Demand Forecasting Strategy for a Water Bottling Company: A Segment-Based Approach

by

Sebastián Villegas
Bachelor of Science in Engineering, Pontifical Catholic University of Chile, 2016

and

Zhaoxia Huang
Bachelor of Arts in English Language and Literature, Xinyang Normal University, 2004

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2025

Signature of Author: _____
Department of Supply Chain Management
May 2, 2025

Signature of Author: _____
Department of Supply Chain Management
May 2, 2025

Certified by: _____ May 2, 2025
Dr. Ilya Jackson
Research Scientist
Capstone Advisor

Certified by: _____
Dr. Benedict Jun Ma
Postdoctoral Associate
Capstone Co-Advisor

Accepted by: _____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Demand Forecasting Strategy for a Water Bottling Company: A Segment-Based Approach

by

Sebastián Villegas
Bachelor of Science in Engineering, Pontifical Catholic University of Chile, 2016

and

Zhaoxia Huang
Bachelor of Arts in English Language and Literature, Xinyang Normal University, 2004

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2025

Signature of Author: _____

Department of Supply Chain Management
May 2, 2025

Signature of Author: _____

Department of Supply Chain Management
May 2, 2025

Certified by: _____

Dr. Ilya Jackson
Research Scientist
Capstone Advisor

Certified by: _____

Dr. Benedict Jun Ma
Postdoctoral Associate
Capstone Co-Advisor

Accepted by: _____

Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Demand Forecasting Strategy for a Water Bottling Company: A Segment-Based Approach

by

Sebastián Villegas

and

Zhaoxia Huang

Submitted to the Program in Supply Chain Management

on May 2, 2025 in Partial Fulfillment of the

Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

This project introduces a segment-based forecast strategy to enhance tactical demand planning for a leading water bottling company. SKUs were segmented based on key demand pattern characteristics—average demand, coefficient of variation, and intermittency—resulting in four distinct clusters that reflect varying levels of demand predictability. A variety of forecasting models, including ARIMA, Exponential Smoothing, XGBoost, TiDE, Croston, and N-BEATS, were evaluated within each cluster using performance metrics such as MAPE, APE, MAE, and RMSE. The analysis also incorporates exogenous variables, such as natural events and holiday periods, to evaluate their impact on forecast accuracy. Testing results showed that implementing a Best-Fit Model strategy improved forecast accuracy by 1.56 percentage points in Absolute Percentage Error (APE). The results indicate that statistical time-series and machine learning models perform well in more stable and high-volume SKU clusters. In contrast, highly intermittent and low-volume SKUs remain challenging to forecast; thus, a more collaborative planning approach such as CPFR is recommended. This segmentation-driven approach improves forecast accuracy, enhances planning efficiency, and provides a scalable and interpretable framework for demand forecasting across the product portfolio.

Capstone Advisor: Dr. Ilya Jackson
Title: Research Scientist
Capstone Co-Advisor: Dr. Benedict Jun Ma
Title: Postdoctoral Associate

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to my advisors, Dr. Ilya Jackson and Dr. Benedict Jun Ma, for their invaluable guidance, encouragement, and expertise. Their unwavering support ensured that our capstone project remained focused and impactful. A heartfelt thanks goes to our sponsor company for providing us with the opportunity to work on this capstone project. Their generosity in granting access to high-quality data laid a solid foundation for our research and allowed us to achieve meaningful outcomes. I am profoundly grateful to my capstone partner, Zhaoxia, for her dedication, creativity, and companionship throughout this journey. Her contributions were instrumental in driving us forward and inspiring me to strive for excellence. Special thanks are due to Pamela for her meticulous attention to detail and consistent encouragement during the writing process, making a significant difference in the development of our work. I would also like to acknowledge the unwavering support of my family, whose belief in me has been a source of strength and inspiration. Without their constant encouragement, I would not be where I am today.

Sebastian Villegas

I would like to express my sincere appreciation to our advisors, Dr. Ilya Jackson and Dr. Benedict Jun Ma, for their unwavering support, expert guidance, and insightful feedback throughout this capstone project. Their expertise and guidance were instrumental in shaping both the direction and depth of the work. I am also deeply grateful to our sponsoring company for the opportunity to engage in a meaningful real-world challenge, for providing high-quality data, and for generously sharing their time during weekly meetings to address our many questions. My heartfelt thanks go to Pamela, our writing coach, whose encouragement, keen eye for detail, and constructive feedback have significantly strengthened the clarity and impact of our writing. I am especially thankful to my capstone partner, Sebastian, who is now a good friend, for his collaboration, critical thinking, consistent contributions, and uplifting spirit throughout this journey. Lastly, I want to thank my family for their unwavering support and belief in me, especially my husband, Callum, who first sparked my interest in supply chain management and Python. Each of them played a meaningful role in the success of this capstone project, and it is their continued support that has made this special journey so rewarding.

Zhaoxia Huang

TABLE OF CONTENTS

1. INTRODUCTION

An effective forecasting strategy helps companies maintain optimal inventory levels to meet customer demand without overstocking. Accurate demand forecasts allow efficient Sales and Operations Planning (S&OP) planning and better allocation of resources, including labor, materials, and capital. Companies with reliable forecasts can outperform competitors by delivering the right products at the right time and cost. Accurate forecasting reduces the risks of overproduction, excess inventory, and stockouts. It enables businesses to operate efficiently, meet customer needs, optimize resources, and improve profitability while minimizing risks from misaligned supply and demand. Demand forecasting is crucial across industries but can be complex, particularly for consumer goods packaging companies with diverse products and customers.

1.1 Background

Our partner company is currently facing challenges in its demand forecasting processes. As a major beverage company in the United States, it has established itself as a low-cost leader in the industry. With a broad manufacturing footprint across the country, this company serves as a key supplier to diverse customers.

1.2 Motivation

With hundreds of customers, SKUs, and thousands of ship-to locations, the company is looking to expand production with more plants. Therefore, the company needs to improve its demand forecasting strategy to support its growth and maintain its status as the low-cost leader in the private-label bottled water industry. The company currently uses a digital supply chain platform to generate statistical forecasting as its tactical forecast within the next 13 weeks. However, there are some limitations to the current platform:

- No advanced machine learning forecasting algorithms.
- No ability to correlate causal factors.
- Limitations with parameter tuning.
- Challenging to manage increasing product mixes and business complexities.

The company is trying to use demand sensing, a forecasting technique that leverages real-time data and analytics to identify short-term changes in customer demand (Chase, 2013). The current forecasting process is overseen by a team of 16 planners whose efforts heavily rely on manual input. In 2024, based on the demand planners' qualitative inputs, 120k out of 189k overrides were made over four weeks for a change of less than 10%. For 2024, the statistical forecast accuracy was 71.8%; with manual interventions, the resultant forecast accuracy was 74.4%. The company is trying to leverage Artificial Intelligence (AI), Machine Learning (ML), and causal factor analysis to improve statistical forecasts. These techniques can be incorporated into the demand sensing process to reduce the number of manual interventions. One of the company's goals is to improve the current statistical forecast accuracy by at least 2% using ML and causal factors correlation.

1.3 Problem Statement and Key Questions

To address these challenges, this project answers the question: “What strategy should our sponsor employ to improve the effectiveness and efficiency of the demand forecasting process?”

The sponsor company asked the capstone team to focus on its plants across the US. The company has three forecast time horizons (Tactical, Strategic, and Long-Range), and it requested that this project to focus on the tactical forecast, which has a time horizon of the next 13 weeks. The scope of the project entailed critiquing the current approach and recommending a new segmentation methodology and best-fit forecasting strategy for every segment.

To address the problem, we focus on the following tasks:

1. Evaluate the forecasting process to determine whether a statistical approach best fits the organization’s entire product line.
2. Identify product segments that may not be suitable for statistical forecasts, such as newly launched products.
3. Determine which product segments are suitable for statistical forecasts.
4. Explore machine learning forecast models to improve the accuracy of statistical forecasts.
5. Investigate the possibility of implementing a hybrid approach, such as statistical forecasts combined with Collaborative Planning, Forecasting, and Replenishment (CPFR).
6. Test the different forecasting methods and compare accuracy metrics to select the best-fit model for each product segment.

We use Python and libraries like Pandas, Scikit-Learn, and Darts as the main tools for our analysis and model building.

1.4 Project Goals and Expected Outcomes

The project's overall goal is to identify the best forecast strategy to improve the effectiveness and efficiency of the company's demand forecasting process.

The expected outcomes include:

1. Product segmentations based on products' demand patterns such as average demand volume, CoV (Coefficient of Variability), and demand intermittency.
2. Improving the current statistical forecast accuracy by testing different statistical and machine learning forecasting models for each product segment.
3. Comparing forecast test results using metrics such as MAPE, MAE, APE and RMSE and selecting the best-fit model for forecastable product segmentations.

4. Recommending tailored forecasting strategies for different SKU segments.

To design an improved forecasting strategy for SKU demand behavior, it was essential to review the current best practices in both industry and research. The following section summarizes demand forecasting practices, including traditional, machine learning, and collaborative approaches.

2. STATE OF THE PRACTICE

Our research focuses on best practices for demand forecasting and the benchmarks of forecast accuracy commonly used in the fast-moving consumer goods (FMCG) sector. The literature review covers various demand forecasting models, including statistical forecasting methods such as Time Series Analysis, ARIMA, Exponential Smoothing, and Causal Analysis. It also compares forecasting metrics, including measurements like Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Absolute Percentage Error (APE), and Root Mean Square Error (RMSE). The topics of leveraging ML and AI in demand forecasting are also researched. Finally, exogenous factors impacting FMCG demand, especially the demand for packaged beverages, are also discussed.

2.1 Forecasting Models

Forecasting demand in push-based (also known as forecast-driven) production systems involves both scientific statistical techniques and intuitive art. Particularly for FMCG, which is characterized by short shelf life and rapidly changing consumer preferences, predicting demand is complex. Traditional methods relying solely on historical sales data often fall short. More sophisticated approaches like causal models, time series analysis, and advanced statistical techniques such as multiple linear regression can potentially improve forecasting accuracy by accounting for diverse influencing factors like marketing activities, consumer behaviors, and external conditions (Farizal et al., 2021).

This section reviews the commonly used statistical forecasting models: Time Series Analysis, Exponential Smoothing, ARIMA, and Causal Analysis. Statistical forecasting models utilize historical data to predict future values in a time series. The main types of statistical forecasting models encompass time series and regression models. (Abraham & Ledolter, 2005).

2.1.1 Time Series Analysis

A time series refers to any data observed in sequence over time. Forecasting time series data aims to predict how the observed sequence will evolve. The simplest forecasting methods focus solely on the predicted variable without attempting to identify the factors influencing its behavior. These methods typically extrapolate trends and seasonal patterns, but they overlook external factors like marketing efforts, competitor activities, and changes in economic conditions (Hyndman & Athanasopoulos, 2018).

Time series is a widely used forecasting method for items with a long history of demand records, such as most products in the FMCG sector. Our sponsor is currently using different time series in its statistical forecasting models. Time series is suitable for mid-range forecasts.

2.1.2 Exponential Smoothing

Compared to time series analyses, Exponential Smoothing treats data differently by applying less weight to older observations. This approach assumes that the newer observations are more influential than older observations. Data becomes less relevant to future forecasts as it ages, so its weight decreases exponentially. This technique helps to smooth out random variations in the data, making it easier to identify trends and patterns.

Exponential Smoothing models are widely used in various industries for demand forecasting due to their ability to capture trends and seasonality. Seasonal fluctuations posed unique challenges for FMCG inventory management. Exponential Smoothing Models can be used to identify trends and seasonal patterns in seasonal FMCG products.

2.1.3 ARIMA (Autoregressive Integrated Moving Average)

Autoregressive Integrated Moving Average (ARIMA) is a well-known linear model for forecasting linear time series data. The model has three parameters: (p, d, q). p is for autoregressive terms. d is for non-seasonal differences, and q is related to moving average terms. A major benefit of the Seasonal ARIMA (SARIMA) model is its capability to account for seasonal patterns in both stationary and non-stationary time series (Falatouri et al., 2022)

ARIMA models offer an alternative method for time series forecasting. Alongside exponential smoothing, ARIMA is one of the two most used forecasting techniques. These methods complement each other; exponential smoothing focuses on modeling trends and seasonality, while ARIMA aims to capture the autocorrelations within the data (Hyndman & Athanasopoulos, 2018).

2.1.4 Causal Analysis

Forecasting demand for FMCG products can be challenging because sales patterns are influenced by factors like marketing activities such as advertising, promotion, and weather conditions rather than just the passage of time. According to a study of 50 companies, many forecasts do not perform as well as expected, mainly because companies rely heavily on historical sales data and often overlook important predictors such as promotions and weather. Causal models, which explore cause-effect relationships among variables, can improve forecast accuracy. These models use linear programming, simulation, and multiple linear regression techniques (Farizal et al., 2021).

When performing causal analysis models, it is important not to confuse correlation with causation and forecasting. While a variable “x” might help predict a variable “y”, it does not necessarily mean that “x” is the cause of “y”. “x” may cause “y”, but it could also be that “y” influences “x”, or their relationship may be more complex than straightforward causation (Hyndman & Athanasopoulos, 2018).

Regression models are helpful when forecasters try to incorporate demand drivers such as promotions, weather, and consumer price index (CPI) into the forecast models. Regression models play a crucial role in forecasting by allowing analysts to quantify the impact of various demand drivers. For example, promotions can lead to short-term increases in demand, and regression analysis enables forecasters to assess the sensitivity of sales to these promotional activities. By incorporating data on past promotions, planners can predict how similar future promotions might influence consumer behavior. However, creating a reliable causal model can be difficult due to challenges in identifying key factors or obtaining necessary data (Farizal et al., 2021).

While classical regression techniques remain important for understanding demand drivers, modern machine learning methods offer greater flexibility in handling complex, nonlinear relationships between exogenous variables and demand outcomes. Machine learning models can automatically detect hidden patterns, interactions, and nonlinear effects that traditional models might miss, especially when working with large volumes of external data such as detailed weather, event calendars, or promotional histories.

2.2 Machine Learning

2.2.1 Machine Learning Methods

Mitra et al. (2024) highlight recent advancements in AI and ML frameworks, particularly with the advent of advanced deep learning algorithms and the abundance of data.

As Feng and Foster (2023) note, machine learning is a process that builds models using large amounts of available data. The greater the amount of data, the better the results that can be achieved. However, machine learning is a broad concept with a wide range of methodologies that can be applied. Mitra et al. (2024) present various projects in different industries, highlighting the corresponding authors and methodologies. A summary of their work is shown in Table 1.

Table 1: Different Methods of ML Applied in Industry (Mitra et al. 2024).

Method	Industry	References
HWES & Hadoop	Walmart stores	Harsoor and Patil (2015)
K-means clustering	Fashion	Tehrani and Ahrens (2016)
LASSO	Retail, Tire Industry	Ma et al. (2016), Sagaert et al. (2018)
CNN	E-commerce	Pan and Zhou (2020), Zhao and Wang (2017)
WaveNet CNN	Grocery	Kechyn et al. (2018)
Deep NN	Fashion	Loureiro et al. (2018)
XGBoost & LightGBM	Product Marketing	Liang et al. (2019)
Dynamic ANN	Food	Adithya Ganesan et al. (2019)
Stacking	Rossmann Store	Pavlyshenko (2019)
Deep-embedded LSTM	Pharmacy	Kraus et al. (2020)
CNN-based meta-learner	Retail	Ma and Fildes (2021)
LSTM	Furniture Industry	Pliszczuk et al. (2021)
Aggregated LightGBM	Steel Industry	Balamwar et al. (2022)
ARIMAX & NN	Steel Industry	Feizabadi (2022)
Multimodal Quasi-AutoRegression	Fashion	Papadopoulos et al. (2022)
Clustering & Aggregated DL	Food and Beverage	Mitra et al. (2024)

Another study by Tseng and Turkmen (2024) also analyzes various machine learning models but focuses on the explainability of these models. According to Tseng and Turkmen (2024), understanding what happens inside the "black box" allows the results to be explained based on the input data. Table 2 illustrates the descriptions, advantages, and disadvantages of various ML and statistical models.

Table 2: Machine Learning Forecasting Models Advantages and Disadvantages.

Model	Type	Description	Advantages	Disadvantages
XGBoost	Machine Learning	Boosted trees that model nonlinear relationships using lagged features and optional covariates.	- Handles nonlinearity well - Accepts covariates - Strong performance with good features	- Requires tuning - No internal notion of time - Needs good lag/feature design
TiDE	Deep Learning	Transformer-inspired model decoding complex temporal relationships using full historical and future context.	- Captures long-term dependencies - Supports multivariate, missing data - Good for complex patterns	- Requires a lot of data - Slower to train - Can overfit if not tuned
TFT (Temporal Fusion Transformer)	Deep Learning	Attention-based model using interpretable gating mechanisms and temporal embeddings for time series.	- Powerful & interpretable - Handles static, historic & future covariates - Designed for multivariate data	- Highly resource intensive - Needs a lot of tuning - Slower training
N-BEATS	Deep Learning	Neural basis expansion analysis for interpretable time series forecasting.	- Captures nonlinear patterns and complex dynamics - No need for feature engineering - Can be trained globally across multiple SKUs - Works well even without covariates	- Requires more training data than classical models - Can overfit on short or noisy time series - Slower to train than XGBoost - May not outperform simpler models on stable/low-variance SKUs

2.2.2 Accuracy in Prediction in Machine Learning

Another important factor when selecting machine learning methods is the overall accuracy of prediction and the role of feature engineering. Chuang et al. (2021) highlight that feature engineering often has a greater impact on forecast accuracy than the specific choice of machine learning algorithm. This finding supports our approach of carefully incorporating external features, such as holiday flags, to enhance model performance where appropriate.

Furthermore, existing literature suggests that no single forecasting methodology is universally optimal across all demand patterns. Mitra et al. (2024) advocate for a mixed framework approach, combining machine learning models with demand segmentation techniques to better tailor forecasts to different behavioral clusters. Their findings reinforce the idea that model flexibility, rather than reliance on a single technique, is critical for effective demand planning across a diverse SKU portfolio.

In line with these insights, we applied K-Means clustering to group SKUs based on demand volume, variability, and intermittency, ensuring that the most appropriate forecasting models could be assigned to each

cluster. This cluster-based strategy recognizes that different segments of SKUs require different modeling solutions to optimize forecast accuracy, balancing model selection with demand behavior characteristics at a tactical planning level.

2.3 Collaborative Planning, Forecasting, and Replenishment (CPFR)

CPFR is a commonly used forecasting strategy in the FMCG industry. It was pioneered by Walmart in the 1990s and has been widely adopted by retailers and CPG manufacturers. In his book *Collaborative Planning, Forecasting, and Replenishment: How to Create a Supply Chain Advantage*, Seifert (2003) defined CPFR as "an initiative among all participants in the supply chain intended to improve the relationship among them through jointly managed planning processes and shared information" (page 30). Walmart is one of the largest customers of our sponsoring company. Therefore, a dedicated CPFR team engages in CPFR activities with Walmart.

The typical benefits of CPFR, as summarized by Seifert (2003), are:

- Improved Reaction Times to Consumer Demand.
- Higher Precision of Sales Forecasts.
- Direct and Lasting Communication.
- Improved Sales.
- Inventory Reduction.
- Reduced Costs.

Even though CPFR has significant potential, many initiatives fall short because of complex implementation challenges. This suggests that collaborative forecasting is not a straightforward, universal solution but a complicated strategic approach that requires careful planning and execution (Folinas & Rabi, 2012).

2.4 Forecasting Metrics

Forecasting metrics offer valuable and actionable insights that can improve forecast accuracy and the business's overall performance. Ultimately, improvement is only possible if it can be measured. Therefore, selecting the appropriate metrics is the first step in any effort to drive forecast improvement (Vandeput, 2023). Three commonly used forecast accuracy measurements are MAE, MAPE, RMSE, and APE.

- **Mean Absolute Error (MAE):** The Mean Absolute Error is the mean of the absolute errors. It is a simple metric that is easy to explain and understand. The downside of minimizing MAE often results in under-forecasting (Vandeput, 2023). The formula is below:

$$MAE = \frac{1}{n}\sum |e_t|$$

- **Mean Absolute Percentage Error (MAPE):** The MAPE is calculated as the average of the absolute errors divided by the demand for each period, reflecting the average of the absolute percentage errors (APE). It is one of the most widely used metrics for assessing forecast accuracy. However, it is also one of the most flawed metrics, according to Vandeput (2023). The reason is that MAPE calculates each error individually by dividing the corresponding demand. This leads to a skewed error weighting. MAPE will be extremely high for low-demand periods. On the other hand, MAPE for high-demand periods will not be limited to a certain percentage error. For these reasons, optimizing MAPE will result in forecasts that will likely be too conservative (Vandeput, 2023). Below is the calculation for MAPE:

$$MAPE = \frac{1}{n}\sum \frac{|e_t|}{d_t}$$

- **Root Mean Square Error (RMSE):** The Root Mean Square Error (RMSE) is the square root of the average squared forecast error. It is calculated as below:

$$RMSE = \sqrt{\frac{1}{n}\sum e_t^2}$$

- **Absolute Percentage Error (APE):** The Absolute Percentage Error is the total absolute errors divided by the total demand. This is the method that our sponsor company is using to aggregate their forecasting errors. The formula is shown below:

$$APE = \frac{\sum |e_t|}{\sum d_t}$$

Vandeput (2023) argues that RMSE is a helpful metric, even though it might be hard to explain and understand. Vandeput lists three main reasons why RMSE is better than the other metrics. First, it puts more weight on the most significant errors. Second, RMSE is related to most inventory optimization formulas for calculating safety stock. Third, RMSE is not a biased measure. Table 3 lists the advantages and disadvantages of the three metrics, according to Vandeput (2023).

Table 3: Pros and Cons of Forecasting Metrics (Vandeput, 2023).

Metric	Simple	Not Biased	Error Weighing	Sensitivity to Outliers
MAE	P	x	x	P
MAPE	P	*	x	x
RMSE	x	P	P	x
APE	P	*	x	x

Vandeput (2023) suggests tracking both MAE and bias while avoiding MAPE to prevent pitfalls, as MAPE is a skewed error indicator that can lead to under-forecasting. In our project, we used MAPE for high-volume, stable SKUs where the forecast errors are unlikely to be skewed. For highly variable and intermittent SKUs,

we use APE (Absolute Percentage Error) as a benchmark to compare model performance.

2.5 Exogenous Variables

Several authors agree that factors such as climate, holidays, weekends, price and promotions are the primary drivers of product consumption in the food and beverage industry. Mitra et al. (2024) emphasize the importance of these factors. Cohen et al. (2017) underscore the significance of promotions, noting that in some European countries, between 12% and 25% of fast-moving consumer goods (FMCGs) are sold as promotions.

Other researchers extend their analyses to include macro-environmental factors such as the economy, demographics, and cultural forces (Feng and Foster, 2023). In this line, Mitra et al. (2024) also highlight socio-economic factors as a trigger for increased consumption of bottled beverages. Geerts et al. (2020) suggest that socio-economic improvements drive the consumption of bottled water over tap water. This factor may be significant for our project's focus on the economic growth of various regions. In their 1998 study, Billings and Agthe also concluded that water demand is directly dependent on the local economy.

Climate's impact on demand is also a critical factor. Keleş et al. (2018) found that “the rising temperature trend is causing an annual increase in liquid refreshment beverage (LRB) demand, with heat waves having a much greater positive impact on beverage demand compared to the negative impact of cold waves.” Feng and Foster (2023) indicate that climate and festival days are key parameters for improving demand estimation accuracy.

In their study, Keleş et al. (2018) found that markets in the Southwest U.S. and Southeast U.S. showed the highest positive temperature trends. In comparison, the North Central U.S. experienced the strongest seasonality. In the same study, Keleş et al. (2018) noted that sports drinks and water are the most impacted items by heat spikes, and that during a heat wave, there can be a demand increase of between 2.3% and 3.1% per additional degree of temperature. Mirasgedis et al. (2014) complement these findings by suggesting that, since summer consumption is already high, it leaves less room for additional growth from higher temperatures. However, they indicate that hot days in winter do trigger demand.

2.6 Synthesis and Research Gap: Contribution and Innovation

This research advances demand forecasting in the fast-moving consumer goods (FMCG) sector, specifically for bottled water, by proposing a hybrid methodological framework that integrates SKU clustering, traditional statistical models, and machine learning (ML) techniques, with selective inclusion of exogenous variables. While prior studies and industry practices often rely on either time series models, causal models, or isolated applications of machine learning, our contribution lies in combining these approaches systematically and evaluating their comparative strengths across distinct demand clusters and forecast

regions.

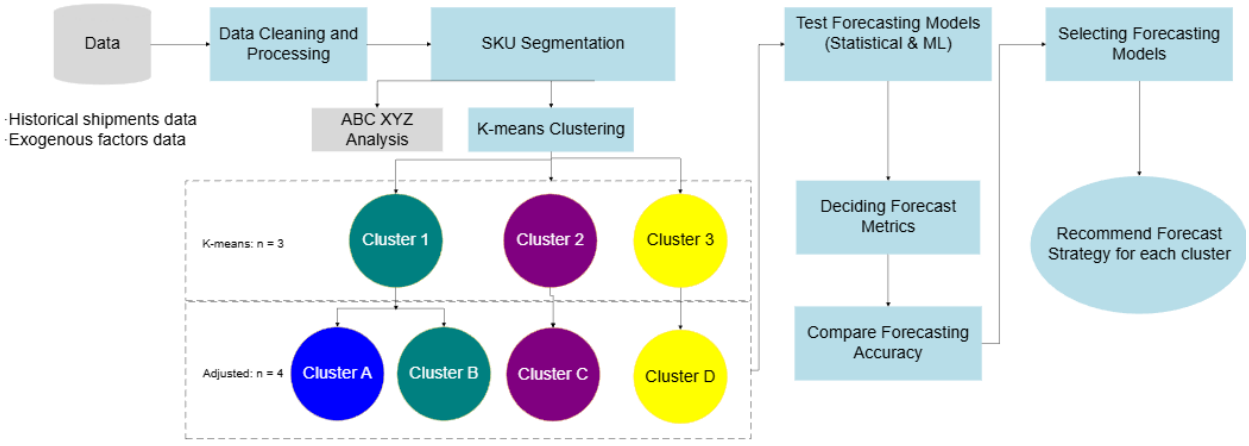
A key innovation of this work is the application of data-driven SKU segmentation using K-Means clustering, allowing forecasting strategies to be tailored to demand behavior characteristics rather than relying on a single forecasting model across the entire product portfolio. Rather than emphasizing model customization through extensive feature engineering, this research focuses on strategic model assignment, selectively testing statistical and machine learning models (such as Exponential Smoothing, ARIMA, XGBoost, TFT, and N-BEATS) based on the behavioral traits of each cluster.

By benchmarking multiple models, evaluating forecast performance using a variety of error metrics (MAPE, APE, MAE and RMSE), and aligning model selection with the sponsor company's specific sales structure and operational needs, this study delivers both methodological innovation and practical, actionable insights for improving tactical demand forecasting in the beverage industry.

3. METHODOLOGY

As outlined in the methodology process map shown in Figure 1, we started by cleaning and processing the data to ensure its quality. This involved gathering raw data, understanding its structure, resolving inconsistencies, and preparing it for analysis. Next, we explored an ABC-XYZ analysis to categorize SKUs based on their sales volume and demand variability. We then included intermittency in our analysis and used k-means clustering techniques to group products with similar demand patterns, so we could apply the forecasting models that best fit each group. Once we established key forecasting metrics, we selected and tested different models to see which performed best. By comparing accuracy across different models, we selected the most suitable model for each cluster. In the end, we identified the best forecasting models for each SKU segment to enhance demand forecasting and support more informed business decisions.

Figure 1: End-to-End Cluster-Based Forecasting Framework



3.1 Data Cleaning and Processing

Our data includes internal shipment data and external variables data such as holidays, temperature and weather events. Our sponsor company provided us with actual shipment data, including 3 years' actual shipment, their system forecast, and resultant forecast. For external data, we created a holiday calendar based on US public holidays. For temperature and weather events, we extracted the average temperature by state and weather events from the National Oceanic and Atmospheric Administration (NOAA)'s published historical data. Then we cleaned both internal and external data and joined the two data sets into one data frame.

3.1.1 Internal Data

Our sponsor company provided us with actual shipment data, including 3 years' actual shipment, their system forecast, and resultant forecast. For our internal data processing, we had weekly meetings with our sponsor company to better understand the data and gather their insights on the forecast levels we should utilize. They requested that we forecast at the levels of Customer, Forecast Region, and Parent SKU, as these are the levels currently used by their demand planners to generate forecasts. We combined these three columns to create a new data key called "Customer_FCRegion_ParentSKU," which became our lowest-level item for forecasting, hereinafter referred to as "ID".

After aggregating the data, we observed that several series contained zero values for the specified time frame. As a result, we eliminated all series with zero values during our clustering process. The company also requested using the most recent full year data for clustering purposes, as some SKUs have become obsolete.

3.1.2 External Data

To process external data, we first gathered information from various sources and created a data frame that matched the date range of our internal data frame. For the weekly temperature data, we converted daily temperatures into weekly values by calculating the minimum, maximum, and average temperatures for each state. We then combined this temperature data with our internal shipments data file to incorporate external variables into our machine-learning models. We treated the holiday data in a similar fashion.

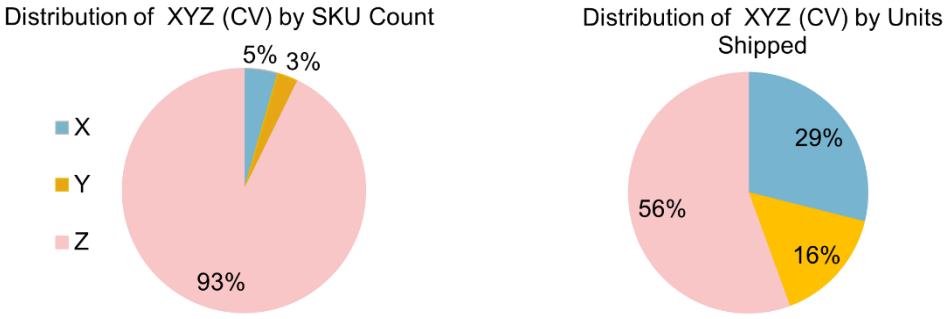
3.2 Demand Segmentation Approach

To improve forecasting accuracy and model interpretability, we adopted a two-stage segmentation approach for our SKU dataset, integrating both traditional and data-driven clustering methodologies.

3.2.1 ABC-XYZ Segmentation

Our initial approach leveraged the well-established ABC-XYZ analysis, which classifies SKUs based on their sales volume (ABC) and demand variability (XYZ). Items in the AX category, for example, are high-volume with low variability, whereas those in CZ represent low-volume, highly erratic demand. This segmentation aimed to prioritize forecasting resources and strategies based on SKU criticality and predictability. Figure 2 illustrates the distribution of SKUs by their XYZ classification based on both unit count and shipment volumes. For this process, the cutting line between one segment and the other is defined by the user.

Figure 2: XYZ Distribution



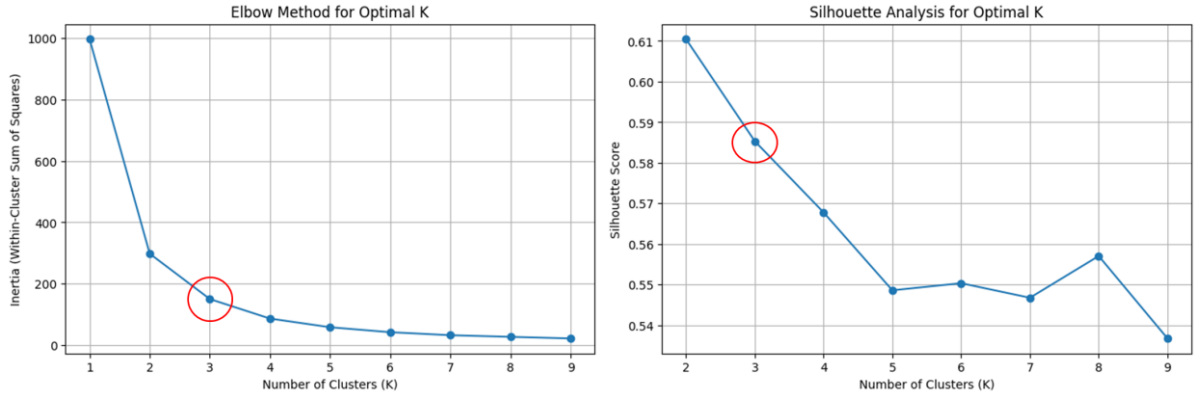
3.2.2 K-Means Clustering-Based Segmentation

To complement and refine the initial segmentation, we implemented an unsupervised learning approach using K-means clustering. Following the methodology outlined by Rožanec et al. (2022), who emphasize the importance of distinguishing between demand occurrence and size in forecasting lumpy and intermittent series, we selected three key features: mean demand, coefficient of variation (CoV), and intermittency. These features reflect the critical dimensions of demand behavior identified in the literature: volume, volatility, and frequency of zero-demand periods.

3.2.3 Elbow Method and Silhouette Analysis

To determine the optimal number of clusters (K), we employed the Elbow method and Silhouette analysis, settling on K=3 for its balance between clustering validity and practical interpretability. The results of both methods are illustrated in Figure 3.

Figure 3: Inertia and Silhouette Score



Below are the justifications for choosing K=3.

- **Elbow Method** (Left Plot): The graph shows a sharp decrease in Within-Cluster Sum of Squares (WCSS) between K=1 and K=3, after which the curve begins to flatten. This "elbow" at K = 3 suggests diminishing returns in variance reduction beyond this point, making it a strong point for optimal cluster count.
- **Silhouette Score** (Right Plot): The silhouette score measures how well each point fits within its cluster. Although K=2 yields the highest silhouette score (≈0.61), two clusters do not provide sufficient granularity to distinguish between different demand patterns. The limited segmentation at K=2 lacks interpretive power for defining meaningful forecasting strategies across product groups. So, we choose K value of 3 with a silhouette score (≈0.59) nearly as high as K = 2 to improve the practical interpretation of each cluster.

In summary, choosing three clusters allows for more straightforward interpretability and downstream segmentation logic (e.g., stable, erratic, lumpy demand), while still maintaining robust performance metrics. This decision supports analytical rigor and practical usefulness when applying differentiated forecasting strategies to each segment.

4. RESULTS AND DISCUSSION

This section presents the demand pattern segmentation results derived from K-means clustering. Initially, a 3-cluster solution was developed to group SKUs based on mean demand, coefficient of variation (CoV), and intermittency. In response to the sponsor's request for greater granularity, the Stable cluster was further segmented into two subgroups based on volume thresholds, resulting in an adjusted 4-cluster model.

4.1 Cluster Results

This section shows the 3-clusters results based on K-means (K=3) and adjusted 4-cluster results (K=4), based on the sponsor's request.

4.1.1 3-Cluster Results

The resulting clusters were characterized as shown in Table 4 below.

Table 4: 3-Cluster Results Characteristics

Cluster	Color	Characteristics
Cluster 1: Stable	Teal	High demand, low CoV, and low intermittency, indicate consistent, regular consumption patterns.
Cluster 2: Erratic & Intermittent	Purple	Low demand, moderate variability, and moderate intermittency, less predictable, with occasional gaps in demand.
Cluster 3: Lumpy	Yellow	Low demand, high variability, and very high intermittency, highly sporadic and unpredictable.

Figure 4 shows three scatter plots that visualize the outcome of K-means clustering (K=3) using three key demand pattern features: Mean Demand, Coefficient of Variation (CoV), and Intermittency. Each cluster is color-coded into one of three, reflecting distinct demand behavior profiles.

Figure 4: 2D Scatter Plot for K-means Clustering

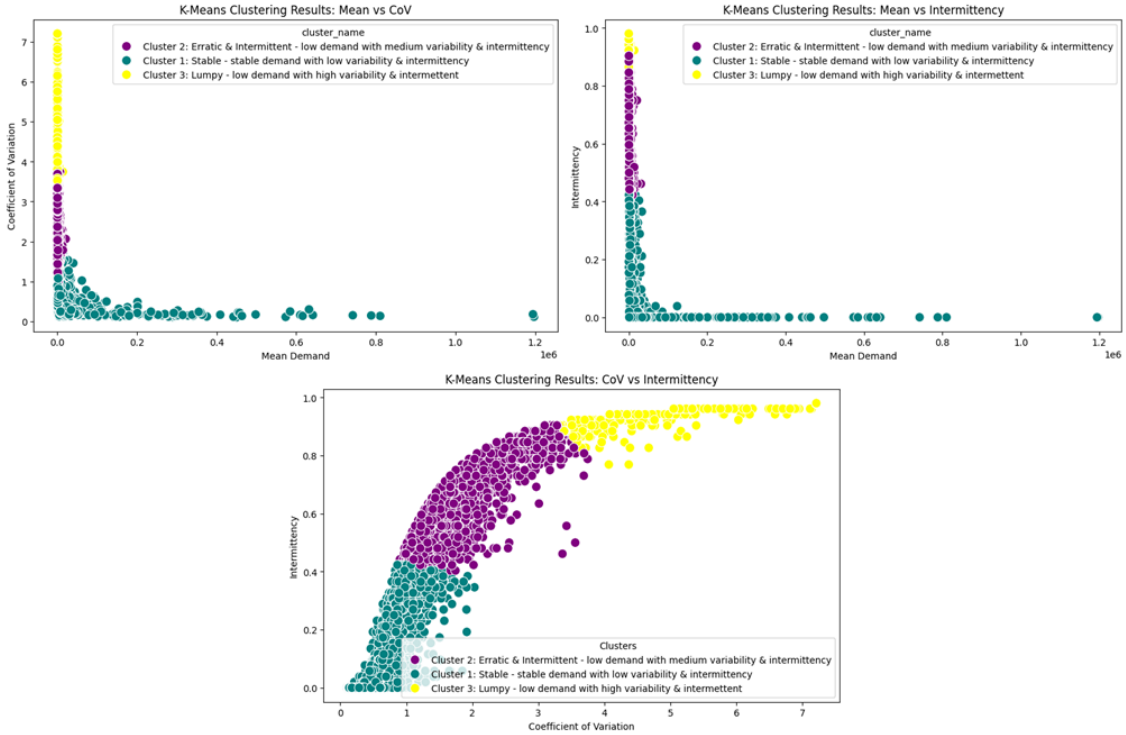
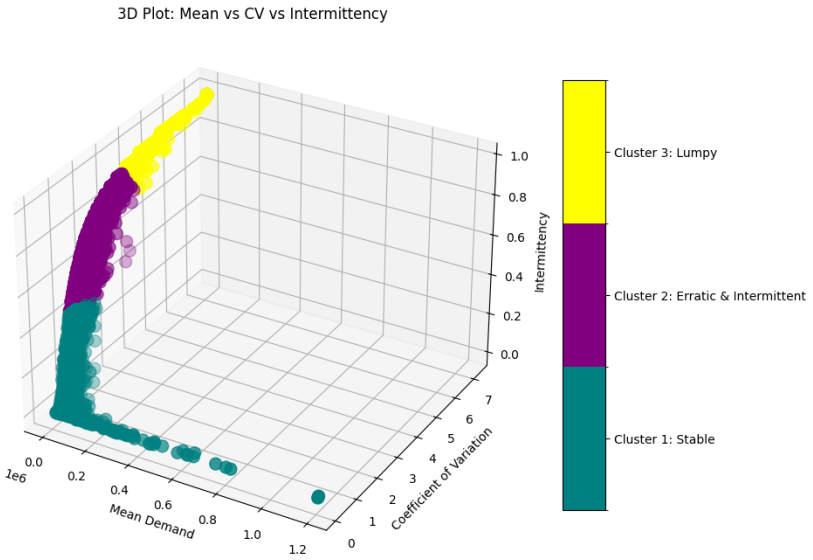


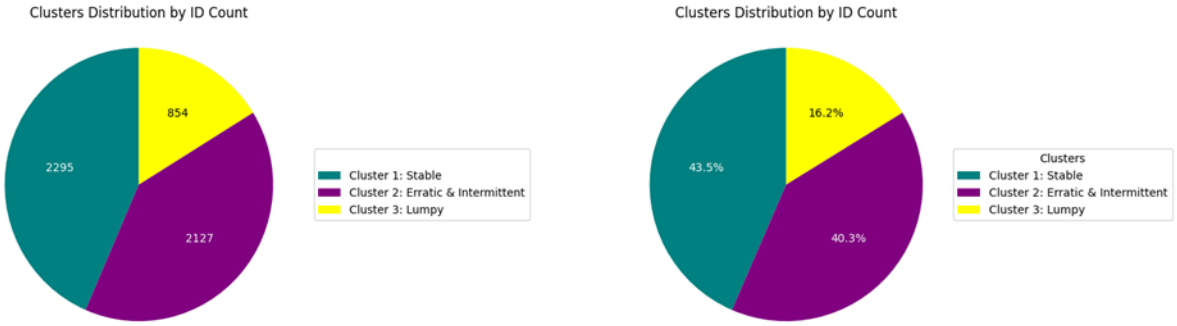
Figure 5 presents a 3D plot that visualizes the results of K-means clustering (with K=3) based on three demand characteristics: Mean Demand, Coefficient of Variation (CoV), and Intermittency in a single plot.

Figure 5: 3D Scatter Plot of Clusters (mean, CV and Intermittency).



The two pie charts in Figure 6 illustrate SKU count distribution across three demand clusters: Stable, Erratic & Intermittent, and Lumpy.

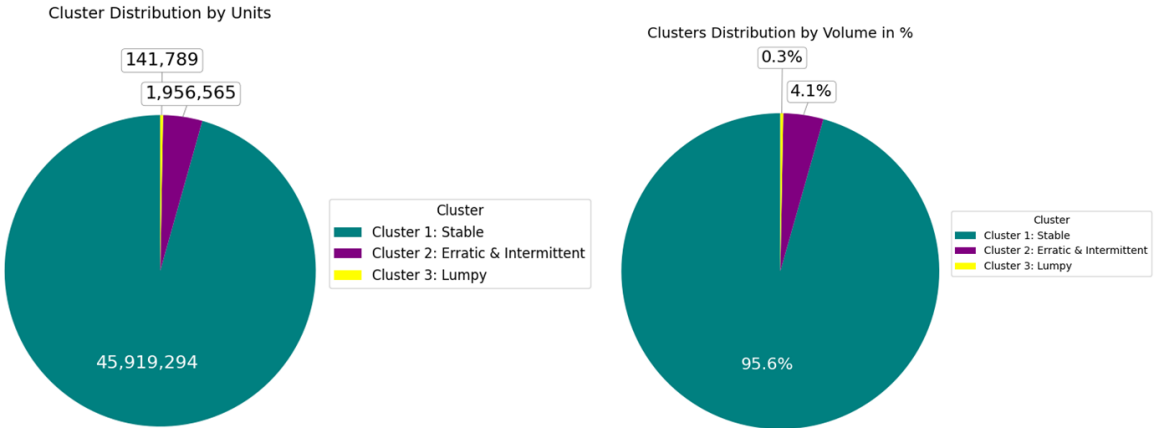
Figure 6: Distribution of SKU ID Count across Clusters



The left chart presents the absolute SKU counts, showing that Cluster 1 (Stable) has the highest number of SKUs (2,295), followed closely by Cluster 2 (Erratic & Intermittent) with 2,127, and Cluster 3 (Lumpy) with 854. The right chart displays these distributions as percentages, where the Stable and Erratic clusters account for 43.5% and 40.3% of the total SKUs, respectively, while the Lumpy cluster makes up 16.2%. This segmentation highlights that most SKUs exhibit stable or moderately erratic demand, while the Lumpy cluster only represents a small portion.

The two pie charts in Figure 7 illustrate the distribution of total shipment volume across three demand clusters: Stable, Erratic & Intermittent, and Lumpy.

Figure 7: Cluster Distribution by Volume



The left pie chart shows the distribution in absolute units, revealing that Cluster 1 (Stable) accounts for most of the volume with over 45.92 million units. In comparison, Clusters 2 and 3 contribute only 1.96 million and 141,789 units, respectively. The chart on the right displays this breakdown as percentages, emphasizing that Cluster 1 alone comprises 95.63% of the total volume, followed by Cluster 2 with 4.07%, and Cluster 3 with just 0.30%. This highlights that although Clusters 2 and 3 contain many SKUs, their contribution to overall volume is minimal, reinforcing the dominance of high-volume stable items in driving total demand.

Given the overwhelming dominance of Cluster 1 (Stable) in both SKU count and total volume, it became clear that this group is the primary driver of total demand and the most operationally critical segment. However, despite sharing the characteristic of stability, SKUs within Cluster 1 display considerable variability in their individual volume contributions. Some SKUs account for a disproportionately large share of the volume, while others contribute only marginally.

Treating Cluster 1 as a homogeneous group would therefore risk misaligning forecasting and inventory management efforts, either over-allocating resources to low-impact SKUs or under-prioritizing high-impact ones. Recognizing this internal heterogeneity is crucial: SKUs with different volume scales pose different levels of operational risk, require different forecast accuracy standards, and have varying impacts on business performance. In response, the sponsor company requested a more granular segmentation within Cluster 1 to enable differentiated forecasting strategies that better align with business impact, ensuring that resource prioritization matches the true operational and financial significance of each SKU.

4.1.2 Adjusted 4-Cluster Results

In response to the sponsor company's request, we further segmented this cluster into two subgroups: Stable_High_Volume and Stable_Low_Volume, using a mean volume threshold of 50,000 units per week. SKUs with a mean volume greater than 50,000 were classified as high volume, while those below the threshold were categorized as low volume. The adjusted clusters were characterized as shown in Figure 8 and Table 5 below:

Figure 8: Examples of Demand Patterns in Each Cluster

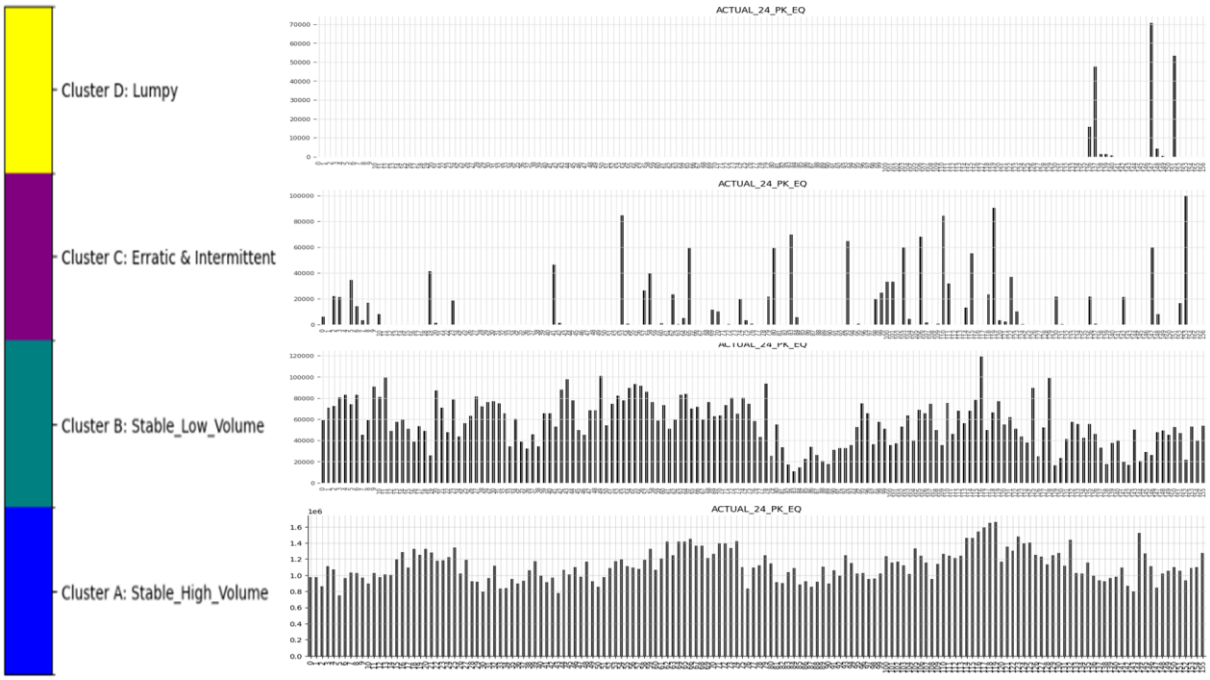


Table 5: 4-Cluster Results Characteristics

Cluster	Color	Characteristics
Cluster A: Stable_High_Volume	Blue	High demand with very low CoV and intermittency — highly predictable, operationally critical SKUs.
Cluster B: Stable_Low_Volume	Teal	Moderate demand with low variability and intermittency — still stable but less impactful in volume.
Cluster C: Erratic & Intermittent	Purple	Low demand with moderate to high variability and moderate intermittency — less predictable with occasional demand gaps.
Cluster D: Lumpy	Yellow	Low demand with very high CoV and high intermittency — highly sporadic and difficult to forecast.

Figure 9 (2D) and Figure 10 (3D) show the updated scatter plots displaying the results of a refined 4-cluster segmentation, where the original Stable cluster has been divided into Stable_High_Volume (blue) and

Stable_Low_Volume (teal), alongside the previously defined Erratic & Intermittent (purple) and Lumpy (yellow) clusters.

Figure 9: 2D Scatter Plot for Adjusted Clustering

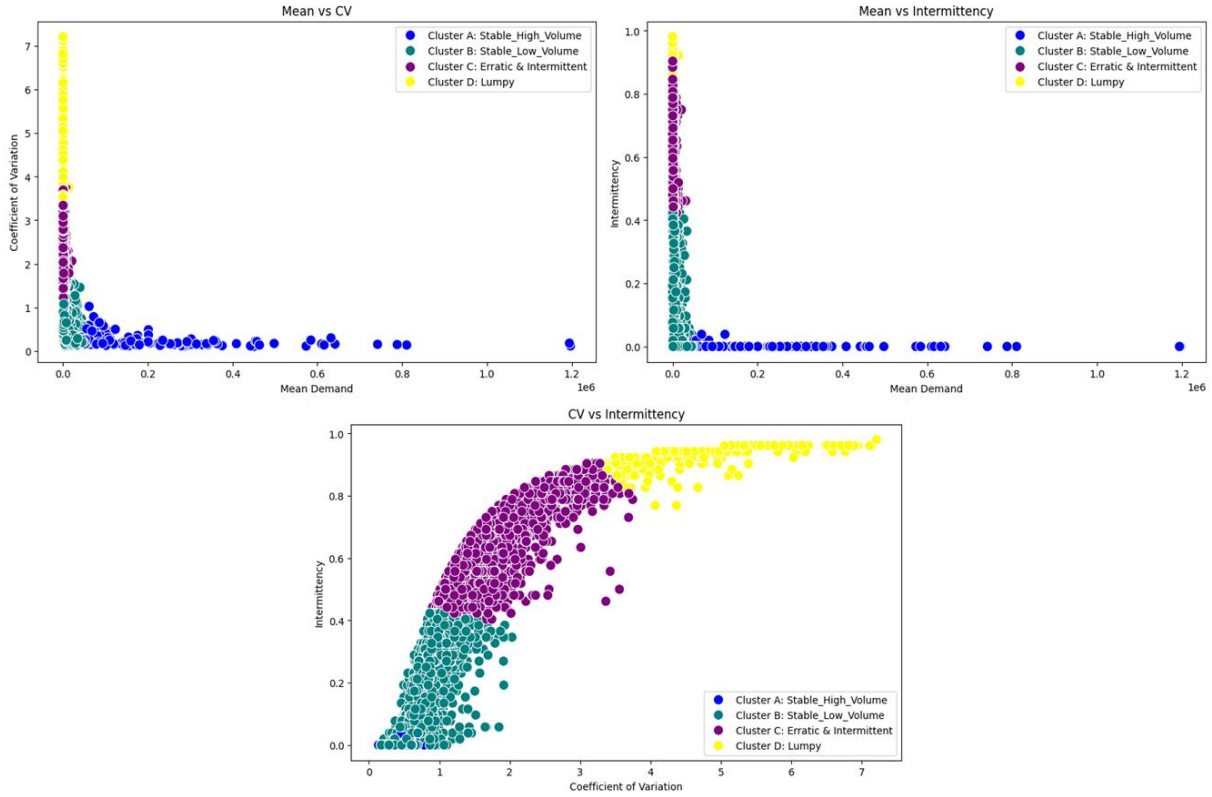
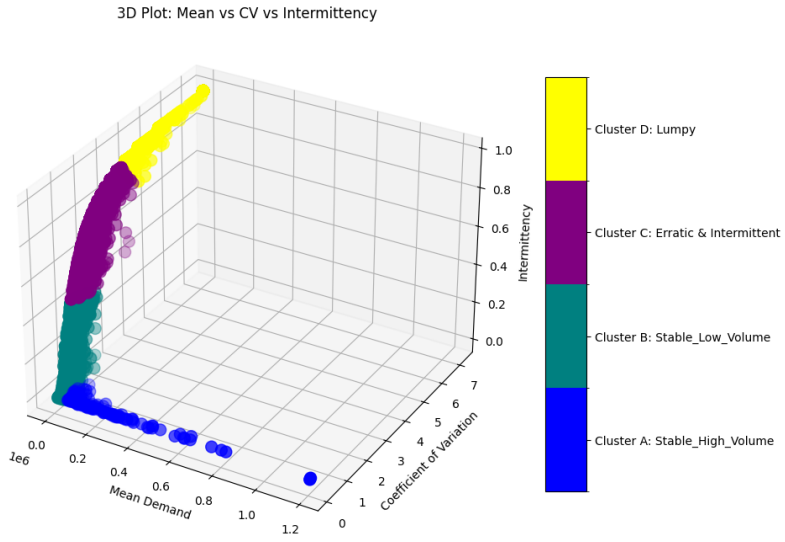
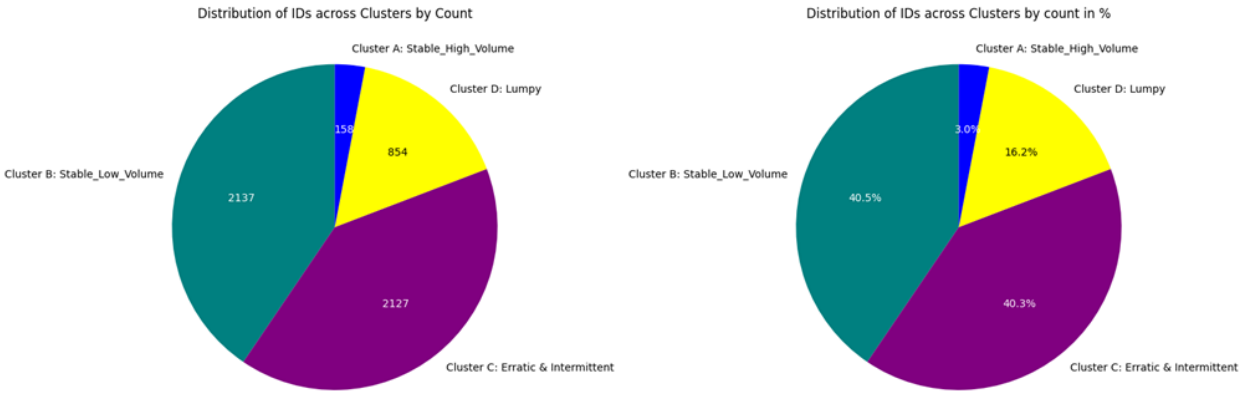


Figure 10: 3D Scatter Plot for Adjusted Clustering



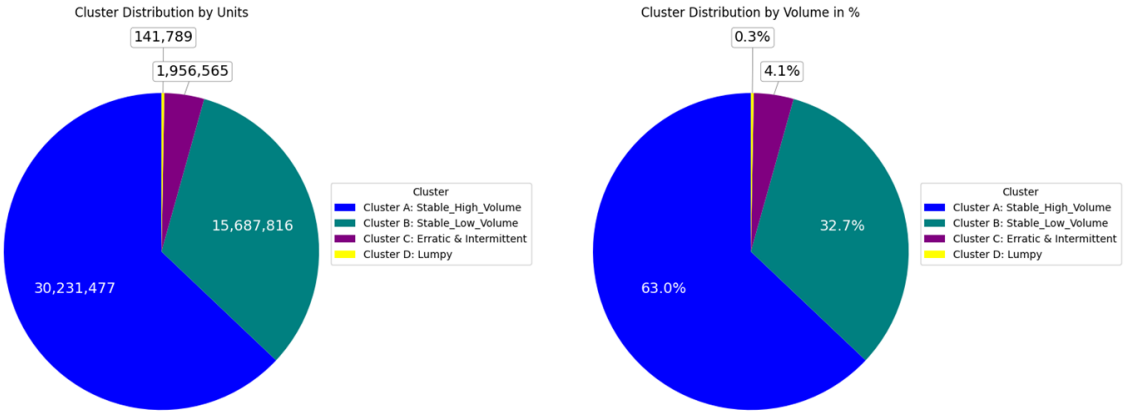
The two pie charts in Figure 11 illustrate the distribution of SKUs (or IDs) across four refined demand clusters following the segmentation of the original Stable group into Stable_High_Volume and Stable_Low_Volume. The left chart presents the absolute SKU counts, showing that most SKUs fall into Stable_Low_Volume (2,137) and Erratic & Intermittent (2,127) clusters, with Lumpy containing 854 SKUs, and only 158 SKUs classified as Stable_High_Volume. The right chart converts these counts into percentages, revealing that Stable_Low_Volume and Erratic & Intermittent each represent approximately 40% of the SKU base, while Lumpy accounts for 16.2%, and Stable_High_Volume comprises just 3.0%. This distribution highlights that while high-volume SKUs are few, most items fall into low-volume or more unpredictable demand categories, underscoring the importance of tailored strategies by cluster type.

Figure 11: Distribution of SKU ID Count across 4 Clusters



The two pie charts in Figure 12 illustrate the distribution of total shipment volume across the four refined demand clusters. The chart on the left presents the data in absolute units, showing that Cluster A: Stable_High_Volume dominates with over 30.2 million units, followed by Cluster B: Stable_Low_Volume with 15.7 million units. In contrast, Cluster C: Erratic & Intermittent and Cluster D: Lumpy contribute significantly less, with volumes of approximately 2 million and 142,000 units, respectively. The chart on the right displays the same distribution in percentage terms, highlighting that Stable_High_Volume alone accounts for nearly 63.0% of total volume, while Stable_Low_Volume represents 32.7%, and the remaining two clusters combined account for just 4.4%. This emphasizes that while high-volume stable SKUs are relatively few in number, they are overwhelmingly dominant in terms of business impact and demand volume.

Figure 12: 4 Cluster Distribution by Volume



To conclude, the adjusted four-cluster segmentation offers a more nuanced understanding of SKU demand behavior by splitting the original Stable cluster into Stable_High_Volume and Stable_Low_Volume, alongside the existing Erratic & Intermittent and Lumpy clusters. This refined view reveals that while Stable_High_Volume SKUs represent only 3% of total items, they contribute a dominant 63% of total shipment volume, highlighting their critical importance to the business. In contrast, Stable_Low_Volume SKUs, which account for over 40% of SKUs, contribute approximately 33% of volume, suggesting they are stable but less impactful. The Erratic & Intermittent and Lumpy clusters, though collectively comprising over 56% of SKUs, contribute less than 5% of volume, underscoring their complexity despite limited contribution. This segmentation provides a strong foundation for applying differentiated forecasting and inventory strategies tailored to the volume and variability characteristics of each group.

4.1 Model Selection and Validation

Our forecast model selection process is structured around four distinct demand clusters, each with tailored modeling strategies based on demand patterns. Below are the reasons why we chose different models for each cluster.

4.1.2 Cluster A: Stable_High_Volume

For SKUs classified as Stable_High_Volume, we selected Exponential Smoothing, ARIMA, XGBoost, and TiDE for evaluation. These items exhibit consistent demand patterns with high volumes, making them ideal candidates for both traditional time series models and advanced machine learning techniques. Exponential Smoothing and ARIMA are time-tested approaches known for effectively capturing trend and seasonality. XGBoost is included to evaluate whether machine learning can uncover non-linear relationships or interactions with external variables. TiDE, a transformer-based deep learning model, is well-suited for modeling temporal patterns in rich, high-frequency datasets, making it a strong candidate for this stable and data-rich cluster.

4.1.3 Cluster B: Stable_Low_Volume

For Cluster B: Stable_Low_Volume, we retained the same core models used in Cluster A—Exponential Smoothing, ARIMA, XGBoost, and TiDE—but also introduced N-BEATS to the candidate pool. While demand remains stable, the lower volume introduces a slightly higher degree of noise, which may challenge some traditional models. N-BEATS is a neural forecasting model that uses backward and forward residual decomposition, making it particularly effective in low-signal environments. This broader model suite ensures we capture both linear and non-linear dynamics while testing the adaptability of neural models to lower-volume series.

4.1.4 Cluster C: Erratic & Intermittent

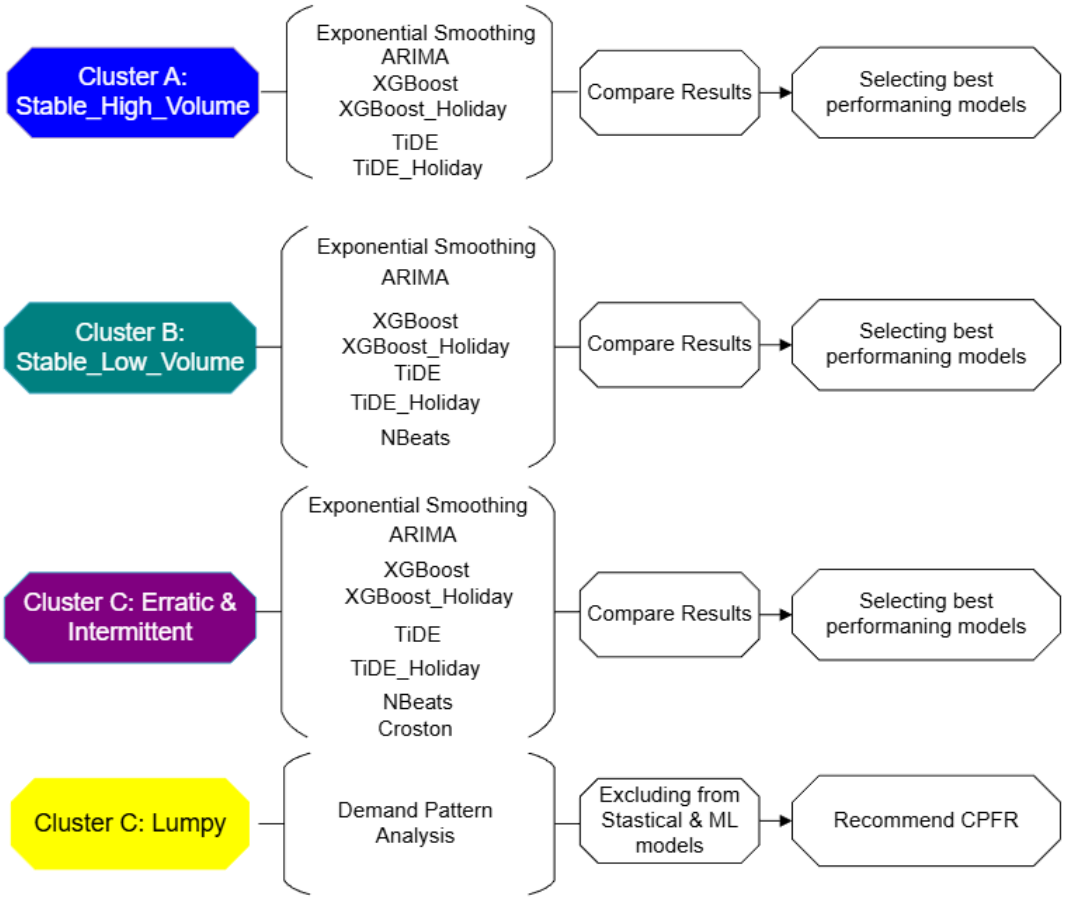
Given the volatile nature of demand in this cluster, we adopted a comprehensive modeling strategy including Exponential Smoothing, ARIMA, XGBoost, TiDE, N-BEATS, and Croston. Including Croston is critical, as it is designed explicitly for intermittent demand patterns characterized by many zero values. While traditional models may underperform in this context, testing them provides useful benchmarks. Meanwhile, advanced models like XGBoost, TiDE, and N-BEATS can potentially capture non-linear signals or hidden structures in erratic data. This cluster requires robust experimentation to identify the few models capable of managing its inherent unpredictability.

4.1.5 Cluster D: Lumpy

Lumpy demand is defined by many periods of zero demand punctuated by large, irregular spikes. These patterns are not conducive to statistical or machine learning models, which assume some degree of pattern regularity or sufficient data frequency. For this cluster, we excluded forecasting models and instead recommend a qualitative, judgment-based approach through CPFR (Collaborative Planning, Forecasting, and Replenishment). CPFR enables proactive alignment between internal planning teams and external partners, ensuring supply is responsive to true demand signals that are otherwise unforecastable.

Based on the above clusters' characteristics, we develop the model selection process as presented in Figure 13.

Figure 13: Model Selection Process



The process begins with applying relevant models to each cluster, followed by a systematic comparison of forecasting performance using accuracy metrics such as MAPE, MAE, APE, and RMSE. Based on this evaluation, we selected the best-performing model(s) for each cluster. For the Lumpy cluster, due to its highly irregular demand patterns, we excluded statistical and ML models and instead recommend a collaborative planning approach (CPFR) as the most suitable forecasting strategy. Notably, for several machine learning and deep learning models—such as XGBoost and TiDE—we also tested variants incorporating a holiday flag to account for demand fluctuations around public holidays. These "holiday" model versions (e.g., XGBoost_Holiday, TiDE_Holiday) allowed us to evaluate the added value of exogenous variables in improving forecast accuracy.

This structured framework ensures that each demand cluster receives a tailored forecasting solution aligned with its unique behavior and business impact. In the next section, we present the results of our model testing and evaluation.

To identify the best-performing forecast models, we evaluated a range of statistical and machine learning methods available in the DARTS library. Forecasts were generated for five selected SKUs in each cluster.

The following sections present detailed results for each cluster, discuss the implications of model performance, and highlight any limitations associated with our recommendations.

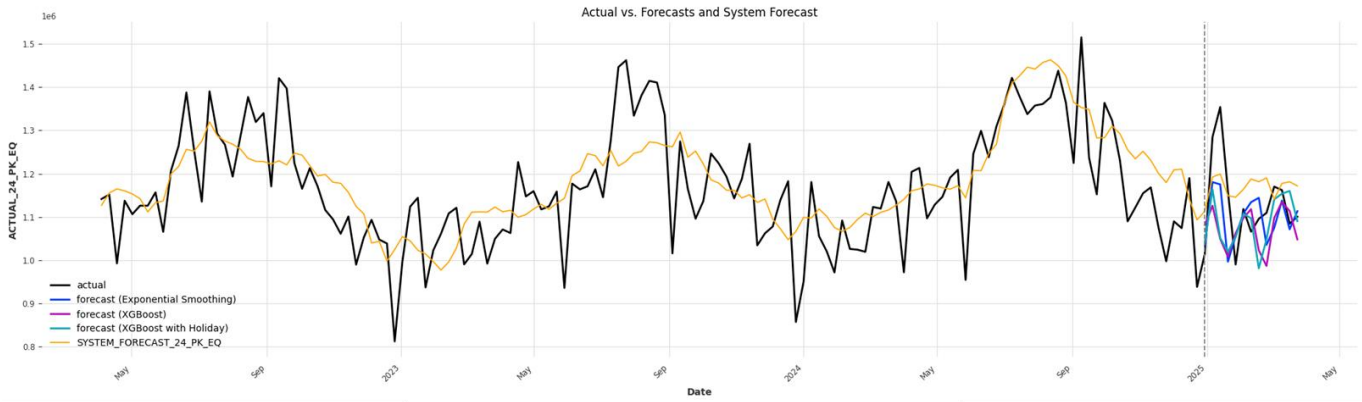
4.2 Cluster A Forecast Results

Cluster A includes SKUs with stable demand, low variability, and minimal intermittency. These characteristics make them suitable candidates for time series models that rely on consistent patterns. For this cluster, we evaluated six forecasting models: Exponential Smoothing, ARIMA, XGBoost, XGBoost with holiday flags, TiDE, and TiDE with holiday flags. In this section, we present the analysis of the forecast performance for two representative SKUs; results for additional SKUs are provided in Appendix A.

Cluster A: SKU1:

Figure 14 compares six forecasting models and the current system forecast against actual demand for Cluster A SKU1.

Figure 14: Forecast results of Cluster A_SKU1

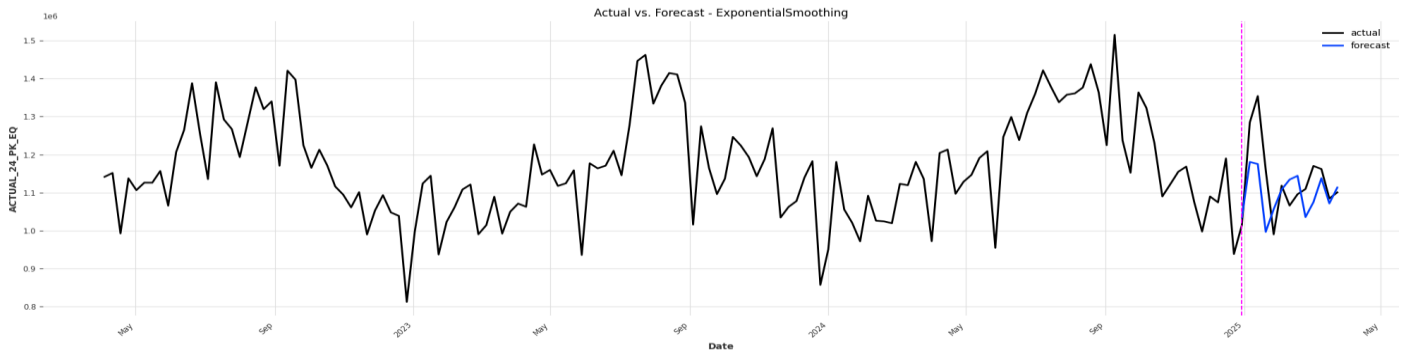


Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	5.76	5.96	67473.71	86138.52
ARIMA	10.88	11.33	128224.51	150570.33
XGBoost	7.86	8.18	92520.15	118787.37
XGBoost_holiday	6.43	6.75	76330.22	109108.92
TiDE	11.55	11.38	128707.08	151006.76
TiDE_holiday	7.74	8.12	91819.50	123092.59
System Forecast	7.28	7.19	81296.06	93061.03

Based on the forecast error metrics, Exponential Smoothing achieved the lowest MAPE at 5.76%, compared to the baseline system forecast at 7.28%, representing an absolute improvement of 1.52 percentage points in forecast accuracy. Incorporating a holiday flag improved the performance of machine learning models like XGBoost; however, the enhancement was not sufficient to outperform Exponential Smoothing.

Figure 15 shows the forecast versus actual results for the best-performing model, Exponential Smoothing.

Figure 15: Best-performing model results of Cluster A_SKU1

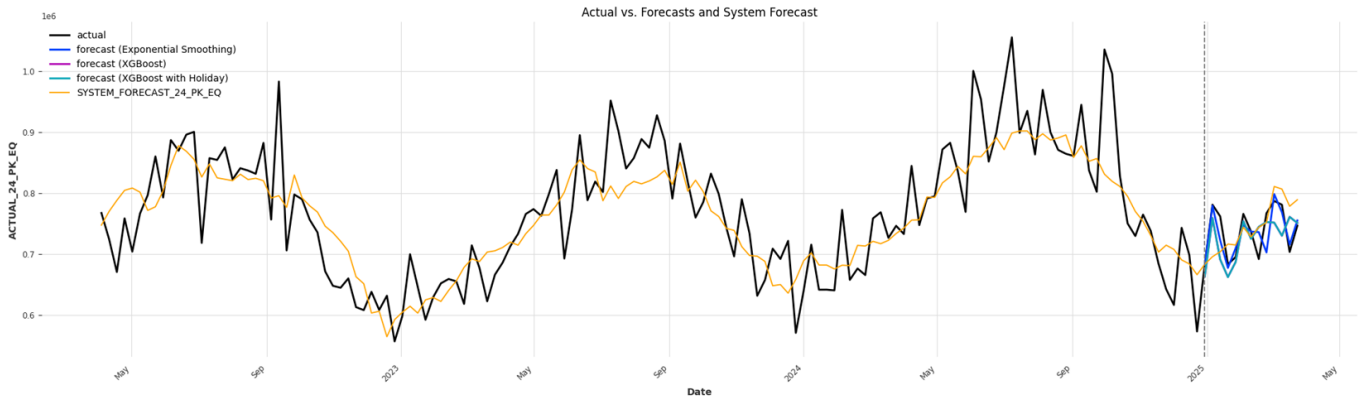


The graph demonstrates that the Exponential Smoothing forecast mostly aligns with actual demand across the 13-week forecast horizon. The model effectively captures both the general level and short-term fluctuations in demand, surpassing the system forecast and other advanced models in terms of accuracy and stability.

Cluster A: SKU2:

Figure 16 compares six forecasting models and the current system's forecast with actual demand for Cluster A SKU2.

Figure 16: Forecast results of Cluster A_SKU2



Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	2.43	2.44	17977.92	25921.56
ARIMA	16.41	16.63	122633.16	129088.24
XGBoost	3.98	3.97	29314.05	36026.36
XGBoost_holiday	3.98	3.98	29318.00	35999.81
TiDE	17.75	17.86	131758.70	156930.63
TiDE_holiday	6.31	6.27	46214.24	58096.63
System Forecast	4.81	4.82	35534.30	43283.00

As shown in Figure 16, Exponential Smoothing again delivered the best performance, achieving a MAPE of 2.43%, and significantly outperforming the system forecast (MAPE = 4.81%) by 2.38 percentage points. While XGBoost and its holiday-enhanced version (both with a MAPE of 3.98%) performed moderately well, they fell short of Exponential Smoothing in both forecast accuracy and consistency.

Figure 17 displays the Exponential Smoothing forecast compared to the actual demand for SKU2.

Figure 17: Best-performing model results of Cluster A_SKU2

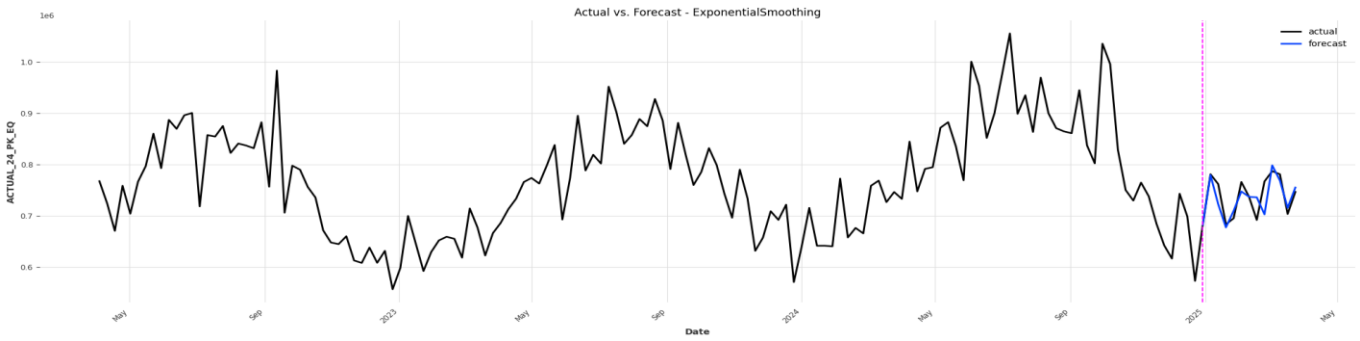
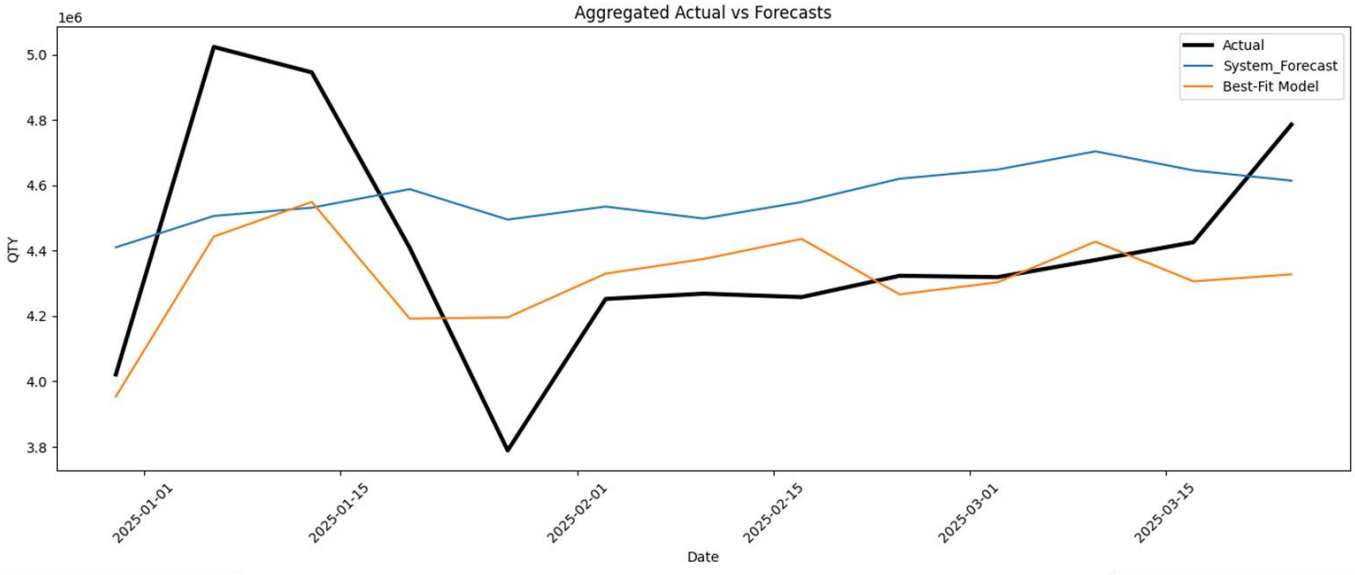


Figure 17 further illustrates how closely Exponential Smoothing tracks actual demand over the 13-week forecast horizon. The model effectively captures both the trend and magnitude of fluctuations, especially during periods of low and gradually recovering demand. This consistency across two representative SKUs reinforces Exponential Smoothing as the most reliable model for stable, high-volume items in Cluster A.

Aggregated Forecast Comparison for cluster A

Figure 18 compares the aggregated forecast results across five selected Cluster A SKUs using the best-fit model per SKU (orange line) against the current system forecast (blue line) and actual demand (black line).

Figure 18: Aggregated Forecast Comparison – Best-Fit Model vs. System Forecast



Model	MAPE	APE	MAE	RMSE
ExponentialSmoothing	11.681686	11.464197	100872.101920	146558.881076
ARIMA	19.700821	19.247786	169358.972055	220185.430119
XGBoost	9.811951	9.959410	87631.663984	128973.433074
XGBoost_Holiday	9.612519	9.555093	84074.123914	126712.155354
TiDE	19.697178	18.653448	164129.460802	202840.139408
TiDE_Holiday	11.186973	10.738054	94482.857345	123937.665648
System_Forecast	10.453895	9.946439	87517.534040	120808.248588
Best-Fit Model	8.601807	8.491803	74718.358903	106333.531924

The best-fit model demonstrates stronger alignment with actual demand, particularly in capturing major demand drops and rebounds in early 2025. In contrast, the system forecast shows a consistent upward bias and underreacts to significant demand shifts.

In terms of aggregated forecast accuracy, the best-fit model achieved a MAPE of 8.60%, outperforming the system forecast (10.45%) by an absolute 1.85 percentage points. This confirms that selecting models based on SKU-level fit results in more accurate and responsive forecasts at the aggregated level.

While our strategy initially focused on tailoring forecasting models at the cluster level, results from Cluster A indicate that incorporating an additional layer of SKU-level customization and leveraging machine learning models with external factors can further enhance forecast accuracy. Exponential Smoothing performed strongly for many individual stable SKUs; however, when forecasts were aggregated, XGBoost with holiday flags (XGBoost_Holiday) achieved the lowest MAPE among all models. This suggests that machine learning models enriched with exogenous variables, such as holiday effects, are particularly effective at capturing subtle demand shifts across multiple SKUs. These findings support a hybrid approach: using cluster characteristics to guide initial model selection while also considering the value of exogenous-driven machine learning models to maximize accuracy at broader planning levels.

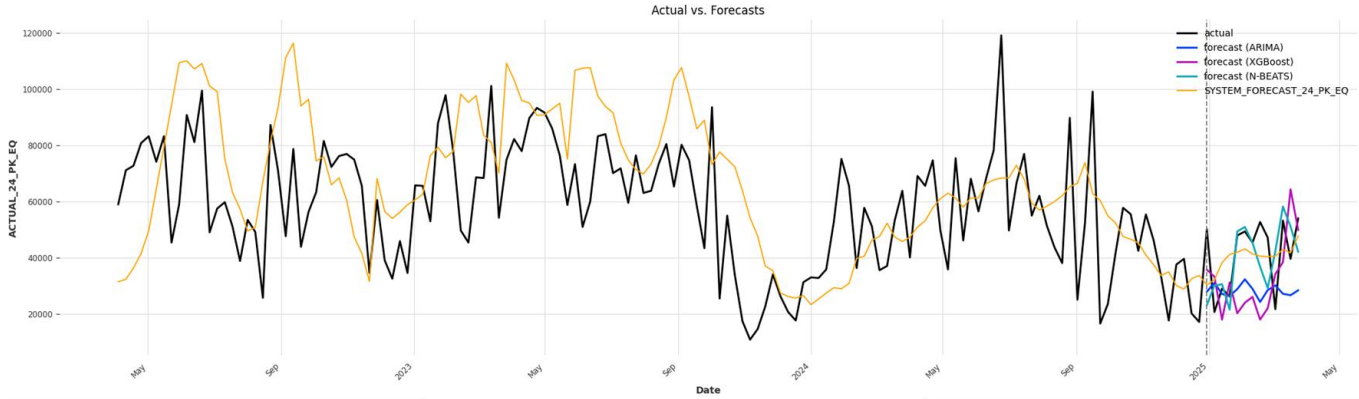
4.3 Cluster B Forecast Results

Cluster B consists of SKUs with relatively low demand volume, variability, and intermittency. These SKUs show some sporadic and intermittent demand patterns. To improve the forecasting of Cluster B SKUs, we have integrated N-BEATS (a deep learning model) into the test. In this section, we evaluate the forecast performance for two representative SKUs; results for additional SKUs in cluster B are included in Appendix A.

Cluster B SKU1:

Figure 19 compares the forecasts generated by seven models and the current system forecast against the actual demand for a representative SKU in Cluster B.

Figure 19: Forecast results of Cluster B_SKU1



Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	52.82	41.53	17143.86	21586.45
ARIMA	36.50	38.66	15959.97	18105.13
XGBoost	44.08	43.09	17788.48	19870.02
XGBoost_holiday	44.08	43.09	17788.51	19870.04
TiDE	75.54	77.83	32130.81	35936.49
TiDE_holiday	60.45	59.13	24411.61	27175.27
N-BEATS	27.15	23.96	9892.11	12854.42
System Forecast	29.07	23.84	9842.92	11151.95

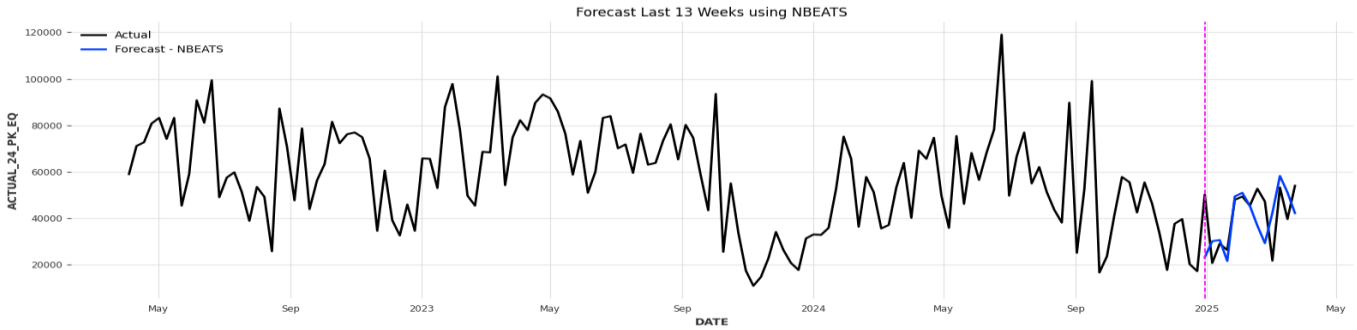
Among the models tested, N-BEATS achieved the best forecast performance, with a MAPE of 27.15%. This slightly outperforms the baseline system forecast, which had a MAPE of 29.07%. Although the system forecast achieved a slightly lower RMSE, N-BEATS delivered lower MAPE and APE, indicating it captured proportional forecast errors more effectively across the 13-week test horizon.

Other models, such as ARIMA (MAPE = 36.50%) and XGBoost (MAPE = 44.08%), showed moderate performance but exhibited higher forecast errors than N-BEATS and the system baseline. Visually, the N-BEATS forecast line closely follows the trend of actual demand, while other models display greater volatility

or overreaction to weekly fluctuations. This highlights N-BEATS' strength in handling SKUs where demand noise is more pronounced relative to volume size.

Figure 20 illustrates the forecast versus actual demand for the best-performing model, N-BEATS.

Figure 20: Best-performing model results of Cluster B_SKU1

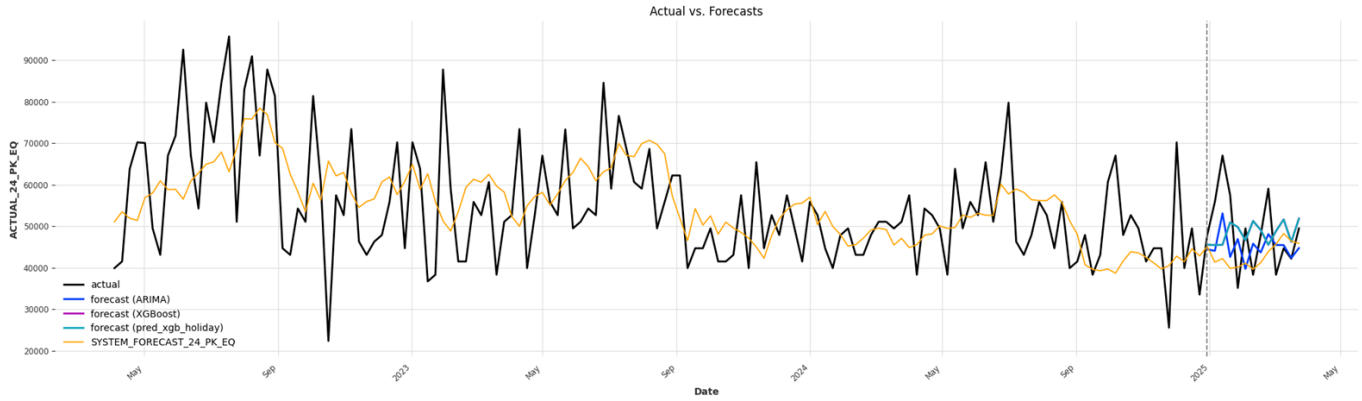


As shown in figure 20, the N-BEATS model captures the overall trend and level of demand reasonably well during the 13-week forecast horizon, despite significant week-to-week variability. The model successfully tracks major directional changes without overreacting to small fluctuations, which is critical for stable lower-volume SKUs where random noise can distort forecasts. This visual confirmation reinforces N-BEATS as the most suitable model for this SKU, outperforming both traditional statistical methods and the current system forecast in proportional accuracy.

Cluster B SKU2:

Figure 21 compares the forecasts generated by seven models and the current system forecast against actual demand for a second representative SKU in Cluster B.

Figure 21: Forecast results of Cluster B_SKU2

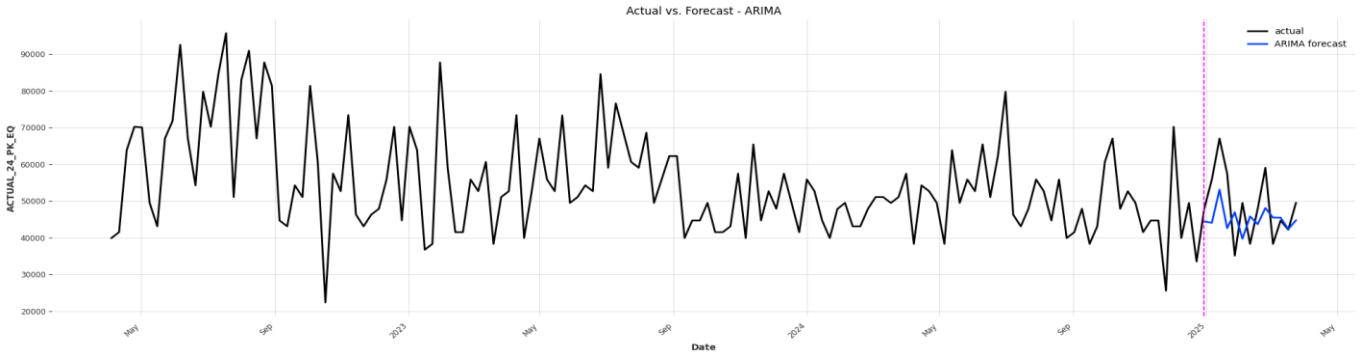


Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	26.68	26.30	12802.15	17272.52
ARIMA	15.82	15.98	7779.42	9069.89
XGBoost	17.79	17.37	8454.67	10287.22
XGBoost_holiday	17.79	17.37	8454.67	10287.22
TiDE	32.32	32.02	15584.64	19138.23
TiDE_holiday	20.86	21.32	10375.45	11990.61
N-BEATS	20.57	20.84	10140.97	12029.38
System Forecast	16.78	18.17	8844.77	11142.47

Among the tested models, ARIMA achieved the best forecast performance, with a MAPE of 15.82%, outperforming all other models, including the baseline system forecast (MAPE = 16.78). XGBoost and its holiday-enhanced variant performed similarly (both with MAPE = 17.79%) but had slightly higher RMSE values compared to ARIMA. Meanwhile, models like Exponential Smoothing, TiDE, and TiDE_holiday exhibited higher forecast errors, indicating that these models struggled with the noisier and lower-volume nature of this SKU.

Figure 22 presents the forecast versus actual demand for the best-performing model, ARIMA.

Figure 22: Best-performing model results of Cluster A_SKU2

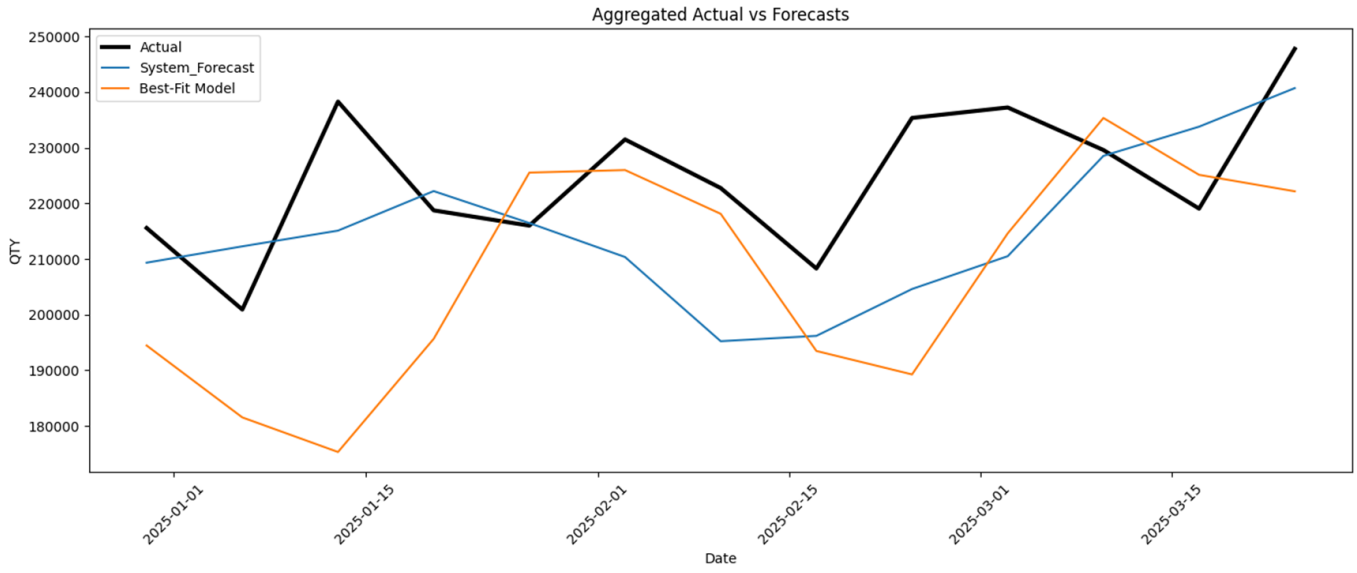


The above graph shows that the ARIMA model reasonably captures the overall level and directional movements of actual demand over the 13-week forecast horizon. While minor deviations are present during sharp demand fluctuations, ARIMA effectively tracks the general trend without introducing excessive lag or volatility. This visual alignment reinforces ARIMA's position as the most appropriate forecasting model for this SKU, outperforming both machine learning models and the current system forecast in proportional forecast accuracy.

Aggregated Forecast Comparison for cluster B

Figure 23 compares actual shipment against the system forecast and the best-fit model, highlighting the improved accuracy of the best-fit model across the 13-week horizon. The table in figure 23 summarizes model performance metrics, showing that the best-fit model achieved the lowest MAPE.

Figure 23: Aggregated Forecast Comparison – Best-Fit Model vs. System Forecast for Cluster B



Model	MAPE	APE	MAE	RMSE
ExponentialSmoothing	28.394291	25.337088	11386.416811	16070.838453
ARIMA	22.268332	22.940097	10309.215656	12942.015796
XGBoost	24.814958	24.140280	10848.574674	13543.172387
XGBoost_Holiday	24.611734	23.990869	10781.429444	13578.732186
TiDE	40.059084	40.293303	18107.698115	23575.593970
TiDE_Holiday	34.581889	33.488692	15049.724742	18899.167320
N-BEATS	20.214571	19.924479	8954.005197	12175.135814
System_Forecast	18.362056	17.066531	7669.651508	9966.972901
Best-Fit Model	18.014299	17.323881	7785.303776	10746.325540

Cluster B includes SKUs with stable but lower-volume demand, which introduces greater week-to-week variability relative to overall volume. Across the five representative SKUs, model performance varied depending on the underlying demand patterns. For individual SKUs, N-BEATS and ARIMA delivered the strongest results, outperforming other models and the system forecast in terms of proportional forecast accuracy (MAPE). Machine learning models such as N-BEATS demonstrated strong ability to capture overall

trends while handling noise, while traditional time series models like ARIMA performed well for SKUs with more regular seasonal structures.

At the aggregated level, the best-fit model strategy—selecting the most accurate model for each SKU—achieved a MAPE of 18.01%, slightly outperforming the system forecast (MAPE = 18.36%) by 0.35 percentage points, and reducing RMSE by approximately 1,220 units. These results reinforce the benefit of applying SKU-level model customization within each cluster to maximize forecast accuracy, particularly when dealing with lower-volume and more variable demand patterns.

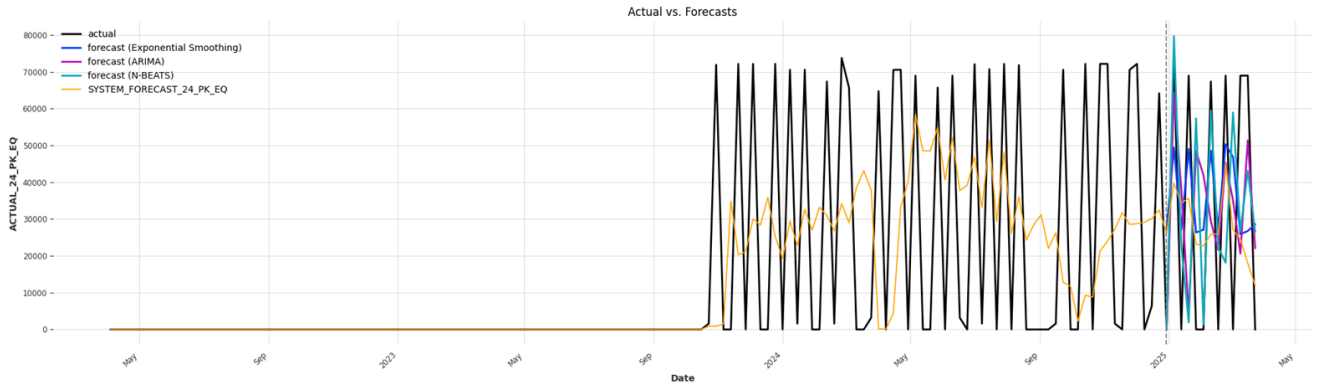
4.4 Cluster C Forecast Results

Having evaluated the forecast performance for stable lower-volume SKUs in Cluster B, we now turn to Cluster C, which consists of SKUs with more erratic and intermittent demand patterns. These SKUs present additional forecasting challenges, requiring adjustments in model selection and evaluation strategies. In this section, we evaluate the forecast performance for two representative SKUs; results for additional SKUs in cluster C are included in Appendix A.

Cluster C SKU1:

Figure 24 compares the forecasts generated by seven models and the current system forecast against the actual demand for a representative SKU in Cluster C.

Figure 24: Forecast results of Cluster B_SKU1



Model	APE	MAE	RMSE
Exponential Smoothing	90.46	28905.00	30265.41
ARIMA	98.07	31336.44	35774.29
XGBoost	108.36	34624.70	39153.69
XGBoost_holiday	107.81	34449.26	39097.77
TiDE	171.48	54794.05	63028.64
TiDE_holiday	272.91	87203.35	114890.66
N-BEATS	94.10	30066.17	37367.75
Croston	107.82	34451.82	34530.00
System Forecast	95.80	30611.23	32219.02

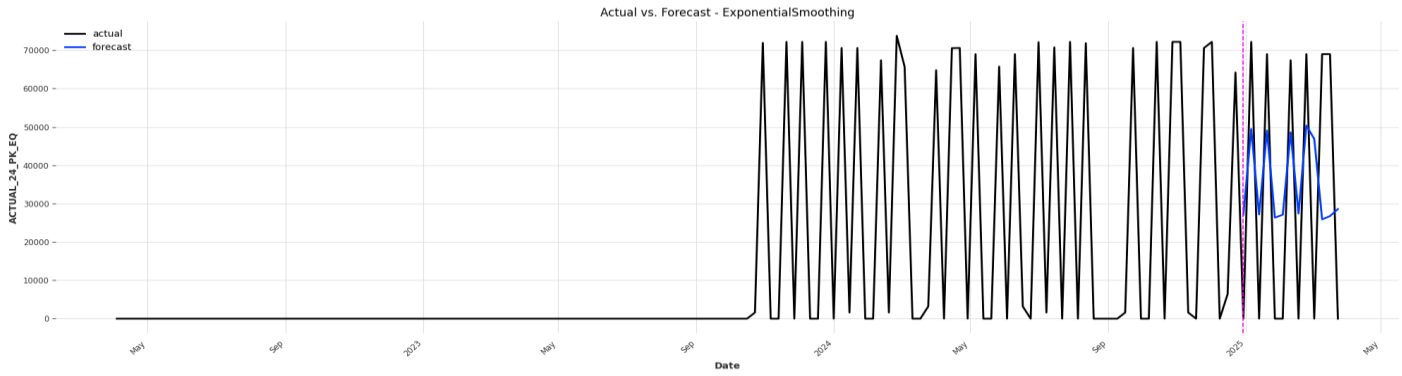
The demand pattern for this SKU is highly intermittent, marked by extended periods of zero demand interspersed with sharp bursts of activity. Given the prevalence of zero demand periods, we used Absolute Percentage Error (APE) instead of MAPE to assess forecast performance and prevent calculation distortions.

Among the models tested, Exponential Smoothing achieved the best forecast performance, with an APE of 90.46. Although the system forecast performed relatively close to Exponential Smoothing (APE = 95.80), Exponential Smoothing consistently showed slightly lower forecast errors across all metrics.

Other models, including ARIMA, XGBoost, and N-BEATS, exhibited higher APEs, and deep learning models like TiDE and TiDE_holiday struggled significantly, producing highly unstable forecasts. While none of the models perfectly capture the extreme variability, Exponential Smoothing offers more stable and consistent forecasts without the large oscillations seen in other model outputs (such as TiDE).

Figure 25 shows the forecast versus actual demand for the best-performing model, Exponential Smoothing.

Figure 25: Best-performing model results of Cluster C_SKU1

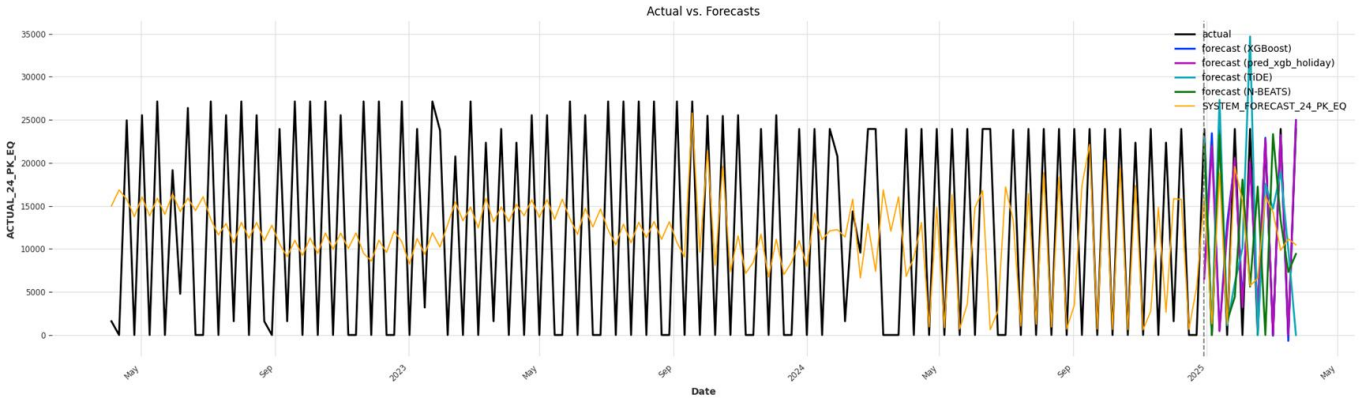


As depicted in Figure 25, Exponential Smoothing captures the overall demand activation periods and avoids the exaggerated oscillations seen in other models. While it does not fully replicate the sharp peaks of actual demand, it provides a stable and reasonable forecast across the 13-week horizon. Given the highly erratic and intermittent nature of the SKU's demand pattern, this smoother and more consistent forecast trajectory offers a practical balance between responsiveness and forecast stability. This visual confirmation supports Exponential Smoothing as the most appropriate model among the methods tested for this intermittent SKU.

Cluster C SKU2:

Figure 26 compares the forecasts generated by eight models and the current system forecast against the actual demand for another representative SKU in Cluster C.

Figure 26: Forecast results of Cluster C_SKU2



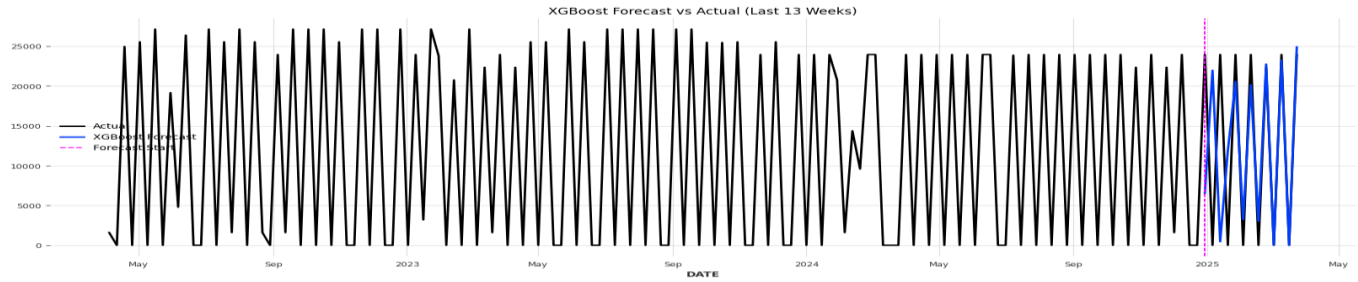
Model	APE	MAE	RMSE
Exponential Smoothing	81.77	10440.79	14095.50
ARIMA	88.18	11258.98	11960.03
XGBoost	56.80	7252.46	11159.06
XGBoost_holiday	54.37	6941.63	10833.78
TiDE	61.98	7914.34	10696.22
TiDE_holiday	171.58	21907.74	22835.88
N-BEATS	93.52	11940.88	14587.76
Croston	93.90	11989.55	12121.73
System Forecast	72.09	9204.38	10660.10

The demand pattern for this SKU exhibits high-frequency seasonality along with intermittent fluctuations, which poses challenges for both traditional and machine learning models. Given the prevalence of zero and irregular demand periods, we again utilize Absolute Percentage Error (APE) instead of MAPE to assess forecast performance.

Among the tested models, XGBoost with holiday flags (XGBoost_holiday) achieved the best forecast performance, with an APE of 54.37%. This outperformed the system forecast, which had an APE of 72.09%. Although the system forecast had a slightly lower RMSE, XGBoost_holiday demonstrated stronger proportional error handling, particularly in managing the SKUs' erratic demand patterns. It was better able to adapt to short-term peaks and valleys without the extreme forecast volatility seen in deep learning models like TiDE_holiday.

Figure 27 presents the forecast versus actual demand for the best-performing model, XGBoost with holiday flags (XGBoost_holiday).

Figure 27: Best-performing model results of Cluster C_SKU1 – XGBoost with Holiday

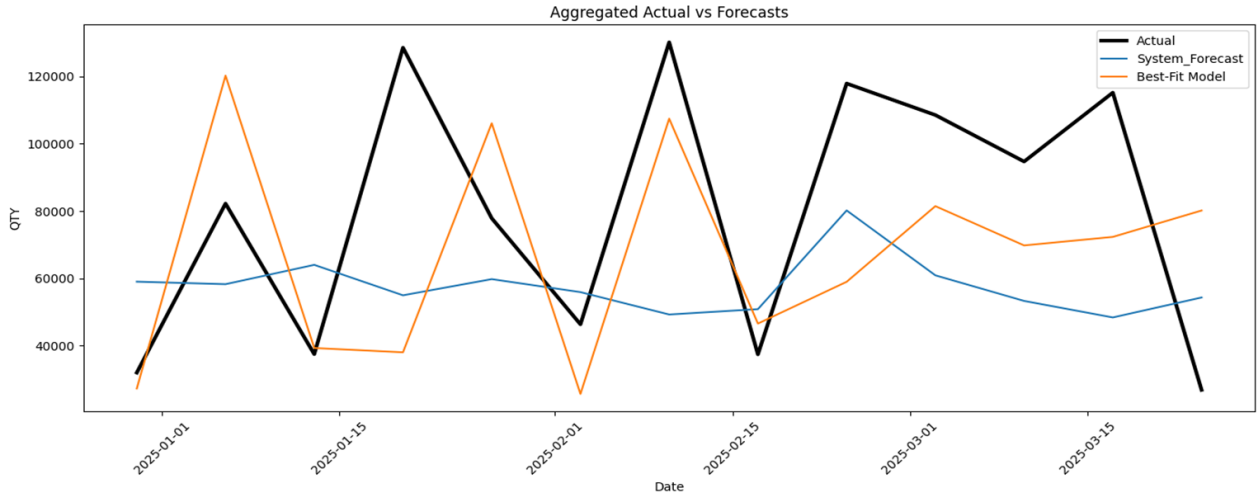


The figure shows that the XGBoost_holiday model is able to reasonably track the frequent and sharp oscillations in demand over the 13-week forecast horizon. While it does not fully replicate the extreme peaks and troughs seen in the actual demand, the model successfully follows the underlying cyclical pattern without introducing excessive volatility. This relative alignment with the SKU's demand patterns, combined with its the lowest APE performance compared to the system forecast and other models, confirms XGBoost_holiday as the most appropriate forecasting model for this SKU.

Aggregated Forecast Comparison for cluster C

Figure 28 compares actual shipments with the system forecast and the best-fit model for a highly variable demand segment. Despite the volatility, the best-fit model consistently outperforms other approaches, achieving the lowest APE, MAE, and RMSE as shown in the performance summary below the chart.

Figure 28: Aggregated Forecast Comparison – Best-Fit Model vs. System Forecast for Cluster C



Model	APE	MAE	RMSE
ExponentialSmoothing	107.998864	17199.741273	23463.847433
ARIMA	107.655872	17145.116769	24138.358341
XGBoost	133.593063	21275.835913	29910.599019
XGBoost_Holiday	132.488941	21099.995042	29875.184435
TiDE	134.085798	21354.308293	33834.915631
TiDE_Holiday	211.573957	33694.959126	58278.638071
N-BEATS	103.235000	16441.055189	24852.173936
Croston	111.059934	17687.242707	23336.620237
System_Forecast	104.126478	16583.030769	23089.586410
Best-Fit Model	91.807829	14621.180665	23793.813769

Cluster C includes SKUs with highly intermittent demand patterns, characterized by frequent periods of zero sales punctuated by unpredictable bursts. Due to the nature of the demand patterns, we used Absolute Percentage Error (APE) instead of MAPE to avoid distortions caused by zero demand periods.

Across the representative SKUs, model performance varied significantly. For SKU1, Exponential Smoothing produced the most stable forecasts, achieving the lowest APE and RMSE among the tested models. For SKU2, XGBoost with holiday flags (XGBoost_holiday) delivered the best proportional forecast accuracy, outperforming both the system forecast and other machine learning models in APE. However, even the best models struggled to accurately capture the sharp peaks and extreme fluctuations typical of Cluster C SKUs.

At the aggregated level, as illustrated in Figure 29, the best-fit model strategy achieved an APE of 91.82%, surpassing the system forecast (APE = 110.06%) by approximately 18.24 points.

These results highlight that statistical forecasting and ML models alone are insufficient for managing highly intermittent SKUs. While forecasts can provide broad directional signals, they are less effective for precise week-to-week planning. Consequently, a hybrid approach that combines statistical models with collaborative planning processes (such as CPFR) and exception-based management strategies is recommended for Cluster C.

4.5 Overall Forecast Metrics Summary for all tested SKUs

We aggregated the forecast results across the 15 representative SKUs selected from Clusters A (Stable High-Volume), B (Stable Lower-Volume), and C (Erratic & Intermittent) in Table 6.

Table 6: Forecast Metrics Summary for all tested SKUs

Model	APE	MAE	RMSE
Best-Fit Model	10.32	32,375	63,215
System_Forecast	11.88	37,257	71,244
XGBoost_Holiday	12.33	38,652	75,571
XGBoost	12.73	39,919	76,838
ExponentialSmoothing	13.76	43,153	86,194
TiDE_Holiday	15.22	47,743	79,821
ARIMA	20.92	65,604	128,104
TiDE	21.64	67,864	119,506
N-BEATS	96.23	301,761	524,043
Croston	100.19	314,172	524,645

Implementing the Best-Fit Model strategy resulted in a 1.56 percentage point improvement in APE and a reduction of approximately 8,028 units in RMSE compared to the baseline system forecast. This demonstrates

that selecting the most appropriate forecasting model per SKU can enhance forecast accuracy and operational planning reliability across a diverse SKU portfolio.

Below are some key insights by Cluster Behavior.

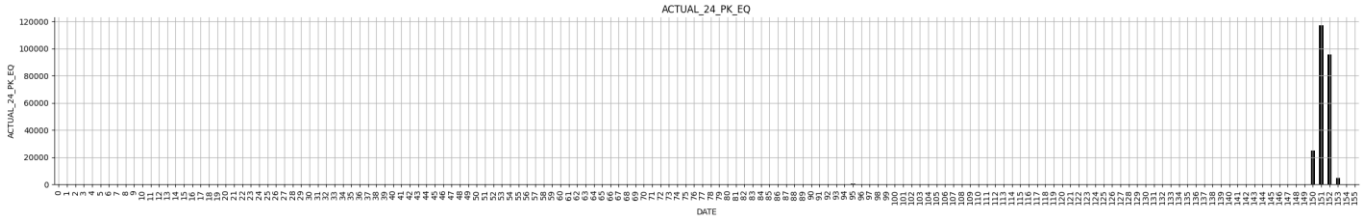
- Cluster A (Stable High-Volume SKUs): Exponential Smoothing consistently provided strong results. Some SKUs benefited slightly from XGBoost_holiday enhancements when holiday effects were significant.
- Cluster B (Stable Lower-Volume SKUs): No single model dominated. Best-fit selection between N-BEATS, ARIMA, and XGBoost models produced better outcomes at the SKU level.
- Cluster C (Erratic & Intermittent SKUs): Forecasting accuracy remained challenging. Even the best statistical and machine learning models showed high proportional errors, requiring supplementary strategies like CPFR and exception management.

4.5 Cluster D Demand Pattern Analysis

Cluster D SKUs exhibit low volume combined with high variability and high intermittency. This Cluster only represents 0.3% of the total weekly volume. These are slow-moving products with low demand or new SKUs that lack sufficient sales history. Figures 29 and 30 are the demand patterns for Cluster D.

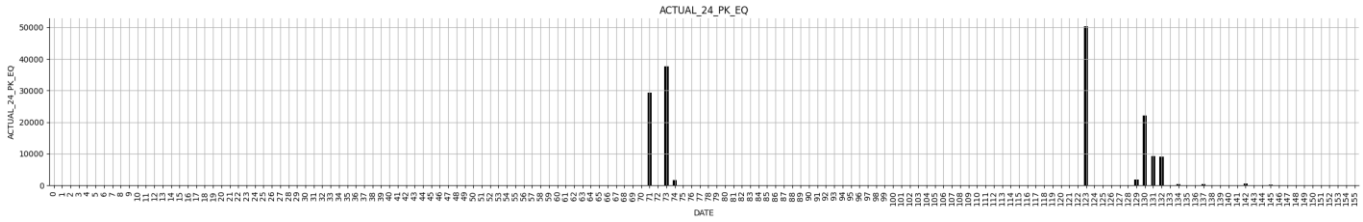
Cluster D SKU1:

Figure 29: Cluster D_SKU1 Demand Pattern



Cluster D SKU2:

Figure 30: Cluster 2_SKU4 Demand Pattern



As shown in the figures above, the demand patterns observed in the plots for highly intermittent SKUs revealed several distinct characteristics. Intermittency was evident, with demand appearing in very few periods and mostly remaining at zero otherwise. Demand bursts were characterized by sudden, large spikes rather than gradual increases. Predictability was extremely low, with no consistent patterns to guide statistical forecasting. Additionally, no clear seasonality or trend was visually detectable, further complicating the ability to model and forecast these SKUs using statistical and ML models. When SKUs display extremely sparse and unpredictable demand and there is no clear pattern or usable history, attempting to forecast demand statistically or through machine learning generates more noise than signal.

The challenges are:

- **No demand pattern:** Forecasting models cannot learn signal.
- **Sparse historical data:** Not enough to train or fit models.
- **No seasonality or trend:** Models default to flat or zero.
- **High variability in spiky size**: High forecast error, even when spikes are predicted.
- **Low frequency:** Does not justify model effort or complexity.

In this case, it makes sense to exclude Cluster D SKUs from statistical and ML forecasting models. Rather than fighting the data, it is better to switch strategies.

4.6 Recommendations and Implications

Based on the analysis and model evaluations across the four clusters, a differentiated forecasting strategy is recommended to optimize forecast accuracy and operational planning as outlined in Table 7.

Table 7: Overall Recommendations for each Cluster

Cluster	Primary Models / Approach	Strategy
Cluster A: Stable High-Volume	-Exponential Smoothing -Second Option: XGBoost_holiday for holiday-affected SKUs	Standardize Exponential Smoothing; selectively enhance with ML models for special cases
Cluster B: Stable Lower-Volume	-N-BEATS -ARIMA	Best-fit model selection at SKU level based on demand pattern characteristics
Cluster C: Erratic & Intermittent	Hybrid Approach: -Statistical models (Exponential Smoothing) -ML models (XGBoost_holiday), supplemented by CPFR, exception-based processes, and monitoring	Use statistical/ML models for directional guidance only; augment with collaborative planning and manual management
Cluster D: Lumpy SKUs: Low Volume, High Intermittence, High Variability	-Collaborative Planning Approach (CPFR)	Forecasting models not suitable; rely on collaboration with customers/sales and manual exception-based forecasting

For Cluster A (Stable High-Volume SKUs), Exponential Smoothing should be standardized as the primary forecasting model due to its consistent strong performance. For SKUs heavily influenced by holiday or promotional events, XGBoost with holiday flags can be selectively applied to capture external demand drivers more effectively. This balanced approach ensures operational efficiency while allowing for incremental improvements where necessary.

For Cluster B (Stable Lower-Volume SKUs), no single model dominated across all SKUs. Therefore, a best-fit model selection approach is recommended at the SKU level, primarily leveraging N-BEATS and ARIMA models depending on the demand pattern. This flexible strategy will allow for more accurate forecasts by aligning model choice to SKU-specific behaviors.

For Cluster C (Erratic and Intermittent SKUs), forecasting models alone were insufficient to achieve high accuracy due to the extreme variability and sporadic nature of demand. A hybrid approach combining statistical models (Exponential Smoothing), machine learning models (XGBoost Holiday), and supplementary processes such as CPFR (Collaborative Planning, Forecasting, and Replenishment), exception management, and monitoring is recommended. Statistical and ML models can provide directional guidance, but collaborative business inputs and manual overrides are critical for managing these SKUs effectively.

For Cluster D (Lumpy SKUs: Low Volume, High Intermittence, High Variability), forecasting models are deemed unsuitable. A full collaborative planning approach (CPFR) with customers and internal stakeholders is recommended. These SKUs should be managed on a case-by-case basis, relying on manual exception-based planning rather than automated statistical forecasting.

Overall, the findings emphasize that a one-size-fits-all forecasting model is not effective across a diversified SKU portfolio. Instead, tailored forecasting strategies aligned to demand characteristics, combined with business process integration where appropriate, offer the best opportunity to enhance forecast accuracy, reduce operational risks, and improve planning efficiency.

4.7 Limitations

One limitation encountered during this project was the computational intensity of advanced deep learning models, particularly N-BEATS. Compared to traditional statistical models such as Exponential Smoothing or ARIMA, and even machine learning models like XGBoost, N-BEATS required significantly more training time and greater computational resources. Due to the model's architectural complexity, running forecasts across multiple SKUs and tuning hyperparameters became increasingly time-consuming and resource-intensive. While N-BEATS showed promise for specific SKU types, its operational scalability may present challenges in real-world deployment without substantial investment in computational infrastructure.

Another limitation was the inability to fully incorporate exogenous variables, such as temperature data, into the forecasting models. Although external factors can meaningfully impact demand patterns, particularly for beverages and seasonal products, access to region-specific temperature data was limited. Available temperature datasets were either too coarse (state-level averages) or not properly aligned with the forecast regions defined in the shipment data. As a result, integrating temperature as a meaningful explanatory variable was not feasible within the project timeline and data constraints, potentially limiting the full predictive power of the machine learning models.

Lastly, while we initially explored the potential impact of weather events (e.g., storms, heat waves) on demand, this analysis was ultimately constrained by data availability and granularity. Specifically, detailed weather event datasets linked precisely to forecast regions were limited. Without sufficient historical coverage or reliable labeling of weather-affected periods, machine learning models could not consistently leverage weather events to enhance forecast performance. Consequently, although the idea of including weather-related features remains attractive, the lack of robust, region-specific datasets limited our ability to validate this enhancement during the project.

5. CONCLUSION

In conclusion, each cluster required a distinct approach based on demand volume, variability and volume. We want to emphasize the importance of aligning forecasting methods with SKU characteristics as follows.

- For Cluster A (Stable High-Volume SKUs), traditional statistical models such as Exponential Smoothing provided the most reliable forecasts, with XGBoost models enhanced by holiday effects performing well for SKUs influenced by promotional or seasonal factors.
- For Cluster B (Stable Lower-Volume SKUs), N-BEATS and ARIMA models delivered the best forecast accuracy depending on SKU-specific demand patterns, highlighting the value of a best-fit model selection strategy.
- Cluster C (Erratic and Intermittent SKUs) required a hybrid approach: using statistical and machine learning models for broad directional guidance, supplemented by Collaborative Planning, Forecasting, and Replenishment (CPFR) processes, exception-based management, and manual overrides to manage operational risks.
- For Cluster D (Lumpy SKUs: Low Volume, High Intermittence, High Variability), statistical and ML models were unsuitable. A pure collaborative planning approach (CPFR) was recommended to manage these highly unpredictable SKUs on a case-by-case basis.

Based on the tested SKUs across Clusters A, B, and C, implementing a Best-Fit Model strategy improved forecast accuracy by 1.56 percentage points in Absolute Percentage Error (APE) compared to the baseline

system forecast, demonstrating the tangible benefits of aligning forecasting models with SKU demand characteristics.

The proposed cluster-based forecasting strategy enables the sponsor company to deploy forecasting models where they are most effective, minimize manual intervention for predictable SKUs, and apply collaborative processes where statistical forecasting is unreliable. This creates a scalable, interpretable, and performance-driven demand planning framework tailored to the company's operational needs.

5.1 Managerial Implications

The project results suggest several implications for demand planning and supply chain management practices.

First, forecasting strategies must be tailored to SKU demand characteristics, rather than applying a one-size-fits-all approach. For stable, high-volume SKUs, managers should focus on standardizing statistical models like Exponential Smoothing and reducing manual interventions wherever possible. Automating forecasts for these predictable SKUs not only improves planning efficiency but also frees up planner time to focus on more complex, high-variability items. At the same time, managers can apply a flexible, SKU-specific modeling strategy for lower-volume or more variable SKUs, ensuring that forecast efforts are prioritized effectively according to demand behavior.

Second, collaborative planning processes are critical for managing erratic and lumpy SKUs. For products with highly intermittent demand, statistical or machine learning forecasts alone are insufficient for reliable planning. Managers should invest in strengthening CPFR practices, building stronger communication channels with sales, marketing, and key customers. Regular review processes and exception-based interventions should be implemented for SKUs where traditional forecasting methods have low predictive power.

Third, technology deployment must balance complexity with operational feasibility. Although advanced models such as N-BEATS and XGBoost can improve forecast accuracy for certain SKU types, they also introduce higher computational requirements and model management complexity. Managers should assess the trade-off between accuracy gains and operational manageability when scaling forecasting systems, ensuring that improvements are sustainable in daily operations without overburdening forecasting teams.

Finally, external data integration (e.g., weather events, temperature effects) remains a promising yet challenging area. While this project demonstrated that incorporating holiday effects improved forecasts, the broader use of exogenous variables is constrained by data accessibility and quality. Managers should plan for the gradual and selective integration of external datasets where strong business justification exists, but avoid overengineering forecasting models when relevant external signals are weak or unreliable.

Overall, the findings emphasize that improving demand planning requires not only better models but also smarter SKU segmentation and process design. Managers must recognize that different SKU types need different forecasting and planning strategies, and that using a uniform process across a diverse portfolio limits forecasting effectiveness. Developing forecasting frameworks tailored to SKU segments, supported by cross-functional collaboration, selective external data integration, and operationally sustainable forecasting practices, will enhance forecast accuracy, resource efficiency, and business scalability.

5.2 Future Work

While implementing advanced deep learning models presents challenges related to computational capacity and data availability, a promising avenue for future development is to conduct small-scale pilot projects focused on high-demand areas. These pilots would create subclusters within each broader demand cluster, allowing for finer segmentation and improved forecasting accuracy where demand concentration is greatest.

By leveraging deep learning with exogenous variables that have been identified as key demand drivers, this approach could enhance baseline predictions while optimizing the use of limited computational resources. As discussed in section 2.5 several studies highlight the importance of such exogenous factors, including climate, holidays, and promotions, in demand forecasting (Mitra et al., 2024; Feng and Foster, 2023). Specifically, climate has been found to significantly impact beverage consumption, with higher temperatures leading to increased demand for bottled water (Keleş et al., 2018; Mirasgedis et al., 2014).

Focusing deep learning efforts on geographic or customer subsegments with strong, documented external demand drivers would not only improve forecast accuracy in pilot areas but also enable better resource allocation. Analysts could prioritize manual adjustments for SKUs outside the pilot zones, ensuring that human forecasting efforts are concentrated where statistical models are less reliable. Over time, successful pilots could be scaled more broadly, supporting a more efficient, targeted, and scalable demand prediction framework.

REFERENCES

Abraham, B., & Ledolter, J. (2009). Statistical methods for forecasting. John Wiley & Sons. https://doi-org.libproxy.mit.edu/10.1002/9780470316610

Adithya Ganesan, V., Divi, S., Moudhgalya, N. B., Sriharsha, U., & Vijayaraghavan, V. (2019, April). Forecasting food sales in a multiplex using dynamic artificial neural networks. In Science and information conference (pp. 69-80). Cham: Springer International Publishing.

Balamwar, A., Mitra, R., Tiwari, M. K., & Verma, P. (2022). Prediction and Analysis of Seasonal Dynamic Metal Consumption using Aggregated LightGBM-A Case Study. *IFAC-PapersOnLine*, *55*(10), 725-730.

Billings, R. B., & Agthe, D. E. (1998). *State-space versus multiple regression for forecasting urban water demand*. *Journal of Water Resources Planning and Management, 124*(2), 113-117.

Chase, C. W. (2013). *Demand-driven forecasting: a structured approach to forecasting*. John Wiley & Sons. https://doi-org.libproxy.mit.edu/10.1002/9781118691861

Falatouri, T., Darbanian, F., Brandtner, P., & Udokwu, C. (2022). Predictive Analytics for Demand Forecasting – A Comparison of SARIMA and LSTM in Retail SCM. Procedia Computer Science, 200, 993–1003. https://doi.org/10.1016/j.procs.2022.01.298

Farizal, F., Dachyar, M., Taurina, Z., & Qaradhawi, Y. "Disclosing Fast Moving Consumer Goods Demand Forecasting Predictor Using Multi Linear Regression." *Engineering and Applied Science Research* 48 (2021): 627636. https://doi.org/10.14456/EASR.2021.64

Feizabadi, J. (2022). "Machine Learning Demand Forecasting and Supply Chain Performance." International Journal of Logistics Research and Applications 25 (2): 119–142. https://doi.org/10.1080/13675567.2020.1803246

Feng, L., and Foster, K. "Demand Forecasting of Consumer Goods for the Indian Subcontinent | Center for Transportation and Logistics," (May 31, 2023). https://ctl.mit.edu/pub/thesis/demand-forecasting-consumer-goods-indian-subcontinent

Folinas, D., and Rabi, S. "Estimating Benefits of Demand Sensing for Consumer Goods Organisations." *Journal of Database Marketing & Customer Strategy Management* 19, no. 4 (December 1, 2012): 245–61. https://doi.org/10.1057/dbm.2012.22

Geerts, R., Vandermoere, F., Van Winckel, T., Halet, D., Joos, P., Van Den Steen, K., Van Meenen, E., Blust, R., & Borregán-Ochando, E. (2020). *Bottle or tap? Toward an integrated approach to water type consumption*. *Water Research, 173*, 115578. https://doi.org/10.1016/j.watres.2020.115578

Harsoor, A. S., and Patil, A. (2015). "Forecast of Sales of Walmart Store Using Big Data Applications." International Journal of Research in Engineering and Technology 4 (6): 51–59. https://doi.org/10.15623/ijret

Hyndman, R., and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice (3rd Ed)*, 2018. https://otexts.com/fpp3/

Kechyn, G., Yu, L., Zang, Y., & Kechyn, S. (2018). "Sales Forecasting Using WaveNet Within the Framework of the Kaggle Competition." Preprint arXiv:1803.04037.

Keleş, B., Gómez-Acevedo, P., & Shaikh, N. I. (2018). The impact of systematic changes in weather on the supply and demand of beverages. *International Journal of Production Economics*, *195*, 186-197.

Kraus, M., Feuerriegel, S., and Oztekin, A. (2020). “Deep Learning in Business Analytics and Operations Research: Models, Applications and Managerial Implications.” European Journal of Operational Research 281 (3): 628–641. https://doi.org/10.1016/j.ejor.2019.09.018

Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019, August). Product marketing prediction based on XGboost and LightGBM algorithm. In Proceedings of the 2nd international conference on artificial intelligence and pattern recognition (pp. 150-153). https://doi.org/10.1145/3357254.3357290

Loureiro, A. L., Miguéis, V. L., & Da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, *114*, 81-93. https://doi.org/10.1016/j.dss.2018.08.010

Ma, S., and Fildes, R. (2021). “Retail Sales Forecasting with Metalearning.” European Journal of Operational Research 288 (1): 111–128. https://doi.org/10.1016/j.ejor.2020.05.038

Ma, S., Fildes, R., and Huang, T. (2016). “Demand Forecasting with High Dimensional Data: The Case of SKU Retail Sales Forecasting with Intra-and Inter-category Promotional Information.” European Journal of Operational Research 249 (1): 245–257. https://doi.org/10.1016/j.ejor.2015.08.029

Mirasgedis, S., Georgopoulou, E., Sarafidis, Y., Papagiannaki, K., & Lalas, D. P. (2014). The Impact of Climate Change on the Pattern of Demand for Bottled Water and Non-Alcoholic Beverages. Business Strategy and the Environment, 23(4), 272-288. https://doi.org/10.1002/bse.1782

Mitra, R., Saha, P., & Kumar Tiwari, M. (2024). Sales forecasting of a food and beverage company using deep clustering frameworks. *International Journal of Production Research*, *62*(9), 3320–3332. https://doi.org/10.1080/00207543.2023.2231098

Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Kompatsiaris, I. (2022). “Multimodal Quasi-AutoRegression: Forecasting the Visual Popularity of New Fashion Products.” International Journal of Multimedia Information Retrieval 11 (1): 1–13. https://doi.org/10.1007/s13735-022-00262-5

Pavlyshenko, B. M. (2019). “Machine-learning Models for Sales Time Series Forecasting.” Data 4 (1): 1–5. https://doi.org/ 10.3390/data4010015

Pan, H., and Zhou, H. (2020). “Study on Convolutional Neural Network and Its Application in Data Mining and Sales Forecasting for E-commerce.” Electronic Commerce Research 20 (2): 297–320. https://doi.org/10.1007/s10660-020-09409-0

Pliszczuk, D., Lesiak, P., Zuk, K., & Cieplak, T. (2021). “Forecasting Sales in the Supply Chain Based on the LSTM Network: The Case of Furniture Industry.” European Research Studies 24 (2): 627–636. https://doi.org/10.35808/ersj/2291

Rožanec, J. M., Fortuna, B., & Mladenić, D. (2022). Reframing Demand Forecasting: A Two-Fold Approach for Lumpy and Intermittent Demand. Sustainability, 14(15), 9295. https://doi.org/10.3390/su14159295

Sagaert, Y. R., Aghezzaf, E. H., Kourentzes, N., and Desmet, B. (2018). “Temporal Big Data for Tactical Sales Forecasting in the Tire Industry.” Interfaces 48 (2): 121–129. https://doi.org/10.1287/inte.2017.0901

Seifert, D. (2003). *Collaborative Planning, Forecasting, and Replenishment: How to Create a Supply Chain Advantage*. AMACOM. https://app.knovel.com/hotlink/pdf/id:kt0097RZZ1/collaborative-planning/efficient-consumer-response

Tehrani, A. F., and Ahrens, D. (2016). “Improved Forecasting and Purchasing of Fashion Products based on the Use of Big Data Techniques.” In Supply Management Research, 293–312. Wiesbaden: Springer Gabler.

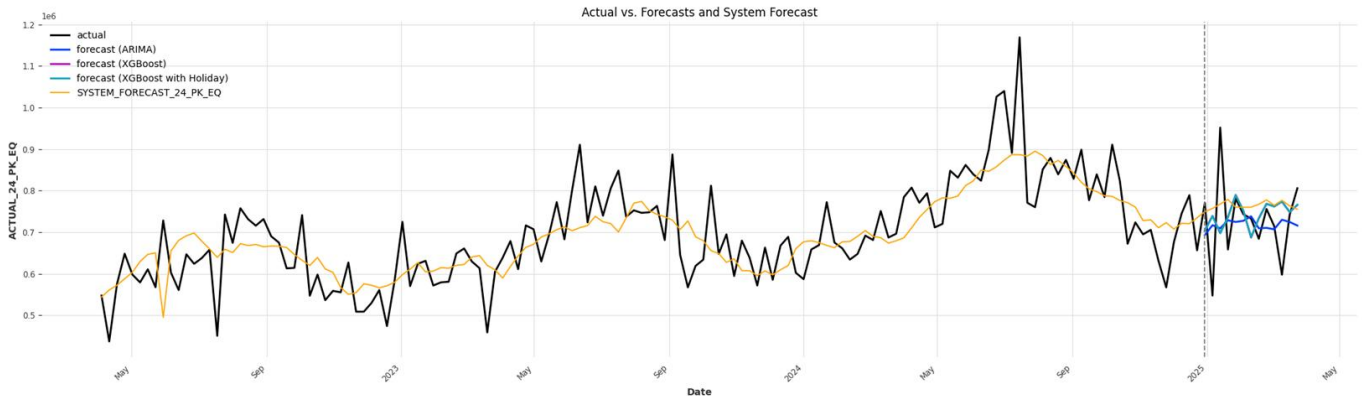
Tseng, C. S., and Turkmen, T. (2024). Demand Forecasting with Machine Learning (Doctoral dissertation, MASSACHUSETTS INSTITUTE OF TECHNOLOGY).

Vandeput, N. (2023). *Demand Forecasting Best Practices*. Simon and Schuster. https://www.oreilly.com/library/view/demand-forecasting-best/9781633438095/

Zhao, K., and Wang, C. (2017). “Sales Forecast in E-Commerce Using Convolutional Neural Network.” Preprint arXiv:1708. 07946. https://learning.oreilly.com/library/view/demand-forecasting-best/9781633438095/h

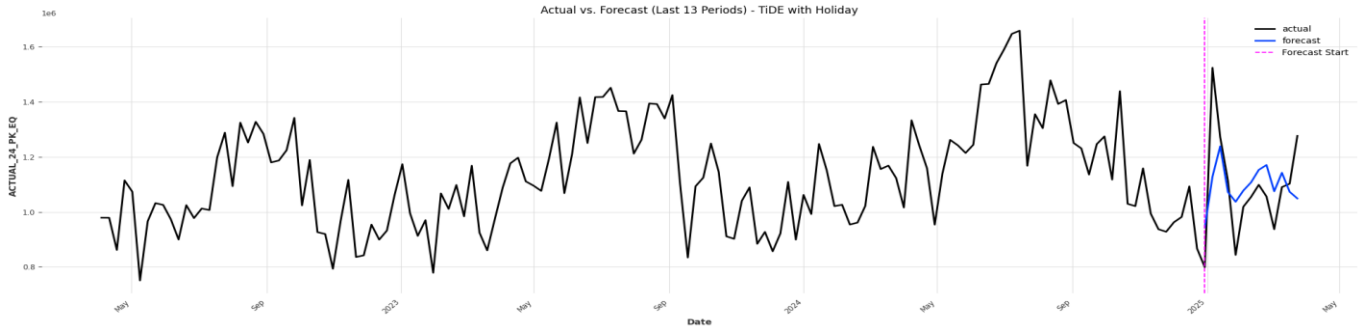
APPENDIX

APPENDIX A.1: Cluster A SKU3: Forecast results of Cluster A_SKU3

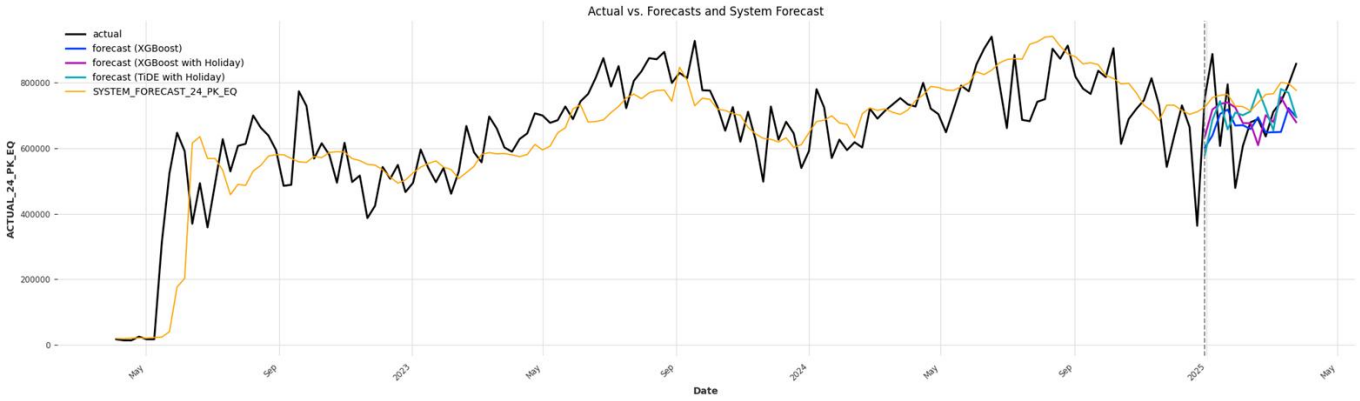


Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	14.50	13.52	98601.24	118931.79
ARIMA	10.29	10.08	73529.56	100260.45
XGBoost	10.88	10.38	75729.64	107939.13
XGBoost_holiday	10.86	10.36	75585.47	107929.07
TiDE	20.80	19.19	139945.93	199389.39
TiDE_holiday	14.54	14.23	103801.84	131767.84
System Forecast	11.44	10.62	77488.74	103522.39

APPENDIX A.2: Cluster A SKU3: Best-Fit Model – TiDE with Holiday

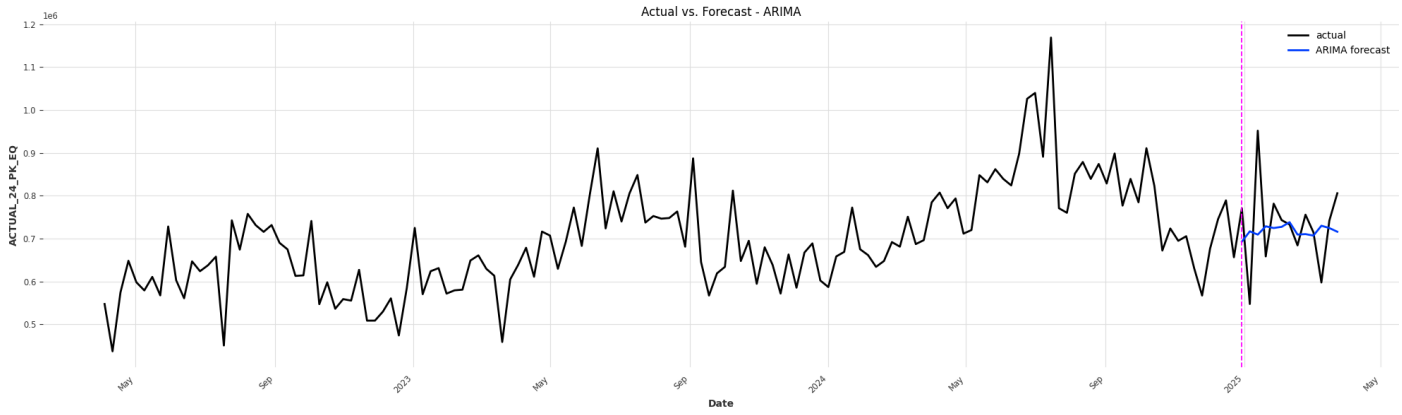


APPENDIX A.3: Forecast results of Cluster A SKU4

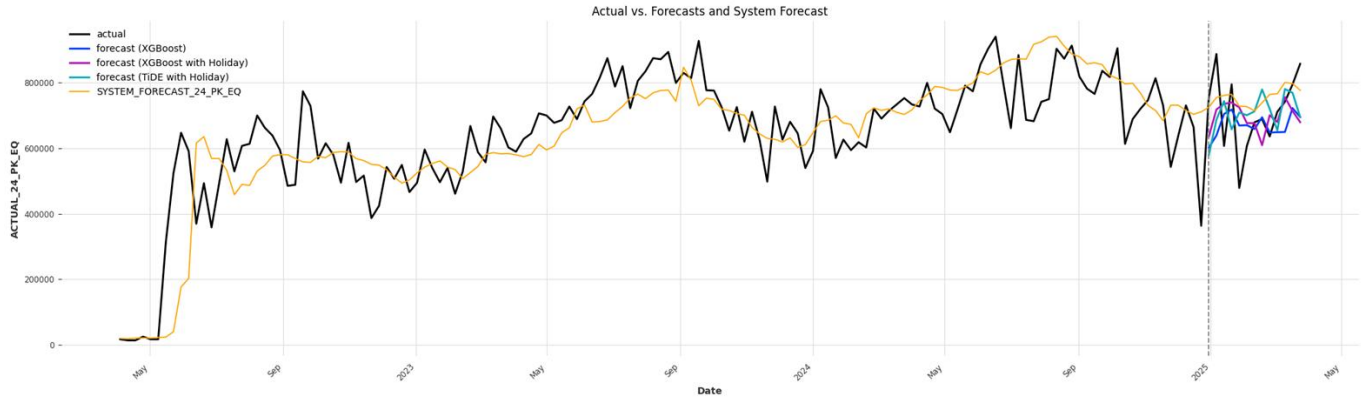


Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	20.58	21.58	153191.74	186070.46
ARIMA	45.03	47.20	335056.79	370218.15
XGBoost	13.78	13.60	96536.79	119240.20
XGBoost_holiday	14.23	13.40	95114.09	116639.01
TiDE	26.90	26.75	189872.62	214465.34
TiDE_holiday	16.58	15.85	112504.89	129387.07
System Forecast	13.69	12.20	86609.28	108362.33

APPENDIX A.4: Cluster A SKU4: Best-Fit Model – ARIMA

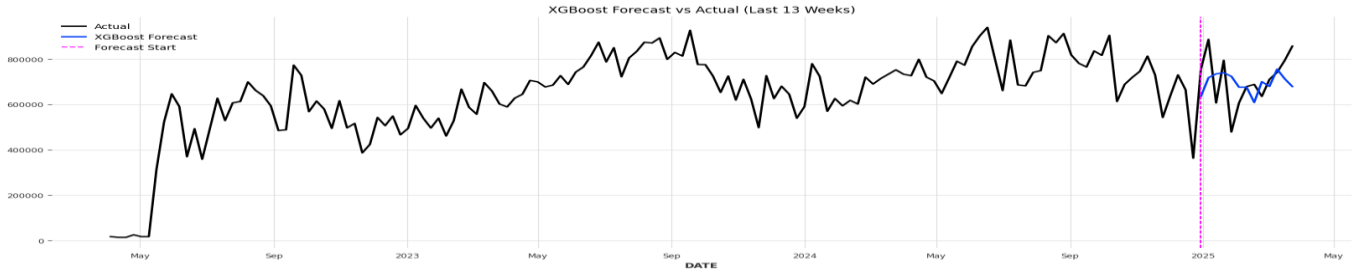


APPENDIX A.5: Forecast results of Cluster A SKU5

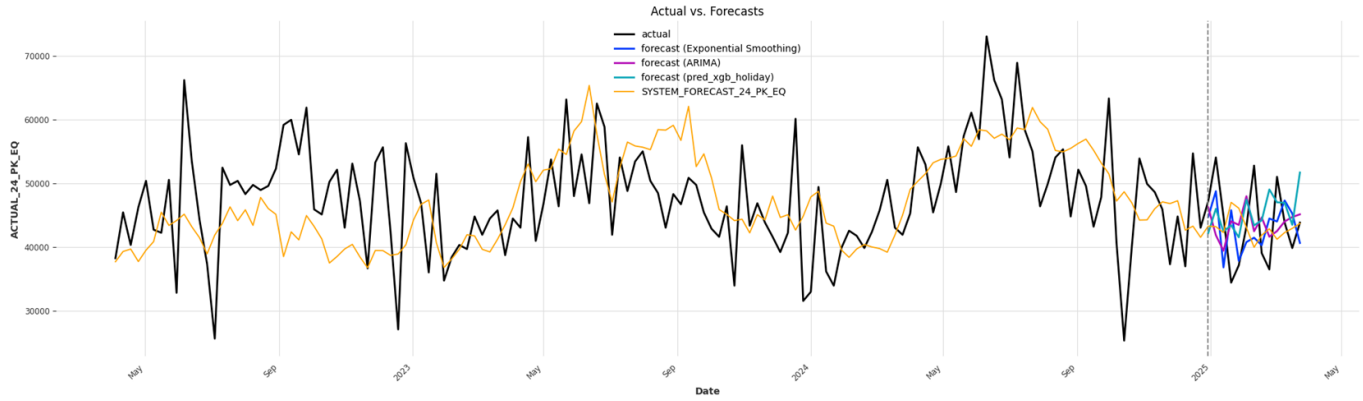


Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	20.58	21.58	153191.74	186070.46
ARIMA	45.03	47.20	335056.79	370218.15
XGBoost	13.78	13.60	96536.79	119240.20
XGBoost_holiday	14.23	13.40	95114.09	116639.01
TiDE	26.90	26.75	189872.62	214465.34
TiDE_holiday	16.58	15.85	112504.89	129387.07
System Forecast	13.69	12.20	86609.28	108362.33

APPENDIX A.6: Cluster A SKU5: Best-Fit Model – XGBoost

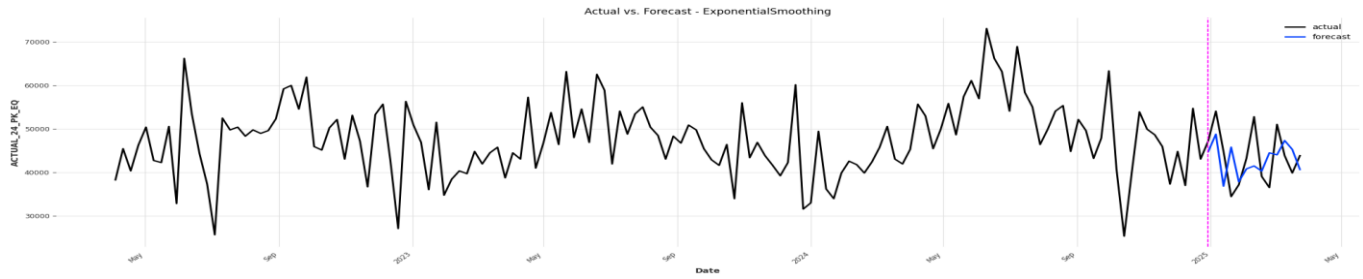


APPENDIX A.7: Forecast results of Cluster B SKU3

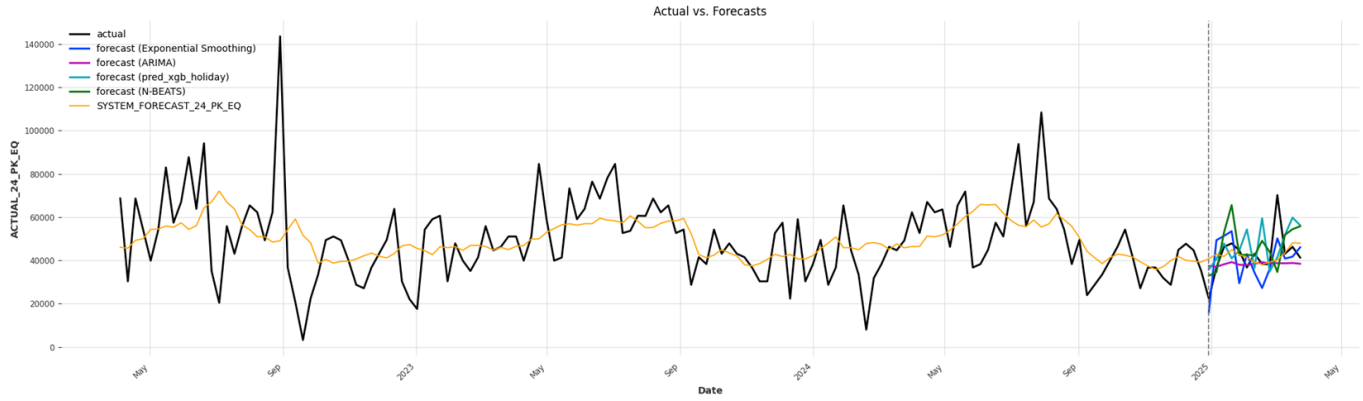


Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	12.51	12.32	5381.54	6381.25
ARIMA	13.32	13.26	5790.52	6782.22
XGBoost	15.17	14.75	6439.67	6984.58
XGBoost_holiday	14.33	13.90	6068.35	6729.57
TiDE	22.47	21.95	9586.18	12080.32
TiDE_holiday	15.58	14.70	6419.79	7526.31
N-BEATS	14.36	13.89	6065.56	6789.06
System Forecast	13.48	13.34	5827.43	7351.29

APPENDIX A.8: Cluster B SKU3: Best-Fit Model – Exponential Smoothing

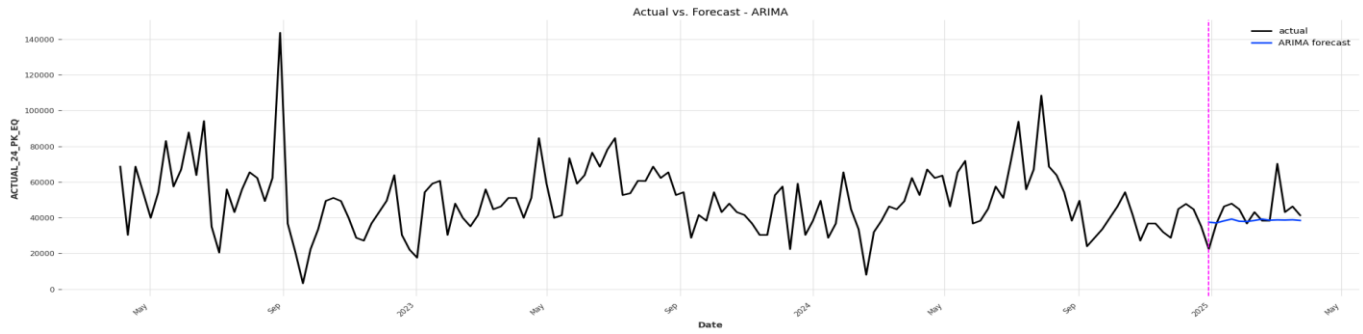


APPENDIX A.9: Forecast results of Cluster B SKU4

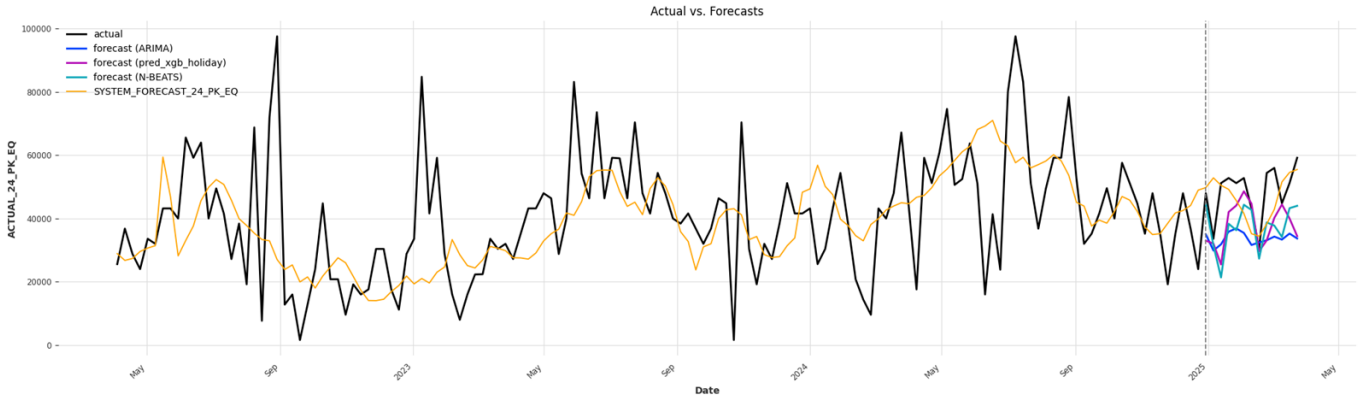


Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	18.99	18.87	8057.80	9533.58
ARIMA	16.47	16.58	7082.80	10782.82
XGBoost	26.41	25.37	10835.61	13508.06
XGBoost_holiday	26.41	25.37	10835.34	13509.17
TiDE	32.83	34.50	14734.06	21253.87
TiDE_holiday	32.68	31.86	13607.08	15234.93
N-BEATS	20.88	23.13	9880.74	14567.33
System Forecast	16.02	14.89	6360.54	10369.84

APPENDIX A.10: Cluster B SKU4: Best-Fit Model – ARIMA



APPENDIX A.11: Forecast results of Cluster B SKU5



Model	MAPE	APE	MAE	RMSE
Exponential Smoothing	30.98	28.01	13546.73	19885.55
ARIMA	29.22	30.87	14933.37	16283.51
XGBoost	20.62	22.17	10724.44	13608.49
XGBoost_holiday	20.45	22.25	10760.26	13909.70
TiDE	37.13	38.25	18502.80	22884.03
TiDE_holiday	43.34	42.25	20434.69	24796.65
N-BEATS	18.12	18.17	8790.64	13150.24
System Forecast	16.45	15.45	7472.60	9306.28

APPENDIX A.12: Cluster B SKU5: Best-Fit Model – NBeats