NOVA IMS Information Management School

MEGI

Master Degree Program in

Statistics and Information Management

Comparative Analysis of Classical Models and Foundation Models for Retail Sales Forecasting

Maria Madalena Mendes de Matos Baião do Nascimento

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Statistics and Information Management

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Comparative Analysis of Classical Models and Foundation Models for Retail Sales Forecasting

by

Maria Madalena Mendes de Matos Baião do Nascimento

Master Thesis presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Data Analytics

Supervised by

Maria Helena Baptista, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 2 of July 2025

Madalena Nascimento

DEDICATION

To my grandmothers.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor, Maria Helena Baptista, PhD, for her invaluable guidance and academic insight.

I am profoundly grateful to my family for their unwavering support, patience, and belief in me. Their constant presence has been a source of strength and motivation throughout the challenges of university life.

I also wish to thank Gfk Portugal for providing the data that made this research possible. Its collaboration contributed significantly to the empirical foundation of this thesis.

ABSTRACT

This study evaluates and compares the performance of classical forecasting models and advanced machine learning approaches for retail sales forecasting. Specifically, the analysis focuses on the applicability of foundation models, namely TimeGPT-1 and Moirai, in addition to Seasonal Autoregressive Integrated Moving Average (SARIMA), Holt-Winters, and Prophet, across three distinct retail product categories, namely camcorders, media tablets, and toys. The models were applied to weekly sales data over two distinct forecasting windows: March-April 2024, representing stable demand, and November-December 2024, characterised by heightened volatility due to the influence of holiday and promotional activities. Forecast accuracy was assessed using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The empirical results demonstrate that foundation models outperformed classical models in terms of absolute accuracy, particularly in contexts characterised by volatile demand conditions. TimeGPT-1 demonstrated a consistent capacity to generate reliable and consistent forecasts across categories. Moirai demonstrated notable efficacy in moderate volatility environments. Prophet produced stable baseline forecasts, but the addition of external regressors did not systematically improve performance. The classical models maintained their competitiveness in stable seasonal contexts. These findings underscore the limitations of classical forecasting models in dynamic retail environments and highlight the potential of foundation models to generalise across different products and timeframes.

KEYWORDS

Forecasting; Foundation Models; Methodological Comparison; Model Performance; Univariate Time Series; Holiday and Promotion Impact

TABLE OF CONTENTS

Statement of Integrity ii
Dedication iii
Acknowledgements iv
List of Figures viii
1. Introduction 1
2. Literature Review 3
2.1. Classical Time-Series Forecasting Models 3
2.2. Prophet Method 6
2.3. Foundation Models 7
2.4. Forecast Accuracy Metrics 12
2.5. Dataset Selection 13
2.6. Comparative Studies 15
3. Conceptual Model 17
4. Empirical Methodology 19
4.1. Data Collection Instrument 19
4.2. Exploratory Data Analysis 19
4.3. Data Split 24
4.4. Forecasting Models 25
4.4.1. Classical Time-Series Forecasting Models 25
4.4.1.1. Arima and Sarima 25
4.4.1.2. Holt-Winters Exponential Smoothing 27
4.4.2. Prophet Method 27
4.4.3. Foundation Models 32
4.4.3.1. TimeGPT-1 32
4.4.3.2. Moirai 33
4.5. Evaluation Metrics 35
5. Results and Discussion 37
5.1. RMSE and MAPE Summary Tables 37
5.2. Camcorders 40
5.3. Media Tablets 44
5.4. Toys 48
5.5. Cross-Model Comparative Analysis 52
6. Conclusions and Future Research 55

Bibliographical References 57
Appendix A 62

LIST OF FIGURES

Figure 2.1 – Recurrent Neural Networks structure (Ensafi et al., 2022) 10
Figure 2.2 - 1D Convolutional Neural Networks architecture (Ensafi et al., 2022) 12
Figure 4.1 – Camcorders sales 20
Figure 4.2 – Media Tablets sales 21
Figure 4.3 – Toys sales 21
Figure 4.4 – Rolling statistics for Camcorders time series 22
Figure 4.5 – Rolling statistics for Media Tablets time series 23
Figure 4.6 – Rolling statistics for Toys time series 23
Figure 5.1 – Forecast by SARIMA configurations for camcorders on March-April 2024 41
Figure 5.2 – Forecast by Holt-Winters for camcorders on March-April 2024 42
Figure 5.3 – Forecast by Prophet configurations for camcorders on March-April 2024 42
Figure 5.4 – Residuals for TimeGPT-1 forecast for camcorders on March-April 2024 43
Figure 5.5 – Forecast by foundation models for camcorders on March-April 2024 44
Figure 5.6 – Forecast by SARIMA configurations for media tablets on November-December 2024 45
Figure 5.7 – Forecast by Holt-Winters for media tablets on November-December 2024 45
Figure 5.8 – Forecast by Prophet configurations for media tablets on November-December 2024 46
Figure 5.9 – Forecast by foundation models for media tablets on November-December 2024 47
Figure 5.10 – Residuals for Moirai forecast for media tablets on November-December 2024 48
Figure 5.11 – Forecast by SARIMA configurations for toys on March-April 2024 49
Figure 5.12 – Forecast by Holt-Winters for toys on March-April 2024 50
Figure 5.13 – Forecast by Prophet configurations for toys on March-April 2024 50
Figure 5.14 – Forecast by foundation models for toys on March-April 2024 51
Figure 5.15 – Residuals for Moirai forecast for toys on March-April 2024 52
Figure A.1 – Forecast by SARIMA configurations for camcorders on November-December 2024 62
Figure A.2 – Forecast by Holt-Winters for camcorders on November-December 2024 62
Figure A.3 – Forecast by Prophet configurations for camcorders on November-December 2024 63
Figure A.4 – Residuals for TimeGPT-1 forecast for camcorders on November-December 2024 63

Figure A.5 – Forecast by foundation models for camcorders on November-December 2024 64
Figure A.6 – Forecast by SARIMA configurations for media tablets on March-April 2024..... 64
Figure A.7 – Forecast by Holt-Winters for media tablets on March-April 2024 65
Figure A.8 – Forecast by Prophet configurations for media tablets on March-April 2024..... 65
Figure A.9 – Forecast by foundation models for media tablets on March-April 2024 66
Figure A.10 – Residuals for Moirai forecast for media tablets on March-April 2024..... 66
Figure A.11 – Forecast by SARIMA configurations for toys on November-December 2024 ... 67
Figure A.12 – Residuals for SARIMA configurations forecasts for toys on November-December 2024..... 67
Figure A.13 – Forecast by Holt-Winters for toys on November-December 2024..... 68
Figure A.14 – Forecast by Prophet configurations for toys on November-December 2024 ... 68
Figure A.15 – Residuals for Prophet forecast for toys on November-December 2024 69
Figure A.16 – Forecast by foundation models for toys on November-December 2024..... 69

LIST OF TABLES

Table 4.1 – Descriptive statistics of the three products 20
Table 4.2 – p-values of the ADF test 23
Table 4.3 – SARIMA hyperparameters for March-April 2024 26
Table 4.4 – SARIMA hyperparameters for November-December 2024 26
Table 4.5 – Optimal Prophet hyperparameter configurations for Camcorders across forecasting windows and model variants 29
Table 4.6 – Optimal Prophet hyperparameter configurations for Media Tablets across forecasting windows and model variants 30
Table 4.7 – Optimal Prophet hyperparameter configurations for Toys across forecasting windows and model variants 30
Table 4.8 – Optimal hyperparameters for the Moirai model (March-April 2024)..... 34
Table 4.9 – Optimal hyperparameters for the Moirai model (November-December 2024)... 34
Table 5.1 – Summary of RMSE Performance on March-April 2024..... 37
Table 5.2 – Summary of MAPE Performance on March-April 2024 38
Table 5.3 – Summary of RMSE Performance on November-December 2024 38
Table 5.4 – Summary of MAPE Performance on November-December 2024..... 39

LIST OF ABBREVIATIONS AND ACRONYMS

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
CNN	Convolutional Neural Network
CV	Computer Vision
GAM	Generalised Additive Model
LLM	Large Language Model
LOTSA	Large-scale Open Time Series Archive
LSTM	Long Short-Term Memory Network
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
NLP	Natural Language Processing
NN	Neural Network
PACF	Partial Autocorrelation Function
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SARIMA	Seasonal Autoregressive Integrated Moving Average
SES	Simple Exponential Smoothing

1. INTRODUCTION

Accurate sales forecasting is critical for businesses in the retail sector, as it influences inventory planning, marketing strategies, and overall profitability. As the volume and complexity of sales data continues to increase, so does the need for more sophisticated forecasting methods. This need is now being addressed by the recent rise of foundation models.

Foundation models have emerged as a powerful tool for analysing time series data, effectively capturing its inherent temporal structure, including trends and seasonality. By leveraging large historical datasets, these models enhance predictive accuracy and adaptability across multiple domains. In recent years, foundation models have transformed time series analysis by providing a sophisticated framework for capturing the complex patterns and dependencies within sequential data. Trained on diverse datasets, foundation models develop a generalised understanding that significantly improves computational analysis, enabling deeper insights and more informed decisions (Awais et al., 2023).

However, despite these advancements, a notable gap in the evolution of their effectiveness in processing long sequences of data remains, particularly in the context of sales (Miller et al., 2024). This study addresses this gap by examining the effectiveness of different foundation model architectures, in addition to traditional and modern time series forecasting methods.

To guide the investigation, the following research questions are proposed: How can foundation models be effectively adapted to capture the complexity of time series sales data? Which specific architectures provide the most accurate forecasts across different product categories? The objectives of the research are threefold: firstly, to develop a robust framework for implementing foundation models in sales forecasting; secondly, to conduct comparative analyses of different model architectures; and thirdly, to identify key features such as holidays and promotions that significantly impact forecast accuracy.

In order to address the aforementioned research questions, the present study employs a comparative methodology involving five forecasting models: Seasonal Autoregressive Integrated Moving Average (SARIMA), Holt-Winters, Prophet, and two foundation models, TimeGPT-1 and Moirai. The analysis involves the training and evaluation of each model on the same weekly sales dataset covering the years 2022 to 2024 across three product categories: camcorders, media tablets, and toys. In order to capture both moderate and peak seasonal demand conditions, two distinct forecasting windows were selected: March to April 2024 and November to December 2024. Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) are employed as evaluation metrics, ensuring a consistent basis for comparison across models, products, and timeframes.

It is anticipated that foundation models will outperform traditional methods, particularly in capturing complex patterns. Key factors such as holiday sales periods and promotions are

likely to have a significant impact on forecasting accuracy. The ability of foundation models to learn from these complexities offers a promising advantage over traditional methods, which often rely on simpler assumptions and linear relationships.

Given that foundation models are very new, there is a critical need for empirical evidence regarding their effectiveness in real-world scenarios. Ultimately, this study seeks to enrich the field with valuable knowledge regarding the implementation of foundation models, thereby encouraging their adoption in settings where traditional methods may fall short.

The structure of the thesis is as follows: Chapter 2 provides a comprehensive literature review, outlining existing approaches to time series forecasting and recent developments in foundation models. Chapter 3 presents the conceptual framework and the rationale behind the proposed approach. Chapter 4 provides a detailed exposition of the empirical study, encompassing the characteristics of the dataset and descriptions of the forecasting models that were evaluated. Chapter 5 provides a detailed exposition of experimental results and their subsequent analysis. Finally, Chapter 6 concludes the study with a discussion of the key findings, the limitations of the study, and directions for future research.

2. LITERATURE REVIEW

This chapter presents a comprehensive review of the key elements in time series forecasting, focusing on the methodologies, evaluation metrics, and datasets that form the foundation of this field. It begins by defining classical and advanced forecasting models, followed by an examination of the metrics used to measure and compare their accuracy. The discussion then shifts to dataset selection, highlighting challenges and opportunities with public and proprietary data. Finally, the chapter concludes with insights from comparative studies in the sales domain.

2.1. CLASSICAL TIME-SERIES FORECASTING MODELS

No single method is universally effective for forecasting time series. Each forecasting problem may require a unique approach, depending on the characteristics of the time series (Ensafi et al., 2022). Among the traditional methods, Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing are widely used, each offering different advantages.

ARIMA integrates autoregressive (AR) and moving average (MA) models. Introduced by Box & Jenkins (1970), ARIMA can be applied to both stationary and non-stationary time series. However, it is usually applied to non-stationary time series because it can transform the series into stationary by taking the difference of the sequences.
This model predicts future values as a linear combination of past observations and random errors, expressed as:

$$y_t = \theta_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}. \quad (2.1)$$

In this formula, y_t denotes the future value, p and q represent the AR and MA polynomials, respectively, and ε_t is the random error at time t. The coefficients θ_i and ϕ_i capture the relationships between variables.

ARIMA is particularly effective for time series that do not show significant seasonal patterns. When seasonality is present, the Seasonal ARIMA (SARIMA) model is more appropriate, as it extends ARIMA by incorporating seasonal components (Box & Jenkins, 1970).

The SARIMA model is represented as $ARIMA(p, d, q)(P, D, Q, s)$, where p, d, and q represent the non-seasonal parameters: p refers to the order of the AR component, d represents the number of differencing operations required to achieve stationarity, and q indicates the order of the MA component. Similarly, P represents the seasonal AR order, D indicates the seasonal differencing order, and Q denotes the seasonal MA order. The parameter s defines the seasonal period, which indicates the length of the cycle in the data. The optimal determination of these parameters is a crucial step in model building to ensure reliable forecasts.

To select ARIMA and SARIMA parameters, it is important to examine the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) plots. The ACF of a MA process

becomes zero beyond the order q, exhibiting a distinct cutoff after lag q. Conversely, the PACF of an AR process is zero beyond lag p, aiding in the identification of the order of the AR component. In the context of an ARIMA or SARIMA process, the ACF typically follows a pattern characterised by a combination of exponential and damped sine waves after the first $q - p$ lags, while the PACF exhibits a similar behaviour after the first $p - q$ lags. Additionally, in SARIMA models, the seasonal period s is often inferred from the peaks observed in the ACF plot.

Beyond visual analysis, hyperparameter tuning for ARIMA and SARIMA models can be optimised using a Grid Search approach. This method systematically explores combinations of (p, d, q) and (P, D, Q) to evaluate model performance based on statistical criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Lower values of AIC and BIC indicate models with better fit and lower complexity (Box & Jenkins, 1970).

ARIMA and SARIMA assume a linear relationship between past values and future predictions, making them less suitable for datasets with complex non-linear patterns. Also, these models require stationarity, which means that the data must undergo transformations such as differencing, which may not always be the optimal solution (Wang et al., 2021).

An alternative to ARIMA-based models is Exponential Smoothing, which was developed by Brown (1963) and extended by Holt (2004). Unlike ARIMA, which applies equal weighting to past observations, Exponential Smoothing assigns unequal weights to observations in a time series. This method gives priority to the most recent data, while gradually reducing the influence of older observations.
Simple Exponential Smoothing (SES) is a forecasting technique that improves predictions by averaging past values of time series data (Brown, 1963). It is particularly effective for stationary series that exhibit random fluctuations around a constant mean. The method assigns an exponentially decreasing weight to each observation. Mathematically, it is expressed as:

$$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1}) = \alpha A_{t-1} + (1 - \alpha)F_{t-1}, \tag{2.2}$$

where F_t is the forecast value for time t, F_{t-1} is the forecast value for the previous period, A_{t-1} is the actual value for period $t - 1$, and α is the smoothing parameter $(0 \leq \alpha \leq 1)$. Larger values of α reduce the degree of smoothing, and at $\alpha = 1$ the forecast is identical to the original series (with a one-period lag).

SES is well suited to univariate data with no trend or seasonality, offering simplicity and reliability for many applications. The method has been successfully applied to a variety of forecasting problems. For instance, Gustriansyah et al. (2019) used SES to forecast a pharmacy's monthly sales across multiple products, achieving high accuracy in their predictions.

The Holt-Winters method extends Exponential Smoothing to handle time series with trends and seasonality. It accounts for three key components: the level, the trend, and the seasonal factor, which can be modelled in two variants, additive or multiplicative, depending on the nature of the seasonal patterns in the data (Winters, 1960). Both models enhance the ability to capture different data patterns, thereby improving forecast accuracy.

The additive model is most appropriate when the seasonal variations have a constant amplitude over time. It represents the series as:

$$Y_t = T_t + S_t + \varepsilon_t, \tag{2.3}$$

where T_t is the trend, S_t denotes the seasonal component, and ε_t represents the random error term. This term accounts for the unpredictable fluctuations in the data that cannot be explained by the trend and the seasonal component. It is assumed to have a zero mean and a constant variance.

On the other hand, the multiplicative model is preferred when the seasonal variation increases proportionally with the level of the series. In this case, the series is expressed as:

$$Y_t = T_t \times S_t \times \varepsilon_t. \tag{2.4}$$

In this formulation, ε_t denotes the random error term, which encapsulates unpredictable fluctuations that are proportional to the level of the series. The error typically has a mean of one, and the magnitude of its variability increases as the series level rises, reflecting the multiplicative nature of the model.

The ability of Holt-Winters method to consider both trend and seasonal components makes it highly adaptable to time series that show consistent upward or downward movements, such as sales trends. Its dynamic nature allows the model to adjust as patterns evolve, making it particularly effective at predicting sales over time. However, its performance is heavily influenced by the suitability of the model chosen (additive or multiplicative) to the underlying data.

For example, a study by Lima et al. (2019) evaluated the effectiveness of additive and multiplicative Holt-Winters models in forecasting e-commerce retail sales in Portugal. The results showed that the multiplicative model outperformed the additive model when seasonal variation increased with the level of the series. This highlights the importance of correctly identifying the type of seasonality in the data when choosing between the two approaches.
Comparative studies have further evaluated the performance of ARIMA, SARIMA and Holt-Winters methods, highlighting their respective strengths and limitations. Puthran et al. (2014) found that both the Holt-Winters multiplicative model and the SARIMA model were effective in forecasting monthly motorcycle sales in India, but the Holt-Winters model was more precise and accurate. Mgale et al. (2021) concluded that the Holt-Winters additive model outperformed ARIMA in forecasting rice prices in Tanzania. A more recently study by Kumar

et al. (2024) also analysed these two models in the context of demand forecasting in dynamic pricing. The researchers found that the Holt-Winters method was particularly effective in capturing seasonality and cyclical trends, while ARIMA was better suited to cases requiring dynamic price adjustments. The recommendation of a hybrid approach, integrating the strengths of both ARIMA and Holt-Winters, was made to enhance forecasting accuracy.
In conclusion, ARIMA and SARIMA are best suited to linear time series and require stationary data, while Exponential Smoothing and the Holt-Winters methods are more flexible but still rely on assumptions of stationarity, trend and seasonality. Therefore, the choice of forecasting method depends largely on the characteristics of the time series being analysed.

2.2. PROPHET METHOD

The Prophet model, developed by Facebook's Core Data Science team, is a time series forecasting tool designed to simplify the application of forecasting techniques across large datasets and large number of users. This open-source model is available in both Python (Python Software Foundation, 2020) and R (R Core Team, 2017).

In their foundational paper, Taylor & Letham (2017) introduced Prophet and its methodology. The model is designed to handle the common characteristics found in business time series data. It offers a user-friendly interface that allows non-experts to adjust model parameters intuitively, without requiring a deep understanding of the model. This feature promotes a feedback loop that allows the model's performance to be continuously improved over time.

The underlying model features a decomposable time series with three main components: growth (or trend), seasonality and holidays. These are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t. \tag{2.5}$$

The trend, represented by $g(t)$, captures non-periodic changes in the value of the time series. To represent this trend, the model uses either a piecewise linear function or a logistic growth curve. This approach provides the flexibility to deal with trends that do not follow a consistent linear pattern. By identifying specific points at which the growth rate changes, the model adapts to shifts in the trend over time. These changepoints can be customised to adjust the trend component to ensure that it fits the characteristics of the data.

The seasonality component, $s(t)$, takes into account recurring patterns in data, such as annual, weekly, or daily variations. To fit and forecast these patterns, the model uses Fourier series, which can handle different seasonal frequencies and magnitudes, capturing the regular cyclical behaviour of the data.

Holiday effects, captured by $h(t)$, model the influence of holidays or significant events on the time series. Unlike regular seasonality, holidays often cause abrupt spikes or dips that do not follow a predictable cycle. For example, Easter's date varies each year because it is celebrated on the first Sunday after the first full moon following the spring equinox. However, its effect

on a time series (such as sales) tends to be consistent, so it is important to include it in the forecast. To take account of such effects, the model introduces dummy variables to denote whether or not there is a holiday at a particular time. This feature is very useful for datasets with several years of historical data and where holidays have a significant impact, like sales data.

The model also includes an error term, ε_t, which represents changes that are not explained by the model. This includes random noise and other unforeseen factors influencing the data. This term is assumed to follow a normal distribution, which helps to filter out noise and ensure a smoother forecast.

The Prophet model can be seen as a Generalised Additive Model (GAM), a regression framework that applies flexible, potentially non-linear smoothers to its predictors. This structure has the advantage of being easily decomposable and adaptable, allowing new components to be added as required. In addition, GAMs are computationally efficient, allowing users to quickly adjust model parameters and interactively refine the forecasting process.

Prophet approaches forecasting as a curve-fitting problem rather than focusing on the temporal dependency structure, which has several practical advantages. This approach allows for flexibility in accommodating multiple seasonal periods, enables rapid model fitting and provides interpretable parameters that allow analysts to easily adjust forecast assumptions. Moreover, Prophet is well suited to real-world forecasting challenges as it can handle irregularly spaced data and does not require filling in missing values.

Empirical applications of Prophet have demonstrated its versatility and effectiveness. For instance, Zunic et al. (2020) used the model to forecast sales dynamics within a product portfolio, generating both monthly and quarterly sales forecasts. The results showed that Prophet not only provided reasonably accurate forecasts but also showed promise in classifying products based on the expected reliability of those forecasts.

However, the performance of Prophet has been questioned in certain applications compared to alternative forecasting methods. Hasan et al. (2022) compared Prophet and ARIMA for forecasting retail sales across three product categories (household, hobby and food) at ten Walmart stores. After model tuning, ARIMA proved to be more accurate, while Prophet struggled to deal effectively with the nuances of higher value sales data.

2.3. FOUNDATION MODELS

The architecture of foundation models has increasingly converged on the Transformer, a model that introduced a revolutionary approach by eliminating recurrence and relying entirely on an attention mechanism to capture global dependencies between input and output (Liang et al., 2024).

An attention function maps a query and a set of key-value pairs to an output, with all the components represented as vectors. The output is computed as a weighted sum of the values, where the weights are determined by a compatibility function that measures the similarity between the query and each corresponding key (Vaswani et al., 2017).

Mathematically, the attention mechanism is expressed as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.6}$$

where Q (query), K (key), and V (value) are matrices of dimensions $T \times d_k$. Here, T represents the number of tokens in the input sequence and d_k denotes the dimensionality of the vector representation for each token.

The scaling factor $\sqrt{d_k}$ is used to moderate the magnitude of the dot product QK^T. Without this normalisation, the resulting values could become excessively large when d_k is high, resulting in extremely sharp or skewed distributions after the application of the softmax function. By dividing the dot product by $\sqrt{d_k}$, the mechanism ensures numerical stability and prevents the attention scores from becoming overly sensitive. This normalisation facilitates efficient computation, allowing the model to dynamically focus on the most relevant parts of the input sequence (Liang et al., 2024).

A key advantage of this architecture is its suitability for parallel processing, which enhances scalability and enables the processing of large datasets. Combined with its ability to model global dependencies, the Transformer has gained widespread adoption beyond its original application in natural language processing (NLP) tasks, extending to fields such as computer vision (CV), time series forecasting and more.

The original Transformer architecture, introduced by Vaswani et al. (2017), is an encoder-decoder model originally designed for machine translation. In this framework, the encoder processes an input sentence to generate a context-rich representation, while the decoder generates the translation token by token, using both the encoded representation and previously generated tokens. A notable example of this architecture in time series forecasting is TimeGPT-1 (Garza et al., 2024), which employs multiple Transformer layers within an encoder-decoder framework to enable efficient zero-shot forecasting across diverse datasets.

Variants of the Transformer architecture, such as decoder-only and encoder-only models, have since been developed for different tasks. Decoder-only models operate without an encoder, generating one token at a time by remembering all previous tokens, making them well suited to autoregressive tasks such as text generation. These models show superior performance on longer input sequences compared to the original encoder-decoder architecture (P. J. Liu et al., 2018). Encoder-only models, on the other hand, focus solely on encoding the input data into a meaningful representation, making them suitable for data comprehension tasks. Moirai (Woo et al., 2024), for instance, exemplifies this architecture

through its masked encoder-based design, which has been tailored for universal time series forecasting. The model in question was pre-trained on the Large-scale Open Time Series Archive (LOTSA), a dataset comprising 27 billion observations from nine distinct domains, including sales.

Recent research has highlighted the strengths and limitations of these variants, particularly for time series forecasting. Y. Liu et al. (2024) compared encoder-only and decoder-only models under different levels of data availability. Their results revealed that encoder-only models tend to perform better when training data is limited, as they are less prone to overfitting. In contrast, decoder-only models excel when pre-trained on large datasets and consistently outperform encoder-only models. These trends are consistent with observations from large language models, where decoder-only models excel at generalising across diverse domains. Consequently, these models appear to be a promising choice for developing large, scalable time series models. Ansari et al. (2024) further explored the versatility of encoder-decoder models and extended their analysis to decoder-only architectures, reinforcing the growing adaptability of Transformer variants.

The Transformer architecture also leverages its ability to model sequences in order to capture temporal dynamics in time series data. One approach is to reuse pre-trained large language models (LLMs) for time series tasks, exploiting their inherent strengths in sequence modelling (Liang et al., 2024).

In language models, the next token, w_{k+1}, in a sequence $w_{1:k} = [w_1, \dots, w_k]$ is predicted by estimating the conditional distribution $p(w_{k+1}|w_{1:k})$. These tokens belong to a vocabulary, V, which can include characters, words, or other units, depending on the tokenisation scheme chosen (Ansari et al., 2024).

PromptCast, introduced by Xue & Salim (2023), illustrates the application of LLMs to time series forecasting through a prompt-based approach. This method transforms time series data into text-based input-output pairs, thereby reframing the forecasting task as a question-answer problem. However, PromptCast relies on dataset-specific templates to effectively convert numerical data into text prompts, which may limit its general applicability.

Alternatively, Transformers can be used as the basis for time series foundation models that are trained from scratch, enabling them to better capture the unique characteristics of time series data (Liang et al., 2024). A prime example is TimeGPT-1 (Garza et al., 2024), a specialised foundation model trained from scratch on over 100 billion data points sourced from publicly available time series spanning diverse domains such as finance, healthcare, energy, weather, IoT sensors (i.e., Internet of Things devices that collect data from physical environments and transmit it to central systems or cloud platforms for processing and analysis), sales, and transportation. Its architecture accommodates varying input sizes, frequencies, and forecast horizons while minimising errors. As a result, TimeGPT-1 consistently outperforms alternative models with minimal complexity. This model represents a paradigm shift in time series

forecasting, providing a solution that is more accessible, accurate, time and computationally efficient.

In parallel, several techniques have emerged to enhance the capabilities of Transformer models in the context of time series applications. A common strategy employed in foundation models is patch-based segmentation, in which the input time series is divided into smaller segments (or patches), thereby enabling the model to capture local temporal dynamics more effectively. Moirai (Woo et al., 2024) follows this patch-based approach to model time series. Furthermore, specialised approaches such as multi-resolution analysis, also demonstrated by Moirai through the employment of varying patch sizes, have been shown to enhance model efficacy substantially.

However, the effectiveness of Transformers is often context dependent, with their advantages becoming more pronounced as dataset size increases (Garza et al., 2024).

Before Transformers revolutionised machine learning, Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) were widely used for pre-training models that excelled at capturing general patterns in complex datasets (Liang et al., 2024).

In recent years, there has been renewed interest in RNN architectures, driven by the need for resource-efficient models capable of processing long sequences (Liang et al., 2024). RNNs are particularly effective in time series forecasting due to their ability to retain information from past events. This ability is made possible by a recurrent loop in their hidden layers, where the output at each time depends not only on the current input but also on the activation values of previous hidden states (Ensafi et al., 2022). This structure makes RNNs to be naturally suited for processing sequential data, as they function as a series of identical networks that pass information forward in time.

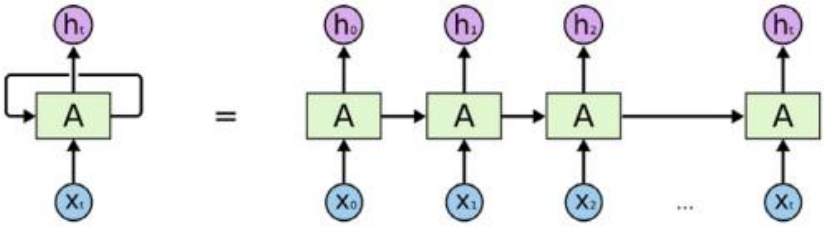

Figure 2.1 – Recurrent Neural Networks structure (Ensafi et al., 2022)

However, RNNs face a significant limitation: as information passes through multiple hidden layers, the effects of previous inputs can diminish or disappear due to repeated multiplication, a phenomenon known as the vanishing gradient problem. This challenge prevents RNNs from effectively capturing long-term dependencies, especially when the gap between relevant information and its required context becomes too large (Hochreiter & Schmidhuber, 1997).

Long Short-Term Memory Networks (LSTMs) have been developed to address this problem. Introduced by Hochreiter & Schmidhuber (1997), LSTMs are a specialised version of RNNs designed to retain information over extended time intervals through a sophisticated gating mechanism. This mechanism comprises three gates: the forget gate, the input gate, and the output gate, which work collectively to manage the flow of information within the network.

The fundamental principle underlying the functionality of LSTMs is the cell state, which facilitates the transmission of information across time steps. This structural design enables the network to selectively update and modify its configuration, thereby determining the retention or discarding of information at each time step. The gating process is governed by sigmoid functions, which constrain their outputs to the range $[0,1]$. A value of 0 implies complete blockage of information, while a value of 1 permits all information to pass through. Specifically, the forget gate determines which information is discarded, the input gate controls the addition of new information, and the output gate selects the information to propagate as the hidden state (*Understanding LSTM Networks -- Colah's Blog*, 2015).

The operations of the gates and the updates to the cell state are expressed mathematically as follows:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
\tilde{C}_t &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
h_t &= o_t * \tanh(C_t)
\end{aligned}
\tag{2.7}
$$

Here, f_t, i_t, and o_t correspond to the forget gate, input gate, and output gate, respectively. σ is the sigmoid function, and the tangent function ($tanh$) maps values to the range $[-1,1]$. The hidden state from the previous time step is denoted by h_{t-1}, and x_t represents the input vector at the current time step. The operator $*$ indicates element-wise multiplication of vectors. The candidate cell state values are represented as $\tilde{C}_t$, while C_t and h_t denote the updated cell state and the hidden state output at time t, respectively. The weight matrices and the bias vectors associated with each gate are represented by W and b, respectively.

The cell state acts as the primary memory mechanism, preserving relevant information, while the hidden state represents the network's output at each time step. Together, these components enable LSTMs to effectively model both short-term and long-term dependencies in sequential data.

The ability of LSTMs to selectively manage memory has proven valuable, for example, Bandara et al. (2019) successfully applied LSTM networks to exploit non-linear demand relationships in the e-commerce sector, improving the accuracy of sales forecasts.

At the same time, CNNs, originally developed for image processing, have gained attention in time series forecasting. Their ability to effectively identify patterns makes them particularly suitable for capturing local temporal dynamics in sequential data.

The architecture of a CNN typically involves multiple layers, with the output of one convolutional layer being passed as input to the next. One of its key components is the max-pooling layer, which reduces dimensionality by selecting the maximum sliding window, helping to minimise the risk of overfitting. Between the pooling and fully connected layers, a flatten layer is introduced to convert the multi-dimensional data into a 1D vector, making it suitable for input to the fully connected layer (Ensafi et al., 2022).

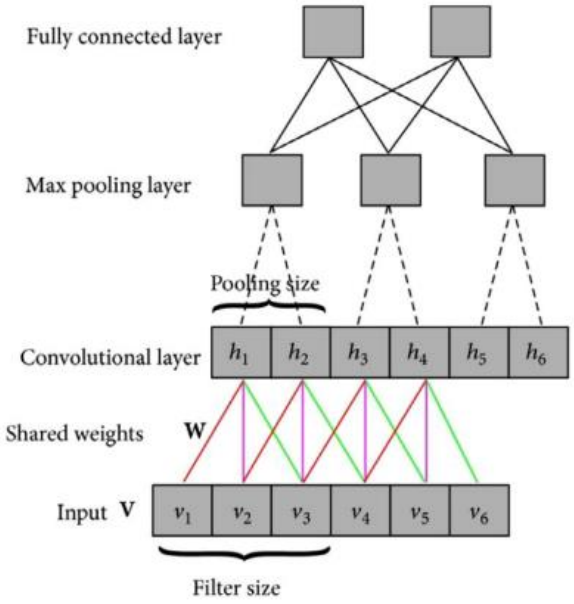

Figure 2.2 - 1D Convolutional Neural Networks architecture (Ensafi et al., 2022)

2.4. FORECAST ACCURACY METRICS

The accuracy of sales forecasting has been the focus of significant research efforts aimed at improving its precision. In the relevant literature, accuracy is usually evaluated using error metrics such as Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE), by comparing actual sales with predicted results.

The forecast error, denoted as $e_t = Y_t - F_t$, represents the difference between the observed value Y_t and the forecast value F_t, regardless of the method used to produce the forecast (Hyndman & Koehler, 2006).

There are some commonly used measures of accuracy whose scale depends on the scale of the data. Scale-dependent measures, such as Mean Absolute Error (MAE), Mean Squared

Error (MSE), and RMSE, are useful for comparing different methods applied to the same dataset. However, these metrics are less suitable for comparisons between datasets of different scales. Among these, RMSE is frequently preferred over MSE because it is expressed on the same scale as the data, making it easier to interpret. Historically, RMSE and MSE have been widely adopted due to their theoretical relevance in statistical modelling. However, their sensitivity to outliers often makes MAE a more robust alternative.

Percentage error metrics, such as MAPE, have the advantage of being scale independent, making them particularly useful for comparing forecast performance across different data series. MAPE is easy to interpret but has some limitations. It becomes infinite or undefined when $Y_t = 0$ for any t, and its distribution can be highly skewed as Y_t approaches zero. Despite these drawbacks, MAPE is often preferred in cases where all the data are positive and significantly greater than zero, due to its simplicity.

When all series are on the same scale, MAE may be favoured for its straightforwardness, as it provides a direct and interpretable measure of error without the complexities associated with percentage-based metrics (Hyndman & Koehler, 2006). Each accuracy metric has unique advantages and limitations, and its suitability depends on the specific characteristics of the dataset and the forecasting context.

2.5. DATASET SELECTION

In the current historical context, where deep learning models have demonstrated undeniable superiority in natural language processing and computer vision, the field of time series analysis remains cautious about the performance of foundation models. This scepticism is primarily driven by several key challenges that hinder the application of foundation models to time series forecasting (Garza et al., 2024).

A major obstacle is the misalignment or poor definition of the evaluation settings. In fields like NLP and CV, benchmark datasets, such as ImageNet for computer vision, have been crucial in standardising model evaluation, enabling fair comparisons and driving progress (Garza et al., 2024). In contrast, publicly available datasets for time series forecasting often lack the scale, diversity, and structure required for foundation models to excel. Most datasets are small, domain-specific, and univariate, limiting the ability of deep learning architectures to capture complex temporal dependencies or exploit multivariate relationships (Ansari et al., 2024). Indeed, smaller datasets can lead to overfitting and poor generalisation. As noted by Menculini et al. (2021), the constraints imposed by limited dataset size are particularly pronounced for Neural Networks models (NNs), which show notable performance gains when trained on larger historical datasets.

The lack of standardised large-scale datasets tailored to the needs of foundation models has led to inconsistencies in dataset selection, pre-processing, and evaluation metrics, making it difficult to fairly assess and compare model performance (Garza et al., 2024). Without such

benchmarks, time series forecasting struggles to establish a clear trajectory for innovation, preventing foundation models from reaching their full potential.

The M5 competition (Makridakis et al., 2022b) was an important step in addressing these issues by providing a large, multivariate dataset focused on hierarchical demand forecasting. It included daily sales for ten representative stores of the global retailer Walmart in three US states: California, Wisconsin, and Texas. The dataset covered the period from 2011 to 2016 and included three product categories: food, household, and hobbies. The data was organised hierarchically, with sales recorded at different levels, ranging from individual products in specific stores to total sales at regional and national levels. This hierarchical structure presented a challenge to the forecasting models, as predictions made at the lower levels (e.g., store or product level) had to be consistent with those made at higher levels (e.g., regional or national level). This required the models to maintain coherence across different levels of aggregation and encouraged innovation in modelling approaches.

In addition to sales data, the dataset included contextual variables such as prices, promotions, and calendar details. The availability of such multivariate data allowed foundation models to capture complex relationships between explanatory variables and target series (Makridakis et al., 2022b). However, while the M5 dataset was an improvement on previous public datasets, it still had limitations. The reliance on contextual variables raised concerns about the generalisability of these models to settings where similar data may not be available. Furthermore, the exclusive focus of the dataset on Walmart raised questions about the applicability of the findings to other retail contexts (Theodorou et al., 2021). In addition, the computational demands of training the underlying models created accessibility barriers, reducing their applicability in resource-constrained settings (Makridakis et al., 2022a).

An alternative to publicly available datasets is the use of proprietary datasets, which provide exclusive access to high-quality, granular data. These datasets allow researchers to perform deeper analyses and develop more robust models by capturing unique patterns and features that may be missing from public datasets. For instance, Qi et al. (2021) demonstrated the effectiveness of proprietary datasets by developing Aliformer, a predictive model enriched with domain-specific knowledge, such as information about upcoming promotions. When evaluated on four public benchmark datasets and a large proprietary dataset from Tmall, Aliformer outperformed state-of-the-art forecasting methods, highlighting the potential of proprietary datasets to improve the accuracy of sales forecasts.

While proprietary datasets offer distinct advantages in terms of data richness and real-world applicability, they also present significant challenges in terms of research transparency and accessibility. As access to proprietary datasets is often restricted, findings derived from them cannot always be independently verified or replicated, which can limit their wider impact on the research community. Ethical and legal considerations around data privacy, consent, and fairness need to be carefully managed to ensure responsible use, especially when long-term

access to the data is uncertain. The risk of data becoming inaccessible further complicates the issue (Vilhuber, 2023).

Advancing the standardisation of datasets and the development of robust model evaluation frameworks is critical to unlocking the full potential of foundation models in time series forecasting. Overcoming these challenges will help bridge the gap between the proven success of these models in other fields and their untapped potential in time series forecasting, ultimately enabling more accurate, reliable, and generalisable predictions.

2.6. COMPARATIVE STUDIES

This section provides an overview of previous studies comparing the above-mentioned models in the sales domain, in line with the scope of this study. The main objective is to adopt a methodological framework similar to that of Ensafi et al. (2022) and Menculini et al. (2021).

Ensafi et al. (2022) compared classical and advanced forecasting methods using a furniture sales dataset, providing valuable insights into their effectiveness in capturing seasonal sales patterns. The study evaluated model performance using three key measures: MSE, RMSE, and MAPE.

Among the classical methods, ARIMA struggled to model seasonality and failed to capture the year-end growth trend, while SARIMA performed better and successfully modelled year-end sales growth. Two Prophet models were tested: one that only considered annual seasonality, and another that incorporated the holiday argument to improve forecasting accuracy. While both models successfully captured seasonality, the second Prophet model proved to be superior to the first, outperforming both ARIMA and SARIMA.

In the advanced methods, two types of LSTMs were evaluated: Vanilla LSTM and Stacked LSTM. The Vanilla LSTM includes a single hidden layer and an output layer, while the Stacked LSTM uses multiple LSTM layers designed to produce sequence-based outputs for more complex patterns. Both models demonstrated strong predictive performance compared to the previous classical methods and Prophet models, with the Stacked LSTM model emerging as the superior approach. A CNN was also evaluated and ranked fifth in performance behind the Stacked LSTM, Vanilla LSTM and both Prophet models. Nevertheless, the CNN and Prophet models were highlighted as promising tools for forecasting seasonal trends.

Menculini et al. (2021) widened the scope of the comparison by evaluating traditional time series models, Prophet, and machine learning techniques for predicting food products prices. The performance of these models was assessed using MAE, RMSE, and MAPE. Their study compared the performance of ARIMA models, Prophet, and LSTM networks, both individually and in combination with CNNs. The diversity of models and input datasets made the comparison particularly revealing.

Prophet proved to be a convenient tool due to its ease of setup and tuning, requiring no prior data pre-processing. However, its predictive performance lagged significantly behind the

ARIMA models and NNs, highlighting the trade-off between simplicity and accuracy in time series forecasting.

The study also looked at Neural Network models: those using only LSTM layers and those combining CNN and LSTM layers. Adding CNN layers enhanced the LSTM's ability to extract features from historical data, improving its capacity to learn complex patterns and resulting in the best overall performance. However, this improvement came at the cost of increased training time and computational requirements, which is a potential limitation for real-time or resource-constrained applications.

A notable finding was the advantage of NNs over ARIMA and Prophet in using multivariate regression, allowing for more effective use of the available dataset. Among the Neural Network models, the pure LSTM architecture offered a compelling balance between accuracy and computational efficiency. Although it slightly underperformed compared to the hybrid CNN-LSTM models, it consistently outperformed ARIMA and Prophet, making it a viable option for applications where both accuracy and resource efficiency are important.

Building on these findings, this research adopts a comparative approach to evaluate traditional and advanced forecasting methods using a proprietary dataset tailored to forecast product sales in Portugal. The dataset was carefully selected to balance key factors such as data availability, privacy concerns, and scale. While this choice ensures access to relevant and comprehensive data, it also highlights the challenges posed by the lack of robust public alternatives, as outlined in the previous section.

3. CONCEPTUAL MODEL

This study adopts a structured hypothesis-testing approach to evaluate the forecasting performance of foundation models in comparison to traditional forecasting methods and Prophet. The hypotheses formulated for this study posit that the null hypothesis assumes no significant difference in forecasting performance between foundation models, traditional forecasting models, and Prophet, while the alternative hypothesis suggests that foundation models outperform both traditional models and Prophet in terms of forecasting accuracy.

The conceptual model is designed to assess forecasting accuracy, focusing on sales data as the primary application domain. Four key forecasting paradigms are considered within this framework. Traditional forecasting methods, such as ARIMA and SARIMA, rely on linear relationships and stationary assumptions, rendering them suitable for structured time series with well-defined trends and seasonality. However, their effectiveness diminishes when faced with complex seasonal patterns or external influences. Exponential Smoothing and Holt-Winters methods also rely on assumptions of stationarity, trend and seasonality, and thus again limit their flexibility in handling highly dynamic time series. Prophet introduces an automated, decomposable approach that integrates trend, seasonality, and external factors. While offering increased flexibility, previous research (Hasan et al., 2022) has highlighted limitations in its ability to handle high-frequency time series data. In contrast, foundation models leverage Transformer architectures to capture complex, non-linear patterns and long-range dependencies, demonstrating particular efficacy in multivariate and large-scale forecasting tasks. However, these models require substantial computational resources and large datasets for optimal performance.

In order to evaluate the effectiveness and robustness of the forecasting models, this study adopts a dual forecasting window strategy. The initial period under consideration encompasses the months of March and April of 2024. It is used to evaluate model performance under consistent patterns and lower external variability. The second window covers the months of November and December 2024. In contrast, it corresponds to a period of significant volatility, characterised by promotions and holidays. The study aims to assess not only their accuracy but also their capacity to generalise across contrasting demand scenarios. To this end, the models are tested in both stable and volatile environments.

The evaluation criteria for this study have been defined based on established research. Forecasting accuracy is measured using two error metrics, namely MAPE and RMSE. The computational efficiency of the models is assessed through an analysis of training and inference time, accounting for the computational demands of foundation models compared to traditional forecasting techniques. Scalability is examined by testing model performance across datasets of varying sizes, thereby determining adaptability to different forecasting environments. Additionally, preprocessing requirements are analysed to measure the extent

of data transformations necessary for each model, given that foundation models often require significant data transformations.

By integrating these factors, the conceptual model provides a structured framework to evaluate the effectiveness of foundation models in comparison to classical forecasting techniques and Prophet. The present study aims to contribute to the ongoing discussion on the applicability, efficiency, and accuracy of foundation models in real-world forecasting scenarios. The findings will offer valuable insights into the practical adoption of these models, highlighting the trade-offs that practitioners must consider in selecting an optimal forecasting approach.

4. EMPIRICAL METHODOLOGY

4.1. DATA COLLECTION INSTRUMENT

This study employs secondary data, specifically weekly sales data in Portugal. The dataset was provided by GfK, a globally recognised market research company known for its expertise in retail market analytics. GfK's proprietary retail tracking system collects and aggregates sales data from a diverse range of retailers and distributors, ensuring accuracy and comprehensive market coverage.

The period covered by the dataset extends from 2022 to 2024, providing a detailed perspective on sales trends and seasonality effects relevant to forecasting. The timeframe was deliberately selected to exclude the anomalous market fluctuations caused by the COVID-19 pandemic, thereby ensuring a more stable and representative dataset for analysis.

It should be noted that, due to the proprietary nature of the dataset, granular details such as geographic segmentation, brand differentiation, and product-level insights are unavailable. Consequently, the dataset is univariate, encompassing key variables, including weekly sales units, temporal indicators and event-based markers.

The structured and systematic nature of the dataset enhances its suitability for forecasting research, providing a solid foundation for assessing model accuracy and predictive performance. Overall, it ensures high-quality and relevant data for evaluating the effectiveness of different forecasting models.

4.2. EXPLORATORY DATA ANALYSIS

The dataset employed in this study comprises weekly sales data for three distinct products, namely camcorders, media tablets, and toys, sold in Portugal between 2022 and 2024. The dataset is complete, containing no missing values, and consists of 156 weekly observations (52 per year). It encompasses five key variables: sales units, product name, the first day of the week within the year, and binary indicators for holidays and promotions. The primary variable of interest in this analysis is the number of units sold.

The binary variable for holidays was set to 1 only for the week preceding Christmas and the week of Christmas, as these periods are typically associated with increased consumer activity. In a similar manner, the promotions variable was assigned a value of 1 exclusively during the weeks of Black Friday and Cyber Monday, which are recognised as major retail events. For all other weeks, both variables were set to 0. This approach enables for a focused assessment of the influence of key seasonal and promotional effects on sales dynamics.

In order to gain an initial understanding of the dataset, summary statistics were computed for each product. The following table presents the key descriptive statistics, including mean sales, standard deviation, minimum, and maximum sales units recorded over the study period.

Table 4.1 – Descriptive statistics of the three products

Product	Mean Sales	Standard Deviation	Minimum Sales	Maximum Sales
Camcorders	496.24	201.01	237	1199
Media Tablets	5255.46	2821.86	2923	18809
Toys	315810.44	231770.36	115552	1287073

The mean sales figures demonstrate significant variation across the three products, reflecting differences in consumer demand and market behaviour. A clear hierarchy emerges when comparing the products in terms of both sales volume and volatility.

The mean sales volume and standard deviation of camcorders are indicative of limited variability and relatively stable consumer demand. The consistency of the sales pattern is further reinforced by the narrow range between the minimum and maximum sales. In contrast, media tablets exhibited considerably higher mean sales, accompanied by greater variability, suggesting responsiveness to external factors such as product releases or promotional activities. Toys, which exhibit an exceptionally high mean and standard deviation, demonstrate the greatest volatility. Their broad sales range reflects significant fluctuations likely driven by seasonality and promotional periods.

To further explore the characteristics of the dataset, time series plots were generated to depict the weekly sales patterns of each product throughout the study period. These plots reveal potential seasonal trends and recurring fluctuations in demand.

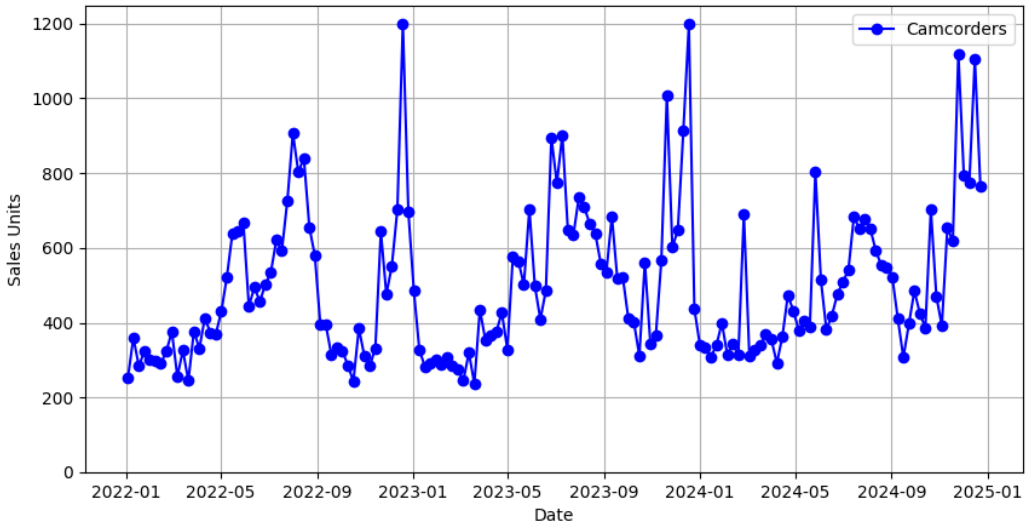

Figure 4.1 – Camcorders sales

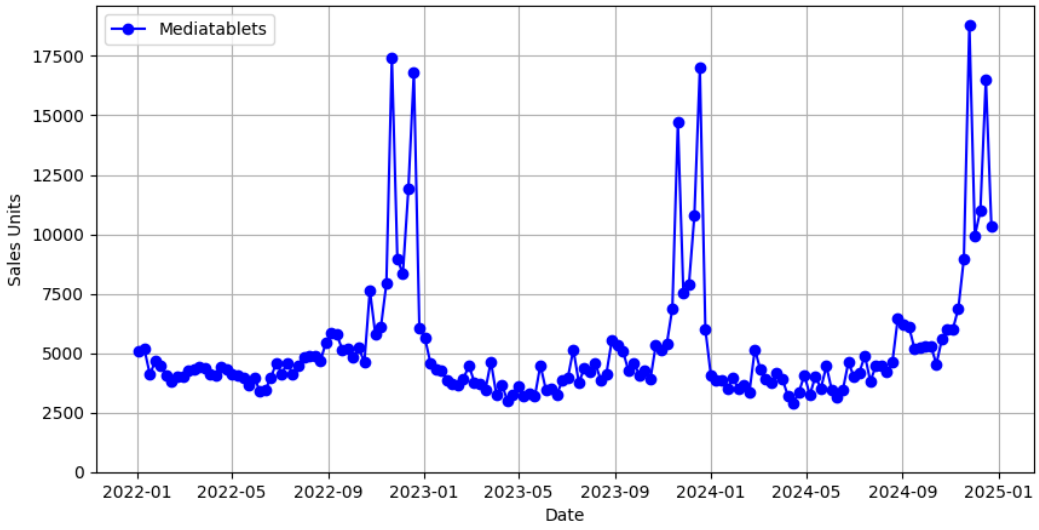

Figure 4.2 – Media Tablets sales

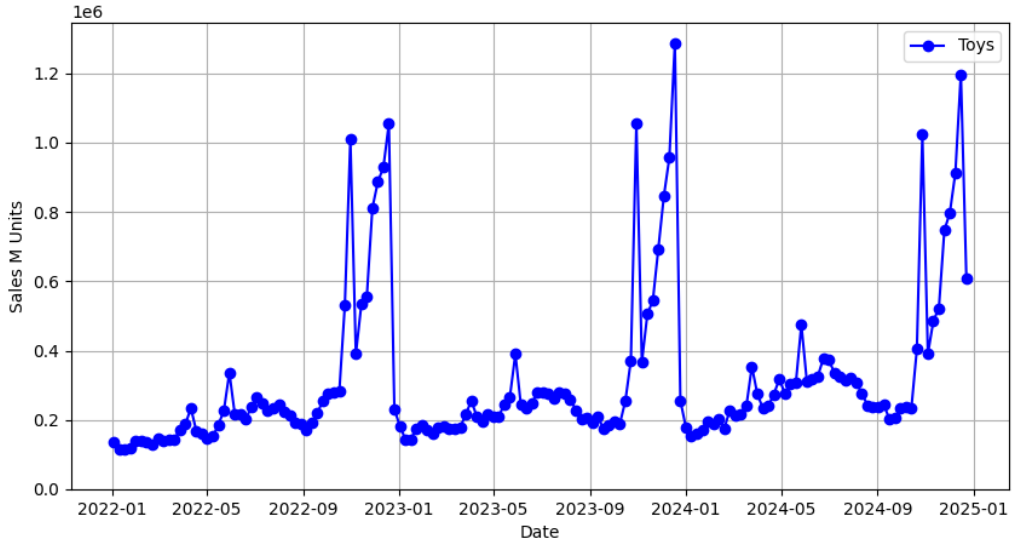

Figure 4.3 – Toys sales

A noticeable pattern observed for all three products is the presence of seasonality, characterised by recurring peaks in sales, particularly towards the end of the year. These periodic surges are likely associated with major shopping events such as Black Friday, Cyber Monday, and Christmas, which significantly influence consumer purchasing decisions.

Despite the presence of fluctuations and periodic spikes in sales data, these variations can be attributed to seasonal effects and promotional activities rather than to an upward or downward trend. Consequently, the overall sales trajectory exhibits stability over time, indicating that the series is stationary during the study period.

A time series is considered stationary if its statistical properties, such as the mean and standard deviation, remain constant over time (Box & Jenkins, 1970). This characteristic is crucial because a stationary time series is more likely to exhibit consistent patterns, thereby facilitating the modelling and prediction of future values (Ensafi et al., 2022).

The stationarity of a time series can be evaluated through various methods, including the Dickey-Fuller test and the analysis of rolling statistics plots. Figures 4.4, 4.5 and 4.6 illustrate the mean and standard deviation of sales over time, showing that these values remain relatively constant. The rolling statistics were calculated using a 52-period window, corresponding to one year of weekly data. This observation provides an initial indication that the time series under analysis may be stationary.

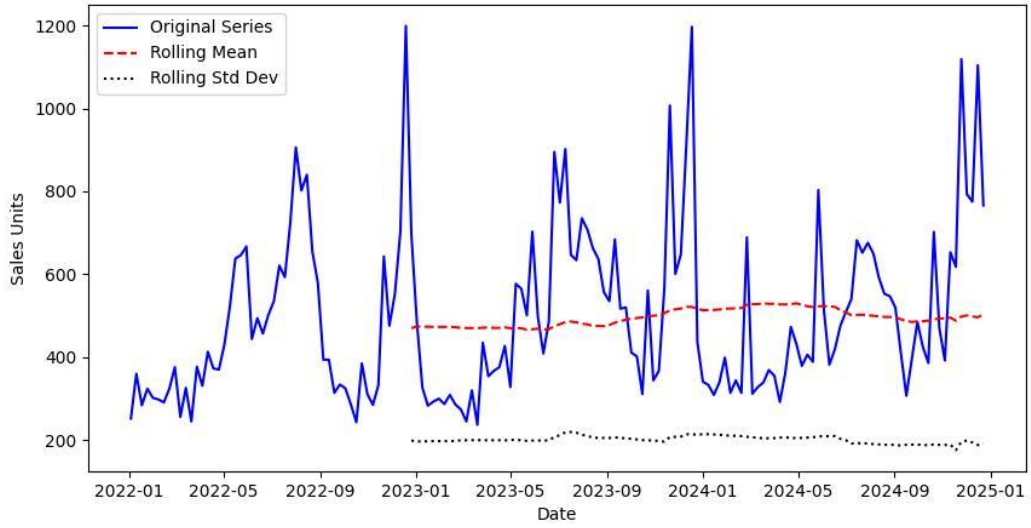

Figure 4.4 – Rolling statistics for Camcorders time series

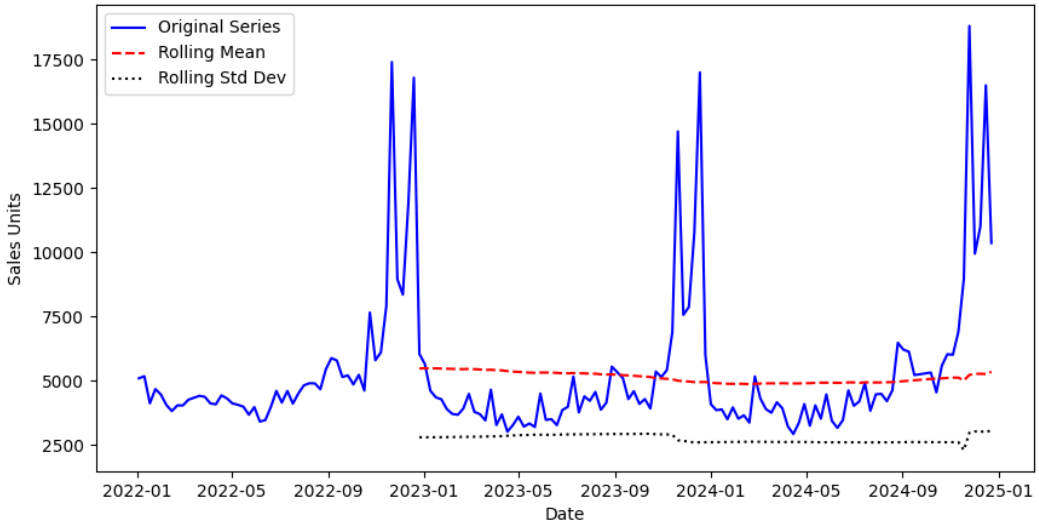

Figure 4.5 – Rolling statistics for Media Tablets time series

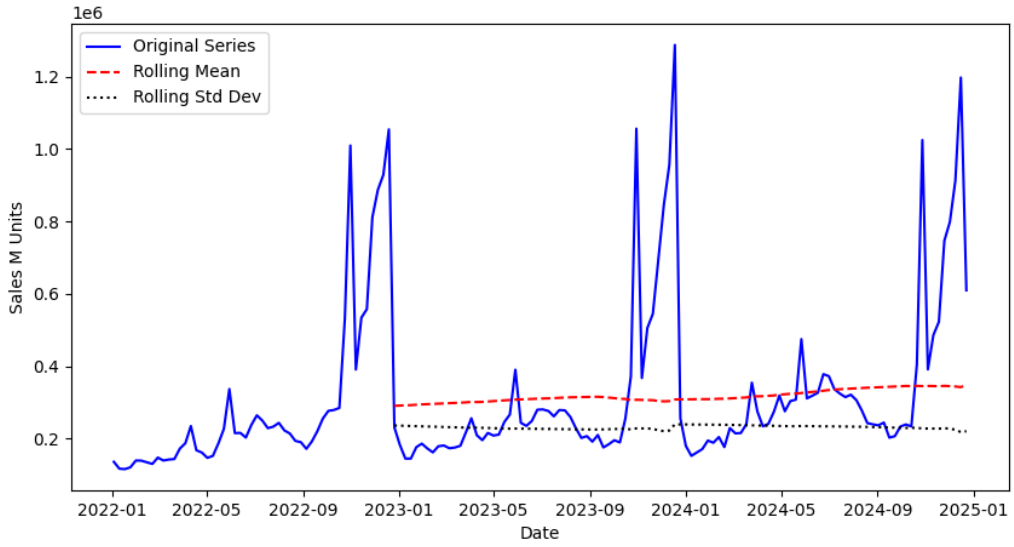

Figure 4.6 – Rolling statistics for Toys time series

The Augmented Dickey-Fuller (ADF) test is a widely used method to determine whether a time series is stationary or non-stationary by detecting the presence of a unit root. The test involves the comparison of the level of the series with its mean. The null hypothesis of the ADF test is that a unit root is present and, consequently, the time series is non-stationary. However, it is important to note that differencing the data can render the series stationary. Conversely, if the null hypothesis is rejected, it suggests that the time series is stationary, with no unit root present (Said & Dickey, 1984).

For the present analysis, the ADF test was conducted using Python (Python Software Foundation, 2020), which outputs both a test statistic and a p-value. A p-value greater than 0.05 (for a 95% confidence interval) indicates that the null hypothesis cannot be rejected, suggesting that the time series contains a unit root and is non-stationary. Conversely, a p-value smaller than 0.05 implies that the null hypothesis is rejected, indicating that the series is stationary. The results of the ADF test applied to the three products (see Table 4.2) demonstrate that the p-values are smaller than 0.05, thereby confirming that the time series for all three products are stationary and that no further data transformations are necessary.

Table 4.2 – p-values of the ADF test

Product	**p-value**
Camcorders	0.0009
Media Tablets	0.0054

Toys	0.0016

4.3. DATA SPLIT

The evaluation of the accuracy of a forecast requires the model to be tested on new, unseen data that was not used during the training process. A common approach in machine learning is to divide the dataset into a training set and a test set to assess the performance of the model.

However, in the context of time series forecasting, the data cannot be split randomly because the observations are time dependent. Unlike machine learning techniques, where methods such as k-fold cross-validation are widely used, time series models must be trained on past data and tested on future data, thereby preserving the natural temporal order (Ensafi et al., 2022).

The size of the training set has been shown to have a critical role in forecasting accuracy. A larger training set allows the model to capture patterns more effectively, leading to more accurate predictions (Medar et al., 2017).

The dataset under consideration in this study encompasses 156 weeks of sales data, extending from 2022 to 2024. In order to rigorously evaluate the performance of the model under different temporal conditions, this study implements two forecasting windows based on a consistent data splitting framework. The initial forecasting period encompasses March and April of 2024, representing a period of relative stability in sales patterns, although it is preceded by a slightly shorter training set. Conversely, the second forecasting window encompasses November and December of 2024, a seasonally volatile period, and benefits from a longer training history.

For each forecasting window, the data has been divided chronologically into three distinct subsets. The training set is used to fit the forecasting models and incorporates a period of more than two years of sales data in both windows, thereby establishing a robust foundation for the analysis of temporal structures, trends and seasonality. The validation test consists of a single week immediately following the training period and is employed exclusively for hyperparameter tuning and model selection, ensuring that the final evaluation remains unbiased. Finally, the test set comprises an eight-week period following the validation week and is used for the evaluation of the generalisability and predictive accuracy of the final model on new, unseen data.

The specific splits for each window are as follows.

For the first forecasting window, which covers March and April 2024:

- Training Set: 112 weeks, from January 2022 to February 2024

- Validation Set: 1 week, corresponding to the last week of February 2024
- Test Set: 8 weeks, from March to April 2024

For the second forecasting window, corresponding to November and December 2024:

- Training Set: 147 weeks, from January 2022 to October 2024
- Validation Set: 1 week, corresponding to the last week of October 2024
- Test Set: 8 weeks, from November to December 2024

This temporal segmentation of the data was deliberately designed to balance two competing objectives: maximising the training period to enhance model learning and preserving temporally distinct test periods that reflect varying levels of forecasting complexity. By implementing the identical structured split across both forecasting periods, the study guarantees that performance comparisons remain valid, controlled, and directly attributable to model capabilities rather than differences in data segmentation.

4.4. FORECASTING MODELS

4.4.1. CLASSICAL TIME-SERIES FORECASTING MODELS

The study included two well-established forecasting approaches: ARIMA/SARIMA and Holt-Winters Exponential Smoothing. These classical methods were selected for their interpretability, theoretical grounding, and longstanding use in forecasting applications. Their inclusion provides a meaningful benchmark against which to assess the added value of more complex models in terms of accuracy and flexibility.

4.4.1.1. ARIMA AND SARIMA

In the implementation of the $ARIMA(p, d, q)$, it is essential to determine the optimal values for the three parameters, autoregressive order (p), differencing order (d), and moving average order (q), to ensure accurate forecasting. A similar principle applies to the SARIMA model, denoted as $ARIMA(p, d, q)(P, D, Q, s)$, which extends the ARIMA framework by incorporating seasonal components, where s denotes the length of the seasonal period. In this study, s is set to 52, corresponding to weekly data with annual seasonality. In cases where seasonality is not detected, the seasonal components (P, D, Q, s) are set to zero, thereby reducing the model to a standard ARIMA model.

There are two main approaches to parameter selection. The first involves a visual analysis of the ACF and PACF plots for each time series. The value of p is inferred from the PACF plot, while q is guided by the ACF plot. The differencing parameter d reflects the number of times the data must be differenced to achieve stationarity. Since the time series in the study were found to be stationary, $d = 0$ was initially considered.

However, it is acknowledged that manual tuning based solely on ACF and PACF may not lead to the most accurate model. To address this limitation, a hyperparameter tuning algorithm

was employed using the *auto_arima* function from the *pmdarima* Python library. This method automates the search for optimal combinations of both non-seasonal (p, d, q) and seasonal (P, D, Q) parameters by evaluating model performance using information criteria such as the AIC (Box & Jenkins, 1970).

In order to ensure a fair comparison with more advanced forecasting models, such as Prophet and foundation models, the SARIMA model was extended to include exogenous variables. The external demand fluctuations linked to calendar effects and promotional activity were accounted for by the introduction of two binary variables, namely "holiday" and "promotion". These variables were passed to the model via the *exogenous* argument in the *auto_arima* function during the training process.

The SARIMA models were fitted separately for each of the two forecasting windows previously described. This design enabled the assessment of model performance in both stable and volatile conditions. Subsequent to the selection of optimal parameters based on validation RMSE, each model was retrained on the full training and validation set. Thereafter, the models were to generate eight-week forecasts on the test data.

The optimal parameters for the selected SARIMA configurations, as identified by the *auto_arima* function, are summarised in Tables 4.3 and 4.4, corresponding to the first and second forecasting windows, respectively.

Table 4.3 – SARIMA hyperparameters for March-April 2024

Product	**p**	**d**	**q**	**P**	**D**	**Q**
Camcorders	1	0	2	1	0	0
Media Tablets	1	0	0	0	0	0
Toys	1	0	0	0	1	0

Table 4.4 – SARIMA hyperparameters for November-December 2024

Product	**p**	**d**	**q**	**P**	**D**	**Q**
Camcorders	1	0	0	1	0	1
Media Tablets	4	0	2	1	0	1
Toys	1	0	0	0	1	0

4.4.1.2. HOLT-WINTERS EXPONENTIAL SMOOTHING

In view of the seasonal fluctuations observed in the weekly sales series, the study employed the Holt-Winters exponential smoothing method to decompose each series into its level, trend, and seasonal components. In this approach, the level, trend and seasonal factors are updated recursively at each time point by weighting the most recent observations through smoothing parameters α, β, γ, respectively each constrained to lie between zero and one. A seasonal period of 52 was specified a priori to reflect the yearly seasonality present in the weekly data. Furthermore, both trend and seasonality were permitted to assume additive, multiplicative or null forms, thereby yielding nine possible model structures.

To determine the most suitable model configuration for each product, every combination of trend (additive, multiplicative, or none) and seasonality (additive, multiplicative, or none) was fitted to the training data. The estimation of smoothing parameters was conducted through maximum likelihood estimation, leveraging the *ExponentialSmoothing* implementation within the *statsmodels* library (Python Software Foundation, 2020).

For each forecasting window, the performance of the model was evaluated on the corresponding validation set using the RMSE metric. The configuration that demonstrated the lowest validation RMSE was then refitted on the combined training and validation sample and used to produce forecasts for the respective test period.

In the initial forecasting period (March-April 2024), the selected model configurations were as follows: for camcorders, an additive trend and multiplicative seasonality; for media tablets, an additive trend and multiplicative seasonality; and for toys, a multiplicative trend and additive seasonality. In the subsequent forecasting period (November-December 2024), the model configurations employed for camcorders and toys were maintained, while for media tablets, the optimal structure was identified as a multiplicative trend and multiplicative seasonality.

This systematic model selection procedure ensures that each product's distinctive growth trajectory and seasonal pattern are modelled in an optimally tailored manner, thereby enhancing the accuracy of out-of-sample forecast.

4.4.2. PROPHET METHOD

Prophet, developed by Facebook, is an additive time series forecasting model designed to capture complex seasonal and trend patterns while remaining interpretable and user-friendly (Taylor & Letham, 2017). It was selected in this study for its flexibility in incorporating external regressors such as holidays and promotions, factors particularly relevant to retail sales forecasting.

The modelling procedure began with the preparation of the data and the specification of the time series parameters, followed by the implementation of the selected model. Once the forecast has been generated, its accuracy is assessed using RMSE and MAPE.

Prior to model fitting, the dataset must undergo reformatting. This involves the renaming of the “week” column to “ds”, and the target variable (“sales units”) to “y”.

Within the Prophet framework, non-periodic behaviour is captured via two distinct trend components, while seasonal components are handled using a Fourier series. The seasonality can be configured as either additive or multiplicative, depending on the characteristics of the data. The model also supports the inclusion of external regressors and allows hyperparameter tuning to control the influence of different components (Taylor & Letham, 2017).

In order to ensure robust evaluation, Prophet was applied separately to each of the two forecasting windows previously described. The model was trained on the corresponding training data, for each window. It was then tested over an eight-week forecast horizon.

The present study compared three Prophet configurations. The first model incorporated only yearly seasonality. The second model added a holiday argument to capture known calendar effects. The third extended the specification further by including a binary promotion variable. The “holiday” and “promotion” variables were both introduced as external regressors, with associated prior scale parameters in order to regulate their influence on the model.

A grid search was performed to identify the optimal combination of hyperparameters for each product and forecasting window.

- $changepoint\_prior\_scale \in [0.01, 0.1, 0.5]$
- $seasonality\_prior\_scale \in [0.01, 0.1, 1.0, 10.0]$
- $seasonality\_mode \in [additive, multiplicative]$
- $yearly\_seasonality \in [5, 10, 15, 20]$
- $holiday\_prior\_scale, promotion\_prior\_scale \in [0.01, 0.1, 1.0, 10.0]$

The *changepoint_prior_scale* parameter controls the flexibility of the model in detecting trend shifts. It has been demonstrated that smaller values (e.g., 0.01) constrain the model to smoother trend changes. Conversely, larger values (e.g., 0.5) permit a greater frequency or abruptness of shifts in trend. These settings enable the model to adapt to structural changes in the sales time series without overfitting. The *seasonality_prior_scale* parameter regulates how strongly the model incorporates seasonal components. Lower values result in smoother seasonal curves, while higher values enable the model to accommodate seasonality with greater intensity. The *seasonality_mode* delineates the manner in which seasonal effects interact with the trend component. The additive mode assumes constant seasonal magnitude over time, while the multiplicative mode assumes the seasonality scales the level of the time series. The *yearly_seasonality* parameter is responsible for determining the number of Fourier terms employed in the modelling of yearly seasonality. Higher values enable the model to capture more complex and high-frequency seasonal fluctuations. The *holiday_prior_scale* and *promotion_prior_scale* parameters determine the influence of binary external regressors (holiday and promotion) on the forecast. Smaller values result in a reduction of the effect of the regressor towards zero, whilst larger values allow for greater freedom in explaining

variation in sales. This range enables the grid search in evaluating whether external events exert a modest or substantial impact on each product category.

In order to address the issues of overfitting and underfitting, Prophet provides a built-in cross-validation procedure (Taylor & Letham, 2017). The purpose of this procedure is to evaluate the performance of forecasts during the training process. In the present study, the performance of the model for each hyperparameter configuration was evaluated using this procedure. This rolling-origin cross-validation simulates multiple historical forecast scenarios by retraining the model on increasingly larger subsets of the data and generating forecasts over a fixed horizon. In this study the forecast horizon was set to 8 weeks to match the length of the test set. For each configuration the RMSE was computed across all forecast windows, and the model with the lowest average RMSE was selected as the optimal configuration for each product. This procedure ensured a robust assessment of model performance across varying historical conditions.

Tables 4.5, 4.6 and 4.7 present the chosen configuration for each product across the three Prophet variants.

Table 4.5 – Optimal Prophet hyperparameter configurations for Camcorders across forecasting windows and model variants

Model	**Hyperparameters**	**Parameter (March-April 2024)**	**Parameter (November-December 2024)**
Prophet	*changepoint_prior_scale*	0.01	0.5
	seasonality_prior_scale	0.1	0.1
	seasonality_mode	additive	additive
	yearly_seasonality	5	5
Prophet	*changepoint_prior_scale*	0.5	0.5
+ holiday	*seasonality_prior_scale*	1.0	0.1
	seasonality_mode	multiplicative	additive
	yearly_seasonality	10	20
	holiday_prior_scale	1.0	0.01
Prophet	*changepoint_prior_scale*	0.01	0.5
+ holiday	*seasonality_prior_scale*	10.0	1.0

+ promotion	*seasonality_mode*	multiplicative	additive
	yearly_seasonality	10	20
	holiday_prior_scale	0.1	10.0
	promotion_prior_scale	1.0	1.0

Table 4.6 – Optimal Prophet hyperparameter configurations for Media Tablets across forecasting windows and model variants

Product	**Hyperparameters**	**Parameter (March-April 2024)**	**Parameter (November-December 2024)**
Prophet	*changepoint_prior_scale*	0.01	0.5
	seasonality_prior_scale	0.1	10.0
	seasonality_mode	additive	additive
	yearly_seasonality	15	15
Prophet	*changepoint_prior_scale*	0.1	0.5
+ holiday	*seasonality_prior_scale*	10.0	10.0
	seasonality_mode	additive	additive
	yearly_seasonality	5	15
	holiday_prior_scale	1.0	0.01
Prophet	*changepoint_prior_scale*	0.1	0.5
+ holiday	*seasonality_prior_scale*	1.0	0.1
+ promotion	*seasonality_mode*	additive	additive
	yearly_seasonality	15	15
	holiday_prior_scale	0.1	0.01
	promotion_prior_scale	1.0	0.01

Table 4.7 – Optimal Prophet hyperparameter configurations for Toys across forecasting windows and model variants

Product	Hyperparameters	Parameter (March-April 2024)	Parameter (November-December 2024)
Prophet	*changepoint_prior_scale*	0.5	0.5
	seasonality_prior_scale	0.1	10.0
	seasonality_mode	multiplicative	additive
	yearly_seasonality	10	20
Prophet	*changepoint_prior_scale*	0.01	0.5
+ holiday	*seasonality_prior_scale*	10.0	0.01
	seasonality_mode	additive	additive
	yearly_seasonality	15	10
	holiday_prior_scale	0.01	0.01
Prophet	*changepoint_prior_scale*	0.01	0.1
+ holiday	*seasonality_prior_scale*	10.0	0.1
+ promotion	*seasonality_mode*	multiplicative	multiplicative
	yearly_seasonality	15	5
	holiday_prior_scale	1.0	1.0
	promotion_prior_scale	1.0	0.01

After hyperparameter selection and cross-validation, the best model for each product was refitted on the corresponding training and validation data, and the resulting forecasts were compared to the actual sales in the test set.

Following the implementation of each model and the generation of forecasts, the results are presented as a data frame. This data frame comprises, amongst other columns, *yhat*, which is the predicted value, *yhat_lower* and *yhat_upper*, which are the corresponding uncertainty levels (Ensafi et al., 2022).

4.4.3. FOUNDATION MODELS

In addition to the classical and Prophet models previously discussed, this study incorporates two foundation models, TimeGPT-1 and Moirai, that represent recent advances in transformer-based time series forecasting.

A key criterion for their selection was the nature of their pre-training data. Both models were pretrained on large-scale time series corpora that included sales data. This domain alignment increases the likelihood of strong zero-shot generalisation and reduces the need for model retraining or extensive customisation. The implementation and evaluation of each model are described in the following subsections.

4.4.3.1. TIMEGPT-1

The study employed TimeGPT-1, a transformer-based model developed by Nixtla. TimeGPT-1 is designed for zero-shot forecasting, meaning it can be applied to previously unseen time series without retraining. However, the model also supports optional fine-tuning, allowing for lightweight adaptation to the specific characteristics of a given dataset using a limited number of gradient update steps. This design offers the flexibility to balance generalisation with precision (Garza et al., 2024).

In the present study, the model was applied separately to each of the three product categories. The dataset was structured in accordance with the model's input requirements. Specifically, the "week" variable was converted to a datetime format and renamed as "ds", while the weekly sales data was renamed as "y", serving as the target variable. The study incorporated two binary exogenous variables, namely "holiday" and "promotion", in both historical and future periods to capture potential external effects on sales behaviour.

For each forecasting window previously delineated, the pretrained model was lightly fine-tuned on the respective training and validation data (113 weeks for the initial window; 148 weeks for the subsequent one). The model was tasked with forecasting sales for the following eight-week period, which corresponds to the test set.

TimeGPT-1 was implemented using the official Nixtla Python client. The model offers the benefit of limited dependence on extensive manual parameter tuning. The forecasting procedure entails the configuration of a limited number of parameters. These include the number of fine-tuning steps required to adapt the pre-trained model to the specific dataset, the desired prediction interval quantiles, and the selection of exogenous regressors. In this implementation, the model was fine-tuned for 20 steps. This choice was informed by Nixtla's practical guidance, which highlights the importance of experimenting with the number of fine-tuning steps to optimise model performance. While an increase in the number of steps can enhance accuracy, it also raises the risk of overfitting and increases training time (Fine-Tuning Tutorial TimeGPT, 2025). During the development of the model, alternative fine-tuning values such as 10, 30, and 50 steps were tested, but these yielded higher RMSE scores on the test

set. Consequently, 20 steps provided the optimal balance between adaptation to the dataset and generalisation to unseen data. The model was further configured to return prediction intervals with a 90% confidence level. In addition, the model incorporated the historical and future values of the holiday and promotion variables as exogenous regressors.

Following the forecasting process, the predicted values were merged with the actual sales figures from the test set to enable evaluation. This merging step ensured that each forecasted observation was aligned with its corresponding real value, thus allowing for a precise assessment of forecasting accuracy. The model's performance was then assessed using the RMSE and MAPE metrics.

4.4.3.2. MOIRAI

The Moirai model is a transformer-based architecture designed for zero-shot time series forecasting. Its core innovation lies in its ability to generalise across highly diverse time series without requiring model retraining, leveraging pretraining on the LOTSA, which includes over 27 billion observations across multiple domains (Woo et al., 2024).

In this study the model was implemented using the *uni2ts* Python library, with pretrained weights loaded via the *MoiraiModule* class (Redoulès, 2024/2025). Forecasts were generated independently for each of the three product time series under analysis.

Although Moirai does not require model training or fine-tuning on the target data, the performance of its forecasts is contingent a set of inference-time configuration parameters. These parameters determine how the model interprets and samples from the data during forecasting. A grid search was conducted to identify the optimal configuration for each product, based on RMSE performance on the test set. In order to ensure consistency, forecasts were generated for the two forecasting windows previously defined, employing the corresponding training and test data splits.

The following hyperparameters were considered:

- $model\_size \in [small, base]$
- $context\_length \in [52, 64, 96, 104, 128]$
- $patch\_size \in [16, 32]$
- $num\_samples \in [20, 50, 100]$

The model size parameter determined the architectural complexity of the network, with two options under consideration: a "small" variant with fewer parameters and reduced computational cost, and a "base" variant with greater capacity for capturing complex patterns (Woo et al., 2024). The context length specified the number of previous time steps (in weeks) that the model used to inform its forecasts. The context lengths of 52, 64, 96, 104, and 128 weeks were tested, thereby enabling the model to incorporate between one and two years of historical data. The patch size defined the granularity with which the model segmented the context window. The values of 16 and 32 were subjected to testing, corresponding to the

number of consecutive time steps that were grouped and processed jointly. Finally, the number of samples determines the number of forecast trajectories generated by the model during the prediction process. The values of 20, 50, and 100 samples were evaluated, with larger counts yielding more stable mean predictions but requiring greater computational effort.

It is important to note that the time series dataset provided to the model included holiday and promotion binary indicators as dynamic real-valued covariates. The aforementioned variables, the values of which are subject to weekly alteration and are known in advance for the forecast horizon, were incorporated into the model input across all configurations to account for exogenous factors influencing sales behaviour. The inclusion of these covariates ensured that the forecasts were conditioned not only on historical sales patterns but also on external events, thereby enhancing the contextual information available to the forecasting process.

Each configuration was evaluated using a prediction length of eight weeks, corresponding to the size of the test set. The training-test split followed the same procedure described previously, with the final eight observations of each forecasting window held out for evaluation. For each configuration under consideration, the RMSE was calculated based on the mean of the predicted values and the actual sales values within the designated test window.

The model that yielded the lowest RMSE for each product was selected as the best-performing configuration. Tables 4.8 and 4.9 provide a concise overview of the optimal hyperparameter settings that have been identified for each product across the two forecasting windows.

Table 4.8 – Optimal hyperparameters for the Moirai model (March-April 2024)

Product	**model_size**	**context_length**	**patch_size**	**num_samples**
Camcorders	base	96	16	20
Media Tablets	base	128	16	100
Toys	small	64	32	20

Table 4.9 – Optimal hyperparameters for the Moirai model (November-December 2024)

Product	**model_size**	**context_length**	**patch_size**	**num_samples**
Camcorders	small	104	16	100
Media Tablets	base	64	16	100
Toys	base	52	16	20

4.5. EVALUATION METRICS

As outlined in the previous section, the performance of each forecasting model was evaluated using the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE).

RMSE is employed to quantify the standard deviation of the prediction errors, providing a measure of the extent to which the predicted values align with the actual observations. This metric evaluates the accuracy of a model on unseen data by returning a numerical value representing the average deviation between the predicted and actual values.

RMSE is formally defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t)^2}, \tag{4.1}$$

where Y_t denotes the actual observed value at time t, F_t is the corresponding forecasted value, and n represents the total number of observations (Hyndman & Koehler, 2006).

RMSE assigns greater weight to larger errors, thereby penalising substantial deviations more heavily than smaller ones. This is achieved by squaring the individual errors prior to averaging and subsequently taking the square root. This characteristic is of particular value in retail contexts, where large forecasting errors, such as underestimating or overestimating demand, can lead to costly operational issues.

Furthermore, RMSE is a reliable performance metric due to its scale-dependence, meaning it is expressed in the same units as the target variable (e.g., sales units). This renders the model both intuitively interpretable and practically relevant when evaluating its performance (Hyndman & Koehler, 2006). As forecasting models are evaluated separately for each product, RMSE provides a consistent and objective basis for comparing their forecasting accuracy across models. This supports the identification of the best-performing model for each unique sales pattern.

In addition to the RMSE, MAPE was used to evaluate the relative accuracy of forecasts. MAPE expresses the prediction error as a percentage, thus making it unit-free and directly interpretable across products and forecasting periods with varying sales volumes.

MAPE is formally defined as follows:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{(Y_t - F_t)}{Y_t}\right| * 100 \tag{4.2}$$

The formula in question calculates the average absolute percentage deviation between predicted and actual values (Hyndman & Koehler, 2006). Although MAPE provides a clear and interpretable metric, it is susceptible to very small or zero actual values, which can result in disproportionately significant percentage errors. Consequently, the use of MAPE in

conjunction with RMSE is recommended to ensure a comprehensive evaluation of model performance.

5. RESULTS AND DISCUSSION

This section presents an evaluation of the forecasting performance of the models under consideration, namely SARIMA, Holt-Winters, Prophet, TimeGPT-1, and Moirai, across the three product categories: camcorders, media tablets, and toys. This analysis directly addresses the main research objectives by comparing model performance, identifying influential features such as holidays and promotions, and assessing the practical implementation of foundation models in a retail forecasting context.

Two distinct evaluation periods were considered: the first covered the months of March and April 2024, serving as a reference period characterised by relative stability in demand patterns; the second encompassed the months of November and December 2024, allowing for an assessment of model sensitivity to seasonal and promotional effects. The accuracy of the forecasts was assessed using the Root Mean Square Error (RMSE) and the Mean Absolute Percentage Error (MAPE), allowing for both individual and comparative evaluations. In both cases, lower values are indicative of superior predictive accuracy.

To complement the evaluation based on RMSE and MAPE values, a residuals analysis was conducted for the forecasts generated by the model that demonstrated the highest performance in each product category. Residuals, defined as the differences between actual observations and forecasted values, provide insight into potential systematic bias, temporal dependencies, or other indications of model miscalibration. Descriptive statistics and residual plots are presented to determine whether the forecast errors are randomly distributed and centred around zero, as would be expected from an effective forecasting model.

5.1. RMSE AND MAPE SUMMARY TABLES

To compare the performance of classical and advanced forecasting models, Tables 5.1 to 5.4 present the RMSE and MAPE values obtained for each model, product category, and forecasting period.

Table 5.1 – Summary of RMSE Performance on March-April 2024

Model	Camcorders	Media Tablets	Toys
SARIMA	123.22	497.98	60446.82
SARIMA + holiday	118.16	901.70	60446.18
SARIMA + holiday + promotion	129.33	1586.27	60442.18
Holt-Winters	335.53	651.91	43062.58
Prophet	103.69	561.82	178956.96

Prophet + holiday	108.55	633.68	40534.55
Prophet + holiday + promotion	107.50	579.39	66762.19
TimeGPT-1	**75.88**	947.97	89438.66
Moirai	96.82	**256.36**	**23512.33**

Table 5.2 – Summary of MAPE Performance on March-April 2024

Model	**Camcorders**	**Media Tablets**	**Toys**
SARIMA	29.80%	11.91%	17.79%
SARIMA + holiday	31.77%	21.43%	17.79%
SARIMA + holiday + promotion	35.05%	39.64%	17.79%
Holt-Winters	90.95%	15.96%	10.29%
Prophet	25.67%	13.06%	67.50%
Prophet + holiday	25.71%	13.63%	12.36%
Prophet + holiday + promotion	25.92%	13.71%	20.56%
TimeGPT-1	**20.85%**	24.36%	28.68%
Moirai	25.14%	**6.58%**	**7.44%**

Table 5.3 – Summary of RMSE Performance on November-December 2024

Model	**Camcorders**	**Media Tablets**	**Toys**
SARIMA	253.61	4878.74	133512.48
SARIMA + holiday	281.93	4339.63	133512.87
SARIMA + holiday + promotion	260.59	5248.69	133512.87
Holt-Winters	303.14	4870.39	161208.59

Prophet	238.39	3248.25	**131076.53**
Prophet + holiday	223.12	3246.69	344692.28
Prophet + holiday + promotion	239.49	3226.12	216725.88
TimeGPT-1	**199.09**	3553.17	140436.48
Moirai	228.83	**2785.70**	226915.80

Table 5.4 – Summary of MAPE Performance on November-December 2024

Model	**Camcorders**	**Media Tablets**	**Toys**
SARIMA	21.54%	26.32%	**12.40%**
SARIMA + holiday	27.08%	25.58%	12.40%
SARIMA + holiday + promotion	22.96%	41.05%	12.40%
Holt-Winters	29.20%	29.57%	12.54%
Prophet	19.65%	22.01%	17.59%
Prophet + holiday	21.46%	22.01%	36.03%
Prophet + holiday + promotion	25.60%	21.66%	33.34%
TimeGPT-1	**19.53%**	21.96%	15.44%
Moirai	20.56%	**20.56%**	34.45%

A preliminary analysis indicates that advanced forecasting methods outperform classical approaches across all three product categories in the March-April 2024 period. TimeGPT-1 attained the lowest RMSE and MAPE for camcorders, while Moirai achieved the best performance for both media tablets and toys, with the lowest RMSE and MAPE values in both categories. This renders these models the most accurate in both absolute and relative terms for their respective products.

In the more volatile November-December 2024 period, TimeGPT-1 again recorded the lowest RMSE and MAPE for camcorders, maintaining its performance under increased seasonal

variability. In the case of media tablets, the performance of Moirai was also maintained, with the model yielding the lowest RMSE and MAPE values. For toys, Prophet without regressors obtained the lowest RMSE, yet its MAPE exceeded that of the SARIMA, Holt-Winters and TimeGPT-1 models.

In order to interpret these RMSE values appropriately, it is essential to consider the scale and variability of product sales during each forecasting period. In March-April 2024, camcorders registered a mean sales volume of 474.77 units, with a standard deviation of 198.49. Conversely, media tablets had an average of 5,005.16 units, with a standard deviation of 2,548.85. Meanwhile, the mean value of toys was 288,132.92 units, with a standard deviation of 220,924.93. These figures provide a necessary reference for evaluating the relative magnitude of forecast errors. To illustrate this point, an RMSE of 75.88 units for camcorders corresponds to approximately 16 percent of their mean sales, while an RMSE of 23,512.33 for toys represents only 8.2 percent of the mean.

In contrast, the November-December 2024 period was distinguished by greater variability in sales, accompanied by a substantial increase in forecast errors across all models. During this period, camcorders recorded a mean of 496.24 units, with a standard deviation of 201.01; media tablets had a mean of 5,255.46 units and a standard deviation of 2,821.86; and toys attained a mean of 315,810.44 units, with a standard deviation of 231,770.36 units. It is notable that the forecast errors were notably higher during this period, with the most accurate models producing RMSE values of 199.09 for camcorders (40.1 percent of the mean), 2,785.70 for media tablets (53.0 percent), and 131,076.53 for toys (41.5 percent). These values indicate a marked deterioration in forecast accuracy.

For a clearer interpretation of the results and the factors that influence the accuracy of the model, each product category is analysed separately in the following sections.

5.2. CAMCORDERS

In this section, detailed outputs for individual model configurations are omitted for reasons of brevity and clarity, with emphasis placed on the comparative performance of key approaches across the two analysed periods. It is important to note that all forecast and residual plots refer to the March-April 2024 period and serve as representative illustrations. The corresponding visualisations for the November-December 2024 period are available in Appendix A.

The SARIMA models demonstrated persistent underperformance across both test windows. The model's overreliance on regular, repeating patterns limited their ability to capture the irregular and event-driven nature of camcorder sales. As evidenced by Figure 5.1, none of the SARIMA configurations for March-April effectively captured the short-term spike in early April. The incorporation of holiday and promotion regressors resulted in observable alterations to the forecast trajectories, most notably a flattening of the predicted sales pattern. However, these modifications did not result in enhanced alignment with the actual sales. Instead, the

forecasts exhibited a persistent structural misalignment with the observed fluctuations, suggesting that the regressors introduced additional noise. In a similar manner, during the period from November to December, all model variants, including those incorporating holiday and promotion regressors, exhibited a consistent tendency to underestimate peak sales. Performance differences across configurations were only marginal (see Figure A.1).

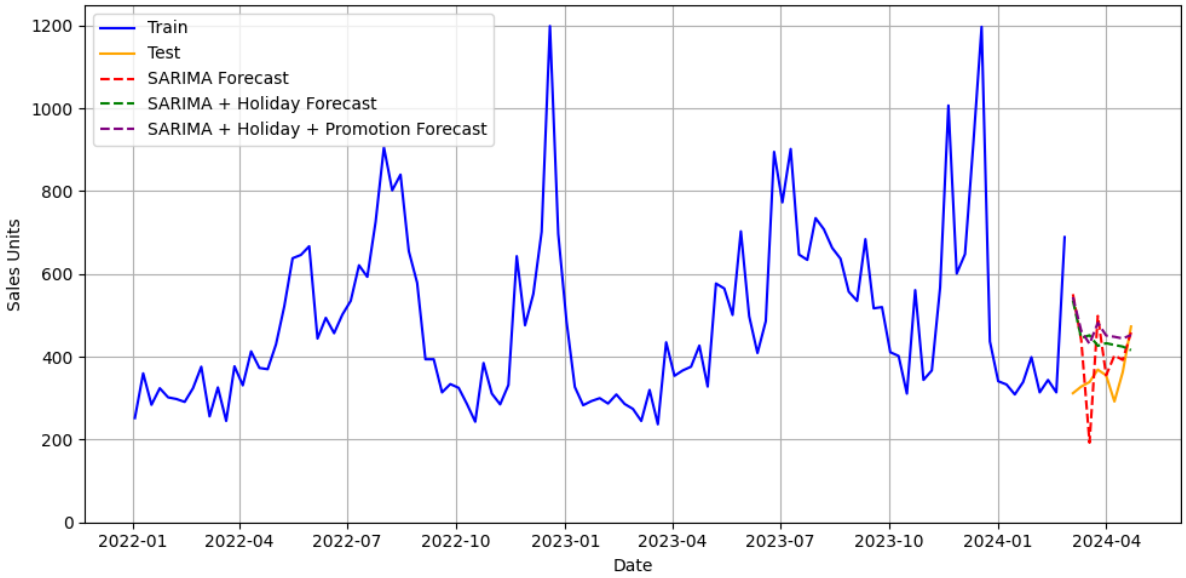

Figure 5.1 – Forecast by SARIMA configurations for camcorders on March-April 2024

The Holt-Winters model demonstrated the poorest overall performance. The model systematically overestimated sales and, during the November-December period, was unable to represent the abrupt shifts associated with promotional activity (see Figure A.2). The model's static structure, built around smooth seasonal and trend components, rendered it ineffective for capturing the complex dynamics observed in camcorder sales during both test periods. As demonstrated in Figure 5.2, the Holt-Winters forecast exhibits a substantial inaccuracy, with the model recording the highest error metrics among all models.

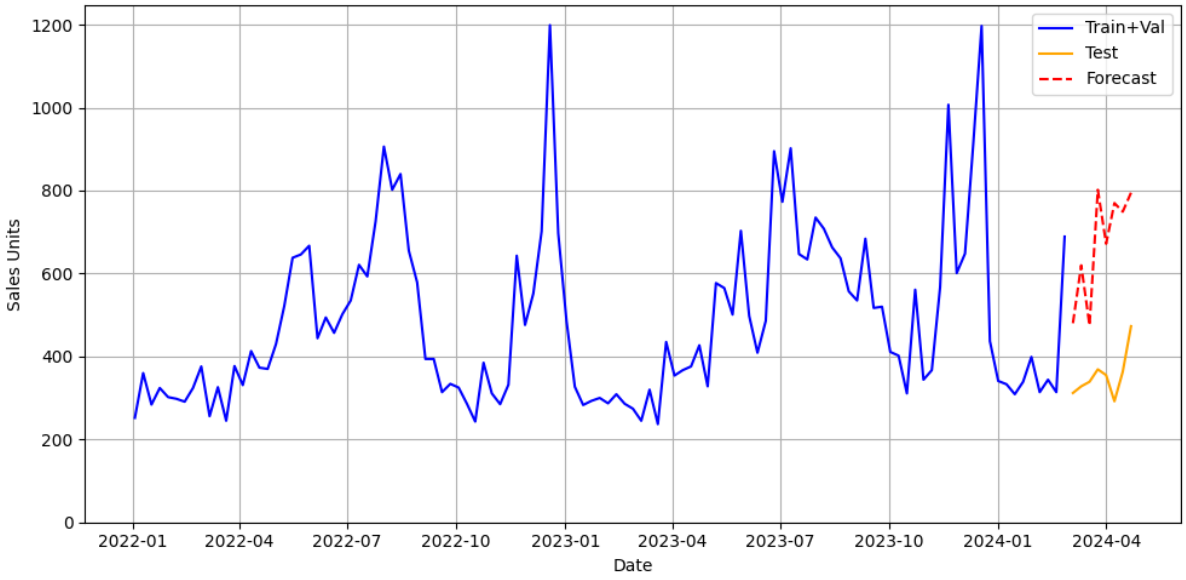

Figure 5.2 – Forecast by Holt-Winters for camcorders on March-April 2024

The Prophet model has been demonstrated to produce more stable and accurate forecasts, achieving low metrics in comparison to SARIMA and Holt-Winters. As illustrated in Figure 5.3, the baseline configuration of the model provided a reasonable approximation of sales patterns during the period of March to April. However, in the November-December period, the baseline configuration tended to underestimate sales volumes, thereby indicating a bias in capturing sharp sales surges, as illustrated in Figure A.3. The incorporation of holiday and promotion regressors yielded only marginal improvements, particularly in the November-December period, while concurrently introducing unnecessary volatility in the March-April period, without improving accuracy.

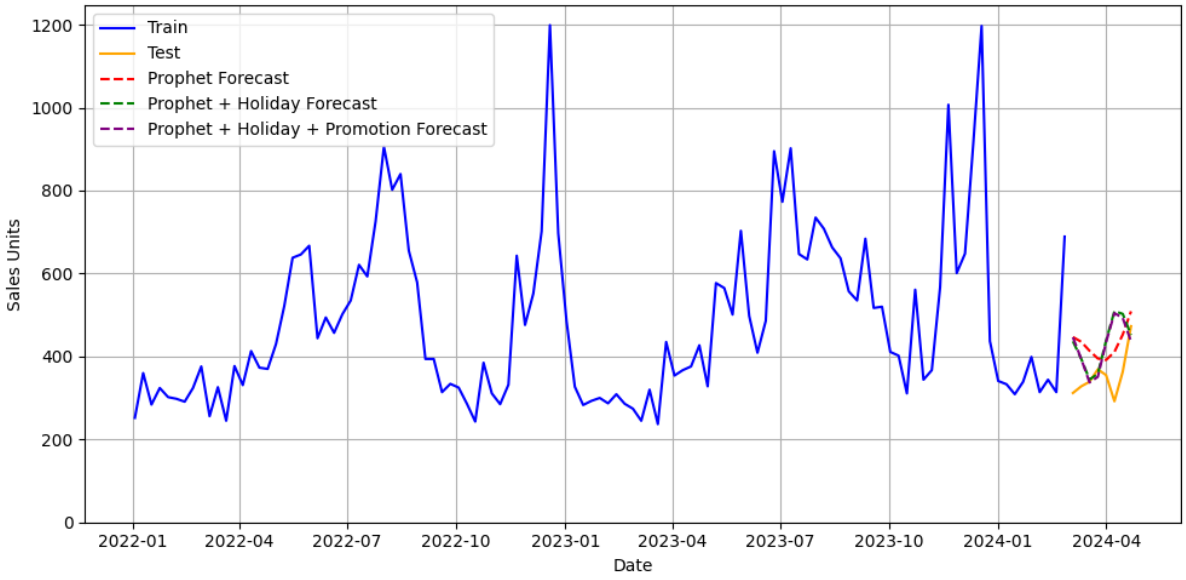

Figure 5.3 – Forecast by Prophet configurations for camcorders on March-April 2024

Among all models, the foundation models demonstrated the strongest performance.

TimeGPT-1 emerged as the most effective model across both periods, in both absolute and relative terms. The model consistently yielded the lowest forecasting errors and was successful in capturing the overall structure of sales trajectories.

In March-April 2024, the model demonstrated an alignment with actual sales levels, exhibiting a tendency towards overestimation, as illustrated in Figure 5.5. This overestimation bias is further confirmed in Figure 5.4, where residuals remain predominantly negative throughout the test window. The mean residual was approximately -63.7 units, indicating systematic bias in the forecasts. The residual plot also highlights periods of larger forecast errors, notably in early April, where sales dynamics were likely more difficult to predict.

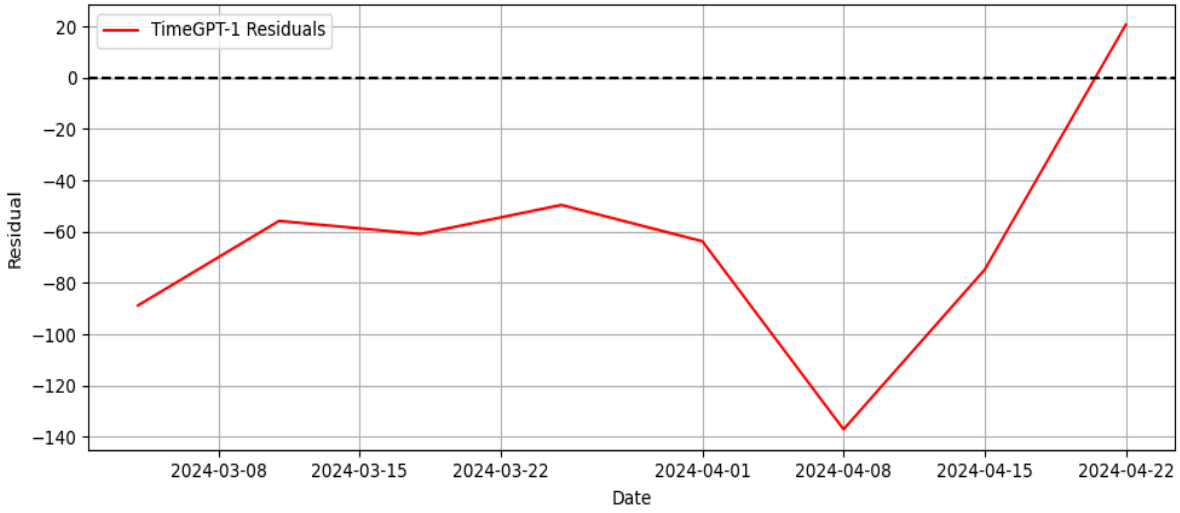

Figure 5.4 – Residuals for TimeGPT-1 forecast for camcorders on March-April 2024

In contrast, during the November-December period, TimeGPT-1 demonstrated a conservative bias, exhibiting a tendency to underestimate the magnitude of major sales peaks, as evidenced by a mean residual of 73.1 units (see Figure A.4). Nevertheless, residual variability remained moderate across both windows, thereby confirming the model's robustness in both stable and promotion-sensitive conditions.

Moirai also demonstrated a strong performance, closely trailing TimeGPT-1 in terms of accuracy. In March-April, the forecast effectively represented the sales trajectory but displayed greater variability. As demonstrated in Figure 5.5, Moirai's forecast curve provides a satisfactory approximation of the sales level, though with reduced precision around local turning points. During the more volatile November-December period, Moirai accurately captured the timing and magnitude of the initial demand peak. However, it completely failed to predict the subsequent surge, resulting in a decline in the forecast trajectory and reduced accuracy in the latter part of the test window.

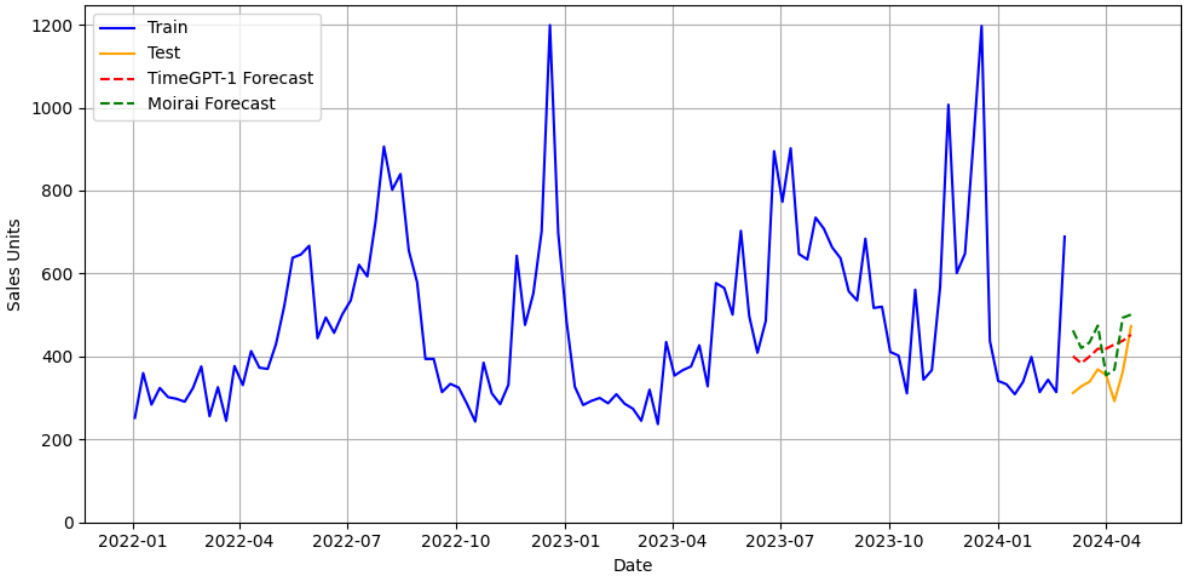

Figure 5.5 – Forecast by foundation models for camcorders on March-April 2024

5.3. MEDIA TABLETS

In order to emphasise the responsiveness of each model to volatile demand conditions, the visual analysis in this section focuses on the November-December 2024 period. The period under consideration is characterised by two pronounced peaks: the initial peak corresponds to the Black Friday and Cyber Monday promotional events, and the subsequent peak is associated with the Christmas holiday season. It is important to note that all forecast and residual plots presented below correspond to this period. The results for the more stable March-April 2024 window, including supporting figures, are provided in Appendix A.

The SARIMA models demonstrated restricted adaptability across the two periods analysed. In March-April 2024, the baseline SARIMA model exhibited a satisfactory capacity to replicate the general sales level, as demonstrated in Figure A.6. However, the inclusion of holiday and promotion regressors significantly impaired performance, indicating that external events played a negligible role in shaping sales dynamics during this timeframe. In contrast, in the November-December 2024 period, SARIMA proved incapable of capturing the irregular and promotion-driven peaks. As illustrated in Figure 5.6, the forecasts demonstrated excessive steepness and volatility, particularly when regressors were included, suggesting overfitting and a lack of responsiveness to rapidly changing sales conditions. In summary, the SARIMA model demonstrated a deficiency in its ability to effectively accommodate both stable and volatile periods.

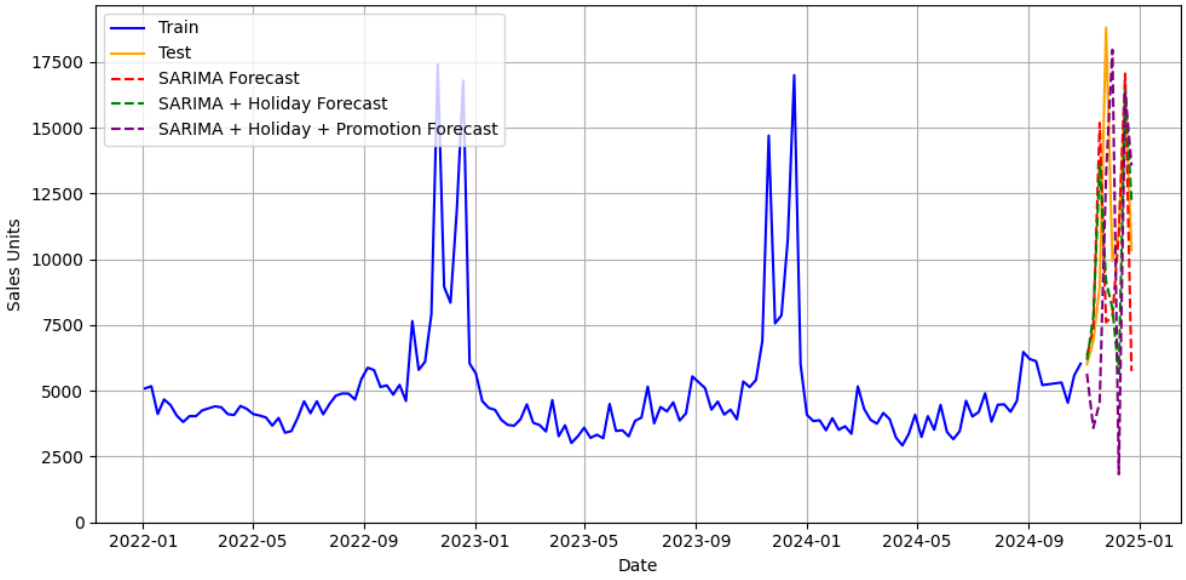

Figure 5.6 – Forecast by SARIMA configurations for media tablets on November-December 2024

The Holt-Winters model produced moderate results in March-April 2024, adequately reflecting the general level of demand but failing to account for short-term shifts (see Figure A.7). The structural simplicity of the model, coupled with its reliance on fixed seasonal and trend components, constrained its capacity to respond to variations in the data. The model demonstrated a substantial decline in performance during the November-December period, where the model consistently overestimated sales, particularly during promotional peaks. The model's rigid forecasting structure was unable to represent the abrupt, event-driven nature of demand, as evidenced in Figure 5.7, rendering it unsuitable for highly volatile periods.

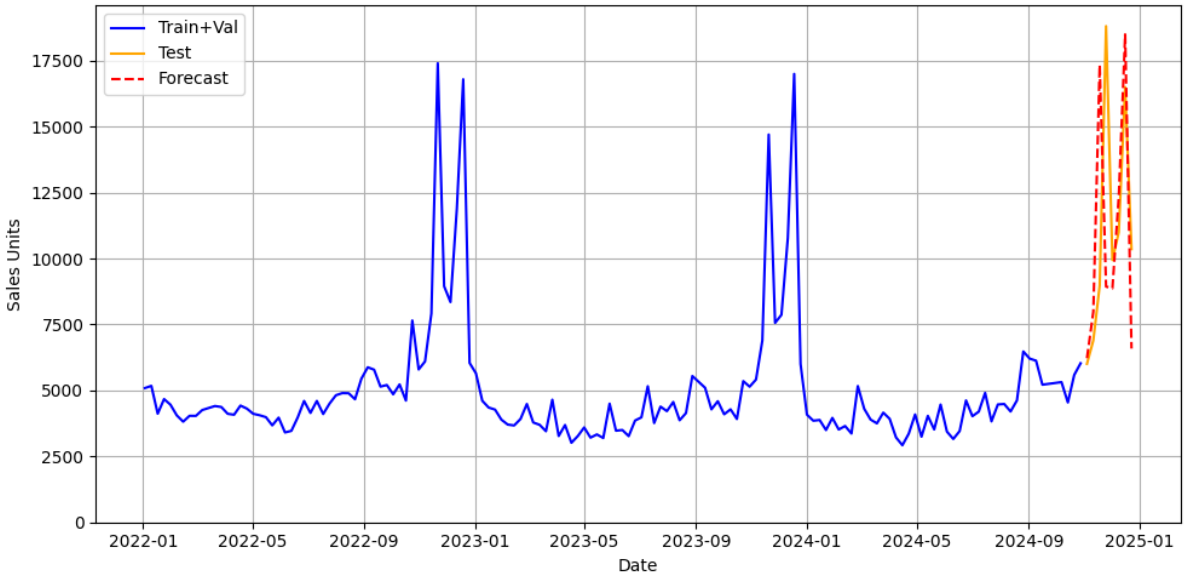

Figure 5.7 – Forecast by Holt-Winters for media tablets on November-December 2024

Prophet delivered stable forecasts across both periods. As demonstrated in Figure A.8, the baseline model captured the general sales pattern effectively during the period of March-April 2024. The implementation of regressors yielded only marginal variations in performance, with no significant gains observed. In November-December, the Prophet also demonstrated consistent behaviour, as illustrated in Figure 5.8. The impact of holiday effects was characterised by a marginal enhancement in responsiveness, without compromising stability. Conversely, the incorporation of promotion effects exerted a modest yet significant influence on forecast accuracy, with the holiday-plus-promotion configuration yielding the lowest RMSE among the Prophet variants.

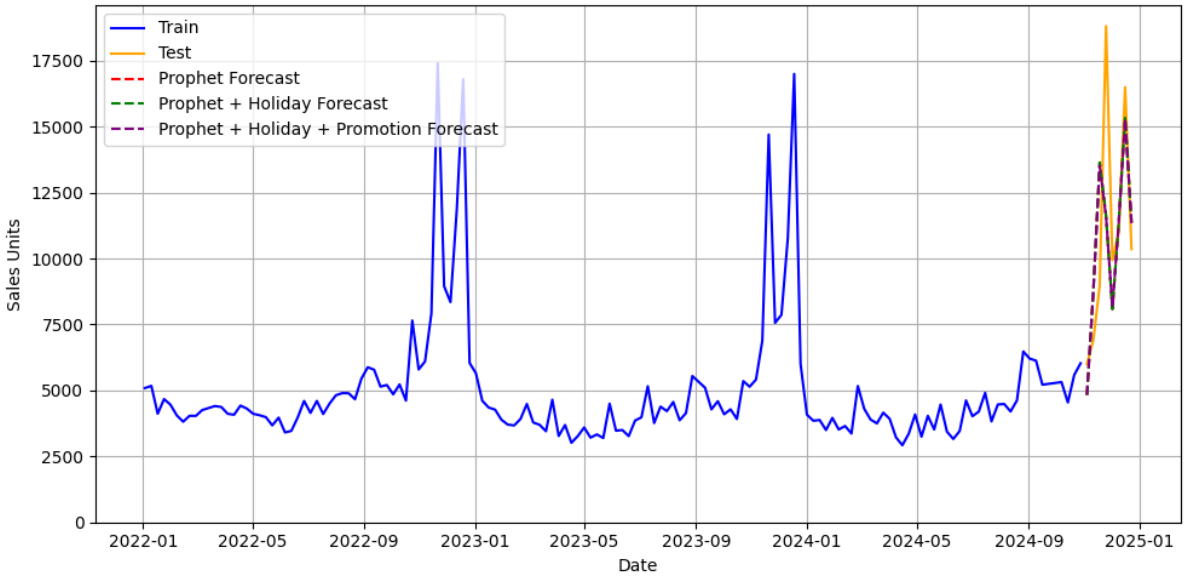

Figure 5.8 – Forecast by Prophet configurations for media tablets on November-December 2024

TimeGPT-1 exhibited divergent outcomes across the two periods. In March-April 2024, it demonstrated substandard performance in comparison to the alternative models. Furthermore, it demonstrated an inability to align with local fluctuations despite capturing the general direction of the series (see Figure A.9). However, there was a significant improvement in performance in November-December 2024, with the model successfully representing the overall trend and outperforming classical methods (SARIMA and Holt-Winters). However, the forecast, presented in Figure 5.9, exhibited a conservative bias and lacked the requisite sensitivity to adequately capture the intensity of promotional demand peaks.

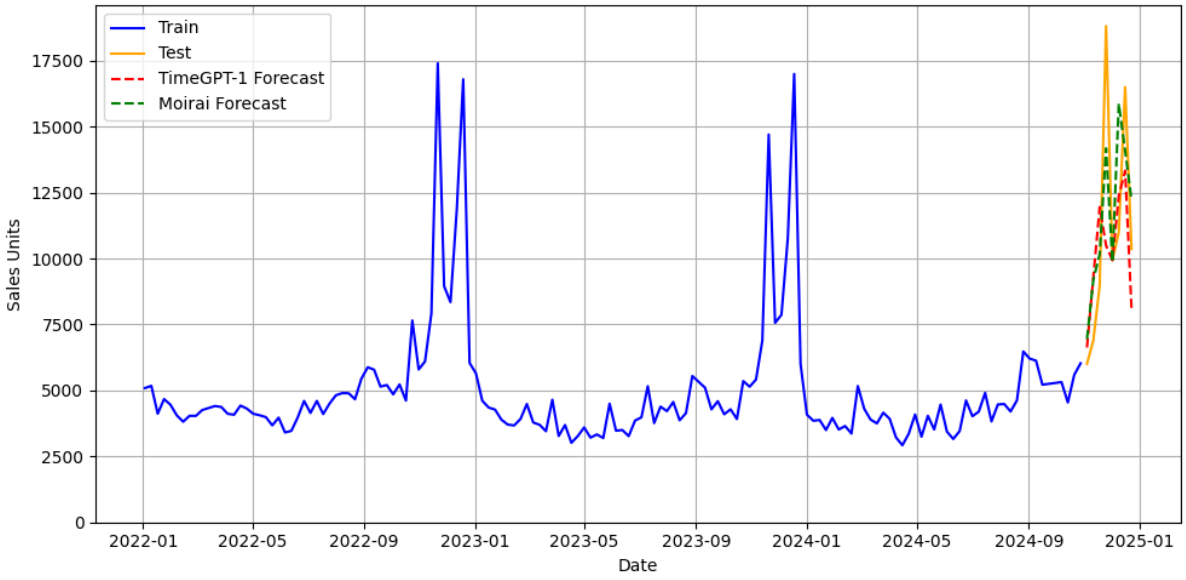

Figure 5.9 – Forecast by foundation models for media tablets on November-December 2024

Moirai was the most accurate and consistently effective model across both periods. In March-April 2024, the model achieved excellent alignment with actual sales, thereby effectively modelling the magnitude and timing of demand shifts (see Figure A.9). In the November-December period, Moirai once again demonstrated its superior performance by outperforming all other models. This was achieved by demonstrating a remarkable capacity to adjust to rapid fluctuations and high-volatility conditions. As depicted in Figure 5.9, Moirai closely followed the sales trajectory and offered more precise estimates of peak levels than TimeGPT-1.

The residual analysis further corroborates the robustness of Moirai's forecasts for both periods. In the March-April period, the residuals fluctuated moderately across the test period, with a mean residual of approximately -18.04 units, indicating a minor overestimation of sales. As illustrate in Figure A.10, the residual dispersion was moderate, with positive residuals early in the period, suggesting underprediction, followed by negative residuals later on, indicating overprediction. This shifting pattern is indicative of the model's adaptation to evolving sales dynamics, with no evidence of persistent bias.

As illustrated in Figure 5.10, for the November-December period, the residuals demonstrated significant fluctuations across the test window, with a mean of -504.79 units, indicating a tendency towards overestimation. The standard deviation of 2,928.74 units is indicative of the inherent volatility of sales. The residual trajectory demonstrated significant deviations around late November and early December, coinciding with peak demand periods. However, the model exhibited reasonable capacity to adjust to the pronounced reversals in sales.

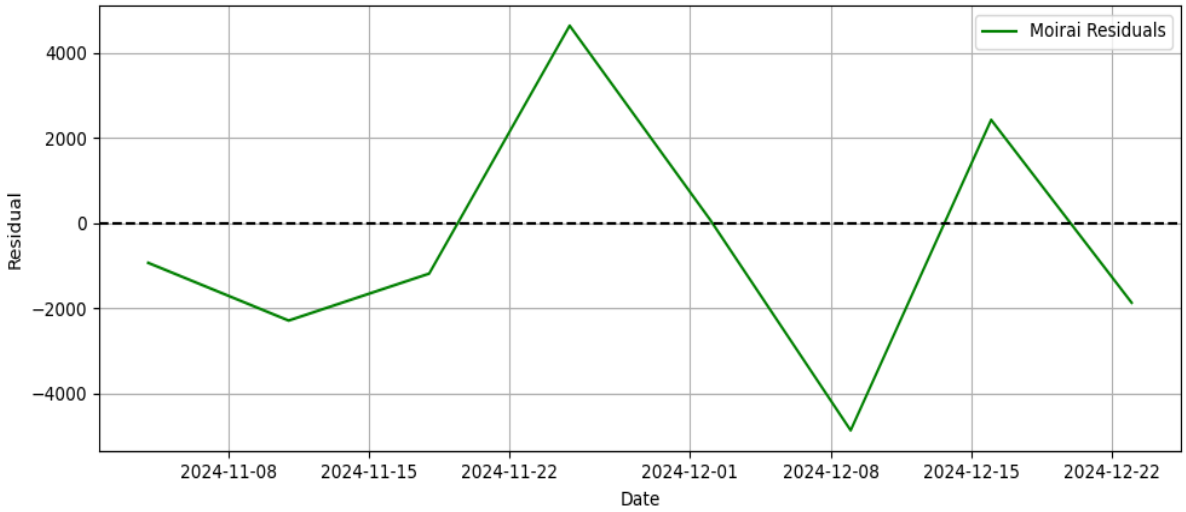

Figure 5.10 – Residuals for Moirai forecast for media tablets on November-December 2024

5.4. TOYS

The March-April 2024 period is utilised for visual analysis in this section, as the toys time series is already inherently volatile due to its high sales volume. By focusing on a comparatively more stable period, it is possible to illustrate the differences in model performance more clearly, without the additional complexity introduced by extreme demand surges. The complete set of forecasts and residuals for the November-December period is presented in Appendix A.

The SARIMA configurations produced stable and relatively consistent forecasts across both periods. In March-April 2024, the baseline configuration was found to adequately capture the general sales trend. However, it consistently underestimated peak values. The incorporation of holiday and promotion regressors exerted a negligible effect. This is clearly demonstrated in Figure 5.11, where there is an absence of distinct forecast lines for the three configurations, as their trajectories overlap. Similarly, during the November-December 2024 period, SARIMA attained the lowest MAPE among all models, thereby signifying its robust relative performance. However, all configurations underestimated the sharp peak in late December, as evidenced in Figure A.11, thereby highlighting the model's limitation in capturing extreme short-term surges driven by holiday demand. The residual analysis corroborated these findings. Across all configurations, the mean residual remained around 31,500 units, and the standard deviation was approximately 138,700 units, indicating large forecast errors consistent with the scale of sales, but without meaningful differences between model variants. The incorporation of external regressors yielded only a minimal increase in value, as the internal seasonal structure of SARIMA predominantly explained the sales dynamics.

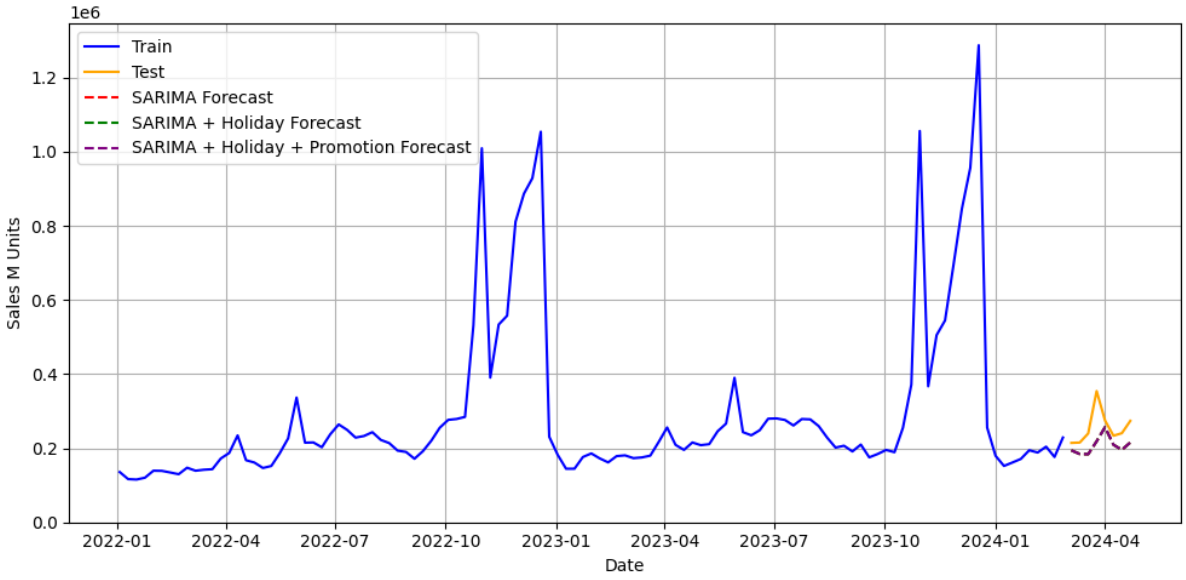

Figure 5.11 – Forecast by SARIMA configurations for toys on March-April 2024

Holt-Winters exhibited competitive results in both periods. In the period from March to April 2024, the model demonstrated robust performance in comparison with competing models. As demonstrated in Figure 5.12, the forecast closely adhered to the sales pattern and exhibited an effective capture of the peak timing. However, as would be anticipated with smoothing-based models, the peak magnitude was slightly attenuated. In the more volatile November-December period, Holt-Winters successfully captured the structure and timing of the demand peak but underestimated its magnitude (see Figure A.13). Despite this, the model demonstrated a high degree of relative accuracy, indicating its effectiveness in capturing seasonal patterns in both stable and volatile periods, particularly when historical sales trends are consistently repeated. However, it exhibited certain limitations in its ability to respond promptly to sudden fluctuations in demand.

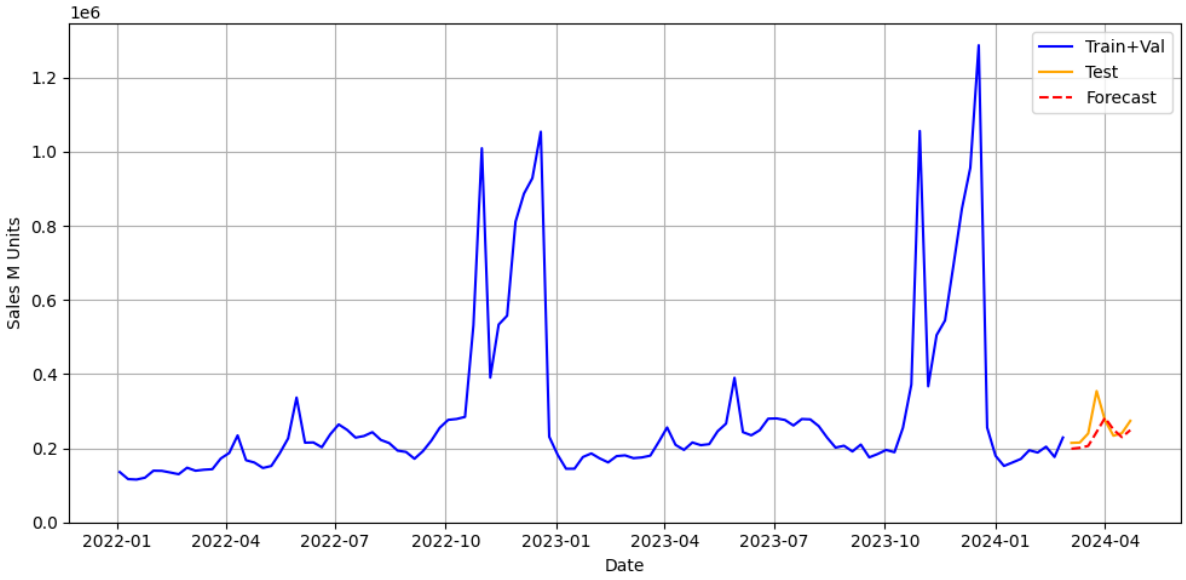

Figure 5.12 – Forecast by Holt-Winters for toys on March-April 2024

The Prophet produced mixed results. In March-April 2024, the baseline model demonstrated substandard performance, exhibiting a marked tendency to overestimate sales. The incorporation of the holiday regressor resulted in significant enhancements, underscoring the significance of holiday effects within this category and this period. The promotion regressor exerted a negligible additional influence, leading to an increase of both error metrics and resulting in the underestimation of the actual sales, as illustrated in Figure 5.13.

In contrast, during the November-December period, the baseline Prophet model attained the lowest RMSE of all configurations. However, this was accompanied by a higher MAPE, attributable to overestimation during weeks of minimal sales (see Figure A.14). This is evidenced by the mean residual of -48,873.22 units, which indicates a systematic tendency to forecast higher sales than those observed. The residual plot (Figure A.15) displays deviations below zero, particularly in early November and mid-December, where the model struggled to adjust to lower-than-expected sales. The incorporation of regressors into the Prophet variants resulted in suboptimal performance, characterised by elevated error levels and diminished stability. This finding indicates that while Prophet can capture the broader sales structure, it may exhibit overfitting when regressors are not aligned with observed demand dynamics.

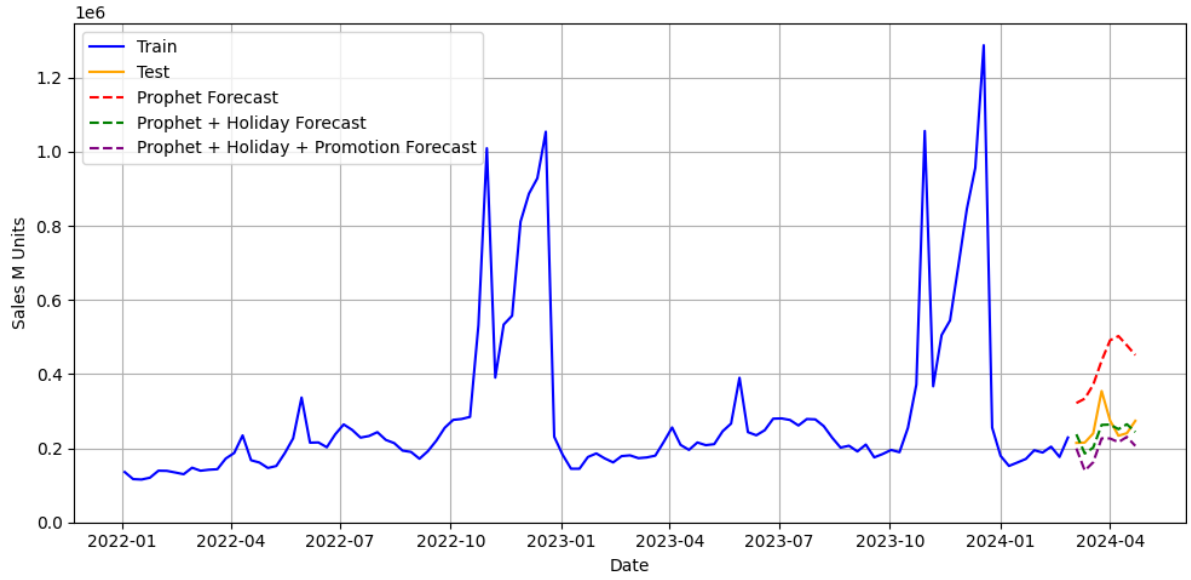

Figure 5.13 – Forecast by Prophet configurations for toys on March-April 2024

The performance of foundation models exhibited significant variation for toys during both periods.

TimeGPT-1 demonstrated substandard performance in March-April 2024, as evidenced by its inability to promptly adapt to fluctuations in sales, as illustrated in Figure 5.14. However, there was a substantial enhancement in performance during the November-December 2024 period. The model demonstrated an ability to capture the overall sales trend and timing of the

seasonal peak (see Figure A.16). However, it continued to underestimate the magnitude of the surge. TimeGPT-1 exhibited moderate adaptability and enhanced consistency in more volatile conditions, although it was outperformed by simpler models.

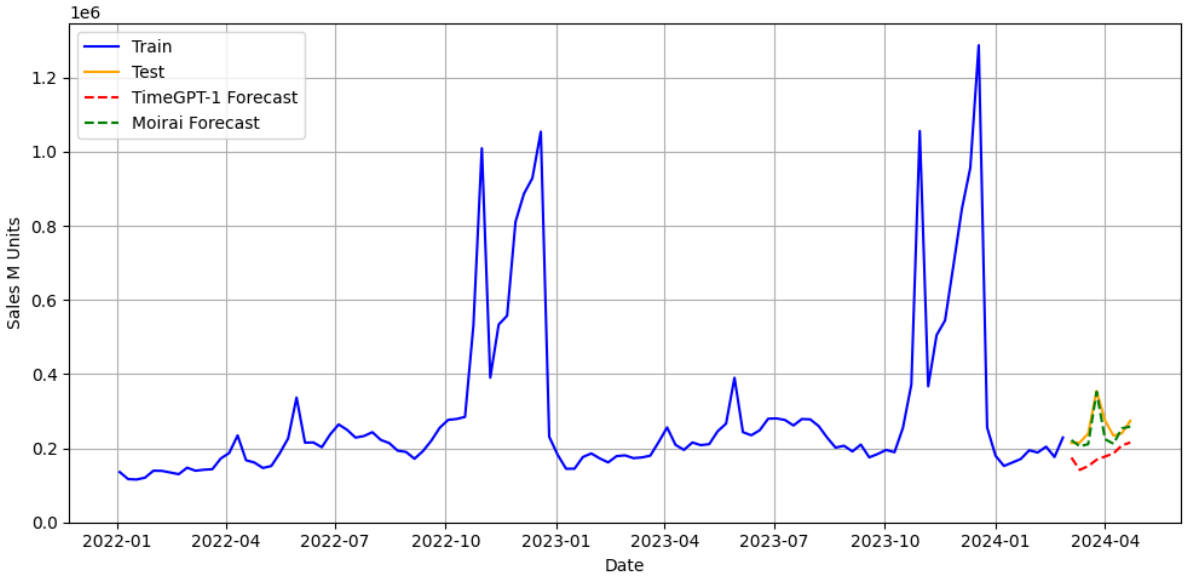

Figure 5.14 – Forecast by foundation models for toys on March-April 2024

Moirai demonstrated a consistent superiority in performance in March-April 2024, producing the most accurate forecasts in both absolute and relative terms (RMSE = 23,512.33; MAPE = 7.44%). The magnitude and timing of sales peak exhibited a high degree of correlation, as demonstrated in Figure 5.14, with the model demonstrating a capacity for adaptation to short-term fluctuations. As presented in Figure 5.15, the residuals confirmed the model's robustness, with minimal bias and controlled variability. The descriptive statistics lend support to this interpretation: the residuals exhibited a mean of 13,532.56 units, indicative of minor forecast bias, consistent with the model's capacity to adapt to varying demand levels.

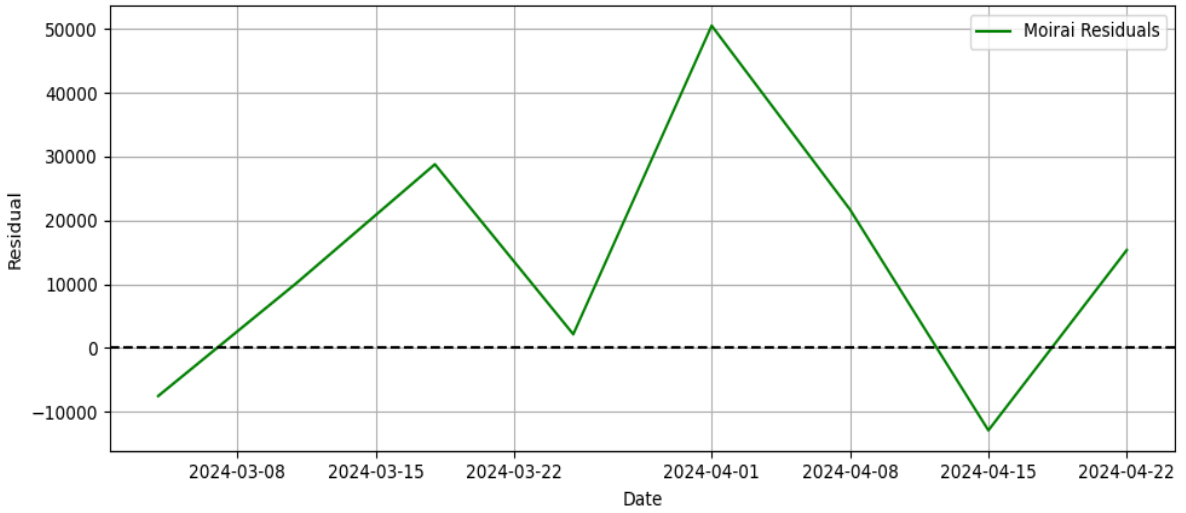

Figure 5.15 – Residuals for Moirai forecast for toys on March-April 2024

However, in the November-December 2024 period, Moirai's performance declined significantly, placing it among the least precise models for this series and period. While the model was able to detect the timing of the peak, it struggled to accurately capture its full scale, and the forecasts were found to be unstable. As illustrated in Figure A.16, the forecasted values demonstrate a clear deviation from the actual sales data. This indicates that, while Moirai demonstrates notable efficacy in stable demand periods, its ability to manage extreme seasonal volatility was comparatively constrained.

5.5. CROSS-MODEL COMPARATIVE ANALYSIS

The performance of different models varied across the product categories and forecasting windows. Each model exhibited distinct strengths and weaknesses, shaped by its capacity to capture seasonal patterns, abrupt demand fluctuations, and the overall sales structure of each series.

The SARIMA models exhibited persistent underperformance across all product categories and periods. In the March-April 2024 period, SARIMA and its variants were found to be inferior to advanced models in terms of accuracy, and they also demonstrated a lack of responsiveness to local fluctuations. The incorporation of regressors did not enhance the accuracy of forecasts, and in certain instances, it led to a decline in their precision. This is likely attributable to the rigid parametric structure of SARIMA, which models seasonality and trends as deterministic and regular. Consequently, it lacks the flexibility required to capture abrupt changes such as promotional surges. This pattern was observed to persist in the November-December 2024 period, during which SARIMA models demonstrated limited sensitivity to peak periods, with marginal differences in performance observed between configurations. The rigidity of their seasonal structure imposed limitations on their adaptability within dynamic sales environments.

The Holt-Winters model exhibited a discernible reliance on stable seasonal patterns, demonstrating efficacy in periods and categories characterised by demand that followed predictable, recurring cycles. The reliability of the results was most evident in the context of the toys series, both during the March-April period and in the more volatile November-December 2024 window. In this latter instance, the model successfully captured the seasonal surge despite the occurrence of a substantial sales increase. Conversely, Holt-Winters encountered challenges in adapting to the volatility in demand for camcorders and media tablets, leading to overestimation or underestimation of sales. These outcomes suggest that Holt-Winters can perform robustly even in volatile periods, when the volatility aligns with historical seasonal trends. However, when volatility is driven by irregular factors or abrupt changes not present in prior periods, the adaptability of Holt-Winters' smoothing structure and fixed seasonal components is limited. Consequently, Holt-Winters was most effective in

environments where volatility was seasonal in nature, but less suited to contexts shaped by unexpected demand shifts.

The Prophet demonstrated consistent, yet product-dependent, performance. The baseline model was generally sufficient to capture the underlying sales patterns, particularly for camcorders and media tablets. Across both forecasting windows, the incorporation of holiday and promotion regressors did not systematically enhance forecast accuracy, and in several cases, it was observed that performance was actually diminished. The performance of Prophet was found to be optimised when demand patterns aligned closely with its internal seasonal structure, and when sales volatility followed predictable seasonal cycles. However, the model encountered difficulties when confronted with sales fluctuations were driven by irregular or short-term factors that were not adequately addressed by its default components. In summary, the Prophet model was found to be a reliable baseline across different products and periods. On occasion, it produced the most optimal absolute results, as evidenced by the toys forecast in the November-December period. However, the model demonstrated limited responsiveness to irregular volatility and did not significantly benefit from the incorporation of external regressors.

The foundation models exhibited distinct patterns of performance across product categories and periods. TimeGPT-1 demonstrated superior consistency in performance across different categories and periods, producing reliable forecasts with reduced absolute errors (RMSE), particularly in camcorders and toys during periods of volatility. It excelled in modelling overall sales trends but showed a consistent tendency to underestimate periods of heightened demand, indicative of a conservative forecast bias. In contrast, Moirai excelled in the March-April 2024 period, exhibiting superior accuracy for media tablets and toys under conditions of moderate volatility and the predominance of seasonal patterns. However, it struggled in the November-December period, particularly in the toys series, during which demand was characterised by significant surges, resulting in volatile and less precise forecasts.

Across the full range of products and periods, the foundation models generally outperformed classical forecasting models in absolute accuracy, particularly during volatile periods and within categories characterised by irregular demand. However, their relative accuracy (MAPE) was occasionally higher than that of the simpler forecasting methods, particularly in periods where demand was driven by recurring seasonal cycles. In such cases, models with fixed seasonal structures, such as the SARIMA and the Holt-Winters models, have been shown to benefit from the alignment between past patterns and observed fluctuations.

While model performance appeared to be robust when assessed using typical error measures, a common limitation across all models was the tendency to produce forecasts that were consistently higher than observed sales, particularly during periods of moderate demand. This tendency persisted even among advanced models, such as TimeGPT-1 and Moirai. Consequently, the forecasts, despite their statistical robustness, posed significant challenges for decision-making. From an operational perspective, inflated forecasts have the potential to

compromise the accuracy of planning, resulting in excess inventory, misaligned marketing efforts, and inefficient resource allocation. Nevertheless, these models remain valuable tools and can be effectively employed in practice, provided their tendency to overestimate is carefully accounted for. These findings underscore the necessity of evaluating models not solely on their predictive accuracy but also on their practical implications.

6. CONCLUSIONS AND FUTURE RESEARCH

This study aimed to evaluate and compare the performance of classical time series forecasting models with more advanced methods, including Prophet and foundation models, in the context of retail sales forecasting. The research was guided by the following core objectives: firstly, to assess the applicability of foundation models to time series sales data; secondly, to determine which model architectures deliver the most accurate forecasts across different product categories; and thirdly, to identify key features that influence forecast accuracy.

To address these objectives, the study applied three variants of SARIMA, Holt-Winters, three configurations of Prophet, TimeGPT-1, and Moirai to weekly sales data from 2022 to 2024 across three product categories, namely camcorders, media tablets, and toys. The research focused on evaluating model accuracy in two forecasting windows: March to April 2024, representing a more stable demand period, and November to December 2024, a period characterised by high volatility due to promotional activity and holiday peaks. The performance of the models was evaluated by the Root Mean Square Error and Mean Absolute Percentage Error, enabling both absolute and relative performance comparisons.

The comparative analysis demonstrated the superiority of foundation models in absolute accuracy, particularly during volatile sales periods and in product categories with irregular demand. TimeGPT-1 demonstrated a consistent capacity to generate reliable and consistent forecasts across categories, particularly in camcorders and toys during periods of volatility. Moirai demonstrated particularly proficiency in moderate volatility, achieving notable results during the March-April 2024 period. Prophet provided stable and moderately accurate forecasts. However, the incorporation of external regressors often proved ineffective in enhancing the performance of the forecasts, and, in certain instances, it resulted in a decline in forecast accuracy. In contexts where sales followed regular seasonal cycles, particularly for the toys series, the SARIMA and Holt-Winters models demonstrated competitiveness. However, these models were unable to adapt to sudden fluctuations.

From a theoretical perspective, these findings emphasise the limitations of classical forecasting methods in retail forecasting, where demand is increasingly influenced by complex, often exogenous factors. SARIMA's rigid seasonal structure and Holt-Winters smoothing mechanisms were found to be inadequate in the context of modelling high-frequency demand changes. In contrast, foundation models demonstrated strong generalisation capabilities, successfully adapting across different categories and temporal contexts. The findings underscore the potential of foundation models to address the escalating intricacy of retail forecasting, a domain in which exogenous factors and event-driven are assuming an increasingly pivotal role.

From a practical perspective, the findings indicate that retail forecasting systems should incorporate foundation models, particularly in environments characterised by demand uncertainty and volatility. It has been demonstrated that these models have the capacity to

reduce forecast error without the necessity of extensive manual adjustment. However, classical models such as SARIMA and Holt-Winters remain relevant for contexts where demand follows stable seasonal patterns. The analysis also demonstrated that the incorporation of external regressors did not consistently enhance forecast accuracy, indicating that a more comprehensive set of contextual features should be incorporated.

Despite the comprehensive nature of the analysis, some limitations must be acknowledged. The dataset was limited in both product scope and temporal coverage, which may have constrained the generalisability of the findings. Additionally, the scope of external regressors was confined to holidays and promotions, which, while informative, may not capture all the relevant drivers of sales fluctuations.

Future research could build on these findings by expanding the product and temporal coverage of the dataset, thereby improving the training of the model and its robustness. Additionally, the incorporation of a wider range of contextual features, including pricing and economic indicators, could enhance model sensitivity to demand drivers. Further comparative evaluations across different forecasting frequencies, such as daily or monthly, may provide further insights into the relative strengths of each model architecture. A comparative analysis involving other foundation models and the exploration of hybrid ensemble strategies could additionally enhance forecasting accuracy.

The present study concludes that foundation models offer a robust and scalable solution for retail sales forecasting, particularly in circumstances where demand is uncertain. However, the superiority of the latter is more evident in volatile environments, whilst classical forecasting models remain applicable in contexts characterised by stable seasonal patterns.

BIBLIOGRAPHICAL REFERENCES

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). *Chronos: Learning the Language of Time Series* (No. arXiv:2403.07815). arXiv. http://arxiv.org/abs/2403.07815

Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., & Khan, F. S. (2023, July 25). *Foundational Models Defining a New Era in Vision: A Survey and Outlook*. arXiv.Org. https://arxiv.org/abs/2307.13721v1

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). *Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Methodology* (No. arXiv:1901.04028). arXiv. http://arxiv.org/abs/1901.04028

Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis Forecasting and Control*.

Brown, R. G. (1963). *Smoothing, Forecasting and Prediction of Discrete Time Series*. Prentice-Hall.

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, *2*(1), 100058. https://doi.org/10.1016/j.jjimei.2022.100058

Fine-tuning Tutorial TimeGPT. (2025, February). TimeGPT Foundational Model for Time Series Forecasting and Anomaly Detection. https://www.nixtla.io/docs/tutorials-fine_tuning

Garza, A., Challu, C., & Mergenthaler-Canseco, M. (2024). *TimeGPT-1* (No. arXiv:2310.03589). arXiv. http://arxiv.org/abs/2310.03589

Gustriansyah, R., Suhandi, N., Antony, F., & Sanmorino, A. (2019). Single exponential smoothing method to predict sales multiple products. *Journal of Physics: Conference Series*, *1175*, 012036. https://doi.org/10.1088/1742-6596/1175/1/012036

Hasan, M. R., Kabir, M. A., Shuvro, R. A., & Das, P. (2022). *A Comparative Study on Forecasting of Retail Sales* (No. arXiv:2203.06848). arXiv. https://doi.org/10.48550/arXiv.2203.06848

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

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kumar, L., Khedlekar, S., & Khedlekar, U. K. (2024). A comparative assessment of holt winter exponential smoothing and autoregressive integrated moving average for inventory optimization in supply chains. *Supply Chain Analytics*, *8*, 100084. https://doi.org/10.1016/j.sca.2024.100084

Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., & Wen, Q. (2024). Foundation Models for Time Series Analysis: A Tutorial and Survey. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6555–6565. https://doi.org/10.1145/3637528.3671451

Lima, S., Gonçalves, A. M., & Costa, M. (2019). *Time series forecasting using Holt-Winters exponential smoothing: An application to economic data*. 090003. https://doi.org/10.1063/1.5137999

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). *Generating Wikipedia by Summarizing Long Sequences* (No. arXiv:1801.10198). arXiv. https://doi.org/10.48550/arXiv.1801.10198

Liu, Y., Zhang, H., Li, C., Huang, X., Wang, J., & Long, M. (2024). *Timer: Generative Pre-trained Transformers Are Large Time Series Models* (No. arXiv:2402.02368). arXiv. http://arxiv.org/abs/2402.02368

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022a). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022b). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *38*(4), 1325–1336. https://doi.org/10.1016/j.ijforecast.2021.07.007

Medar, R., Rajpurohit, V. S., & Rashmi, B. (2017). Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning. *ResearchGate*. https://doi.org/10.1109/ICCUBEA.2017.8463779

Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). *Comparing Prophet and Deep Learning to ARIMA in Forecasting Wholesale Food Prices* (No. arXiv:2107.12770). arXiv. http://arxiv.org/abs/2107.12770

Mgale, Y. J., Yan, Y., & Timothy, S. (2021). A Comparative Study of ARIMA and Holt-Winters Exponential Smoothing Models for Rice Price Forecasting in Tanzania. *Open Access Library Journal*, *8*(5), Article 5. https://doi.org/10.4236/oalib.1107381

Miller, J. A., Aldosari, M., Saeed, F., Barna, N. H., Rana, S., Arpinar, I. B., & Liu, N. (2024, January 25). *A Survey of Deep Learning and Foundation Models for Time Series Forecasting*. arXiv.Org. https://arxiv.org/abs/2401.13912v1

Puthran, D., Prasad H C, S., Kumar, K., & Muniyappa, M. (2014). *Comparing SARIMA and Holt-Winters' forecasting accuracy with respect to Indian motorcycle industry*.

Python Software Foundation. (2020). *Python* (Version 3.9.0) [Computer software]. https://www.python.org/

Qi, X., Hou, K., Liu, T., Yu, Z., Hu, S., & Ou, W. (2021). *From Known to Unknown: Knowledge-guided Transformer for Time-Series Sales Forecasting in Alibaba* (No. arXiv:2109.08381). arXiv. https://doi.org/10.48550/arXiv.2109.08381

R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Redoulès, G. (2025). *Redoules/moirai* [Computer software]. https://github.com/redoules/moirai (Original work published 2024)

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599–607. https://doi.org/10.1093/biomet/71.3.599

Taylor, S. J., & Letham, B. (2017). *Forecasting at scale*. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.3190v2

Theodorou, E., Wang, S., Kang, Y., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). *Exploring the representativeness of the M5 competition data* (No. arXiv:2103.02941). arXiv. https://doi.org/10.48550/arXiv.2103.02941

Understanding LSTM Networks—Colah's blog. (2015). https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vilhuber, L. (2023). Reproducibility and Transparency versus Privacy and Confidentiality: Reflections from a Data Editor. *Journal of Econometrics*, *235*(2), 2285–2294. https://doi.org/10.1016/j.jeconom.2023.05.001

Wang, S., Li, C., & Lim, A. (2021). *Why Are the ARIMA and SARIMA not Sufficient* (No. arXiv:1904.07632). arXiv. https://doi.org/10.48550/arXiv.1904.07632

Winters, P. R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, *6*(3), 324-324–342. JSTOR Journals.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). *Unified Training of Universal Time Series Forecasting Transformers* (No. arXiv:2402.02592). arXiv. https://doi.org/10.48550/arXiv.2402.02592

Xue, H., & Salim, F. D. (2023). *PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting* (No. arXiv:2210.08964). arXiv. http://arxiv.org/abs/2210.08964

Zunic, E., Korjenic, K., Hodzic, K., & Donko, D. (2020). *Application of Facebook's Prophet Algorithm for Successful Sales Forecasting Based on Real-world Data* (No. arXiv:2005.07575). arXiv. http://arxiv.org/abs/2005.07575

APPENDIX A

This appendix provides the complete set of forecast and residual plots for all evaluated models across the two forecasting windows: March-April 2024 and November-December 2024. These results complement the main analysis presented in Sections 5.2 to 5.4. The visualisations in this appendix offer a comprehensive overview of each model's behaviour under varying levels of demand volatility.

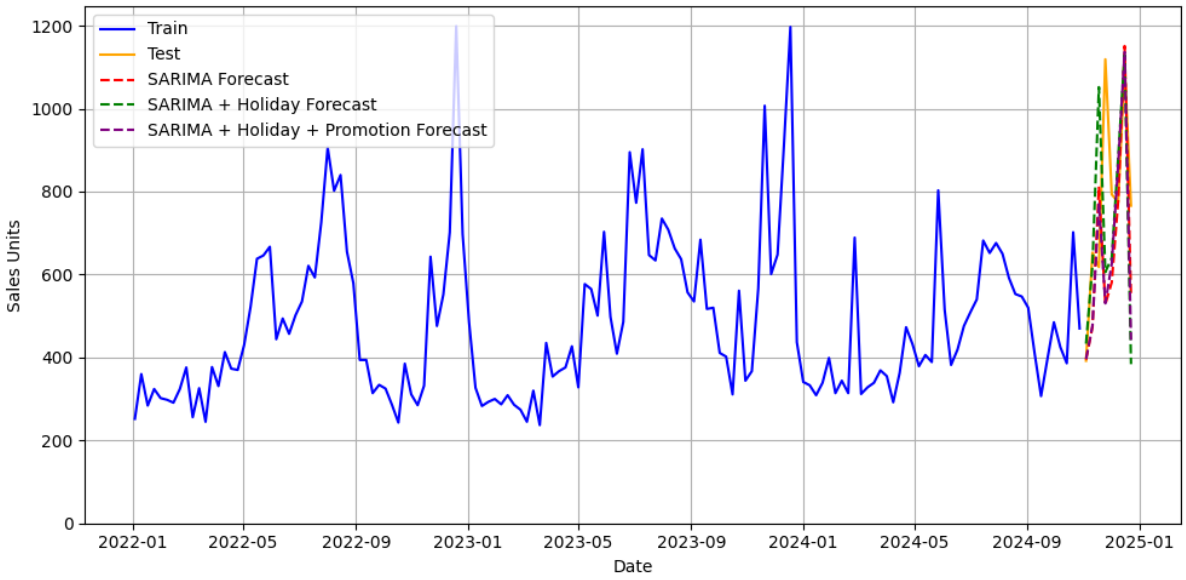

Figure A.1 – Forecast by SARIMA configurations for camcorders on November-December 2024

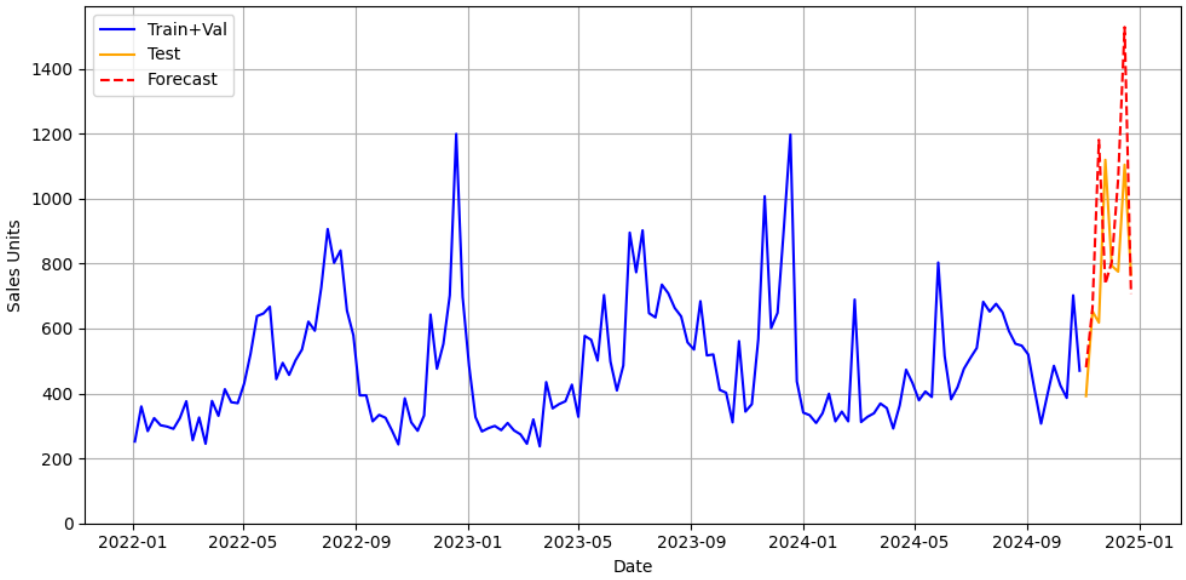

Figure A.2 – Forecast by Holt-Winters for camcorders on November-December 2024

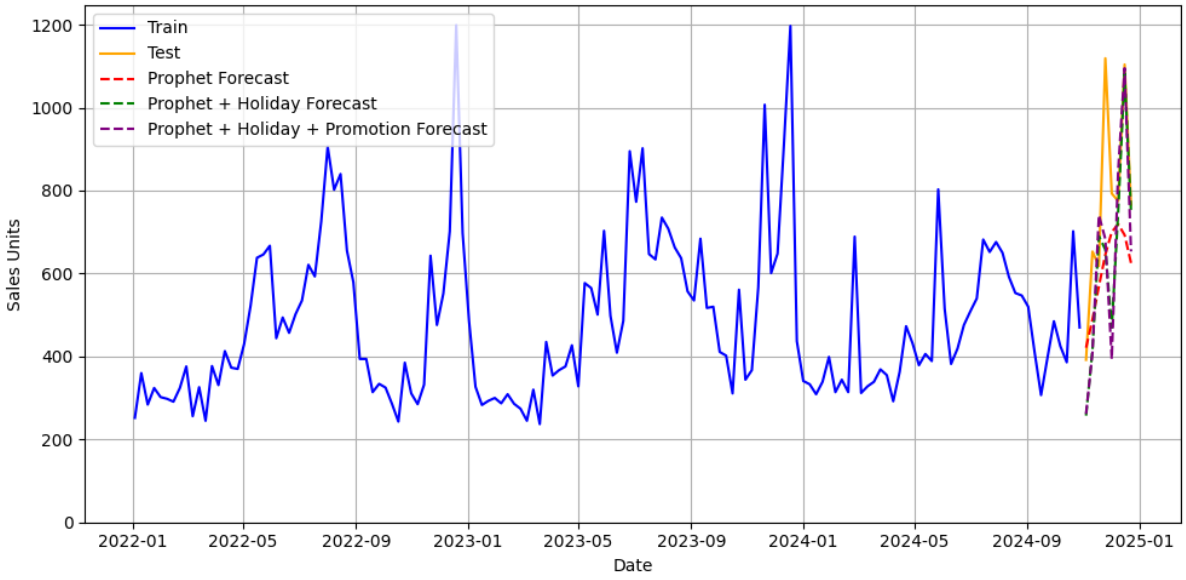

Figure A.3 – Forecast by Prophet configurations for camcorders on November-December 2024

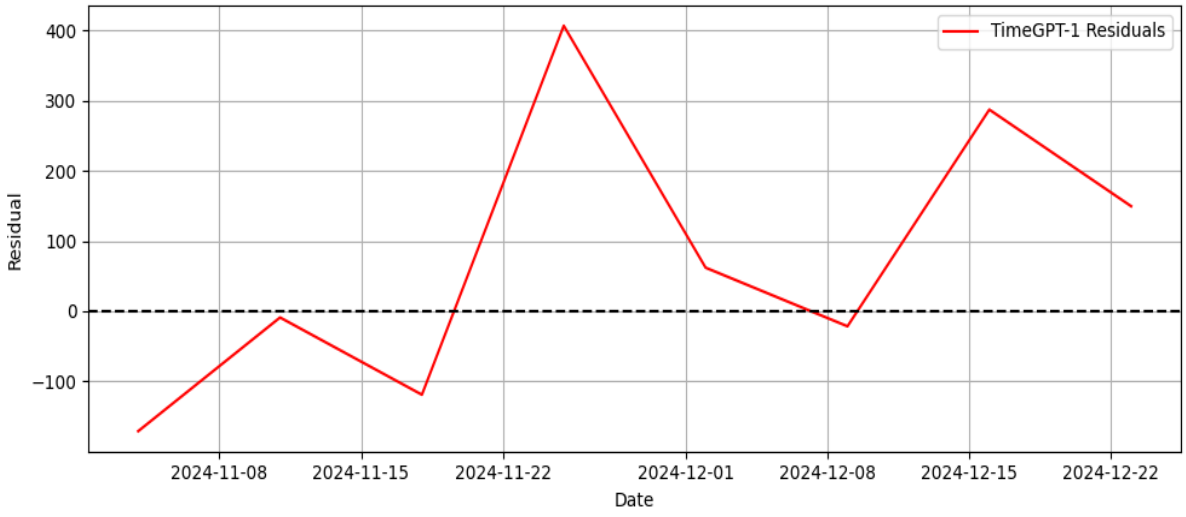

Figure A.4 – Residuals for TimeGPT-1 forecast for camcorders on November-December 2024

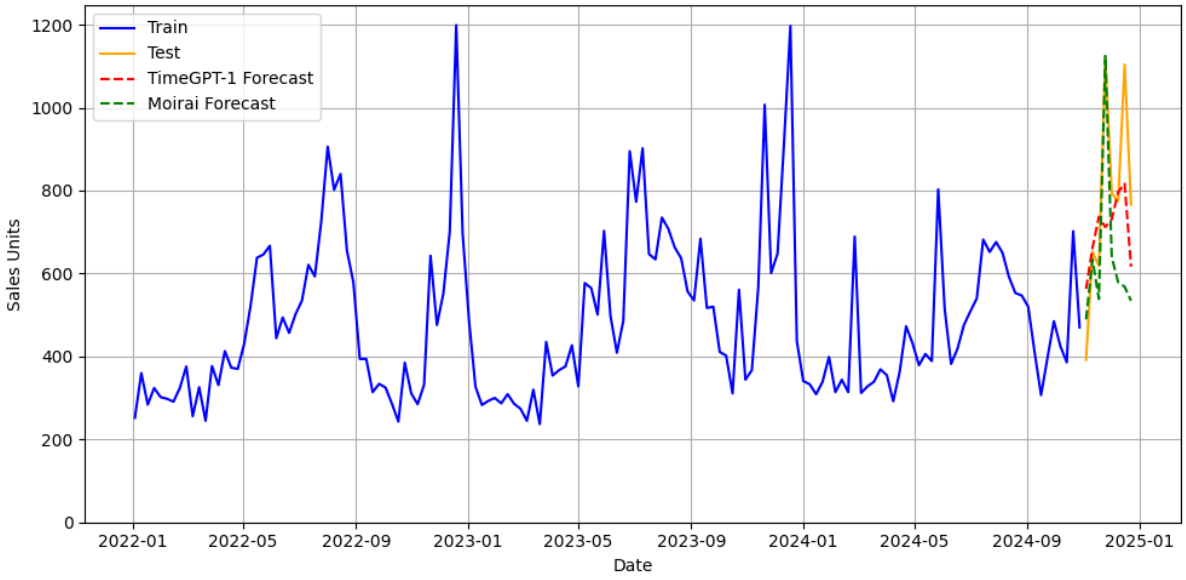

Figure A.5 – Forecast by foundation models for camcorders on November-December 2024

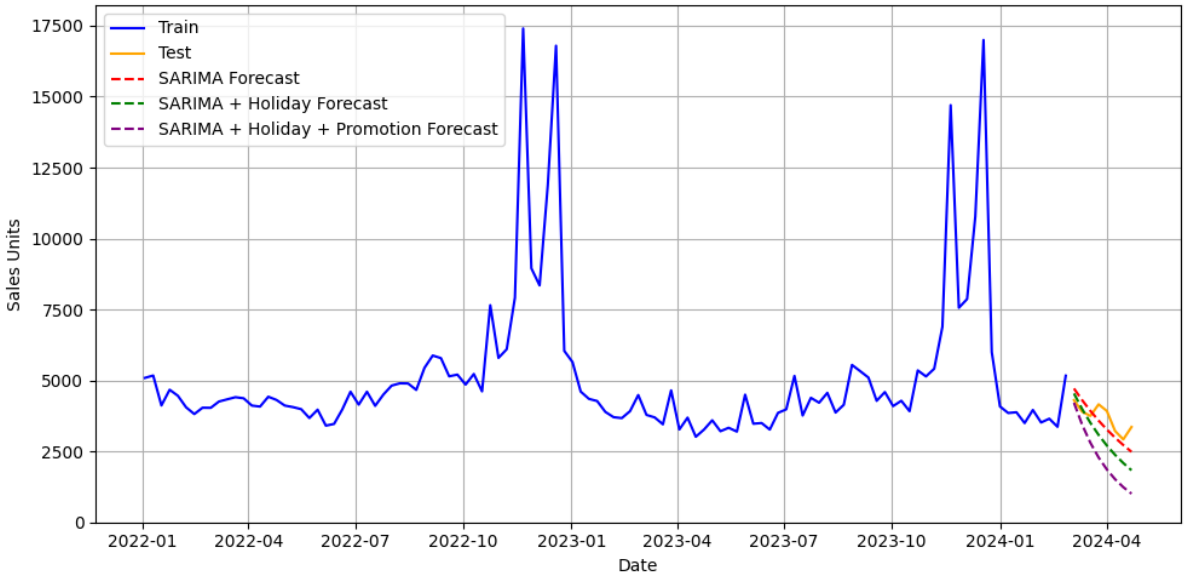

Figure A.6 – Forecast by SARIMA configurations for media tablets on March-April 2024

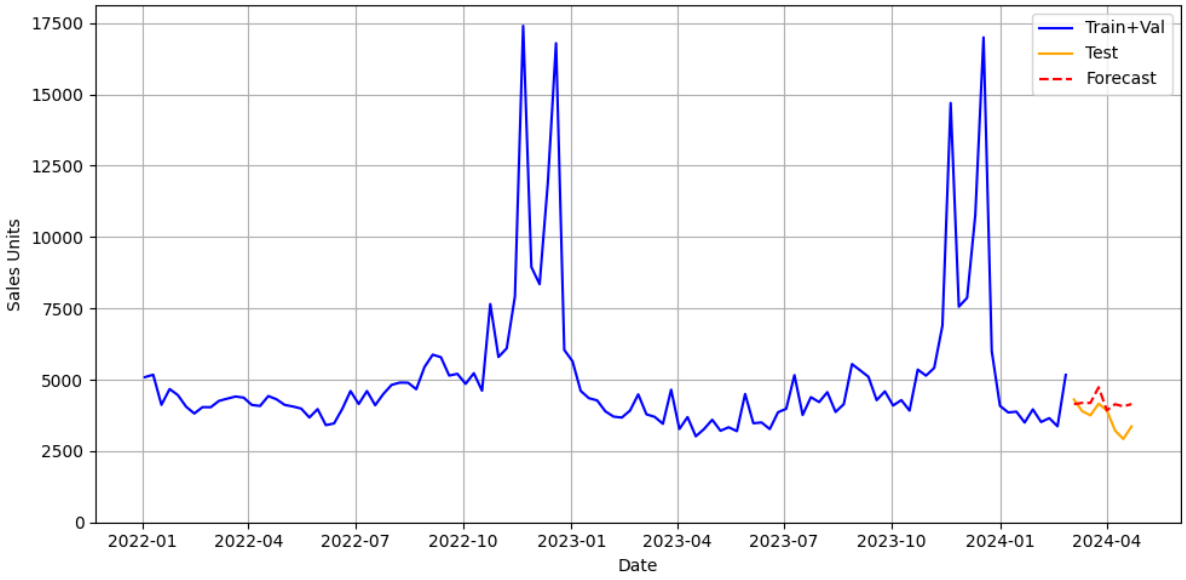

Figure A.7 – Forecast by Holt-Winters for media tablets on March-April 2024

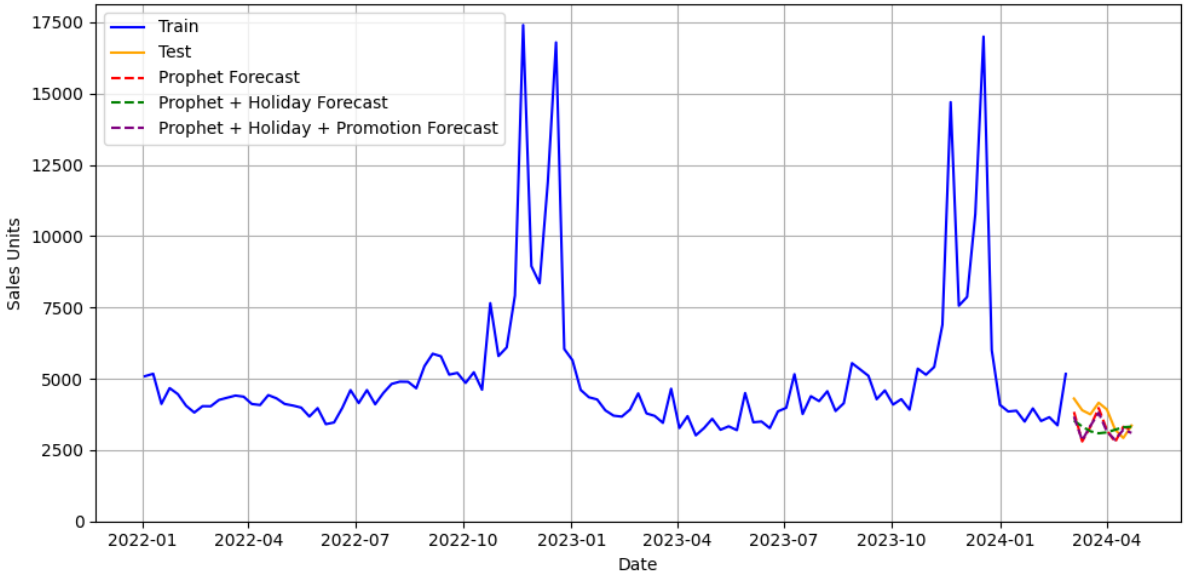

Figure A.8 – Forecast by Prophet configurations for media tablets on March-April 2024

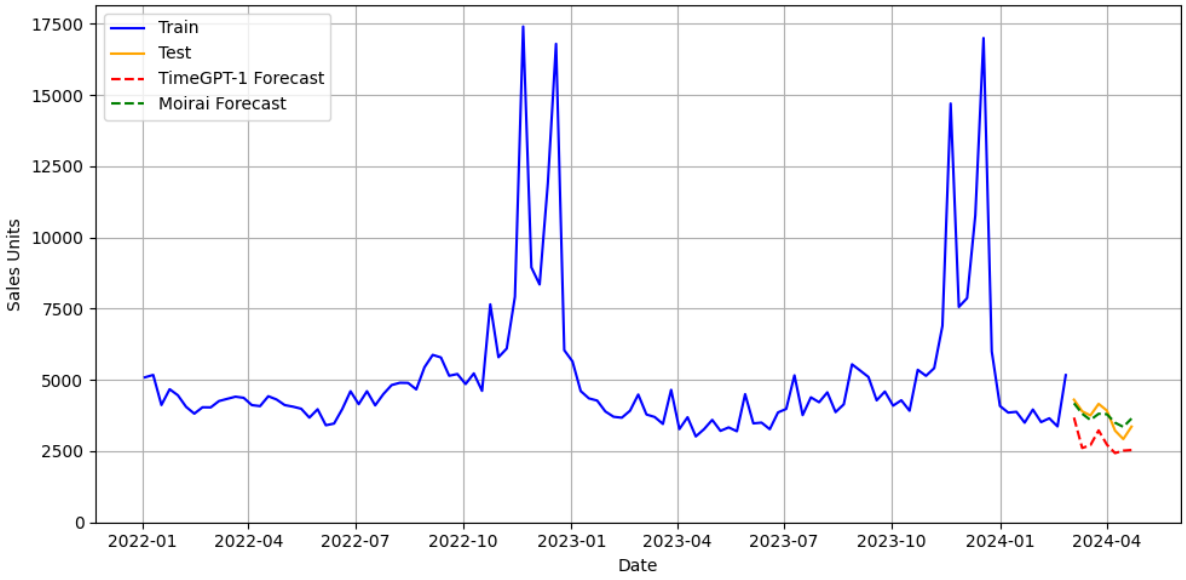

Figure A.9 – Forecast by foundation models for media tablets on March-April 2024

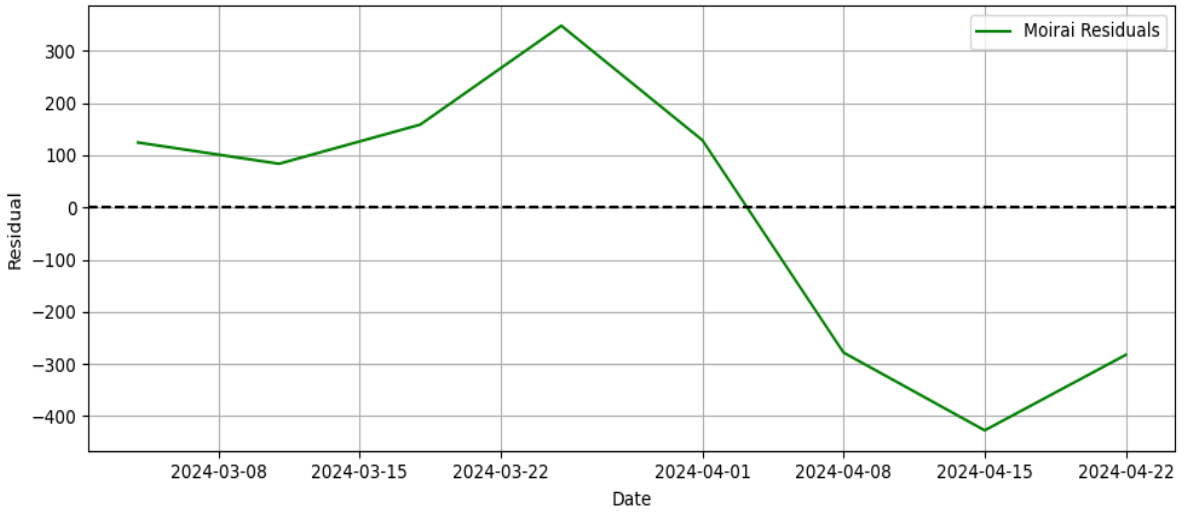

Figure A.10 – Residuals for Moirai forecast for media tablets on March-April 2024

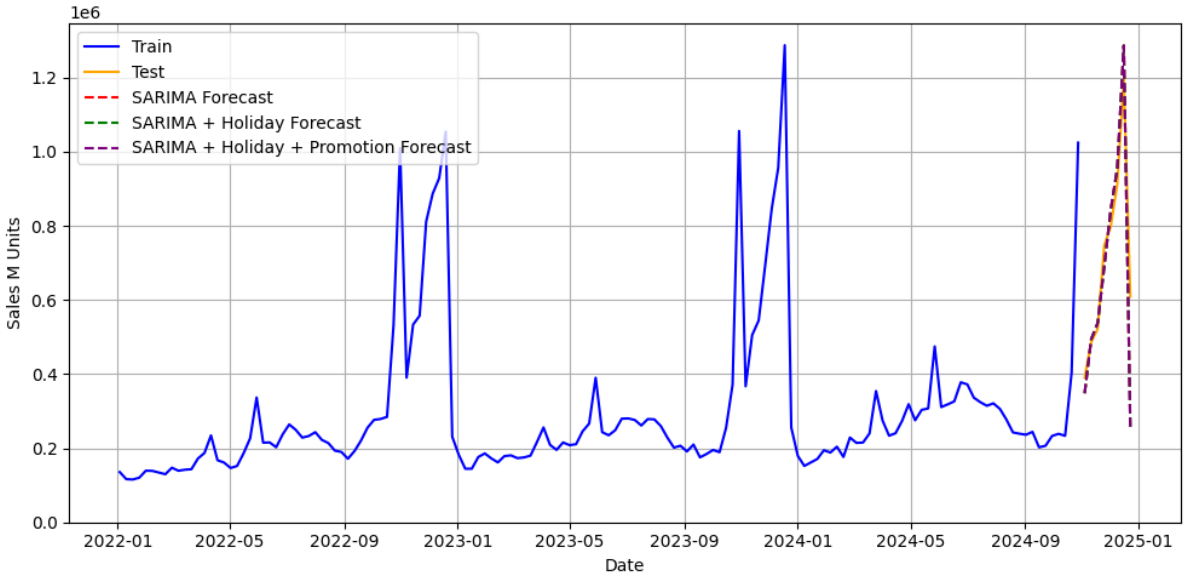

Figure A.11 – Forecast by SARIMA configurations for toys on November-December 2024

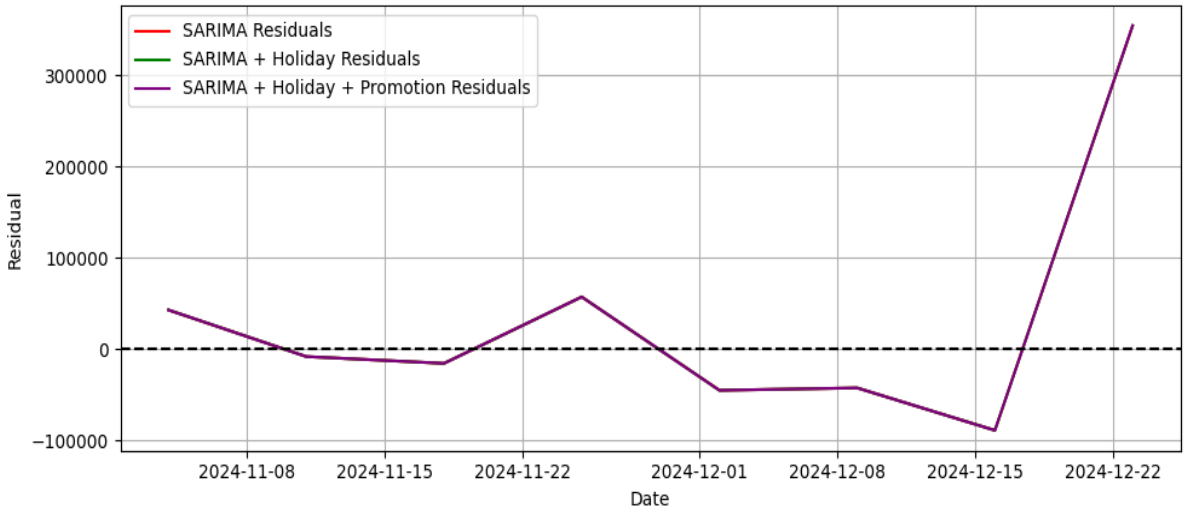

Figure A.12 – Residuals for SARIMA configurations forecasts for toys on November-December 2024

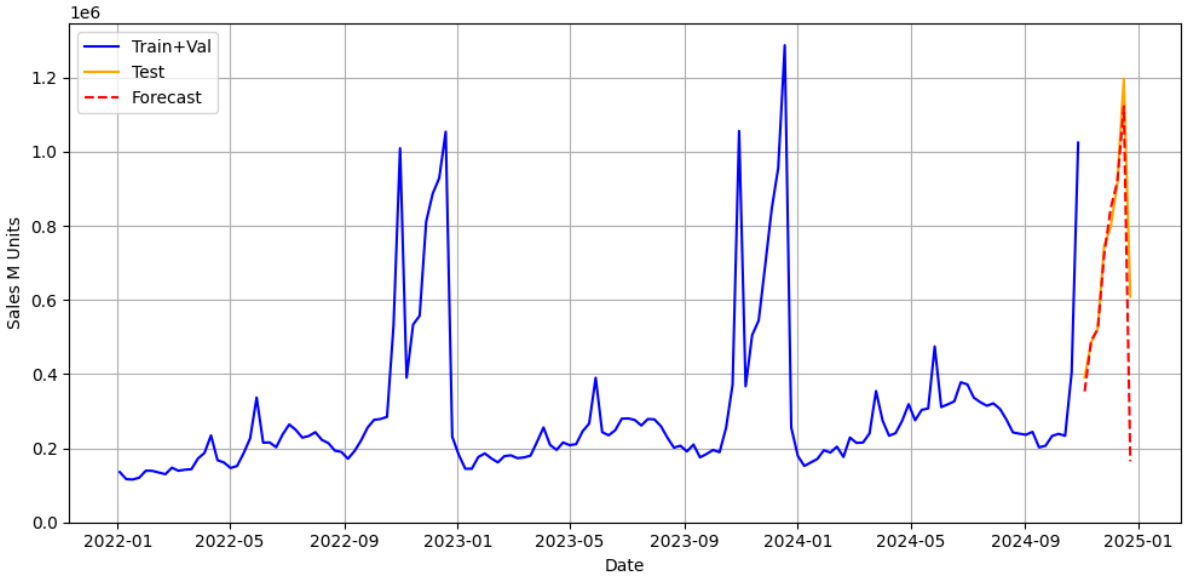

Figure A.13 – Forecast by Holt-Winters for toys on November-December 2024

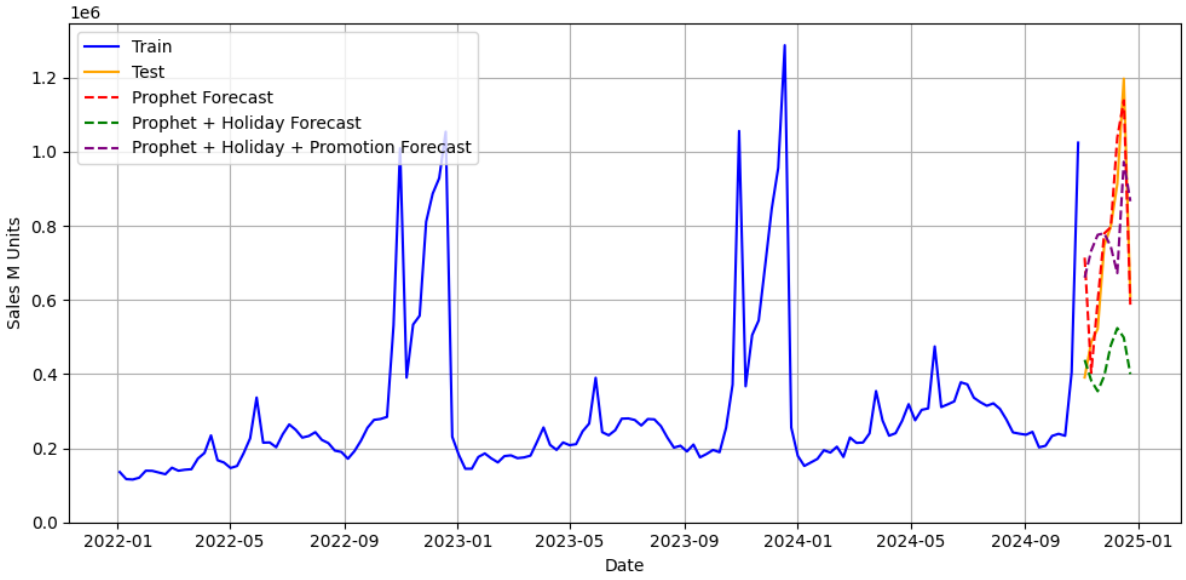

Figure A.14 – Forecast by Prophet configurations for toys on November-December 2024

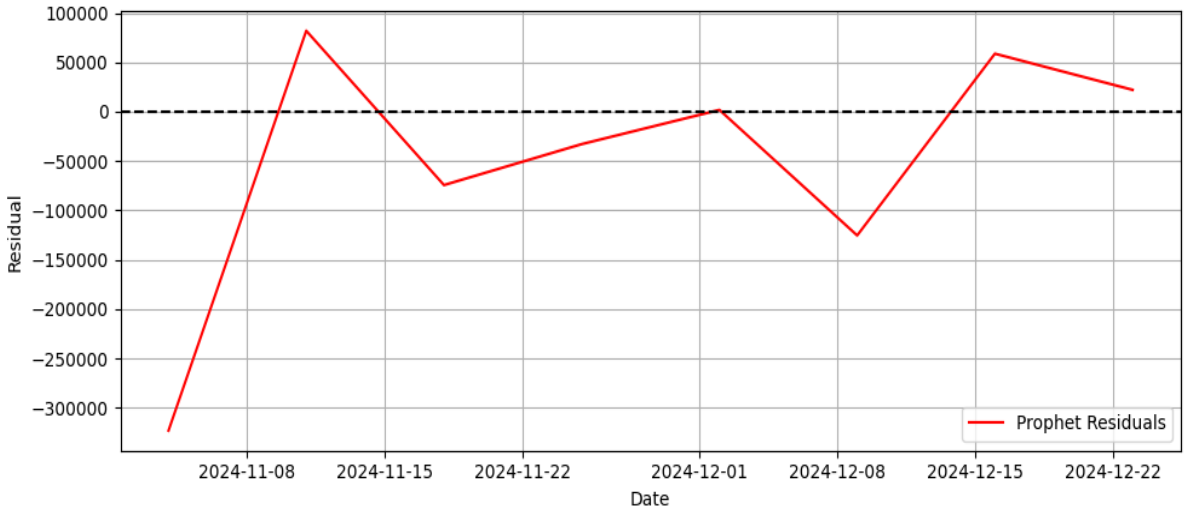

Figure A.15 – Residuals for Prophet forecast for toys on November-December 2024

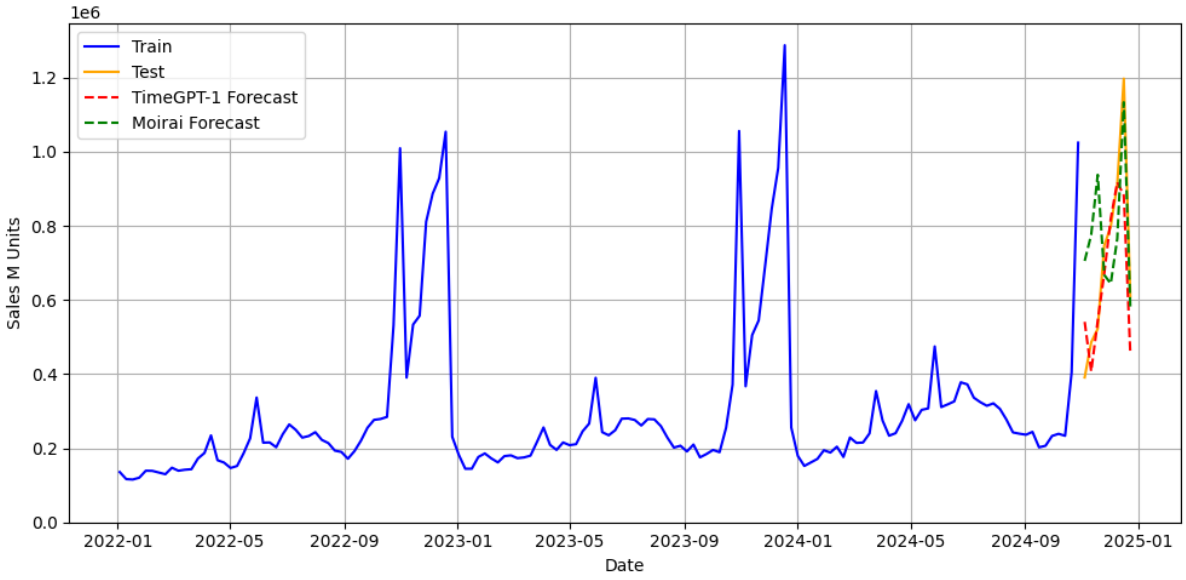

Figure A.16 – Forecast by foundation models for toys on November-December 2024

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa