

Vrije Universiteit Amsterdam



Bachelor Thesis

Optimizing Sales Forecasting: The Role of Weather and Seasonality in SARIMAX Models

Author: Bodale Lavinia Maria (2735694)

Supervisor: Sieuwert van Otterloo
Company supervisor: Martijn Matena
2nd reader: Jacco van Ossenbruggen

July 9, 2024

Acknowledgements

Before starting I would quickly like to thank the following people without whom I would not have been able to complete this research: my supervisor Sieuwert for his enthusiasm for the project, for his support and guidance over the past months; Marijn, Nina, and Erik from the CB for offering me the chance to grow a business point of view.

Now, I would like to thank the people without whom I would not have made it through my bachelor's degree. First of all, my family for all the love and unconditional support; my friends for always being there; and, as Snoop Dogg said, "...last but not least I wanna thank me, for believing in me, for doing all this work, for having no days off, for never quitting".

Abstract

This research investigates the enhancement of sales forecasting by incorporating weather variables and volatile events such as Easter, Ramadan, and Carnival into the SARIMAX model. SARIMAX is a time series model that incorporates both seasonal patterns and external factors. The study reviews the literature on the impact of weather and events on consumer behavior and sales, highlighting both supporting and contradictory findings. The dataset is divided into training and testing sets, with 83.5% for training and the remainder for testing, focusing on data from the beginning of 2023. Model performance is evaluated using metrics like MAE, MSE, and RMSE, with particular attention to the percentage of days with prediction errors below 5%.

Preliminary results showed mixed improvements with different events as exogenous variables and that integrating weather would not lead to significant conclusions. The model with Easter data performed better than those with Ramadan or Carnival, yet integrating all events simultaneously did not yield significant accuracy gains. Future research should consider the duration of these events, as their varying lengths may impact sales differently. The findings aim to contribute to more refined predictive models that account for contextual factors, offering better sales forecasting and strategic insights for businesses that could be used to anticipate and adapt to changing demand.

Contents

1	Introduction	6
1.1	Case study: CB	6
1.2	Problem and motivation	7
1.3	Scientific contribution	7
1.4	Research objective	7
1.5	Related studies	8
1.6	Research questions	11
1.7	Used model	12
1.8	Libraries	13
1.9	Dataset	14
2	Methods	14
2.1	Preliminary steps	14
2.1.1	Problem analysis	14
2.1.2	Explore user requirements	15
2.2	Phase 1: Base SARIMA model construction	15
2.3	Phase 2: Integration of exogenous variables into SARIMAX model	16
2.4	Phase 3: Model comparison and evaluation	16
3	Expected deliverables	16
4	Data preprocessing	17
4.1	Data Selection	17
4.2	Feature Extraction	17
4.3	Data Exploration	17
4.4	Rolling mean and standard deviation	19
4.5	Lags	19
5	Data split	20
6	Models	20
6.1	SARIMA Base Model	20
6.2	SARIMA Base Model Optimized	21
6.3	SARIMAX Model: weather	21
6.4	SARIMAX Model: volatile events	22
6.4.1	Easter	23
6.4.2	Ramadan	23
6.4.3	Carnival	23
6.4.4	Easter, Ramadan and Carnival	23
6.5	Combined model: Optimized SARIMA and Seasonality	24
7	Model Evaluation	25
7.1	Evaluation Metrics	25
7.2	SARIMA Base Model	25
7.3	SARIMA Base Model Optimized	26

7.4	SARIMAX Model: volatile events	26
7.4.1	Easter	26
7.4.2	Ramadan	26
7.4.3	Carnival	26
7.4.4	SARIMAX: Easter, Ramadan, and Carnival	26
7.5	Combined model: Optimized SARIMA and Seasonality	27
8	Discussion	27
9	Conclusion	28
	References	30

1 Introduction

In today's rapidly changing market conditions, the ability to accurately predict incoming orders becomes not just an asset but a necessity for maintaining competitive advantage and ensuring the flawless execution of business processes. Order forecasting is an important analytical process businesses use to predict future demand for their products or services over a specified period. By analyzing historical sales data, market trends, customer behavior, and various other influencing factors, companies aim to estimate the volume of future orders. This multidimensional approach facilitates resource optimization, allowing businesses to maintain the right balance of stock levels to minimize storage costs and reduce waste while ensuring demand is met without delay.

Order forecasting supports strategic planning, empowering companies to allocate their budget, conduct targeted marketing campaigns, and make strategic decisions that align with predicted market demand. Furthermore, fulfilling orders on time and accurately is key to achieving customer satisfaction and retention. Forecasting helps anticipate customer needs, leading to timely product availability and enhanced service quality. In competitive markets, the ability to anticipate and adapt to changing demand can provide a significant competitive advantage. Moreover, the integration of advanced forecasting models into the logistics strategy empowers companies to align their production schedules, workforce allocation, and distribution plans with anticipated demand, thereby enhancing operational agility and responsiveness.

1.1 Case study: CB

This research will focus on a specific case study, which involves analyzing the CB company. CB is a well-renowned logistics company currently operating in the Netherlands and Flanders. Generally, due to the costly and inefficient process of ordering books, the bookstores would have to write several publishers when ordering. For this reason, in 1871 the CB company was founded. CB is an abbreviation for Centraal Boekhuis, a centralized storage, processing, and delivery center that supplies bookstores in the Netherlands and Flanders.

The CB is the link between the publisher, bookseller, and the consumer, and over the years, it has grown into the largest distribution center for books and e-books. Even though they have expanded into the industry of e-books and book portals physical books still have a large fanbase due to some people's preferences. Authors can register their books with the CB and they will immediately be visible to approximately 2,500 booksellers in both the Netherlands and Flanders. The sellers can order the books directly from the CB's stock. In addition, the book will appear in the web shops of all online platforms such as Bol.com. Hence, CB offers both B2B(business-to-business) and B2C(business-to-customer) services.

The company has recently heavily invested in (POD) print-on-demand services. This service offers reduced stock risk, sustainable use, and fast deliveries - within 24 hours. This could be quite convenient for the end customer but quite a hassle for CB as they have a limited period to satisfy their end consumer. CB offers their client a dashboard for analytics to measure the turnover of the books and guide the bookstore's ordering process. Their data storage contains information about the orders since 2006. They also offer (B2C) solutions, where clients can order a book through bol.com and they would fulfill the order

on demand. However, their core business is B2B, namely selling to bookstores.

1.2 Problem and motivation

This research aims to address the challenges faced by the CB company which recently acquired a new printing machine and is experiencing difficulties in fulfilling orders punctually. Without a predictive system in place, CB encounters disruptions due to the unpredictability of order arrivals. This study focuses on developing a model to forecast the likelihood and timing of incoming orders, thereby assisting CB in anticipating demand more effectively.

It is important to distinguish between CB's business-to-business (B2B) and business-to-consumer (B2C) sectors. For the B2B sector, the goal is to predict the timing of the last incoming order, while for the B2C sector, the objective is to examine consumer behavior. Specifically for the B2C sector, the research delves into examining consumer behavior and predicting the number of orders, focusing on the influence of weather conditions and seasonality. This approach acknowledges the role that environmental factors play in shaping consumer purchasing patterns. By integrating weather and seasonal data into the predictive model, the research intends to offer CB more accurate insights into the timing and likelihood of incoming orders, thereby facilitating better preparation and response strategies to meet customer demand.

1.3 Scientific contribution

Solving the problem of forecasting incoming orders would bring benefits to CB. It would improve operational efficiency by allowing better planning and utilization of resources, including the efficient use of their new printing machine. This, eventually, would increase customer satisfaction through the fulfillment of orders on time, an important factor in the B2C sector where consumer expectations are high. If the problem remains unresolved, CB may continue to face operational disruptions due to irregularities in order fulfillment, leading to inefficiencies and increased operational costs. This could result in customer dissatisfaction, affecting CB's reputation, moreover, without a clear understanding of demand patterns, especially those influenced by weather conditions and seasonality, CB risks missing out on sales opportunities during peak periods.

From solving this problem, professionals and scientists can learn about the advancements in predictive analytics and its application in forecasting demand based on complex factors like weather, seasonality, and consumer behavior. This challenge also offers interdisciplinary insights, combining data science, logistics, and consumer psychology, and highlighting the real-world impact of data science in driving business improvements and strategic decisions. This contributes not only to CB's operational efficiency and customer satisfaction but also advances the field of predictive analytics through practical applications.

1.4 Research objective

The main objective of this research is to improve the operational efficiency of CB by analyzing the impact of business (B2B) and consumer (B2C) behavior on order patterns. This study aims to identify trends, inefficiencies, and opportunities within CB's current

order processing and fulfillment system, focusing on both their traditional B2B model, primarily servicing bookstores, and their expanding B2C operations through platforms like bol.com. By analyzing order data since 2017, alongside the specifics of print-on-demand (POD) services, the goal of the research is to offer practical insights that can boost customer satisfaction, accelerate procedures, and eventually raise turnover rates.

In addressing the operational challenges faced by CB, this research will focus on understanding both short-term and long-term trends within its B2B and B2C segments. For B2B interactions, particularly with bookstores, the study will identify areas for immediate improvement in order processing and fulfillment efficiency and forecast future demands to ensure sustained profitability and growth. In the B2C sector, the emphasis will be on analyzing consumer behavior trends for quick operational improvements and customer satisfaction boosts. By identifying patterns that impact CB's service quality and operational efficiency, the research aims to develop targeted strategies to address these challenges, thus optimizing CB's business model for both present and future scenarios.

Effectively forecasting incoming orders for CB's B2C sector, influenced by weather conditions and consumer behavior, requires a multidisciplinary approach that combines data science, machine learning, time series analysis, and specialized knowledge in logistics and retail. Such an approach ensures the development of predictive models that can anticipate demand fluctuations, incorporating external factors to enhance decision-making and operational efficiency. Understanding the dynamics of consumer behavior and how external factors like weather impact purchasing decisions is important. Furthermore, the development of precise predictive models requires expertise in statistical modeling and forecasting methods.

1.5 Related studies

Previous studies have significantly contributed to the field of order forecasting. For instance, Hyndman and Athanasopoulos (2021) provide a comprehensive overview of forecasting methods, including time series analysis, which is central to order forecasting. The authors delve into various forecasting models, including ARIMA and exponential smoothing, and extend their discussion to more advanced techniques like dynamic regression and hierarchical forecasting. An interesting aspect of their approach is the emphasis on using R, an open-source statistical software, which allows practitioners to apply complex forecasting methods through accessible coding examples. Chase (2013) in "Demand-Driven Forecasting: A Structured Approach to Forecasting," explores the concept of demand-driven forecasting, which marks a shift from traditional forecasting methods towards more agile and adaptive approaches. This addresses the limitations of conventional forecasting techniques in responding to volatile market conditions and emphasizes the importance of incorporating real-time data and market signals. The structured approach outlined in the book includes practical guidelines for implementing demand-driven forecasting in organizations, supported by case studies and examples from different industries. This work highlights the evolving nature of forecasting in the digital age and underscores the need for businesses to adopt more flexible and data-informed strategies to anticipate demand more effectively.

The influence of weather on consumer behavior is well-documented, with significant implications for retail sales forecasting. Murray, Di Muro, Finn, and Popkowski Leszczyc

(2010) found that weather conditions, especially sunlight, can affect consumer spending by altering mood states. Their research indicates that increased sunlight reduces negative effects, leading to higher consumer spending. Additionally, Buchheim and Kolaska (2017) explored how current weather conditions influence consumer decisions to purchase outdoor movie tickets in advance. They found that consumers are heavily influenced by the weather at the time of purchase, which is explained by psychological biases such as projection bias and salience. This further highlights the importance of considering current weather conditions when forecasting future sales. Rose and Dolega (2022) quantified the impact of various weather conditions on retail sales using data from a major UK retailer. Their findings reveal that weather impacts are most pronounced in the summer and spring, with wind speed being the most influential factor. These insights underscore the importance of incorporating weather variables into sales forecasting models to enhance accuracy and reliability. Štulec, Petljak, and Naletina (2019)'s paper explores how weather conditions significantly influence retail sales. The authors discuss the use of weather derivatives as financial instruments to mitigate the adverse impacts of weather variability on retail performance. The paper found peak sales during Easter and Christmas. This highlights the interest in examining data behavior during both fixed events like Christmas and variable dates like Easter. Bahng and Kincade (2012) also analyzes the relationship between temperature and retail sales of seasonal garments. The research focuses on branded women's business wear in the Seoul-Kyunggi area of South Korea. It uses descriptive statistics, correlation analysis, and interviews with merchandisers to understand how temperature fluctuations impact sales.

There is also previous research relevant to CB's challenges specifically. The paper by Tarapata, Nowicki, Antkiewicz, Dudzinski, and Janik (2020), "Data-Driven Machine Learning System for Optimization of Processes Supporting the Distribution of Goods and Services – a case study", introduces a novel machine learning system aimed at improving the distribution processes of goods and services. The paper treats a similar problem as CB, focusing on optimizing distribution processes and predicting staffing needs and order volumes in logistics companies through a data-driven machine learning system. This work offers solutions, including different algorithms and a smart tracking device (IUM), to optimize logistics and distribution. Hlupic, Orescanin, and Petric (2020) paper presents a case study on using the ARIMA model for sales predictions in the wholesale industry, emphasizing the importance of precise sales forecasts for optimizing both sales and logistics processes. This research is relevant to CB's need for predicting the timing of incoming B2B orders, providing a methodological framework for forecasting based on historical sales data.

The paper by Alqatawna, Abu-Salih, Obeid, and Almiani (2023) focuses on employing time-series analysis techniques to accurately forecast the resources needed for logistics companies to meet their objectives and sustain growth. This study uses different models such as SARIMAX, ARIMA, AR, and LSTM to optimize the prediction of order volume during specific periods and determine the staffing requirements for the company. The research is important because it addresses a similar optimization problem as CB, in terms of forecasting staffing needs and order volume in logistics companies. Hirche, Haensch, and Lockshin (2021) study investigates the effects of weather and holidays on retail sales of alcoholic beverages, applying SARIMAX models to forecast sales variations. This approach is relevant to CB's goal of examining consumer behavior in the B2C sector, especially re-

garding how external factors like weather conditions and seasonality influence purchasing patterns. These studies underscore the importance of ongoing research and innovation in developing more accurate, responsive, and efficient forecasting models to meet the evolving needs of businesses and the market.

The use of SARIMA models for capturing seasonal and trend components in time series data and the extension to SARIMAX for incorporating exogenous variables has been well documented. For instance, Kumari and Muthulakshmi (2024) highlights the effectiveness of SARIMA in weather forecasting and suggests future improvements by integrating machine learning techniques. This paper discusses the application of the SARIMA model for weather forecasting, emphasizing the importance of capturing seasonal and trend components in time series data. It outlines defining model parameters (p, d, q, P, D, Q) using autocorrelation and partial autocorrelation functions. Key limitations include handling non-linear patterns and integrating external factors such as atmospheric pressure systems. Elshewey et al. (2022) introduces a hybrid WD-SARIMAX model for temperature forecasting using daily climate data from Delhi. The dataset spans from 2013 to 2017, with 80% used for training and 20% for testing. The WD component reduces data volatility, improving predictability and stability. The SARIMAX model is enhanced by incorporating multi-dimensional components from the WD process. Experimental results demonstrate the superior performance of the WD-SARIMAX model compared to other recent models, with evaluation metrics showing significant improvements.

Regarding the evaluation of the models, studies have shown that among the most used metrics of evaluation, RMSE is the most common one. Liemohn et al. (2021) extensively discusses various metrics used for validating models in the field of geospace environment modeling, with a particular emphasis on accuracy and precision metrics. The Root Mean Square Error (RMSE) is highlighted as one of the most common metrics for accuracy due to its ability to emphasize larger discrepancies between model predictions and observations by squaring the differences before averaging and taking the square root. The paper notes that while RMSE is widely used, it is often compared against other metrics such as Mean Absolute Error (MAE) and standard deviation to provide a more comprehensive understanding of model performance. In "Metrics for Evaluating Data-Model Comparisons in Space Weather", Chase (2013) also discusses RMSE as a primary metric for assessing the accuracy of model predictions. Similar to the previous paper, this work acknowledges the importance of complementing RMSE with other metrics like correlation coefficients to capture the relationship between observed and predicted trends. For these reasons, this research will also use RMSE to validate the model due to its robustness and wide acceptance. However, a business perspective necessitates a more practical metric. Therefore, the percentage of days when the prediction error is less than 5% will also be utilized as it directly aligns with business goals of reliability and consistency in forecasts, providing a clearer indication of model performance in a real-world context. None of the studies treated in the background study used this metric, instead, they focused more on traditional ones such as RMSE, MAE, and MSE.

Other relevant works include Zhou (2023)'s paper presents a method for improving retail sales forecasting by combining trend and seasonality decomposition with the LightGBM algorithm. The approach leverages both LightGBM for capturing complex non-linear relationships and the Prophet model for handling irregularities and seasonality in sales data. Key techniques include hierarchical time series analysis, feature engineering, and the use

of a Tweedie-based loss function to better manage skewed sales distributions. The results show significant improvements in forecasting accuracy, making it a valuable tool for retail sales prediction. Furthermore, Flynn and Greenberg (2012) analyzes the impact of weather on tipping behavior using two years of data from a restaurant in Poughkeepsie, New York. Contrary to prior studies, their findings show no significant relationship between sunshine and tipping rates, suggesting that tipping behavior is more influenced by social norms rather than weather-induced mood changes. This study provides a comprehensive examination of the factors affecting tipping, emphasizing the limited role of weather in this specific economic behavior.

1.6 Research questions

In the pursuit of improving sales forecasting methods within the B2C sector, this research focuses on the complex interplay between weather conditions, seasonality, and sales dynamics. The investigation is structured around a series of key research questions, designed to systematically explore whether integrating weather data into forecasting models can enhance prediction accuracy. Each research question is created with a distinct objective:

1. *What is the impact of weather conditions on sales in the B2C sector?*

This question aims to uncover the extent to which various weather parameters (such as temperature, precipitation, and humidity) influence consumer purchasing behavior and sales trends. Understanding this relationship will enable the development of more accurate forecasting models that can adapt to the dynamic nature of weather-related influences on market demand.

2. *How can the current SARIMAX model, utilizing different features, be modified to incorporate weather data effectively?*

This question involves analyzing the current model's structure, identifying its limitations in handling weather-related data, and proposing modifications to the model to improve its predictive accuracy in the context of sales forecasting.

3. *Which weather-related features have the most significant impact on the accuracy of B2C sales forecasts when incorporated into a SARIMAX model?*

This question seeks to pinpoint which specific weather variables are most influential in predicting B2C sales, thereby guiding the selection of features that should be prioritized for inclusion in the forecasting model to optimize its accuracy.

4. *What preprocessing steps are necessary for weather data to ensure compatibility with the SARIMAX model?*

The aim here is to delineate the preprocessing requirements for weather data before its integration into the SARIMAX model. This includes identifying the necessary steps to clean, normalize, and prepare weather data, ensuring it is in a format that is compatible with the model and conducive to accurate forecasting.

5. *Can the inclusion of variables representing volatile events, such as Easter, Ramadan, and Carnival, improve the accuracy of sales forecasting?*

This question aims to examine whether taking into account different holidays, national events have an impact on sales to improve the organizational measures during these periods.

By addressing these research questions, the study aims to develop an understanding of the weather's impact on sales and contribute to the advancement of forecasting models used in the B2C sector. The goal is to create a more efficient forecasting tool that uses weather data to better predict changes in customer demand.

1.7 Used model

As mentioned, this research will be mainly done utilizing the SARIMAX model since the ultimate goal is to be able to integrate weather data accurately with it. SARIMAX stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous factors model. This model is an extension of the ARIMA (AutoRegressive Integrated Moving Average) model, which is used for time series forecasting. SARIMAX incorporates both seasonal components and exogenous variables, making it highly adaptable for forecasting tasks where seasonal patterns are evident and external factors significantly influence the target variable. The model is composed of different key components:

- AR (AutoRegressive): This part of the model captures the relationship between an observation and some lagged observations.
- I (Integrated): This component refers to the differencing of the observations to make the time series stationary, which means that the properties of the series do not depend on the time at which the series is observed.
- MA (Moving Average): It models the relationship between an observation and a residual error from a moving average model applied to lagged observations.
- Seasonal Elements: These are similar to the AR, I, and MA components but are applied to the seasonal components of the time series.
- X (Exogenous factors): This part allows the model to incorporate external information that could affect the target variable but is not derived from the time series itself.

The model can be expressed with the following equation:

$$\Phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^DY_t = \Theta_q(B)\Theta_Q(B^s)\epsilon_t + X_t\beta \quad (1)$$

where:

- Y_t is the time series at time t .
- B is the backshift operator, such that $BY_t = Y_{t-1}$.
- $\Phi_p(B)$ is the non-seasonal AR polynomial of order p :

$$\Phi_p(B) = 1 - \phi_1B - \phi_2B^2 - \dots - \phi_pB^p \quad (2)$$

- $\Phi_P(B^s)$ is the seasonal AR polynomial of order P with seasonality s :

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps} \quad (3)$$

- $(1 - B)^d$ is the non-seasonal differencing operator of order d .
- $(1 - B^s)^D$ is the seasonal differencing operator of order D .
- $\Theta_q(B)$ is the non-seasonal MA polynomial of order q :

$$\Theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q \quad (4)$$

- $\Theta_Q(B^s)$ is the seasonal MA polynomial of order Q with seasonality s :

$$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \dots - \Theta_Q B^{Qs} \quad (5)$$

- ϵ_t is the white noise error term at time t .
- X_t is the vector of exogenous variables at time t .
- β is the vector of coefficients for the exogenous variables.

SARIMAX is chosen for this research due to its comprehensive approach to modeling time series data that shows seasonal patterns and is influenced by external factors such as weather. Weather data often involves external factors such as geographical location, altitude, and human activities that can influence the outcomes. SARIMAX can incorporate these exogenous variables into the forecasting model, providing more accurate predictions. This model offers the flexibility to model relationships between the target variable and external factors, which is important for dealing with the complexities of weather data.

1.8 Libraries

In order to implement SARIMAX, some libraries will be needed. The primary libraries along with a small description of each are:

- **Statsmodels:** This is the main library for fitting the SARIMAX model, offering extensive functionalities for time series analysis and econometrics.
- **Pandas:** It is essential for data manipulation and preparation, enabling easy handling of time series data before and after model fitting.
- **NumPy:** This library provides support for mathematical functions and operations on arrays, needed for the computational aspects of time series modeling.
- **Matplotlib and Seaborn:** These libraries are used for data visualization, helping in plotting time series, and forecasting results to assess model performance.

1.9 Dataset

The dataset provided by the CB is divided into three distinct parts:

- Sales Data File: This file contains the total sales data, which was the primary focus of this research since the analysis was centered on understanding overall sales trends and forecasting future sales numbers.
- Annual Order Files (2017-2024): Separate files exist for each year from 2017 through 2024, detailing orders placed within each specific year.
- Articles Data File: A file containing details about all the books.

As mentioned, this research mainly focused on using the daily sales data for books spanning from 2017 to 2024. The CSV file is structured into three main columns: "ONT-VANGST DATUM", which records the date of the transaction; "DISTRIBUTION CHANNEL", specifying the sales channel as either B2B (Business to Business) or B2C (Business to Consumer); and "EXEMPLAREN AANT", detailing the number of copies sold on that particular day. The entries are a mix of dates, reflecting the non-sequential daily sales activities, indicating the dataset likely includes various peak sales periods and possibly promotional impacts.

Figure 1 depicts how the file is structured. In the data preprocessing session, this dataset is filtered for B2C transactions only for the analysis of direct consumer purchases.

Filtering this dataset for B2C transactions allows for the analysis of direct consumer purchases. This approach enables the identification of trends, seasonal peaks, and overall consumer demand patterns across the specified years. Integrating this data with weather information can then provide insights into whether environmental factors, such as weather conditions, have a significant impact on consumer behavior. This combined analysis could be fundamental in understanding the external influences on sales performance and refining marketing and operational strategies accordingly.

2 Methods

The methodology of this study is structured into three distinct phases, each aimed at refining the understanding of order forecasting and the capability of weather data to influence sales in the B2C sector. In this part, a breakdown of the methods used in each phase will be provided. However, before diving into the structured three-phase approach to model development, two preliminary steps were taken to lay the foundation for an effective forecasting solution.

2.1 Preliminary steps

2.1.1 Problem analysis

The forecasting challenge, specifically the prediction of order volume, was subjected to an in-depth analysis. An extensive review of existing literature related to the problem of forecasting order volume was conducted. This review encompassed the analysis of various scholarly papers that addressed similar forecasting challenges, focusing particularly on

those that implemented SARIMAX models as part of their methodology. Motivated by these findings, it was decided to also employ a SARIMAX model for this research to evaluate the potential impact of integrating weather data on the accuracy and reliability of order volume forecasts. This exploration aims to uncover whether the inclusion of weather variables could meaningfully improve or possibly degrade the performance of the forecast.

2.1.2 Explore user requirements

Parallel to the problem analysis, a detailed engagement with company stakeholders was initiated to gain an understanding of how the forecasting model would be operationalized. Through discussions and meetings, information regarding the specific needs, preferences, and constraints of the end-users of the forecasting model was collected. This engagement ensured that the model developed would not only address the technical forecasting challenge but also be practically useful and readily adoptable by the business.

These initial steps were important in shaping the direction and focus of the model development effort. Following this foundational groundwork, the research progressed into a structured three-phase approach, each phase designed to build upon the insights gained and to refine the forecasting model progressively. The phases are as follows:

2.2 Phase 1: Base SARIMA model construction

In this initial phase, the focus is on constructing a base model using SARIMA (Seasonal Autoregressive Integrated Moving Average). This model serves as a foundation for further analysis to capture intrinsic sales patterns, trends, and seasonality without the influence of external variables. This is done through:

- **Literature Study:** An extensive review of existing literature was conducted to gather general information, established methodologies, and best practices related to time series forecasting and the specific application of SARIMA models. This foundational knowledge facilitated the informed development of the baseline model.
- **Exploratory Data Analysis (EDA):** The EDA was performed on historical sales data to discover any existing patterns, seasonal trends, and anomalies. The objective was to develop an understanding of the data, which included statistical summaries, visualization of sales trends, and identification of seasonal patterns and outliers. This phase also aimed to generate new research questions that could be explored through model development.
- **Baseline Model Construction:** A SARIMA model was built as a baseline to model the B2C sales data without weather data. The construction of this model involved identifying the appropriate differencing orders (d), autoregressive terms (p), and moving average components (q), as well as assessing the seasonality components (P , D , Q , S) through ACF and PACF plots.

2.3 Phase 2: Integration of exogenous variables into SARIMAX model

Upon the completion of the base model, the second phase focuses on integrating exogenous variables in the SARIMAX model. This phase requires a selection of weather data variables that are hypothesized to impact sales and their subsequent integration into the forecasting model. The process involves a careful examination of the weather data to ensure compatibility with the sales data in terms of frequency and format. After analyzing the weather data, the research will proceed with the integration of particular special events such as Easter, Ramadan, and Carnival into the model to see whether they improve sales performance.

- **Model Construction with Exogenous Variables:** Weather data and other events were integrated into the SARIMAX model as exogenous variables to investigate the impact on sales forecasting accuracy. This involved preprocessing weather data and storing special days to align it with sales data frequency and format, as well as the selection and transformation of relevant weather features based on their potential impact on sales as suggested by the literature and EDA.

2.4 Phase 3: Model comparison and evaluation

The final phase is dedicated to comparing and evaluating the base SARIMA model and the enhanced SARIMAX model. This comparison aims to quantify the added value of integrating weather data and volatile events into the forecasting process and is done through:

- **Model Validation:** The enhanced SARIMAX model is subjected to validation to verify the correctness and relevance of its forecasts in the operational context. This involves both quantitative assessments, such as backtesting against historical data, and qualitative assessments through stakeholder review.
- **Model Evaluation:** Models are evaluated on their predictive performance using metrics such as Mean Absolute error(MAE), Mean Squared Error(MSE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). Additionally, from a business perspective, the proportion of days with a forecast error of 5% or less is calculated.
- **Comparison of Models:** A comparative analysis between the baseline SARIMA and the enhanced SARIMAX models is conducted to determine the added value of incorporating exogenous variables. The final selection of the model was based on a balance of complexity, accuracy, and operational applicability.

This methodological approach, with its phased structure, ensures a systematic and comprehensive assessment of the role of weather data in sales forecasting and allows for an informed decision on the integration of such exogenous variables in the SARIMAX model.

3 Expected deliverables

In this research, the primary output will be the development of a robust forecasting model that integrates weather data and exogenous events to predict future B2C sales orders.

The objective is to construct a working model that not only forecasts sales with high accuracy but also delineates the specific influence of weather variables on these predictions. By carefully selecting and incorporating relevant weather features such as temperature, humidity, and precipitation into the SARIMAX model, the goal is to be able to quantify their impact on sales dynamics.

The expectation is that this model will provide clear and actionable insights, demonstrating whether and how meteorological conditions and seasonality affect consumer purchasing behavior and sales outcomes. The model's accuracy and predictive quality are important, as they must be sufficient to confidently assess the influence of weather. A well-calibrated model will allow us to find out whether changes in weather patterns contribute significantly to variations in sales, thereby validating the initial hypotheses. Lastly, the success of this model will be measured by its ability to produce reliable sales forecasts that can inform and optimize business strategies in the B2C sector.

4 Data preprocessing

4.1 Data Selection

The initial stage of data preprocessing involved filtering the dataset to focus on business-to-consumer (B2C) transactions. This filtering was based on the need to analyze consumer sales patterns, which are distinct from business-to-business (B2B) transactions in terms of volume, frequency, and influencing factors. For this reason, only the data where the `DISTRIBUTION_CHANNEL` was identified as 'B2C' was kept for further analysis.

4.2 Feature Extraction

Following the selection of relevant data, additional time components were extracted from the transaction dates to aid in the time series analysis. The `ONTVANGST_DATUM` field, originally in string format, was converted into a `DateTime` object to facilitate the extraction of temporal components. These components included:

- Year: Extracted to analyze annual trends and accommodate any year-over-year changes in consumer behavior.
- Month: Important for assessing seasonal trends and monthly performance fluctuations.
- Day: Used to examine potential patterns occurring on specific days within the month.

These components were added as separate columns in the dataset, improving the data's granularity for the analysis.

4.3 Data Exploration

An initial exploration of the data was conducted to understand the underlying patterns within the B2C sales data. This foundational analysis aimed to identify the key trends and seasonality in the sales data, which spans, as previously mentioned, from 2017 to

2024. This preliminary exploration was important in identifying critical periods of sales activity and in assessing the overall trajectory of the business's performance over the period. By aggregating sales data on a daily and monthly basis, and employing time series decomposition techniques, the structure of the data was defined, highlighting significant seasonal peaks, and determining long-term trends, preparing the ground for the creation of the SARIMAX model.

Here is an overview of how the data exploration was conducted:

- **Analysis of Sales Over Time (Figure 2):** This initial visual representation of daily B2C sales from 2017 through 2024 provides an in-depth look at the operational volatility of sales. This view is essential for identifying outlier events and understanding daily sales variability, which can impact supply chain decisions and inventory management.
- **Monthly Sales Aggregation (Figure 3):** To further explore seasonality, B2C sales from all the years were aggregated monthly, providing a visual representation of the total copies sold each month across the years. The goal was to identify any consistent trends that occur throughout the year. From the plot, it can be seen that there is a significant increase in sales towards the end of the year, particularly in November and December, likely due to the holiday season which traditionally boosts consumer spending. On the other hand, sales seem to be at their lowest around February and March, reflecting a post-holiday decrease in consumer purchasing. This pattern repeats annually, suggesting strong seasonal influences on consumer behavior in the B2C segment.
- **Decomposition of Sales Data (Figure 4):** The sales data was decomposed into three key components: trend, seasonality, and residuals. This decomposition provides some strategic insights as:
 1. **Trend Component:** Shows whether the sales volume is increasing, stable, or decreasing over the years. As it can be seen, there has been a gradual increase over the years, indicating steady growth in sales volumes especially with the beginning of the Covid pandemic. There appears to be a slight flattening or even a slight decline towards the end of 2023 into 2024, which could suggest a plateau or a slowdown in growth.
 2. **Seasonal Component:** Highlights specific times of the year when sales increase or decrease significantly. There is a clear and consistent seasonal pattern in sales volumes that repeats annually, especially towards the end of each year.
 3. **Residuals:** Analyzing the residuals helps identify dates when actual sales deviated from predicted values. The observed residuals do not show any systematic pattern over time, suggesting that the model has captured most of the systematic information. However, there are some larger residuals observable particularly in later years (post-2022). These could be due to unexpected events changes in the market or consumer behavior that are not captured by the trend or seasonal components.

4.4 Rolling mean and standard deviation

In the visualization presented in Figure 5, the analysis of the sales data incorporates the calculation and plotting of the used mean and standard deviation alongside the original sales data. This graph is useful in assessing the temporal dynamics of sales, highlighting trends and volatility over time. The blue line represents the actual number of copies sold daily from 2017 through 2024. The data exhibits significant fluctuations, indicating variability in sales volume across the observed period.

The red line, namely the rolling mean, calculated over a defined window, smoothens the original data to reveal underlying trends by averaging out short-term fluctuations. In particular, the rolling mean helps in identifying broader yearly trends, showing periods of increase or decrease in sales that are not immediately apparent from the original data.

Lastly, the green line is the rolling standard deviation and it measures the variability in sales data within the same moving window used for the mean. It provides insights into the consistency of the sales volume; higher values indicate greater volatility, suggesting significant changes in sales volume during those periods.

4.5 Lags

Figure 6 displays scatter plots comparing sales data at different time lags, specifically $y(t)$ vs. $y(t+1)$, $y(t)$ vs. $y(t+12)$, and $y(t)$ vs. $y(t+24)$. These plots are important for understanding the autocorrelation in the sales data, which indicates how values are related to their previous values over different time lags. This kind of analysis helps in identifying patterns that can be relevant for forecasting and modeling the time series data.

- $y(t)$ vs. $y(t+1)$: This scatter plot shows the relationship between the sales data at time t and the sales data at the next time point $t + 1$. The plot appears to show a positive correlation, as expected because sales from one day are likely to be somewhat similar to sales from the following day. This immediate lag can often capture short-term trends or daily fluctuations in the data.
- $y(t)$ vs. $y(t+12)$: The second plot compares the sales data at time t with the sales data 12-time units later ($y(t+12)$). This plot is particularly interesting as it does not exhibit as clear a pattern as $y(t)$ vs. $y(t+1)$, reflecting a more dispersed distribution of points. Given the context of monthly data, this lag likely represents the sales from the same month in consecutive years (if assuming monthly data), revealing the year-over-year stability or variability in sales, as it was seen in Figure 3.
- $y(t)$ vs. $y(t+24)$: The third scatter plot shows the relationship between sales at time t and sales two years later ($y(t+24)$). This plot shows a broader spread but still maintains a positive correlation, suggesting that while sales figures two years apart are related, the variance increases with the length of the lag. This indicates more significant changes over a two-year period, which could be due to broader market trends, economic cycles, or changes in consumer behavior.

5 Data split

In this section, the process of splitting the data into training and testing sets before applying the model is explained. The data was split using a scheme designed to roughly allocate 80% of the data for training and the remaining 20% for testing, similar to the splitting done by Elshewey et al. (2022). This approach was chosen to ensure that a substantial portion of the data is used for training the model, while still retaining a significant amount of data for robust testing.

The split function calculates the total number of days in the dataset and allocates approximately 83.5% of these days for training. The end date of the training set is determined based on this percentage, and the dataset is subsequently divided into training and testing sets accordingly. The choice of 83.5% for training ensures that the model is trained on a significant portion of the data, while the testing is conducted on data starting from the beginning of 2023. This allows the assessment of the model’s performance on more recent data, ensuring that the testing phase evaluates the model’s predictive capabilities on data that it has not seen during training.

6 Models

6.1 SARIMA Base Model

In the development of the SARIMA model, initial analyses were focused on understanding the underlying patterns within the sales data. The autocorrelation function (Figure 7) and partial autocorrelation function (Figure 8) were meaningful in identifying the nature of the data’s temporal structure. These functions highlighted significant autocorrelations at specific lags, suggesting the influence of both immediate past values and seasonal effects, which recur over regular intervals. This insight was relevant in determining the appropriate parameters for the SARIMAX model, ensuring that both short-term dependencies and long-term seasonal trends were accounted for in the model’s formulation.

As previously mentioned, an exploratory analysis was conducted using lag plots (Figure 6). By visually assessing the correlation between sales data at different time lags, it was possible to refine the model’s specifications. These plots validated the selection of a seasonal period of 12 months, aligning with the apparent annual sales cycle, which is particularly influenced by key retail periods and consumer behavior patterns.

The base SARIMAX model was configured with parameters determined from these analyses, specifically a non-seasonal order of (1,1,1) and a seasonal order of (1,1,1,12). This model structure was chosen to capture both non-seasonal and seasonal effects adequately, with the order parameters representing autoregressive terms, differences, and moving average components respectively.

Finally, the model’s forecasting capability was evaluated through a rolling forecast approach where predictions are made 7 days ahead based on the past 30 days of data. The resultant Figure 9 shows the model’s effectiveness in capturing the general trend and fluctuations in sales, though there is room for improvement in peak prediction accuracy. This approach is useful in dynamic business environments where frequent updates to forecasts are necessary to adapt to new data and changing market conditions. The plots showed

that while the model captured the overall trends and seasonality effectively, the peaks and troughs were not always perfectly aligned with the actual data, suggesting areas for further model refinement.

6.2 SARIMA Base Model Optimized

After evaluating the performance of the base SARIMA model, an exploration to optimize the model was conducted by employing the ‘auto_arima’ function to systematically search for the most effective parameters. To streamline the search process and ensure computational efficiency, the range for potential values for both p and q was restricted from 0 to 2. This limitation was based on the reason that extending the parameter range further would likely yield diminishing returns in model performance improvement relative to the increased computational cost and complexity.

The ‘auto_arima’ function determined the optimal non-seasonal parameters to be (1,0,0) and the seasonal parameters as (0,0,1,12). The choice of these parameters suggests that the best model involves a simple autoregressive component of order 1 for the non-seasonal part and a moving average component of order 1 on the seasonal cycle of 12 months, reflecting yearly patterns without the need for differencing either in seasonal or non-seasonal components. These parameters aim to capture the essential temporal structure in the data while maintaining model simplicity and efficiency.

Indeed, the optimized model parameters led to improved performance (Figure 10), which will be discussed in detail in the results section. This section will also explore how the adjustments to the model parameters affected the overall forecasting accuracy and will assess the impact of these changes through different performance metrics.

6.3 SARIMAX Model: weather

In this part of the research, the analysis extends to address the questions related to the impact of weather conditions on sales. Before integrating weather data into the predictive model, a preliminary analysis was conducted to explore potential correlations between various weather conditions and the number of orders. Utilizing data accessed via the Open Weather API, this study examines correlations within the past year.

The following regression plots display the relationship between some features that could be important such as daily mean temperature, humidity, precipitation, and wind speed against the number of orders.

- Temperature vs. Sales (Figure 11): The analysis shows that as temperature increases, there appears to be a slight decline in sales, which aligns with previous observations that sales peak towards the cooler, end-of-year months, a period characterized by holiday shopping and seasonal promotions.
- Humidity vs. Sales (Figure 12): The plot indicates a moderate upward trend, suggesting that higher humidity levels could correlate with increased sales. However, the scatter of data points is wide, implying that while there may be a relationship, it is not strong enough to conclusively affect sales predictions.
- Precipitation vs. Sales (Figure 13): There is a flat trend line across the precipitation data, indicating no clear impact of rain or snowfall on sales volumes. This lack of

correlation further diminishes the argument for weather as a direct sales driver in this market context.

- Wind Speed vs. Sales (Figure 14): Similarly, wind speed shows no significant correlation with sales, suggesting that unless extreme weather conditions prevail, normal fluctuations in wind do not influence buying behaviors.

The findings from these regression analyses indicate that while certain weather conditions show very mild correlations with sales, these are not sufficiently robust to guarantee a direct inclusion of weather data in the primary sales forecasting models. This insight aligns with the observed seasonality and annual sales trends, where factors such as consumer behavior during specific times of the year, marketing efforts, and holiday seasons have a more significant impact on sales. These conclusions suggest that future model improvements should focus less on integrating weather data and more on refining the understanding of seasonal patterns and promotional impacts. This strategic focus will enable more accurate forecasting and resource allocation, enhancing the company's ability to capitalize on predictable peak periods without unnecessary weather variables, which have shown limited predictive value.

6.4 SARIMAX Model: volatile events

Since the previous analysis revealed no significant correlation between sales and weather variables, it was decided not to pursue it further, as it was unlikely to yield actionable insights for the company. Instead, the focus shifted towards understanding seasonality, particularly by integrating events with variable dates such as Easter, Ramadan, and Carnival. These events, due to their fluctuating occurrence each year, introduce a level of volatility that makes them particularly interesting for enhancing the model's ability to predict significant deviations in sales patterns. Exploring these special events is also important because they can significantly impact consumer behavior and sales trends, making them valuable predictors in sales forecasting models. This is the reason for choosing volatile events instead of events such as Christmas or Sinterklaas that have fixed dates and therefore are easier to predict.

To integrate the impact of special events into the forecasting model, the initial step involved augmenting the dataset with additional columns representing the occurrence of volatile events such as Easter, Ramadan, and Carnival. These events, which occur on different dates each year, were mapped out from 2017 to 2024. Dates were first identified for each event and then converted into the datetime format to ensure accuracy in temporal data handling. Afterwards, for each event, a new column was created in the dataset. On the exact dates these events occurred, the column was populated with a '1' to indicate the presence of the event, otherwise, it remained '0'. This methodical approach enables the incorporation of these significant but variably timed events into the predictive modeling, facilitating a nuanced analysis of their potential impact on sales dynamics.

This enrichment of the dataset was needed for capturing the temporal variability of these events, which do not occur on fixed calendar dates each year. By marking each event in the dataset for the respective dates it occurred over the years, it was possible to directly correlate sales data with these events, providing a structured way to assess their impact on sales trends.

6.4.1 Easter

Incorporating Easter as a volatile event in the forecasting model is important due to its significant and irregular impact on sales patterns. Unlike fixed-date holidays, Easter varies each year, falling between late March and late April. This variability introduces a unique challenge in forecasting, as consumer behavior around Easter can significantly deviate from typical patterns. Ignoring this effect could lead to inaccuracies in the model, as the sales data during the Easter period would not follow the usual trends and seasonality.

Figure 15 shows how the model behaves after the integration of Easter as an exogenous variable. By explicitly including Easter, the model can better account for these irregular spikes in sales, therefore enhancing the accuracy and reliability of the forecasts. This adjustment ensures that the model does not underestimate or overestimate sales during this period, providing more precise predictions that reflect the true impact of this significant holiday.

6.4.2 Ramadan

Another cultural event that could be analyzed because of its volatility is Ramadan. Ramadan is a month-long period of fasting, reflection, and community, with its timing based on the Islamic lunar calendar, causing it to shift approximately 10 days earlier each year on the Gregorian calendar. Sales might increase in the days leading up to Ramadan and during the Eid al-Fitr celebration at its conclusion, driven by the purchase of food, gifts, and other items.

Figure 16 illustrates the performance of the SARIMAX model after including Ramadan as an exogenous variable. An improved alignment between the predicted and actual sales can be seen and this indicates the effectiveness of accounting for such volatile events in enhancing the forecasting accuracy when compared to the base model that did not include any exogenous variables.

6.4.3 Carnival

Carnival, often characterized by parades, festivals, and public festivities, typically occurs in the weeks leading up to Lent. The timing of Carnival varies each year, dependent on the liturgical calendar, making it a movable feast with dates that can shift by several weeks. For this reason, it was decided to also integrate Carnival into the forecasting method to find some insights regarding its influence on sales. Figure 17 depicts how the model behaves after integrating Carnival as an exogenous variable. Even though the red line, representing the predicted forecast, doesn't seem to follow the blue line (testing set) properly, the model performed better than the base model. This will be further discussed in the following results section. Overall, ignoring Carnival's impact could result in forecasting errors, as the sales data during this period would exhibit atypical patterns not captured by standard seasonal trends since the event is volatile.

6.4.4 Easter, Ramadan and Carnival

After analyzing the models with each individual event (Easter, Ramadan, and Carnival) as exogenous variables, it was considered valuable from a business perspective to examine

the model's performance when all these events are integrated simultaneously as exogenous variables.

Figure 18 shows the comparison between the testing set and the forecasted values when Easter, Ramadan, and Carnival are included as exogenous variables in the model. The blue line represents the actual number of orders, while the red line represents the forecasted values. By including all three events, the model captures the combined effects of these significant periods, potentially leading to a more accurate forecast. This integrated approach can provide businesses with a comprehensive understanding of how multiple events influence sales simultaneously, helping to more effectively decision-making and planning.

6.5 Combined model: Optimized SARIMA and Seasonality

In the pursuit of refining the forecasting model, a combined approach was implemented that merged the seasonally adjusted predictions with the optimized SARIMA model, as seen in Figure 19. The inclusion of both seasonal adjustments and optimized parameters aimed to create a model that not only captures the general sales trends but also adapts to seasonal peaks and troughs effectively. This dual approach allowed for a more nuanced understanding of the sales dynamics over the period under study, potentially offering more accurate and actionable insights for strategic planning and decision-making in the business context.

To account for seasonal variations, a seasonal index using the training data was calculated. The average number of orders for each month was computed and compared to the overall monthly mean to derive a multiplicative seasonal index. These seasonal adjustments were then applied to the optimized SARIMA model predictions in two ways:

- **Multiplicative Adjustment:** Directly adjusting predictions based on the seasonal index.
- **Dampened Multiplicative Adjustment:** A dampening factor of 0.5 was used in the dampened adjustment. This factor reduces the influence of the seasonal adjustment by half, ensuring that the seasonal pattern is considered without overly skewing the predictions.

The graph 19 illustrates a comparison between actual sales data, original predictions, and predictions adjusted for seasonal variations over a specified period. The actual data, represented by the solid blue line, shows the real-world sales figures recorded, providing a baseline for assessing the accuracy of the forecasting models.

The original predictions, marked by the dashed red line, represent the output from the forecasting model before any seasonal adjustments were applied. These predictions capture the general trend and cyclical behavior of the sales data but may not align closely with the actual peaks and troughs observed, particularly during periods of significant seasonal influence.

The adjusted predictions, shown with the solid green line, incorporate seasonal adjustments to better align with the known seasonal patterns in the data, such as increased sales around specific holidays or events that recur annually. This adjustment appears to enhance the model's accuracy, as indicated by the green line's closer adherence to the actual sales peaks compared to the original predictions.

This visual comparison underscores the importance of accounting for seasonal variations within predictive models, especially when dealing with data that exhibits strong seasonal trends. The adjusted model not only provides a more accurate reflection of expected sales figures but also offers more reliable insights for planning and resource allocation in response to anticipated demand fluctuations.

7 Model Evaluation

The focus of this section is to validate different models, including those that integrate exogenous variables, to determine if they enhance the accuracy of sales predictions. Specifically, the validation will cover the effectiveness of the models in forecasting sales over a 7-day horizon based on data from the preceding 30 days. By evaluating metrics such as RMSE, MAE, and MAPE, this analysis aims to establish whether the inclusion of variables like Easter, Ramadan, and Carnival dates contributes to more precise sales forecasts.

From a business perspective, the model validation will also include an analysis based on the days where the error is less than 5%. This specific measure helps to provide a clear and tangible understanding of the model's accuracy in practical, in business terms. This approach aligns the technical evaluation of the model with business outcomes, making the results more relevant and actionable for decision-making within the company.

7.1 Evaluation Metrics

Here are some additional insights regarding the evaluation metrics that were used to validate the models :

- Mean Absolute Error (MAE): This metric represents the average absolute difference between the forecasted values and the actual sales numbers.
- Mean Squared Error (MSE): MSE is particularly critical as it gives a sense of the error magnitude by squaring the differences between predicted and actual values, therefore emphasizing larger errors more than smaller ones due to the squaring component.
- Root Mean Squared Error (RMSE): This is the square root of MSE and provides error terms in the same units as the data, making it more interpretable.
- Mean Absolute Percentage Error (MAPE): It calculates the average of the absolute differences between predicted and actual values, expressed as a percentage of the actual values.
- Percentage of Days with Error within 5%: These are the number of days for which the absolute percentage error between the forecasted and actual values is 5% or less.

7.2 SARIMA Base Model

As it can be seen in table 1 the base SARIMA model, configured with initial non-optimized parameters, exhibited an MAE of 9215, an MSE of approximately 134774207, and an RMSE of 11609. These figures highlight significant variability in the model's predictions

compared to actual sales data, which is also reflected in a MAPE of 47.52%, indicating substantial average errors relative to the actual sales values. Furthermore, this model only achieved a prediction accuracy within 5% of actual sales for approximately 15% of the days, underscoring a need for parameter optimization to enhance model performance.

7.3 SARIMA Base Model Optimized

Upon optimizing the parameters using `auto.arima`, the model showed improved performance. The MAE was reduced to 8259, the MSE to 123410261, and the RMSE to 11109. The MAPE significantly decreased to 28.00%, reflecting closer alignment with the actual sales figures. Notably, the percentage of days with prediction errors within a 5% margin increased to approximately 18%, indicating a better fit of the model to the sales data.

7.4 SARIMAX Model: volatile events

The inclusion of Easter, Ramadan, and Carnival as exogenous variables aimed to capture sales fluctuations associated with these events. Table 1 also shows how each model integrating volatile events performed and in this following part a brief summary will be provided.

7.4.1 Easter

The model incorporating Easter alone shows a relatively lower MAE (8248), MSE (123076860), and RMSE (11094) compared to other models, indicating it performs better in absolute terms. The percentage of days with prediction errors within 5% is also the highest at 18% days, and the MAPE is 27.92%, showing a better fit in terms of percentage error.

7.4.2 Ramadan

The model with Ramadan as the exogenous variable has slightly higher MAE, MSE, and RMSE compared to the Easter-only model, indicating a somewhat less accurate prediction. The number of days with prediction errors within 5% is lower at 17% days, and the MAPE is slightly higher at 28.20%, suggesting a marginally less accurate model compared to the Easter-only model.

7.4.3 Carnival

When incorporating Carnival as the exogenous variable, the model exhibits a noticeable increase in MAE, MSE, and RMSE, indicating less accurate predictions. The number of days with prediction errors within 5% is almost 17% days, and the MAPE is higher at 30.81%, showing a decrease in prediction accuracy compared to both Easter and Ramadan models.

7.4.4 SARIMAX: Easter, Ramadan, and Carnival

Further expanding the model to include Carnival along with Easter and Ramadan resulted in an MAE of 9238, an MSE of 153680491, and an RMSE of 12397. The MAPE also increased to 30.99%, and the model achieved prediction accuracy within 5% for about

16% of the days. This indicates that while the inclusion of additional seasonal indicators may capture broader variability, it could introduce complexity that slightly degrades the model’s overall predictive accuracy.

7.5 Combined model: Optimized SARIMA and Seasonality

The results from this integrated model underscore its effectiveness in capturing the dynamic aspects of the sales data. The mean absolute error (MAE) was reported at 8,710.31, and the root mean squared error (RMSE) stood at 12,857.44, which indicates a substantial reduction in prediction error compared to the initial base model. The mean squared error (MSE) was calculated at approximately 165,313,862.91. Furthermore, the mean absolute percentage error (MAPE) improved to 28.47%, demonstrating a more accurate model performance relative to the scale of the data.

Additionally, the model’s reliability in day-to-day operations is evidenced by the fact that only about 16.67% of the days exhibited prediction errors exceeding 5%. This metric is particularly critical from a business point of view as it provides a benchmark for assessing the model’s operational viability.

8 Discussion

The results from the various SARIMAX models tested demonstrate a clear trajectory of improvement and adaptation as modifications are made to the model’s structure and parameters. Initially, the base SARIMA model’s performance metrics indicated substantial room for improvement, particularly in terms of predictive accuracy as reflected by high MAE, MSE, RMSE, and MAPE values. The introduction of optimized parameters markedly enhanced the model’s efficiency, reducing error metrics and increasing the days where the model’s predictions fell within a 5% error margin.

The investigation into the integration of external variables such as weather conditions and significant events into sales forecasting models for the B2C sector has yielded several insights and implications for future research. Despite initial considerations to include weather variables in the forecasting model, the preliminary analysis showed a negligible correlation between weather conditions and B2C sales. This lack of significant correlation led to the decision not to integrate weather data into the final models, underscoring the importance of data relevance in forecasting practices. The SARIMAX models, while robust for time series forecasting, demonstrated limited efficacy in integrating non-time-series external data such as weather conditions. The study highlights the model’s limitations in accommodating external variables that do not directly interact with typical sales patterns.

In contrast to weather data, the integration of cultural events (e.g., Easter, Ramadan, and Carnival) that have non-fixed dates throughout the year showed a slight improvement in forecasting accuracy. This improvement suggests that such events, which directly influence consumer purchasing behaviors, are more relevant and should be prioritized in forecasting models. However, the subsequent addition of Carnival alongside Easter and Ramadan did not yield the expected improvement and, in some metrics, reversed some of the gains made by previous models. This suggests that while the inclusion of specific exogenous factors can potentially refine the model’s outputs, there is a complex interplay

between these variables that may not always produce linear improvements in forecasting accuracy.

From a business perspective, the key takeaway is the importance of continuous model refinement and the cautious integration of exogenous variables based on thorough testing and validation. The objective remains not only to minimize error in forecasting but also to enhance the model's practical applicability by increasing the number of days with minimal prediction errors. Businesses should focus on refining forecasting models to better capture the effects of specific events or holidays that have a proven impact on consumer behavior. This is important for effective inventory management, resource allocation, and strategic planning in a sales-driven environment.

Research could explore integrating a wider range of external variables, such as economic indicators or social trends, which may offer more substantial insights into sales fluctuations. For future studies, it would also be valuable to consider that the volatile events that were analyzed in this paper, namely Easter, Ramadan, and Carnival, do not have the same duration. For instance, Ramadan spans over an entire month, while Carnival lasts for about three days, and Easter typically affects a few days around the holiday itself. Incorporating the varying durations of these events could lead to more accurate and nuanced forecasting models. By accounting for the different time spans, future research can better capture the true impact of each event on the metrics being studied. This approach will help in refining the predictive capabilities and enhance the robustness of the models.

This study illustrates the complexities involved in integrating external variables into sales forecasting models and emphasizes the necessity of empirical support for the variables used. The exclusion of weather data from the final models was based on its demonstrated lack of impact on sales predictions, shifting the focus towards variables with more direct relevance to consumer behavior. Overall, while advancements have been made, the results suggest a need for ongoing adjustments and possibly exploring additional modeling techniques or data inputs to further refine the accuracy and reliability of sales forecasts. This iterative approach is essential for developing a robust forecasting model that aligns closely with the dynamic nature of sales data and the strategic goals of the business.

9 Conclusion

The comprehensive analysis of the models and outcomes presented in this research offers valuable insights into the influence of external variables on B2C sales forecasting, particularly concerning weather conditions and volatile events like Easter, Ramadan, and Carnival.

The initial research question aimed to explore the relationship between weather variables and sales trends. Despite rigorous testing and the integration of various weather-related data into the SARIMAX model, the results consistently showed a negligible correlation between weather conditions and B2C sales. This lack of significant impact suggests that while weather may influence individual consumer behaviors to some extent, it does not translate into a discernible pattern in overall sales data that could be leveraged to enhance forecasting accuracy.

Regarding the modification of the SARIMAX model to effectively incorporate weather data, due to the lack of correlation in the preprocessing phase, the research was not

continued in this direction. Therefore, the integration of weather data in the current modeling approach and data set does not significantly enhance forecast precision.

The third research question focused on identifying which weather-related features could potentially impact the accuracy of B2C sales forecasts. Although initial hypotheses suggested that variables such as temperature, humidity, and precipitation could be influential, the empirical analysis did not support this. None of the tested weather features consistently improved the model's predictive accuracy, suggesting a more complex interplay of factors influencing consumer purchasing decisions that might not be captured solely through weather data.

Preprocessing steps for weather data integration posed significant challenges, primarily due to compatibility issues with the SARIMAX model. Ensuring data cleanliness, normalization, and appropriate formatting were essential steps undertaken; however, as mentioned, even well-preprocessed data did not yield the expected results in model performance.

Lastly, the inclusion of variables representing significant volatile cultural events like Easter, Ramadan, and Carnival provided a more promising direction. The models that integrated these variables demonstrated a slightly improved accuracy over the base model, indicating that such events, which directly affect consumer behavior, could be more relevant predictors for sales forecasting in the B2C sector. This suggests a potential avenue for future research to explore other significant events and their impact on sales to refine the forecasting models further.

In summary, while the exploration of weather data integration into sales forecasting models did not yield the expected results, the analysis has underscored the importance of targeting more directly influential variables such as major cultural events that do not have fixed dates during the year. This shift in focus could guide future studies to improve the accuracy of forecasting models, providing businesses with more reliable tools for strategic planning and operational adjustments during key sales periods.

References

- Alqatawna, A., Abu-Salih, B., Obeid, N., & Almiani, M. (2023, Jul). Incorporating time-series forecasting techniques to predict logistics companies' staffing needs and order volume. *Computation*, *11*(7), 141. doi: 10.3390/computation11070141
- Bahng, Y., & Kincade, D. H. (2012, May). The relationship between temperature and sales. *Int. J. Retail Distrib. Manag.*, *40*(6), 410–426.
- Buchheim, L., & Kolaska, T. (2017, November). Weather and the psychology of purchasing outdoor movie tickets. *Manage. Sci.*, *63*(11), 3718–3738.
- Chase, C. (2013). *Demand-driven forecasting: A structured approach to forecasting*. Wiley.
- Elshewey, A. M., Shams, M. Y., Elhady, A. M., Shohieb, S. M., Abdelhamid, A. A., Ibrahim, A., & Tarek, Z. (2022, December). A novel WD-SARIMAX model for temperature forecasting using daily delhi climate dataset. *Sustainability*, *15*(1), 757.
- Flynn, S. M., & Greenberg, A. E. (2012, March). Does weather actually affect tipping? an empirical analysis of time-series data¹. *J. Appl. Soc. Psychol.*, *42*(3), 702–716.
- Hirche, M., Haensch, J., & Lockshin, L. (2021, Jan). Comparing the day temperature and holiday effects on retail sales of alcoholic beverages – a time-series analysis. *International Journal of Wine Business Research*, *33*(3), 432–455. doi: 10.1108/ijwbr-07-2020-0035
- Hlupic, T., Orescanin, D., & Petric, A.-M. (2020, Sep). Time series model for sales predictions in the wholesale industry. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. doi: 10.23919/mipro48935.2020.9245255
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting principles and practice*. Otexts, Online Open-Access Textbooks.
- Kumari, S., & Muthulakshmi, P. (2024). SARIMA model: An efficient machine learning technique for weather forecasting. *Procedia Comput. Sci.*, *235*, 656–670.
- Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., & Mukhopadhyay, A. (2021, July). RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *J. Atmos. Sol. Terr. Phys.*, *218*(105624), 105624.
- Murray, K. B., Di Muro, F., Finn, A., & Popkowski Leszczyc, P. (2010, November). The effect of weather on consumer spending. *J. Retail. Consum. Serv.*, *17*(6), 512–520.
- Rose, N., & Dolega, L. (2022, March). It's the weather: Quantifying the impact of weather on retail sales. *Appl. Spat. Anal. Policy*, *15*(1), 189–214.
- Štulec, I., Petljak, K., & Naletina, D. (2019, July). Weather impact on retail sales: How can weather derivatives help with adverse weather deviations? *J. Retail. Consum. Serv.*, *49*, 1–10.
- Tarapata, Z., Nowicki, T., Antkiewicz, R., Dudzinski, J., & Janik, K. (2020). Data-driven machine learning system for optimization of processes supporting the distribution of goods and services – a case study. *Procedia Manufacturing*, *44*, 60–67. doi: 10.1016/j.promfg.2020.02.205
- Zhou, T. (2023). Improved sales forecasting using trend and seasonality decomposition with LightGBM.

Appendix

Figures

	ONTVANGST_DATUM	DISTRIBUTION_CHANNEL	EXEMPLAREN_AANT
0	16-06-2020 00:00	B2C	44427
1	05-08-2019 00:00	B2B	201430
2	08-11-2017 00:00	B2B	193427
3	28-08-2017 00:00	B2B	182087
4	08-07-2020 00:00	B2C	38662
5	21-05-2021 00:00	B2B	162215
6	24-07-2023 00:00	B2C	39406
7	18-06-2020 00:00	B2C	40266
8	13-11-2022 00:00	B2C	43671
9	18-10-2023 00:00	B2B	120570

Figure 1: Example of data

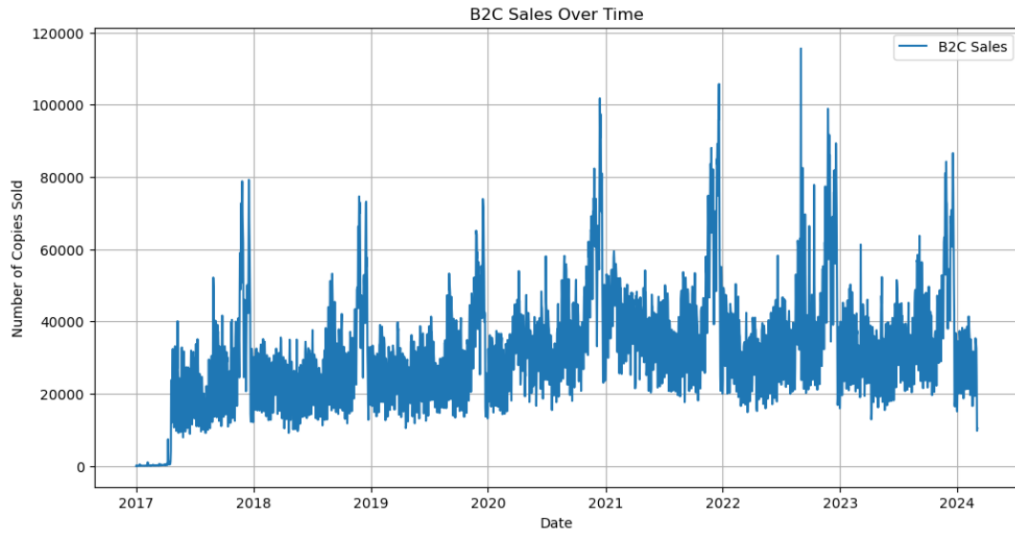


Figure 2: Sales Data Over Time

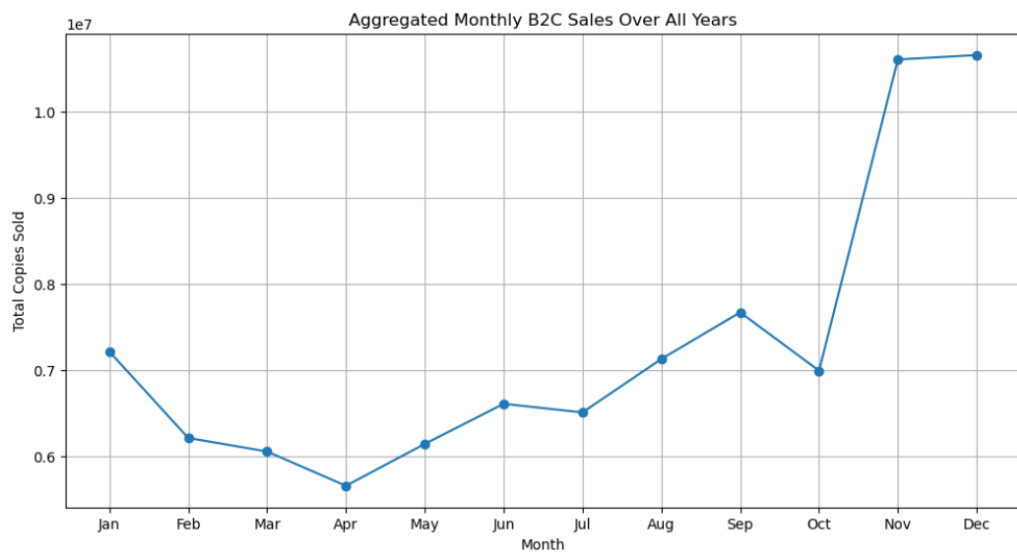


Figure 3: Monthly Sales Over Time

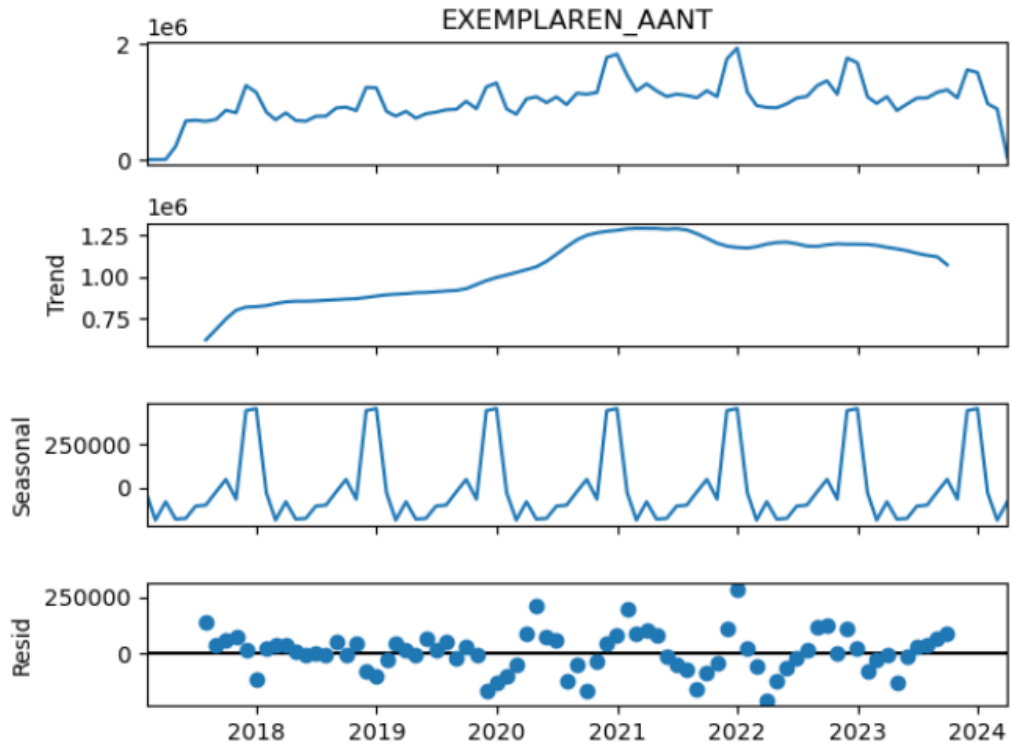


Figure 4: Decomposition: trend, seasonality, and residuals

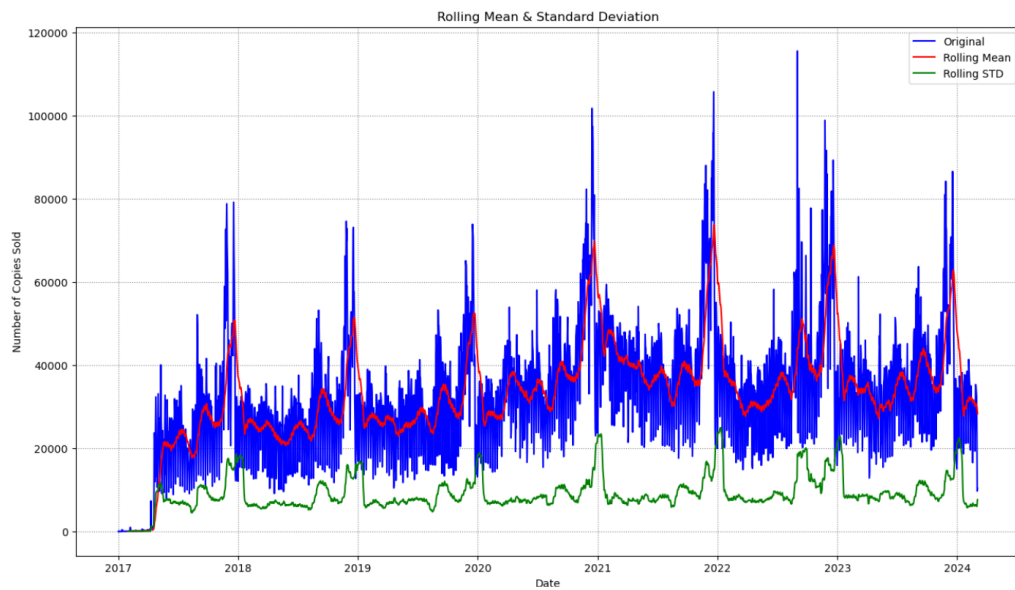


Figure 5: Rolling mean and standard deviation

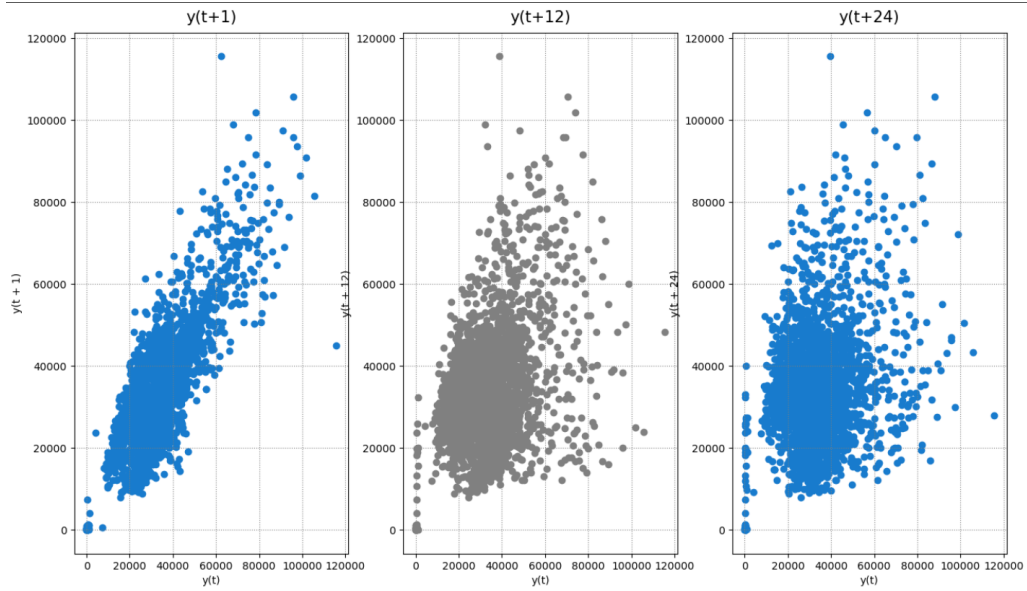


Figure 6: Lag analysis

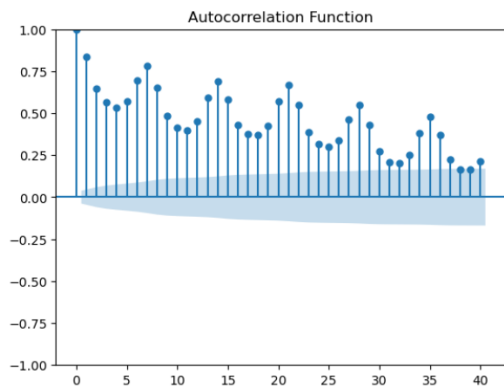


Figure 7: ACF

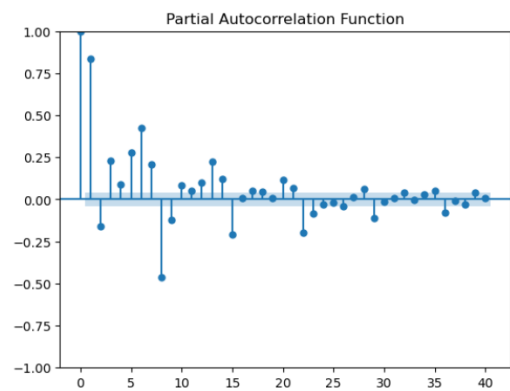


Figure 8: PACF

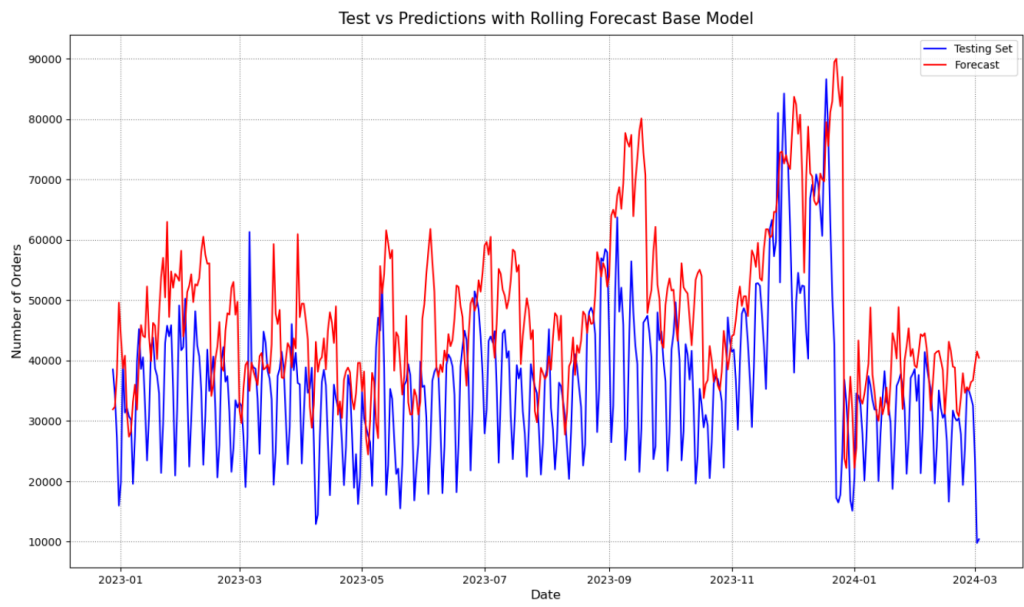


Figure 9: SARIMA Base Model

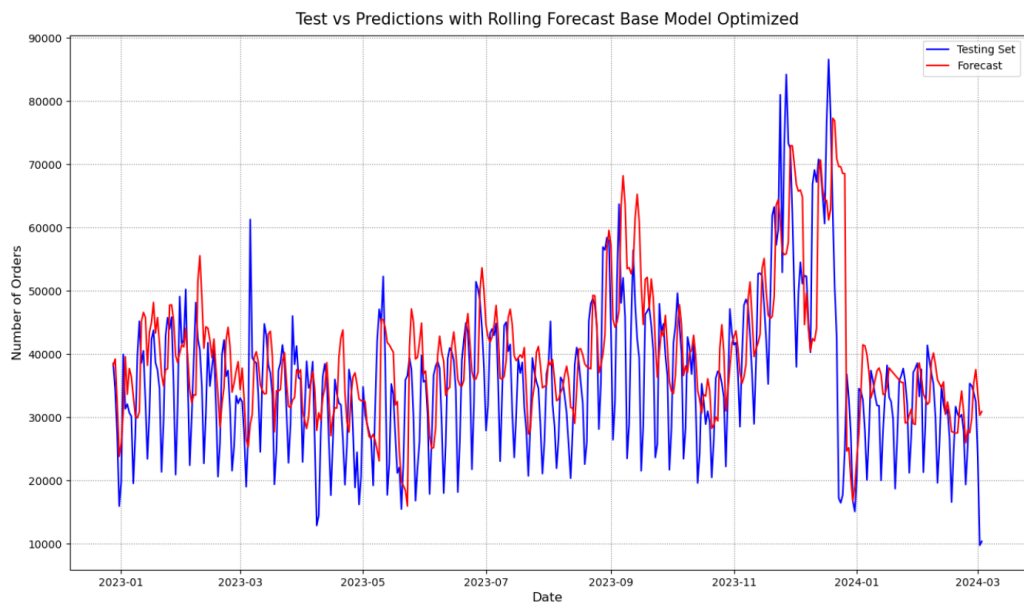


Figure 10: SARIMA Optimized Model

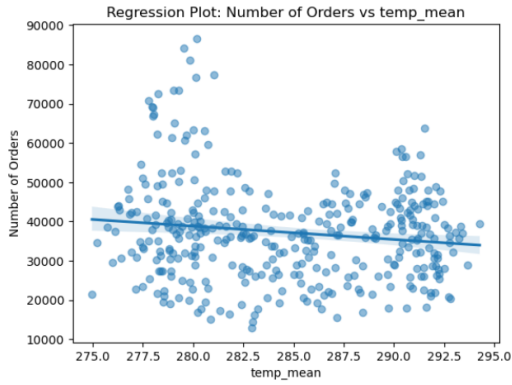


Figure 11: Sales vs Temperature

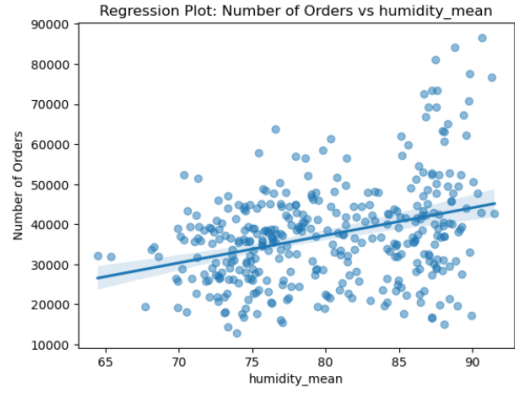


Figure 12: Sales vs Humidity

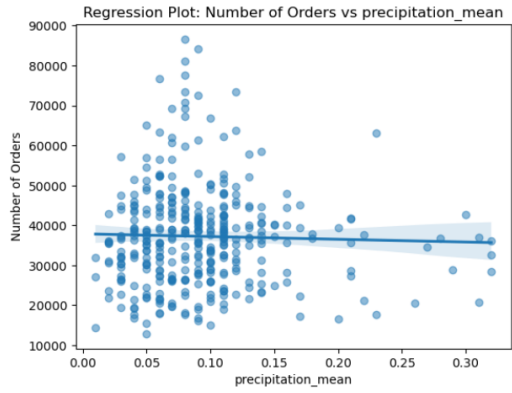


Figure 13: Sales vs Precipitation

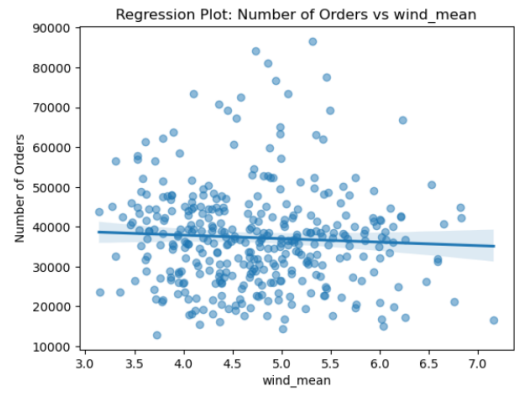


Figure 14: Sales vs Wind

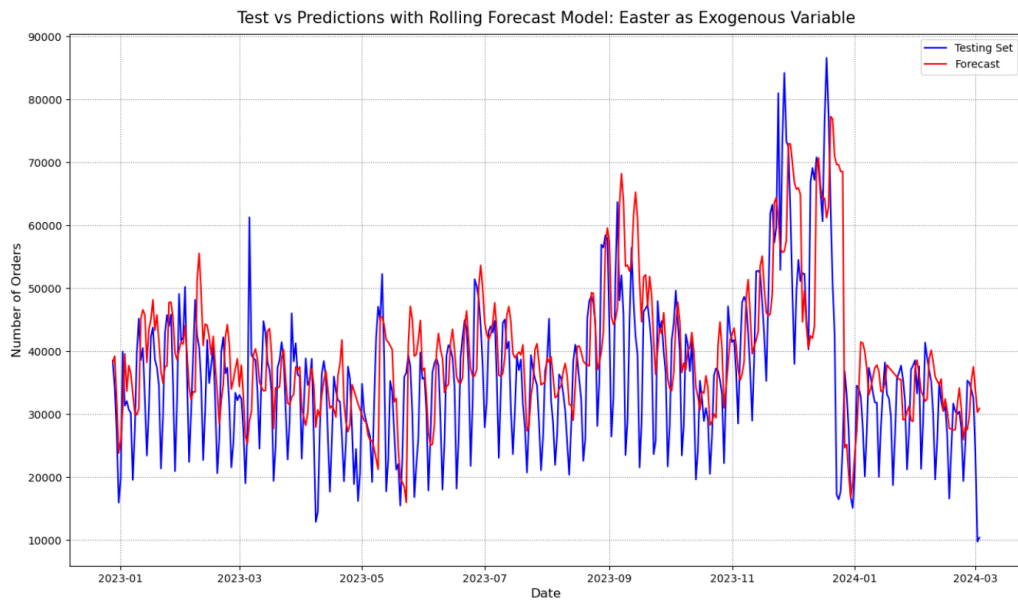


Figure 15: SARIMAX Model: Easter

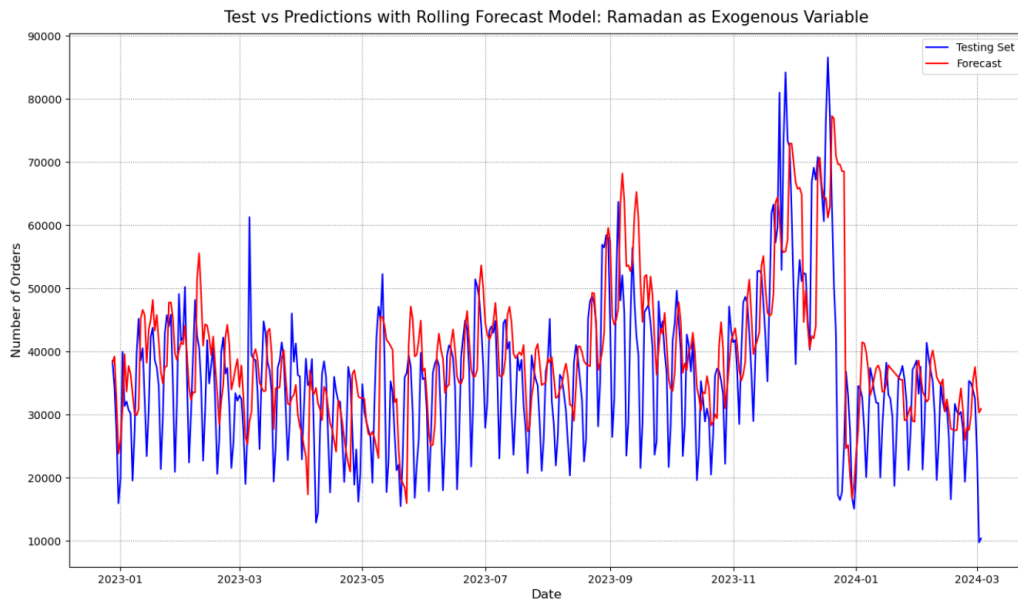


Figure 16: SARIMAX Model: Ramadan

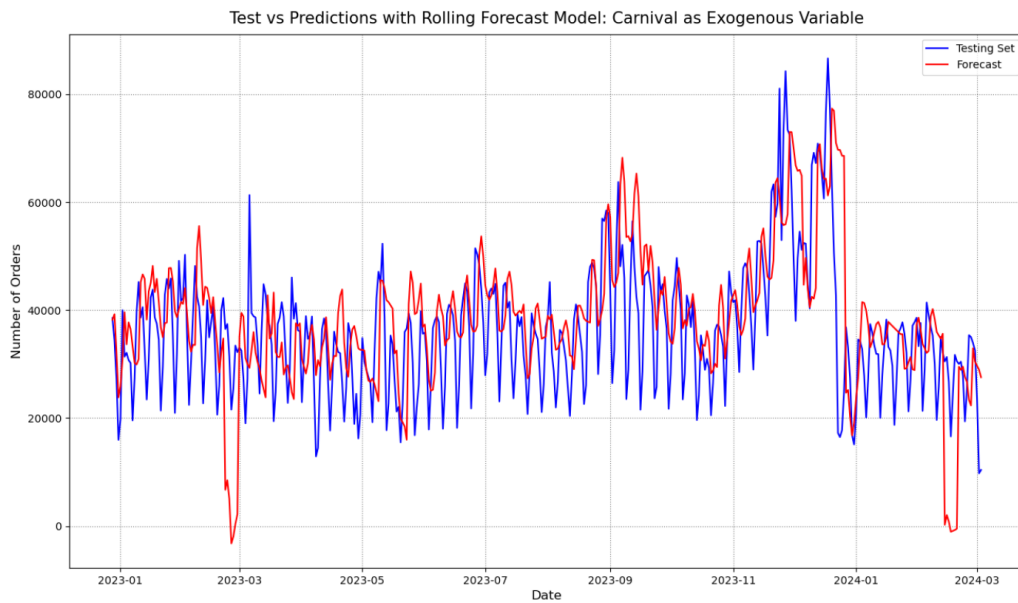


Figure 17: SARIMAX Model: Carnival

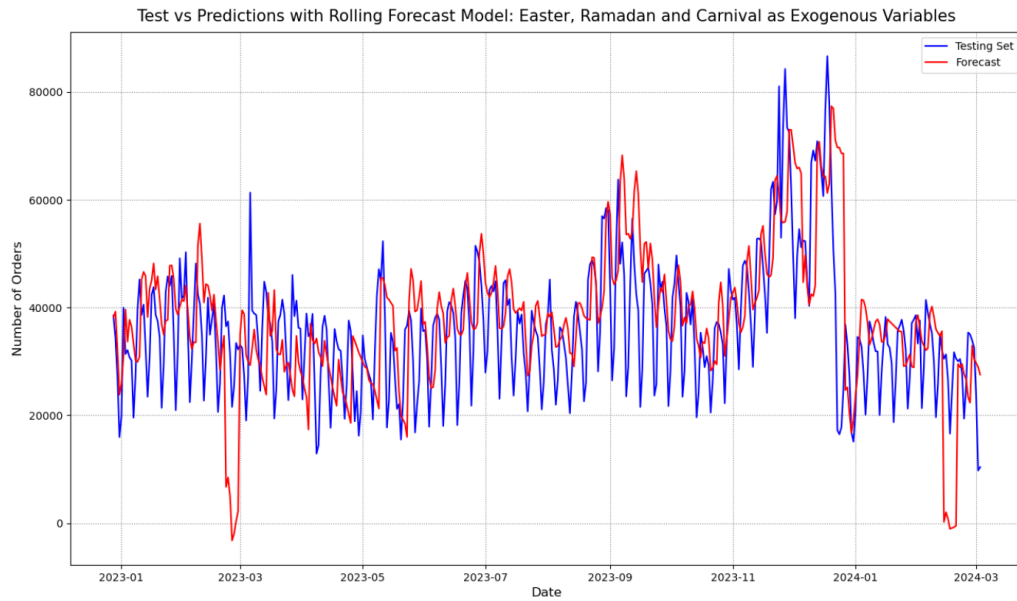


Figure 18: SARIMAX Model: Easter, Ramadan, and Carnival

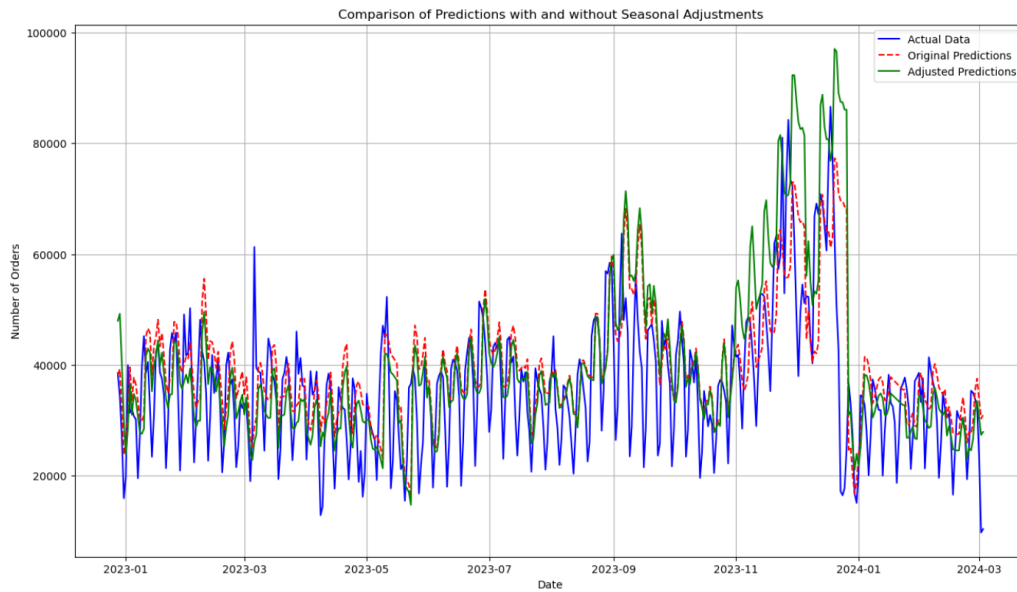


Figure 19: Combined model: Optimized SARIMA and Seasonality

Tables

Model	MAE	MSE	RMSE	MAPE	ERROR WITHIN 5%
SARIMA Base model	12949	314164184	17724	47.52%	15.05%
SARIMA Optimized model	8258	123410261	11109	27.99%	17.59%
SARIMAX Easter	8247	123076859	11094	27.92%	18.10%
SARIMAX Ramadan	8340	125063152	11183	28.20%	17.12%
SARIMAX Carnival	9134	151682548	12315	30.80%	16.67%
SARIMAX 3 volatile events	9237	153680491	12397	30.98%	16.20%
Combined prediction	8710	165313862	12857	28.47%	16.67%

Table 1: Results Table