
Binita K C and Rajib Dangol

# A Comparative Analysis of Statistical and Machine Learning Methods for Retail Demand Forecasting to Support Operational Planning

School of Technology and Innovations
Master's Thesis, Industrial Systems Analytics
Master of Science in Technology

Vaasa 2026

**UNIVERSITY OF VAASA**
**School of Technology and Innovations**
**Author:** Binita K C and Rajib Dangol
**Title of the thesis:** A Comparative Analysis of Statistical and Machine Learning Methods for Retail Demand Forecasting to Support Operational Planning
**Degree:** Master of Science in Technology
**Discipline:** Industrial Systems Analytics
**Supervisor:** Petri Helo
**Year:** 2026 **Pages:** 107

**ABSTRACT:**

Retail demand forecasting is essential for businesses in their operational planning and supply chain management. Since the rise in the level of competitiveness in the markets and the adoption of data-driven decisions is high, there is a growing need for forecasting in businesses. For retailers, forecasting is essential because it aids in planning the workforce, distribution, reorder levels, and inventories. In contrast to the conventional statistical approaches that have dominated forecasting, the machine learning algorithms have introduced alternative models for complex demands. There is, however, little certainty about the relative effectiveness and viability of these forecasting approaches in the retail environment.

This study aims to compare statistical and machine learning techniques to retail demand forecasting and the relevance of the approaches in retail operational planning. The research will compare the accuracy of the forecasted values, the behaviour of the models, and how helpful the approach is to the decision-making process in the retail sector.

This research work has been conducted based on quantitative and empirical techniques. The empirical analysis has been conducted based on publicly available sales transaction data obtained from the UCI Machine Learning Repository. The dataset has been pre-processed through data cleaning, data aggregation, outlier handling, and feature engineering processes. In addition to that, the fixed-origin-hold-out validation technique has been used for model comparison for the Naive, ETS, SARIMA, XGBoost, Random Forest and Neural Network models. Forecasting accuracy has been calculated using forecasting accuracy metrics MAE, RMSE, and WMAPE.

The findings from the analysis prove that machine learning approaches outperform conventional statistical methods in managing the data with high volatility in the retail industry. From the results obtained, Random Forest emerged to be the most reliable method in generating unbiased predictions as it performs well in balancing bias and variance. The study shows that precise forecast results could improve operational planning through reduced risks while decision-making on inventory management. The research is an addition to the existing forecasting and retail analytics body of knowledge through employing a holistic analysis approach using public retail datasets to benchmark various forecast methods. Future studies are recommended to incorporate external exogenous factors in the prediction model along with a hybrid architecture system.

**KEYWORDS:** Operational Planning, Time-Series Analysis, Decision Support, Forecast Metrics, Machine Learning.

## Contents

## Figures



## Tables

## Disclaimer on Use of Artificial Intelligence (AI) Tools in Thesis Preparation

Following the academic integrity principles of the University of Vaasa, the authors would like to state the deliberate and strategic utilization of Generative Artificial Intelligence (AI) technologies, namely ChatGPT developed by OpenAI and Gemini by Google, in the process of writing this thesis.

The above-mentioned AI tools were primarily applied in the research process by:

**Language Refinement:** These AI tools were used for improving the language used in the document and making sure that the language is grammatically precise, clear, and academic in nature.

**Structural Organization:** The tools offered suggestions for the logical order of sub-sections as well as the flow from theoretical and empirical findings.

**Code Refinement:** AI tools were employed to improve the efficiency of the forecasting models' code that were implemented using Python Programming Language.

The authors state that all the activities of primary research goals, methods, specific configuration of the forecasting models and final interpretation of the results are still the original work of the authors. The authors compiled all the data and final manuscript was critically reviewed and verified for accuracy. The authors accept full responsibility for the content and conclusions in this thesis.

# 1 Introduction

## 1.1 Background

In modern highly competitive and data-driven retail business environment, accurate demand forecasting has emerged as an important element of operational planning and supply chain management. Retailers should keep on expecting customer demand to make sound decisions regarding inventory management, production planning, workforce allocation, distribution, and pricing strategies. Failure to produce accurate forecasts might result in a number of operational inefficiencies, including stockouts, excess inventory, lost sales, increased holding costs, and lower customer satisfaction (Chopra & Meindl, 2013). However, with the rapid growth of omnichannel retailing, shorter product life cycles, and unpredictable customer behaviour, traditional forecasting approaches are being challenged to deliver reliable predictions under complex conditions (Makridakis et al., 2018).

Traditionally, classical statistical tools like time-series models and regression techniques have been very important in retail demand forecasting. The techniques such as exponential smoothing and the autoregressive integrated moving average (ARIMA) models have been broadly applied because of their readability, relatively small data demands and solid theoretical bases (Box et al., 2015). These techniques work well when demand trends are only stable and linear in nature, and when the historical data has some apparent trends or seasonality (Gardner, 2006). Nonlinear patterns in retail demand are however likely to include promotions, holidays, changed prices, weather conditions, and outside market forces which the usual statistical methods are unlikely to be effective in capturing.

The recent years have seen the advent of advanced computational capabilities and the large-scale availability of data that have made it easy to adopt the machine learning techniques in demand forecasting. Decision-trees, random forests, gradient-boosting,

support-vector-machines, and neural networks can be used to model thought complex nonlinear relationships and interactions with multiple variables (Breiman, 2001). Such models are able to include various data points, such as transactional data, customer behaviour, promotional activities and external indicators which may enhance the accuracy of a forecast within a dynamic retail environment. However, machine learning models can be demanding in terms of data, need tuning, and not as transparent as the traditional statistical methods, which is a concern with the interpretability and usability in operational settings (Hyndman & Koehler, 2006).

Although machine learning application has gained considerable attention, the issue of whether these advanced approaches can always perform better than the classical statistical procedures in various retail situations is still a subject of debate. There are studies indicating that the traditional models are still competitive, especially in the short-term forecasting or constant demand of a product (Makridakis et al., 2018), whereas there are studies indicating the effectiveness of machine learning models in complex and highly volatile environments (Hyndman & Athanasopoulos, 2021). Besides, the trade-off that exists between the accuracy of the forecast, computational cost, interpretability, and ease of deployment presents a major issue to the practitioners seeking the best forecasting method suitable in operational planning (Fildes et al., 2022).

Therefore, it is necessary to determine the strengths and weaknesses of both statistical and machine learning methods to facilitate the successful decision-making in the retail practice. Retail companies need forecasting tools that do not only generate valid predictions, but also fit well into the planning systems, scale well into thousands of products, and offer managers to act upon them (Chopra & Meindl, 2013). Nevertheless, detailed comparative studies, which assess these approaches under realistic retail scenarios are still scarce, especially as far as their practical implication in operational planning system including inventory choice, replenishment timing, and resource dispensation are concerned (Makridakis et al., 2018).

This study aims at carrying out the comparative analysis of the statistical and machine learning techniques in the retail demand forecasting and investigating how well they perform in assisting operational planning. This study will attempt to find out which methods of forecasting are the most accurate and useful based on varying demand trends and planning horizons through analysis of the various forecasting methods based on the available retail data. This research will focus on the relative accuracy, strength as well as usability of these methods in the retail setting in the real world.

## 1.2 Research Gap

Despite the extensive research on demand forecasting, some key gaps still remain in the context of retail demand forecasting and operational planning. Majority of the research is conducted on either statistical models or machine learning techniques separately, rather than directly comparing the two under same conditions. This leaves practitioners without clear evidence on which approach performs better in practical retail settings. Although competitions such as the M4 study have compared statistical and machine learning methods, results vary across datasets and domain (Makridakis et al., 2018).

Furthermore, comparative studies do not focus much on the practical implications of the performance of the forecast but only concentrates on accuracy. In the operations of retail businesses, the process of making forecasts is very important, especially in decision-making for activities like inventory management, replenishment, staffing, and logistics. However, the link between forecast accuracy and operational decision quality remains underexplored in retail settings (Fildes et al., 2022).

The other problem is the use of exclusive proprietary datasets, which limits the transparency of the process. Differences in data availability, preprocessing methods, and evaluation criteria make it difficult to generalize results across studies. As a result, there

is an important need for studies that use different approaches for making forecasts under an analytical framework that uses public datasets.

Finally, the previous studies often fail to consider the trade-offs between model performance, interpretability, computational cost, and ease of deployment. For example, although machine learning techniques may provide higher accuracy, they may be difficult to interpret, making them less useful for many retail organizations.

This research aims to fill these gaps by providing a systematic comparison of traditional statistical models and modern machine learning techniques using the same dataset, metrics, and decision support perspective. This study hopes to provide academically robust yet practically useful insights by connecting forecasting performance to its operational relevance.

## 1.3 Research Question, Aims and Objectives

**Research Question**

How do statistical and machine learning forecasting methods compare in terms of predictive accuracy and operational planning relevance in retail demand forecasting?

**Research Aims and Objectives**

The general aim of this study is to conduct a comprehensive comparative analysis of different statistical and machine learning methods for retail demand forecasting and to assess their effectiveness in supporting operational planning decisions.

To achieve this aim, the study pursues the following specific objectives:

- To identify commonly used statistical and machine learning methods for retail demand forecasting.
- To develop statistical and machine learning forecasting models using the same

dataset.

- To evaluate and compare the performance and accuracy of the models.
- To evaluate the usefulness of forecasting results for operational planning decisions.

## 1.4 Definitions and scope of the study

Demand forecasting is a phenomenon that involves the prediction of customer demand in future through historical data and analytical models. In the retail industry, the demand forecasting process is important in facilitating the operational planning tasks like inventory, replenishment, staffing as well as distribution scheduling. Proper predictions help the retailers to match supply with the demand of consumers hence minimizing the cost of operation and enhancing the service standards.

This paper is concerned with the comparative research on statistical and machine learning when it comes to retail demand forecasting. The use of statistical forecasting techniques has been commonplace because of their interpretability, and capability to formulate trend and seasonal tendencies of time-series data. Conversely, machine learning methods have also become increasingly popular over the past few years due to their capability to represent complicated nonlinear associations in the big data and feature numerous driving forces.

The study also has a narrow scope of testing the chosen forecasting models based on historical retail transaction data. The research seeks to examine the predictive capability of the two methods of analysis, statistical and machine learning, in the same frame of analysis. Using various forecasting methods on the same data, the study will be looking at the variation in the performance of the model, their strength and usefulness in application in operational planning.

Besides, the paper interprets the applicability of forecasting outcomes in terms of an operational approach to planning. The retail organizations need forecasting tools that do not only demonstrate good predictive accuracy but also contribute to the process of making practical decisions. Consequently, the research results of this study are expected to present evidence on the appropriateness of the various forecasting techniques to be used in real-life retail setting.

The research however is limited in its scope to the demand forecasting using mainly historical sales data. Owing to the constraints in data, the external variables like promotion campaigns, weather conditions, or the macroeconomic indicators are not explicitly factored into the analysis. Consequently, the analysis of demand trends and predicting performance based on transactional data and temporal features acquired through transactional data can be considered the aim of the study.

On the whole, the proposed work will contribute to academic research as well as practical retail management, as it will offer a comparative analysis of the forecasting strategies and will examine how these strategies can be used to enhance the operational planning decisions.

## 1.5 Structure of the Thesis

This research contributes to the existing literature in several ways. Firstly, this study makes a systematic comparison between classical statistical models and modern machine learning methods using a single methodology. Secondly, this work not only evaluates predictive performance but also takes into consideration operational relevance, thus, emphasizing how predictive accuracy can contribute to better decision-making. Finally, the study offers practical insights for retail managers and supply chain professionals regarding the proper selection and implementation of forecasting techniques as per their operational needs.

The structure of this thesis consists of five chapters. The first chapter starts with introducing the research problem, aims and objectives, definition, scope and significance of the research. In the second chapter, a thorough review of the literature is conducted on the topic of forecasting methods used by retailers for operational planning. The third chapter describes the research methodology, data sources and analysis approaches, as well as model development and evaluation criteria. Empirical findings are provided in the fourth chapter, where results are discussed in relation to the existing literature, and recommendations are formulated. Finally, the fifth chapter concludes the thesis by summarizing key contributions, outlining limitations, and suggesting directions for future research.

## 1.6 Division of Work Between Authors

The thesis has been developed as a joint research effort of Binita K C and Rajib Dangol. Both authors contributed to all aspects of the research process from information retrieval, literature review, methodology, data collection, data analysis, interpretation of results, and thesis writing, according to the requirement of Master of Science in Industrial Systems Analytics degree program. The academic integrity and research process structure have been guaranteed through allocation of major responsibilities according to their respective expertise.

The theoretical framework and literature review (Chapters 1 and 2) were divided into two parts, where, Rajib Dangol studied the statistical approach to prediction and advancement in the theory of time series in retail industry, while Binita K C focused on development of architecture of machine learning in recent times and research gaps identified in existing literature. Both chapters were written in collaboration between the two authors for ensuring seamless transition from statistical theory to machine learning theory.

During the empirical stage (Chapters 3 and 4), there was a clear division of roles. Rajib Dangol was in charge of the data engineering aspect, which included the process of data cleaning, data aggregation, and feature engineering. Binita K C was involved in designing and implementing the basic statistical models (Naive, ETS, and SARIMA) and machine learning models (Random Forest, XGBoost, and Neural Network), along with optimizing the parameters. The assessment of results and validation of performance through the measures of WMAPE and RMSE, as well as the discussion of findings, was done collaboratively.

Regarding the discussion of operational aspects, including the determination of safety stocks and inventory management (Chapter 5), a collaborative approach was adopted, where mathematical results from both modeling strategies were integrated within a supply chain context. Lastly, the conclusion and recommendations for future work were collaboratively written by the two authors, who contributed equally in editing the final version of the thesis paper.

# 2 Literature Review

This chapter provides a thorough analysis of theory and related academic literature concerning retail demand forecasting. Initially, the chapter outlines the significance of predictive analytics in the contemporary environment of retail activities and logistics management. Then, a brief historical perspective is presented in terms of classic statistical models, including Naive, SARIMA and Exponential Smoothing, followed by an examination of modern ML algorithms, such as Random Forest, XGBoost, and Neural Network in time-series engineering. Thus, through reviewing the history of development and the current advances in data science methods, the chapter reveals the theoretical deficiencies which this comparative study will seek to address.

## 2.1 Demand Forecasting in Retail

Demand forecasting is one of the most important analytical processes in retail operations and supply chain management. The retail organizations place a lot of importance on accurate forecasts to make operational decisions associated with inventory management, planning of replenishment, organization of working schedule, and logistics. The accuracy of the forecast has a direct impact on the efficiency of operations and customer satisfaction since either an inventory shortage or holding costs of overstock can result due to errors in the forecasts (Charbonneau et al., 2008).

Retail demand forecasting is a process which tries to predict future demand of a product in relation to the historical sales patterns and explanatory variables like promotions, holidays and pricing plans. Retail demand data normally have complex properties such as trend, seasonality, volatility and irregular demand variations. Such characteristics complicate the forecasting process significantly and demand very advanced modeling techniques that would enable the capture of the temporal dependencies and nonlinear associations.

Traditionally, statistical time-series forecasting models have been frequently applied in the retail demand forecasting because of their high theoretical base and interpretability. The exponential smoothing and the autoregressive integrated moving average (ARIMA) models are classical methods in demand forecasting studies that have long been regarded as standard ones (Hyndman and Athanasopoulos, 2021). These models use past observations in a time series to find trends and seasonality which can be extrapolated to forecast the demand in the future, but modern retail demands models that incorporate additional factors such as pricing actions, seasonal events, promotions, and external economic indicators (Falatouri et al., 2022).

Moreover, the recent retail environments are more dynamic and evolving due to the shifting consumer preference, omnichannel shopping behaviours, promotion and short product lifecycle. These complications increase the difficulty of forecasting and its significance to the success of operations (Ganguly & Mukherjee, 2024). In addition to that, large retailers have to deal with thousands of SKUs (stock-keeping units) distributed in different locations each having different demand patterns. As indicated by Petropoulos et al. (2025), forecasting should be of product, store, and regional scale and responsive to local differences. The retail companies are currently acquiring huge volumes of transactional information that include information regarding the individual purchases, product classes, customer behaviours and promotions. These mass datasets have provided the possibility to use the complex data-driven forecasting methods.

The machine learning methods have thus been acquiring growing concern in demand forecasting studies. Machine learning models can directly learn complex nonlinear relationships based on data without making any strong assumptions on the underlying structure of the time series. Consequently, the models have demonstrated promising results in most retail forecasting tasks (Makridakis et al., 2018).

Empirical results have recently confirmed that machine learning algorithms can outperform traditional statistical models in some forecasting problems, especially when

one has to work with large datasets with complicated demand patterns. As an example, the decision-tree-based ensemble, e.g. Random Forest and Gradient Boosting, are broadly used in retail forecasting because they can include nonlinear interactions among variables (Nasseri et al., 2023).

However, there are several challenges that are also associated with the implementation of machine learning models in retail forecasting (Molnar, 2022). The models tend to involve a lot of data preprocessing, feature engineering, and hyperparameter tuning. Moreover, several machine learning models can be seen as “black box” models and therefore, they might lack interpretability as offered by conventional statistical models.

Systematic comparative studies, therefore, still need to be done in the evaluation of both machine learning and statistical methods of forecasting under consistent conditions.

## 2.2 Statistical Time-Series Forecasting Models

Statistical forecasting models have been applied in business forecasting over many years and they have been proven to be quite reliable in terms of facilitating decisions based on facts. The models are based on mathematical and statistical methods to process historical time-series data and provide predictions while taking into account the assumption that the data generating process does not change substantially over time (Box et al., 2015).

Although a large number of statistical models have been developed in academia in this regard, this research will look specifically at the three models that lie at the foundation: The Naive model, the Exponential Smoothing (ETS), and the SARIMA architecture. Through the comparison between these traditional statistical approaches and contemporary machine learning techniques, this study seeks to verify if higher complexity provides more accurate predictions (Hyndman & Athanasopoulos, 2021).

### 2.2.1 Naive Forecasting Models

Naive forecasting is one of the most basic forecasting methods. In this approach, the next period forecast is an equal to the latest observed value. Although simple, the naive model is often used as a benchmark model in forecasting studies (Hyndman & Athanasopoulos, 2021).

The Naive model is mathematically defined as:

$$\hat{y}_{T+h|T} = y_T, \tag{1}$$

where:

$\hat{y}_{T+h|T}$ is the forecast for *h* periods into the future given the history up to time *T*, and $y_T$ is the actual observed value at the most recent time point.

The role of benchmark models in forecasting research is significant as they can be used to evaluate more advanced forecasting models. When the complex forecasting model is not doing better than a simple naive benchmark then there may be no justification in the extra complexity of the model.

### 2.2.2 Exponential Smoothing

One of the most commonly used methods of forecasting in business and industry is exponential smoothing. The underlying concept of exponential smoothing is that past observations are given exponentially declining weights so that the more recent data point is the more powerful in terms of the influence it has on the forecasts (Hyndman & Athanasopoulos, 2021).

A number of exponential smoothing model variants have been built to be able to model different components of time-series. Simple exponential smoothing is suitable to use when the data has neither trend nor seasonality whereas the linear trend model developed by Holt introduces the trend components. The Holt-Winters approach also extends the model by incorporating seasonal factors (Gardner, 2006; Box et al., 2015).

The **Additive Holt-Winters Triple Exponential Smoothing** model is used in this study. It was selected because the seasonal variations in the retail dataset appeared relatively constant regardless of changes in the underlying demand level. It is defined by the forecast equation:

$$\widehat{y_{t+h|t}} = l_t + hb_t + s_{t+h-m(k+1)}, \quad (2)$$

where:

$\widehat{y_{t+h|t}}$**:** This is the Predicted Value for time *t+h*, given what we know about time *t*,

$l_t$ **:** The baseline average at the end of training period,

$hb_t$**:** The Trend $b_t$ times the number of weeks ahead *h*, giving us the sloped line of the forecast,

$s_{t+h-m(k+1)}$**:** This goes back to the training data to get the past high/low seasonality value, adding it to the trend line.

Exponential smoothing models have been widely used in retail forecasting since they are computationally efficient and comparatively simple to use. Moreover, these models can be interpreted well and give information on the underlying demand patterns.

### 2.2.3 ARIMA Models

Autoregressive Integrated Moving Average (ARIMA) models represent another major class of statistical forecasting techniques. ARIMA models capture relationships between

current observations and past values through autoregressive, differencing, and moving average components to capture autocorrelation and temporal structure.

ARIMA was created by Box and Jenkins and has evolved to be one of the most popular time-series modeling techniques (Box et al., 2015). ARIMA models can be used to model a vast number of time-series behaviour. Lindfors (2021) found that time-series approaches such as ARIMA and Holt-Winters often perform competitively in retail settings where demand patterns are stable and predictable.

The **seasonal ARIMA (SARIMA**) models are the extensions of ARIMA framework with seasonal elements which are used in this comparative analysis.

The SARIMA model has seven parameters:

$$(\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q})) * (\boldsymbol{P}, \boldsymbol{D}, \boldsymbol{Q})_{\boldsymbol{m}} ,$$

where:

***p, P:*** The orders of the autoregressive terms. These measure how an observation is related to its lagged (previous) observations.

***d, D:*** Orders of the integrated terms. They describe the number of times differencing is applied to make the data stationary.

***q, Q:*** Orders of the moving average terms. These measure the association between an observation and the error term of previous forecasts.

***m:*** The period of the seasonality. It is equal to 12 in this study.

SARIMA models are also very effective in cases where demand trends are regular in terms of seasonal variations, which is common in retail sales data (Box et al., 2015).

### 2.2.4 Strengths and Limitations of Statistical Forecasting in Retail Operations

The statistical forecasting methods also remain important for retail operations due to their transparency, their simplicity of use, and their relatively low requirements of computing. Many retail businesses are still using statistical forecasting methods for forecasting short-term sales, managing inventory and warehouse operations (Silver et al., 2016). In reality, the methods like moving averages, exponential smoothing, and ARIMA are still popular and continue to be used in ERP systems because they are easy to understand and perform well in structured demands.

The one of the biggest strengths of statistical forecasting models is that they deliver forecasting logic that is easy to understand to decision makers. Forecasting systems that are able to clearly communicate the reasons for the change in demand can be preferred by retail managers when forecasts are an integral part of the purchase and replenishment process. Unlike the complex machine learning methods, the use of statistical models will give a planner the ability to explicitly observe the trend structure, seasonality, and autocorrelation structures in the data that will be used for forecasting purposes (Hyndman & Athanasopoulos, 2021).

An additional advantage of statistical approaches is that they can be very efficient with small amounts of historical data (Makridakis et al., 2018). Some small and medium-sized businesses might not have the vast amounts of data needed for advanced machine learning approaches. In these cases, statistical forecasting models are still very useful as they can produce reasonably accurate predictions with limited data and less computing power (Hyndman & Athanasopoulos, 2021).

Retail demand conditions, however, have become far more volatile, as a result of promotion, the multi-channel buying habits of consumers, fast product rotation, and shifting customer tastes (Lindfors, 2021). The complexities make it difficult to use these types of purely statistical models, which are frequently based on assumptions of linear

and stable temporal relationships. The sudden demand changes that result from external events or the non-linear interdependency of factors may, however, be a challenge for statistical models (Makridakis et al., 2020).

These are only a few of the limitations of statistical forecasting techniques, but they are still important benchmark models in statistical forecasting research. They are simple and robust and can serve as benchmarks for assessing the performance gains of machine learning models. As a result, there are still a number of forecasting research studies that compare advanced machine learning methods with traditional statistical learning models, including ETS and SARIMA models.

## 2.3 Machine Learning Approaches to Retail Demand Forecasting

Machine learning (ML) methods have gained popularity due to the rapid growth of data availability and their ability to model nonlinear relationships and include a wide range of input features. The machine learning algorithms can handle large, complex datasets, and automatically extract patterns in data and generate predictive models without depending on explicit statistical assumptions (Makridakis et al., 2020). Traditional forecasting models are univariate in nature; however, the current state of retail forecasting necessitates cross-sectional modeling using multiple data sources, which is best handled by machine learning (Petropoulos et al., 2022).

The effectiveness of machine learning approaches in a retail setting was established through the M5 Forecasting Challenge, where winning models made use of ensemble gradient boosting to account for the intermittency and hierarchy of retail operations (Makridakis et al., 2022). Recent studies conducted by Suryawanshi et al. (2024) indicate that although conventional approaches such as ARIMA serve as a good starting point, ML models like XGBoost tend to perform better with big data in the retail industry, due to their ability to detect patterns that conventional models cannot identify. This study

shows that model accuracy heavily depends on the nature of data, thus calling for a comparative approach to methodology. This study takes into account three most common machine learning methods: Random Forest, XGBoost and Neural Network.

### 2.3.1 Random Forest

Random Forest is an ensemble technique of learning that builds several decision trees and integrates their forecasts to enhance the accuracy of forecasting. The algorithm brings in randomness in which subsets of features and training samples are randomly chosen in the construction of individual trees (Breiman, 2001).

Random Forest models are specifically beneficial in making predictions of tasks with nonlinear relationship and multidimensional data (Fawagreh et al., 2014). This approach is also quite resistant to overfitting and therefore this makes the approach applicable to most real-world applications.

Random Forest has been extensively used in retail demand forecasting because it can model complex interactions between the explanatory variables including price, promotions and seasonal factors. Research has indicated that the Random Forest models have the ability to perform better than the traditional regression models in predicting the pattern of retail sales (Breiman, 2001).

**Mathematical Representation**

Random Forest regression model operates by making use of B decision trees and summing their outputs. This output, which is the final value of $\hat{y}$ is defined as follows:

$$\widehat{\boldsymbol{y}} = \frac{\mathbf{1}}{\boldsymbol{B}} \sum_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{B}} \boldsymbol{T_b}(\boldsymbol{x}), \tag{3}$$

where:

$\widehat{\boldsymbol{y}}$**:** Final retail demand,

***B*****:** Total number of trees in the forest; for instance, there can be 100 or 500 trees,

$\boldsymbol{T_b(x)}$**:** The prediction generated by a single decision tree *b*, taking into consideration input *x*.

### 2.3.2 Gradient Boosting and XGBoost

Another ensemble learning algorithm is gradient boosting which constructs forecasting models one by one. The successive models are trained to correct the mistakes committed by the previous model, gradually improving predictive performance.

XGBoost is a highly optimized gradient boosting implementation, which has become very popular in machine learning competitions and real-world applications (Chen & Guestrin, 2016). XGBoost has regularization methods and parallel computing features which enable it to have high predictive performance and remain computationally efficient.

Gradient boosting models have proved to be very predictive in retail forecasting applications because they can predict complex nonlinear relationships between sales and influencing factors. In a retail setting, this has been proven to be accurate since tree-based models have a considerable impact in minimizing forecasting errors particularly during promotional periods or during the peak season (Niemelä, 2025).

**Objective Function of XGBoost Model**

The XGBoost model is defined by an objective function which incorporates prediction precision and model complexity (regularization). The prediction of the model at step *t* can be expressed as:

$$\hat{\boldsymbol{y}}^{\boldsymbol{t}} = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{t}} \boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{i}}) = \hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(\boldsymbol{t}-\boldsymbol{1})} + \boldsymbol{f}_{\boldsymbol{t}}(\boldsymbol{x}_{\boldsymbol{i}}), \quad (4)$$

The objective function minimized by the model in order to determine the best function is expressed as follows:

$$\{\boldsymbol{L}\}(\boldsymbol{\phi}) \;=\; \sum\_\{\boldsymbol{i}\}\, \boldsymbol{l}(\hat{\boldsymbol{y}}_{\boldsymbol{i}}, \boldsymbol{y}_{\boldsymbol{i}}) \;+\; \sum\_\{\boldsymbol{k}\}\, \boldsymbol{\Omega}(\boldsymbol{f}_{\boldsymbol{k}}), \quad (5)$$

where:

$\boldsymbol{l}(\hat{\boldsymbol{y}}_{\boldsymbol{i}}, \boldsymbol{y}_{\boldsymbol{i}})$: The loss function, which measures the difference between the predicted demand $\hat{y}_i$ and the actual retail sales $y_i$,

$\boldsymbol{\Omega}(\boldsymbol{f}_{\boldsymbol{k}})$**:** The regularization term, that restricts tree complexity and prevents overfitting.

### 2.3.3 Neural Networks and Deep Learning

The use of artificial neural network in solving demand forecasting problems has also been on the rise. A recurrent neural network (RNN) is considered effective in time-series prediction due to its ability to represent a sequential dependency in data, and The Long Short-Term Memory (LSTM) models have been very effective in the forecasting tasks that involve complex temporal dependencies. Classical RNN models have challenges in learning long-term dependencies because of vanishing gradients and so on. LSTM networks were developed to overcome these shortcomings, in that they added memory cells with the ability to remember information in a long time horizon (Hochreiter and Schmidhuber, 1997).

In the context of this research, the Multilayer Perceptron (MLP) model is considered the representative of deep learning methods. The MLP model is a type of feedforward artificial neural network that includes an input layer, hidden layers, and an output layer (Hornik et al., 1989).

While recurrent neural networks operate based on sequential processing, the MLP neural network operates on a feature vector from each observation independently. Thus, the MLP neural network is well suited for forecasting based on structured retail data because the temporal dependency is accounted for by the engineered features, such as the lagged values and seasonal indicators (Zhang et al., 1998). Each neuron in the hidden layer performs a weighted sum over the inputs, and the result is further passed through an activation function, which can be non-linear, for example, the rectified linear unit (ReLU) or sigmoid activation function (Goodfellow et al., 2016).

**MLP Mathematical Model**

The MLP produces the output from the input variables using the non-linear transformations of the weighted sums at the hidden layers. This can be formulated for an MLP with one hidden layer as:

$$\boldsymbol{y} = \boldsymbol{\phi}\left(\sum_{\boldsymbol{j=1}}^{\boldsymbol{q}} \boldsymbol{w}_{\boldsymbol{j}}^{(\boldsymbol{2})} \cdot \boldsymbol{\sigma}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{w}_{\boldsymbol{ji}}^{(\boldsymbol{1})} \boldsymbol{x}_{\boldsymbol{i}} + \boldsymbol{b}_{\boldsymbol{j}}^{(\boldsymbol{1})}\right) + \boldsymbol{b}^{(\boldsymbol{2})}\right), \quad (6)$$

where:

$\boldsymbol{x_i}$ **:** Input variables (e.g., lagged sales, seasonality factor, price),

$\boldsymbol{w_{ji}^{(1)}}$ **:** Weights in the first layer from variable *i* to hidden unit *j*,

$\boldsymbol{b_j^{(1)}}$**:** Bias for the hidden layer,

$\boldsymbol{\sigma}$**:** The non-linear activation function such as ReLU or Sigmoid that gives MLP ability to learn non-linear relationships,

$\boldsymbol{w_j^{(2)}}$ **and** $\boldsymbol{b^{(2)}}$**:** Output layer weights and bias respectively,

$\boldsymbol{\phi}$**:** Output layer activation function, usually a linear function.

While deep learning models require large-scale data sets, the MLP network offers the best balance for retail forecasting. In contrast to regression models, MLP networks offer greater flexibility in forecasting while requiring less computational power.

### 2.3.4 Feature engineering in Machine Learning Forecasting

The quality of machine learning models and the accuracy of the forecasting results rely on the quality and relevance of the variables used in the model. Feature engineering is one of the most important steps in machine learning-based forecasting. Machine learning algorithms can use a combination of explanatory variables, unlike classical statistical forecasting methods that focus on the historical time-series structure (Kuhn & Johnson, 2019).

Typical feature engineering tasks in retail demand forecasting include the calculation of lagged demand variables, rolling averages, moving standard deviations, seasonal indicators, calendar variables and promotional features (Kuhn & Johnson, 2019; Mustapha & Sithole, 2025). Lagged demand variables are used to allow models to learn the time dependency between previous sales observation and future demand values. Likewise, statistics features are rolled to capture short-term trends and volatility of demands.

Additionally, retail demand is often highly seasonal, which is why month, week number, day of the week and holiday indicators are commonly added to retail forecasting models (Hyndman & Athanasopoulos, 2021). For instance, there can be very large swings in sales on weekends compared to weekdays and periods of holidays can result in very high sales volumes.

It has been demonstrated that the predictive abilities of machine learning models like Random Forest and XGBoost can be greatly enhanced by well-designed temporal features (Niemelä, 2025). Tree-based ensemble methods are especially useful when the data is rich in features, due to the ability to capture nonlinear interactions between the demand variables and the seasons and between seasonal and lagged demand variables.

As noted by Zheng & Casari (2018), well-engineered temporal and categorical features can significantly improve the accuracy of machine learning models for prediction, even if relatively simple models are employed. Likewise, Makridakis et al. (2020) contended that modern forecasting models may succeed less due to their complexity but because of the combination of good feature engineering, data preprocessing, and appropriate validation methods.

But too much feature generation can also lead to overfitting, particularly when the data set is small. Hence, feature selection and validation get importance in machine learning forecasting pipelines as per Kuhn and Johnson (2019). When developing a forecasting system, researchers are frequently faced with the challenges of getting the system to be simple enough to be deployed in retail operations while maintaining predictive performance.

### 2.3.5 Strengths and Limitations of Machine Learning Models

ML models are effective in nonlinearities, promotion effects, and incorporating exogenous variables. Still, they may lack interpretability, which makes it more difficult to learn about the drivers of forecasts in the case of operational planners. They also require considerable data preparation and tuning (Makridakis et al., 2020).

In the retail sector, machine learning models have several key benefits. Machine learning algorithms are capable of modelling nonlinear relationships and interactions between many variables, without assuming strong properties about the distribution of the data in question. This feature is particularly well suited for today's retail stores that are subject to several dynamic factors affecting demand, including promotions, holidays, seasonal changes, pricing, and evolving consumer buying habits (Makridakis et al., 2020).

One of the key advantages of machine learning models is their ability to handle both structured and unstructured data. Retail companies have huge volumes of transactional and operational information coming in and out of their systems via registers, e-commerce websites, inventory, and customer interactions. These various data sources can be combined with forecasting models that help increase the accuracy of forecasts and provide more responsive operational planning systems (Chowdhury et al., 2025).

In the field of retail forecasting, ensemble learning methods like Random Forest and Gradient Boosting have shown exceptional success, especially in handling nonlinear demand dynamics and interactions among various features. Tree models also tend to be relatively less sensitive to noise and outliers, often seen in retail sales data. In turn, deep learning models like Long Short-Term Memory (LSTM) networks have proven effective for predicting tasks with extended temporal dependencies and temporal sequences of demand (Hochreiter & Schmidhuber, 1997).

One more benefit of the machine learning models is their adaptability. A lot of machine learning algorithms are capable of learning as new data comes in and adapting their forecasting patterns over time (Mitchell, 1997; Bull et al., 2024). This adaptability is particularly critical in the fast-paced retail landscape, where dynamics like customer preferences and buying habits are constantly evolving.

While there are many benefits of machine learning forecasting models, they have a number of drawbacks and practical issues as well. Limited interpretability of many machine learning algorithms is one of the big concerns. More complex forecasting models include neural networks and boosted ensembles that are generally referred to as “black box” models because they are not easy to explain (Kuhn & Johnson, 2019). Such lack of transparency may lead to reduced managerial confidence and restrict the implementation of forecasting in practical operational planning contexts where decisions makers need clear and easily comprehensible forecasting logic (Carbonneau et al., 2008).

High volume of high-quality historical data is also needed for effective training of machine learning models. Forecasts may be less accurate if the data used is incomplete or inconsistent, or if it includes inaccurate or redundant data. This makes data preprocessing and feature engineering crucial components of a machine learning forecasting process. Creating appropriate input variables, lag features, seasonal indicators, and rolling statistics requires significant technical expertise and computation (Kuhn & Johnson, 2019).

In addition, machine learning algorithms often are very complex to optimize and validate their model, which is usually necessary to obtain the best result. If the model is not tuned correctly, it could lead to overfitting, producing poor results on future forecasting periods. This challenge is one of the most peculiar ones in the domain of time-series forecasting, as the demand for retail products could vary over time depending on the external market conditions.

Another key constraint of advanced machine learning techniques is the computational cost. Zheng and Casari (2018) highlight that ensemble forecasting systems and deep learning models can be more expensive in terms of computing power and require longer to train, when compared to traditional statistical models like ETS or SARIMA. So, small retailers face challenges in implementing advanced forecasting systems into practice (Mustapha & Sithole, 2025).

Overall, the literature indicates that machine learning models have the potential to offer significant forecasting power in complex demand environments, especially when there are a large dataset and multiple explanatory variables. But there are several factors that need to be taken into account when choosing the right forecasting tools for operational retail planning, such as interpretability, complexity, computational needs and data quality.

## 2.4 Forecast Accuracy Evaluation

In forecasting research, evaluation of the accuracy of the forecast is a necessary step. Any forecasting technique requires objective assessment in terms of its effectiveness before it can be employed for practical purposes. Evaluation of forecasts allows comparing various forecasting techniques based on their performance and identifying the models that yield the best results in terms of forecast reliability in varying business settings (Hyndman & Athanasopoulos, 2021). The importance of accurate forecast evaluation becomes especially evident in demand forecasting in retail as forecasting inaccuracies have a direct impact on inventory planning, scheduling, staffing, and the efficiency of logistics operations.

Forecasting results are typically evaluated based on various quantitative measures of forecast error that enable comparison of the predicted value and observed demand. Since various evaluation metrics reveal different features of forecasting errors, multiple measures are usually adopted to assess the accuracy of forecasts (Hyndman & Koehler, 2006).

### 2.4.1 Mean Absolute Error (MAE)

MAE is one of the most commonly used measures that quantifies the average magnitude of the errors made in a series of predictions irrespective of their direction. It is computed as the average of the absolute deviations between the forecasted and observed values.

The MAE metric is mathematically calculated using the following formula:

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} |\boldsymbol{yi} - \widehat{\boldsymbol{yi}}|, \tag{7}$$

where $\boldsymbol{n}$ is the number of observations in the test set, $\boldsymbol{yi}$ is the actual value for period *i*, and $\widehat{\boldsymbol{yi}}$ is the forecasted value at period $\boldsymbol{i}$.

MAE gives an intuitive measure of forecast accuracy because it gives errors in the same units as the original data (Makridakis et al., 1998). Therefore, the MAE is an ideal measure for communicating error results because the measurement unit will be expressed in either sales or currency. Moreover, as opposed to the RMSE, the MAE avoids squaring the differences (Willmott & Matsuura, 2005).

### 2.4.2 Root Mean Squared Error (RMSE)

RMSE is used to measure the square root of the mean squared errors between the predicted and actual values. RMSE uses larger errors as a bigger penalty than MAE hence it is applicable when a large error in forecasting is very undesirable.

The mathematical formula for RMSE is as follows:

$$\boldsymbol{RMSE} = \sqrt{\boldsymbol{1/n}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \widehat{\boldsymbol{y_i}})^{\boldsymbol{2}}}, \quad (8)$$

In terms of forecasting within retail environments, RMSE is especially useful since it makes larger forecast errors much costlier to the business than multiple smaller ones, as they may cause either stockouts or overstocking (Hyndman & Athanasopoulos, 2021). It is also essential to mention that, due to RMSE being a squared measure, its values may be more susceptible to extreme values when compared to MAE (Willmott & Matsuura, 2005). So, employing both metrics would allow for an overall assessment where larger discrepancies between MAE and RMSE imply a model making a few extreme forecast errors (Chai & Draxler, 2014).

### 2.4.3 Weighted Mean Absolute Percentage Error (WMAPE)

Weighted Mean Absolute Percentage Error (WMAPE) is one of the metrics that is independent of scale and is used to assess the level of accuracy in the forecasts for datasets characterized by different levels of demand. While using the classic MAPE, which might be influenced by small volume, WMAPE calculates the errors taking into consideration the sum of actual demands, thereby improving the accuracy in evaluating retail portfolio forecasts (Hyndman & Athanasopoulos, 2021).

Mathematically, the formula for the WMAPE can be described as:

$$\boldsymbol{WMAPE} = \frac{\sum_{i=1}^{n} |yi - \widehat{yi}|}{\sum_{i=1}^{n} |yi|} * \mathbf{100}\%, \quad (9)$$

In retail forecasting, the use of WMAPE prevents the errors as "division by zero" or an excessively large percentage for a certain period with low demand. Given the fact that higher-value items (with a higher financial value or inventory risk) have a greater influence on the forecast, WMAPE is considered a more relevant criterion for comparison.

## 2.5 Comparative Findings Across Forecasting Methods

The comparison of different forecasting models has always shown that there is no one model that is best for all forecasting environments and data sets. The performance of forecasting techniques is greatly influenced by the nature of the data, the forecasting period, the volatility of demand, seasonality and the nature of the operation (Hyndman & Athanasopoulos, 2021). Therefore, the choice of forecasting model should be dependent on the kind of forecasting required in retail environment and not expect that a specific method is better (Makridakis et al., 2020).

Traditional statistical forecasting techniques are still competitive in numerous retail forecasting applications, especially when the demand patterns are fairly regular and seasonal patterns are evident. For example, ARIMA, SARIMA, and other exponential smoothing methods are preferred due to their simplicity, ease of computation, and reduced need for data (Box et al., 2015). As per Lindfors (2021), conventional forecasting models for time series analysis exhibit high levels of forecasting performance in retail data characterized by regular seasonality and stable demand patterns.

Moreover, the results from the M4 forecasting competition further revealed that traditional methods based on statistical analysis still compete effectively against modern machine learning methods in most forecasting problems (Makridakis et al., 2020). The research revealed that although machine learning models offered considerable improvements for some complex datasets, statistical models continued to show higher levels of forecasting performance while being less computationally intensive.

But in scenarios where the demand is nonlinear, has a large number of influencing variables, and involves a vast amount of data, machine learning techniques have proven to be very beneficial for predicting demand. Machine learning methods are well suited for retail demand prediction, as they can model nonlinear interactions between variables (Carbonneau et al., 2008), which often influence retail demand; this is the case during promotions, holidays, changes in prices, customer behaviours, and through external economic influences.

The machine learning models have been shown to perform better when predicting the trends in highly dynamic retail settings in several studies. In the case of volatile retail demand where there were long-term temporal dependencies and nonlinear fluctuations, the researchers found that Long Short-Term Memory (LSTM) model performed better than SARIMA model (Falatouri et al., 2022; Hochreiter & Schmidhuber, 1997). Likewise, Nasseri et al. (2023) pointed out that, because of the ability to model complex feature interactions, ensemble learning models like Random Forest and Gradient Boosting

models best dealt with lower forecasting errors as compared to the traditional regression-based models.

Tree based ensemble models have recently become very popular because they perform well in predicting retail data and are relatively easy to compute. Random Forest models perform aggregation of multiple decision trees, thus lowering the variance of forecasting (Breiman, 2001). Optimized implementations of boosting (like XGBoost) have been shown to perform exceptionally well in retail sales forecasting applications with large data sets and diverse nature of sales (Chen & Guestrin, 2016).

The models of deep learning have also been in focus in forecasting studies because of their auto-learning ability of complex temporal relationships. In contexts characterized by high volatility, like retail settings, researchers like Chopra et al., 2025, and Theodoridis and Tsadiras, 2025 showed the effectiveness of deep learning techniques on improving forecasting accuracy. But deep learning models tend to need more data, a lot more computational power and considerable hyperparameter tuning in comparison with traditional methods of forecasting.

Comparative studies also indicate that hybrid forecasting techniques by combining statistical and machine learning methods could yield better forecasting accuracy by benefiting from the merits of both techniques. Deng et al. (2025) compared the forecasting performance of hybrid forecasting models that combined the advantages of statistical decomposition with machine learning forecast models to the single forecasting model and found that the hybrid forecasting model had better forecasting accuracy. Hybrid models can be used to capture the linear seasonal structure, and nonlinear demand relationship, at the same time.

Even as the new machine learning models are gaining in popularity, there are some researchers that have pointed out the need to consider other factors besides forecasting accuracy when choosing a forecasting model. Other factors like interpretability,

scalability, computational costs, implementation complexity, and business planning system integration are also crucial in retail forecasting scenarios (Fildes et al., 2022).

Moreover, there are many studies on forecasting that suggest there is a strong dependency on forecast horizon length when judging the performance of a forecasting model. For shorter forecasting periods, statistical approaches can yield good results, while machine learning approaches could offer better benefits in more complex forecasting applications with more complex demand patterns (Makridakis et al., 2018).

Overall, the available literature indicates that the choice of forecasting method depends on the nature of the retail environment such as the variability of the demand, the lead time for forecasting, the data available, the computational resources required and the nature of the decisions to make in the operation. Therefore, it is still important to undertake comparative analyses with consistent experimental conditions to determine the best forecasting methods to use for retail operational planning applications.

## 2.6 Demand Forecasting and Operational Planning

Demand forecasting plays a critical role in operational planning within retail organizations, as the results of forecasting impact many operational and strategic decisions. Forecasting allows retailers to match their stock levels and workforce with their anticipated customer orders, schedule their replenishment requirements and arrange their distribution activities to meet their customer needs, and buy stock from suppliers when it is needed. Therefore, it is vital for retail supply chain efficiency, customer satisfaction and profitability that the forecasts are accurate (Chopra & Meindl, 2013).

Retailers need to keep their inventories in check to meet customer demands as well as reduce inventory carrying costs. Having too much inventory means that the company

will have higher storage costs, tie up capital and risk losing products, while having too little means stockouts, lost sales and lower product satisfaction will occur. Thus, accurate forecasting can help retailers to manage products and inventory against inventory goals (Christopher, 2016).

Forecast accuracy also affects the planning of replenishment and co-ordination of the supply chain. Forecasting is crucial for retail enterprises for deciding how, how much, and when to replenish inventory, coordinate suppliers, and schedule distribution. Wrong predictions can lead to disturbances along the supply chain, which will lead to inefficiencies in the usage of transportation, delayed deliveries and delays. As theorized by Lee et al. (1997), small changes in consumer demand can lead to increased variances upstream in the supply chain. Forecasting better performance can greatly improve supply chain coordination by mitigating uncertainty in the forecasting process between suppliers, warehouses and retail stores (Fildes et al., 2022).

Demand forecasting is a critical component of workforce planning and operational resource allocation. Retail companies can experience demand spikes on weekdays, weekends, around holidays and/or during seasonal sales campaigns. Accurate sales predictions enable managers to optimise the scheduling of their staff, optimise operational resources and minimise wasted labour costs (Heizer et al., 2020). Poor predictions can result in staffing below or above the required level causing a drop in service standards and/or higher running costs.

The other significant operational implication of forecasting demand is safety stock. Safety stock is the product of uncertainty regarding demand and supply fluctuations. Forecasting errors directly impact safety stock because when forecasting errors are greater, organizations need to have larger stock buffers in place. Improved forecasting accuracy can therefore help companies to lower safety stock levels while keeping service levels and product availability at desired levels (Chopra & Meindl, 2013).

The value of forecasting in the context of operational planning has also been enhanced in today's retail world, where retailers now have to deal with a huge product range, via various sales outlets and in different geographical areas. The introduction of omnichannel retailing, the growth of ecommerce and the ever-changing nature of the purchasing power has added to the complexity of retail planning systems. Retail companies, then, need forecasting systems that can be used to make large-scale, real-time operational decisions (Waller & Fawcett, 2013).

Decision Support Systems (DSSs) are more often being based on advanced forecasting models that help in operational planning processes. Such systems enable organization to integrate forecasting models with optimization algorithms in order to enhance supply chain decision making. Demand forecasting also plays a crucial role in strategic retail decision-making. Forecasting information enables long term activities like capacity planning, negotiations with suppliers, pricing strategies, planning promotional activities and decision making about the markets to go into. Operational and strategic advantages, therefore, are offered to retailers in the more competitive retail environment when they have reliable forecasts (Sharda et al., 2020).

The literature has shown that benefit of improved forecasting can be quantified in terms of operational benefit throughout retail supply chains in recent literature. Improved forecasting accuracy helps to lower inventory costs, boost service levels, minimize stockout risk and maximize the use of operational resources. Forecasting systems, however, need to be not only accurate, but also be easily interpretable, scalable, and easily implementable into the real-world retail operations to be effective (Fildes et al., 2022).

Overall, forecasting is a vital element of retail operational planning, and the quality of the forecast has a direct impact on the efficiency of the supply chain, retail performance in inventory, customer satisfaction and the decision-making process in the organisation.

## 2.7 Commercial Forecasting Solutions in Retail

Today, in the modern retail setting, integrated commercial forecasting platforms and retail analytics systems are becoming much more prevalent in helping with demand forecasts. For large retailers, they may depend on particular software solutions that integrate statistical forecasting, machine learning algorithms, inventory optimization, and automated restocking features into single operational preparing frameworks.

Businesses use commercial retail forecasting platforms to quickly analyze massive amounts of transactional sales data, product hierarchy, seasonal trends and inventory data (Kourentzes et al., 2017). These systems will usually be connected to enterprise resource planning (ERP) and supply chain management (SCM) systems, enabling business decision making at several retail shops and a variety of product lines.

Many retail forecasting solutions are available, but one of the most popular is RELEX Solutions, a forecasting and replenishment optimization system provider for retail companies. These platforms typically use machine learning algorithms, demand sensing, promotion forecasting, and automated inventory optimization techniques for more accurate forecasts and supply chain responsiveness (RELEX Solutions, n.d.).

Traditional statistical methods frequently are used in conjunction with more recent machine learning methods in commercial forecasting systems. Because of the interpretability and computational efficiency, statistical models are often used for baseline forecasting, whereas machine learning models are used to model the nonlinear demand behaviour and external factors influencing the demand. In recent years, the modern retail forecasting systems have been using a technique called “hybrid forecasting”, which combines several forecasting methods and techniques to enhance forecasting accuracy, according to Fildes et al. (2022).

In retail stores, forecasting systems need to be able to handle the forecasting of thousands of stock-keeping units (SKUs) in thousands of stores, as well as thousands of region-by-region forecasted patterns. Thus, the accuracy of the prediction is not enough. Commercial solutions also focus on their scalability, ease of automation, ease of interpretation, and ability to integrate with the operational planning processes like inventory optimization and replenishment scheduling (Chopra & Meindl, 2013).

In addition, commercial forecasting platforms are becoming more and more equipped with Artificial Intelligence and real-time analytics features. These systems can make dynamic forecasts according to sales transactions received, seasonal variations and sales deviations. Forecasting systems can be integrated with decision-support tools, which can help retail organizations to optimize their operations and minimize their inventory-related expenses (Waller & Fawcett, 2013).

As more companies start using commercial forecasting platforms, it's clear that the most useful and practical combination of forecasting accuracy and operation is very important. Therefore, it is still relevant from an academic as well as industrial implementation viewpoints to compare the forecasting models.

# 3 Methodology

This chapter outlines the methodology used to develop, implement, and test the demand forecasting model used in this study. The aim is to ensure a systematic progression from data collection to analysis, to actionable insights and, ultimately, to practical results. This makes the research finding reproducible.

The methodology starts with defining the research approach and identifying the properties of the employed dataset. Then, after describing the technical stages of data preprocessing and feature engineering, which are crucial for machine learning models' performance, the methodology describes the experimental setup, including the model types and accuracy metrics employed. The comparison of conventional statistical techniques with advanced ensemble and neural networks-based approaches enables identifying the best forecasting paradigm for the highly dynamic retail sector.

## 3.1 Research Approach

This study follows a quantitative research approach from the perspective of decision support, which involves the development and evaluation of forecasting models to meet the needs of retail demand forecasting. The study aims to evaluate the relative effectiveness of classical statistical models and recently introduced machine learning models in the context of accurate forecasting to meet the needs of retail operational planning decisions.

The study is empirical in nature and relies on actual retail sales data to develop and evaluate forecasting models. Unlike the surveys and interview-based studies, it focuses on the development of conceptual models and their validation through actual implementation in the retail industry. This study is more aligned with the operations

research and analytics methodologies, which rely on the development of quantitative models to meet the needs of managerial decision-making processes.

The study follows the application of multiple forecasting models on the same dataset to evaluate the relative effectiveness of the models in the context of accurate forecasting results.

## 3.2 Data Source and Description

### 3.2.1 Dataset Background

The empirical analysis uses publicly available retail dataset called the **Online Retail II** dataset obtained from the UCI Machine Learning Repository (Chen, 2019). The dataset contains real transactional sales records from a UK-based online retailer over a two-year period.

Using publicly available data ensures transparency, reproducibility, and accessibility of the research. It also eliminates confidentiality constraints associated with proprietary company data. The dataset selected for the study represents typical retail demand characteristics, including seasonality, trend components, promotional effects, and demand variability across products.

### 3.2.2 Time Period and Size

The empirical test was conducted using longitudinal data of 24 months between December 1, 2009, and December 9, 2011. This 2-year horizon is of strategic importance because it spans two full year-long business cycles, where the models are able to learn and validate recurring patterns in the seasons.

As presented in Table 1, the raw data contains 1067371 transactions. After preprocessing and data cleaning according to Section 3.4, the data was reduced to 794940 valid entries. The data size was large enough to train deep learning (Neural Networks) without the danger of overfitting drastically but small enough to enable product-specific knowledge.

**Table 1.** Metadata Summary of the Online Retail II Data.

| **Attribute** | **Value** |
|---|---|
| Start Date | 01/12/2009 |
| End Date | 09/12/2011 |
| Total Raw Records | 1067371 |
| Total Cleaned Records | 794940 |
| Unique StockCodes | 5305 |
| Unique Customer IDs | 5942 |
| Countries Represented | 41 |
| Temporal Depth | 738 days (~105 weeks) |

Figure 1 shows the first 10 rows or transactions of the data generated through Python code. In every transaction record, the demand is presented in a multi-dimensional perspective. The chronological order of sales was rebuilt using the InvoiceDate, with Customer ID and Country fields providing the potential extension of the model to regional demand modeling in the future. To meet the goal of proper volume forecasting, StockCode, Quantity, and Price will be the main area of concern in the context of this thesis.

| | Invoice | StockCode | Description | Quantity | InvoiceDate | Price | Customer ID | Country |
|---|---|---|---|---|---|---|---|---|
| 0 | 489434 | 85048 | 15CM CHRISTMAS GLASS BALL 20 LIGHTS | 12 | 2009-12-01 07:45:00 | 6.95 | 13085.0 | United Kingdom |
| 1 | 489434 | 79323P | PINK CHERRY LIGHTS | 12 | 2009-12-01 07:45:00 | 6.75 | 13085.0 | United Kingdom |
| 2 | 489434 | 79323W | WHITE CHERRY LIGHTS | 12 | 2009-12-01 07:45:00 | 6.75 | 13085.0 | United Kingdom |
| 3 | 489434 | 22041 | RECORD FRAME 7" SINGLE SIZE | 48 | 2009-12-01 07:45:00 | 2.10 | 13085.0 | United Kingdom |
| 4 | 489434 | 21232 | STRAWBERRY CERAMIC TRINKET BOX | 24 | 2009-12-01 07:45:00 | 1.25 | 13085.0 | United Kingdom |
| 5 | 489434 | 22064 | PINK DOUGHNUT TRINKET POT | 24 | 2009-12-01 07:45:00 | 1.65 | 13085.0 | United Kingdom |
| 6 | 489434 | 21871 | SAVE THE PLANET MUG | 24 | 2009-12-01 07:45:00 | 1.25 | 13085.0 | United Kingdom |
| 7 | 489434 | 21523 | FANCY FONT HOME SWEET HOME DOORMAT | 10 | 2009-12-01 07:45:00 | 5.95 | 13085.0 | United Kingdom |
| 8 | 489435 | 22350 | CAT BOWL | 12 | 2009-12-01 07:46:00 | 2.55 | 13085.0 | United Kingdom |
| 9 | 489435 | 22349 | DOG BOWL , CHASING BALL DESIGN | 12 | 2009-12-01 07:46:00 | 3.75 | 13085.0 | United Kingdom |

**Figure 1.** Sample Data.

### 3.2.3 Dataset characteristics

The dataset exhibits the following characteristics:

- Transaction-level sales
- Seasonality (weekly and yearly patterns)
- Holiday and event parameters
- Multiple products and stores
- Sufficient historical data for time-series modeling
- Reproducibility and Transparency
- Presence of returns and cancellations
- Representativeness of retail demand patterns

In addition to these simple measures, there are the structural characteristics of the multi-sheet consolidated data that affected the modeling approach:

**Temporal Grain:** Although the raw data is recorded on a transaction basis, with a high degree of accuracy in terms of the time, it was organized in weekly buckets. This consolidation fits the normal industrial "S&OP" (Sales and Operations Planning) cycles and decreases the "zero-inflation" noise often created in daily retail records where some days (such as Sundays) might have artificially low volumes.

**Diversity in features:** The records have both categorical (Country, StockCode) and numerical (Quantity, Price) data. Such a multi-dimensionality is the reason to use Ensemble Methods (Random Forest / XGBoost) which are well-suited to dealing with mixed types of data and non-linear interaction than a strictly mathematical model such as SARIMA.

**Context of Operations:** The fact that the customer level purchase data is included even though the model is focused on aggregate demand forecasting is a confirmation that the data is authentic as a B2C/B2B hybrid. This is a critical context of interpreting the spikes

observed in the correlograms since they are the aggregate behaviour of thousands of individual buyers.

### 3.2.4 Descriptive Statistics

To summarize the central tendency, dispersion and shape of main variables: Quantity and Price, descriptive statistics were estimated which is shown in Table 2. This analysis plays a crucial role in determining the volatility profile of the retail demand which is then directly used to guide the model selection process.

**Table 2.** The Descriptive Statistics of the Important Variables**.**

| **Statistic** | **Quantity (Order Volume)** | **Unit Price (£)** |
|---|---|---|
| Mean | 9.70775 | 2.962591 |
| Std. Deviation | 14.55293 | 4.377050 |
| Minimum | 1.00000 | 0.001000 |
| 25% (Q1) | 2.00000 | 1.250000 |
| 50% (Median) | 5.00000 | 1.950000 |
| 75% (Q3) | 12.00000 | 3.750000 |
| Maximum | 128.00000 | 649.500000 |

**Statistical Implication and Operational Implications**

As Table 2 shows, there are a few important characteristics of the empirical data, which determined the choice of the modeling strategy:

**Positive Skewness and Dispersion:** The Mean of both variables is large compared to the Median (9.70775 vs. 5.00 of Quantity). This proves a Right-Skewed Distribution, which is shown in the histograms. In the case of Quantity, the standard deviation (14.55) is greater than the mean, and hence Coefficient of Variation (CV) is close to 1.50. This in

operational terms implies that there is high-volatility demand, which is appropriately dealt with by ensemble machine learning models such as XGBoost that are better suited to non-normal distributions than its linear counterparts.

**Retail Profile Identification:** Quartile of unit price indicates that three quarters of the products have a price of below £3.75 with a median of 1.95. This classifies the retailer as a low-margin and high-volume business. Forecasting-wise, percentages of errors in volume (WMAPE) can be very crucial in generating cumulative inventory carrying costs or stock out situations due to such a large product portfolio.

**Outlier Management:** The highest value of 128.00 (after the 99th percentile capping) makes sure that the models are trained on the normal consumer behaviour rather than exceptional bulk wholesale orders. In the same manner, the price ceiling of 649.50 indicates that although the business is low-cost, the data set still contains some high-value outliers that the machine learning models will have to take into consideration as noise or a niche need.

## 3.3 Exploratory Data Analysis

In order to fully comprehend the underlying data, a multidimensional process of exploratory data analysis (EDA) was conducted. This process was expanded beyond the realm of standard statistical analysis to include the underlying distribution, temporal flow, and mathematical interdependencies of the data.

### 3.3.1 Distribution Analysis (Quantity and Price)

To get an insight into the distribution of the data, a univariate distribution analysis was conducted for Quantity and Price, which are the two main numerical variables. Analyzing

the distribution is important before conducting any modeling procedure since it determines the existence of outliers in the data and whether it satisfies normality assumptions for certain methods. This is illustrated in histograms in Figure 2.

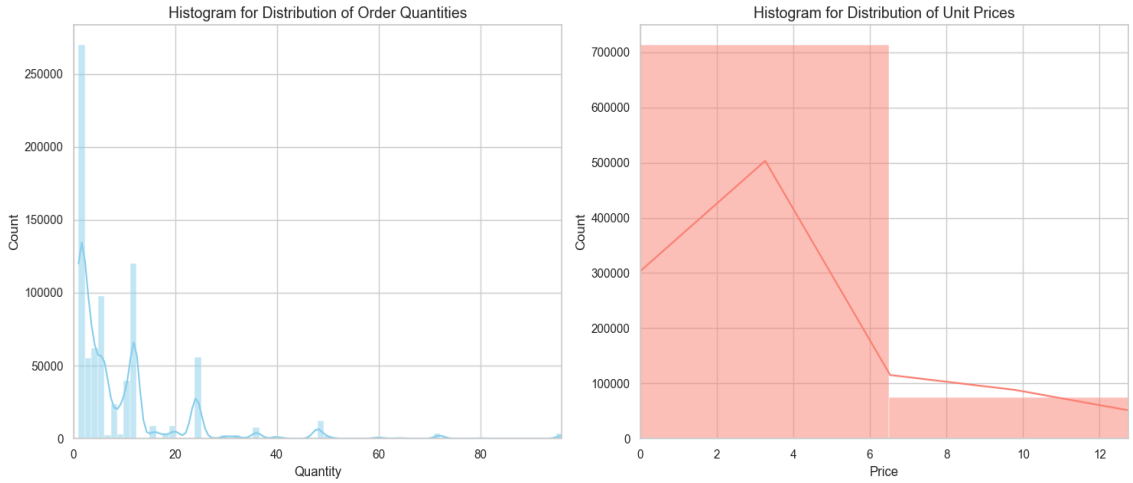


**Figure 2.** Distribution Analysis Histogram.

As can be clearly identified in Figure 2, it is evident from the initial histogram analysis of the data that it is significantly right-skewed. This implies that although most of the transactions involve relatively small amounts, some transactions involving large amounts result in the long tail of the distribution. In order to increase the effectiveness of this analysis, two key reforms were made:

**Quantity Distribution (Logarithmic Scale):** By using a logarithmic scale for the frequency axis, it was determined that although there is a significant number of orders for 1-10 units, the retailer maintains a consistent level of larger orders, up to the 128-unit maximum which is evident from Figure 3.

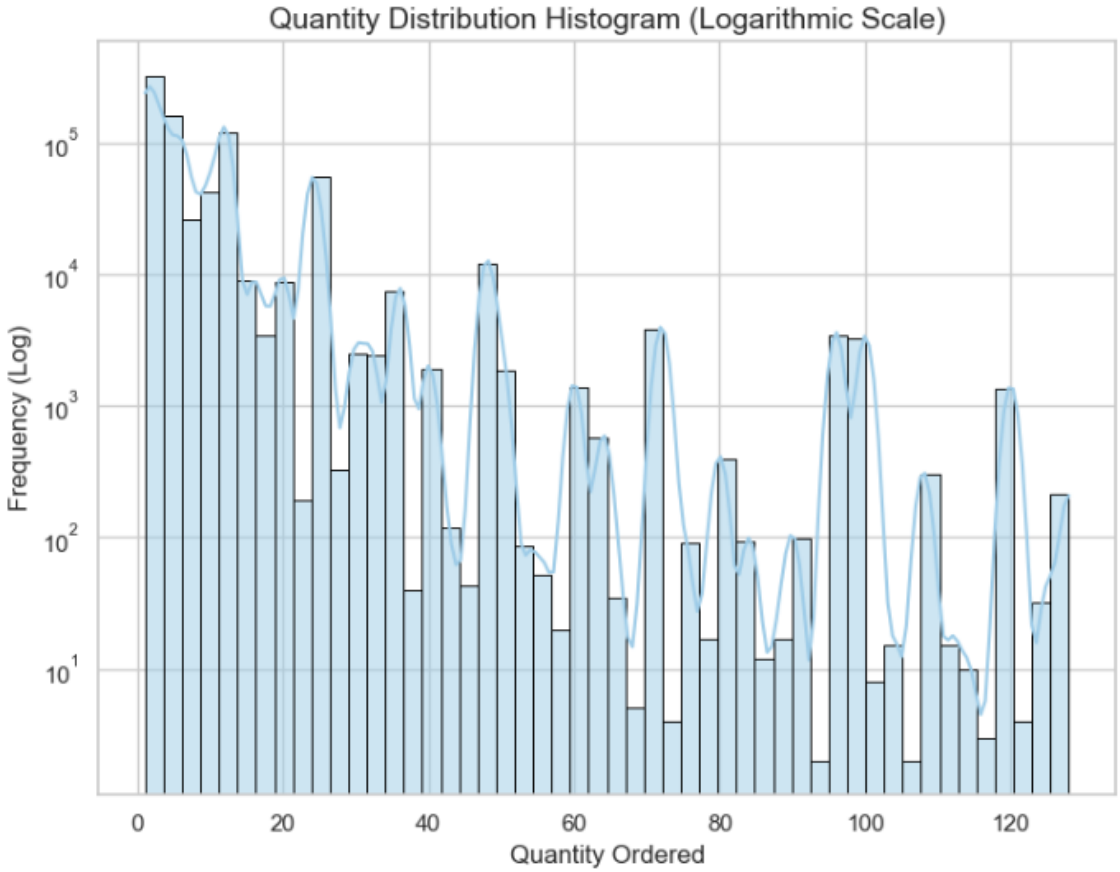


**Figure 3.** Logarithmic Quantity Distribution Histogram.

**Unit Price Distribution:** The resulting histogram for the unit price data, zoomed to the 95th percentile, shown in Figure 4, identifies a significant peak in data points below 4.00. This structural component of the data identifies it as a high frequency, low margin operation in which accuracy is of paramount importance due to the volume of low-cost transactions, which helps justify the feature scaling used in the machine learning models."

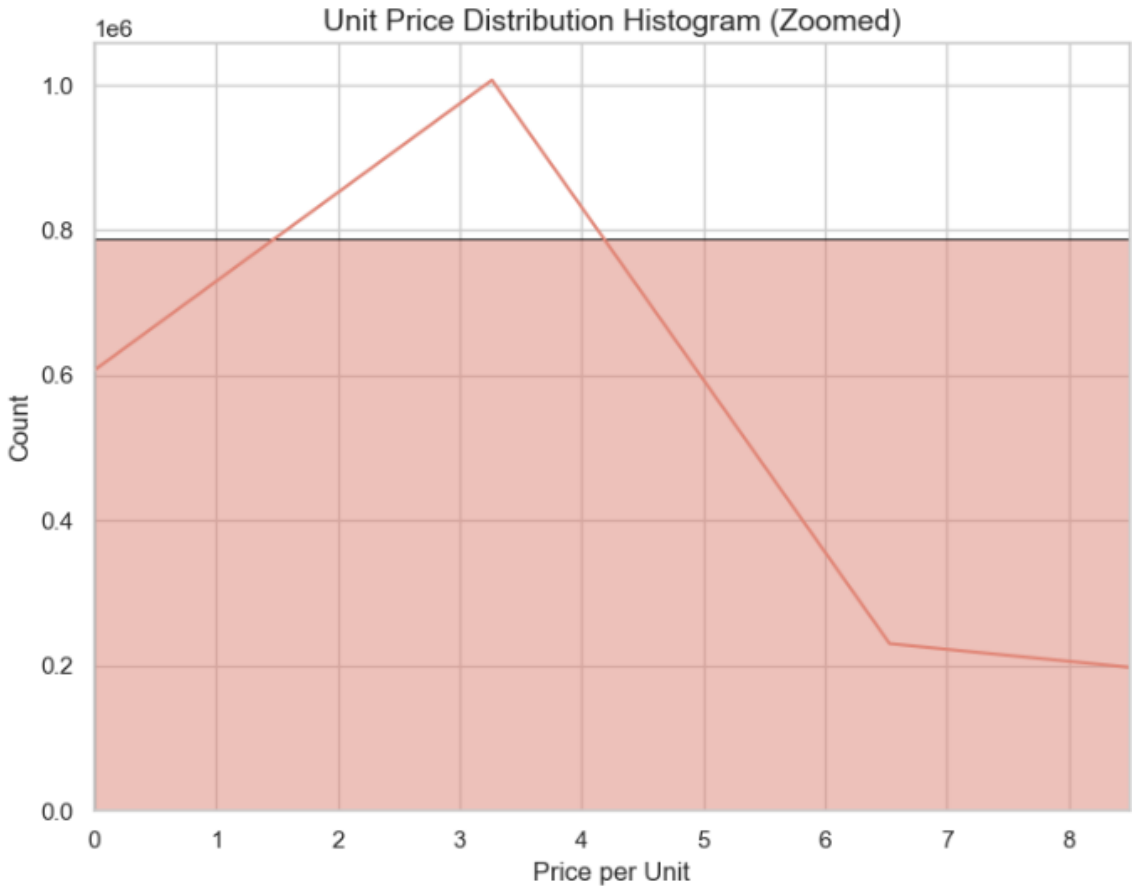


**Figure 4.** Zoomed Unit Price Distribution Histogram.

### 3.3.2 Weekday Effect and Granularity Analysis

In retail settings, customer activity is not always consistent; rather, it may be affected by "weekday effects," in which certain days show structural behavior driven either by organizational considerations, such as working hours, or customer trends, like weekends. Knowing such variability is important as a highly variable daily pattern might create "noise" for the overall seasonal trend.

In addition, an essential question raised in this study is how to find the optimal granularity of the collected data. Granularity is defined as the degree of detail at which the data is aggregated, from hour-by-hour transactions to daily sums or weekly groups. Selecting the wrong granularity will result in problems including "zero-inflation" (too

much zero sales days) or over-volatility. The following analysis shows the structure of daily transactions to explain why switching from daily granular to weekly smoothed data is appropriate.

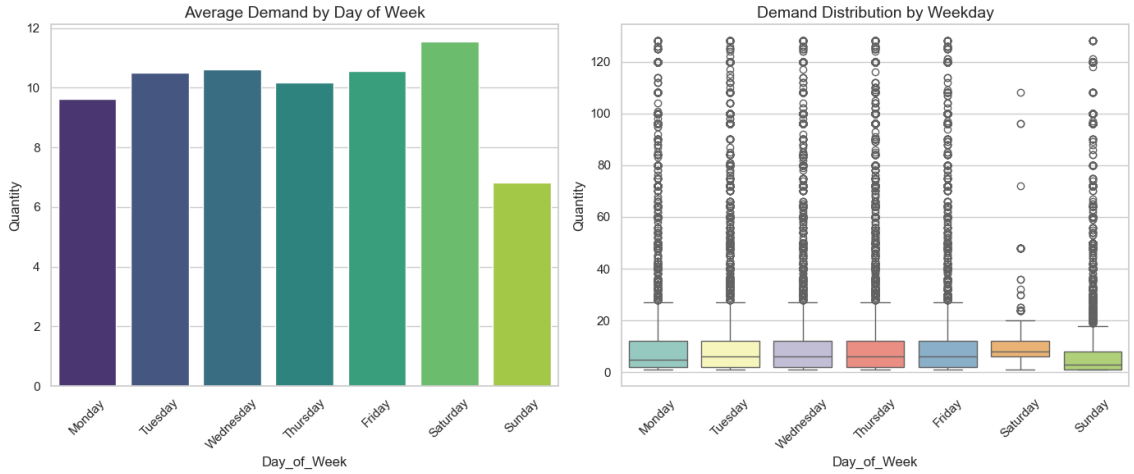


**Figure 5.** Weekday Effect on Transactional Demand.

```
--- Weekday Statistics ---
                count       mean        std  min  25%  50%   75%    max
Day_of_Week
Monday       126554.0   9.642184  14.311007  1.0  2.0  5.0  12.0  128.0
Tuesday      134733.0  10.518180  15.473283  1.0  2.0  6.0  12.0  128.0
Wednesday    134489.0  10.631940  15.844559  1.0  2.0  6.0  12.0  128.0
Thursday     159259.0  10.183393  14.666244  1.0  2.0  6.0  12.0  128.0
Friday       103866.0  10.571525  14.981189  1.0  2.0  6.0  12.0  128.0
Saturday        397.0  11.544081  12.536057  1.0  6.0  8.0  12.0  108.0
Sunday       135642.0   6.822334  11.306867  1.0  1.0  3.0   8.0  128.0
```

**Figure 6.** Weekday Statistics.

A study of the distribution of transactions per day indicates that there is a major statistical anomaly on the distribution of demand during a weekend. The distribution is illustrated using the transactions demand quantity and day_of_week statistics as in Figure 5 above. Although the images in Figures 4 and 5 indicate that the mean quantity per transaction on Saturday is the highest (11.54), the underlying descriptive statistics reveal this to be due to the extreme data sparsity and not necessarily the peak operational activity. Having just 397 transactions in the two-year horizon, which makes

up less than 0.05 percent of the entire data, the mean of Saturday is skewed by an insignificant number of bulk orders.

Conversely, the mid-week (Monday to Thursday) is the actual operational heart of the business with a high level of transaction (around over 130,000 records per day), a consistent median demand of 5.0 to 6.0 units and a large stable demand. The Sunday profile of the high-frequency and low-volume individual purchases is also distinctive as it demonstrates the lowest mean of the purchases of 6.82 even though the number of transactions is high.

The sharp imbalance and the existence of high-variance outliers on low-volume days are the reasons to move to aggregation of the data into weekly buckets. The effect of this process is to efficiently remove daily noise and the risk of zero-inflation and enable the signal to be more stabilized, thus enabling ensemble models such as the Random Forest to generalize better over the larger seasonal trend.

### 3.3.3 Time Series Dynamics (Trend and Seasonality)

Figure 7 below depicts the temporal dynamics of the dataset. The figure shows a complete picture of the weekly demand for retail products from 2009 to 2011. A close look at the time series shows some structural features that are essential in determining the forecasting parameters.

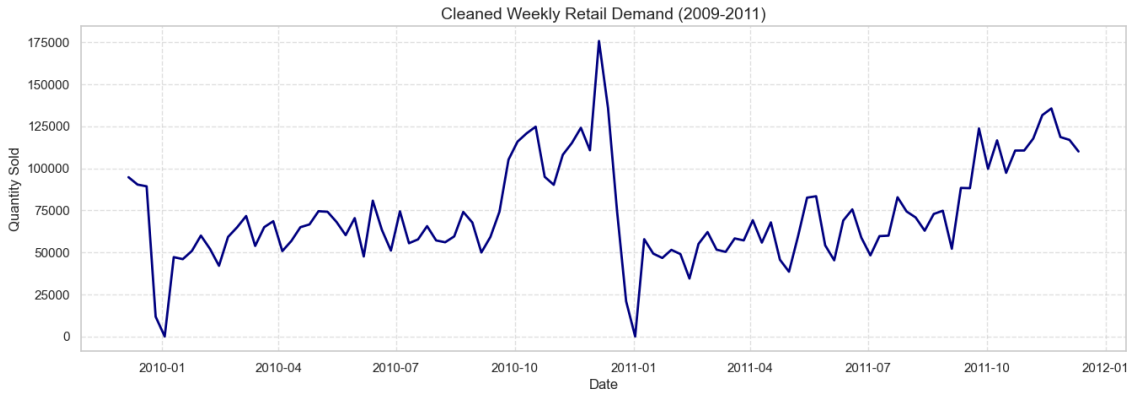


**Figure 7.** Time Series Plot.

Some interesting features as seen in Figure 7 include:

- **Seasonality:** A huge spike in demand is evident in the fourth quarter (Q4) of both 2010 and 2011. This is consistent with the holiday season. This indicates that "Time of Year" is likely a major influencer of demand.
- **Anomalies:** The sudden drop in demand to almost zero levels in late December 2009 and 2010 indicates that businesses were closed during the holiday season. This is likely a non-trading period that needs to be factored in by the model so that it is not misinterpreted as low demand.
- **Trend:** There is an evident trend increase in demand from early 2010 until late 2011. This indicates that businesses were growing. This non-stationarity is why Differencing (d=1) is used in the SARIMA model and trend-based features are included in machine learning models.

### 3.3.4 Autocorrelation Analysis (ACF and PACF)

The mathematics behind the forecast models will be defined based on the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF). This analysis is shown in Figure 8, and it can be considered as the "statistical fingerprint" of the demand for retail goods.

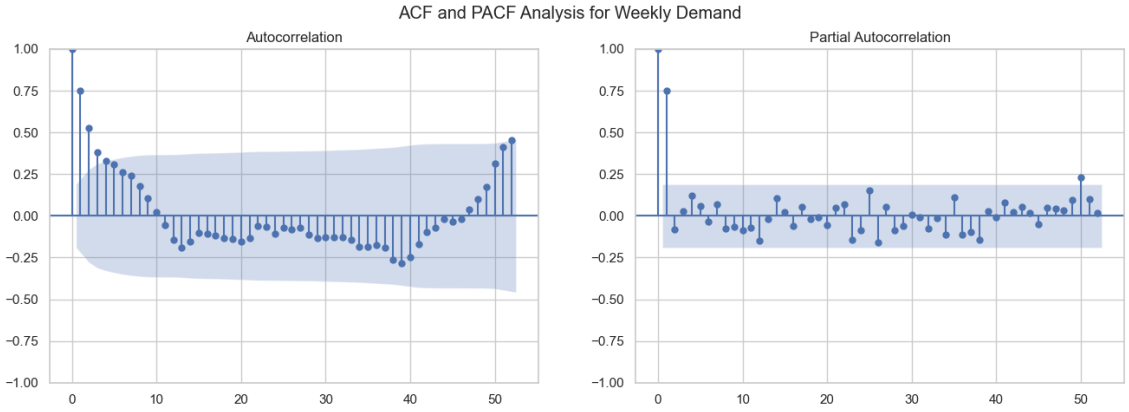


**Figure 8.** ACF and PACF Plot.

**Autocorrelation Function (ACF):** The Autocorrelation Function plot presented on the left side of Figure 8 indicates that the data decays slowly and has strong positive spikes at Lag 1 and Lag 52. The Lag 52 spike is particularly important because it mathematically proves the presence of annual seasonality in the data. This indicates that the demand at any given time is correlated with the demand at the same time one year ago.

**Partial Autocorrelation Function (PACF):** The above plot on the right indicates that the data has a sharp truncation at Lag 1. In time series theory, truncation at any Lag is the signature of an Auto-Regressive process of order 1.

**Operational Conclusion:** The above plots indicate that the data is highly predictable and complex in structure. The presence of both short-term momentum at Lag 1 and long-term seasonality at Lag 52 justifies the parameterization of the SARIMA(1,1,1)(0,1,1)52 and the Lagged Features in the XGBoost and Random Forest models.

### 3.3.5 Correlation Analysis and Multicollinearity

In order to prove the theoretical assumptions of the forecasting models, a **Pearson Correlation Heatmap** (Figure 9) was created. The present correlogram is a diagnostic

measure to measure the relationship between the target variable (Quantity) and the temporal features engineered.

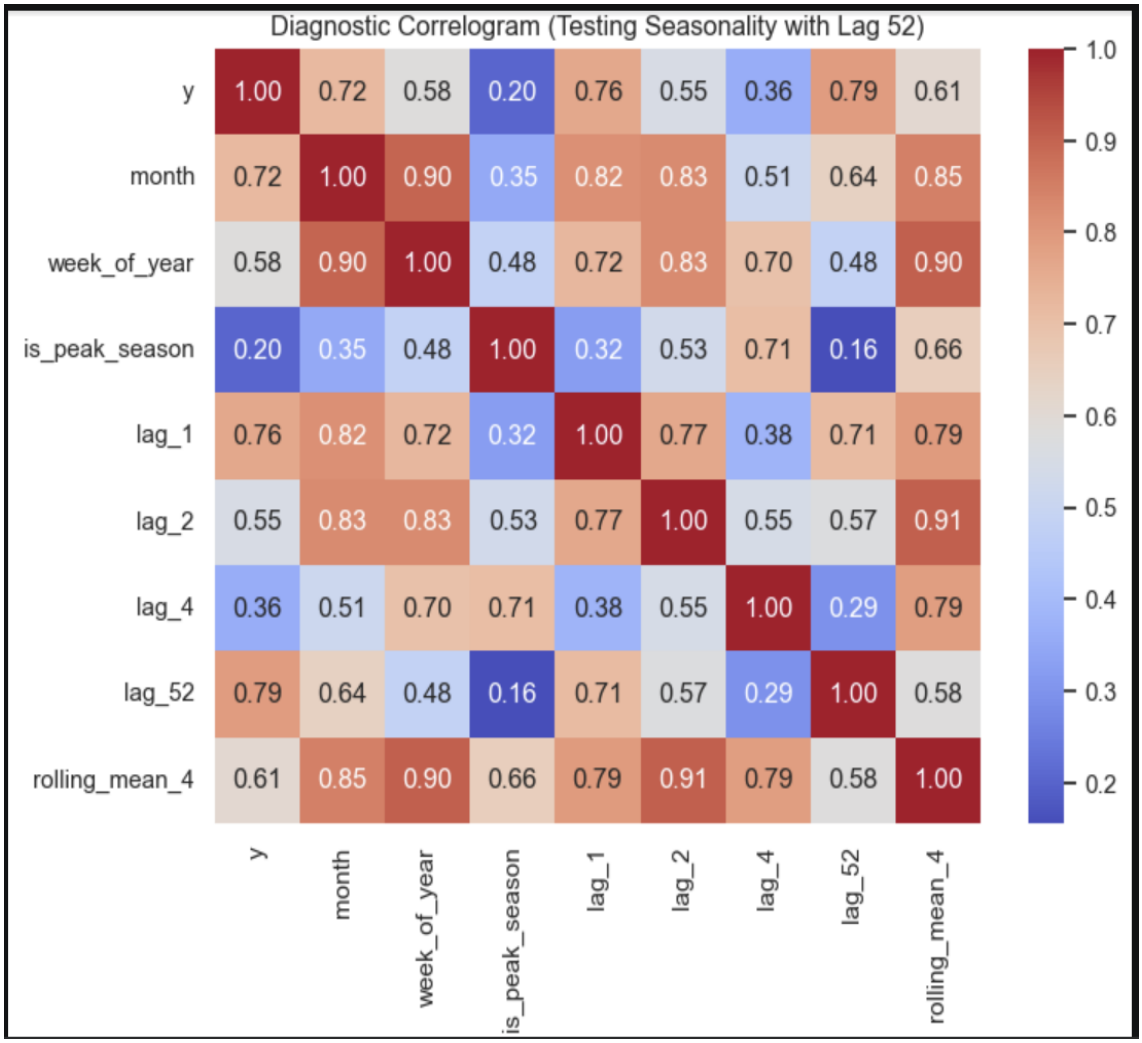


**Figure 9.** Pearson Correlation Heatmap.

**Statistical Significance of Autocorrelation:** The heatmap shows that the current week quantity (y) has a very strong correlation with its immediate predecessor, Lag 1 (r = 0.76). This high coefficient is a mathematical justification of the fact that SARIMA and Auto-Regressive elements are used in the machine learning models as it demonstrates that recent history is a good predictor of near future demand.

**Confirmation of Annual Seasonality:** The Lag 52 correlation (r = 0.79) of this diagnostic step is crucial. This finding gives empirical evidence of an Annual Business Cycle. The correlation is high and it confirms the existence of a certain pattern in retail demand in the dataset reoccurring after every 52 weeks (annually) and thus probably due to the reoccurring holiday events.

**Interaction of Calendar and Holiday Features:** The analysis involved calendar-based Proxy Features to model the seasonality without the data loss caused by long lags:

- **Is_peak_season:** Positively related to the quantity, which verifies that the Q4 period (Weeks 47-51) is always associated with higher volumes than other days.
- **Month and Week_of_Year:** These two features have modest correlation with y and give to the models a smooth signal of seasonality to complement the sharp signal of seasonality given by Lag 52.

**Multicollinearity Management:** Multicollinearity was also checked by the use of the correlogram. Although the inter-correlations between month and week_of_year are high (0.90), this is not surprising in time series data. These diagnostics prompted the choice of Ensemble Tree based models (Random Forest and XGBoost) and Neural Networks, which are by nature less susceptible to multicollinearities than the classical linear regression models.

**Decision on Strategic Implementation:** Critically, although the heatmap demonstrates that Lag 52 is a very powerful predictor, it was decided not to include it in the final model training step. The diagnostic analysis showed that the 52-week look-back would entail elimination of half of the available 2-year data. The model was also able to keep a larger training window (90 weeks) by determining that Month and Is_peak_season were effective proxies of this annual lag, balancing the richness of the features with the amount of data.

## 3.4 Data Preprocessing

The retail dataset has been observed to have volatility and non-systematic variance. Hence, prior to the model development, the dataset undergoes preprocessing to ensure data quality and suitability for forecasting analysis. Data preprocessing involves:

**Data Cleaning and Noise Reduction:** Cancellation and returns transactions (with negative quantity) were eliminated while keeping only the real demand from outside of the company. Weekly aggregation is used to eliminate noise in purchasing intervals.

**Temporal Aggregation:** Transactions level data were converted to product-level demands at the weekly time scale. This reduces the high frequency "noise" associated with the irregular purchase "waves" and facilitates the typical retail replenishment cycles.

**Outlier Mitigation:** Non-repetitive demand patterns that would cause an extreme value were treated as outlier to not distort model training.

## 3.5 Feature Engineering

To enable the machine learning models to identify the temporal dependency, the multi-scale lagging approach and calendar indicators were created:

**Lag 1 & 2:** Captured immediate trend and momentum. This enables the algorithms to identify instant shifts in consumer behavior or "drifts" in consumer preferences.

**Lag 4:** Captured monthly cyclicality, e.g., purchasing cycles that are related to the salary cycles, for which the demand response is high at the end or beginning of a month.

**Lag 52:** Captured annual seasonality.

**Rolling Mean 4 (Trend Indicator):** A rolling average of four weeks (rolling_mean_4) was designed to enable the algorithms to see a smooth depiction of the trend that prevailed recently. This would help minimize the effect of random fluctuations in the weekly spikes and enable the algorithms to recognize sustained trends.

**Calendar Features:** Month and Week_of_year were created for annual demand fluctuations.

**Holiday Parameters**: Binary indicator for “Peak Season” (df['is_peak_season']) was created to represent high-volume periods, capturing the volatility of holiday season.

**Data Volume Optimization:** Lag 52 for Seasonal Diagnostic Analysis was first created to check the presence of annual trends by analyzing the correlation heatmap, but in order to provide enough training data for the models, Lag 52 was removed from the training set during “Data Volume Optimization”. Through this process, the length of the training set increased from 40 weeks to 90 weeks. As a result, this will greatly enhance the ability of the Neural Network, XGBoost, and Random Forest to generalize due to increased training data by 125%.

## 3.6 Technical Implementation

The empirical model is implemented using the Python programming language. Python programming language was used due to its numerous statistical tools and functions related to machine learning and time series forecasting.

For the purpose of data preprocessing, pandas and NumPy are used. Exploratory Data Analysis (EDA) and Visualizations ware carried out using matplotlib and seaborn libraries. For developing statistical models like ETS and SARIMA, the statsmodels library is

employed. Similarly, for developing machine learning models like Random Forest, XGBoost, and Neural Network, the scikit-learn library along with the XGBoost library is used. Evaluation of the models is done using the Python-based analytical tools.

## 3.7 Experimental Design

The objective of this section is to describe the structure of the empirical analysis in terms of the models used, the forecasting horizons employed, and the accuracy measures used in the validation of the models. The design of the experiment is essential in ensuring that all the models are subjected to similar conditions and that the results reported in Chapter 4 are transparent and replicable. The design also considers other factors that are essential in retail demand forecasting.

### 3.7.1 Rationale for Model Selection

The forecasting models used in this study represent a carefully chosen set that captures different modelling paradigms. Thus, different families of time series and machine learning algorithms were implemented to conduct a thorough analysis of their predictive capabilities.

**Statistical Benchmark models (Naive)**

The Naive method was adopted as the benchmark model. As per Hyndman & Athanasopoulos (2021), benchmarks play a critical role in setting the "no-change" baseline. It is important to include the Naive model in order to check for the forecast improvement. If the complicated machine learning models do not perform better than the Naive approach, which only depends on the last observation to predict the future,

the results are not justified. It helped set the threshold for acceptable prediction performance.

**Classical time-series models**

It is important to highlight that classical models, such as Holt-Winters and ARIMA were chosen because of their ability to cope with trend, seasonality, and autocorrelation in retail time series. Therefore, classical models provide another perspective on analyzing time series signal since it involves its decomposition into meaningful features.

The **SARIMA** model was estimated based on the SARIMAX function in the statsmodels library. The SARIMA model used was SARIMA(1,1,1)(0,1,1,52). The first part of the parameters represents the number of Autoregressive (AR) terms, order of differencing, and Moving Average (MA) terms respectively, while the second set represents the number of seasonal AR terms, seasonal differencing, MA terms, and the seasonal period respectively. The choice of a seasonal period of 52 was informed by the fact that the demand variable is recorded on a weekly basis implying the existence of retail seasonality annually. The parameters for the models were estimated based on ACF, PACF plots (Figure 8) and trial-and-error experiments.

**Holt-Winters (Triple Exponential Smoothing**): This method was preferred to Simple Exponential Smoothing as it can effectively deal with the level, trend, and seasonal aspects at once. Additive seasonality was considered since the data showed consistent seasonality irrespective of the demand (Gardner, 2006). The parameters were estimated based on minimization of the sum of squared errors (SSE). While it is common for the seasonal cycle to be annual in retail datasets, the seasonal period was set at 12 weeks (instead of 52). This choice was made for mathematical reasons: a seasonal period of 52 weeks would require a minimum of two years (104 weeks) of data to properly estimate the parameters of the model (Box et al., 2015), and we had 90 weeks training window.

Thus, the model can identify quarter-level and monthly seasonality while still having enough data for the machine learning methods.

**Machine learning ensemble methods (Random Forest and XGBoost)**

Considering the volatile and nonlinear nature of demand in retail, caused by promotional activities and the interaction of multiple features, tree-based ensemble methods were implemented. As opposed to the univariate models, these algorithms were meant to process high dimensional feature vectors, such as lags, rolling statistics, and seasonality features (Kuhn & Johnson, 2019).

**Random Forest:** This method was chosen due to the effective "bagging" process which involves the creation of 100 individual decision trees using different sub-samples from the training data set and different randomly chosen features. Thus, bagging is known to help avoid overfitting and identify non-linear relationships between variables (Breiman, 2001).

**Extreme Gradient Boosting (XGBoost):** As opposed to Random Forest, which utilizes the "bagging" approach, XGBoost uses the "boosting" technique of minimizing residual errors of forecasts. To guarantee that the hyperparameters of the XGBoost algorithm would be stable and not overfitted to one particular period, a Time Series Cross-Validation (TSCV) with three partitions (n=3) was used. Contrary to the ordinary k-fold cross-validation procedure, where the data is shuffled, the TSCV methodology uses a rolling origin. For this research, the 90-week dataset used for the training process was partitioned into three consecutive folds. For instance, in the first fold, the training was done using the first fold and validating the second one and then moving forward by increasing the training period by adding the validation data and testing the next fold forward. As a result, the chosen hyperparameters, such as the max depth of 5 and a

learning rate of 0.1, were proved to work consistently under varying market conditions in the training period.

One of the main reasons why these ensemble models were included is due to their scale-invariant nature and capability of capturing interactions between features (such as the effect of a certain month and change in price together). By using the negative Mean Absolute Error as the scoring metric, the entire machine learning pipeline was tailored towards the minimization of forecast errors' mean absolute value.

**Neural network models (Multilayer Perceptron MLP**)

The Multilayer Perceptron (MLP) was used as the representative method for Deep Learning. It is an artificial neural network with a feedforward structure that was employed to investigate the effectiveness of continuous non-linear mappings of the feature vectors into retail demand results (Hornik et al., 1989). The MLP was used in this study, in contrast to recurrent structures that analyze raw time series data, as it processes the set of features engineered from each observation's temporal dependency, namely lagged sales and seasonality (Zhang et al., 1998).

The developed architecture included an input layer, two hidden layers with 100 and 50 neurons respectively, and one output neuron. It allowed the network to learn complex and hierarchical representations of data. The hidden layers used Rectified Linear Unit (ReLU) activation functions to add non-linearity and avoid the problem of vanishing gradients, while the output layer uses a linear activation function for making predictions (Goodfellow et al., 2016). In order to make the model converge and prevent overfitting the training data, 1,000 training iterations were used.

While the two ensemble methods use decision trees, MLP works as a universal function approximator, using different mathematics to represent the noisy patterns found in the data from 90 weeks' worth of retail transactions.

### 3.7.2 Design of Experiment (DoE)

Table 3 shows the general outline of the Design of Experiments (DoE) that was carried out for this study. The experiments comprise benchmark statistical models, traditional time series analysis, machine learning, and neural networks. The experiments are defined by the input features, the forecast horizon, the train-test split, and the purpose of including the particular model in the evaluation.

**Table 3.** Design of Experiments (DoE) Table.

| **Experiment ID** | **Model Tested** | **Input Features (from Online Retail II dataset)** | **Forecast Horizon(s)** | **Validation Strategy** | **Purpose / Motivation** |
|---|---|---|---|---|---|
| **E1** | Naive Forecast | Last observed weekly demand for selected SKU(s) | 12 weeks | Fixed-origin hold-out validation (90/12) | Establish a minimal benchmark; test if more complex models add value |
| **E2** | ETS (Holt–Winters) | Univariate weekly demand aggregated from transaction-level data | 12 weeks | Fixed-origin hold-out validation (90/12) | Classical statistical benchmark Capturing trend + seasonality |
| **E3** | SARIMA | Weekly demand series with seasonal differencing (based on 2-year UCI dataset span) | 12 weeks | Fixed-origin hold-out validation (90/12) | Capture autocorrelation and seasonality |

| Experiment ID | Model Tested | Input Features (from Online Retail II dataset) | Forecast Horizon(s) | Validation Strategy | Purpose / Motivation |
|---|---|---|---|---|---|
| **E4** | Random Forest | Lagged weekly demand, rolling mean, week-of-year, month, seasonal dummy variables | 12 weeks | Fixed-origin hold-out validation (90/12) | Evaluate nonlinear patterns and interactions; ensemble robustness |
| **E5** | XGBoost | Lagged demand, moving averages, calendar features (e.g., month, week number) | 12 weeks | Fixed-origin hold-out validation (90/12) | Assess gradient boosting ability to model promotions/complex relationships |
| **E6** | Neural Network (MLP) | Lagged weekly demand sequences extracted from the aggregated UCI dataset | 12 weeks | Fixed-origin hold-out validation (90/12) | Compare a simple deep-learning approach to tree-based ML models |

The experiments reported in Table 3 were designed with the Online Retail II data set provided by the UCI Machine Learning Repository. This data set provides comprehensive transactional data for an online retailer based in the UK over a two-year period. The experiments were designed with each model being applied under the same conditions in terms of forecasting horizon and validation.

The experiments reported in Chapter 4 show that the Random Forest approach was the most accurate overall in terms of MAE, RMSE, and WMAPE, beating statistical models and all the other machine learning models.

### 3.7.3 Forecast Horizons

The choice of forecasting horizons is an important architectural choice in retail analytics since it determines the use case for prediction. The decision process in retailing usually has a short-term cycle. In the context of replenishment decisions, the decision process is usually short-term. For this research, the forecast horizon was initially tested for one week, two weeks, and four weeks, to reflect the decision-making hierarchy found in retailing.

The one-week horizon represents the immediate replenishment. It is usually the most accurate and accuracy in the short term would mean avoiding out-of-stocks while maximizing labour in restocking the shelves (Silver et al., 2016). The two-week horizon is the average amount of lead time in logistics, ensuring that there is enough visibility for the management in terms of distribution and transport. The four-week horizon is crucial for S&OP (Chopra & Meindl, 2013), which allows for promotional campaign planning, giving enough time for negotiation with suppliers and reducing the Bullwhip Effect.

While the above-mentioned forecast horizons pertain to particular business processes, this research applied the technical approach to a longer forecasting horizon of 12 weeks. The purpose of the experiment was conducting a "Quarterly Stress Test" for the models. As a matter of fact, for a retailer, a period of 12 weeks corresponds to a full seasonal quarter period when financial budgeting and inventory management occur in practice.

Statistically speaking, testing models on a 12-week horizon constitutes a challenging forecasting stability test. To put it simply, the variance in the error distribution tends to increase as the distance in time from the latest observed point increases—a phenomenon called error propagation. Through generating the 12-week forecast in the experimentation phase, the study aims at assessing how well the learned knowledge of each particular model retains its validity once the horizon moves beyond the 4-week operational period. However, unlike linear models (SARIMA and ETS), the machine

learning models (Random Forest and XGBoost) have more chances to maintain forecast stability based on the use of higher lags.

### 3.7.4 Data Partitioning and Validation Strategy

Given that the data set is a time series of transactions, it is important that the data be maintained in chronological order during the training and validation of the models. To maintain the integrity of the evaluation process, to avoid over-optimistic results due to data leakage, a chronological split was used. The data was split at a particular cutoff point, with the first 90 weeks used for model training and tuning, and the most recent 12 weeks reserved for a 'Hold-out' test set. This is to mimic a real-world scenario, where the model needs to forecast future demand using only past data.

**The 90/12 Temporal Split**

Since the entire dataset contains approximately 105 weeks, the following specific time interval was selected to ensure sufficient depth of history during the training phase and at the same time, a sufficiently powerful period for out-of-sample testing:

- **Training Set (Weeks 1–90):** The initial period of 90 weeks from January 3, 2010 (week ending on January 3, Sunday) to September 18, 2011 (week ending on September 18, 2011) was used as a training period. This allowed us to train models using demand and seasonality patterns of holidays in 2010.

- **Test Set (Weeks 91–102):** Next 12 weeks from September 25, 2011 to December 11, 2011 were taken as a "Hold-out" testing sample. In fact, the selection of this testing period was quite difficult since it includes year-end “Peak Season”.

One of the technical aspects involved in the data splitting process is handling lagged variables. To have each model, especially the Machine Learning algorithms, work on a fully featured dataset since day one of training, a “buffer” strategy had been adopted.

The research will focus on 102 weeks of modeling period; however, for the initial weeks of 2009 raw data set, we used them in order to create features like 4-week moving average (rolling_mean_4) and Lag-1 (lag_1). As the calculation of the 4-week rolling mean needs four prior periods to obtain the first value, the initial period in the year 2009 was used as the initialization period. By pushing the starting point of the modeling period from December 1st, 2009 to 4 weeks forward, the research would ensure that all the features are populated in the model without any missing values (NaNs). This is why the period has shifted from 105 weeks to 102 weeks.

**Validation Strategy**

The validation strategy that was used in this study is: **Fixed-origin-hold-out** validation. In the context of the Fixed-Origin approach, prediction models were trained using the data in Weeks 90. The last 12 weeks of the data set remained unseen to the models throughout the training process and then the 12-week forecast was made based on Week 90. The hold-out period worked to provide an unbiased standard against which the models could be assessed in terms of their accuracy levels. Through avoiding the danger of overfitting and testing the algorithms using out-of-sample data only, the study presents the best possible performance of each algorithm in the real world.

### 3.7.5 Model Evaluation Metrics

The performance of each forecasting model was evaluated using three standard accuracy metrics commonly applied in demand forecasting studies. These metrics give different insights into the forecasting error and are well-established in the literature

related to retail forecasting. These include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Weighted Mean Absolute Percentage Error (WMAPE). Using multiple metrics provides a comprehensive assessment of model performance, as each measure captures different aspects of forecasting errors (Hyndman & Athanasopoulos, 2021).

**Mean Absolute Error (MAE)**

MAE measures the average magnitude of the forecasting error. It is easily understandable and is helpful in comparing the absolute accuracy of the models. MAE can be used for **workforce and logistics planning**.

**Root Mean Squared Error (RMSE)**

RMSE penalizes the models more if there are larger errors. This is helpful in identifying models that are likely to make occasional extreme forecasts, which is important in the context of the problem where larger errors are more costly. It is used to calculate **Safety Stock**.

**Weighted Mean Absolute Percentage Error (WMAPE)**

WMAPE is a percentage error measure that is scale-independent and is relevant to the context of the problem. This is helpful in avoiding the divide-by-zero problems that are common with the Mean Absolute Percentage Error, which is more commonly used in the context of demand forecasting. WMAPE is particularly relevant to the context of the Online Retail II dataset, where there are many intermittent and low-demand items. It is used for **Performance Benchmarking and KPI's**.

In addition to predictive accuracy, the study examines model robustness and stability across different forecasting horizons. The results allow identification of which methods perform best under varying demand patterns and operational contexts.

## 3.8 Linking Forecasting Accuracy to Operational Planning

The value of forecasting accuracy can only be seen by how it can be applied in operational decision making, even though the ultimate aim of this study is to test forecasting algorithms to see how accurate they are. According to Christopher (2016) the accuracy of the demand signal is the basic factor for the efficiency of the supply chain. It provides the framework to analyse the benefits of the increased accuracy of machine learning models in the real world of industry.

There is a critical relationship between forecast errors (RMSE) and the calculation of safety stock with regards to optimization of safety stock and capital efficiency. Based on the inventory logic as developed by Heizer et al., (2020) safety stock (SS) can be a function of the uncertainty associated with the demand during the lead time. In the mathematical sense, the buffer needed is directly proportional to the standard deviation of forecast error, namely:

$$\textbf{\textit{Safety Stock (SS)}} = Z \times \sigma d \times \sqrt{L}, \tag{10}$$

where:

***Z:*** The service level factor (e.g., 1.65 for a 95% service level),

***σd:*** The standard deviation of forecast error (RMSE),

***L:*** The lead time (number of days to receive an order).

The organization can then mathematically support decreasing the safety stock levels due to the decrease in error metrics (RMSE and WMAPE). This enables the retailer to keep

service level (e.g. 95% availability of products) unchanged with a substantial reduction in working capital locked up in inventory.

In e-commerce, minor inaccuracies in the downstream demand forecast for retail orders create a significant downstream impact on orders in the upstream business, which is called the Bullwhip Effect (Lee et al.,1997). Advanced models such as Random Forest or XGBoost are used so that they provide better information of non-linear spikes and promotion effects, resulting in reduced information distortion passed to suppliers. Better predictions at retail node add clarity to the chain, minimizing the need to have excess inventory at suppliers and to make emergency deliveries.

Resource Allocation and Workforce Planning are also based on forecasting accuracy. Labor scheduling to manage the growing number of orders and customer service requests during the "Peak Seasons" defined in this study is important for retail operations. Under-forecasting causes a shortage of labour force, causing delays of delivery and drop in service quality. Labor costs and wasted operational overhead occur when there is over-forecasting. The use of forecasting models that give very high speed and real-time forecasts helps management to synchronize capacity of the workforce with the expected consumer flow.

The use of WMAPE goes beyond being just a method of measuring errors; it serves as an important tool in Performance Benchmarking and in the creation of Key Performance Indicators (KPIs). Since it is not dependent on any scale, it helps in carrying out performance evaluations in different inventory classes and allocating important resources where forecast accuracy varies the most."

Finally, the findings of this empirical research can support the development of a strong Decision Support System (DSS. This research gives a guideline for "Model Selection" in strategic planning, since the different models work best in different situations, such as the intermittent demand versus high-value seasonality. Accurate forecasts help to better

negotiate with suppliers about volume discounts and serve as the basis for long-term market growth strategies.

Thus, the connection of the results of the forecast to operations is expected to offer insights that transcend accuracy and are significant in operations management in retail settings.

# 4 Results and Discussion

The chapter shows the research findings through empirical evidence from the comparative demand forecasting study. The analysis aimed to test how well traditional statistical forecasting models and machine learning methods performed when evaluated with identical dataset and validation method and accuracy evaluation system. The test period consists of twelve weeks (weeks 91 through 102), representing the peak season of the retail calendar, characterized by very volatile demand. The main aim is to identify the model framework that provides strong support for replenishment decisions. The study results evaluate numerical performance but also examine how forecast accuracy and error patterns impact retail decision-making through operational planning assessment.

## 4.1 Visual Comparison of Forecasts

Figure 10 displays the actual weekly demand data alongside the forecasted results from models tested during the twelve-week evaluation period. The figure demonstrates how various forecasting methods react to seasonal changes and ongoing trends and fluctuating demand patterns.

As seen in Figure 10, there is a visible difference between the two groups of models. Although the statistical measures (Naïve, ETS) do not react at all to the trend increase, the machine learning methods (Random Forest, Neural Networks) adequately reproduce the momentum of the real sales figures. This observation justifies the shift towards data-driven modeling.

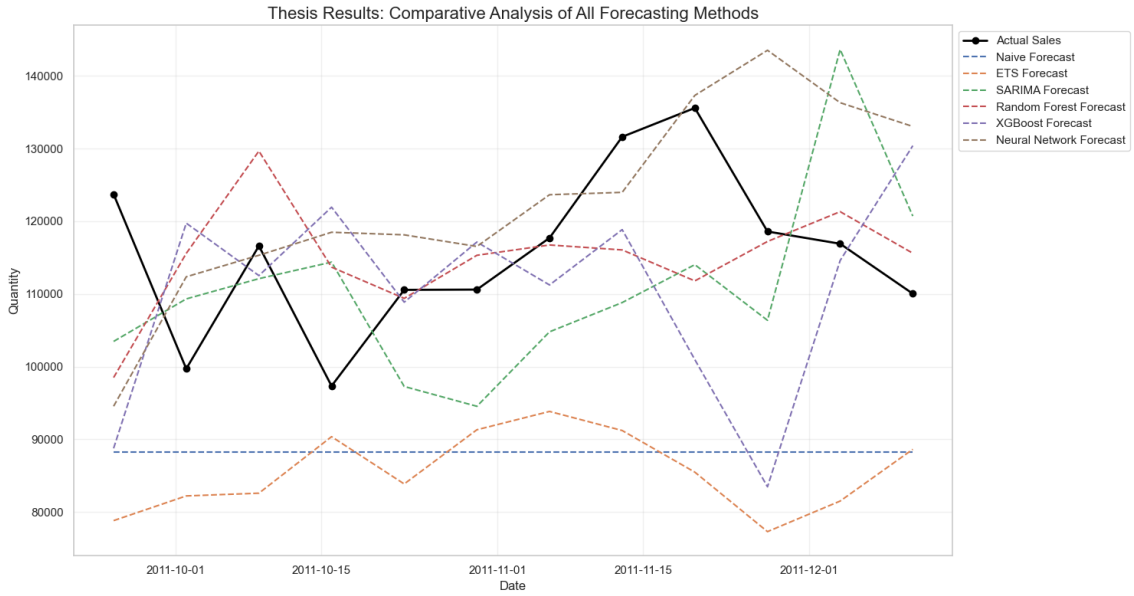


**Figure 10.** Actual vs Model Forecasting Comparison.

## 4.2 Performance Analysis of Individual Models

Even though the aggregated outcomes of all forecasting models taken for this study can offer a comparison perspective, a closer look at how each model works is needed to understand how different architectures manage the Online Retail II dataset in detail. Demand forecast on an industrial level is more than just the problem of optimizing for an objective function; it involves considering how a model behaves when introduced with trends, seasonality, and random variations.

### 4.2.1 Random Forest

The Random Forest method of all tested models showed the highest ability to predict actual demand patterns from Figure 11. The forecast maintains close accuracy to the peak season trend which follows an upward path, and they successfully capture general level shifts without exhibiting excessive short-term variations. The ensemble structure

provides effective noise reduction while maintaining its ability to detect seasonal patterns that occur at regular intervals.

The model performs extreme value forecasting at a conservative level. The Random Forest system underestimates peak demand during sharp demand spikes because it combines decision trees to achieve stability instead of accurately modeling rare extreme events. The system produces operational benefits because it reduces overstock risk but it needs proper safety stock management during periods of high customer demand.

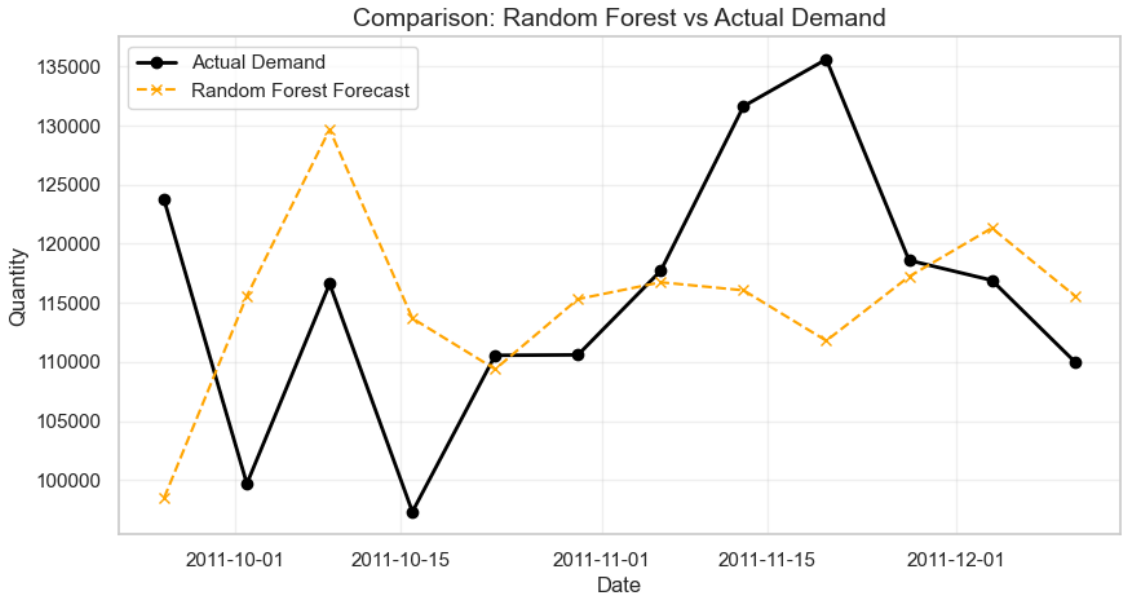


**Figure 11.** Actual vs Random Forest Model Forecasting C**omparison.**

### 4.2.2 XGBoost and Neural Network Models

The XGBoost and neural network models show enhanced ability to respond to immediate changes in demand patterns as per figure 13 and 14 respectively. The models demonstrate accurate change direction tracking while they quickly adapt to recent shifts because of their strong reliance on previous data elements. The models demonstrate increased volatility because they both experience rapid fluctuations in their behaviour. The forecasts for multiple weeks either exceed or fall short of actual demand, especially

during sudden market changes. The forecast variance increases because of this sensitivity which results in higher RMSE values when compared to Random Forest. The process of operational planning will lead to unstable replenishment choices because forecasts will be used without any smoothing methods implemented.

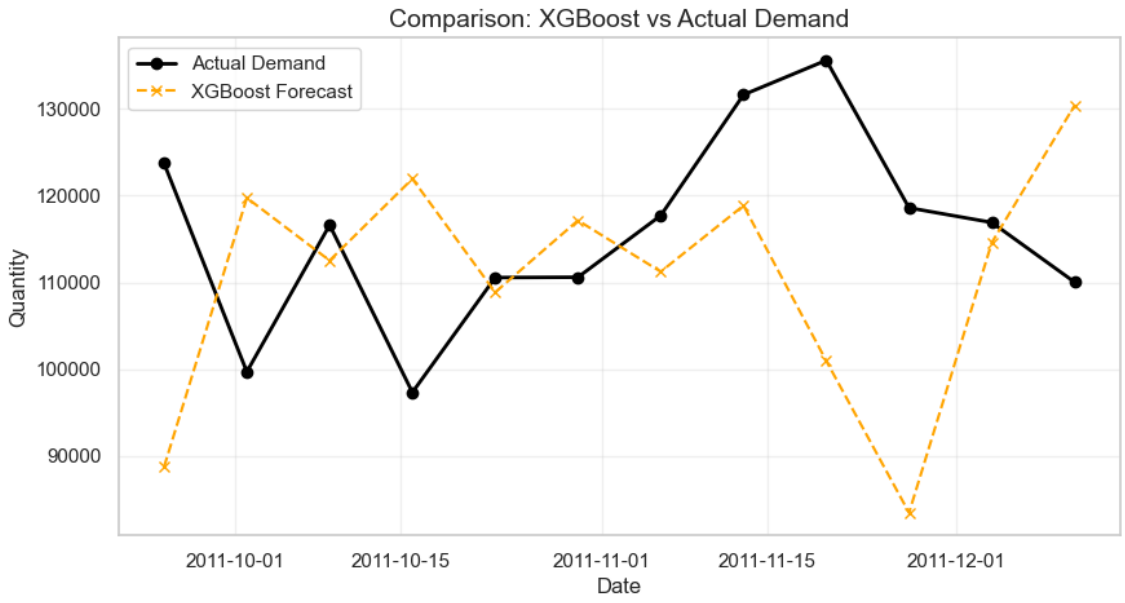


**Figure 12.** XGBoost vs Actual Demand Comparison.

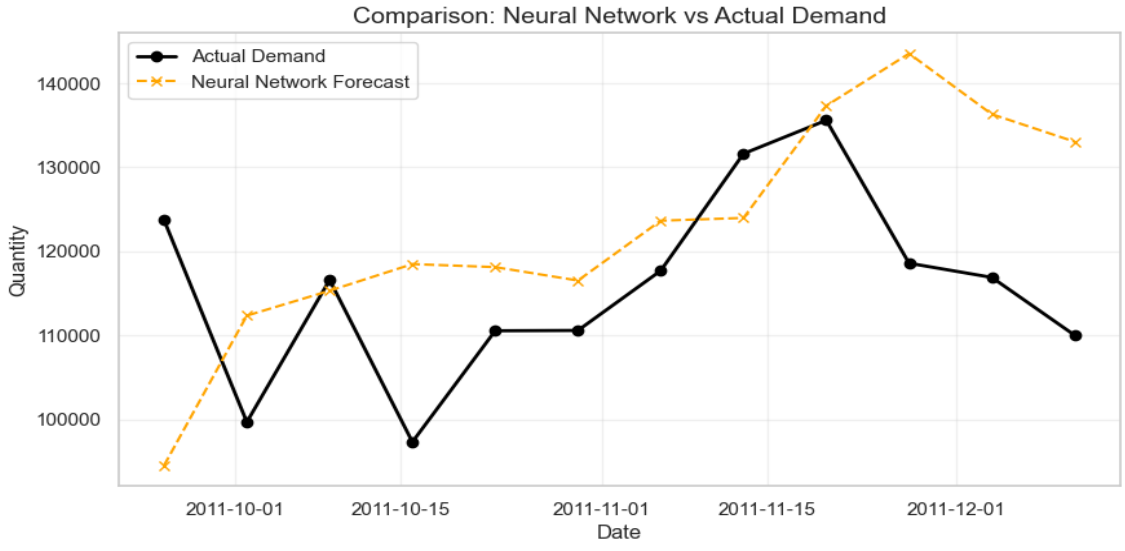


**Figure 13.** Neural Network vs Actual Demand Comparis**on.**

### 4.2.3 SARIMA

The SARIMA model successfully captures the recurring seasonal structure of weekly demand and reflects the underlying trend pattern reasonably well as seen in Figure 14. The model performs satisfactorily during stable periods with regular seasonality.

However, SARIMA exhibits limited adaptability to non-linear changes. Forecasts tend to lag behind sudden demand surges, especially during holiday-driven peaks. This delay is explained by the linear structure of the model, which relies primarily on past average behaviour rather than contextual or calendar-based signals.

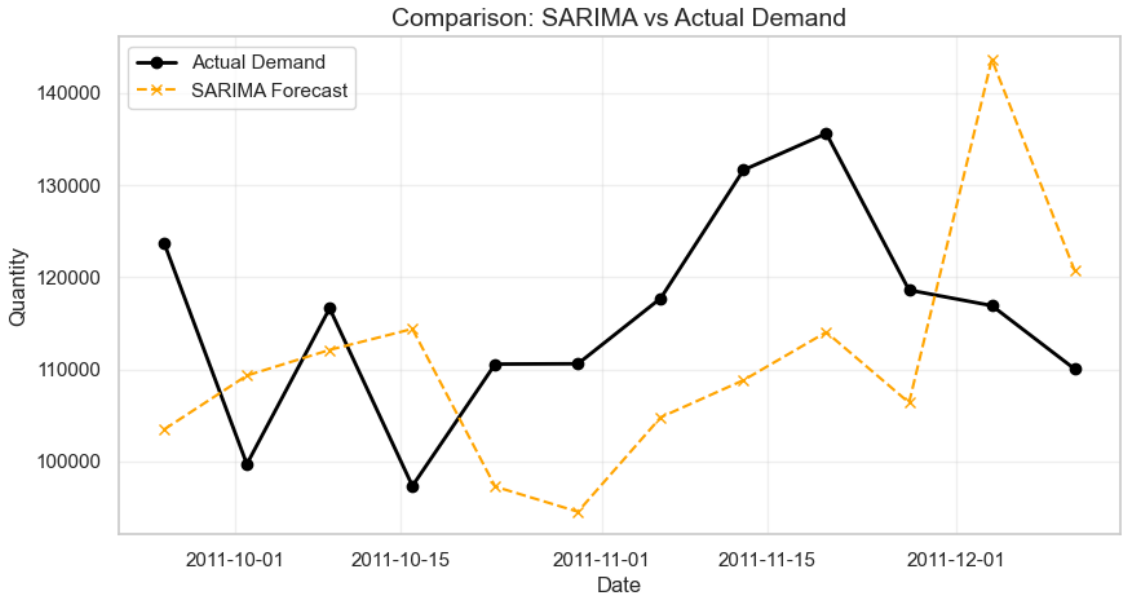


**Figure 14.** SARIMA vs Actual Demand.

### 4.2.4 Naive and Exponential Smoothing Models

The visual assessment in Figure 15 and Figure 16 shows that both the naive benchmark and the exponential smoothing models exhibit their worst performance. The naive forecast creates an unchanging path which depends only on the latest data point, and it fails to predict seasonal growth. Exponential smoothing improves upon this by capturing

trend components but still underestimates the magnitude of peak demand periods. The results show that basic time-series methods provide simple implementation and interpretation, but these methods fail to deliver accurate forecasts for retail environments which experience both strong seasonal patterns and unpredictable demand changes. These models are found to be high-bias (underfitting) models that are too rigid to capture the momentum of real sales figures.

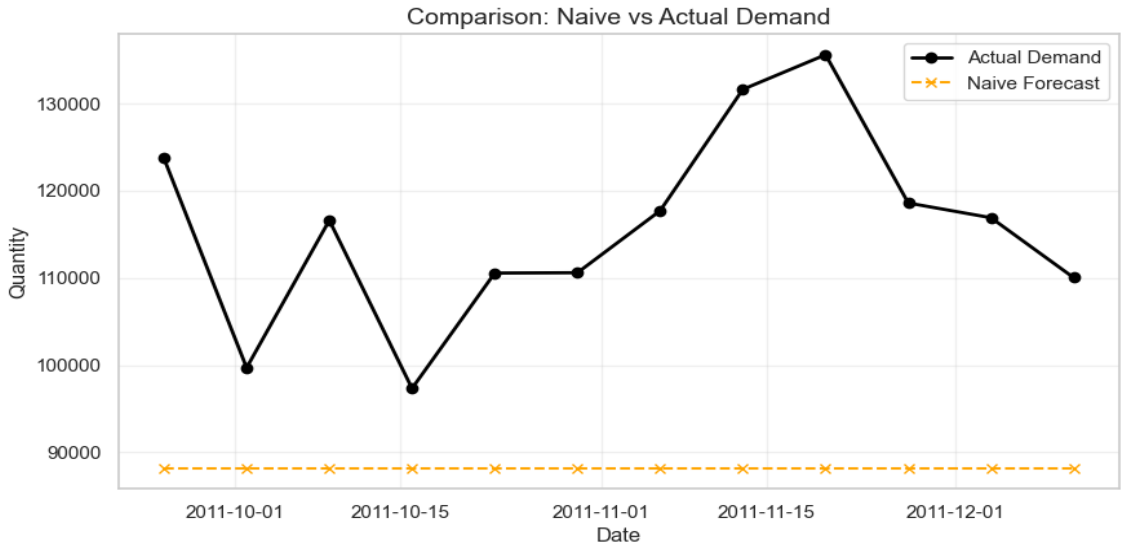


**Figure 15.** Naive vs Actual Demand.

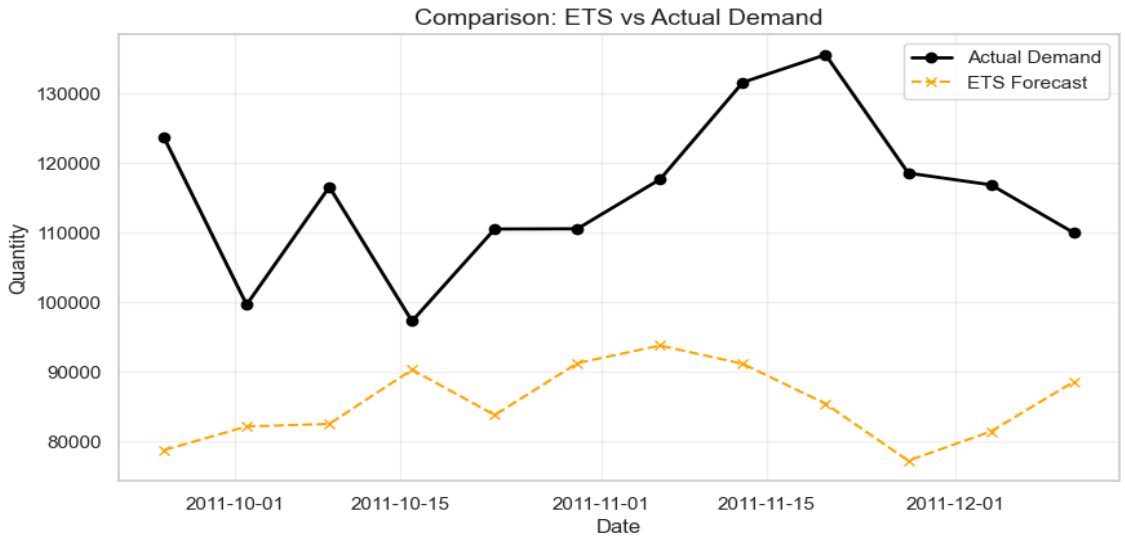


**Figure 16.** ETS vs Actual Demand.

## 4.3 Quantitative Model Performance Evaluation

The main aim of this quantitative analysis is developing a mathematically sound foundation for model selection. Since demand behavior in complex retail environments is characterized by a non-linear relationship and can contain outliers, the use of a single parameter is not sufficient to reliably measure the effectiveness of models (Hyndman & Athanasopoulos, 2021). Consequently, for this research, all forecasting models have been analyzed based on their forecasting capabilities through three different measures: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Weighted Mean Absolute Percentage Error (WMAPE).

By using these metrics, the assessment goes beyond just evaluating accuracy to focus on three key factors: the average value of the error (MAE), high-volatility outliers and "tail risks" (RMSE), and relative importance of the errors from the business perspective (WMAPE). Such a methodology guarantees that the most accurate model will also be the most stable and reliable one for real-world supply chain decision making (Fildes et al., 2022).

### 4.3.1 Statistical Metrics Summary

Figure 17 shows the value of statistical error metrics generated such as MAE, RMSE, and WMAPE as a result of model implementation for all selected statistical and machine learning methods. The three metrics establish different ways to measure accuracy which also show how the system behaves with major faults and how it performs regardless of measurement size. MAE and WMAPE indicate overall accuracy, RMSE is used as a "risk measure" showing the behavior of models when exposed to extreme data points. Through this combination of the findings, this research will reveal what models can be applied to retail industry.

--- FINAL PERFORMANCE METRICS ---

| | Model | MAE | RMSE | WMAPE (%) |
|---|---|---|---|---|
| 3 | Random Forest | 10676.66 | 116.53 | 9.22 |
| 5 | Neural Network | 13374.79 | 127.59 | 11.55 |
| 2 | SARIMA | 15645.28 | 129.55 | 13.52 |
| 4 | XGBoost | 16970.90 | 145.33 | 14.66 |
| 0 | Naive | 27563.00 | 172.08 | 23.81 |
| 1 | ETS | 30169.15 | 180.56 | 26.06 |

**Figure 17.** Final performance metrics of different models' results table.

**Detailed Analysis of Results**

**1. Weighted Mean Absolute Percentage Error (WMAPE)**

This is the most important "Management Metric". As seen in Figure 17, only Random Forest has WMAPE under 10% (9.22%), which usually serves as the industry's standard for "Excellent" forecasting for high-volatility retail operations.

Neural Network (11.55%) and SARIMA (13.52%) also demonstrate excellent results, keeping the forecast error well within acceptable parameters.

The difference between Random Forest (9.22%) and ETS (26.06%) models is almost three-fold, meaning that, in terms of this dataset, even simple exponential smoothing models have poor mathematical justification.

**2. Mean Absolute Error (MAE)**

MAE shows the average error measured in the same units of demand. Random Forest fails by an average of 10,676 units. Naive fails by 27,563 units.

With the use of the Naive method, the management will under/overestimate its weekly inventory requirements by more than 27,000 items. Using Random Forest helps minimize this "planning loss" by more than 61%.

**3. Root Mean Squared Error (RMSE)**

RMSE is referred to as "Risk Metric" since the error is squared before taking the mean, and hence, is sensitive to large mistakes. Random Forest has the most stable results with an RMSE of 116.53.

Despite having decent value of WMAPE (14.66%), the RMSE of XGBoost is notably high (145.33) compared to Neural Network (127.59) and SARIMA (129.55).

It implies that although the model captures trends accurately, it makes some forecasts that may not reflect the demand very well, i.e., the error might be high. In a supply chain operation, low RMSE would be better as it implies consistency and reduces the risk of predicting an abnormal value, which could cause a huge shortage or excess inventory. Hence, when working within a supply chain domain, the safer option would be Random Forest.

**Summary**

From the quantitative results obtained, it is evident that the Machine Learning algorithms, in particular, the Random Forest and Neural Network, outperform traditional

techniques. From the consistency observed in the three metrics (MAE, RMSE, and WMAPE), it is justified to state that Random Forest is the best fit for this retail dataset.

### 4.3.2 Analysis of Model Performance Tiers

Comparative analysis classified the six forecasting models implemented into three performance classes which are the baseline models, conventional statistical models, and machine learning algorithms. This classification illustrates the growing complexity needed for high volatility of demand in retail operations.

**1. Baseline Performance**

The naive model serves as a minimal benchmark. The historical carry-forward methods show their inability to capture complex retail demand patterns because their error rates exceed acceptable limits. The advanced model needs to exceed this benchmark through its performance in providing real-world benefits.

**2. Statistical Models**

The naive baseline performance shows significant improvements through the use of exponential smoothing and SARIMA. The SARIMA model demonstrates better performance than exponential smoothing because it can represent both autoregressive and seasonal patterns at medium forecast time frames. The statistical models demonstrate rising prediction errors when their forecast time periods are extended. The modeling system cannot accurately represent changing demand patterns because it depends on linearity and constant seasonal patterns.

**3. Machine Learning Models**

Machine learning models surpass traditional statistical methods in all evaluation accuracy metrics. The following list includes the three models that achieve their performance with the following results: Random Forest achieves the lowest overall MAE and WMAPE which shows its ability to maintain accurate performance across all demand conditions. XGBoost delivers competitive performance but exhibits larger RMSE values which show its tendency to respond to extreme forecast deviations. Neural Network models show learning ability, but they need additional tuning work, and they produce unstable performance results during short-term prediction periods.

## 4.4 Comparative Assessment

The comparative results demonstrate that no model achieves optimal performance across all testing conditions yet specific usage patterns between models show clear superiority. Statistical models from the traditional approach need to maintain their usefulness operational capacity when businesses experience constant customer traffic and predictable seasonal patterns. The planning process requires transparent models which need minimal computational resources to operate within environments that experience predictable demand patterns and low operational changes.

Machine learning models deliver better results through their Random Forest system which handles complex retail environments that experience seasonal demand patterns and trend changes and short-term demand fluctuations. The steady superiority of Random Forest compared to statistical models and also single ML algorithms such as the MLP algorithm indicates that "Ensemble Learning" is exclusively tailored for e-commerce datasets. Through creating multiple decision trees, Random Forest manages to mitigate any transactional noise in the Online Retail II dataset. On the other hand, XGBoost tends

to concentrate on correcting previous mistakes, thus overfitting the noise, as is evident from the increased RMSE.

Traditional approaches like SARIMA and ETS were successful in detecting a consistent seasonal pattern; however, they could not be used for prediction during the peak season due to the lack of "contextual awareness". Machine learning algorithms benefited from calendar-based features such as Month and Week-of-Year as well as demand lags to predict the volume of the holiday peak before it occurred.

An important thing to learn from this comparison is that accuracy and stability are different concepts. Although the Neural Networks were found to be very accurate, they did not perform as well as the Random Forest algorithm in terms of stability of its performance. For retail businesses with a large resource management requirement, the Random Forest system has high operational efficiency. In terms of business, stability of errors may be optimum instead of optimum accuracy because stable error patterns can be easily managed by safety stocks.

## 4.5 Operational Implications

The operational planning process is enhanced by better forecasts as it generates measurable operational gains. Minimizing forecast errors allows companies to keep a reduced level of safety stocks, thus allowing them to reduce the stock-out cases and setting up regular restocking procedures. The demand estimates which exceed the actual demands add to storage costs, and the demand estimates which are delayed during peak periods add to the operational risk. The results show that the appropriate development and testing of machine learning models have the potential to markedly improve retail planning performance. The ensemble tree-based methods are the most reliable decision-making support methods for real retail demand scenarios within the tested methods.

### 4.5.1 Inventory Optimization and Safety Stock

The first operational benefit is the reduced safety stocks. It can be explained using the well-known inventory management formula: **Safety Stock (SS) = Z × σd × √L.** The value of the "uncertainty parameter" (σd) in the equation is depends on the forecasting error.

**Cost Effectiveness:** As a result of lowering the RMSE and WMAPE by more than 60% compared to the benchmark, the company can prove mathematically that there should be a considerable decrease in safety stock values while keeping the same service level. Consequently, the cash flow locked in additional inventory will be released.

**Waste Prevention:** The MAE decrease by 16,887 units per week on average (Naïve Baseline – Random Forest = 27,563 – 10,676) means lower chances of creating surplus inventories, especially for seasonally sold products that depreciate after December.

### 4.5.2 Strategic Decision Support System (DSS)

The results offer guidelines on how to develop the automated DSS.

**Tiered Modeling Strategy:** The research suggests the use of the "Hybrid Approach" for retail planning. Ensemble machine learning algorithms should be applied to items with high volume and value when the WMAPE equal to 9.22% returns the highest Return of Investment (ROI). SARIMA model can still be employed for items with stable and low-volume behavior, since it is computationally simpler.

**Risk-Adjusted Planning**: Since the Random Forest regression has been shown to have lower RMSE compared to the other models, it can become a solid "baseline" for negotiating with suppliers about ordering the quantities with confidence. It will reduce the risks of emergency orders for the company.

### 4.5.3 The Use of WMAPE for Performance Management

WMAPE can become a key performance indicator (KPI) for management within the organization. Since the Random Forest model has scored an excellent result (WMAPE<10%), it can now serve as a new benchmark of performance. Such approach opens doors for benchmarking by other organizational units.

# 5 Conclusion and Future Research

This final chapter compiles the research findings and theories that have been discussed throughout this study to deliver a conclusive judgment on the effectiveness of statistical and machine learning algorithms in retail forecast prediction. First, an overview of the research process and its achievement rate towards the proposed research aims are provided. This chapter then moves on to the impact of the findings on practice, paying special attention to how the improved prediction precision leads to efficient inventory management and decision-making strategies. Lastly, the limitations in the present study are highlighted, along with the gap analysis for future research potential.

## 5.1 Summary of the Study

The thesis aimed to conduct a detailed analysis which compared traditional statistical forecasting methods with modern machine learning techniques to assess their usefulness for retail demand forecasting operational planning. The study was motivated by the increasing complexity of retail demand which occurs because of seasonal patterns and market fluctuations and changes in consumer buying patterns together with the availability of extensive retail transaction data. The researchers created multiple forecasting models within a unified analytical system to address the research question, and they evaluated these models using a publicly accessible retail dataset. The research study assessed traditional statistical methods which include naive forecasting and exponential smoothing and SARIMA together with Random Forest and XGBoost and a neural network model. The researchers assessed the models using established forecasting accuracy evaluation methods which they tested over short-term forecasting periods that matched actual retail planning processes. The research study evaluated predictive accuracy through its direct impact on inventory management and replenishment scheduling and stock-out reduction in operational planning decisions.

The research study aims to provide academic research results which will help businesses through its combination of analytical assessment and real-world business requirements.

## 5.2 Key Findings

The empirical analysis showed that machine learning models outperformed the traditional statistical models as machine learning models were able to forecast more accurately when it came to demand of systems during the seasonal peak periods as they are able to deal with complex demand patterns. The Random Forest model showed superior overall performance as it had the least forecasting errors that were measured based on various accuracy criteria and forecasting timelines. The traditional statistical models SARIMA and exponential smoothing showed good performance when they were required to identify permanent seasonal patterns and to measure long time trends. The model encountered issues when handling unexpected peak demands with no clear trend and during holidays. The machine learning models achieved better success rates, as they were able to incorporate calendar-based data and historical demand data, allowing them to adapt to fluctuations in customer needs. The research demonstrated that accurate forecasting increases operational planning efficiency. The study found that retailers achieve better forecasting accuracy through actual store operations which enables them to decrease safety stock levels while protecting against stock shortages during peak demand times and developing better restocking processes. The study discovered a major real-world implementation problem because machine learning models provide better accuracy than traditional statistical forecasting methods but require more effort to use and understand.

## 5.3 Contributions of the Study

The results obtained from the analysis performed during the comparative study will be highly valuable not only from the academic point of view but also for practical application in a supply chain. First, the aim of the present paper was to adapt the theoretical knowledge of data science that originated from the machine learning side, as well as from the statistical modeling approach. Thus, different advanced machine learning pipelines were compared with traditional statistical models, and a vast body of transactional data was analyzed. There are two main types of contributions to the present research: on the one hand, this work contributes methodologically to the quantitative time-series approach; on the other hand, it also gives some valuable empirical insights regarding improving the efficiency of retail ecosystems.

### 5.3.1 Methodological contribution

An important methodological contribution of this research is the creation of a benchmarking scheme for the assessment of various modeling approaches which is highly transparent and entirely replicable, employing the same validation parameters. In academic studies, these comparisons can differ based on several factors like the chosen statistical model, the selected machine learning algorithm, the used training data, the pre-processing scheme and the employed assessment metric. These differences have been accounted for in this study, by ensuring that all models undergo the same pre-processing procedure and follow the same train/test split scheme, which is consistent across all models tested, at a ratio of 90/12.

In addition, this research incorporates WMAPE as the key performance indicator, which along with the traditional error metrics such as MAE and RMSE gives a methodology that scales the errors according to the number of transactions. This overcomes a common problem in time-series engineering of having a low volume of items and getting

deceptive percentage errors. The whole engineering process, feature transformation, and model implementation and comparison processes are stored in a public version-controlled repository, which is an open standard that can be used in future empirical verification in retail analytics.

### 5.3.2 Empirical contribution

The research demonstrates that ensemble machine learning methods, especially the Random Forest model, provide better forecasting accuracy in retail stores which experience both seasonal demand patterns and unstable customer demand. The results show that machine learning models can better understand complex relationships in retail sales data than traditional statistical techniques.

The study shows that accurate demand forecasting enables retail operations to function correctly. The research demonstrates that operational planning results will improve when forecasts become more accurate which leads to better inventory management and efficient restocking processes and decreases the chances of stock shortages and excess product inventory. The advanced forecasting methods provide managerial benefits which extend beyond their statistical performance capabilities according to the broader perspective.

## 5.4 Theoretical Reflection

The results of the empirical evaluation of the comparative study conducted offer crucial benchmarks to measure the practical applicability of classical inventory models and current data-driven solutions. In comparison to the existing theories from operations management literature, the behavior exhibited by the SARIMA and ETS models in case of sudden spikes in demand fits well within the constraints of linear time-series models,

as described by Box et al. (2015). The reason being that the nature of linear autoregressive and moving average models is determined by the fixed parameters obtained from the previous historical realizations, making it difficult for them to capture abrupt and non-linear fluctuations in the market (Hyndman & Athanasopoulos, 2021). For this reason, when a structural change or shift occurs, or the trends change, the linear time-series models will not be able to deal with the multi-dimensional nature of issues being discussed in the supply chain and will not provide us with visibility in the supply chain during the peak time of consumer demand (Lee et al., 1997).

This question concerning the dominance of one approach over another has come a long way since then. Previous research, like Makridakis et al. (2018 and 2020), has indicated that conventional statistical techniques have an edge over machine learning in both predictive performance and computational efficiency. This study, in contrast to them, finds that its outcomes resemble those discovered in the M5 Competition that the best performing algorithms used today, in particular during modern retail competitions, rely on machine learning structures (Makridakis et al., 2022). This can be attributed to the granularity of the datasets used, since Makridakis relied on a monthly sample, whereas the analysis presented herein benefits from the application of ensemble techniques such as Random Forest, which are capable of identifying the non-linear dependencies that cannot be recognized by linear models like SARIMA. Therefore, the 'computational complexity' highlighted by Makridakis should be considered a compromise."

Furthermore, “responsiveness vs. efficiency” of the supply chain is the well-known dilemma that has been pointed out by Chopra & Meindl (2013) and is realistically resolved by Random Forest's accuracy and stability of predictions. In a very volatile retail market, there is an expected amplification of minor shifts in the signals of demand further upstream, from retail to wholesale and wholesale to manufacturer, through the retailers' downstream terminals. This effect is called the Bullwhip Effect (Lee et al., 1997; Christopher, 2016) and typically would be enhanced by local inventory planners using high variance forecasting models.

This shows that through the production of a robust prediction signal with less than a 10% WMAPE during the peak period when volatility was high, the newly constructed Random Forest architecture can indeed manage to eliminate this upstream disturbance factor. By reducing the amount of localized variability occurring at the very first transaction node in the forecasting algorithm, one eliminates the need for artificially increasing the amount of safety stocks along the entire network. From an analytical perspective, this design approach confirms the fundamental concepts discussed in machine learning models (Kuhn & Johnson, 2019).

The findings presented above serve to provide empirical evidence for digital concepts in modern supply chains by showing how switching from a linear historical method to machine learning leads to optimized processes and proactiveness in inventory management.

## 5.5 Limitations

The research provides important results, but multiple limitations exist which need acknowledgement because these limitations will reduce the research findings' ability to apply in different situations.

The study conducted its empirical analysis through the Online Retail II dataset which contains all transaction records of a UK-based giftware retailer. The dataset shows typical retail demand patterns which include seasonal changes and unpredictable demand and times of highest demand. The specific product characteristics and business operations of the company restrict the research results to particular retail markets which do not include grocery and fashion and durable goods retailing. Industry demand patterns show significant differences because different product categories lead to various buying patterns and marketing techniques.

The research examined a few of the most common statistical techniques and machine learning forecasting techniques, but it does not study more advanced techniques such as deep learning LSTM networks and a combination of statistical and machine learning forecasting techniques. The models perform much better when it comes to predicting future events as they are able to manage longer relationship patterns and their ability to deal with complex demand scenarios. These models were excluded since the researchers wanted to maintain the simplicity of the research methodology that could be easily carried out and would yield easily understandable research results.

The study assessed forecasting accuracy through three statistical accuracy measures which included MAE and RMSE and WMAPE. Although these measures are commonly used in forecasting research, the study did not include a direct quantitative analysis of operational or financial outcomes such as inventory holding costs, service-level improvements, or stock-out costs. The study presents operational advantages of forecasting accuracy improvements through theoretical discussions instead of using precise optimization techniques or financial assessment models.

Machine learning models achieved better prediction results through their higher predictive accuracy but their use in actual retail environments remains difficult to implement. The adoption of these systems in actual operations faces obstacles from their limited interpretability and governance requirements and their need for high-quality data and maintenance support and system compatibility with current planning frameworks. The study failed to acknowledge the organizational and technical obstacles which hold critical importance for the successful application of advanced forecasting models in real-world settings.

## 5.6 Directions for Future Research

The research offers an organized evaluation which looks at different statistical methods as well as machine learning techniques to predict retail demand, but it also leaves a lot of research possibilities for researchers to build on existing research.

First, future studies could explore the use of more advanced forecasting architectures, particularly deep learning and hybrid models. The techniques which include Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Temporal Fusion Transformer (TFT) and hybrid statistical–machine learning frameworks enable better detection of complex nonlinear demand patterns together with extended temporal dependencies which exceed the model limitations present in this investigation.

The present study depends mainly on the Online Retail II dataset which contains transactional sales records. Forecasting performance could be further enhanced by incorporating external and contextual variables such as promotional activities, pricing changes, weather conditions, holidays, macroeconomic indicators, and social media signals. The inclusion of exogenous factors would enable models to create more authentic retail demand patterns which they can use to handle demand drivers that exist outside their historical sales data.

The study primarily was based on short-term forecasting scenarios, based on aggregated weekly demand data. Multi-horizon forecasting could be explored in future studies, along with the performance of models in daily, weekly and monthly forecasting. Hierarchical forecasting approaches require testing to determine the retail demand for each level that includes product categories and customer segments, as well as geographical areas, to inform better decision making.

The research needs to explore how forecasting models affect actual business operations and their managerial outcomes which go beyond their statistical accuracy. The future

studies should focus on operational performance metrics to measure the benefits of improved forecasting accuracy across multiple operational areas including inventory holding costs and stock-out rates together with service levels, replenishment efficiency and complete supply chain responsiveness. The analysis will help establish greater linkage between forecasting research and retail operations management practice.

Another potential research direction is related to model interpretability and explainable artificial intelligence (XAI). When retailers are choosing which machine learning models to use, they are looking for models that give them improved predictive results, but they also need to put in place transparent systems that users need to trust and find easy to use. Further research might explore the use of XAI methods to improve the understanding of model output by managers, and the improved decision making in operation planning scenarios that can be achieved with such understanding.

The research should examine the way businesses are implementing their forecasting system as businesses must have accurate forecasts to get the expected business results. Organizations maximize the benefit of their advanced forecasting model by designing their processes and decision-making structures to use it. Future research should focus on the process organizations take to prepare their data, the organizational governance system and whether there are partnerships between organizational analytics and operations teams. The longitudinal case studies will discuss the journey of retail companies to move from traditional statistical to machine learning and hybrid forecasting systems. Research will improve upon the current quantitative research by providing insight into the ways that forecasting systems provide value to organizations when they are used in business operations.

## References


Bousqaoui, H., Slimani, I., & Achchab, S. (2021). Comparative analysis of short-term demand predicting models using ARIMA and deep learning. *International Journal of Electrical and Computer Engineering*, 11(4), 3319–3328. https://doi.org/10.11591/ijece.v11i4.pp3319-3328. Accessed 31 Oct. 2021.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley. https://doi.org/10.1111/jtsa.12194

Breiman, L. (2001). Random forests. *Machine Learning,* 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bull, L., Knocky, L., & John, A. (2024). *Adaptive machine learning algorithms: Improving efficiency through dynamic model adjustment.* ResearchGate. https://www.researchgate.net/publication/386453002

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, 184(3), 1140–1154. https://doi.org/10.1016/j.ejor.2006.12.004

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, D. (2019). *Online Retail II* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5CG6D.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

Chopra, M., Chopra, R., Reddy, R., & Chopra, S. (2023). Leveraging LSTM neural networks and ARIMA models for enhanced real-time sales forecasting in dynamic retail environments. (2025). *Journal of Artificial Intelligence and Machine Learning Research*. https://www.joaimlr.com/index.php/v1/article/view/35

Chopra, S., & Meindl, P. (2013). Supply chain management: Strategy, planning, and operation (5th ed.). Pearson. https://sfaaz.org/assets/documents/library/RECOMMENDED%20TEXTBOOKS/Chopra_Meindl_SCM.pdf.

Chowdhury, A., Paul, R., & Rozony, F. Z. (2025). A systematic review of demand forecasting models for retail e-commerce enhancing accuracy in inventory and delivery planning. *International Journal of Scientific Interdisciplinary Research*, *6*, 1–27. https://doi.org/10.63125/mbbfw637

Christopher, M. (2016). *Logistics & supply chain management* (5th ed.). Pearson Education.

Falatouri, T., Kazemi, S. M. R., & Zadeh, M. H. (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

Fatima, A., & Salam, M. A. (2026). A data-driven predictive framework for inventory optimization using context-augmented machine learning models. *arXiv*. https://arxiv.org/abs/2601.05033

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Diversified random forests using random subspaces. In E. Corchado, J. A. Lozano, H. Quintián, & H. Yin (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2014* (Vol. 8669, pp. 82–89). Springer, Cham. https://doi.org/10.1007/978-3-319-10840-7_11

Fildes, R., Kolassa, S., & Ma, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting, 38*(4), 1283-1318. https://doi.org/10.1016/j.ijforecast.2019.06.004

Forecasting demand in retail supply chain management: A comparison of deep learning and machine learning methodologies. (2024). In *Advances in Supply Chain Analytics*. Bentham Science Publishers. https://www.eurekaselect.com/chapter/26738

From ARIMA to LSTM: Evaluating traditional and AI-based models for accurate retail sales forecasting. (2025). *International Journal for Research in Applied Science*

*and Engineering Technology*. https://www.ijraset.com/research-paper/ai-based-models-for-accurate-retail-sales-forecasting

Ganguly, P., & Mukherjee, I. (2024). *Enhancing retail sales forecasting with optimized machine learning models*. arXiv. https://doi.org/10.48550/arXiv.2410.13773

Gardner, E. S. (2006). Exponential smoothing: The state of the art—Part II. *International Journal of Forecasting, 22(4),* 637–666. https://doi.org/10.1016/j.ijforecast.2006.03.005

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Heizer, J., Render, B., & Munson, C. (2020). *Operations management: Sustainability and supply chain management* (13th ed.). Pearson.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9(8),* 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). Melbourne, Australia. Retrieved from https://otexts.com/fpp3/

Hyndman, R.J. & Koehler, A.B. (2006) Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting, 22,* 679–688. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001

Jewel, R. M., Linkon, A. A., Shaima, M., Badruddowza, Md Shohail Uddin Sarker, Rumana Shahid, Norun Nabi, Md Nasir Uddin Rana, Md Ahnaf Shahriyar, Mehedi Hasan, & Md Jubayar Hossain. (2024). Comparative Analysis of Machine Learning Models for Accurate Retail Sales Demand Forecasting. *Journal of Computer Science and Technology Studies, 6(1),* 204–210. https://doi.org/10.32996/jcsts.2024.6.1.23

Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research*, *78*, 1–9. https://doi.org/10.1016/j.jbusres.2017.04.016

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315108230

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, *43*(4), 546–558. https://doi.org/10.1287/mnsc.43.4.546

Lindfors, A. (2021). Demand forecasting in retail: A comparison of time series analysis and machine learning models.

Makridakis, S., Petropoulos, F., & Spiliotis, E. (2022). The M5 competition: Conclusions. *International Journal of Forecasting*, *38*(4), 1346–1354. https://doi.org/10.1016/j.ijforecast.2022.04.006

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE, 13*(3). https://doi.org/10.1371/journal.pone.0194889

Mitchell, T. M. (1997). *Machine learning.* McGraw-Hill.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Mustapha, O., & Sithole, D. (2025). Forecasting retail sales using machine learning models. *American Journal of Statistics and Actuarial Sciences, 6,* 35–67. https://doi.org/10.47672/ajsas.2679

Nasseri, M., Falatouri, T., Brandtner, P., & Darbanian, F. (2023). Applying machine learning in retail demand prediction—A comparison of tree-based ensembles and long short-term memory-based deep learning. Applied Sciences, 13(19), 11112. https://doi.org/10.3390/app131911112

Niemelä, T. (2025). *Demand forecasting in the retail environment: A comparative study of LightGBM, XGBoost, and MLP models*.

Pagidoju, R. T. (2025). Optimizing LSTM neural networks for resource-constrained retail sales forecasting: A model compression study. In *Proceedings of the 5th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)* (pp. 166–170). IEEE. https://doi.org/10.1109/ICUIS67429.2025.11380599

Petropoulos, F., Akkermans, H., Aksin, O. Z., Ali, I., Babai, M. Z., Barbosa-Povoa, A., Bendoly, E., Blossey, G., Borreguero, S., Boute, R., Boylan, J. E., Brintrup, A., Burnard, K., Cano, A., Carter, C. R., Chae, B. K., Chao, G. H., Chen, C., Chen, L., ... Zheng, T. (2025). Operations & supply chain management: principles and practice. *International Journal of Production Research*, *64*(1), 1–184. https://doi.org/10.1080/00207543.2025.2555531

Reddypogu, V. B., & Prasad, D. P. U. (2025). Demand forecasting in e-commerce fashion retail: A comparative study of generative AI, LSTM and ARIMA models. *Journal of Information Systems Engineering and Management, 10*(18s), 23–28. https://doi.org/10.52783/jisem.v10i18s.2876

RELEX Solutions. (n.d.). Living retail platform: AI-driven retail planning. https://www.relexsolutions.com/

Schmid, L., Roidl, M., Kirchheim, A., & Pauly, M. (2024). Comparing Statistical and Machine Learning Methods for Time Series Forecasting in Data-Driven Logistics—A Simulation Study. *Entropy, 27(1)*. https://doi.org/10.3390/e27010025

Sharda, R., Delen, D., & Turban, E. (2020). *Analytics, data science, & artificial intelligence: Systems for decision support* (11th ed.). Pearson.

Sharma, A., Verma, R., & Patel, S. (2021). Enhancing retail sales forecasting through LSTM networks and ARIMA models: A comparative analysis of AI methodologies. *European Advanced AI Journal, 10*(2). https://eaaij.com/index.php/eaaij/article/view/7

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2016). *Inventory and production management in supply chains* (4th ed.). CRC Press. https://doi.org/10.1201/9781315374406

Singh, M., & Dayama, A. (2025). Leveraging spatiotemporal graph neural networks for multi-store sales forecasting. *arXiv*. https://arxiv.org/abs/2511.19267

Suryawanshi, R., Musale, S., & Bhosale, S. (2024). Comparative analysis of use of machine learning algorithm for prediction of sales. *Journal of Electrical Systems*, *20*(3s), 851–863. https://doi.org/10.52783/jes.1383

Theodoridis, G., & Tsadiras, A. (2025). Retail demand forecasting: A comparative analysis of deep neural networks and the proposal of LSTMixer, a linear model extension. *Information, 16*(7), 596. https://doi.org/10.3390/info16070596

Upadhyay, H., Shekhar, S., Vidyarthi, A., Prakash, R., & Gowri, R. (2023). Sales prediction in the retail industry using machine learning: A case study of BigMart. *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, 1–6. https://doi.org/10.1109/ELEXCOM58812.2023.10370313

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics,* 34(2), 77–84. https://doi.org/10.1111/jbl.12010

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, *30*(4), 1030–1081. https://doi.org/10.1016/j.ijforecast.2014.08.008

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

Zhang, W., Liu, H., & Chen, Y. (2024). A study on promotional shipment forecasting for e-commerce merchants based on ARIMA time series and LSTM models. *Highlights in Science, Engineering and Technology, 98*, 280–286. https://doi.org/10.54097/yp1bjc76

Zhao, S., Li, X., & Wang, T. (2025). Optimization of deep learning models for dynamic market behaviour prediction. *arXiv*. https://arxiv.org/abs/2511.19090

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists.* O'Reilly Media.

# Appendices

## Appendix 1. Technical Implementation Snippets

### Data Cleaning

```
# 1. Remove missing customer IDs and Descriptions (Critical
for data quality)
df = df.dropna(subset=['Customer ID', 'Description'])

# 2. Remove Cancellations (Invoices starting with 'C')
df['Invoice'] = df['Invoice'].astype(str)
df = df[~df['Invoice'].str.startswith('C')]

# 3. Filter for valid sales only, removes negative prices
and quantities
df = df[(df['Quantity'] > 0) & (df['Price'] > 0)]

# 4. Stock codes like 'POST' (Postage) are service fees,
not product demand.
junk_codes = ['POST', 'D', 'M', 'DOT', 'BANK CHARGES',
'ADJUST', 'ADJUST2']
df = df[~df['StockCode'].astype(str).isin(junk_codes)]

# 5. Outlier Handling (Capping at 99th percentile to remove
noise)
q_limit = df['Quantity'].quantile(0.99)
df = df[df['Quantity'] <= q_limit]

print(f"Cleaning complete. Records remaining: {len(df)}
(Removed {initial_count - len(df)} rows)")
```

### Feature Engineering Logic

```
# ---Engineering Features for ML Models---

def create_features(data):
    df = data.copy().set_index('Datestamp')

    # Calendar & Holiday Features
    df['month'] = df.index.month
    df['week_of_year'] =
df.index.isocalendar().week.astype(int)
    df['is_peak_season'] = df['week_of_year'].apply(lambda
x: 1 if x >= 47 or x <= 1 else 0)
```

```
    # Multiple Lags for Momentum and Trend
    df['lag_1'] = df['y'].shift(1)
    df['lag_2'] = df['y'].shift(2)
    df['lag_4'] = df['y'].shift(4)

    # Seasonal Diagnostic Lag (Used for Heatmap ONLY)
    df['lag_52'] = df['y'].shift(52)

    # Trend Indicators
    df['rolling_mean_4'] =
df['y'].shift(1).rolling(window=4).mean()

    return df

# A. Create data for the Heatmap (Diagnostic phase)
diag_data = create_features(weekly_data).dropna()

# --- DISPLAY HEATMAP FOR CORRELATION ANALYSIS ---

plt.figure(figsize=(8, 6))
sns.heatmap(diag_data.corr(), annot=True, cmap='coolwarm',
fmt=".2f")
plt.title('Diagnostic Correlogram (Testing Seasonality with
Lag 52)')
plt.show()

# B. Create training data WITHOUT Lag 52 to keep row count
high
# Dropping rows only based on lag_1 and rolling_mean_4
(losing only 4 weeks, not 52)
ml_ready_data =
create_features(weekly_data).drop(columns=['lag_52']).dropn
a()

print(f"Total available weeks for modeling:
{len(ml_ready_data)}")
```

**Train-Test Split**

```
# --- Explicit 90/12 Train-Test Split ---
#
==========================================================
===

# We take exactly 90 weeks for training
X_train = ml_ready_data.drop(columns=['y']).iloc[:90]
y_train = ml_ready_data['y'].iloc[:90]

# We take the next 12 weeks for testing
X_test = ml_ready_data.drop(columns=['y']).iloc[90:102]
```

```
y_test = ml_ready_data['y'].iloc[90:102]

print(f"Split complete:")
print(f"- Training: {len(X_train)} weeks
({X_train.index.min().date()} to
{X_train.index.max().date()})")
print(f"- Testing:  {len(X_test)} weeks
({X_test.index.min().date()} to
{X_test.index.max().date()})")

# Ensure no data leakage
if X_train.index.max() < X_test.index.min():
    print("Check passed: No overlap between training and
testing dates.")
```

**Model Implementation**

```
# --- Running Models & Hyperparameter Tuning ---
# ==========================================

print("...Running Models & Performance Metrics...")

# 1. Unified Metric Function
def calculate_metrics(actual, predicted, model_name):
    mae = mean_absolute_error(actual, predicted)
    # RMSE: Penalizes larger errors, critical for inventory
safety planning
    rmse = np.sqrt(root_mean_squared_error(actual,
predicted))
    # WMAPE: Industry standard for retail volume-weighted
accuracy
    # Handle cases where actual sum might be zero to avoid
division by zero
    if np.sum(actual) == 0:
        wmape = np.nan # Or 0, depending on desired
behavior for zero actuals
    else:
        wmape = (np.sum(np.abs(actual - predicted)) /
np.sum(actual)) * 100

    return {
        'Model': model_name,
        'MAE': round(mae, 2),
        'RMSE': round(rmse, 2),
        'WMAPE (%)': round(wmape, 2)
    }

results = {}
metrics_list = []
```

```
# --- Naive Forecasting (Baseline) ---
results['Naive'] = np.full(shape=12,
fill_value=y_train.iloc[-1])
metrics_list.append(calculate_metrics(y_test,
results['Naive'], 'Naive'))

# ---  Exonential Smoothing (ETS) ---
# Using seasonal_periods=52 for weekly data as per thesis
plan
ets_model = ExponentialSmoothing(y_train, seasonal='add',
seasonal_periods=12).fit()
results['ETS'] = ets_model.forecast(12)
metrics_list.append(calculate_metrics(y_test,
results['ETS'], 'ETS'))

# --- SARIMA ---
sarima_model = SARIMAX(y_train, order=(1,1,1),
seasonal_order=(0,1,1,52)).fit(disp=False)
results['SARIMA'] =
sarima_model.get_forecast(steps=12).predicted_mean
metrics_list.append(calculate_metrics(y_test,
results['SARIMA'], 'SARIMA'))

# --- Random Forest ---
rf_model = RandomForestRegressor(n_estimators=100,
random_state=42).fit(X_train, y_train)
results['Random Forest'] = rf_model.predict(X_test)
metrics_list.append(calculate_metrics(y_test,
results['Random Forest'], 'Random Forest'))

# --- XGBoost with Grid Search ---
tscv = TimeSeriesSplit(n_splits=3)
param_grid = {'max_depth': [3, 5], 'learning_rate': [0.05,
0.1], 'n_estimators': [100]}
xgb_search = GridSearchCV(XGBRegressor(random_state=42),
param_grid, cv=tscv, scoring='neg_mean_absolute_error')
xgb_search.fit(X_train, y_train)

results['XGBoost'] =
xgb_search.best_estimator_.predict(X_test)
metrics_list.append(calculate_metrics(y_test,
results['XGBoost'], 'XGBoost'))

# --- Neural Network (MLP) ---
mlp_model = MLPRegressor(hidden_layer_sizes=(100, 50),
max_iter=1000, random_state=42).fit(X_train, y_train)
results['Neural Network'] = mlp_model.predict(X_test)
metrics_list.append(calculate_metrics(y_test,
results['Neural Network'], 'Neural Network'))

print("All models processed successfully.")
```

**Plot All Models and Actual Demand**

```
print("...Generating Final Results Comparison...")

# 1. Create and Display the Comparison Table
metrics_df = pd.DataFrame(metrics_list)
# Sort by WMAPE to show the best performing model at the
top
metrics_df = metrics_df.sort_values(by='WMAPE (%)')

print("\n--- FINAL PERFORMANCE METRICS ---")
display(metrics_df)

# Save metrics for thesis appendix
metrics_df.to_csv(f'{RESULTS_DIR}/final_performance_metrics
.csv', index=False)

# 2. Visual Comparison Plot
plt.figure(figsize=(15, 8))

# Plot Actual Data
plt.plot(y_test.index, y_test.values, label='Actual Sales',
color='black', marker='o', linewidth=2)

# Plot all model predictions stored in results
for name, preds in results.items():
    plt.plot(y_test.index, preds, label=f'{name} Forecast',
linestyle='--')

plt.title('Comparative Analysis of All Forecasting
Methods', fontsize=16)
plt.xlabel('Date')
plt.ylabel('Quantity')
plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.savefig(f'{RESULTS_DIR}/complete_model_comparison.png')
plt.show()
```

## Appendix 2. Source Code Repository

In order to ensure transparency, integrity, and reproducibility of the research results, the full source code and implementation used for the research has been published on a GitHub repository and made publicly accessible.

**GitHub Repository (Full Implementation):**

https://github.com/BinitaaKC/Master-s-Thesis

This repository contains:

- Data cleaning and preprocessing pipeline
- Weekly demand aggregation
- Feature engineering functions and lag variable creation
- Train-test split logic for forecasting
- Statistical models (Naive, ETS, SARIMA)
- Machine learning models (Random Forest, XGBoost, Neural Network)
- Hyperparameter tuning and optimization
- Forecast evaluation metrics and comparison
- All visualizations used in the thesis
- Time-series analysis, seasonality analysis, and correlogram generation