchmarking studies and provide a methodological foundation for contemporary empirical comparisons. For a comprehensive treatment of intermittent demand forecasting, including its context,

various methods, and practical applications, readers are referred to the textbook by Boylan and Syntetos [24].

Demand forecasting plays a critical role in enhancing operational efficiency within dynamic and competitive environments such as the retail sector. As retail systems have become increasingly complex, recent research has shown a growing tendency to integrate artificial intelligence (AI) and machine learning (ML) techniques with traditional time series models to improve forecasting accuracy. Accordingly, the demand forecasting literature encompasses a broad spectrum of methodologies and explanatory variables, reflecting both methodological evolution and the increasing availability of high-dimensional retail data [1].

Traditional time series analysis remains a widely used approach in demand forecasting, relying primarily on historical demand patterns to predict future sales. For example, Braun et al. [5] employ Seasonal Autoregressive Integrated Moving Average (SARIMA) and Support Vector Regression (SVR) models for short-term demand forecasting, while Aydın and Yazıcıoğlu [25] compare ARIMA and Artificial Neural Networks (ANNs), reporting that ARIMA can outperform ANN models under certain data conditions. Although such approaches are effective for relatively smooth demand patterns, their performance may deteriorate in the presence of intermittency and zero inflation, motivating the exploration of more flexible modeling paradigms.

More recently, ML-based methods have gained traction due to their scalability, flexibility, and ability to capture nonlinear relationships in high-dimensional datasets. Regression-based models and ensemble learning algorithms are particularly prominent in the literature. Jain et al. [6] demonstrated that Extreme Gradient Boosting (XGBoost) outperforms conventional regression techniques in retail sales forecasting, while Sekban [26] evaluated several ML algorithms—including support vector machines, decision trees, random forests, and XGBoost—and reported high predictive accuracy across multiple model types.

The retail sector, characterized by dynamic and large volumes of data, presents a fertile ground for machine learning-based demand forecasting. Studies by Abdelfatah [27] and Karim [28] have further demonstrated that ML approaches in retail supply chains yield significant benefits in areas such as inventory management, reduction of stockouts, and enhancement of operational efficiency. These approaches improve forecasting accuracy by integrating diverse data types, including historical sales, promotions, seasonality, and exogenous factors. Beyond general retail, ML has also shown promise in specific areas like new product demand forecasting, where historical data is scarce. Wang [29] proposed a dynamic dual-phase forecasting model that combines machine learning and statistical control to address the 'cold-start' problem for new products, leveraging insights from similar products and product attributes.

Large-scale forecasting competitions, particularly the M4 and M5 competitions [30], have provided systematic evidence regarding the comparative performance of statistical and machine learning methods in retail SKU-level forecasting. The M5 competition, in particular, emphasized scale-free evaluation metrics (e.g., WRMSSE, RMSSE) and rigorous rolling-origin validation protocols. These findings highlight the importance of transparent benchmarking frameworks and careful metric selection in empirical forecasting research [31].

Furthermore, recent studies have demonstrated the efficacy of machine learning approaches in specific retail contexts. For instance, Giri and Chen [32] proposed an intelligent forecasting system utilizing deep learning and image feature attributes to predict weekly sales of new fashion apparel, effectively addressing challenges posed by the industry's transient behavior. Similarly, Ahmadov and Helo [33] provided empirical evidence that deep neural networks (DNNs) can significantly outperform classic methods like Moving

Average, Exponential Smoothing, Croston's method, and ARIMA for forecasting intermittent online sales. Large-scale empirical studies further indicate that ML approaches can outperform classical statistical models under certain retail demand conditions when evaluated using rigorous benchmarking protocols [34].

Ensemble learning approaches, which aggregate predictions from multiple models, have also been shown to enhance forecasting accuracy. Loureiro et al. [7] reported improved performance from ensemble methods relative to individual algorithms such as decision trees and support vector machines. In fashion retail environments—where demand volatility is driven by seasonality, rapidly changing consumer preferences, and limited historical data—machine learning approaches have shown considerable promise. Anitha and Neelakandan [35] provided a comprehensive review of AI- and ML-based demand forecasting techniques for new fashion products and highlight the potential of advanced learning architectures to address industry-specific challenges.

Deep learning methods have emerged as a powerful alternative for demand forecasting, particularly due to their ability to process large-scale datasets and model complex nonlinear relationships. Kaneko and Yada [36] develop a sales forecasting model based on point-of-sale data and report that deep learning models significantly outperform logistic regression. Saha et al. [37] applied a deep learning framework in a multinational retail context and found consistent improvements over traditional approaches across multiple forecasting horizons. The integration of exogenous variables further enhances the effectiveness of deep learning models. Liu and Ichise [3] for instance, incorporated meteorological data into a Long Short-Term Memory (LSTM) network to predict weather-driven sales variations, while Aci and Doğansoy [38] demonstrated the superior performance of deep learning models in e-retail environments characterized by complex consumer behavior.

In fashion retail settings, deep learning models have also been used to capture visual and product-level attributes that influence consumer preferences. Li et al. [39] proposed a hybrid predictive framework combining merchandise image data with transaction records to forecast consumer color preferences, demonstrating that product attributes can play a significant role in demand formation. More recently, de Castro Moraes et al. [40] introduced a CNN–LSTM architecture that integrates exogenous variables and captures both seasonal patterns and inter-product correlations, substantially improving forecast accuracy in retail applications.

The evolution of machine learning algorithms in demand forecasting has been significantly shaped by the emergence of deep learning (DL) models. Habib and Hossain [41] illustrated how advanced deep learning and strategic feature engineering approaches, even in diverse fields like wind power prediction, provide valuable methodological insights applicable to demand forecasting. More recently, Transformer-based models, initially successful in natural language processing, have demonstrated superior performance in time series forecasting. Oliveira and Ramos [42] evaluated several Transformer architectures—including Vanilla Transformer, Informer, Autoformer, PatchTST, and Temporal Fusion Transformer (TFT)—for retail demand forecasting, finding them to significantly outperform traditional methods. Among these, the Temporal Fusion Transformer (TFT) stands out for its ability to process complex time series data, capture multi-variate interactions, and enhance model interpretability [43].

Hybrid forecasting models that integrate traditional time series methods with ML and deep learning techniques have attracted increasing attention due to their ability to capture both linear and nonlinear demand patterns. Yücesan [44] reported strong forecasting performance using ARIMAX–ANN and SARIMAX–ANN models, while Huber and Stuckenschmidt [11] showed that hybrid neural network and gradient-boosted decision tree models consistently outperform purely linear approaches in seasonal forecasting tasks.

Additional studies confirm the effectiveness of hybrid methods in high-dimensional retail environments, including promotional and SKU-level forecasting problems [45–48].

Recent research has further explored hybrid architectures that integrate unsupervised learning and ensemble techniques. Van Steenbergen and Mes [16] proposed the Demand-Forest model, combining K-means clustering, Random Forests, and Quantile Regression Forests to address cold-start demand forecasting. Ensafi et al. [13] compared classical time series models with LSTM networks for seasonal demand forecasting and reported superior performance from deep learning approaches in complex seasonal settings.

Artificial neural networks (ANNs) and deep neural networks (DNNs) have also been widely applied across multiple industries, reinforcing their potential as robust demand forecasting tools. Güven and Şimşir [49] achieved high accuracy in apparel sales forecasting using ANNs and support vector machines, while Seyedan and Mafakheri [12] proposed a multi-step forecasting framework combining customer segmentation, LSTM networks, and Prophet models. Beyond retail, deep learning models have demonstrated strong performance in automotive [50], energy [51], and transportation demand forecasting [52,53], highlighting their general applicability.

A parallel line of research explores the decoupling of intermittent demand into occurrence and magnitude components. Recent studies suggest that separating demand occurrence from demand size may offer practical advantages in specific intermittent-demand settings, particularly when zero inflation dominates the data-generating process [54]. However, this perspective is not universally accepted, as demand occurrence and magnitude are jointly realized in practice and zero observations may carry informative signals. Accordingly, decoupled frameworks are best viewed as pragmatic alternatives whose effectiveness must be assessed empirically rather than as inherently superior solutions. To further enhance the practical utility of such frameworks, understanding the distinct feature drivers for each stage (occurrence vs. magnitude) through interpretability techniques becomes paramount.

While deep learning models offer powerful predictive capabilities, their increasing complexity often leads to a 'black box' nature, posing significant challenges for interpretability [41]. This lack of transparency can hinder trust and adoption by corporate decision-makers, who require understanding not just what a model predicts, but why. In this context, models like the Temporal Fusion Transformer (TFT) represent a crucial advancement by integrating strong predictive power with enhanced interpretability [55]. For tree-based machine learning models, post hoc interpretability methods such as SHAP (SHapley Additive exPlanations) values provide a robust framework to quantify feature contributions and reveal model logic, directly addressing the 'black box' concern and translating predictive power into actionable insights. Zhang et al. [43], in their study on live-streaming cross-border e-commerce sales forecasting, demonstrated how TFT's gating mechanism and variable selection network dynamically reveal which features contribute most to a prediction. This interpretability supports actionable business decisions, such as optimizing marketing budgets, implementing dynamic inventory management strategies, and refining real-time decision-making. For instance, TFT's attention mechanism can show that short-term forecasts rely more on recent historical data, while long-term forecasts emphasize more stable signals like logarithmic sales and seasonality [43].

Beyond model interpretability, the effective implementation of ML/DL forecasting solutions hinges on meticulous technical considerations. The comprehensive review by Giannopoulosa et al. [21] delves into critical aspects such as hyper-parameter tuning, data partitioning strategies, training methodologies, and feature engineering. They emphasized that these elements are not merely peripheral but fundamentally impact model performance. For example, robust feature engineering (temporal, contextual/expert-elicited, market-

driven) is crucial for capturing the diverse and variable nature of time series problems, especially in intermittent and lumpy demand scenarios. The authors also highlighted the importance of systematic approaches to hyper-parameter optimization (e.g., automated tools like Optuna or meta-heuristic algorithms) and appropriate data partitioning (e.g., k-fold cross-validation or specific train/validation/test splits based on data characteristics) to ensure generalizability and prevent overfitting. These detailed technical analyses provide insights into not only how ML models work but also why they produce certain predictions, thereby fostering a more informed and strategic decision-making process in business.

In line with this literature, the present study evaluates two complementary modeling scenarios for intermittent retail demand under a unified benchmarking framework. The first is a one-stage approach, in which demand is forecast directly over the full dataset. The second is an established two-stage approach, in which a classification model first predicts demand occurrence and a regression model then estimates demand magnitude conditional on positive demand. The empirical value of this decomposition is not assumed a priori, but is instead assessed against single-stage machine learning models and classical intermittent-demand baselines under identical validation settings. Following the M5 Competition standards and the recent intermittent demand literature, model performance is evaluated primarily using WRMSSE (Weighted Root Mean Squared Scaled Error), which provides appropriate scaling for intermittent demand patterns while accounting for series-level heterogeneity.

By evaluating these approaches within a unified benchmarking framework and against classical intermittent-demand baselines, the study seeks to provide incremental empirical evidence on their relative performance in a product range-level fashion retail context characterized by rich explanatory features.

3. Materials and Methods

The data used in this study are proprietary and cannot be shared publicly due to commercial confidentiality agreements. The code used in this study is not publicly available because it was developed within the scope of a proprietary industrial application; however, methodological details can be provided by the corresponding author upon reasonable request.

This study did not involve human participants or animals and therefore did not require ethical approval.

During the preparation of this manuscript, generative artificial intelligence tools were used to assist with language refinement, text editing, and the drafting of preliminary non-final text. No generative artificial intelligence tools were used to generate data, perform formal analysis, produce figures, or make scientific conclusions. All outputs were critically reviewed, revised, and approved by the authors, who take full responsibility for the final content of this manuscript. All machine learning and deep learning models were implemented in Python (version 3.9.13). The deep learning benchmarks were implemented using PyTorch (version 2.8.0), and the Temporal Fusion Transformer (TFT) model was developed using PyTorch Forecasting (version 1.4.0).

This study adopts a systematic empirical methodology to evaluate alternative modeling strategies for intermittent demand forecasting in retail. The workflow begins with data collection and preprocessing to ensure the quality and consistency of input variables, followed by the implementation and comparison of two modeling frameworks. The first framework applies a one-stage approach, in which demand quantities are predicted directly. The second follows established practices in intermittent demand forecasting by decomposing the problem into two sequential tasks: (i) classification of demand occurrence and (ii) regression of demand magnitude conditional on positive demand.

In addition to machine learning models, classical intermittent-demand benchmark methods—namely, Croston's method and the Syntetos–Boylan Approximation (SBA)—were implemented to ensure compliance with established field standards. All models were evaluated under identical data splits and validation settings to provide a fair and transparent comparison. For the machine learning component, ensemble-based models—namely, Gradient Boosting Machine (GBM), XGBoost, LightGBM, and Random Forest (RF)—were implemented within both one-stage and two-stage configurations.

The overall workflow is summarized in Figure 1.

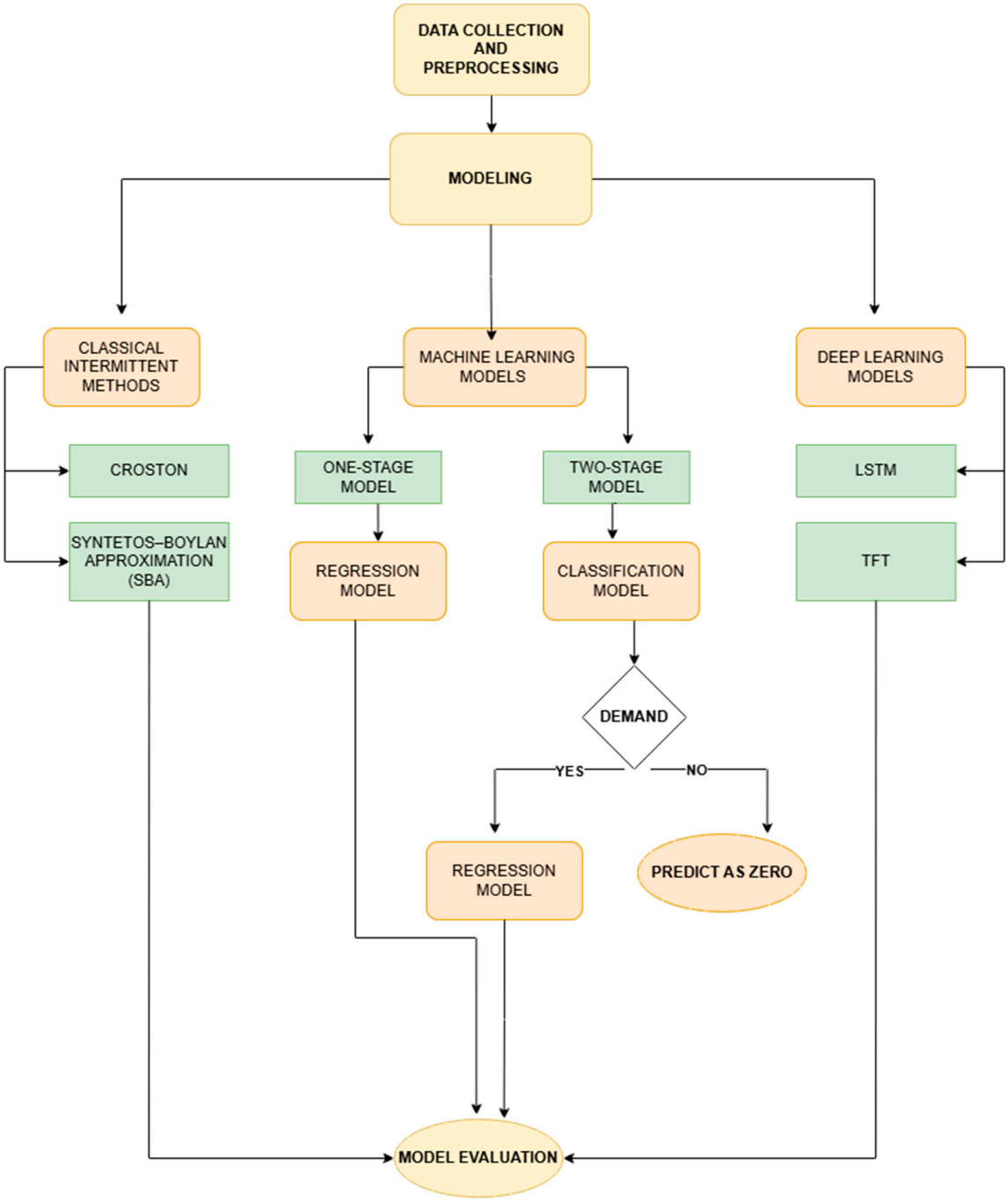

Figure 1. Overall modeling framework. Light orange/beige boxes indicate the main process steps and modeling stages, green boxes denote the forecasting methods/models, the diamond represents a decision point, and the ovals indicate outputs/evaluation stages.

3.1. Data Source and Retail Context

The primary dataset for this study consists of retail data obtained from the Iraqi operations of one of Turkey's leading clothing store chains, including factors influencing demand, such as weather information. The dataset provides extensive historical sales data from 30 retail stores of the clothing chain located in Iraq over several years. Due to its comprehensive nature, it is particularly suitable for developing and testing time series forecasting models. The dataset includes weekly sales figures, detailed product information, discount rates, special event indicators, and weather-related variables. The retailer's identity is anonymized due to data confidentiality agreements. The dataset used in this study contains no personally identifiable information and was utilized solely for academic research purposes. While data is collected at the individual store level, our forecasting unit aggregates sales across all 30 stores for specific product groups defined by their attributes.

The operational context of Iraq presents several unique characteristics that significantly influence fashion retail demand and differentiate this study from those conducted in more commonly researched Western or East Asian markets. Firstly, the region experiences a distinct climate, with extremely hot and dry summers (often exceeding 45 °C) and milder, sometimes rainy, winters. This leads to pronounced seasonal shifts in clothing demand, often more extreme than in temperate climates, and makes weather anomalies particularly impactful on fashion choices and purchasing behavior. Secondly, socio-economic factors, including the ongoing post-conflict recovery, an oil-dependent economy, and a young, urbanizing population, contribute to a dynamic and sometimes volatile consumer purchasing behavior. The prevalence of cash-based transactions and the timing of salary disbursements (as noted later in this section) play a more critical role in weekly sales patterns compared to credit-driven economies in developed markets. Thirdly, the strong influence of local cultural and religious events, such as Eid al-Fitr and Eid al-Adha (beyond Ramadan Eid, which is already incorporated), creates distinct demand spikes and troughs that may differ in timing, duration, and intensity from major secular holidays observed in other regions. Finally, the developing retail infrastructure and supply chain challenges in the region can introduce additional complexities to inventory management, making accurate and context-aware forecasting even more crucial. These specific environmental and behavioral factors underscore the importance of integrating rich exogenous information, as demonstrated in our feature set, to effectively model demand in such a unique market.

Sales data offer detailed insights into demand patterns, necessary for understanding fluctuations over time. Product information, including details such as color, fit, collar type, and thickness, allows for contextualizing sales trends. Discount data highlight the impact of promotions on sales, emphasizing the importance of promotional activities. Economic factors provide a broader context, illustrating how external factors affect consumer purchasing behavior and retail sales.

3.2. Product Group Selection and Feature Construction

The selection of factors used in this study is critical for constructing accurate forecasting models. The factors influencing demand include Color Group, NOS (Never Out of Stock—indicating basic products intended to be continuously available), Fit, Collar Type, Thickness, Total Option Count, Discounted Option Count, Lowcategory Option Count, Midcategory Option Count, Topcategory Option Count (see Section 3.2.7), Special Event Count, Special Event Label, Temperature, Precipitation Conditions, and sales quantities for the previous four weeks. Weekly sales serve as the target variable, reflecting the sales figures recorded each week.

3.2.1. Category Group and Merch Brand Age Group Determination

In the worldwide (domestic and international) subtotal, the average stock percentage, total sales percentage, and gross profit percentage values for a one-year period from the first week of 2024 to the first week of 2025 were examined across thirty different category groups (CG). Among these, the category groups 'Knitted Body T-Shirt Short Sleeve' and 'Knitted Pants' were identified as the best-performing groups company-wide in terms of sales and profit generated relative to the stock held, while also considering their cover (stock/sales) values. These two groups, when combined, accounted for approximately 18% of the company's overall total sales and 15% of the company's overall total profit within the analyzed period. At the same time, selecting one topwear and one bottomwear product group allowed the observation of the effects of different product range characteristics.

3.2.2. Range Determination

The identification of relevant product attributes, or range criteria, is crucial for defining the specific product variations within each selected group. These criteria are determined for the identified Category Group—Merch Brand Age Group (KG-MMYG) pairs. KG-MMYG is a hierarchical category that combines a broad product category (e.g., 'Knitted Body T-Shirt Short Sleeve') with a specific brand and age demographic (e.g., 'Men's Casual Trendy'). For each KG-MMYG combination, range criteria are selected based on their operational significance. This process involves calculating the ratio of 'NUMBER OF OPTIONS ENTERED' (the count of unique SKUs for which a specific range criterion's data is recorded by the product team) to the 'TOTAL OPTION COUNT' (the total number of unique SKUs within that KG-MMYG combination). A threshold value is then applied to these percentage values, and only criteria above this threshold are considered. This data-driven approach ensures that we include features that product teams consistently prioritize for data entry, signifying their critical importance in defining a product's identity and market position (e.g., NOS, as a basic product attribute, is always recorded). For instance, for the first sample group, 'Knitted Body T-Shirt Short Sleeve' category group—'Men's Casual Trendy' age group, the determined range criteria were NOS, Basic Color, Fit, Collar Type, and Thickness. Further evaluation was conducted across 25 brands within these category groups. It was found that 'Men's Casual Trendy' and 'Women's Vision' age groups generated the highest sales for both the 'Knitted Body T-Shirt Short Sleeve' and 'Knitted Pants' category groups. Specifically, these age groups contributed approximately 24% of the total sales within the 'Knitted Body T-Shirt Short Sleeve' category group and 21% within the 'Knitted Pants' category group. In terms of profit, the two category groups together contributed to approximately 25% of the total profit generated by these age groups. Examples of products corresponding to the 'Men's Casual Trendy' and 'Women's Vision' age groups are presented in Figures 2 and 3.

The four sample groups on which the finalized models will be applied are as follows:

- Knitted Body T-Shirt Short Sleeve—Women's Vision
- Knitted Body T-Shirt Short Sleeve—Men's Casual Trendy
- Knitted Pants—Men's Casual Trendy
- Knitted Pants—Women's Vision

Table 1 illustrates the range determination process for the 'Knitted Body T-Shirt Short Sleeve—Men's Casual Trendy' sample group. It showcases multiple range criteria (product attributes) identified for this specific group, along with their associated counts and distinct values. This table serves as an illustrative example of how several relevant features are selected, rather than implying that only a single feature is chosen per group. It is crucial to note that the TOTAL OPTION COUNT (e.g., 2079 for 'Knitted Body T-Shirt Short Sleeve—Men's Casual Trendy') refers to the total number of individual Stock Keeping

Units (SKUs) within one specific product category–merch brand age group combination (KG-MMYG) over the entire historical period. This count is specific to each individual KG-MMYG combination and varies significantly across other product groups, reflecting their distinct SKU compositions. Crucially, this value does not represent the number of forecasted time series in our study.

Figure 2. Men's casual trendy.

Figure 3. Women's vision.

Table 1. Range determination of knitted body—men's casual.

CATEGORY GROUP	MERCH BRAND AGE GROUP	RANGE CRITERIA	CRITERIA DISTINCT VALUE COUNT	NUMBER OF OPTIONS ENTERED	TOTAL OPTION COUNT	RATIO
Knitted Body T-Shirt Short Sleeve	Men's Casual Trendy	NOS	2	2079	2079	1.00
Knitted Body T Shirt Short Sleeve	Men's Casual Trendy	BASIC COLOR	24	2079	2079	1.00

Table 1. *Cont.*

CATEGORY GROUP	MERCH BRAND AGE GROUP	RANGE CRITERIA	CRITERIA DISTINCT VALUE COUNT	NUMBER OF OPTIONS ENTERED	TOTAL OPTION COUNT	RATIO
Knitted Body T Shirt Short Sleeve	Men's Casual Trendy	FIT	6	2077	2079	1.00
Knitted Body T Shirt Short Sleeve	Men's Casual Trendy	THICKNESS	3	2053	2079	0.99
Knitted Body T Shirt Short Sleeve	Men's Casual Trendy	COLLAR TYPE	6	2054	2079	0.99

Footnotes for Table 1: RANGE CRITERIA: Specific product attributes identified as critical for defining the product range and influencing demand within the given category Group and Merch Brand Age Group combination. These are features for which product teams consistently ensure data entry for each SKU, indicating their essential nature and operational importance. CRITERIA DISTINCT-VALUE COUNT: The number of unique values a specific RANGE CRITERIA attribute takes for the items within the listed category Group and Merch Brand Age Group combination (e.g., NOS has 2 distinct values, typically 0 or 1; Basic Color has 24 distinct color values). NUMBER OF OPTIONS ENTERED: The count of unique Stock Keeping Units (SKUs) for which data for the specific RANGE CRITERIA is recorded within the given group by product teams. TOTAL OPTION COUNT: The total number of unique SKUs within the given Category Group and Merch Brand Age Group combination. RATIO: The ratio of 'NUMBER OF OPTIONS ENTERED' to 'TOTAL OPTION COUNT', indicating the completeness of data entry for that criterion within the group, used to determine its relevance as a range criteria. The high cardinality of some categorical variables (i.e., having many unique values) can indeed affect model training and performance, potentially leading to increased dimensionality with encoding methods like one-hot encoding. However, the 'RANGE CRITERIA' selected through our methodology (e.g., NOS, Basic Color, Fit) have inherent business significance and are crucial for demand modeling.

Using the same method, the range criteria for the other three sample groups, for which models will be tested, are as follows:

- Knitted Body T-Shirt Short Sleeve—Women's Vision: NOS, Fit, Color, Collar Type, Length
- Knitted Pants—Men's Casual: NOS, Fit, Color, Thickness, Leg Detail
- Knitted Pants—Women's Vision: NOS, Fit, Color, Length, Thickness, Leg Detail

3.2.3. Color Grouping

Hierarchical clustering was used for the purpose of grouping colors. The colors were converted to the CIELAB color space, which is a color space better suited for distinguishing colors based on how the human eye perceives them. This transformation allows for a more accurate assessment of color similarities.

Data Preparation: A total of 24 different colors were defined in RGB format for the project. These colors were selected to create a visual dataset.

Color Space Transformation: The colors were transformed into the CIELAB color space. This transformation is necessary to more accurately calculate the distances between colors.

Hierarchical Clustering: The colors were grouped using the hierarchical clustering method. In this approach, similar colors were brought together, and a dendrogram (tree structure) was created. The dendrogram visualizes the relationships between the colors, showing which colors are closer to each other.

The color groups determined through hierarchical clustering and the colors included in each group are provided in Table 2.

Table 2. Color groups.

COLOR GROUPS	COLORS
1	Red, Burgundy, Brown, Orange, Coral
2	Green, Yellow

Table 2. *Cont.*

COLOR GROUPS	COLORS
3	Purple, Navy Blue, Lilac, Plum, Fuchsia
4	Turquoise, Blue
5	Anthracite, Black, Petrol, Gray, Indigo, Khaki
6	Beige, White, Ecru, Pink

3.2.4. Special Event Labeling

Instead of directly including the special days and the weeks in which they occur in the dataset, the relevant special days were labeled and incorporated into the dataset. For the labeling process, the sales graph of each special day was examined on a weekly basis, and it was observed that the effect of the special day on sales could start from one or two weeks prior and often extend for one or two weeks after the event, reflecting a temporal ripple effect on demand. In this way, the weeks when the effects of special days on sales began were identified and labeled with values ranging from -2 to +2, with 0 representing the event week itself or weeks where no specific event's significant sales impact was observed within this defined window. This 'Special Event Label' feature was subsequently treated as a numerical covariate in our machine learning models.

We acknowledge that treating such an ordinal temporal indicator numerically might raise concerns about implying a linear relationship where none exists. However, this numerical representation is not intended to imply a linear relationship, but rather to encode the ordinal temporal proximity and direction relative to an event. Our primary models (Random Forest, XGBoost, LightGBM, and Gradient Boosting) are tree-based. These models are highly effective at learning nonlinear relationships and identifying optimal splitting points on numerical features without assuming strict linearity. The numerical scale (-2 to +2) inherently captures the sequential phases of an event's impact (pre-event, event, post-event), allowing tree-based algorithms to effectively partition the data based on these distinctions (e.g., distinguishing between different lead/lag times) and model their unique and dynamic impact on demand.

The following graph shows the total sales quantities for the "Ramadan Eid" special day on a year-week basis. The weeks corresponding to Ramadan Eid are 2022-19, 2023-17, and 2024-15. The weeks corresponding to Ramadan Eid are highlighted in red in the graph, while other special days falling within the relevant effect period are highlighted in orange. From this graph, it can be observed that the effect of Ramadan Eid on sales started increasing two weeks prior and continued to decrease two weeks after the event. In this manner, sales graphs for all special days in the country were examined, and the effect weeks were determined and labeled.

Figure 4 presents the sales graph for the special day of Ramadan Eid. Beyond Ramadan Eid, other significant cultural and religious observances in Iraq, such as Eid al-Adha and various local festivities, also contribute to distinct patterns in consumer spending and fashion demand. The timing and impact of these events are carefully tracked and labeled, reflecting their unique influence on the retail calendar, which can differ markedly from secular holiday cycles in Western markets and may not be adequately captured by generic holiday indicators.

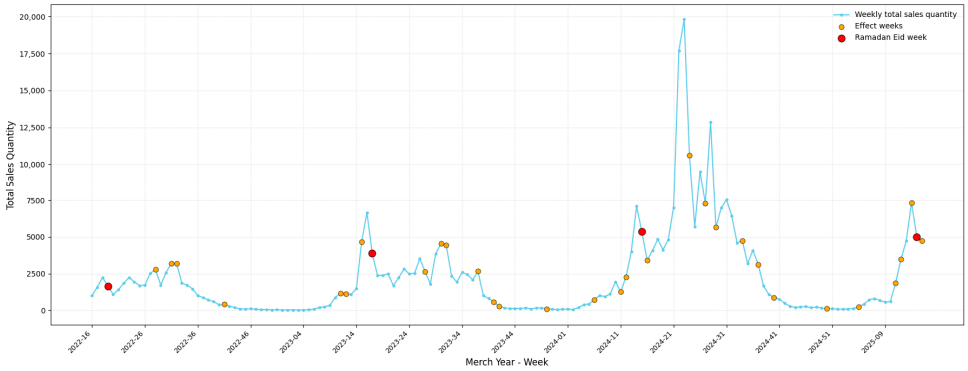

Figure 4. Ramadan Eid sales chart.

3.2.5. Meteorological Data

Meteorological data such as Temperature and Precipitation are also utilized in the dataset. For each week-city observation, a historical baseline was established by calculating the average temperature and precipitation for the last five years (excluding the current year to avoid look-ahead bias) for that specific week-of-year and city. The 'Temperature Anomaly Status' and 'Precipitation Anomaly Status' labels—categorized as 'Low Anomaly', 'High Anomaly', or 'No Anomaly'—were then determined by comparing the actual observed temperature and precipitation for the respective Week-City against this historical baseline. For the training period, these anomaly labels were directly derived from the actual observed weather data. Crucially, for forecasting into the 13-week holdout period, these anomaly status features were populated based on the most frequently observed anomaly status (mode) for the corresponding week-of-year and city, derived from the preceding five years of historical data. This approach ensures that no future actual weather data was leaked into the forecasting process, as the anomaly labels for the forecast horizon are entirely based on historical patterns known at the time of forecast. In the labeling process, common threshold values for anomaly detection in climate-related studies are used: a temperature threshold of 3.5 °C and a precipitation threshold of 0.5 mm.

3.2.6. Price Data

To incorporate price information, our approach is specifically designed to provide relevant features for our product group-level forecasting objective. A product group, which serves as the unit of our demand forecasting, is defined by a unique combination of attributes including Year, Week, Color Group, NOS, Fit, CollarType, and Thickness. Sales for this product group are aggregated across all 30 retail stores.

For each individual SKU within a given MerchWeek, the Last Sales Price based on cash sales is determined. This Last Sales Price is then compared to the weekly average Last Sales Price within its MerchWeek to assign a Price Category ('Low Price', 'Medium Price', or 'High Price'). Additionally, the discount status of each SKU is tracked.

However, these individual SKU-level price details (Last Sales Price, discount status, or assigned price category) are not directly used as features in the models. Instead, we aggregate this information to create five distinct numerical count features at the product group level. These final features, which serve as covariates in our models, are:

- Total Option Count: The total number of unique SKUs available within a specific product group for that week.
- Discounted Option Count: The number of SKUs within the product group that are currently under discount.
- Lowcategory Option Count: The number of SKUs classified as 'Low Price' within the product group.

- Midcategory Option Count: The number of SKUs classified as 'Medium Price' within the product group.
- Topcategory Option Count: The number of SKUs classified as 'High Price' within the product group.

This methodology, using these five numerical count features, provides a robust representation of the pricing distribution, promotional activity, and breadth of offering within a product group. This is highly relevant for group-level demand forecasting, as it captures the overall price positioning and competitive landscape of the group. Furthermore, this approach helps to mitigate the risk of overfitting that could arise from using numerous highly granular numerical averages at this aggregated level, while still providing valuable insight into the group's overall price composition. These aggregated count features are then used as covariates in our models.

3.2.7. Determination of Variables Used for Demand Forecasting

During the Range Determination phase, different range features were defined for the four sample groups mentioned, and categorical variables such as collar type, thickness, leg detail, and height varied across the four sample groups. Attributes of the 'Knitted Body T-Shirt Short Sleeve—Men's Casual' dataset and the data types of the corresponding attributes are illustrated in Table 3.

Table 3. Variables Used for Demand Forecasting.

#	VARIABLE	VARIABLE TYPE
1	Year	INT
2	Week	CATEGORY
3	Week of Month	CATEGORY
4	Color Group	CATEGORY
5	NOS	INT
6	Fit	CATEGORY
7	CollarType	CATEGORY
8	Thickness	CATEGORY
9	Total Option Count	INT
10	Discounted Option Count	INT
11	Lowcategory Option Count	INT
12	Midcategory Option Count	INT
13	Topcategory Option Count	INT
14	Sales Quantity	INT
15	Previous Week Sales Quantity	INT
16	Previous 2 Week Sales Quantity	INT
17	Previous 3 Week Sales Quantity	INT
18	Previous 4 Week Sales Quantity	INT
19	Special Event Count	INT
20	Special Event Label	INT
21	Temperature Anomaly Status	CATEGORY
22	Precipitation Anomaly Status	CATEGORY

Note: INT indicates integer data type.

Our final dataset comprises a total of 64 unique attribute-defined product group time series that were actively forecasted after preprocessing and filtering of inactive series. To quantitatively characterize the demand patterns within this dataset and provide context for our forecasting models, we computed three widely recognized intermittency metrics for each of these 64 active product group time series: the Average Demand Interval (ADI), the Squared Coefficient of Variation (CV^2), and the Intermittent Demand Index (IDI). These metrics are crucial for understanding the nature of demand and comparing our findings with other studies in the intermittent demand forecasting literature.

The Average Demand Interval (ADI) measures the average number of time periods between successive non-zero demand observations. A higher ADI indicates more infrequent demand. The Squared Coefficient of Variation (CV^2) quantifies the variability in the size of non-zero demands. A higher CV^2 suggests greater volatility in demand magnitudes. The Intermittent Demand Index (IDI), defined as ADI $\times$ CV^2, provides a comprehensive measure of intermittency, with values typically classifying demand into categories such as smooth, intermittent, erratic, or lumpy [24].

The summary statistics for these metrics across our 64 unique product group time series are presented in Table 4.

Table 4. Summary Statistics for Intermittency Metrics.

METRIC	COUNT	MEAN	STD	MIN	25%	50% (MEDIAN)	75%	MAX
ADI	64.00	1.49	0.80	0.76	1.03	1.20	1.62	5.07
CV^2	64.00	1.46	0.60	0.41	1.01	1.40	1.81	2.93
IDI	64.00	2.18	1.40	0.31	1.14	1.83	2.57	6.76

As shown in Table 4, our dataset exhibits a wide range of demand intermittency. The mean ADI of 1.49 weeks indicates that, on average, demand occurs roughly every 1.5 weeks across our series, pointing to significant demand sparsity. Concurrently, the mean CV^2 of 1.46 highlights substantial variability in the size of non-zero demands.

Based on the standard category framework (where ADI < 1.32 and $CV^2 < 0.49$ for 'Smooth'; ADI ≥ 1.32 and $CV^2 < 0.49$ for 'Intermittent'; ADI < 1.32 and $CV^2 \geq 0.49$ for 'Erratic'; ADI ≥ 1.32 and $CV^2 \geq 0.49$ for 'Lumpy'), the majority of our series fall into the more challenging categories: 39 series (61%) are classified as Erratic, 23 series (36%) as Lumpy, and only 2 series (3%) as Smooth. Notably, none of our series were classified as purely 'Intermittent' within this framework. This diverse range and the high prevalence of erratic and lumpy demand patterns confirm the challenging nature of this forecasting problem and underscore the necessity of robust forecasting methods that can effectively handle zero-inflation and high demand variability.

After the relevant data was collected and the necessary preprocessing steps were applied, encoding techniques such as one-hot encoding, label encoding, or target encoding were applied to the categorical variables based on the structure of the variable

The Week variable, representing the week of the year (1–52/53), is a crucial temporal feature for capturing recurring seasonal demand effects. While periodic representations (e.g., sine and cosine transformations) are common for cyclical data, we made a deliberate choice to treat Week as a categorical variable, which was then target-encoded. This approach was chosen because fashion retail demand often exhibits sharp, non-smooth, and irregular peaks/troughs (driven by discrete events like promotions or specific collection launches) that may not be accurately captured by smooth periodic functions. Given its high cardinality (up to 52 values), Target Encoding allowed us to efficiently capture the unique historical sales "signature" of each specific week, providing a data-driven representation without

introducing excessive dimensionality (as One-Hot Encoding would). Our tree-based models are highly robust to such target-encoded categorical features, effectively learning these nonlinear, week-specific demand patterns.

Among these variables, the "Week of the Month" variable indicates which week of the month the data corresponds to (ranging from 1 to 5). Employees in Iraq receive their salaries at the beginning of the month, which influences their purchasing behavior. For instance, sales tend to increase during the first week when salaries are deposited, while a decline is observed in the last week of the month. Therefore, this variable was treated as an ordinal categorical variable and processed using the label encoding method. As seen in Figure 5, average sales quantities vary significantly across different weeks of the month. Other categorical variables with relatively limited cardinality—such as color group, fit, product-attribute categories (e.g., collar type, thickness, or length, depending on the sample group), and anomaly indicators—were represented using one-hot encoding.

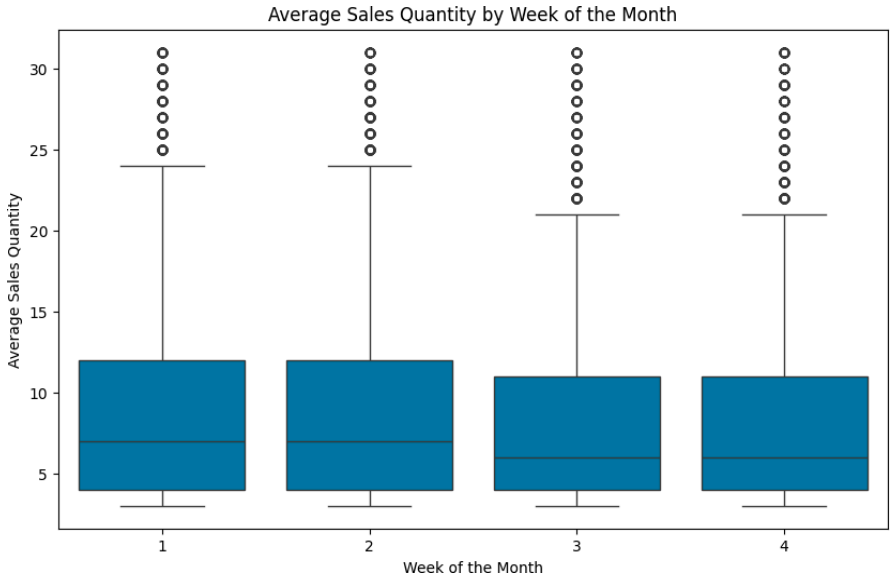

Figure 5. Box plot of sales quantity by week of the month. The circles denote outliers, the box represents the interquartile range, the central line indicates the median, and the whiskers show the data spread.

This visual trend suggests a practical importance for understanding consumer behavior, particularly concerning the influence of salary disbursement cycles on purchasing patterns, which is crucial for operational planning and inventory management. It is important to note that this figure presents a descriptive visual trend and does not imply statistically significant differences in sales, as no formal statistical testing was conducted for this visualization.

Observed weekly sales distribution across the month, visually illustrating the observed variation in average sales quantities and potential patterns related to payment cycles. This descriptive figure highlights practical implications for demand but does not imply statistically significant differences.

3.3. Forecasting Framework and Validation Strategy

The dataset consists of weekly observations from 2022 to the end of 2025. To ensure rigorous out-of-sample evaluation, the final 13 weeks of the dataset were reserved as a fixed holdout period. All models were trained exclusively on observations preceding this holdout window. The same fixed holdout design was preserved for the additional deep learning benchmarks to ensure consistent out-of-sample comparison across all model families.

The 13-week forecasting horizon was selected to reflect a typical quarterly planning cycle in fashion retail, which aligns with seasonal assortment and replenishment decisions. To preserve temporal ordering and prevent information leakage, hyperparameter tuning and model selection were conducted using rolling-origin cross-validation (TimeSeriesSplit) within the training data. In each fold, training observations strictly preceded validation observations.

Series that exhibited zero sales throughout the 13-week holdout period were treated as inactive and excluded from the benchmark. This filtering resulted in a final panel of 64 active attribute-defined product group weekly time series for evaluation. This filtering rule was also applied to the deep learning benchmarks so that LSTM and TFT were evaluated on the same active-series panel used throughout the benchmark design. This filtering was applied consistently across classical intermittent methods and machine learning models to ensure comparability and operational relevance.

All feature transformations, including target encoding for high-cardinality categorical variables, were performed exclusively within the training folds. Encoding mappings learned from training data were applied to validation and holdout data without exposure to future information, thereby eliminating temporal leakage.

Two distinct modeling strategies were examined: one-stage and two-stage modeling. In the one-stage approach, regression models—including Random Forest, Gradient Boosting, LightGBM, and XGBoost—were applied directly to the full training dataset without preliminary filtering.

In the two-stage approach, a Random Forest classifier was first trained to distinguish between zero-demand and positive-demand observations. Specifically, this classifier was trained on the full training dataset where the target variable (SalesQuantity) was binarized (0 for zero demand, 1 for positive demand). To determine the optimal point for classifying demand occurrence, the classifier's predicted probabilities were analyzed using a precision–recall curve on the validation set, and a threshold was selected to maximize the F1-score. This data-driven approach ensures the threshold is optimized to balance precision and recall, which is crucial for intermittent demand where both correctly identifying demand and avoiding false positives (over-forecasting zeros) are important. For instances predicted as zero demand by this optimized classifier, the final forecast was set to zero. For instances predicted as positive demand, regression models were subsequently applied to estimate demand magnitude. These regression models (Random Forest, Gradient Boosting, XGBoost, and LightGBM) were exclusively trained on the subset of the training data where actual positive demand (SalesQuantity > 0) had occurred. This approach ensures that the magnitude estimation is conditioned solely on historical instances of positive sales, allowing the models to learn the characteristics of non-zero demand more effectively. During the prediction phase for the holdout period, the final forecast for each observation was then derived by combining the classifier's decision (demand occurrence or not) with the magnitude prediction from the respective regressor, with all final forecasts clipped at zero to ensure non-negative values. Final performance was evaluated on the fixed 13-week holdout period using the metrics described in Section 3.10.

In addition to the machine learning configurations, classical intermittent-demand benchmarks—Croston's method and the Syntetos–Boylan Approximation (SBA)—were implemented as smoothing-based reference models. These methods were trained on the same pre-holdout training data and evaluated over the identical 13-week holdout period to ensure direct comparability.

3.4. Classical Intermittent-Demand Methods

To establish baseline performance, two classical intermittent-demand forecasting methods were implemented:

- Croston's Method [22];
- Syntetos–Boylan Approximation (SBA) [19].

Croston's method separates the demand process into two components: demand size and inter-arrival interval. Exponential smoothing is applied independently to both components. The forecast is computed as:

$$\hat{y}_t = \frac{\hat{z}_t}{\hat{p}_t}$$

where $\hat{z}_t$ represents the smoothed demand size and $\hat{p}_t$ denotes the smoothed inter-demand interval.

The SBA method introduces a bias correction factor to Croston's estimator:

$$\hat{y}_t^{SBA} = \left(1 - \frac{\alpha}{2}\right)\frac{\hat{z}_t}{\hat{p}_t}$$

where α is the smoothing parameter.

For both Croston and SBA, the smoothing parameter was fixed at $\alpha = 0.1$, following common practice in intermittent-demand benchmarking. Forecasts were generated recursively over the 13-week holdout period using parameters estimated on the pre-holdout training data. Forecasts were produced at the same aggregation level (product range weekly series) and evaluated over the identical 13-week holdout window. This alignment ensures comparability between classical intermittent-demand benchmarks and the proposed machine learning approaches under identical forecast horizon and evaluation metrics.

3.5. Gradient Boosting Machine (GBM)

Gradient Boosting constructs a strong learner by iteratively adding weak learners (typically decision trees) [56]. Each new model aims to minimize the residual errors of the previous models. In each iteration, the model corrects the prediction errors of the current ensemble.

Mathematically, the addition of each new model can be expressed as:

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

where $F_m(x)$ is the cumulative prediction, $F_{m-1}(x)$ is the prediction from the previous model, $h_m(x)$ is the prediction of the newly added weak learner, and η represents the learning rate.

The new model $h_m(x)$ is trained at each step to minimize the errors of the previous model. This typically increases the accuracy of the model but can also lead to overfitting, so proper parameter adjustments are crucial.

In this study, hyperparameters of the GBM model were optimized using randomized search combined with rolling-origin cross-validation within the training period to preserve temporal ordering.

3.6. Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM) is an efficient and scalable implementation of gradient boosting, specifically designed to handle large-scale datasets with high-dimensional features [57]. It builds upon the foundational principles of gradient boosting while introducing key optimizations such as histogram-based binning, leaf-wise tree

growth, and second-order gradient approximation, all of which contribute to faster training speed and reduced memory consumption.

The objective function optimized by LightGBM can be expressed as:

$$L_{GBM} = \sum_{i=1}^{N} L(y_i, f(x_i)) + \Omega(f)$$

where $L(y_i, f(x_i))$ is the loss function, and $\Omega(f)$ is the regularization term.

These terms enhance model accuracy while mitigating overfitting. LightGBM also employs optimization techniques to accelerate the learning process and minimize memory usage.

All LightGBM models were trained exclusively on the pre-holdout training data and evaluated on the fixed 13-week holdout period to ensure consistency with classical intermittent-demand benchmarks.

3.7. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a scalable and efficient implementation of gradient boosting designed to optimize both predictive accuracy and computational performance [58]. It extends traditional gradient boosting by incorporating regularization techniques (L1 and L2 penalties), second-order gradient optimization, and tree pruning mechanisms that improve generalization and reduce overfitting.

The objective function minimized by XGBoost can be formulated as:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $L(y_i, \hat{y}_i)$ is the loss function measuring the discrepancy between observed and predicted values, and $\Omega(f_k)$ is the regularization term controlling tree complexity, defined as:

$$\Omega(f) = \gamma T + 1/2\lambda ||w||$$

where T denotes the number of leaves and w represents leaf weights.

In this study, XGBoost models were trained under the same rolling-origin cross-validation framework and evaluated on the identical 13-week holdout period to ensure consistency with other ensemble-based models.

3.8. Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees, each trained on a randomly sampled subset of the training data using bootstrapping and feature randomness [59]. This diversity among trees enhances the model's generalization capability and mitigates the risk of overfitting.

The final prediction for regression tasks is computed as the average of the predictions from all individual trees:

$$\hat{y}_{RF} = \frac{1}{N} \sum_{i=1}^{N} h_i(x)$$

where $\hat{y}_{RF}$ is the final prediction, $h_i(x)$ is the prediction of the i-th tree, and N is the total number of trees. For classification tasks, the final decision is made by majority voting.

3.9. Deep Learning Benchmarks

To address the need for more contemporary forecasting baselines, two deep learning benchmarks were additionally implemented: a Long Short-Term Memory (LSTM) network

and a Temporal Fusion Transformer (TFT). These models were included as benchmark comparators rather than as proposed methods. Both were evaluated using the same fixed 13-week holdout horizon and the same active-series filtering rule applied in the machine learning benchmark.

3.9.1. Long Short-Term Memory (LSTM)

LSTM networks are recurrent neural architectures designed to capture temporal dependencies in sequential data while alleviating the vanishing-gradient problem associated with conventional recurrent neural networks [60]. Owing to their ability to model nonlinear sequential structure, LSTMs are widely used as benchmark models in forecasting applications with dynamic covariates.

In this study, the LSTM benchmark was implemented as a global multivariate sequence model trained across the active weekly series. Input sequences were constructed using an 8-week historical window and included lagged sales values, temporal indicators, and exogenous retail covariates. The final architecture consisted of a single LSTM layer with 64 hidden units, followed by dropout and fully connected layers for demand prediction. Training was conducted in PyTorch (version 2.8.0) using the Adam optimizer and mean squared error loss, with early stopping applied to reduce overfitting. Final forecasts were generated for the fixed 13-week holdout period.

3.9.2. Temporal Fusion Transformer (TFT)

The Temporal Fusion Transformer (TFT) is a multi-horizon deep learning architecture that combines recurrent processing, attention mechanisms, variable selection networks, and gating layers to model complex temporal dependencies while preserving a degree of interpretability [55]. It is particularly suited to global forecasting problems involving multiple related time series with static and time-varying covariates.

TFT was implemented using the PyTorch Forecasting framework (version 1.4.0), built on PyTorch (version 2.8.0), as a global multi-series model. The specification employed a 16-week encoder length and a 13-week prediction horizon, aligned with the external holdout design of the study. The series identifier was included as a static categorical feature, while calendar and assortment-related variables—such as year, week, week-of-month, option counts, and special event indicators—were included as time-varying known real covariates. Past sales were treated as time-varying unknown reals. Series-level normalization was applied through a GroupNormalizer, and the model was trained using QuantileLoss. This choice is integral to TFT's design as a multi-horizon probabilistic forecasting model, as QuantileLoss (also known as Pinball Loss) is the appropriate objective function for optimizing models to generate accurate quantile predictions. This approach allows the model to capture the inherent uncertainty in demand, providing a more comprehensive view of potential future outcomes crucial for robust decision-making in areas like inventory planning and stock risk mitigation, which often rely on probabilistic forecasts rather than just point estimates. Even though our primary evaluation metrics (e.g., WRMSSE) are point forecast-based, training with QuantileLoss enables the TFT to learn a richer representation of the conditional distribution of demand, from which more robust point forecasts can also be derived. Final forecasts were then evaluated over the same fixed 13-week holdout period used for the classical and machine learning models.

3.10. Model Evaluation

Hyperparameter tuning is critical for achieving optimal performance for machine learning algorithms. Randomized search combined with rolling-origin cross-validation (TimeSeriesSplit) was employed within the training period. This approach preserves

temporal order and prevents future information leakage while efficiently searching the parameter space.

All final performance metrics were computed exclusively on the fixed 13-week holdout period. For the additional deep learning benchmarks, forecasts were generated for the same active weekly series retained in the main benchmark, and evaluation was performed using the same metrics to preserve comparability under the study's benchmark design.

The predictive performance was evaluated using:

- Weighted Root Mean Squared Scaled Error (WRMSSE)—primary metric
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2) (reported as a secondary descriptive measure)

RMSE provides the square root of mean squared error, facilitating interpretability by expressing errors in the original units of the target variable. MAE quantifies the average absolute difference between predictions and actual values, offering a straightforward measure of prediction accuracy. However, as noted in the forecasting literature, MAE can be misleading in some settings, particularly when demand is sparse or intermittent; therefore, in this study it is reported only as a complementary metric alongside WRMSSE, which is adopted as the primary evaluation criterion [31].

WRMSSE (Weighted Root Mean Squared Scaled Error) serves as the primary evaluation metric, following M5 Competition standards for intermittent demand forecasting [30]. The selection of an appropriate error measure is crucial, especially for intermittent demand, as the notion of the 'best' point forecast is highly dependent on the chosen metric [31]. WRMSSE provides appropriate scaling for intermittent demand patterns by normalizing errors against naive forecast performance while weighting series by their sales volume, making it particularly suitable for sparse retail demand settings and accounting for series-level heterogeneity [61]. In parallel, advanced modeling frameworks, including two-stage deep learning approaches, have been increasingly utilized to capture complex consumer preferences and optimize forecasting accuracy within multi-channel retail environments [62].

The coefficient of determination (R^2) quantifies the proportion of variance in the target variable explained by the model. However, its interpretation in sparse, intermittent-demand contexts requires careful consideration. Due to the prevalence of zero observations, the overall mean of sales can be very low. Consequently, models that fail to accurately predict these numerous zeros, often overestimating demand during inactive periods, can generate a sum of squared residuals (SS_{res}) that is larger than the total sum of squares (SS_{tot}) of the actual values around their mean. This leads to negative R^2 values, as observed for classical methods like Croston and SBA in our results, indicating that these models perform worse than a simple naive forecast based on the historical mean. Therefore, while R^2 can still offer a general indication of a model's ability to capture overall variability, it is considered less informative and potentially misleading as a primary performance metric in such settings. Nonetheless, for the sake of completeness and as a supplementary diagnostic measure that allows for comparison with studies using similar metrics, we report R^2 for all evaluated models. The mathematical formulations are as follows:

Root Mean Squared Error (RMSE)

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}}$$

Mean Absolute Error (MAE)

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} |\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}|$$

Coefficient of Determination (R^2)

$$\boldsymbol{R^2} = \mathbf{1} - \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{y_i} - \boldsymbol{\overline{y}})^{\mathbf{2}}}$$

Weighted Root Mean Squared Scaled Error (WRMSSE)

$$\boldsymbol{WRMSSE} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{w_i} * \boldsymbol{RMSSE_i})}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}}(\boldsymbol{w_i})}$$

where $\boldsymbol{RMSSE_i}$ is the Root Mean Squared Scaled Error for series i, and $\boldsymbol{w_i}$ represents the weight based on series volume.

3.11. Simple Benchmarks

To provide a comprehensive set of baselines, we included the historical mean as a fundamental reference point for comparison across different methodological categories. This benchmark represents the average demand observed throughout the respective training periods and data processing contexts.

- Classical Historical Mean: For the classical methods, this benchmark employs the overall historical mean of sales quantities calculated directly from the raw training data. This constant value is then used as the forecast for every instance in the holdout period, serving as a straightforward and robust lower bound for performance.
- Machine Learning Historical Mean: Within the machine learning framework, a historical mean benchmark was also established. This involved calculating the historical mean of sales quantities using the training data as it was prepared and processed for the machine learning models. This benchmark helps to contextualize the performance of complex ML models against a simple, consistent prediction within their specific data pipeline, accounting for any relevant feature engineering or aggregation steps.
- Deep Learning Historical Mean: Similarly, for the deep learning benchmarks, a historical mean was employed. This mean was computed using the training data as processed for the deep learning models, accounting for any specific normalization or scaling applied before model input. This allows for a fair comparison of the deep learning architectures against a simple baseline within their unique data transformation context.

These three variants of the historical mean benchmark, while conceptually similar, are evaluated under the specific data preparation and evaluation protocols pertinent to their respective model categories (Classical, Machine Learning, Deep Learning). This approach ensures that performance comparisons are made against appropriate baselines, accounting for the nuances of each framework's data handling.

3.12. SHAP-Based Feature Importance Methodology

To provide a comprehensive understanding of feature contributions and address the 'black box' nature of complex machine learning models, SHAP (SHapley Additive exPlanations) values were employed. SHAP is a game-theoretic approach that quantifies the contribution of each feature to a model's prediction for a specific instance, aligning with the principles of Shapley values from cooperative game theory. By calculating the average

marginal contribution of each feature value across all possible feature coalitions, SHAP provides a robust and consistent measure of local and global feature importance.

This analysis was specifically applied to our best-performing model, the XGBoost Two-Stage model, to distinguish between features driving demand occurrence and those influencing demand magnitude.

- Demand Occurrence (Classifier Stage): For the demand occurrence classifier (XGBoost Classifier), a shap.TreeExplainer instance was initialized using the trained classifier model. SHAP values were then computed for the positive class (i.e., demand occurrence). To derive global feature importance, the mean absolute SHAP value for each feature was calculated across all instances, providing an aggregate measure of its impact on the classifier's output.
- Demand Magnitude (Regressor Stage): Similarly, for the demand magnitude regressor (XGBoost Regressor), a separate shap.TreeExplainer was utilized with the trained regressor model. Global feature importance for this stage was also determined by computing the mean absolute SHAP value for each feature across all instances, reflecting its overall impact on the predicted sales quantity.

The results of this analysis are visualized using SHAP summary plots, which effectively depict the distribution of SHAP values for each feature, indicating both their overall importance and the direction of their effect on the model output. This detailed SHAP analysis was crucial for providing transparency into the model's decision-making process, specifically differentiating the feature drivers for demand occurrence versus demand magnitude, and enabling the translation of model insights into actionable inventory management strategies.

4. Results

4.1. Comparative Forecasting Performance

The predictive performance of all benchmark models was evaluated on the fixed 13-week holdout period using R^2, MAE, RMSE, and WRMSSE. WRMSSE was employed as the primary comparison metric following M5 Competition standards for intermittent demand forecasting, as it provides appropriate scaling and weighting for sparse retail demand patterns, whereas R^2 was used only as a secondary descriptive indicator. The comparative results are presented in Table 5.

Table 5. Model Performance Results.

Method	Model	WRMSSE	R^2	MAE	RMSE
ML	XGBoost Two-Stage	**0.332**	0.689	5.265	**11.819**
ML	LightGBM One-Stage	0.336	0.657	6.027	12.417
ML	XGBoost One-Stage	0.339	0.617	5.251	13.108
ML	LightGBM Two-Stage	0.342	0.635	5.481	12.802
ML	Random Forest One-Stage	0.344	0.621	5.422	13.039
ML	Gradient Boosting One-Stage	0.347	0.656	6.069	12.422
ML	Gradient Boosting Two-Stage	0.348	0.653	5.528	12.478
ML	Random Forest Two-Stage	0.357	0.588	**5.247**	13.608
Classical	Historical Mean	1.178	−2.922	35.157	36.394
ML	Historical Mean	1.327	−4.643	49.030	50.334

Table 5. *Cont.*

Method	Model	WRMSSE	R^2	MAE	RMSE
Classical	SBA	2.528	−76.436	70.162	161.725
Classical	Croston	2.668	−85.408	74.155	170.837
Deep Learning	TFT	3.634	**0.724**	28.135	75.126
Deep Learning	LSTM	3.964	0.492	45.902	101.807
Deep Learning	Historical Mean	5.544	−0.023	61.178	144.533

Note: Bold values indicate the best performance in each metric column (lowest for WRMSSE, MAE, and RMSE; highest for R^2). Models are ranked by WRMSSE performance. R^2 values are provided as supplementary indicators. Due to known limitations of R^2 in sparse intermittent-demand settings (especially when models perform worse than a naive mean forecast, leading to negative values), it is interpreted as a supplementary diagnostic metric rather than the primary evaluation criterion. Among all evaluated models, the Two-Stage XGBoost achieved the lowest WRMSSE (0.332), followed by LightGBM (0.336), One-Stage XGBoost (0.339), and Two-Stage LightGBM (0.342). Machine learning approaches dominated the top rankings, with all ML models achieving WRMSSE values below 0.36. The classical intermittent-demand benchmarks, Croston and SBA, produced significantly higher WRMSSE values (2.668 and 2.528, respectively), while deep learning models performed poorly with WRMSSE values exceeding 3.6.

4.2. One-Stage Versus Two-Stage Results

Within machine learning approaches, the comparison between one-stage and two-stage configurations showed mixed results. While Two-Stage XGBoost achieved the best overall performance (WRMSSE: 0.332), One-Stage XGBoost also performed competitively (WRMSSE: 0.339). Similarly, One-Stage LightGBM (0.336) outperformed its two-stage variant (0.342).

Conversely, for Random Forest and Gradient Boosting, their one-stage variants slightly outperformed their two-stage counterparts. This suggests that the effectiveness of demand occurrence decomposition varies across different algorithmic families.

4.3. Comparison Across Classical, Machine Learning, and Deep Learning Benchmarks

The results reveal a clear performance hierarchy across methodological approaches. Machine learning models achieved superior performance with WRMSSE values ranging from 0.33 to 0.36, representing a 3.5× improvement over the best classical method (Historical Mean: 1.18) and an 11× improvement over the best deep learning model (TFT: 3.63). Classical intermittent-demand methods showed mixed results. While Historical Mean provided a reasonable baseline (WRMSSE: 1.18), specialized methods like Croston (2.67) and SBA (2.53) performed significantly worse. This suggests that simple averaging may be more effective than complex smoothing approaches in this retail context. Deep learning models, despite achieving competitive R^2 values (TFT: 0.72, LSTM: 0.49), failed to deliver effective intermittent demand forecasting. Both models substantially overestimated demand during zero-demand periods, resulting in poor WRMSSE performance.

4.4. Illustrative Forecast Behavior

Figure 6 presents an illustrative comparison of one-stage and two-stage forecast behavior during intermittent-demand periods for a representative product group time series, using Two-Stage XGBoost which achieved the best overall performance among all evaluated models. The plot spans 29 weeks, with weeks 1–16 representing the training data and weeks 17–29 constituting the 13-week holdout (evaluation) period. The forecast origin is therefore set at the beginning of week 17. As observed, the two-stage forecasts are more closely aligned with zero-demand observations, whereas the one-stage forecasts tend to generate small positive values during some zero-demand weeks. The large spikes observed around weeks 21 and 28 are influenced by the model's effective incorporation of special event indicators (e.g., Eid al-Adha, local festivities) and seasonal features, which are

crucial contextual drivers in the Iraqi market. This representative product group time series exhibits a typical intermittent-demand pattern characterized by frequent zero observations and occasional positive spikes. This visual pattern is consistent with the lower WRMSSE values observed for the two-stage configurations in Table 5.

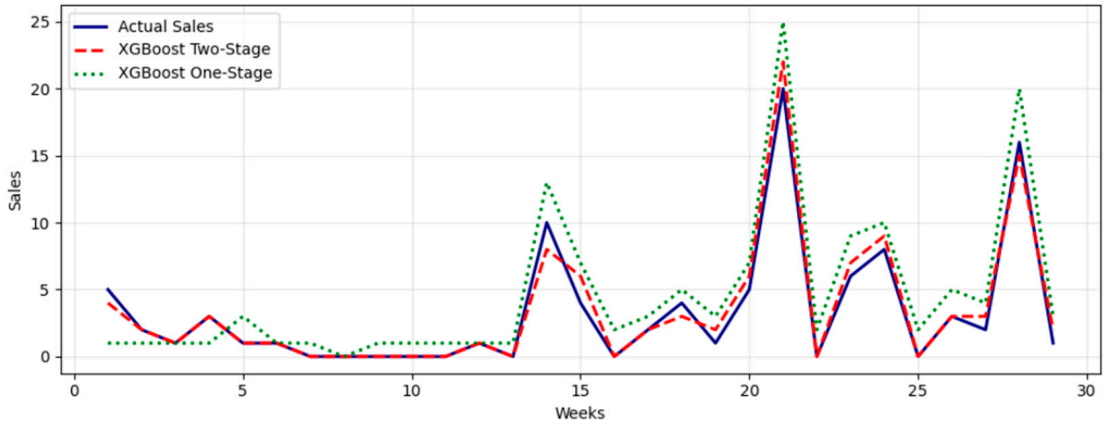

Figure 6. Actual sales versus Predicted sales for a representative product group.

Furthermore, to assess the robustness and stability of our best-performing model (Two-Stage XGBoost) over the forecasting horizon, we analyzed its weekly error performance (MAE and RMSE) across the 13-week holdout period. This analysis, presented in Figure 7, reveals the week-by-week evolution of forecast accuracy. While minor fluctuations in error metrics are observed, consistent with the inherent volatility and event-driven dynamics of the Iraqi fashion retail market, there is no evidence of a systematic performance degradation or an increasing trend in errors as the forecast horizon extends. This stability suggests that the model, leveraging rich exogenous features like Week (for seasonality), Special Event Label, and Temperature Anomaly Status, effectively adapts to intra-seasonal variations and maintains consistent predictive power throughout the planning cycle.

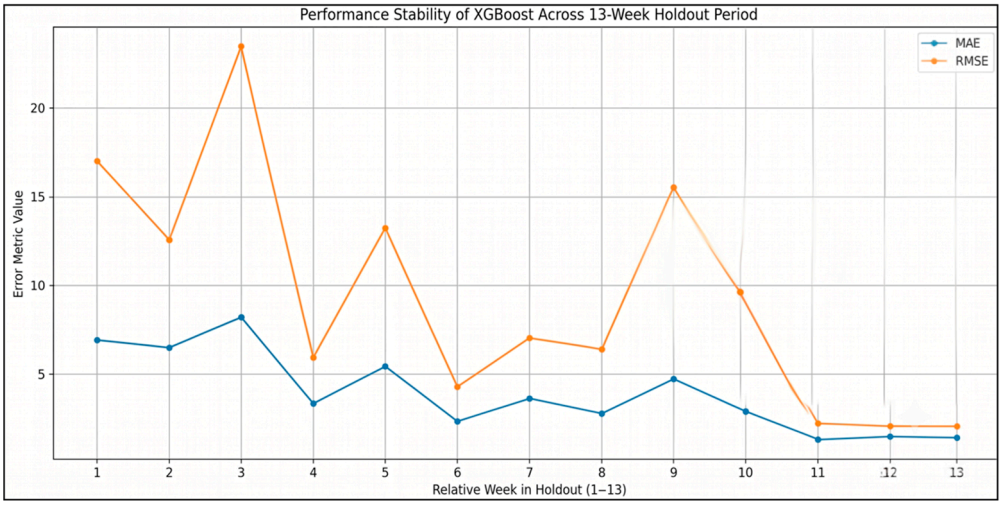

Figure 7. Performance Stability (MAE, RMSE) of Two-Stage XGBoost.

4.5. SHAP Feature Importance Analysis

To further elucidate the distinct mechanisms of our two-stage model and address the reviewer's request regarding feature contributions, this section presents the SHAP (SHapley Additive exPlanations) analysis for the best-performing XGBoost Two-Stage model. This analysis clearly differentiates which features primarily drive demand occurrence (classifier stage) versus those that influence demand magnitude (regressor stage).

4.5.1. Demand Occurrence (Classifier Stage)

Figure 8 presents the SHAP summary plot for the demand occurrence classifier, illustrating the global importance and impact direction of each feature. Table 6 further details the top 10 most influential features for predicting whether demand will occur in a given week.

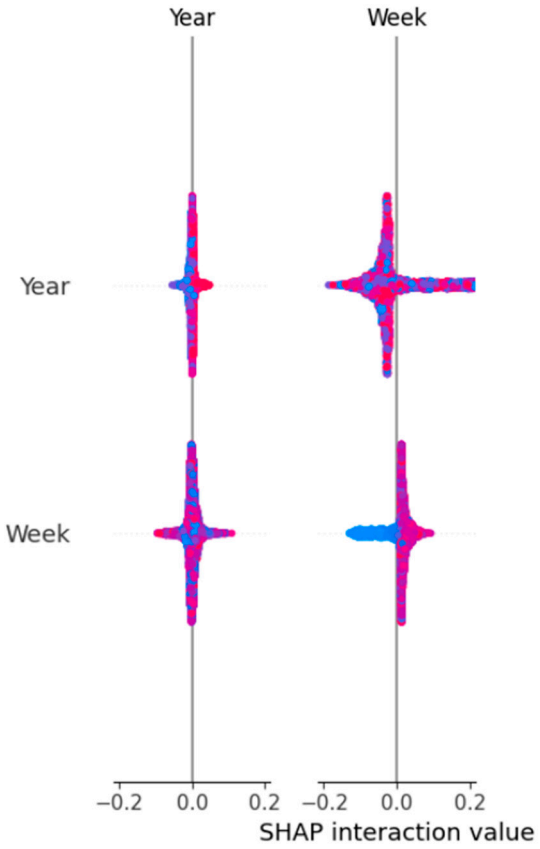

Figure 8. SHAP summary plot for Demand Occurrence (Classifier). Color indicates the feature value, with blue representing lower values and pink/red representing higher values.

Table 6. Top 10 Features for Demand Occurrence.

Feature	Importance
Previous Week Sales Quantity	0.093
Previous 2 Week Sales Quantity	0.056
Previous 3 Week Sales Quantity	0.042
Week	0.031
Previous 4 Week Sales Quantity	0.027
Total Option Count	0.026
Discounted Option Count	0.016
Topcategory Option Count	0.014
Special Event Label	0.013
Midcategory Option Count	0.011

The analysis reveals that lagged sales quantities (Previous Week Sales Quantity, Previous 2 Week Sales Quantity, Previous 3 Week Sales Quantity, and Previous 4 Week Sales Quantity) are the most critical features for predicting demand occurrence. This underscores

the model's reliance on recent historical activity as a primary signal for whether a product will be active in the upcoming week. Week (Week of Year) also exhibits significant importance, indicating the strong influence of seasonal patterns on the initiation of demand. Other influential features include various Option Count indicators (Total Option Count, Discounted Option Count, Topcategory Option Count, Midcategory Option Count), suggesting that product availability and pricing strategies play a role in triggering demand. The Special Event Label is also a key contributor, highlighting the distinct impact of cultural and religious events on demand initiation. While the SHAP interaction plot for Year and Week (as depicted in the full SHAP output, not explicitly shown here) indicates their individual contributions, their complex interaction effects with other features for demand occurrence prediction are generally minor, with most interaction SHAP values clustering around zero.

4.5.2. Demand Magnitude (Regressor Stage)

Figure 9 displays the SHAP summary plot for the demand magnitude regressor, detailing the features influencing the predicted sales quantity given that demand occurs. Table 7 lists the top 10 features for this stage.

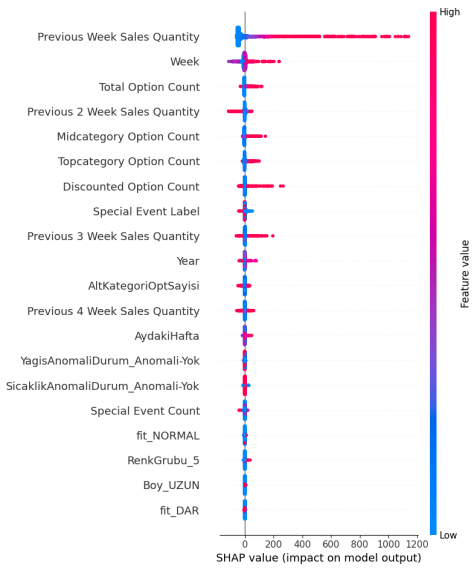

Figure 9. SHAP summary plot for Demand Magnitude (Regressor).

Table 7. Top 10 Features for Demand Magnitude.

Feature	Importance
Previous Week Sales Quantity	68.852
Week	7.501
Total Option Count	5.347
Previous 2 Week Sales Quantity	5.243
Midcategory Option Count	4.764
Topcategory Option Count	4.542
Discounted Option Count	2.961
Special Event Label	2.839
Previous 3 Week Sales Quantity	2.727
Year	2.100

For quantifying the sales volume, Previous Week Sales Quantity emerges as overwhelmingly the most dominant feature, with a SHAP importance value of 68.85. This demonstrates that the most recent actual sales volume is the strongest indicator for the expected sales volume in the next period, provided demand has occurred. Week (Week of Year) is also highly influential, reflecting the significant impact of seasonality on sales magnitude. Other key contributors include Total Option Count and Previous 2 Week Sales Quantity, which provide additional context on product offering and historical trends. Special Event Label also plays a crucial role, indicating that special events not only trigger demand but also significantly amplify its volume. This differentiated analysis highlights that while some features (e.g., lagged sales, week, special events) are important for both stages, their relative contribution and dominance vary significantly between predicting demand occurrence and estimating its magnitude.

5. Discussion

The empirical results indicate that the two-stage modeling strategy is particularly well suited to intermittent demand forecasting in the examined fashion retail setting. A likely explanation is that demand occurrence and demand magnitude reflect partially different underlying processes. By first classifying whether demand occurs and then estimating demand size conditional on positive sales, the two-stage framework appears to better accommodate the zero-inflated structure of the data than direct one-stage regression.

This interpretation is consistent with the nature of intermittent demand discussed in prior studies, where the separation of occurrence and magnitude has been identified as a pragmatic strategy for sparse demand environments. In this regard, the present findings align with recent work emphasizing decoupled forecasting structures for intermittent demand and support the view that such decomposition can provide empirical gains when zero-demand periods are frequent and operationally important.

Another notable finding is that machine learning models consistently outperformed the classical intermittent-demand benchmarks. While Croston-type methods remain important reference models, their smoothing-based structure is less capable of exploiting the rich exogenous information available in the present retail dataset, including product attributes, special event labels, price categories, and weather-related variables. The advantage of the machine learning models therefore appears to arise not only from algorithmic flexibility but also from their ability to integrate multiple contextual demand drivers within a unified predictive framework. In contrast, the two-stage machine learning framework employed in this study provides a more theoretically grounded approach for complex intermittent settings. By explicitly decoupling the demand occurrence (modeled via category) from the demand magnitude (modeled via regression), our framework directly addresses the theoretical premise that these are two distinct stochastic processes, each potentially influenced by different sets of features and relationships. The power of machine learning in each stage allows for the flexible integration of a wide array of exogenous variables, enabling a more nuanced and accurate representation of the underlying demand generation process. This explicit modeling of occurrence and magnitude, empowered by rich feature integration, constitutes a significant theoretical advantage over classical methods, which is empirically validated by the superior performance observed in our study. This decomposition allows the model to learn context-specific drivers for 'when' demand occurs and 'how much' is demanded, offering a more granular and theoretically consistent approach to intermittent demand forecasting.

This theoretical grounding is empirically supported by our detailed SHAP analysis (Section 4.5), which clearly elucidated the distinct feature contributions to demand occurrence versus demand magnitude. For the demand occurrence classifier, lagged sales

quantities (Previous Week Sales Quantity, Previous 2 Week Sales Quantity), Week (of year), and Special Event Label were identified as primary drivers (Table 6, Figure 8). This highlights the model's ability to leverage both historical momentum and critical contextual factors to predict if a sale will happen. For demand magnitude, while Previous Week Sales Quantity was overwhelmingly dominant (68.85% SHAP importance), reflecting the inherent inertia of sales, other features such as Week, Total Option Count, and Special Event Label still contributed significantly (Table 7, Figure 9). This indicates that while recent sales provide a strong baseline, contextual factors play a crucial role in fine-tuning the volume of demand. The ability of the two-stage model, powered by machine learning, to integrate and interpret these diverse features across both stages is key to its superior performance, enabling it to go beyond mere historical patterns to capture complex demand dynamics.

The comparison with the deep learning benchmarks provides an additional, critical insight. Although TFT and LSTM represent contemporary sequence modeling architectures, they performed significantly worse than all machine learning models, and even underperformed the simple Classical Historical Mean benchmark under WRMSSE. This suggests that greater architectural complexity does not necessarily guarantee superior forecasting performance in sparse retail demand settings [61]. Specifically, in contexts characterized by highly intermittent and sparse demand, deep learning models, with their extensive parameter sets, face increased challenges related to data scarcity and potential overfitting. Their capacity to learn complex hierarchical representations can be diminished when the underlying data patterns are not consistently present across observations, leading to a risk of memorizing noise rather than generalizable features. Furthermore, the computational demands and training efficiency of these complex architectures become a significant consideration in practical retail scenarios, especially where data volumes, while large overall, may be effectively "small" for individual attribute-defined product groups due to intermittency. In such 'small-data' or highly sparse environments, the overhead of training and tuning deep learning models may outweigh their predictive benefits, thus highlighting the continued competitiveness and robustness of carefully feature-engineered traditional machine learning approaches. In highly intermittent environments, carefully engineered explanatory variables and explicit occurrence modeling may remain more valuable than architecture-driven gains alone.

The Iraqi fashion retail context further strengthens the practical relevance of these findings. Demand in this setting is shaped by pronounced seasonality, extreme weather conditions, local cultural and religious events, and specific purchasing cycle dynamics such as salary timing. Under such conditions, the use of exogenous variables becomes particularly important, and the empirical advantage of the two-stage models suggests that context-aware machine learning systems may be highly beneficial for inventory and replenishment decisions in volatile retail markets.

From a managerial perspective, the granular insights provided by the SHAP analysis (Section 4.5) significantly enhance the actionable implications. By understanding which specific features primarily drive demand occurrence (e.g., the high importance of Week, Special Event Label, and various Option Counts) versus those that refine demand magnitude (where Previous Week Sales Quantity provides the baseline, but Week and Special Event Label still modulate the volume), retailers can develop more targeted and proactive strategies. For instance, the strong influence of Special Event Label on both occurrence and magnitude suggests that proactive planning around cultural events is paramount. Similarly, Total Option Count and Discounted Option Count being important for both stages highlights the impact of product assortment breadth and pricing strategies. This detailed feature importance allows for a more informed allocation of resources and strategic interventions. The findings imply that retailers operating in zero-heavy demand environments

may benefit from forecasting systems that explicitly distinguish between no-sale weeks and positive-sale weeks. Even when the final point forecast (derived from the two-stage model's hard classification) is zero, the underlying demand occurrence probability remains a valuable output. Specifically, the demand occurrence probability, which is a direct output of the two-stage model's classification component, offers a granular and actionable insight for inventory optimization. This probability can be directly leveraged to dynamically adjust safety stock levels: a low probability of demand occurrence could justify reducing safety stock, while a high probability might signal the need for a higher buffer, particularly for fast-moving or critical items. Furthermore, it can inform the setting of more responsive replenishment triggers and reorder points, allowing for more precise timing of orders. By incorporating this explicit probability, retailers can move beyond static inventory policies towards more adaptive strategies that significantly reduce overestimation during inactive periods and prevent stockouts during demand spikes, thereby improving replenishment planning, inventory allocation, and overall stock risk management. In fashion retail, where product life cycles are short and assortment decisions are highly time-sensitive, even moderate gains in forecast accuracy can translate into meaningful operational improvements.

Several limitations should also be acknowledged. First, the empirical analysis is based on a single retail company operating in one national market, which may limit the generalizability of the findings. Second, the benchmark focuses on a selected set of machine learning and deep learning models rather than an exhaustive set of modern architectures. Third, the evaluation is based on point forecasting metrics and does not consider probabilistic forecasting performance. Future research may therefore extend this benchmark to larger retail panels, additional deep learning architectures, and probabilistic forecasting settings.

6. Conclusions

This study presented a structured empirical comparison of one-stage and two-stage machine learning frameworks for intermittent demand forecasting in fashion retail. Using weekly product range sales data from an international fashion retailer operating in Iraq, across 64 unique attribute-defined product group time series, the analysis benchmarked Random Forest, Gradient Boosting, XGBoost, and LightGBM models against classical intermittent-demand methods and additional deep learning baselines.

The results showed that while the Two-Stage XGBoost achieved the best overall performance (WRMSSE: 0.332), the superiority of two-stage machine learning models over their one-stage counterparts was not consistent across all algorithms under WRMSSE, which served as the primary evaluation metric for this zero-inflated forecasting setting.

The findings therefore indicate that explicitly separating demand occurrence from demand magnitude can improve forecast accuracy in intermittent retail demand environments. Furthermore, a detailed SHAP value analysis provided critical insights into the distinct feature contributions for each stage, highlighting the specific drivers of demand occurrence (e.g., lagged sales and special events) and demand magnitude (e.g., immediate past sales volume), offering a transparent and actionable framework for inventory management.

The study also showed that machine learning models (both one-stage and two-stage configurations) significantly outperformed all classical intermittent-demand benchmarks and deep learning baselines included in the analysis. Notably, deep learning models performed worse than even the simple Classical Historical Mean benchmark under WRMSSE, highlighting their limitations in this highly intermittent context. This result suggests that, in sparse retail demand settings, a feature-engineered two-stage learning framework may

remain more effective than both conventional intermittent-demand estimators and more complex sequence modeling architectures.

Beyond its empirical benchmarking contribution, the study provides practical relevance through its focus on the Iraqi fashion retail market, where demand is shaped by strong seasonality, weather variability, cultural event effects, and volatile purchasing patterns. These findings highlight the importance of context-aware forecasting systems for improving inventory planning and replenishment decisions in under-researched and operationally challenging retail environments. Specifically, the demand occurrence probability, a direct output of our two-stage framework, offers a concrete mechanism for optimizing safety stock settings and dynamically adjusting replenishment triggers, thereby enabling more adaptive and efficient inventory management strategies.

Future research may extend this benchmarking framework by incorporating additional forecasting architectures, probabilistic loss functions, and broader retail panels to further assess the robustness and generalizability of two-stage forecasting strategies under intermittent demand. It is important to acknowledge that advanced inventory management and stock risk mitigation strategies, including the optimization of safety stock levels and service rates, often rely on quantile forecasts. Therefore, future work should specifically focus on extending this framework to include probabilistic forecasting, which would allow for a more direct and robust integration with sophisticated inventory optimization models.

Author Contributions: Conceptualization, B.Y.S., Ö.F.B. and F.K.; methodology, B.Y.S. and Ö.F.B.; software, B.Y.S.; validation, B.Y.S. and Ö.F.B.; formal analysis, B.Y.S.; investigation, B.Y.S.; resources, Ö.F.B. and F.K.; data curation, B.Y.S.; writing—original draft preparation, B.Y.S.; writing—review and editing, B.Y.S., Ö.F.B. and F.K.; visualization, B.Y.S.; supervision, Ö.F.B. and F.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not available due to strict commercial confidentiality agreements with the data provider. Therefore, the raw data cannot be shared or made available upon request.

Acknowledgments: The authors gratefully acknowledge the support of the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the BIDEB 2211 Domestic Graduate Scholarship Program.

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

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
GBM	Gradient Boosting Machine
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
NOS	Never Out of Stock
RF	Random Forest
RMSE	Root Mean Squared Error
SBA	Syntetos–Boylan Approximation
SVR	Support Vector Regression
TFT	Temporal Fusion Transformer
XGBoost	Extreme Gradient Boosting
WRMSSE	Weighted Root Mean Squared Scaled Error

References

1. Fildes, R.; Adam, S.; Allen, G. The evolution of forecasting research: A review of the last 40 years. *Int. J. Forecast.* **2022**, *38*, 1–20.
2. Bohanec, M.; Kljajić Borštnar, M.; Robnik-Šikonja, M. Explaining machine learning models in sales predictions. *Comput. Ind.* **2017**, *82*, 103–114. [CrossRef]
3. Liu, X.; Ichise, R. Food sales prediction with meteorological data: A case study of a Japanese chain supermarket. *Expert Syst. Appl.* **2017**, *80*, 58–67. [CrossRef]
4. Xue, J.; Jiang, X.; Cai, Y.; Yuan, X. Research on demand forecasting of retail supply chain emergency logistics based on NRS-GA-SVM. *Comput. Ind. Eng.* **2018**, *125*, 300–310. [CrossRef]
5. Braun, S.; Potharst, R.; Heiser, W.J. 24-hours demand forecasting based on SARIMA and support vector machines. *J. Forecast.* **2014**, *33*, 456–474. [CrossRef]
6. Jain, R.; Mehrotra, S.; Roy, A. Sales forecasting for retail chains. *Int. J. Bus. Anal.* **2015**, *3*, 14–25. [CrossRef]
7. Loureiro, A.; Torgo, L.; Soares, C. Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decis. Support Syst.* **2018**, *114*, 81–93. [CrossRef]
8. Sun, Z.L.; Choi, T.M.; Au, K.F.; Yu, Y. Sales forecasting using extreme learning machine with applications in fashion retailing. *Decis. Support Syst.* **2008**, *46*, 411–419. [CrossRef]
9. Priyadarshi, R.; Patnaik, S.; Panda, S. Demand forecasting at retail stage for selected vegetables: A performance analysis. *Int. J. Agric. Econ.* **2019**, *4*, 31–45. [CrossRef]
10. Turatti, D.E.; Mendes, F.H.d.P.e.S.; Mazzeu, J.H.G. Combining Volatility Forecasts of Duration-Dependent Markov-Switching Models. *J. Forecast.* **2025**, *44*, 1195–1210. [CrossRef]
11. Huber, J.; Stuckenschmidt, H. Daily retail demand forecasting using machine learning with emphasis on calendric special days. *Eur. J. Oper. Res.* **2020**, *287*, 507–518. [CrossRef]
12. Seyedan, M.; Mafakheri, F. Cluster-based demand forecasting using Bayesian model averaging: An ensemble learning approach. *Comput. Ind. Eng.* **2022**, *167*, 107999. [CrossRef]
13. Ensafi, Y.; Hassanzadeh Amin, S.; Zhang, G.; Shah, B. Time-series forecasting of seasonal items sales using machine learning—A comparative analysis. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100058. [CrossRef]
14. Nacar, S.; Erdebilli, B. Makine öğrenmesi algoritmaları ile satış tahmini. *J. Turk. Artif. Intell.* **2021**, *5*, 14–28. [CrossRef]
15. Lingelbach, F.; Das, D.; Lee, C. Demand forecasting using ensemble learning for effective scheduling of logistic orders. *J. Logist.* **2021**, *28*, 45–58. [CrossRef]
16. Van Steenbergen, A.; Mes, M. Forecasting demand profiles of new products. *Oper. Res. Perspect.* **2020**, *7*, 100143. [CrossRef]
17. Arif, M.; Sany, S.; Nahin, K.; Rabby, M. Comparison study: Product demand forecasting with machine learning for shop. *Int. J. Forecast.* **2020**, *36*, 78–91. [CrossRef]
18. Huber, J.; Thorsten, M.; Wiedmann, C. Cluster-based hierarchical demand forecasting for perishable goods. *Supply Chain Manag. Rev.* **2017**, *12*, 34–45. [CrossRef]
19. Syntetos, A.A.; Boylan, J.E.; Croston, J.D. The accuracy of intermittent demand estimates. *Int. J. Forecast.* **2005**, *21*, 303–314. [CrossRef]
20. Petropoulos, F.; Spiliotis, E.; Makridakis, S. Forecasting: Theory and practice of benchmarking and evaluation. *Int. J. Forecast.* **2022**, *38*, 1234–1250. [CrossRef]
21. Giannopoulosa, E.; Spiliotis, E.; Petropoulos, F. Machine learning algorithms in intermittent and lumpy demand forecasting: A review. *Comput. Oper. Res.* **2025**, *162*, 106450. [CrossRef]
22. Croston, J.D. Forecasting and stock control for intermittent demands. *J. Oper. Res. Soc.* **1972**, *23*, 289–304. [CrossRef]
23. Ramanathan, R. The Aggregate-Disaggregate Intermittent Demand Approach (ADIDA) for forecasting. *Int. J. Prod. Econ.* **2006**, *103*, 675–684. [CrossRef]
24. Boylan, J.E.; Syntetos, A.A. *Intermittent Demand Forecasting: Context, Methods and Applications*; Wiley: Online, 2021. [CrossRef]
25. Aydın, M.; Yazıcıoğlu, H. Yapay sinir ağları ile talep tahmini. *Perakende. Bilgi İletişim Teknol. Derg.* **2019**, *4*, 50–60.
26. Sekban, J. Applying Machine Learning Algorithms in Sales Prediction. Master's Thesis, Kadir Has University, Istanbul, Türkiye, 2019.
27. Abdelfatah, O.S.M. Machine Learning Approaches for Improving Demand Forecasting Accuracy in Retail Supply Chains. *SocArXiv* **2023**. Available online: https://osf.io/preprints/socarxiv/4z9be_v1 (accessed on 10 March 2026).
28. Karim, S. Artificial Intelligence-Enhanced Predictive Analytics for Demand Forecasting in U.S. Retail Supply Chains. *ASRC Procedia Glob. Perspect. Sci. Scholarsh.* **2025**, *1*, 43. [CrossRef]
29. Wang, C. Dynamic Dual-Phase Forecasting Model for New Product Demand Using Machine Learning and Statistical Control. *Mathematics* **2025**, *13*, 1613. [CrossRef]
30. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. M5 accuracy competition: Results, findings, and conclusions. *Int. J. Forecast.* **2022**, *38*, 10–22. [CrossRef]
31. Kolassa, S. Why the "best" point forecast depends on the error or accuracy measure. *Int. J. Forecast.* **2020**, *36*, 208–211. [CrossRef]

32. Giri, C.; Chen, Y. Deep Learning for Demand Forecasting in the Fashion and Apparel Retail Industry. *Forecasting* **2022**, *4*, 565–581. [CrossRef]
33. Ahmadov, Y.; Helo, P. Deep learning-based approach for forecasting intermittent online sales. *Discov. Artif. Intell.* **2023**, *3*, 45. [CrossRef]
34. Spiliotis, E.; Assimakopoulos, V.; Petropoulos, F. Forecasting retail sales: A large-scale empirical comparison of statistical and machine learning methods. *Int. J. Forecast.* **2022**, *38*, 1251–1267.
35. Anitha, S.; Neelakandan, S. A Demand Forecasting Model Leveraging Machine Learning to Decode Customer Preferences for New Fashion Products. *J. Forecast. Artif. Intell. Rev.* **2024**, *57*, 8425058. [CrossRef]
36. Kaneko, M.; Yada, K. A deep learning approach for the prediction of retail store sales. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016; pp. 531–537. [CrossRef]
37. Saha, P.; Gudheniya, N.; Mitra, R.; Das, D. Demand Forecasting of a Multinational Retail Company using Deep Learning Frameworks. In Proceedings of the IFAC-PapersOnLine, Virtual, 11–15 July 2022; Volume 55, pp. 100–105. [CrossRef]
38. Aci, M.; Doğansoy, G.A. Makine öğrenmesi ve derin öğrenme yöntemleri kullanılarak e-perakende sektörüne yönelik talep tahmini. *J. Fac. Eng. Archit. Gazi Univ.* **2022**, *37*, 1325–1339. [CrossRef]
39. Li, Y.; Li, Y.; Wang, Y. A hybrid predictive framework combining merchandise image data with transaction records to forecast consumer color preferences. *Decis. Support Syst.* **2020**, *131*, 113264.
40. de Castro Moraes, F.; Yuan, X.-M.; Chew, E.P. Hybrid convolutional long short-term memory models for sales forecasting in retail. *J. Forecast.* **2024**, *43*, 1278–1293. [CrossRef]
41. Habib, M.R.; Hossain, M.S. Revolutionizing Wind Power Prediction—The Future of Energy Forecasting with Advanced Deep Learning and Strategic Feature Engineering. *Energies* **2024**, *17*, 1215. [CrossRef]
42. Oliveira, A.M.; Ramos, M. Evaluating the Effectiveness of Time Series Transformers for Demand Forecasting in Retail. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *20*, 2728. [CrossRef]
43. Zhang, Y.; Li, X.; Chen, T. Forecasting Sales in Live-Streaming Cross-Border E-Commerce in the UK Using the Temporal Fusion Transformer Model. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 92. [CrossRef]
44. Yücesan, M. Forecasting Monthly Sales of White Goods Using Hybrid ARIMAX and ANN Models. *Atatürk Üniversitesi Sos. Bilim. Enstitüsü Derg.* **2018**, *22*, 2603–2617.
45. Ma, S.; Fildes, R.; Huang, T. Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *Int. J. Forecast.* **2016**, *38*, 1343–1365. [CrossRef]
46. Sönmez, O.; Zengin, K. Yiyecek ve içecek işletmelerinde talep tahmini: Yapay sinir ağları ve regresyon yöntemleriyle bir karşılaştırma. *Eur. J. Sci. Technol.* **2019**, *17*, 302–308. [CrossRef]
47. Türk, E.; Kiani, F. Yapay Sinir Ağları ile Talep Tahmini Yapma: Beyaz Eşya Üretim Planlaması için YSA Uygulaması. *İstanbul Sabahattin Zaim Üniversitesi Fen Bilim. Enstitüsü Derg.* **2019**, *1*, 30–37.
48. Arslankaya, S. Bir lojistik firmasında zaman serileri analizi ve yapay sinir ağları ile talep tahmininin karşılaştırılması. *Eur. J. Sci. Technol.* **2019**, *16*, 239–245. [CrossRef]
49. Güven, İ.; Şimşir, F. Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods. *Comput. Ind. Eng.* **2020**, *147*, 106678. [CrossRef]
50. Kaya, S.; Yildirim, O. A prediction model for automobile sales in Turkey using deep neural networks. *Endüstri Mühendisliği* **2020**, *31*, 57–74. Available online: https://dergipark.org.tr/tr/pub/endustrimuhendisligi/article/669930 (accessed on 10 March 2026).
51. del Real, A.J.; Dorado, F.; Durán, J. Energy Demand Forecasting Using Deep Learning: Applications for the French Grid. *Energies* **2020**, *13*, 2242. [CrossRef]
52. Liu, L.; Chen, R.C. A novel passenger flow prediction model using deep learning methods. *Transp. Res. Part C Emerg. Technol.* **2017**, *84*, 74–91. [CrossRef]
53. Zhu, K.; Xun, P.; Li, W.; Gao, X.; Zhou, R. Prediction of Passenger Flow in Urban Rail Transit Based on Big Data Analysis and Deep Learning. *IEEE Access* **2019**, *7*, 131872–131881. [CrossRef]
54. Rožanec, J.M.; Gams, M.; Dvornik, M. Reframing Demand Forecasting: A Two-Fold Approach for Lumpy and Intermittent Demand. *Sustainability* **2022**, *14*, 9295. [CrossRef]
55. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
56. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
57. 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. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 3146–3154.
58. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
59. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

60. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
61. Kolassa, S. Evaluating predictive count data distributions in retail sales forecasting. *Forecasting* **2026**, *8*, 24. [CrossRef]
62. Wu, J.; Liu, H.; Yao, X.; Zhang, L. Unveiling consumer preferences: A two-stage deep learning approach to enhance accuracy in multi-channel retail sales forecasting. *Expert Syst. Appl.* **2024**, *257*, 125066. [CrossRef]