Decoding Pharmaceutical Demand: A Scalable Framework for Benchmarking Internal Forecasts

by

KEVIN PILZ

and

YI-EN WONG

Submitted on May 8th, 2026 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

at the Massachusetts Institute of Technology

ABSTRACT

Pharmaceutical demand forecasting is often constrained by limited visibility into true market demand at the country level and other phenomena, which reduces the effectiveness of planning, increases costs, and potentially impacts companies' ability to meet patient health needs. This project addresses the question of how to extract meaningful demand signals and identify the most appropriate forecasting models across different product portfolios and countries. A scalable, data-driven forecasting framework was developed integrating demand decomposition, statistical methods, and machine learning models to process up to 36 months of historical data and benchmark performance of internal forecasts. Results show that while overall forecast accuracy is generally strong, significant improvements can be achieved for low-accuracy products, with alternative models delivering substantial gains and reducing mis-forecasted volumes. The findings support a targeted deployment approach, recommending that our sponsor company apply the framework to identify high-impact opportunities, combine model outputs with local expertise, and tailor forecasting strategies based on product maturity and data availability.

Capstone Advisor: Jarrod Goentzel

Title: Director, MIT Humanitarian Supply Chain Lab Principal Research Scientist Lecturer

TABLE OF CONTENTS

1. INTRODUCTION 3

1.1 Motivation 3

1.2 Problem Statement and Key Questions 4

1.3 Project Goals and Expected Outcomes 4

2. STATE OF THE PRACTICE 5

2.1 Pharmaceutical Forecasting and Demand Planning 5

2.2 Time Series Decomposition 6

3. METHODOLOGY 7

3.1 Methodology Selection 7

3.2 Data Collection and Preparation 8

3.3 Methodology Application 9

4. RESULTS AND DISCUSSION 10

4.1 Baseline Forecast Performance and Problem Diagnosis 11

4.2 Model Benchmarking Results (Spain and Portugal) 11

4.3 Scaling Results Across 28 Countries in Europe 14

4.4 Quantifying Business Impact 16

4.5 Limitations 17

5. RECOMMENDATIONS 17

5.1 Management Recommendations 18

5.2 Future Work 19

REFERENCES 20

1. INTRODUCTION

With over a hundred years of history and a global workforce of over 100,000 employees, the sponsor company, referred to as "Kevian Pharma" (hereafter, Kevian to preserve confidentiality), is a multinational healthcare company operating in more than 90 countries. Kevian focuses on both pharmaceuticals and diagnostics, with approximately 75% of its revenue generated from pharmaceuticals and 25% from diagnostics. In 2024, Kevian delivered treatments and diagnostic solutions to more than 25 million patients worldwide. The company also continued to expand its laboratory business, investing $10 billion dollars in research and development and conducting over millions diagnostic tests.

This project is sponsored by the Supply Chain Insights Department within Kevian's pharmaceutical division. The department plays a critical role in supporting Kevian's complex, multi-tier global supply chain by analyzing data to identify trends and insights that can improve business performance before issues arise. In this project, we partner with the department to strengthen how pharmaceutical sales and forecast data from country affiliates are used to improve global demand planning.

Kevian's global supply network consists of two primary market structures. In affiliate markets, Kevian operates through in-country affiliates that sell to local wholesalers and distributors. In direct markets, Kevian sells directly to wholesalers without an affiliate intermediary. In both cases, Kevian has visibility into "sell-in" data (sales to wholesalers) but limited access to "sell-out" data (sales from wholesalers to hospitals and pharmacies). This lack of downstream visibility creates challenges in accurately understanding true market demand.

Kevian's manufacturing process further adds complexity. Production begins with Drug Substance (active pharmaceutical ingredients), which are produced at chemical or biotechnological sites. These materials are then transferred to Drug Product facilities, where they are formulated and packaged into bulk products. Finally, products are sent to Finished Goods centers for labeling, serialization, and distribution to affiliates or wholesalers. This multi-stage value chain includes several decoupling points, where planning shifts from forecast-driven (push) to demand-driven (pull). These transitions contribute to variability and noise in demand signals, which will be discussed further in Section 1.2.

1.1 Motivation

As a global pharmaceutical manufacturer, Kevian must balance supply and demand across diverse and dynamic markets. Accurate demand forecasting is critical, not only for financial performance but also for ensuring patients receive timely access to medicines. Underestimating demand can lead to product shortages, delayed patient treatment, lost revenue, and increased operational costs due to emergency production and expedited shipments. Overestimating demand, on the other hand, can result in excess inventory, product expiration, inefficient use of manufacturing capacity, and increased reallocation costs.

A key challenge lies in limited visibility across Kevian's multi-tier supply chain, particularly within affiliate markets, which account for the majority of its global operations and revenue. This project focuses exclusively on these affiliate markets and does not include direct markets where Kevian sells directly to wholesalers. In affiliate markets, demand signals originate at the country level and are shaped by orders from wholesalers, pharmacies, and patients. However, Kevian does not directly observe downstream consumption. Instead, forecasting relies on data processed by Supply Leads and Forecasting Leads, who must interpret affiliate orders that are often irregular and batch-driven. These ordering patterns can distort true demand and exhibit characteristics of the bullwhip effect.

The motivation for this project is to improve how Kevian captures and uses demand signals from country affiliates to enhance forecasting accuracy at both global and market levels. By strengthening the use of

affiliate data, this project aims to improve forecast accuracy, enhance service levels, and increase cost efficiency across Kevian's global supply chain.

1.2 Problem Statement and Key Questions

The central objective of this project is to determine how to extract meaningful and reliable demand signals from affiliate data to improve market-level and global demand planning. This includes developing analytical approaches that enhance forecast accuracy and supply chain responsiveness.

This project is structured to focus on defining a manageable scope by specific markets and products. The analysis will examine historical sales data to identify patterns and evaluate how different variables influence demand signals derived from affiliate orders. The goal is to develop descriptive insights and initial analytical models that better represent true demand. Based on the insights, this project further expands the analysis to additional markets and identifies the methods and models that perform best for specific countries and products.

The project is guided by two key questions: How can true market demand be extracted from country demand signals to improve forecast accuracy (FA)? What signal decomposition and forecasting approaches work best for predicting future demand?

Overall, the project follows a structured workflow that moves from data exploration and model development to validation and scalability.

1.3 Project Goals and Expected Outcomes

Kevian currently achieves relatively strong forecasting performance at an aggregate level. When all product stock keeping units (SKUs) are combined, average forecast accuracy ranges from 60% to 90% across European countries. The goal of this project is to improve how demand data is processed and utilized across Kevian's global supply chain, with a focus on enhancing forecast accuracy at the country and product level. By refining affiliate-level demand signals into more accurate and actionable inputs, the project supports more effective upstream planning and overall supply chain responsiveness.

To achieve this, the project develops a scalable forecasting framework that integrates data pipelines including data decomposition, statistical and time series forecasting methods as well as machine learning models. Using historical data from the past 36 months, the data pipeline generates monthly forecasts for the following year. These forecasts are then compared against actual outcomes to evaluate performance and identify the most suitable models for different SKUs across different countries.

The Initial phase focuses on Portugal and Spain, where the analysis examines product portfolios including both lower-accuracy SKUs and higher-accuracy SKUs. This allows for a more targeted understanding of where improvements are most needed and which modeling approaches are most effective.

Once the forecasting framework is validated and streamlined, it is scaled and applied to 28 countries across Europe. The objective is to demonstrate that the approach can be consistently and efficiently implemented across different markets and product families, as long as sufficient data is available.

Ultimately, the project delivers a flexible and scalable analytical framework that enables Kevian to systematically identify opportunities to improve forecast accuracy. Additionally, the outputs of this work can be combined with the deep market knowledge of country affiliates (including insights into product characteristics, epidemiology, and local market dynamics) to support more informed decision-making and stronger alignment between global planning and local execution.

2. STATE OF THE PRACTICE

We review current practices and research relevant to pharmaceutical demand forecasting, focusing on two key areas: (1) pharmaceutical forecasting and demand planning, and (2) time series decomposition. These areas provide the foundation for the project's objective of improving how demand signals from country affiliates are interpreted and used in forecasting. Pharmaceutical forecasting is particularly complex because observed demand is often distorted by operational processes such as batch ordering and replenishment cycles. Prior research shows that decomposition techniques can help separate meaningful demand signals from noise, providing a clearer view of underlying demand patterns (Hyndman & Athanasopoulos, 2021; Petropoulos et al., 2022).

The literature also indicates that no single forecasting method is sufficient for pharmaceutical demand planning. Studies such as Bilal et al. (2025) and Sazvar et al. (2022) highlight the need to combine multiple approaches due to demand variability and operational complexity. Traditional time series models provide a strong and interpretable baseline (Hyndman & Athanasopoulos, 2021), while machine learning models have been shown to deliver improved performance in more complex and data-rich environments (Van Belle et al., 2021; Fourkiotis & Tsadiras, 2024). In addition, decomposition techniques play a critical role in improving signal quality before forecasting by separating trend, seasonal, and residual components (Hyndman & Athanasopoulos, 2021; Petropoulos et al., 2022).

These insights are consistent with the findings of this project. Model performance varies depending on product lifecycle, data availability, and demand variability. Based on this foundation, the project adopts a structured framework to systematically evaluate and identify the best-performing model for each scenario. This approach enables more accurate interpretation of demand signals and supports improved forecast accuracy across Kevian's global supply chain.

2.1 Pharmaceutical Forecasting and Demand Planning

The pharmaceutical supply chain has unique characteristics that make forecasting particularly challenging. These include short product shelf life, strict service-level expectations, and significant variability across markets and products. In addition, affiliate ordering behavior often reflects internal processes (such as replenishment cycle, inventory policies, and ordering constraints) rather than true market demand.

Research shows that relying only on historical data is often insufficient in this environment. Studies such as Bilal et al. (2025) demonstrate that incorporating operational and contextual factors, such as replenishment cycles, inventory processes, and seasonality, can significantly improve forecast accuracy.

Similarly, Sazvar et al. (2022) emphasize the importance of accounting for uncertainty and variability in pharmaceutical demand. Their work shows that demand patterns are often influenced by external and structural factors, making scenario-based thinking and sensitivity analysis important for decision making.

These findings are consistent with the challenges observed in this project. Interviews and data analysis indicate that demand signals from affiliates are often delayed, incomplete, or influenced by operational processes. That, in turn, limits forecast accuracy. As a result, this project adopts a combined approach that integrates statistical models, machine learning methods, and signal processing techniques to better capture the true demand patterns which are described in Sections 2.1.1 through 2.1.3.

2.1.1 Time-Series and Context-Augmented Approaches

Traditional time series models remain widely used due to their simplicity and interpretability. Common methods include autoregressive integrated moving average (ARIMA), exponential smoothing, and moving averages. These models perform well when demand patterns are relatively stable and consistent over

time (Zahra & Putra, 2019). ARIMA models are particularly useful for capturing trends and autocorrelation in time series data. They adjust for changes in demand patterns through autoregressive and differencing components, making them effective in structured environments.

However, their limitations become more visible in environments with irregular demand patterns, such as those created by batch ordering. Since these models rely heavily on historical data, they may not fully capture operational or market-driven changes. This limitation is reflected in the findings. While ARIMA performs well for products with limited history and with strong autocorrelation (where recent trends are self-reinforcing), its performance varies across different product segments and market conditions.

2.1.2 Forecast Accuracy and Uncertainty Diagnostics

Evaluating forecast performance requires consistent and comparable metrics across products with different demand levels. Literature recommends using scale-independent metrics to enable meaningful comparisons, as well as measures of bias such as evaluating the accuracy of the forecast distribution to reward models that are both precise and well-calibrated (Hyndman & Athanasopoulos, 2021).

These metrics are directly applied in this project's framework to compare model performance across countries and product segments. This enables consistent benchmarking and supports the identification of the most suitable model for each scenario.

2.1.3 Machine Learning Forecasting Models

Machine learning models are increasingly used in pharmaceutical forecasting due to their ability to handle complex and nonlinear relationships. These models are particularly useful in environments with high variability and large datasets. Studies such as Van Belle et al. (2021) and Fourikiotis and Tsadiras (2024) demonstrate that tree-based models, such as extreme gradient boosting (XGBoost) and Random Forest, can outperform traditional approaches in volatile demand conditions.

XGBoost builds a sequence of decision trees where each iteration improves on previous errors, making it effective in capturing complex patterns, handling missing data, and reducing overfitting. Similarly, Random Forest constructs multiple decision trees using different subsets of data and features, and aggregates their predictions to improve robustness and reduce variance. These characteristics make both models well suited for forecasting in environments with noisy and irregular demand patterns.

These findings align with the results observed in this project. Machine learning models, particularly tree-based methods, perform well for products with sufficient historical data and more stable demand patterns. They outperform traditional methods in a majority of cases, especially for established product portfolios.

However, these models also require more data and are less transparent, which can limit their applicability in early product lifecycle stages or data-constrained environments.

2.2 Time Series Decomposition

Time series decomposition refers to breaking down a series of data points collected over time into underlying time series, or "components." Decomposition of time series data dates back to the 1800s (Petropoulos et al., 2022) and has a wide array of applications. Decomposition is used to better understand a signal's meaning and how it changes over time, enabling better prediction of future events. Time series are typically decomposed into three components: trend, or the overall increase or decrease in signal; seasonality, or a cyclical pattern of change in the signal; and the remainder or residual, which is the remaining term that is not explained by the other components (Hyndman & Athanasopoulos, 2021).

A key advantage of decomposition is that it relies only on historical data, making it applicable even when external data is limited. It can be used as a preprocessing step before applying forecasting models to

better detect true signals and differentiate them from residual noise. This is directly reflected in the project's methodology. The forecasting pipeline tests the value of decomposition techniques to clean and structure the demand signal before running forecasting models.

Among the most common decomposition approaches used for forecasting are Seasonal Trend Decomposition using Loess (STL) and the Kalman filter. STL repeatedly applies smoothing for seasonality and trend to provide more accurate and robust estimates of those components (Petropoulos et al., 2022). The Kalman filter produces the optimal smoothed estimate of the true underlying signal under standard assumptions (Hyndman et al., 2002) and is well suited for Kevian's affiliate order and sales data because it naturally accommodates missing observations and irregular data patterns. These methods are effective for the affiliate order data (which contains noise, irregular patterns, and data gaps) used in this project.

3. METHODOLOGY

Building on the challenges identified in the Introduction and State of Practice (particularly the challenges in capturing accurate demand signals at the affiliate level), the methodology develops a unified modeling approach that integrates data preprocessing, signal decomposition, statistical forecasting, and machine learning. This approach is operationalized through a structured and scalable framework implemented in Python. Because of the data pipeline in this framework, data is systematically cleaned, transformed, and organized before multiple forecasting models are applied and evaluated. The process is designed to be consistent and repeatable across product–country combinations using historical demand data. Furthermore, raw affiliate data is transformed into clearer demand signals, enabling robust comparison across multiple modeling techniques and contexts. As a result, the methodology not only improves forecast accuracy but also provides a data-driven foundation for selecting the most appropriate forecasting approach for each product and market.

3.1 Methodology Selection

Consistent with the literature reviewed in the State of Practice, this project adopts a comparative modeling framework rather than relying on a single forecasting technique. Prior research indicates that pharmaceutical demand patterns are highly heterogeneous across products and markets, and therefore require flexible modeling approaches. In response, this methodology integrates time series models, machine learning models, and demand decomposition to capture different demand patterns. Statistical models such as ARIMA and Exponential Smoothing provide interpretable baselines, while machine learning models such as Random Forest and XGBoost are capable of capturing nonlinear relationships and complex temporal dependencies. Signal decomposition methods, including STL and Kalman filtering, are incorporated to improve signal quality by separating trend, seasonality, and noise components.

Based on our discussion with Kevian point-of-contacts (POCs) and subject matter experts (SMEs), our methodology is designed to (1) understand Kevian's forecasting, demand, and ordering processes at the country and global level; (2) develop analytical processes that reveal meaningful demand information embedded in affiliate ordering patterns; (3) evaluate uncertainty and diagnose noise sources through decomposition; and (4) design and validate forecasting analytics that are capable of supporting upstream planning and decision-making at multiple decoupling points.

The overall methodology is operationalized as a modular data pipeline implemented in the forecasting framework. This data pipeline consists of sequential stages, including data preprocessing, feature engineering, decomposition, model training, forecast generation, and performance evaluation. A key design objective of the framework is scalability, allowing it to be applied consistently across products and

countries. This aligns with the project's goal of developing a reusable analytical framework that can be deployed across Kevian's global supply chain to identify forecasting improvement opportunities.

In summary, this methodology establishes a scalable and data-driven forecasting framework that integrates decomposition, statistical modeling, and machine learning to improve pharmaceutical demand forecasting. By systematically transforming raw affiliate data into cleaner demand signals and benchmarking multiple models across different contexts, the approach enables the identification of the most appropriate forecasting method for each product and market. This framework not only improves forecast accuracy but also provides a structured way to evaluate and deploy forecasting models at scale.

3.2 Data Collection and Preparation

The analysis is based on structured pharmaceutical demand data extracted from Kevian systems and processed through a Python-based forecasting framework. The dataset includes country, product family, product name, presentation, snapshot month, demand month, quantity in packs, and covers different time horizons across regions. For the prototype countries, Spain and Portugal, the data spans from 2021 to 2028 and includes 138 SKUs (not all of which are used in the analysis), while the broader European dataset spans from 2023 to 2029 and includes 32 SKUs.

During data ingestion, time-related fields are standardized and two key variables are constructed. The snapshot date is defined as the point in time at which the data extract is taken from the planning system, representing the state of forecasts and information available at that moment. The sales date represents the demand period being observed or forecasted. Using these 2 variables, each observation is classified as either historical sales or a system forecast depending on whether the demand period occurs before or after the snapshot date. This separation ensures that only historical sales are used for model training, thereby preventing data leakage and preserving real-world forecasting conditions.

Several data preprocessing steps are implemented to address common issues in pharmaceutical demand data. Negative values, typically driven by product returns, are adjusted by clipping them to zero to reflect non-negative demand. The dataset is also validated at the full granularity of country, product, presentation, snapshot date, and sales date to ensure uniqueness of observations.

Missing values are handled carefully: instead of being systematically imputed, they are preserved in the raw time series to avoid distorting the underlying demand signal. Interpolation is applied only in specific cases where model requirements necessitate a continuous time index. For example, certain models such as ARIMA require evenly spaced time steps and therefore require missing months to be explicitly represented, in which case missing periods are introduced with zero values. In contrast, models such as Prophet can natively accommodate gaps in the time series and do not require such adjustments. This conditional handling ensures that data is modified only when required for model stability, which minimizes unnecessary smoothing. Additionally, outliers are not automatically removed, as extreme values often reflect real operational behavior, such as batch ordering, rather than data errors.

To construct the modeling dataset, the data pipeline generates a monthly time series for each country–product–presentation combination using historical sales observations. Each series is reindexed to a complete monthly calendar, allowing missing months to be explicitly identified. For each time series, summary statistics such as mean, standard deviation, and coefficient of variation are computed to characterize demand variability. To ensure sufficient data quality for modeling, a minimum threshold of historical observations is required, and the proportion of missing months is calculated for each series. These metrics are later used to determine whether a series is suitable for seasonal analysis.

The dataset is then split into training and testing sets using a time-based approach. Models are trained on historical data of up to 36 months and evaluated on the most recent 12 months. This structure mirrors

real forecasting practice and ensures that model performance is assessed on unseen future data. For machine learning models, additional features are engineered from the time series, including lag variables, rolling statistics, and calendar-based features such as month indicators. These features allow the models to capture temporal dependencies, short-term fluctuations, and potential seasonal effects, thereby enhancing predictive performance.

3.3 Methodology Application

The application of the methodology follows a structured sequence of analytical steps, beginning with exploratory analysis and progressing through decomposition, forecasting, and model evaluation. The exploratory analysis examines key statistical properties of the demand data, including central tendency, variability, and temporal patterns. Particular attention is given to identifying seasonality and volatility, as these characteristics strongly influence model performance. Forecast accuracy is evaluated using Mean Absolute Percentage Error (MAPE), defined as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where A_t represents actual demand and F_t represents forecasted demand. This metric is implemented directly in the data pipeline and applied consistently across all models to ensure comparability.

A key component of the methodology is the classification of demand patterns based on seasonality. Rather than assuming all products exhibit seasonal behavior, the data pipeline evaluates each time series using a structured diagnostic approach. A series is considered eligible for seasonal modeling only if it has sufficient historical length (at least 24 months) and acceptable data completeness (no more than 25% missing months). For eligible series, seasonality is assessed using multiple indicators, including seasonal strength, consistency of monthly patterns, and autocorrelation at a 12-month lag. Based on these criteria, each product is classified as seasonal, weakly seasonal, or non-seasonal. This classification directly determines the appropriate decomposition and forecasting approach for each case.

To address the issue of noisy demand signals, signal decomposition is applied as an intermediate step. For non-seasonal products, a Kalman filter is applied to estimate a smoothed underlying trend. For seasonal and weakly seasonal patterns, two different decomposition methods are independently applied and tested for impact. For products exhibiting clear seasonal patterns, both the STL decomposition and the Kalman seasonal filter are applied to separate the time series into trend (T_t), seasonal (S_t), and residual (R_t) components, following the formulation:

$$Y_t = T_t + S_t + R_t$$

For weakly seasonal products, both seasonal and non-seasonal Kalman models are tested to determine whether explicitly modeling seasonality improves performance. These decomposition methods are implemented within the data pipeline and applied conditionally, ensuring that model inputs are tailored to the characteristics of each time series. The primary objective of decomposition is to segment the variation to improve the signal quality before forecasting.

The forecasting stage applies multiple models to both raw and decomposed data, enabling a comprehensive comparison of modeling approaches. Statistical models include Exponential Smoothing, ARIMA, and Prophet. Exponential Smoothing captures level and trend dynamics, while ARIMA is driven by linear temporal dependencies and autocorrelation structures. Prophet provides a flexible data pipeline for handling trend and seasonal components. In addition, machine learning models (Linear Regression, Random Forest, and XGBoost) are implemented using engineered features derived from the time series. These models are trained using a recursive forecasting strategy, where predicted values are fed back into

the model to generate multi-step forecasts. This approach reflects real forecasting conditions and allows machine learning models to capture both short- and longer-term dependencies.

To ensure robustness, the data pipeline includes a back-testing step for machine learning models, where a portion of historical data is held out to evaluate model performance before generating final forecasts. Performance metrics such as mean absolute error, root mean squared error, and MAPE are calculated during this stage. This intermediate validation step provides additional confidence in model performance prior to final evaluation.

Final model evaluation is conducted by comparing forecast outputs to actual sales from a later snapshot, ensuring that evaluation is based on realized demand rather than the same dataset used for training. In addition to MAPE, the methodology incorporates Kevian's internal forecast accuracy metric, defined as:

$$Kevian's\ Forecast\ Accuracy = max\left(0, 100 - 100 \times \frac{\sum|A_t - F_t|}{\sum A_t}\right)$$

This metric provides a volume-weighted measure of forecast error and aligns the evaluation framework with Kevian's internal performance standards. Model performance is assessed at the individual product–country level, and the best-performing model is selected for each combination rather than applying a single model globally.

Beyond model selection, the methodology also examines the relationship between baseline forecast accuracy and improvement potential. By comparing the best-performing model against Kevian's forecast, the analysis shows that products with lower baseline accuracy exhibit significantly higher improvement potential. This finding highlights the importance of targeted model selection and suggests that advanced forecasting methods deliver the greatest value when applied to products with higher demand uncertainty.

4. RESULTS AND DISCUSSION

The analysis highlights key findings from the forecasting framework and their implications for improving pharmaceutical demand forecasting at Kevian. Organized in line with the methodology, the results move from an assessment of current forecast performance to model benchmarking, expansion across countries, and an evaluation of business impact, providing a clear view of both current strengths and areas with the greatest potential for improvement.

Overall, Kevian's forecasting performance is perceived to be high at an aggregate level but exhibits significant variability at the product level. This creates substantial opportunities for improvement, particularly among low-performing SKUs. The results also confirm that no single forecasting model is universally optimal. Instead, model performance depends on product characteristics, data availability, and market dynamics. By systematically benchmarking with multiple models, the framework enables a scalable and data-driven approach to identify the best performing forecasting method for each context.

Importantly, improvements in forecast accuracy can translate into meaningful operational impact, particularly in markets where baseline forecast accuracy is lower and product volumes are substantial. In these cases, even moderate improvements in forecasting performance can lead to significant reductions in excess inventory or missed demand. At the same time, the findings highlight the importance of combining quantitative model outputs with domain expertise to ensure forecasts remain aligned with real-world market conditions. Together, these insights establish a clear foundation for the recommendations that follow, focusing on how this approach can be operationalized to improve forecasting performance across the global supply chain.

4.1 Baseline Forecast Performance and Problem Diagnosis

The initial analysis evaluates Kevian's existing forecast performance using dashboard data across product families and countries. At an aggregate level, forecast accuracy often exceeds 70%. However, a more granular analysis at the product-family level reveals a different reality (see Table 1). More than half of product families fall below this threshold, indicating substantial variability that is masked when performance is averaged across portfolios.

Product Families	Average of Forecast Accuracy
A	84.0%
B	76.7%
C	76.1%
D	74.2%
E	68.6%
F	66.0%
G	59.8%

Table 1. Forecast Accuracy Distribution Across Product Families (Across 28 Countries in Europe)

To ensure a more accurate assessment, product–country combinations with reported forecast accuracy of 0% were excluded from the analysis, as these cases typically reflect missing sales data or the absence of a forecast rather than true performance. Including these observations would systematically lower forecast accuracy at the product-family level by approximately 2–20%, potentially distorting the interpretation of results.

This discrepancy highlights a structural issue: aggregated metrics obscure meaningful variation at the SKU level, which is where forecasting decisions are ultimately made. In practice, this means that while overall performance appears acceptable, a subset of product–country combinations operate with forecast quality that falls below Kevian's expectations. Cases where this impacts product-country combinations with high-volume are important, as inaccuracies will drive the majority of inventory imbalance and service-level risk, whereas low-accuracy forecasts in low-volume segments have a comparatively limited operational impact. This finding reinforces the central hypothesis of the project, namely, that demand signals derived from affiliate batch ordering distort true market demand and limit forecasting effectiveness.

From a strategic perspective, this result is important for executive decision making. It suggests that improving forecast accuracy does not require a complete system overhaul, but rather a targeted approach focused on underperforming segments. It also underscores the importance of shifting from aggregate key performance indicators (KPIs) to more granular diagnostics when evaluating forecasting performance.

4.2 Model Benchmarking Results (Spain and Portugal)

Building on the baseline analysis, the second stage evaluates the performance of alternative forecasting models developed within the data pipeline. These include statistical approaches, machine learning models, and decomposition-enhanced methods. The objective is not to identify a single superior model, but to understand how model performance varies across different demand profiles.

A consistent pattern emerges when comparing model performance against Kevian's forecasts. The magnitude of improvement is strongly linked to baseline forecast accuracy (See Figure 1). The y-axis represents the relative improvement in forecast accuracy. This is calculated as the difference between the best-performing model and Kevian's baseline, which is divided by Kevian's forecast accuracy. A value of 2 indicates that the alternative model achieves 200% higher accuracy, while values near zero indicate little

improvement. Negative values reflect cases where alternative models underperform the baseline, which is typically observed in already high-accuracy SKUs.

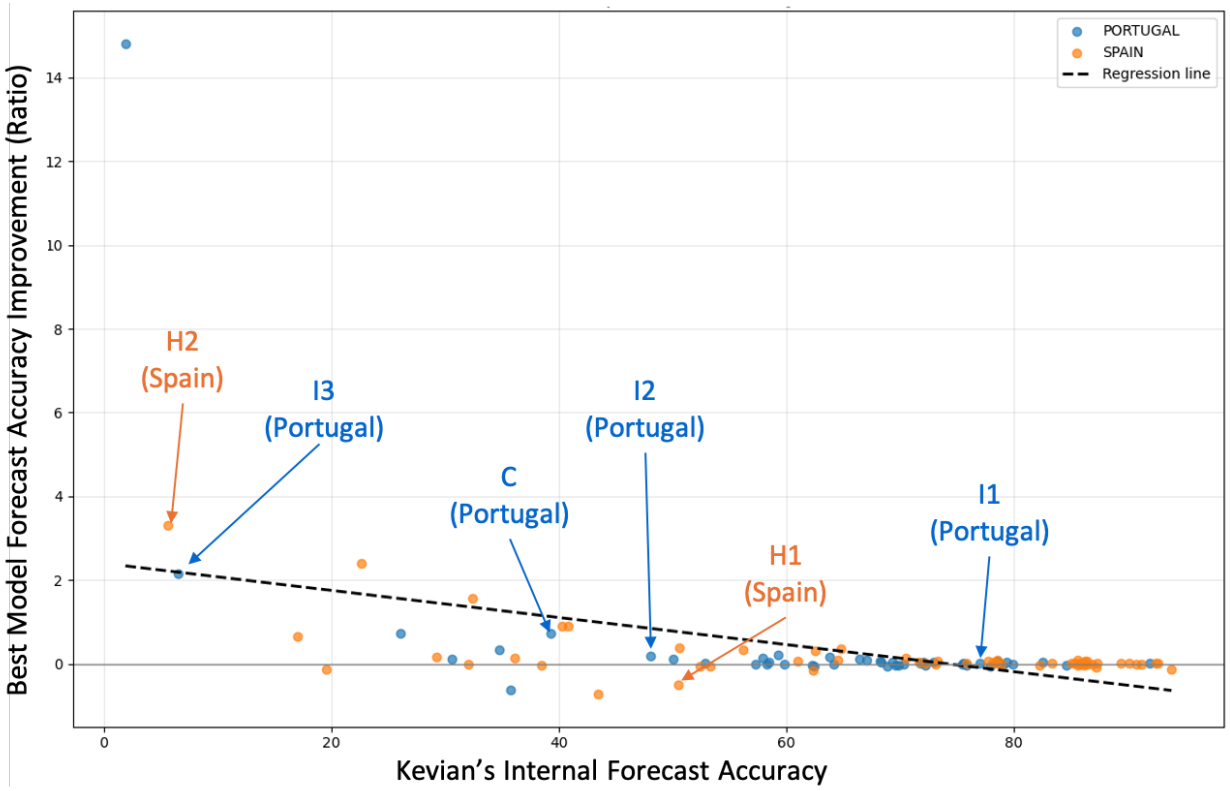


Figure 1. Best Model Improvement vs. Kevian Forecast Accuracy

For products with low forecast accuracy (typically below 50–60%) alternative models deliver substantial improvements. In contrast, for products already performing well, incremental gains are limited. This relationship is both intuitive and strategically significant. It indicates that forecasting improvements are not evenly distributed across the portfolio; rather, they are concentrated in specific segments where demand signals are more difficult to interpret. The data indicate that, even for low-accuracy SKUs, there is valuable demand signal within the historical data that can be captured and utilized to inform more accurate predictions of future sales.

This finding has direct implications for resource allocation. Instead of attempting to optimize forecasting performance uniformly across all products, organizations can achieve greater impact by focusing on low-performing SKUs, where the return on improvement efforts is highest.

4.2.1 Product-Level Insights

To better understand the drivers of model performance, several representative products were analyzed. For product C1 in Portugal as shown in Figure 2, a non-seasonal product that was identified as challenging to forecast, the use of alternative models significantly improved accuracy. This demonstrates the data pipeline's ability to extract meaningful signals even when demand patterns are relatively stable but obscured by noise. The bar chart on the left indicates that nearly all alternative forecasting models (grey) outperformed Kevian's own forecast (navy blue), with Exponential Smoothing on data decomposed with the Kalman filter (ETS kalman, in red) being the highest performing. The line chart at right shows that Kevian's own forecast (navy blue) underestimated future sales by a far greater degree than ETS kalman.

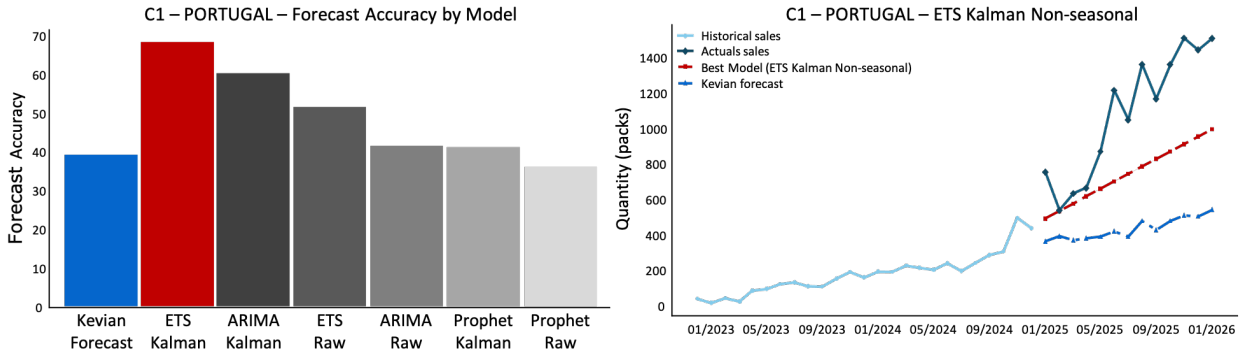


Figure 2. Forecast Comparison for C1 (Portugal)

In contrast, product family H in Spain, as shown in Figure 3, illustrates that some SKUs exhibit moderate forecast accuracy under existing methods, others (such as product H2) show extremely low accuracy, with all alternative models outperforming the baseline.

A closer comparison across SKUs further highlights this variation. For product H1, which has moderately high forecast accuracy (49%), none of the alternative models (grey and red bars) outperform Kevian's own forecast (blue bar). In this case, the existing forecasting approach already captures the underlying demand pattern effectively, leaving limited room for improvement. However, the situation differs significantly for the related product H2. Here, Kevian's forecast performs poorly, and all alternative models (grey bars) deliver superior accuracy. Among them, the Random Forest model applied to Kalman-filter–decomposed data (red bar) achieves the best overall performance.

This contrast within the same product family highlights the importance of differentiating not only across products but also within product families, as forecasting performance (and the value of alternative models) can vary significantly even among closely related SKUs.

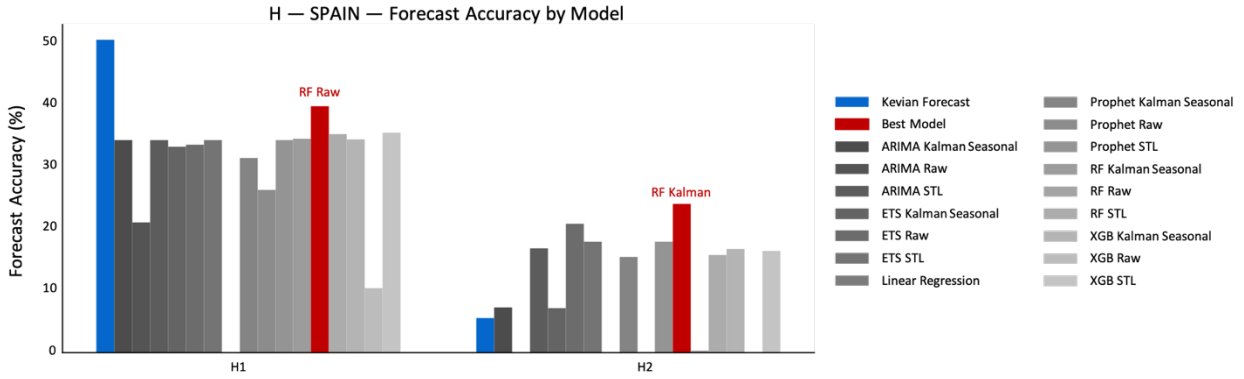


Figure 3. Forecast Comparison for H (Spain)

Product family I in Portugal, as shown in Figure 4, provides an additional perspective by illustrating three distinct performance scenarios within a single product family: high accuracy with limited improvement potential, moderate accuracy with incremental gains, and low accuracy where alternative models consistently outperform.

The top chart compares Kevian's forecast with alternative benchmarking models across SKUs within this product family. Consistent with earlier findings, the value of alternative models is inversely related to baseline performance. When Kevian's forecast accuracy is high, alternative models provide limited incremental benefit, whereas lower baseline accuracy creates greater opportunities for improvement.

This pattern is evident for product I3. In this case, supply constraints drive sharp fluctuations in observed sales; demand drops when the product is unavailable but rebounds quickly once supply resumes. While alternative model does not perfectly align with every peak and valley, it captures the overall directional movement and variability of demand more effectively than Kevian's relatively flat forecast. Among the models evaluated, ARIMA's autoregressive structure is best able to reflect this dynamic behavior, resulting in a forecast accuracy of approximately 20%, a meaningful improvement over Kevian's baseline.

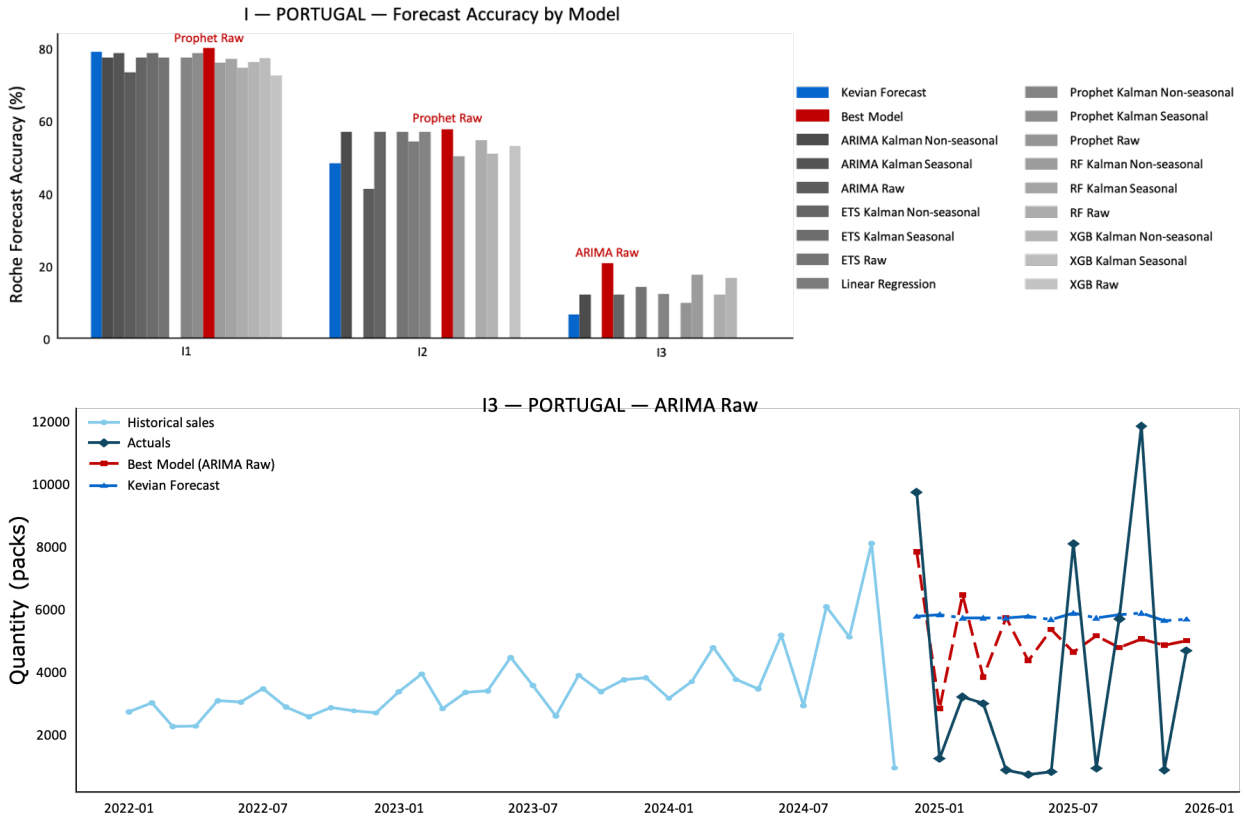


Figure 4. Forecast Comparison for I (Portugal)

Overall, these results highlight the importance of aligning model selection with underlying demand characteristics. In scenarios driven by supply constraints and abrupt demand shifts, classical time-series approaches such as ARIMA can outperform more complex models by effectively capturing short-term dependencies that are not reflected in baseline forecasts.

These examples reinforce a key insight: model effectiveness depends heavily on the underlying demand structure. Machine learning models tend to perform well in complex, data-rich environments, while classical statistical models remain highly effective in more constrained scenarios with less historical data.

4.3 Scaling Results Across 28 Countries in Europe

Following validation in Spain and Portugal, the forecasting framework was applied across 28 European countries to assess its scalability and generalizability. Extending the analysis across a broader set of markets allows for a more comprehensive evaluation of forecasting performance under varying demand patterns, data availability, and operational contexts.

Average forecast accuracy varies significantly across countries, ranging from approximately 61% to 89%. This variation reflects differences in market dynamics, product mix, and data completeness. Importantly,

it also highlights where improvement efforts can be most effectively targeted. In particular, large markets with relatively low forecast accuracy represent high-impact opportunities, as even modest improvements in these regions can translate into meaningful operational benefits.

While model selection must ultimately be tailored to each product-country combination, additional analysis was conducted to identify broader performance patterns across different forecast accuracy tiers. Table 2 does not report country-level averages; rather, it summarizes results at the product-country level using a "win rate" metric, which is defined as the percentage of cases in which an alternative model outperformed Kevian's baseline forecast in direct comparison. Performance is evaluated across low-, mid-, high-accuracy tiers based on Kevian's baseline forecast accuracy.

As shown in Table 2, the results are consistent with earlier findings from Spain and Portugal. Alternative models deliver the greatest value in low-accuracy scenarios, where win rates are highest, indicating frequent outperformance of Kevian's forecasts. In mid-accuracy cases, improvements are more moderate, while in high-accuracy scenarios, alternative models rarely outperform the baseline. This reinforces that the benefit of model benchmarking is concentrated in lower-performing product-country combinations rather than uniformly across all markets.

Low (<50 FCA)		
RF	+13.5 pp	80% win
XGB	+11.2 pp	77% win
ARIMA	+11.4 pp	72% win
ETS	+4.2 pp	56% win
Lin.Reg.	+3.8 pp	54% win
Prophet	+3.3 pp	60% win

Mid (50–75 FCA)		
XGB	+3.1 pp	67% win
RF	+2.2 pp	62% win
ARIMA	+1.8 pp	64% win
ETS	-3.9 pp	56% win
Prophet	-2.8 pp	55% win
Lin.Reg.	-13.9 pp	37% win

High (>75 FCA)		
XGB	-0.4 pp	54% win
RF	-0.7 pp	49% win
ARIMA	-1.9 pp	49% win
ETS	-3.1 pp	44% win
Prophet	-3.8 pp	48% win
Lin.Reg.	-18.0 pp	31% win

Table 2. Average Product Forecast Accuracy Across 28 Countries in Europe

Across the full dataset, machine learning models outperform Kevian's forecasts in approximately 60% of cases and are particularly effective for mature products with sufficient historical data. Among these, XGBoost consistently delivers the strongest performance, with Random Forest closely following across all accuracy tiers. In contrast, ARIMA remains the most effective classical forecasting approach, outperforming baseline forecasts in approximately 58% of cases. Its strength is most evident in early-stage or data-limited scenarios, where simpler autoregressive structures are better suited to capturing short-term demand dynamics.

A more detailed view of performance by accuracy tier further reinforces these patterns. For products in the low forecast accuracy tier, Random Forest improves forecast accuracy by approximately 13% relative to Kevian's baseline and achieves a win rate of around 80%. In the mid-accuracy tier, XGBoost delivers more modest gains (approximately 3% improvement) with a win rate of 67%. For products already in the high-accuracy tier, no alternative model consistently outperforms Kevian's forecasts, indicating that existing methods are already near optimal for these cases.

Taken together, these findings validate the design of the forecasting pipeline, which systematically evaluates multiple models rather than relying on a single methodology. While model selection should remain context-specific, the results suggest that machine learning approaches (particularly for mature products) offer significant untapped potential. At the same time, the continued effectiveness of classical models such as ARIMA underscores the importance of aligning model choice with product lifecycle and data availability. This combination of flexibility and structure provides a practical framework for improving forecasting performance across Kevian's diverse portfolio.

4.4 Quantifying Business Impact

While improvements in forecast accuracy are valuable, their ultimate significance lies in their operational impact. To assess this, the forecast performance from Kevian's baseline model was compared against the best-performing alternative model for each product-country combination over the 13-month test period. The resulting differences between forecasted and actual demand were aggregated to estimate the total reduction in forecast error, expressed in terms of corrected product volumes (packs).

The results in Figure 5 show that improvements are not evenly distributed. Across the top five markets in Europe (Group 1), the use of alternative models reduces the total forecast error by approximately 23,641 packs, meaning that this volume of over- or under-forecasted demand is corrected relative to Kevian's baseline. In contrast, the second tier of markets (Group 2) achieves a substantially larger correction of approximately 84,616 packs. This difference is driven by lower baseline forecast accuracy in many Group 2 countries, which creates greater chances for improvement. As a result, although these markets are smaller individually, they offer a high potential for operational gains when forecast accuracy is improved.

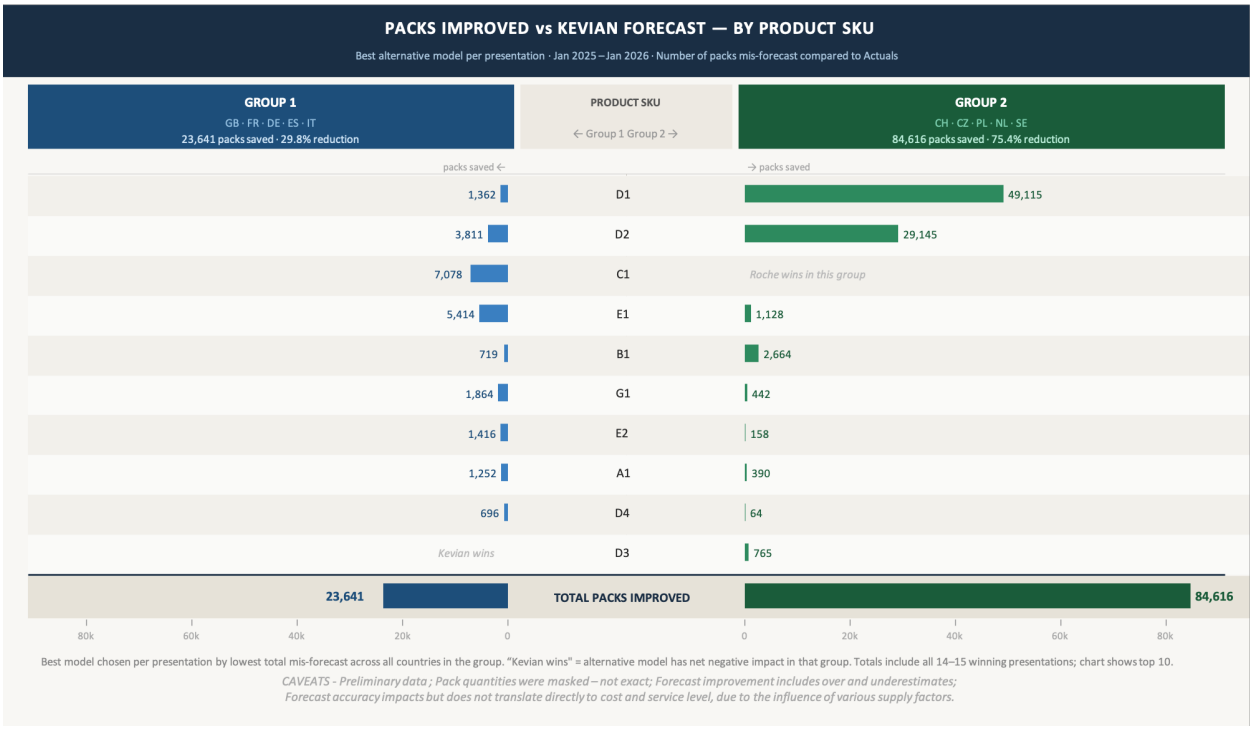


Figure 5. Units Improved vs. Kevian Forecast by Country Group

A closer examination of specific country-product combinations within Group 1 markets reveals targeted opportunities for improvement with meaningful volume impact. In particular, consistent patterns of over-forecasting are observed for Product E1 in Italy and Product D2 in Germany (see Figure 6), each resulting in excess volumes of more than 4,000 packs over the 13-month test period. Identifying such systematic bias is operationally significant: persistent over-forecasting drives excess inventory and associated holding costs, while under-forecasting risks service-level degradation and lost revenue.

Figure 6 illustrates these dynamics in detail, with the left panel showing Product E1 in Italy and the right panel showing Product D2 in Germany. In both cases, Kevian's forecasts consistently overestimate future demand relative to actual sales, whereas alternative models more closely track observed patterns. This gap suggests that the baseline forecasting approach does not fully capture underlying demand behavior in these scenarios. As a result, the observed over-optimism likely translates into unnecessary inventory

buildup at the country level. More broadly, these examples demonstrate how the data pipeline can be used to systematically identify and correct directional forecast bias, enabling more balanced inventory management and improved alignment between supply and true market demand.

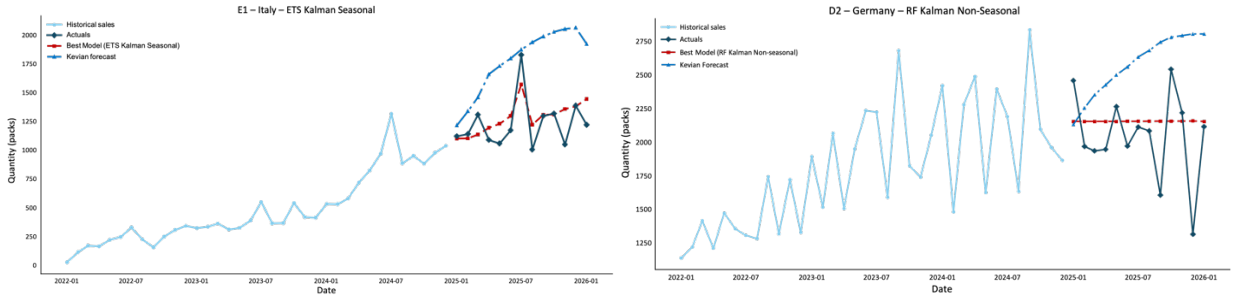


Figure 6. Forecast Comparison for E1 (Italy) and D2 (Germany)

From a business perspective, these findings indicate that improved forecasting can lead to meaningful corrections in both over- and under-forecasting. Across the dataset, however, over-forecasting is more prevalent, particularly in higher-volume markets where baseline forecasts tend to be systematically optimistic. As a result, many of the observed improvements are driven by reducing excess inventory rather than addressing stockouts. This distinction is important, because the operational benefits are more closely tied to lowering inventory holding costs and improving working capital efficiency. However, it is important to recognize that forecast accuracy improvements do not directly translate into cost savings or service-level improvements. The relationship between forecasting and operational outcomes is mediated by many factors such as inventory policies, production constraints, and supply chain responsiveness.

4.5 Limitations

Several limitations should be considered when interpreting these results.

First, data availability remains a constraint. In some cases, only 24 months of historical data were available, limiting the ability to fully capture seasonal patterns. While the availability of 36 months of data improves model performance, this is not consistent across all products and countries.

Second, the treatment of returns introduces noise into the dataset. Returns are recorded as negative sales rather than as a separate variable, which complicates the interpretation of demand and prevents the explicit modeling of return behavior.

Third, the model selection process relies on retrospective evaluation using known future data. While this approach is appropriate for benchmarking, it represents a best-case scenario and may overstate achievable performance in real-world forecasting contexts.

Finally, improvements are measured in terms of forecast accuracy and volume correction, rather than financial or service-level outcomes. As a result, the direct business impact must be interpreted with caution, as it depends on broader supply chain dynamics that are beyond the scope of this analysis.

5. RECOMMENDATIONS

This project developed and implemented a scalable forecasting framework that integrates signal decomposition, statistical models, and machine learning to improve pharmaceutical demand forecasting at Kevian. By benchmarking alternative models against existing forecasts across multiple products and countries, the analysis confirms that Kevian's overall forecast accuracy is generally strong. However, the

results also reveal meaningful opportunities for targeted improvement, particularly for products with lower baseline accuracy. The findings demonstrate that forecast performance is highly context-dependent, and that applying the right model to the right product and market conditions can significantly improve demand visibility and support more effective supply chain decision-making.

5.1 Management Recommendations

A key insight from this study is that forecasting improvement should not be approached uniformly across the portfolio. The results show that for products with already high forecast accuracy, alternative models provide only marginal incremental value, if any. In contrast, for products with lower forecast accuracy, alternative models can deliver substantial gains, with improvements ranging from 10% to over 200% relative to existing forecasts. This indicates that Kevian should adopt a targeted deployment strategy, focusing its forecasting enhancement efforts on products and markets where baseline performance is weak, and particularly for high-volume, high-impact products and markets. By prioritizing low-accuracy and high-impact products, Kevian can maximize the return on analytical investment while avoiding unnecessary complexity in well-performing segments.

To operationalize this approach, Kevian should scale the forecasting framework developed in this project across its country affiliates. The framework enables systematic training, evaluation, and benchmarking of multiple forecasting models across product-country combinations, allowing teams to identify the best-performing model based on historical data. Embedding this capability into routine planning processes would significantly reduce manual effort in model selection, accelerate the identification of underperforming forecasts, and create a standardized framework for continuous improvement. From a strategic perspective, this supports Kevian's broader digital transformation objectives by institutionalizing advanced analytics within its global supply chain operations.

At the same time, it is critical to recognize that model outputs alone are not sufficient for decision making. The benchmarking process used in this study benefits from hindsight, because models are evaluated against known outcomes. In real-world forecasting, uncertainty remains inherent, and local market dynamics play a significant role. Therefore, Kevian should adopt a hybrid forecasting approach that combines quantitative model outputs with qualitative inputs from country affiliates and forecasting experts. Specifically, this could include implementing a formal forecast review layer in which model-generated forecasts are first produced automatically and then reviewed through standardized workflows, such as monthly demand consensus meetings or exception-based adjustments where forecasts deviate beyond defined thresholds.

In addition, Kevian could introduce guided adjustment frameworks, such as predefined adjustment categories (e.g., epidemiological shifts, competitor launches, and policy changes) to ensure that qualitative inputs are consistent, documented, and traceable over time. Local expertise related to epidemiology, market access, and competitive dynamics can complement statistical forecasts and improve their practical relevance. This integration is essential to ensure both accuracy and organizational adoption, as it aligns data-driven insights with the experience and judgment of local stakeholders.

The results also highlight the importance of prioritizing high-impact markets and products when deploying improved forecasting approaches. The analysis shows that forecast improvements translate into meaningful volume reductions in mis-forecasted demand, including approximately 23,600 packs in the top five European markets and over 84,000 packs in second-five largest European markets. These gains are not evenly distributed but are concentrated in specific products and geographies, such as product families D and E in selected countries. By focusing deployment on these high-impact areas, Kevian can achieve significant operational benefits, including reduced excess inventory, improved product availability, and more efficient allocation of supply. While pack-level forecast improvements do not directly translate

into improved financial outcomes and product availability, they strongly indicate the potential for meaningful cost savings and service level improvements.

Underlying these recommendations is a broader structural insight: forecast accuracy is fundamentally constrained by the quality of the underlying demand signal. The study confirms that there are significant opportunities to increase the capture of accurate demand signal from historical sales data. As a result, improving forecasting performance requires not only better models but also better signal processing. Kevian should therefore strengthen its capabilities in demand signal decomposition, such as using techniques like STL and Kalman filtering to extract underlying demand patterns from noisy data. In addition, it should expand implementation of modern forecasting models, such as machine learning. By improving the quality of input data, data processing, and the models applied, Kevian can enhance the performance of all downstream forecasting models and enable more reliable upstream planning decisions.

5.2 Future Work

While this project demonstrates clear improvements in forecast accuracy and demand signal interpretation, further work is required to fully quantify and operationalize the value of these findings. A critical next step is to translate forecast accuracy improvements into financial and operational metrics, such as revenue impact, inventory cost reduction, and service level improvements. This will require deeper integration with financial data and a clearer understanding of how forecast errors propagate through the supply chain. Quantifying this value is essential for prioritizing investments and securing executive support for broader implementation.

Another important area for future development is the integration of sell-in data to enhance demand signal quality. The current analysis relies primarily on affiliate sales data, which reflects realized demand but does not fully capture batched ordering behavior from downstream customers. Incorporating sell-in data, including order timing and quantities, would provide a more reliable demand signal and earlier visibility into demand changes while improving short-term forecasting accuracy and responsiveness. This is relevant for managing replenishment cycles and reducing the impact of batch ordering distortions.

In addition, future work could focus on improving the alignment between forecasting outputs and upstream production planning. The study highlights that even with improved forecasts, there remains a gap in how affiliate demand is translated into manufacturing requirements. Addressing this gap will require a deeper review of planning processes and decision rules used to convert forecasts into production plans. Strengthening this linkage is critical to ensuring that improvements in forecast accuracy translate into tangible operational benefits.

Finally, there is an opportunity to enhance product segmentation using data-driven approaches. Applying clustering techniques, such as K-means based on demand variability metrics, can provide a more systematic way to classify products and match them with appropriate forecasting models. This would improve scalability and consistency in model selection, enabling Kevian to move toward a more automated and adaptive forecasting framework.

In a broader context, this study demonstrates the value of combining signal decomposition, statistical modeling, and machine learning within a unified and scalable framework. While developed for Kevian, the approach is applicable to other pharmaceutical companies and industries facing similar challenges with noisy demand signals and decentralized planning structures. By improving how demand data is interpreted and used, this work contributes to more responsive, data-driven supply chains, ultimately supporting better operational performance and improved access to critical therapies.

REFERENCES

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., … Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting, 38*(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Bilal, A. I., Rostami-Tabar, R., Hewage, H. H., Bititci, U. S., & Fenta, T. G. (2025). The missing puzzle piece: Contextual insights for enhanced pharmaceutical supply chain forecasting. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2025.2546028

Sazvar, Z., Zokaee, M., Tavakkoli-Moghaddam, R., Salari, S. A., & Nayeri, S. (2022). Designing a sustainable closed-loop pharmaceutical supply chain in a competitive market considering demand uncertainty, manufacturer's brand and waste management. *Annals of Operations Research, 315*(2), 2057–2088. https://doi.org/10.1007/s10479-021-03961-0

Fourkiotis, K. P., & Tsadiras, A. (2024). Applying machine learning and statistical forecasting methods for enhancing pharmaceutical sales predictions. *Forecasting, 6*(1), 170–186. https://doi.org/10.3390/forecast6010010

Van Belle, J., Guns, T., & Verbeke, W. (2021). Using shared sell-through data to forecast wholesaler demand in multi-echelon supply chains. *European Journal of Operational Research, 288*(1), 466–479. https://doi.org/10.1016/j.ejor.2020.05.059

Zahra, I. A., & Putra, Y. H. (2019). Forecasting methods comparation based on seasonal patterns for predicting medicine needs with ARIMA method, single exponential smoothing. *Materials Science and Engineering, 662*(2), 022030. https://doi.org/10.1088/1757-899X/662/2/022030

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting, 18*(3), 439–454. https://doi.org/10.1016/S0169-2070(01)00110-8

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach*. Springer. https://doi.org/10.1007/978-3-540-71918-2

Abolghasemi, M., Gerlach, R., Tarr, G., & Beh, E. (2020). Demand forecasting in supply chains: The impact of demand volatility in the presence of promotion. *Computers & Industrial Engineering, 142*, 106380. https://doi.org/10.1016/j.cie.2020.106380

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*(1), 3–33. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-decomposition-procedure-based-on-loess.pdf

Hochenbaum, J., Vallis, O., & Kejariwal, A. (2017). Automatic anomaly detection in the cloud via statistical learning. *arXiv*. https://arxiv.org/abs/1704.07706

Lei, T., Gong, C., Chen, G., Ou, M., Yang, K., & Li, J. (2023). A novel unsupervised framework for time series data anomaly detection via spectrum decomposition. *Knowledge-Based Systems, 280*, 111002. https://doi.org/10.1016/j.knosys.2023.111002