UNIVERSITY OF TWENTE

BACHELOR'S THESIS

OPTIMIZING INVENTORY MANAGEMENT INCORPORATING DELIVERY DEMAND UNCERTAINTY AT TURFF

Author:
ERIKA MÄKELÄ

Supervisors:
DR. AMIN ASADI
DR. MARCOS MACHADO
BART ASBERG
SARAH BEN ABID

A thesis submitted for completion of the bachelor's program of Industrial Engineering and Management.

7 July 2024

Optimizing inventory management incorporating delivery demand uncertainty at Turff

DEPARTMENT
Behavioural, Management and Social Sciences

DATE
7 July 2024

VERSION
Version 1

AUTHOR
Erika Mäkelä
S2850990

EMAIL
erika.makela01@gmail.com

POSTAL ADDRESS
P.O. Box 217
7500 AE Enschede

WEBSITE
www.utwente.nl

FILE NAME
BSc Industrial Engineering and Management Thesis

Preface

Dear reader,
You are beginning to read my thesis “Optimizing inventory management incorporating delivery demand uncertainty at Turff”. It has been developed for the completion of the bachelor’s program Industrial Engineering and Management at The University of Twente, conducting research for a Dutch start-up Turff. The research addresses the improvement of demand forecasting and inventory decision making. During the research, I gained valuable experience and insights on applying the knowledge learned throughout the study in practice. Before going into the research, I would like to thank the people who have contributed to the completion of the graduation assignment.

I am grateful for the guidance I have received during the thesis from several people. First, I want to thank my first supervisor Amin Asadi for the guidance, ideas, and helpful feedback you have provided me throughout the work. I greatly appreciate the time you took every week to answer my questions and discuss my progress. I want to thank my second supervisor Marcos Machado for valuable feedback and perspectives. A special thank you also to my supervisors at Turff, Bart Asberg and Sarah Ben Abid, for your professional guidance, support, and insightful advice throughout the process. Your help has been extremely valuable for the completion of the assignment. Additionally, all the other employees at Turff who warmly welcomed me since the beginning and made my time at the office enjoyable each week. Last, I would like to thank my friends for motivation and good company throughout the process, and my family for support.

I hope you enjoy!

Erika Mäkelä
Enschede, July 2024

Management Summary

Context

Turff is a start-up providing student houses with tablets for easily tracking finances and consumption of shared foods and drinks, as well as making fast-scheduled drink deliveries. The company wishes to optimize and increase automation of their inventory processes, as demand of the delivery products is forecasted and replenishment decisions made with experience based calculations. The current approach often leads to non-optimal inventory levels, time wasted on manual decision making, and dependency on one employee. The company believes it can benefit from a more accurate demand forecasting method, which can be utilized again in the future. Furthermore, Turff wants to obtain optimal order quantities to assist with decision making for replenishment orders.

The focus of the research is solely on Heineken, which accounts for 75-80% of all delivery sales highlighting the importance of an individual inventory plan. The replenishment orders are also made more frequently (weekly) compared to other delivery products (monthly). The demand forecast is thus made weekly and aggregated through the delivery cities, as one central warehouse is used. The main aim is to improve forecasting accuracy by applying a scientific method to allow for a more automated process and more optimal inventory levels. Following the steps of the CRISP-DM methodology, we address the following research question: *How can inventory management be improved at Turff for their best selling delivery product, Heineken, in order to increase forecasting accuracy from about 75-78% to 85-90% and optimize inventory levels?*

Methods

We conduct literature review to identify methods that suit the event-driven seasonality of the data, and choose to focus on time series and causal methods. From time series methods we can expect accurate results from seasonal naive, Holt-Winters, and ARIMA with its extensions of seasonality and exogenous variables. From machine learning techniques we propose multiple regression and random forest. The main limitation with the use of the methods is the availability of data. We exclude data before August 2022 as it is not representative of the current situation. However, having only one historical observation for some weeks can pose challenges for data with yearly seasonality.

For an inventory policy we have two main choices: continuous or periodic review, and cost minimization or service level approach. For continuous review the replenishment orders are placed every time the inventory drops below a predetermined level, while for periodic, the inventory is checked and new orders placed at equal intervals. A benefit of a continuous policy is the need for a lower safety stock, however, a periodic policy is easier to manage. With cost minimization, different costs are weighed for optimizing inventory levels, while in a service level approach, the focus is on ensuring that all demand can be met in a certain percentage of periods. We choose to use a periodic review (R, S) inventory policy with a service level approach as it better matches the company's current situation and objectives.

Results

Results of the forecasts are evaluated with the KPIs MAPE, RMSE, bias, and R^2, giving us a comprehensive view of accuracy. Table 1 summarizes the results of the best performing models. We find that SARIMAX and semi-supervised random forest outperform the other tested models. Random forest scores slightly better on MAPE, indicating that the overall percentage accuracy is higher. SARIMAX, on the other hand, is superior with regards to RMSE. RMSE is the difference between the forecasted and actual values in the same unit as the demand. It can be determined as a more important metric for Turff because it penalizes large errors more significantly, as the cost of a stockout is high and thus, large errors should be aimed to eliminate. The low R^2 value of the semi-supervised random forest also tells us that the model does not capture the patterns of the data well. Additional advantages of SARIMAX are that it is less complex and better accounts for seasonal patterns of time series data.

KPIs	Seasonal Naive	Holt-Winters (no trend)	ARIMA	SARIMA	SARIMAX	Random Forest, Semi-supervised
MAPE	14.0%	14.5%	14.1%	14.5%	12.1%	10.3%
RMSE	198.3	200.5	195.1	194.8	165.6	174.2
bias/R^2	93.8/-	100.6/-	61.4/-	101.9/-	35.3/-	-/0.16

Table 1 Summary of performance results of the most accurate forecasting models tested.

Based on the SARIMAX forecast, we aim to optimize inventory levels by making an ordering policy. This is done by combining the demand forecast at each time i with added dynamic safety stocks by utilizing the (R, S) inventory policy, where every period R we place an order to bring the on-order inventory up to S_i. Developing the ordering policy helps to reduce time spent on manual decision making, optimize inventory levels to reduce costs, and reduce dependency on experience based calculations.

Conclusions and Recommendations

Through answering each of the sub-research questions, we can conclude that we are able to increase the forecasting accuracy from about 75-78% to 87.9% with the chosen SARIMAX model and to 89.7% with the tested random forest with semi-supervised learning, excluded due to other characteristics. Currently, the forecast closely reflects the observations of 2022-23 and we can expect the accuracy to improve when more representative data is available. As the forecast does not by itself determine replenishment order quantities, we have developed an (R, S) ordering policy for more direct implementation. Thus, we have met our goal of developing a more scientific, more automated, and more optimized method for managing the stock of Heineken.

For further improving the forecast accuracy, we recommend to gather more information on:

- The times a stockout occurs
- Price changes of Turff as well as competitors for Heineken and products of the same category
- Promotions of Heineken and products of the same category
- The relationship between demand and each event and exam period

We used interpolation to account for sales figures that do not represent demand due to stockouts, while more details can help improve the validity of the estimations. Price changes and promotions can be implemented as additional independent variables for the model with the aim of increasing accuracy, when there is more data and knowledge on their effect on the demand. This can help the model understand the demand fluctuation and allow for a better ability to predict future patterns. Analyzing the relationship between each event and exam period can help to adjust the scale and scores to be more optimal. We advise the company to regenerate the forecast more frequently than the eight-month horizon used in this research, as short term forecasts tend to be more accurate.

For longer term recommendations, we mention that the forecast and inventory policy could be improved by automating inventory management to more extent, i.e. with an ERP system. This can help to utilize all information live to make more accurate inventory decisions, optimize shift scheduling based on the forecast, and allow for a possibility of a continuous review ordering policy for decreasing costs.

Table of Contents

Reader's Guide

The reader's guide gives an overview of what is discussed in each chapter of the thesis to help navigate the material.

Chapter 1 - Research Design
The first chapter explains the methodological approach to the thesis. It gives context to the company where we complete the graduation assignment. We identify the problems the company is facing and choose the problem to focus on based on a brief cost-benefit analysis. We use the CRISP-DM methodology to set out the steps we will take to approach the thesis in a structured way and define the sub-research questions based on the methodology. Last, we choose KPIs to assess the models and touch upon the validity of the used methods.

Chapter 2 - Current Methods and Data Analysis
In the second chapter we explore the current methods of forecasting and inventory decision making of the company. We explain the availability of data and what limitations come along with it, as well as its reliability. We divide the data into a training and testing set to allow us to analyze performance of the models and explore the characteristics of the data.

Chapter 3 - Theoretical Framework
The third chapter includes the theoretical framework of the thesis, which we obtain by conducting a literature review. We explore the importance of accurate demand forecasting and its typical characteristics. Based on criteria we define for suitable forecasting methods, we decide to aim our focus mainly to time series analysis and causal methods, and further investigate different methods that fall into these categories. Last, we explore suitable order quantity policies.

Chapter 4 - Forecasting and Performance Evaluation
In Chapter 4 we test several different models and analyze their performance on the historical data divided into a training and testing set. We explain the process of parameter optimization and selection for each model. By comparing the results assessed by KPIs, we choose the best performing one and develop the ex-ante forecast.

Chapter 5 - Ordering Policy
In the fifth chapter we utilize the developed forecast to make a dynamic ordering policy. We choose between a cost minimization and a service level approach based on suitability and objectives of the company, and use calculations of safety stock for attempting to optimize order quantities.

Chapter 6 - Conclusion and Outlook
In Chapter 6 we discuss the results and the assumptions that were made for the models. Furthermore, we make recommendations on how the forecast's accuracy can be increased in the future, leading to a more optimal inventory policy as well.

Appendices
The report includes nine appendices for further information. The appendix includes details on the research design, specifics on the used KPIs, comparison of characteristics of forecasting methods, addressing the assumption of data stationarity for ARIMA, results of the ex-post and ex-ante forecasts and their parameters, the resulting ordering policy, an explanation of the python libraries utilized for data analysis and forecasting, and an explanation of the use of AI during the research.

Chapter 1 - Research Design

In the first chapter we introduce the company and the problems they are facing. We identify the core problem we want to focus on based on a developed problem cluster and a cost-benefit assessment. We define the methodological approach to solving the problem based on the CRISP-DM methodology. Following the steps of CRISP-DM we define the sub-research questions, which help us structurally approach the research.

1.1 Context

Turff is a student-led start-up that originated from The Delft University of Technology. The company provides student houses with tablets that allow the members to track for each person in real time the consumption of shared foods and drinks, money balances regarding the shared goods, and the people who sign up to do groceries, cook, or eat per day. The tablet can be obtained with a one time payment of €99.99. In addition, a free mobile app is available. When the tablets are not being used, advertisements are shown as screensaver. The advertisements currently bring the company their main source of revenue, with over 2500 student houses around The Netherlands using the tablets.

In addition to the tablets, Turff makes fast-scheduled drink deliveries to student houses, which is the focus of this research. Turff allows students to order beer, hard seltzers and sodas, and get them delivered to their front doors even as fast as in a few hours for a delivery fee of €3.90. The delivery fee is not dependent on the size of the order or the customer's location. The tablets, the app, and the webshop can be used to place orders. The orders made via the tablet and the app give the customer a €1 discount per each product to encourage more users on these platforms. There is no minimum or maximum limit to the size of the orders. However, the delivery vans can take up to 80-90 crates at once until the weight limit is exceeded.

Turff started out with delivering to multiple cities around The Netherlands with several warehouses for storing the drinks. In the last about one and a half years the number of warehouses has decreased from three (Delft, Utrecht, Enschede) to one (Delft). In the early stages of the company, even more warehouses were used for delivery. The reason why Turff gave up on the Enschede warehouse was because there was not enough demand in the city for it to be profitable. With the shutting down of the warehouse, deliveries to Enschede were also stopped. Utrecht, on the other hand, had and continues to have enough demand, but the problem was with the warehouse and not enough staff working in the city. The more optimal solution was to make the deliveries to Utrecht from the Delft warehouse. Currently, the central warehouse in Delft is enough to meet the orders of Delft, Leiden, Rotterdam, and Utrecht. The location of Delft is ideal for the warehouse, as it is the city where Turff originated from and the city with the most amount of orders.

The company currently utilizes up to three vans each day to make the deliveries. Via a program the company has developed, the drivers can easily generate the optimal route to take based on the location of the customers, starting and ending in the warehouse in Delft. Each of the four delivery cities have some differences in the delivery possibilities regarding time slots, as explained below. Since the warehouse is in Delft, the city has the most flexibility. On the delivery day, the customers receive a message of a more specific one hour time slot to not have to wait at home several hours for the delivery to arrive.

Delft

- Weekdays: 2 time slots of 3 hours. Weekends: 1 timeslot.
    - Orders can be placed until 1 hour before the time slot closes throughout the week.

Leiden

- Weekdays: 2 time slots of 2-3 hours. Weekends: 1 timeslot.
    - Orders can be placed until 1 hour before the time slot begins throughout the week.

Rotterdam

- Weekdays: 1 time slot of 3 hours. Weekends: 1 timeslot.
    - Orders can be placed until 1 hour before the time slot begins throughout the week.

Utrecht

- Weekdays: Mondays, Wednesdays, and Fridays 1 time slot of 3 hours.
    - Orders can be placed until 2 hours before the time slot begins.
- Tuesdays, Thursdays, weekends: No deliveries.

Turff wants to improve the order forecasting, as currently it is manual with no optimized method used. The supply chain manager of the company makes weekly decisions for the size of the replenishment order with experience based calculations. In Chapter 2.1 we give a practical example of the current method.

A more accurate and longer-term forecast would help save time and money for the company. As calculations for demand estimations would not have to be made weekly, time would be saved. Furthermore, the manual calculations currently give a base for how many delivery drivers are scheduled to work per day. Improving the forecasting accuracy can help with having a more accurate number of drivers each day, and thus, reduce the excess number of shifts that the company needs to pay for. Currently, stockouts of Heineken, the focus of the research, do not happen often as the replenishment orders are made large enough to assure the sufficiency to meet demand. However, Turff does not want to order significantly higher amounts of the product per week than needed as it ties up capital. The opportunity cost of capital is a significant factor for the company, which is one of the reasons why they want to have a more accurate forecast. The idea is to try different forecasting methods to assess which one returns the most accurate results. In addition, it allows the company to use the same method in the future more easily, giving them a longer term solution as the model will be clearly presented.

Having an accurate demand forecast is important for making inventory decisions. The demand forecast can be used to develop an inventory policy for Turff, giving us valuable insights on, for example, the size of a safety buffer that should be used. A safety buffer reduces risks of shortages (Kim & Jeong, 2018), making it important for Turff that loses sales in the case of a stockout instead of having backorders. When a cost minimization approach is taken for developing an ordering policy, it can help to reduce inventory costs. When a service level approach is adopted, it helps the company better ensure a desired percentage of all demand can be met, minimizing lost sales.

1.2 Problem Identification

In this section we will introduce the different problems Turff is facing regarding warehouse management and develop a problem cluster to visualize the cause-effect relationships of the issues. Based on the problem cluster we will identify the candidate core problems, the problems we could tackle, and analyze the costs and benefits of them in order to select one to solve that brings the most benefit to the company.

1.2.1 Problem Cluster

From first conversations with the company representatives, it became clear that there are several points of improvement regarding warehouse management at Turff. We began by developing a list of various issues caused partly by the current working methods, as seen in the table in Appendix A1. The problem cluster in Figure 1.1 is made based on the cause-effect relationships the problems have to see how they link together (Heerkens & van Winden, 2017). In addition, it shows that problems are not isolated, but usually affect one another. Problems that we cannot influence are left out of the cluster, for example, the issue of broken beer crates during delivery drives.

Developing the problem cluster is essential for understanding what issue to focus on. The candidate core problems, the root causes of the problems, are marked with orange colors on the cluster in Figure 1.1.

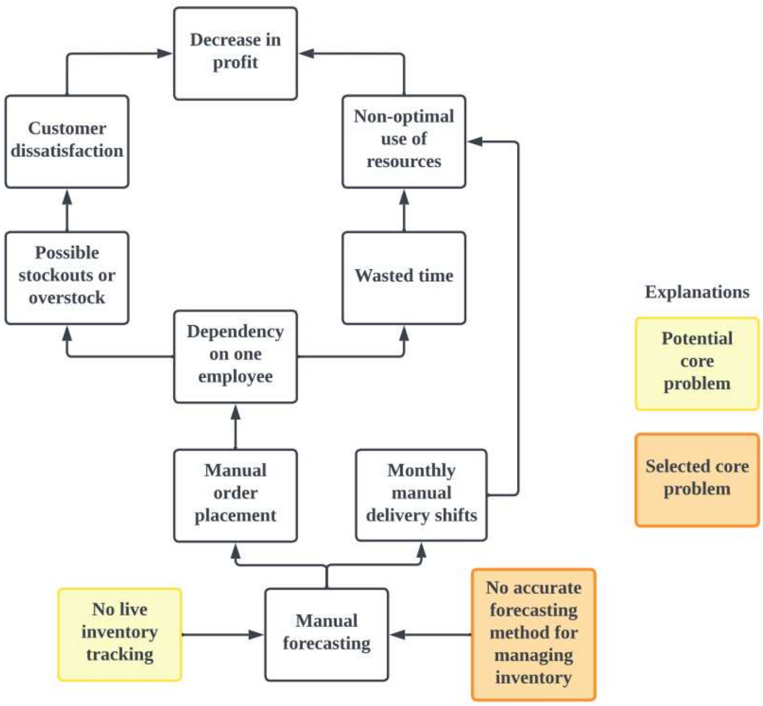


Figure 1.1 Problem cluster.

As can be seen from the cluster in Figure 1.1, many aspects of inventory management are done manually at Turff. The delivery drivers calculate the inventory, the supply chain manager calculates the forecasted demand in order to make replenishment decisions as well as determine how many delivery drivers are required per day. Due to this, the company is quite dependent on the employee for these decisions. This is also not an efficient use of the company's resources, as the tasks could be more automated and that way save time for other work. Even with experience in making the decisions the lack of structured approach reduces the accuracy, which is why they often have some overstock that ties up capital of the company.

Overall, we can see that there are several possibilities for improvement. Choosing the problem with the most relevance and benefit for the company is vital for a good research design.

1.2.2 Core Problem

Looking at the problem cluster in Figure 1.1 we can identify the candidate core problems to be no live method of inventory tracking and no accurate forecasting method, from which we select one to solve. The selected core problem will be the one that by solving it brings the most benefit to the company from the cluster of problems. Both of the mentioned problems can be influenced and are causes, instead of consequences, of other problems, making them candidates for the core problem (Heerkens & van Winden, 2017). We analyze the costs and benefits of the two problems in order to choose the more relevant one.

By choosing to solve the problem of no live inventory tracking, time would possibly be saved by the delivery drivers who are currently counting the inventory manually. The inventory is calculated three times a week, taking around 15-20 minutes per time. However, an investment would need to be made to implement a new system, for example, a barcode beeper and an Enterprise Resource Planning (ERP) system in order to automate the system more. In addition, solving this problem in itself would have little effect on its consequences. As an example, having more accurate inventory count could improve the issue of stockout or overstock, but due to the already quite high accuracy of the current system the effect would be small. Therefore, if the issue of live inventory tracking would be tackled it would be beneficial to combine it with one of its consequences, broadening the scope of the assignment significantly.

There are no specific costs for developing the company's approach for demand forecasting. The benefits of a more accurate forecast are significant for its consequent issues. It will allow the company to not spend time on estimating weekly demand manually, it can help determine the number of delivery drivers needed per day, reduce dependency on the supply chain manager, reduce overstock and stockouts, and consequently increase customer satisfaction. Improving all of the aforementioned aspects can also lead to an increase in the revenue of the company.

By analyzing the problem of no accurate forecasting method, we can conclude that costs are lower and the benefits higher compared to the other candidate core problem. Due to these reasons, the chosen core problem is defined as: **The demand for Heineken is forecasted without a structured or optimized approach.** We take action on this problem in order to optimize the stock levels and improve the use of the company's resources. The forecast accuracy will be assessed through several Key Performance Indicators (KPIs) in order to measure whether we are able to improve on the problem. The KPIs are defined more specifically in Section 1.3.6. The current reality of the forecasting accuracy is about 75-78%, however, there is not extensive data to determine the exact percentage. The company wants to achieve an accuracy of about 85-90% (norm).

1.3 Research Approach

In this section we define the scope of the research, as well as what we hope to accomplish with it. We introduce the CRISP-DM methodology, which will be used throughout the research for a structured approach. We define the main research questions, and the sub-questions that need to be answered in order to complete the research. Last, we define the KPIs that help us measure the performance of the forecasts and the inventory policy and touch upon the validity of the methods.

1.3.1 Scope

The scope of the research is narrowed to one product, on a weekly basis, and analyzing the demand as a whole, not separately per each of the four delivery cities. There are two main reasons why the research is focused solely on Heineken. First, Heineken is the best selling delivery product of the company, making up on average about 75-80% of the sales of the delivery department, highlighting the importance of the product. Based on the ABC inventory classification system, Heineken belongs to class A. Winston (2004) explains that in most companies, 5%-20% of all sold products generate 55%-65% of the overall sales. Furthermore, an individual demand forecast should be made for type A items, as well as developing an effective inventory policy may produce substantial savings. Second, replenishing the stock of Heineken is usually done once a week by Turff, twice in exceptional cases. This is different to all other delivered drinks, which are reordered maximum once a month due to lower demand and minimum order quantity regulations placed by the suppliers.

Turff's supplier acts as a third party for the orders of Heineken and is currently allowing one delivery day due to the volume of their deliveries. In some cases, there is a possibility for Turff to get a second delivery date if the supplier is informed early. The extra day is sometimes used by Turff to separate the payment into two parts and keep the cash flow more constant.

The demand forecast will be made per week, analyzing the historical data aggregated per week as well. As mentioned, most weeks the replenishment orders are made once a week, which is why more frequent, such as daily, demand data is not necessary for Turff. In addition, because the company is using one central warehouse, developing a forecast per city (Delft, Leiden, Rotterdam, Utrecht) does not bring additional value for Turff.

The decision of demand forecasting methods that will be used is affected by the time frame we have for the completion of the research. Methods that take from many weeks to months to develop are out of the scope.

1.3.2 Research Goal

The research aim is to address the core problem of the overall issues Turff is facing: demand forecasting of deliveries of Heineken. As visualized by the problem cluster in Figure 1.1, by improving the forecast we can make significant improvements to the issues it is causing.

Overall, we aim to develop an accurate forecast and transparently present the method used in order for the company to be able to easily apply it again in the future. The forecasting technique should be such that applying it fits the time frame of 10 weeks of the thesis assignment. We want to test the accuracy of different methods testing them with several KPIs to find one that suits Turff the best. In addition, we aim to make recommendations, for example for data gathering, that can be later on implemented in the forecast for increased accuracy. While the forecast gives valuable information on the expected demand and its patterns, it does not directly indicate how much the size of replenishment orders should be. We want to ensure that the forecast can bring the most benefit to the company by including calculations of optimal safety buffers and order up to levels.

A good forecast is essential for business planning (Ivanov et al., 2019). If demand is forecasted higher than actual, the unnecessary products increase costs and take up storage space. However, too low estimates may result in stockouts and lost sales. The more accurate the forecast the less money has to be tied up in excess inventory.

The main deliverables will be a demand forecast for Heineken until the end of January 2025. In addition, deliverables include forecast quality assessment, an inventory policy for the same time period, and recommendations on what data to gather to improve the forecast in the future. An explanation of how the selected forecasting method works and was modeled will be given, as well as obtaining the optimal parameters. This will ensure that the company can use the forecasting method again once the made forecast expires, allowing to save time on figuring out how to use the methods themselves.

1.3.3 Research Question

Based on the defined context and research aims, we formulate the following research question to solve the core problem:

How can inventory management be improved at Turff for their best selling delivery product, Heineken, in order to increase forecasting accuracy from about 75-78% to 85-90% and optimize inventory levels?

1.3.4 Sub-research Questions

In this section, we define the sub-research questions and explain the method of approaching each of the questions. Further details on each sub-question can be found in the table of research design in Appendix

A3. The sub-research questions show the steps needed to take in order to complete the research as a whole, and structures the order of the tasks.

1. *What is the current forecasting method?*

As the aim of the research is to improve the method of forecasting, it is important to first be familiar with the current methods. We obtain the knowledge by conversations with the supply chain manager of the company in charge of the current decision making. This helps to understand what factors are already implemented in the forecast and what could possibly be added. Explanation with a practical example is given in Section 2.1.

2. *What historical data is available and what characteristics does it possess?*

We identify the available data that can be used for the forecast and analyze it to get a better understanding of its characteristics. We evaluate its reliability and the limitations that it introduces, and prepare it for forecasting. We especially address the trends and seasonality that the data contains, as well as the effect of external factors, such as exams and events. For more details, refer to Section 2.2.

3. *What are the most relevant forecasting methods available in literature?*

We conduct literature review with the aim of gaining insights on the forecasting methods that can be applied to the case. Different methods account better for different characteristics of data and thus, do not lead to the same results. We focus our search on quantitative methods, more specifically two common methods, time series and causal methods, and conclude which are the most relevant models. We evaluate this based on the characteristics of each method, the objectives of the company, and the available data. This allows us to tackle the main deliverable of developing the forecast for the upcoming months. The findings are presented in Chapter 3.

4. *How can the chosen forecasting methods be applied to the case of Turff?*

Based on the different forecasting methods identified through the literature search as well as the conducted data analysis, we develop the models to fit the data of Turff. This includes careful parameter selection for ensuring the model to capture the data patterns as precisely as possible. We apply several methods in order to find the one with the highest accuracy, and present the process and results in Chapter 4.

5. *How can we assess the accuracy of the chosen methods?*

The quality of the forecasts is evaluated with different KPIs defined in Section 1.3.6 to assess the accuracy and determine the most suitable method. This allows us to provide Turff with recommendations on what methods they should use now and in the future as well as validating the used techniques.

6. *How can we optimize inventory levels by developing an ordering policy for Heineken?*

To utilize the forecast in a more practical manner, we develop a policy to determine optimal safety stocks and order up to levels for Heineken. In Chapter 3, we introduce different approaches for the ordering policy, including options of a continuous or a periodic review, and a cost minimization or service level approach. In Chapter 5, we choose the most appropriate approach and develop the policy.

7. *How can the forecast be improved and applied in the future?*

By analyzing the results of the forecasts, we make conclusions on the findings and identify the ways the forecast and ordering policy can be improved. This includes additional data to gather as well as further automation and research opportunities. We answer the last sub-research question in Section 6.2.

1.3.5 Research Methodology

As the main aspect of the research is data, we choose to structure the research based on the Cross Industry Standard Process for Data Mining (CRISP-DM). To choose the method, we compared different methodologies for conducting research, mainly the Managerial Problem Solving Method (MPSM) and CRISP-DM. MPSM is more theory oriented and does not specify the use of data, whereas CRISP-DM is data-centric. CRISP-DM is fitting for data science processes (Hotz, 2023). This is why the steps of the model can well be used for designing the research on demand forecasting and inventory management.

The research design is based on the CRISP-DM methodology, with each of the sub-question reflecting a step of the process visualized in Figure 1.2. The first step is business understanding, which is why we start with getting familiar with the working methods of the company, as well as understanding how the demand forecasting is currently being done. We explore what data is available and how we can access it. We mainly analyze the data on the aspect of what factors affect it the most and how, as well as identifying trends and seasonality. As the focus is on Heineken, beer can be associated with different happenings such as parties, exams, and special days. The demand is highly affected by these factors, therefore, it is of high importance to analyze it. This entails the steps of data understanding and data preparation of the CRISP-DM methodology.

After analysis of the data, we move on to researching what methods are available in literature that can be applied with the given data. We pick the most suitable methods based on how well they match the data's characteristics and use them for developing the forecast. This concludes the modeling step of the methodology. The next part is evaluation, where we conduct forecast quality assessment by utilizing several KPIs such as the mean absolute percentage error (MAPE) and the root mean squared error (RMSE). The way of deployment is to be decided by the company, however, recommendations on applying the solutions in the most beneficial way are given. We make an ordering policy based on the forecast to more directly assist with optimal decision making for replenishment orders. We make recommendations on how the forecast can be improved and applied in the future by gathering additional data and automation.

The design of the research is visualized in Appendix A2, showing the steps needed to solve the problem based on the CRISP-DM methodology. The steps also correspond to the sub-research questions defined in Section 1.3.4.

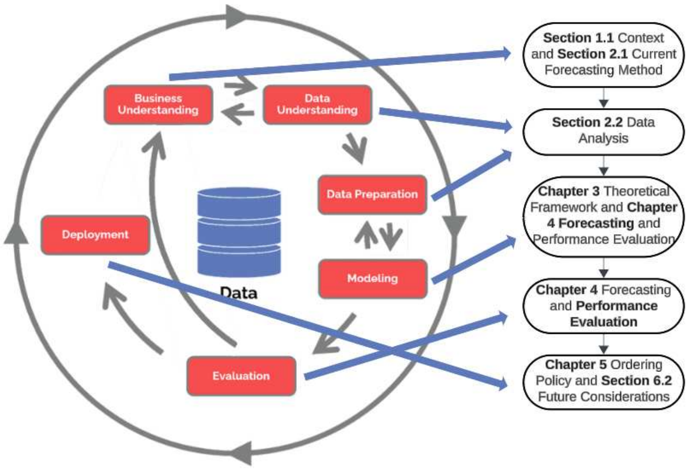


Figure 1.2 Steps of the CRISP-DM methodology corresponding to the research.

1.3.6 Selection of KPIs

We want to evaluate the forecast and inventory policy with several KPIs. Measuring the performance of the forecast is of high importance for determining to what extent the forecast can be relied on for decision making. Common performance indicators for accuracy assessment of forecasts include Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) (Da Veiga et al., 2014; Onder & Wei, 2022). Furthermore, the bias of the forecast, the overall simple difference of the forecast and the real observed sales, can give us insights on whether the forecast is on average over or underestimating the sales and by how much.

We choose to use MAPE, which shows the absolute value of the error instead of the direction of the error. We can use it for comparing the accuracy of the developed forecast and the current method. The smaller the MAPE the more accurate the forecast. A resulting MAPE that is lower than 10% may indicate a potentially very accurate forecast, below 20% potentially a good forecast, and higher than 30% potentially an unreliable forecast (Da Veiga et al., 2014).

MSE measures the quadratic deviation of the forecast and the actual data (Ivanov et al., 2019). Because the method squares all errors, large errors are penalized more considerably compared to smaller errors, making it a useful measure when the cost of large errors is substantial (Chopra & Meindl, 2016). RMSE is the square root of MSE. We choose to use RMSE, as it gives the result in the same unit as the forecast, making it more straightforward to interpret. The lower the obtained value is, the more reliable the forecast.

For evaluating the effectiveness of the ordering policy, we can calculate the resulting cycle service level (CSL) for the testing set, defined as the percentage of periods with no stockouts. Additionally, as the testing set is short, determining the percentage of products the company is able to deliver (fill rate) may give a more accurate performance estimate. Last, it is important to analyze optimality of the possible overstock. For instance, minimal overstock may be risky and cause stockouts if we observe unexpected demand fluctuations. Conversely, large amounts of excess stock ties up capital of the company. For specific formulas for calculation of the KPIs, refer to Appendix B.

1.3.7 Validity of the Methods

We take measures to address the validity of the research process in each of its steps. This includes the use of data, the use of literature, the selection of models and determination of parameters, and the assessment of the performance of each forecasting model and the inventory policy through KPIs and statistical tests.

The main measure we take to address validity of the forecasts and ordering policies is dividing the data into a training and testing set to be able to realistically assess the models' performances on unseen data. This gives us an accurate estimate of the performance of the model on predicting new values, which is the end goal after the models have been compared for selection.

For the forecasting models, we assess each of their accuracy on the testing set with KPIs. Validity of the chosen performance measures is of high importance for accurate reporting. One aspect is interpretability, since it is essential that the measures are simple to understand by the readers. We aim for interpretability by using KPIs that are common in literature. By this we can help to ensure that the metrics are widely studied to give valid indication of performance and be easier to verify. In addition, the KPIs should directly measure the topic of study: forecasting and ordering accuracy. Construct validity measures how well the KPIs actually measure the model performance (Cooper & Schindler, 2014). We use several different KPIs to gain a comprehensive understanding of accuracy, allowing us to better find the most suitable model.

For different forecasting models we take into account different measurements. For time series methods we mainly use the KPIs MAPE and RMSE, which give different perspectives on the accuracy, explained in Section 1.3.6. In addition, for machine learning models we take into account their R^2 values to estimate how much of the data's variability is explained by the used independent variables, which tells us about the explainability of the model.

We aim to mitigate possible biases that may occur in forecasting by taking several precautions. First, the nature of the data can cause bias on the forecasts. The data is analyzed to be consistently tracked promoting reliability, however, the lack of information on stockouts creates the need to adjust these values to make them better reflect the real demand. Data reliability is explored further in Section 2.2.2 and data preparation in Section 2.2.3. Regarding bias mitigation in model selection, we exclude models that do not have the ability to capture the seasonality of the demand, and carefully select parameters for the used models with various optimization methods. For example, for the ARIMA models we utilize the autocorrelation of lags as well as several information criteria, and for random forest we use a grid search for parameter selection that tests combinations of parameter values and uses cross-validation to evaluate them. These are elaborated on in Chapter 4. Last, we test the possibility of combining two forecasting models to aim to improve the overall accuracy and reduce possible biases.

To further compare the performance of some selected forecasting models with multiple parameter options, we utilize k-fold cross-validation. The technique is useful for obtaining a more comprehensive view of the models' performances, as it divides the data into k subsets and tests their performances iteratively on different sets. Finally, to help validate the performance of the final forecast we analyze its residuals to check that there is no crucial correlation.

To assess the validity of the ordering policy, we evaluate its performance by first using the training set to develop a policy for the period of the testing set. From this, we can calculate the resulting service level, the fill rate, and the values of over or understock per week. We conduct a simple sensitivity analysis by calculating the KPI values for a range of service levels to help choose the most appropriate percentage.

Overall, we want to make sure we validate the performance of each of the forecasting models in order to be able to choose the best one for the company, as well as the accuracy of the ordering policy to help with decision making. The overall design of the research is developed based on a careful literature search to define the theoretical framework. This ensures the methods are well thought out and validated in literature by numerous other studies.

Chapter 2 - Current Methods and Data Analysis

In this chapter, we explain the current method of forecasting with a practical example. We analyze the available data, its limitations and reliability, and inspect whether it contains trends and seasonality.

2.1 Current Forecasting Method

In this section, we want to answer the first sub-research question defined in Section 1.3.4: *What is the current forecasting method?*

Currently, estimation of demand for each week is done by the supply chain manager of Turff with experience based calculations. These calculations provide her with a rough idea of what the demand might be but is not based on a scientific method. This often leads to overstock, as that is less costly than lost sales due to a stockout. This takes time from the employee as well as ties up extra capital from the company. In addition, it makes the company more dependent on the employee.

As the employee is experienced in making these decisions, some self-estimation is done on the results of the calculation. Therefore, the quantity ordered does not always directly match the calculation's result. Next, we give a practical example of the current method of forecasting. The numbers used for demand are partly based on real data but multiplied by an anonymous factor for confidentiality reasons.

We are currently in week 12 of 2024 and want to determine the order quantity for week 13. If the order is made in week 12, it will arrive in the beginning of week 13. Table 1.1 shows the demand for the weeks that are used for the forecast:

Week #/Year	*Delft*	*Leiden*	*Rotterdam*	*Utrecht*	*Total*
11/2023	*856*	*320*	*235*	*222*	*1633*
12/2023	*824*	*228*	*484*	*319*	*1855*
13/2023	*673*	*434*	*431*	*390*	*1928*
11/2024	*601*	*396*	*477*	*211*	*1685*
12/2024 (exp.)	*812*	*308*	*509*	*263*	*1892*

Table 2.1 Demand of Heineken (partly based on real data but multiplied by an anonymous factor for confidentiality).

Based on the amounts shown in Table 2.1, it can be determined that the increase from week 11 to 12 in 2023 was about 14% (1855/1633=1.136) and similarly, the increase from week 12 to 13 was about 4% and from week 11 to 13 about 18%. From this it can be seen that the demand currently has a small upwards trend. The increase in week 11 between the two years was about 3%.

Looking at the data, a similar increase from week 11 to week 12 in 2024 can be seen as compared to the increase observed in 2023, approximately around 12% (exp). This suggests that for week 13 of 2024, the estimation can be calculated by multiplying the value of week 11 by 16% (a decrease of 2% compared to 2023 since the increase is 12% in 2024 and 14% in 2023), aiming for a result for week 13 in 2024 that closely matches the 3% growth in demand observed between week 11 in 2023 and the value for week 11 in 2024.

Overall, the deviation from the real demand is significant in some weeks, showing the relevance of aiming for a more accurate forecast.

2.2 Data Analysis

As the focus is on quantitative demand forecasting, understanding the data is an essential step of the process. This section is related to the second step of the CRISP-DM methodology called Data Understanding. We address the second sub-research question: *What historical data is available and what characteristics does it possess?*

2.2.1 Data Availability and Limitations

The first aspect is to understand what data is available and what limitations it comes with. Turff has collected data of the deliveries for several years. The data is collected per product, per city, and per day. However, as the forecast will be made for one product (Heineken) per week, and all cities aggregated, we transform the data by looking at the total weekly demand of Heineken per each city, and adding them up to receive the total amount. The demand is measured as 1 unit = 1 crate of Heineken.

Although data is available for 2021 and earlier, we only use it from August 2022 onwards. Turff is a four-year-old start-up and therefore many aspects have changed and developed in the past years. Thus, after discussing with the supply chain manager, we can conclude that data from 2021 and beginning of 2022 is not representative of the current situation. In this time period, Turff was still a very new company and the amount of customers (student houses) was much lower. In addition, Turff was delivering to other cities in addition to the four current cities, had several warehouses around The Netherlands, and different time slots for deliveries. Due to this, the earlier data does not reflect the current situation.

The availability of more data could increase the accuracy of the forecasts. The amount of historical data is limited for a start-up, bringing a challenge of not having observations from two yearly seasonality cycles. This restricts or complicates the use of some forecasting methods. However, as the aim is to test various models to determine the most accurate one, the model can be applied again in the future as more data is gathered and be beneficial in the long run.

For privacy reasons of the company, the actual numbers of the demand are not shown in this report. All data points are either multiplied by an anonymous number or the axes with demand or forecast values hidden on the graphs in order to be able to see the patterns of the data, while still ensuring confidentiality of the company.

2.2.2 Data Reliability

The main aspect of reliability is that the data is consistent, which helps to avoid interpretation errors and mistakes in converting the data into a consistent format. The data has been gathered by the company in a consistent manner throughout the years. Furthermore, the data is very simple as we are taking it from one source for one product, reducing the chance of interpretation errors.

We must also take into account other factors affecting the data that may slightly decrease the accuracy of the data set. The data of promotions and marketing campaigns is not explicitly gathered. However, these have mostly included coupons for free delivery, which according to the supply chain manager have little effect on the overall demand of Heineken. As the delivery fee, regardless of large order quantities, is €3.90, the error it might bring to the forecast is expected to be small. Furthermore, small price changes from the company's side or from other sellers can affect the demand, which is difficult to account for as detailed information is not gathered on the topic. In addition, until the end of November 2023, a second warehouse was used, located in Utrecht. Closing this caused the ordering time slots for Utrecht to be one hour shorter, which may have had a small effect on the demand for the city. Since these mentioned factors overlap, it is not clear what effect they had on the demand.

For the data to be reliable, it should also be complete. The data is complete in a sense that we have data from each city for each week of the full time frame of the used data. This means there are no gaps in tracking demand. However, the data is not complete in the sense of having information of enough seasonal cycles to be able to accurately estimate the patterns, as we use less than two years of observations. Due to having only 91 weeks of data available that represents the current situation, we are not able to assess whether we can fully rely on the past demand patterns to reflect on to the future, decreasing the reliability of the data set. In addition, in the case of a stockout, the data represents the number of sales, which we must distinguish to be different from demand. In these cases we convert the observation into a more reliable figure, avoiding decreasing the accuracy of the forecast due to these outliers. This is addressed in Section 2.2.3.

2.2.3 Division and Analysis of Data

First, we want to address the issue of outlier values in the data set caused by stockouts. There are a few clear occurrences of Heineken stockouts in the data set, some only in Utrecht in August 2022 and in May 2023 and one stockout in all cities in February 2024. It is a mistake to assume historical sales always represent demand, and thus, we must adjust the sales affected by stockouts to make the values better represent reality (Chopra & Meindl, 2016). There is no data on the real demand as the product is taken out of the online shop when the stockout is realized. Estimating these values will slightly decrease the validity of the data set, however, the effect is not significant as the occurrences are rare.

For the stockouts in Utrecht, we do not have representative data from the previous years to use. Thus, we use interpolation to make a rough estimation. For the stockout in February 2024, we do the same but also compare the percentage change between the same and previous weeks in 2023 in order to estimate the value. This simple method can give us a relatively safe estimation for these time periods, and help remove the outliers that would have a more significant decrease in the forecast accuracy.

As we do not have time to gather new data of demand for testing the models, we must divide the available data into a training and testing set. We do this to test the performance of the forecasts on unseen data, giving us more reliable information on their accuracies. Common divisions include using 2⁄3 or ⅘ of the data for training depending on the availability of data and forecast horizon. Because we have less than two years of data and want to make a forecast for about 8 months, we must deviate from a rule of thumb that the testing set should be at least as large as the forecast horizon. We leave about 20% for testing and realize that this compromise may not give a fully comprehensive understanding of the models' performance. To conclude, data from August 2022 until the end of 2023 will be used for training the models.

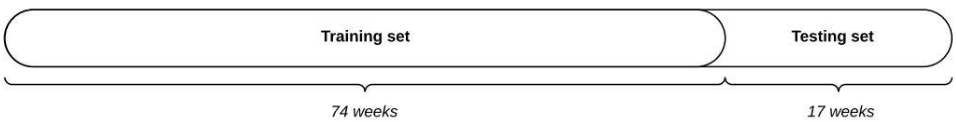


Figure 2.1 Division of available data into a training and testing set.

Next, we analyze the data to check whether it contains trend and/or seasonality. A trend is defined as a long-lasting pattern with a certain direction that can be increasing, decreasing, or constant (Singh et al., 2024). Seasonality is a cycle occurring on a seasonal basis, for example, annually or quarterly.

We start with visual inspection of Figure 2.2. There is no clear overall upwards or downwards pattern, but a possibility for a subtle increasing trend is present. We also observe this when fitting a trend line to the data. Seasonality is also not straightforward but we can identify peaks and declines in the same weeks of different years. We cannot identify seasonal patterns in shorter time intervals. We know that the demand of Heineken for Turff is very dependent especially on events and exam periods in the delivery cities. This is called event-driven seasonality, where seasonality is not observed in regular intervals but dependent on another factor. We can see that, for example, strong peaks are present in August and in the beginning of September due to schools' kick-off periods. As the customer base consists of student houses, it is naturally dependent on the behavior of students.

Event-driven seasonality can make forecasting more difficult, decreasing the accuracy as peaks are not observed in constant time intervals. What also increases the difficulty of assessing event-driven seasonality is that from weeks 18-30 we only have one year's observations of representative data. In Appendix D we go further into detail on trends and seasonality of the data by addressing stationarity.

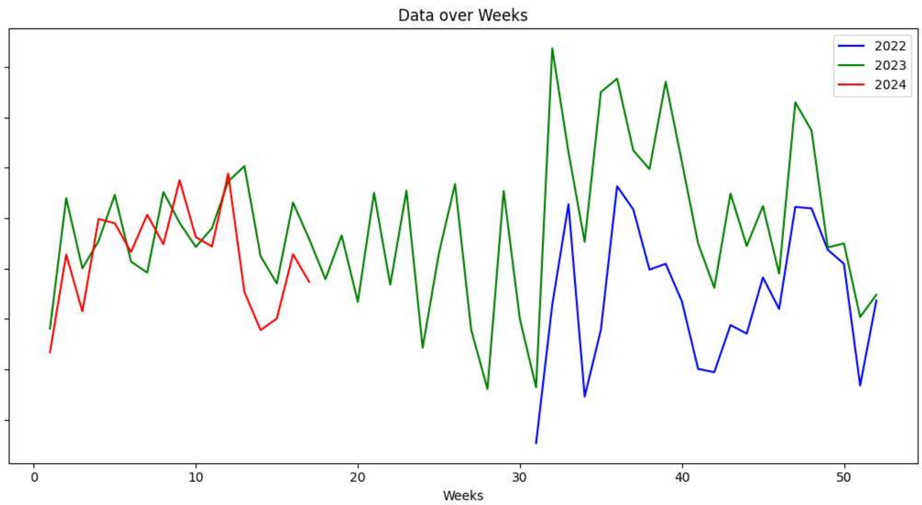


Figure 2.2 Weekly demand of Heineken per each year of the data set (values hidden for confidentiality).

In this chapter, we have introduced the current method of forecasting as well as explored the issues it may bring along. We have concluded that the main limitation for the research is the little amount of available data, as we also further have to divide it into a training and testing set. We have explored the data to be consistent with no missing values, however, the lack of data on promotions, price changes, and stockouts slightly decreases its reliability. Last, we have analyzed the data to contain some trends and seasonality, which affect the choice of forecasting methods.

Chapter 3 - Theoretical Framework

In this chapter, we conduct literature review to explore the importance of demand forecasting in supply chain management and the issues it contributes to solve. We investigate various different forecasting methods available to be able to choose the most suitable ones. We direct our focus to quantitative forecasting methods, more specifically, time series analysis and causal methods. Last, we introduce methods for using the forecast to obtain an optimal order quantity policy. In this chapter, we answer the sub-research question: *What are the most relevant forecasting methods available in literature?*

3.1 Integration of Theory

3.1.1 Impact and Characteristics of Demand Forecasting

Because the company orders inventory based on predicted demand, it can be seen as a push process (Chopra & Meindl, 2016). This highlights how essential demand forecasting is for Turff. It is one of the major challenges of supply chain management, due to its strong effect on optimizing stocks, reducing inventory related costs, and increasing sales and loyalty of customers (Kilimci et al., 2019). Stockouts as well as overstock are significant issues to a company. If a product is out of stock it can lead to lost sales and a decrease in customer satisfaction, in which case a customer may start ordering from another company (Kilimci et al., 2019). On the other hand, overstock can cause a loss of revenue as the capital of the company is tied up in excess inventory. Having a more accurate forecast will allow for lower safety stocks (Chambers et al., 1971). Simply, demand forecasting is essential to meet the needs of customers (Kim & Jeong, 2018). It helps with decision making, not only regarding replenishment orders, but also with, for example, staffing decisions, revenue management, and warehouse space and location needs.

Chambers et al. (1971), in line with other studies, divide forecasts into three categories based on the time period the forecast is made for. Short term forecasts are for a period of 0 to 3 months, medium term forecasts are for 3 months to 2 years, and long term forecasts are for 2 years and longer. The focus of the thesis is on a medium term forecast, due to the wishes of the company as well as allowing us to better see whether the yearly seasonality is captured by the model.

Chopra and Meindl (2016) introduce four important characteristics that supply chain managers of a company should be aware of with forecasting. First, forecasts are never fully accurate and thus, must also include a measure of the error of the forecast in addition to the forecast's expected value itself. Second, short-term forecasts are usually more accurate than long-term ones. More specifically, the error's standard deviation relative to the mean is often larger for long-term forecasts. Third, aggregate forecasts often have higher accuracy compared to disaggregated ones. If the demand of many products is forecasted at once, the forecast is often more accurate. This is also related to the time-period, as forecasting per week aggregates more demand points than forecasting per day, for example. Last, Chopra and Meindl (2016) mention that the farther away in the supply chain the company is from the customer, the more difficulties may arise in developing an accurate forecast. This can be due to information distortion which can cause, for example, the bullwhip effect.

In the case of Turff, some of the above mentioned characteristics apply. The first characteristic implies the importance of measuring forecast error. This will be taken into account as we conduct a quality assessment to be informed about the extent of trust that can be placed on the forecast. The second characteristic shows that the medium-term forecast that will be made for Turff may be less accurate compared to a short-term forecast. This must be considered in longer term decision making, as well as the performance continuously monitored. As we are only focused on one product of the company, Heineken, the forecast is very disaggregated in that sense. However, from the time perspective, the forecast will be made on a weekly basis, aggregating the demand from a longer period. The forecast is also aggregated for the four delivery cities. Reflecting on the fourth characteristic, Turff directly delivers to its customers, which is why information distortion is not a considerable factor.

The selection of a forecasting method is dependent on several aspects (Chambers et al., 1971). These aspects include the relevance and availability of past data, the desired accuracy, the time period of the forecast, and the time available for developing the forecast. Thus for the case of Turff, we define the following limitations for choosing appropriate models based on the scope of the research, the available data, and the company objectives:

- The forecasting method must be suitable for medium-term forecasting (about 8 months)
- The method must be able to deal with some seasonality and trends
- The method must be suitable to use with a limited amount of historical data of about two years
- The method must be simple enough to develop in a timespan of a few weeks
- The method must be easily understandable for the possibility for the company to apply it again in the future

3.1.2 Introduction to Common Forecasting Methods

We first make a distinction between qualitative and quantitative forecasting techniques. When there is very little or no data to use, (i.e. when a product is new) or when there are experts with market intelligence that may impact a product's demand forecast, qualitative methods are used (Chopra & Meindl, 2016; Ivanov et al., 2019). Furthermore, qualitative methods are often subjective and dependent on human knowledge and judgment. Quantitative forecasting methods use existing data to make predictions of the future. Because Turff has representative historical demand data to use for the forecast and we do not have specific expert knowledge on qualitative forecasting, we will rely on quantitative methods.

Winston (2004) makes a distinction between two common quantitative forecasting techniques: extrapolation methods and causal methods. Extrapolation methods use historical values to predict future observations, and are based on a time series. Thus, we will call these time series methods, as used in much of the literature. In time series methods, past trends and patterns are expected to continue into the future, but what caused the past fluctuations are not taken into consideration (Winston, 2004). In other words, historical demand is used to make forecasts of the future (Chopra & Meindl, 2016). Time series can be defined as a set of successive data points in a chronological sequence with equally timed intervals between each observation (Agung, 2011, Frechtling, 2012, Montgomery et al., 2015, as cited in Onder & Wei, 2022). Time series methods are appropriate when the demand pattern does not vary significantly between different years. Many time series methods well account for trends and seasonality, making them excellent candidates for demand forecasting of Heineken.

One of the simplest time series forecasting methods include the (seasonal) naive method. The naive method simply defines the forecast as the observed value of the demand of the previous period. Furthermore, the seasonal naive method can be used when historical data is highly seasonal, making the method more suitable for the case of Turff. The seasonal naive method suggests that the value of the forecast is equal to the last observed value of the corresponding season of the period before (Onder & Wei, 2022). As the demand of Turff is event-driven, it is good to test the seasonal naive method with a seasonal period of one year.

Causal methods, on the other hand, forecast future values of a dependent variable by one or multiple independent variables (Winston, 2004). Past data is used to estimate the relationships between the dependent and independent variables. In other words, these methods try to identify causal relationships between variables (F. et al., 2021). To use causal models, careful analysis must be made to explicitly understand the relationships between the dependent variable and the independent variables affecting it (Chambers et al., 1971). Furthermore, causal models are the most sophisticated kinds of forecasting tools. They are suitable for medium to long term forecasts, but are complex and can take long to develop.

From causal methods we introduce two machine learning (ML) techniques. ML is referred to as "using and developing computer systems that can learn and adapt by using algorithms and models to analyze and draw inferences from patterns in datasets", (Taparia et al., 2023, p. 4). Various ML techniques have been developed that can be utilized for demand forecasting, such as (linear) regression, support vector regression and random forest. ML models are more complex compared to many traditional forecasting methods but can generate more accurate results, especially when dealing with non-linear trends (Taparia et al., 2023). In this research, we focus on (linear) regression and random forest.

Time series further divides into different methods. The two common techniques are exponential smoothing and moving average methods. We will go into more detail first on exponential smoothing methods, then on moving average methods, and last, introduce regression models and random forest from ML techniques.

3.2 Exponential Smoothing Methods

We start by investigating the three different exponential smoothing techniques. Exponential smoothing techniques work in such a way that the weight of the data points are decreasing exponentially throughout the time series (Onder & Wei, 2022). They can be divided into the following categories: simple (or single) exponential smoothing (SES), double exponential smoothing (Holt's method) and triple exponential smoothing (Holt-Winters method) (Shiwakoti et al., 2023). Each category is an extension of the previous one, bringing another factor into the model.

3.2.1 Simple and Double Exponential Smoothing

SES and Holt's method are the most basic forms of exponential smoothing. SES does not account for seasonality or trends. It is used for univariate time series forecasting, meaning that the focus is only on a single variable (Profillidis & Botzoris, 2019). Holt's method includes trends as an additional component to SES. It is used when data shows no seasonality but some trend (Shiwakoti et al., 2023). Furthermore, it utilizes two separate equations to develop the forecast, one for the level component and one for the trend component of the time series. For specific formulas, refer to Shiwakoti et al. (2023, p. 2).

As these methods do not account for seasonality, they are likely not suitable for our data. We move forward with methods that better include trends and seasonality and thus, show more promise capturing the patterns of the data of historical demand.

3.2.2 Triple Exponential Smoothing (Holt-Winters Method)

Holt-Winters method includes trend and seasonality (Matsumoto & Komatsu, 2015; Shiwakoti et al., 2023; Tirkeş et al., 2017). The seasonality can be taken into account in two different ways, either additively or multiplicatively. The method with additive seasonality is used when seasonality is constant across all the levels of the time series. Multiplicative is suitable when the seasonal variations of the data change proportionally to the level of the time series. To put it in simple terms, additive seasonality should be used when the seasonal fluctuations stay approximately the same each time compared to the average observation. On the other hand, if the fluctuations get larger as the average observation gets larger, multiplicative seasonality is the better option. The following formulas can be used for the Holt-Winters model with additive seasonality (Da Veiga et al., 2014; Matsumoto & Komatsu, 2015; Shiwakoti et al., 2023):

Level: $\mathcal{L}_t = \boldsymbol{\alpha}(y_t - \delta_{t-s}) + (1 - \boldsymbol{\alpha})(\mathcal{L}_{t-1} + \mathcal{b}_{t-1})$ (1)

Trend: $\mathcal{b}_t = \beta(\mathcal{L}_t - \mathcal{L}_{t-1}) + (1 - \beta)\mathcal{b}_{t-1}$ (2)

Seasonality: $\delta_t = \gamma(y_t - \mathcal{L}_{t-1} - \mathcal{b}_{t-1}) + (1 - \gamma)\delta_{t-s}$ (3)

Forecast n periods in the future: $F_{t+n} = \mathcal{L}_t + n\mathcal{b}_t + \delta_{t-s+n}$, (4)

where $\boldsymbol{\alpha}$, β, and γ are the smoothing constants between 0 and 1 with the description of "If alpha is equal to 0 then the current observation is ignored entirely and if it is equal to 1 then the previous observations are ignored entirely" (Kalekar, 2004, as cited in Shiwakoti et al., 2023, p. 2).

The formulas with multiplicative seasonality are (Da Veiga et al., 2014; Shiwakoti et al., 2023):

Level: $\mathcal{L}_t = \boldsymbol{\alpha}\frac{y_t}{\delta_{t-s}} + (1 - \boldsymbol{\alpha})(\mathcal{L}_{t-1} + \mathcal{b}_{t-1})$ (5)

Trend: $\mathcal{b}_t = \beta(\mathcal{L}_t - \mathcal{L}_{t-1}) + (1 - \beta)\mathcal{b}_{t-1}$ (6)

Seasonality: $\delta_t = \frac{\gamma y_t}{\mathcal{L}_{t-1} + \mathcal{b}_{t-1}} + (1 - \gamma)\delta_{t-s}$ (7)

Forecast n periods in the future: $F_{t+n} = (\mathcal{L}_t + n\mathcal{b}_t)\delta_{t-s+n}$, (8)

where $\boldsymbol{\alpha}$, β, and γ are the smoothing constants between 0 and 1 (Shiwakoti et al., 2023).

3.3 Moving Average Methods

The second part of time series analysis are the moving average methods. There are several different extended methods for the moving average that we describe, each one a more complex extension of the previous method, taking into account more aspects of the historical data.

3.3.1 Simple and Weighted Moving Average

The most basic moving average technique indicates that the demand for the next period will be determined by the running means (the mean updated as new data becomes available) of the demand's historical values with equal weights (Profillidis & Botzoris, 2019). It does not take into account trends or seasonality. Furthermore, weights can be added to the moving average in order to place more importance on recent data. For more details and formulas refer to Profillidis and Botzoris (2019, p. 20-22).

3.3.2 ARMA

The Autoregressive Moving Average (ARMA) method combines characteristics of autoregressive as well as moving average methods (Kim & Jeong, 2018; Profillidis & Botzoris, 2019). The autoregressive (AR(p)) aspect suggests that the value of the forecast y_t depends linearly on two factors: it's previous p observation's weighted average (y_{t-1}, y_{t-2},..., y_{t-p}) and on a fully random factor called white noise (the error). Furthermore, the error term has zero mean and constant variance. The autoregressive aspect of the ARMA method requires the data to be stationary, indicating it has a constant mean and variance over time and thus, does not contain trends or seasonality.

The moving average (MA(q)) aspect of the ARMA method is different from the simple moving average method described in Section 3.3.1. It suggests that the value of the forecast y_t is linearly dependent on the current and the past q values of white noise, so a random process (Profillidis & Botzoris, 2019). The moving average process is always stationary.

The ARMA(p, q) model depends on its own p historical values and q historical values of white noise, with the general formula as follows (Profillidis & Botzoris, 2019):

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + ... + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - ... - \theta_q \varepsilon_{t-q} , \tag{9}$$

where ε_t = white noise (error)

c = intercept related to the mean μ of the time series and to the parameters $\varphi_1, ..., \varphi_p$ with the equation: $\mu = \frac{c}{1-\varphi_1-\varphi_2-...-\varphi_p}$ (10)

φ = parameter of the AR part

θ = parameter of the MA part

Because the autoregressive part of the method requires the data to be stationary, this should be inspected before application of the method (Profillidis & Botzoris, 2019).

3.3.3 ARIMA

The Auto Regressive Integrated Moving Average (ARIMA) is one of the most frequently used forecasting models of time series analysis (Matsumoto & Komatsu, 2015). It is an extension of the ARMA model, allowing the use of non-stationary data (Da Veiga et al., 2014). The extra step compared to the ARMA(p, q) model is that the non-stationary data must be differenced to make it stationary. The d in the ARIMA(p, d, q) model is defined as the number of times the series must be differenced in order to make it stationary (Da Veiga et al., 2014; Profillidis & Botzoris, 2019). The parameter d is usually 1 or 2, as most non-stationary time series data take one or two times of differencing to become stationary.

Differencing means taking the difference of each consecutive observation. This can help to get rid of seasonality or trends that the data contains. After differencing, in addition to visual inspection of the data, we can use statistical tests to determine whether the data is stationary. A commonly used test is the Augmented Dickey-Fuller (ADF). It checks whether the data is stationary based on the presence of a unit root in the data, with a null-hypothesis of non-stationarity (Onder & Wei, 2022). If the null-hypothesis cannot be rejected, the data must be differenced again and tested for stationarity. Presence of a unit root in data indicates some kind of systematic and unpredictable pattern, which is not suitable for forecasting techniques such as ARIMA.

3.3.4 SARIMA(X)

The ARIMA model can be expanded to the seasonal ARIMA(p, d, q)(P, D, Q, m) model, which adds a seasonal component to the method. The added P, D, and Q are the seasonal autoregressive, differencing, and moving average parameters, and m is the number of seasonal periods in a seasonal cycle (i.e. m=52 for weekly data) (Huang & Petukhina, 2022; Profillidis & Botzoris, 2019). If seasonality can be removed from the data by differencing, the basic ARIMA may be sufficient. The seasonal ARIMA explicitly considers seasonal aspects of the data, and is suitable when seasonality is clear and occurring at specific intervals.

The formula for SARIMA is the following (Matsumoto & Komatsu, 2015):

$$\Phi(B^m)\varphi(B)(1 - B^m)^D(1 - B)^d y_t = c + \Theta(B^m)\theta(B)\varepsilon_t \quad , \tag{11}$$

where ε_t = white noise (error)

B = backshift operator

c = intercept related to the mean μ of the time series and to the parameters $\varphi_1, ..., \varphi_p$ with equation (10)

φ, Φ = parameter of the non-seasonal and seasonal AR part, respectively

θ, Θ = parameter of the non-seasonal and seasonal MA part, respectively

SARIMAX (Seasonal Autoregressive Integrated Moving Average with EXogenous Variables) is a further extension of the model, incorporating exogenous variables by adding βX_t terms to the SARIMA formula based on the number of independent variables used. The exogenous variables are external regression components (Huang & Petukhina, 2022), allowing us to estimate the relationship between the demand and external influences.

3.4 Regression Models

Regression models are probably the most commonly used causal forecasting methods, modeling relationships between a dependent variable and one or multiple independent variables.

3.4.1 Simple Linear Regression

Simple linear regression allows us to model the linear relationship between a dependent variable and an independent variable (Winston, 2004). The regression model helps us understand how changes in the independent variable, also called predictor, affects the dependent variable. We can use simple linear regression to try to model the linear relationship between the events (independent variable) and the demand (dependent variable). For the specific formula, refer to Winston (2004, p. 1303). Since we have more than one independent variable that affects the demand of Heineken, we focus on multiple regression.

3.4.2 Multiple Regression

When more than a single independent factor is relevant for forecasting the dependent variable, we can use multiple regression (Winston, 2004). The principle is the same compared to simple linear regression, but with additional variables taken into account. The following formula can be used (Hyndman & Athanasopoulos, 2021; Winston, 2004):

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_k x_{kt} + \varepsilon_t \, , \tag{12}$$

where k = number of independent variables

$\beta_1, ..., \beta_k$ = measure of effect of independent variables, after accounting for effects of other predictors in the model.

The model is referred to as multiple linear regression, if the correlation of independent and dependent variables is considered to be linear (F. et al., 2021). A requirement of the method is that independent variables are not correlated to each other and that there is a linear relationship between the dependent and independent variables. In addition, the residuals should be random and normally distributed.

3.5 Random Forest

Random forest is a machine learning technique that combines multiple decision trees (Singh et al., 2024). A decision tree recursively splits the data into smaller subsets (Taparia et al., 2023). A tree contains leaf nodes with possible prediction outcomes, and root nodes that determine which leaf node the observation will end up in by determining its path. An observation from the testing set is guided through the tree's root nodes, at last arriving at a leaf node for the prediction (Taparia et al., 2023).

Decision trees can be used both for classification and regression tasks. In this research, however, we focus only on regression. To determine the prediction of a decision tree for regression tasks, the values of the trained data points in the specific leaf node are usually averaged. A visualization of a simple decision tree is depicted in Figure 3.1.

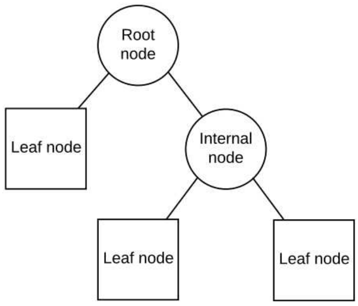


Figure 3.1 Visualization of a simple decision tree.

To develop each tree in a random forest, bootstrap aggregating (also called bagging) is used to randomly choose subsamples of the data for training (Singh et al., 2024). The final forecast of the random forest is obtained by averaging the results of the individual trees, leading to a more robust forecast with improved accuracy compared to utilizing just one decision tree.

3.6 Selection of Forecasting Methods

Based on the conducted literature review we have identified several suitable methods for demand forecasting that fit the seasonal delivery data of Turff. We assess how well the methods take into account the characteristics needed for forecasting such data and conclude what methods to test.

An overview of how well the introduced time series and causal forecasting methods deal with the suitability criteria defined in Section 3.1.1 is shown in Appendix C. It indicates that the Holt-Winters, SARIMAX, multiple regression, and random forest may be the most suitable methods for the criteria. We can see that the simple and double (Holt's) exponential smoothing are too simple for the data of Turff, as they do not capture seasonality and are more suitable for short-term forecasting. The ARIMA model well meets most of the criteria, but the Holt-Winters model scores better on the characteristics of understandability and better accounting for seasonality. Holt-Winters is simple and can generate outcomes that are comparable to methods with more complexity (Tirkeş et al., 2017). However, extending the ARIMA model incorporates seasonal factors and external components, adding detail and possibly more accuracy but complexity as well. Last, multiple regression and random forest are suitable for event-driven seasonality, as it allows to well capture dependencies of several variables. They do not specifically account for trends and seasonality but can be incorporated by adding the right variables to the model.

Based on analyzing the criteria, we choose to test Holt-Winters, ARIMA and its extensions, and multiple regression. In addition, we want to test the seasonal naive method despite its simplicity, as the seasonal pattern seems to well replicate itself over the years. We choose to exclude the simple and double exponential smoothing methods, as well as the simple (weighted) moving average methods, because they are too simple to capture the fluctuations of the data of Turff. Last, we want to test the introduced machine learning technique random forest. The model does not meet the criteria of simplicity but is able to effectively capture more complex patterns in the data and possible non-linear relationships.

Testing multiple methods allows us to gain a better understanding of the characteristics of the data, and find a method that provides the company with the most optimal results. This will help with improving the accuracy of inventory decisions. In addition, there is a possibility of combining different forecasting methods by taking, for instance, a simple or weighted average of them for a chance of obtaining higher accuracy (Chopra & Meidl, 2016; Onder & Wei, 2022).

3.7 Order Quantity Policies

Developing a demand forecast gives the company insights on the expected future behavior of the product, however, does not directly inform them on decision making based on the results. Thus, we also want to address another research question: *How can we optimize inventory levels by developing an ordering policy for Heineken?* In this section, we will investigate different policies for obtaining the optimal order quantities based on the forecast. This also includes insights on the safety stock the company should have. Determining the correct order quantities will help with effective inventory management and can lead to higher customer satisfaction, less costs, and a better flow of cash.

There are many variations of ordering policies, we distinguish between continuous and periodic policies. In continuous inventory models, replenishment orders can be made any time, while in periodic models replenishment orders are made after a fixed period of time. Due to the agreement Turff currently has with its supplier, they are using a periodic ordering policy, which is why these models should be explored. We also distinguish between deterministic and probabilistic models. In deterministic models demand is known and usually occurs at a constant rate, while in probabilistic models demand is uncertain. We start by explaining the most basic economic order quantity (EOQ) model that is continuous with deterministic demand. We extend to explore other models that are periodic with probabilistic demand.

3.7.1 Basic EOQ Model

First, some assumptions need to be clarified for the basic EOQ model to hold (Chopra, 2019; Winston, 2004):

- Demand is deterministic and constant.
- Independent of the replenishment order size, an ordering and setup cost K is incurred.
- The lead time is 0.
- Shortages are not allowed.
- The per-unit holding cost of inventory is h per period of time.

The decision of the EOQ is made based on minimizing the total costs (Chopra, 2019). Due to the assumptions required for the model that lead time is zero, we should in no case order when inventory $I > 0$ (Winston, 2004). This is because inventory incurs an unnecessary holding cost, leading to the observation that orders should be placed at $I = 0$.

We define:
D = Demand per period of time
Q = Order quantity
K = Ordering cost
p = Per unit cost

h = Holding cost per unit per period of time

The total costs we want to minimize consist of ordering costs, holding costs, and purchasing costs (Chopra, 2019; Winston, 2004). First, we define a formula for total costs. We take the derivative of it, set it to zero, and solve for Q in order to find the EOQ.

$$TC(Q) = ordering\ costs + purchasing\ costs + holding\ costs = \frac{KD}{Q} + pD + \frac{hQ}{2} \tag{13}$$

$$TC'(Q) = -\frac{KD}{Q^2} + \frac{h}{2} = 0 \tag{14}$$

$$Q^* = \sqrt{\frac{2KD}{h}} \tag{15}$$

Due to the required assumptions of the model, it is often not realistic. Demand rarely occurs at a constant and known rate, lead time is often not zero, and the possibility of shortages is often present. Thus, we investigate models where all of the defined assumptions do not need to hold.

3.7.2 (*R, S*) Model

Next, we discuss the (*R, S*) inventory policy with periodic review and probabilistic demand. The policy works in a way that we check the on-hand inventory every R unit of time (weeks for Turff) and make a replenishment order to raise the on-order inventory level up to *S* (Winston, 2004). On-order inventory is the sum of on-hand inventory and inventory that has been ordered but has not been delivered yet (Winston, 2004). The (*R, S*) model often results in higher holding costs compared to continuous models, but the (*R, S*) policy is easier to manage for the company, as we know when new replenishment orders are placed. As Turff has a set once a week agreement with its supplier for replenishment orders, a periodic review policy should be investigated. First, we introduce the cost minimization method and then a method with a service level approach.

We use the part of the (*R, S*) model that helps us determine the order up to level with weekly replenishment orders and a known lead time. In the case of a stockout, Turff's orders are not backlogged but the sales are lost as the product is taken out of the online shop. Thus, we introduce the case of lost demand. Assumptions of the calculation that we make in the research include that demand is a continuous random variable that follows a normal distribution, and the lead time is constant.

We use the model to calculate the on-order inventory level, which uses the costs of lost sales and holding costs. These are the most important costs for Turff, as there are no specific ordering costs from the supplier that can be affected. Holding costs can be seen as the opportunity costs of tying up capital, which are relevant for Turff as it is important for them to not make big investments at once to maintain the flow of cash.

We define:
L = Lead time for each order (constant)
D = Random variable of yearly demand, with mean $E(D)$ and standard deviation σ_D
c_{LS} = Cost incurred for lost profit
R = Time between reviews (in years)
S = Order up to level
h = Costs of holding an item in inventory for a year
D_{L+R} = Random variable of demand during time $L+R$, with mean $E(D_{L+R})$ and standard deviation $\sigma_{D_{L+R}}$

To determine the order up to level we use marginal analysis to calculate *S* such that it minimizes the sum of the expected holding and lost sales costs (Winston, 2004). That value of *S* should satisfy:

$$P(D_{L+R} \geq S) = \frac{Rh}{Rh+c_{LS}} \tag{16}$$

We can then use the *NORM.INV* function in Excel to compute the on-order inventory level. The function determines the inverse of the normal cumulative distribution for a determined probability. It requires inputs of the mean and standard deviation, as well as the mentioned probability, and returns the value at which the probability of the variable being less than or equal to this value equals the specified probability. From the received number, we must still subtract the amount in inventory to determine the order quantity.

Next, we introduce the cycle service level (CSL) approach for determining *S* and safety stock (ss). CSL is the percentage of replenishment cycles where all customer demand is satisfied, in other words, the probability of no stockout in the review period (Chopra & Meindl, 2016). As sales of Heineken account for 75-80% of all deliveries and costs of excess stock are cheaper than the costs of a stockout, ensuring that close to all demand is met is very important.

Since a stockout will occur if the demand during $L+R$ exceeds S, the following must be true (Chopra, 2019):

$$P(E(D_{L+R}) \leq S) = CSL \tag{17}$$

We define (Chopra, 2019):

$$S = E(D_{L+R})+ss \tag{18}$$

The relationship with safety stock is shown by (Chopra, 2019; Silver et al., 2017):

$$ss = F_s^{-1}(CSL) * \sigma_{D_{L+R}} = NORMS.INV(CSL) * \sigma_{D_{L+R}} \tag{19}$$

When we determine a desired CSL, we can use the *NORMS.INV* function in Excel to calculate the safety stock. The function simply requires a probability for an input and returns the inverse of the standard normal cumulative distribution. It is used for determining the z-score for the specified probability. In inventory management, the z-score is utilized as a “safety factor”, corresponding to a desired service level.

Again, we need to subtract the inventory level from the resulting *S* to determine the order quantity.

Chapter 4 - Forecasting and Performance Evaluation

In this chapter, we answer the fourth sub-research question: *How can the chosen forecasting methods be applied to the case of Turff?* and the fifth sub-research question: *How can we assess the accuracy of the chosen methods?*

Forecasting can be divided into ex-post and ex-ante forecasts (Huang & Petukhina, 2022). Ex-post forecasting refers to forecasting for the testing set consisting of already observed data. This is done in order to fit and evaluate the model. Ex-ante forecasting means using the historical data to forecast beyond the present. We start by testing seven different forecasting methods with ex-post forecasting. In Section 2.2.3, we divided the data into a training and testing set. In this chapter, we develop the forecasts, test their performance with KPIs and choose the most accurate one for a final ex-ante forecast.

An essential part of developing the forecast is "fitting" them. This means determining the different parameters for the forecasts that yield us with the most optimal results. This is done differently for each forecasting method, which is why details can be found in Sections 4.1-4.6.

4.1 Seasonal Naive

We start by testing one of the simplest time series forecasting techniques called the seasonal naive method. The seasonal naive method, suggests that the forecast for the current period is the observed demand one seasonal period ago, in our case, one year ago. As the demand pattern seems to approximately repeat yearly and there is only a subtle trend, we can expect the seasonal naive method to return quite accurate results despite its simplicity.

The accuracy of the methods are depicted in Table 4.1. The accuracy is about 86%, which is already quite high. However, the average deviation between the forecast and the observed demand is almost 200 units (1 unit = 1 crate of Heineken). From the graph in Figure 4.1 we can observe that the seasonal cycle quite constantly repeats itself over the years. However, we can expect more accurate forecasts from methods that incorporate other factors that influence the demand.

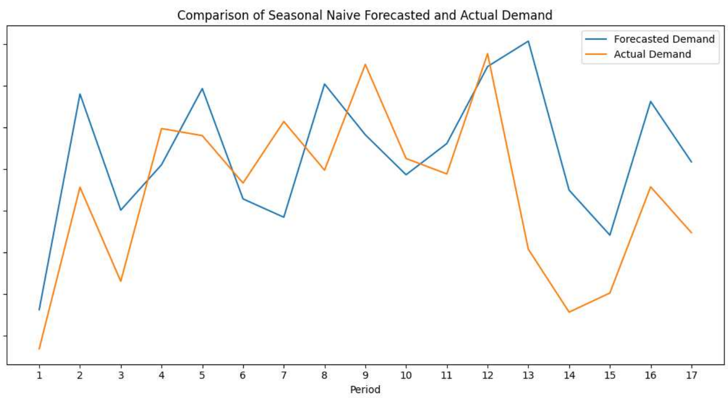


Figure 4.1 Comparison graph of the seasonal naive forecast and the actual demand (values hidden for confidentiality).

4.2 Holt-Winters

To continue, we test the Holt-Winters (triple exponential smoothing) method, as it well accounts for trends and seasonality. We first need to make a choice of additive or multiplicative seasonality (Matsumoto & Komatsu, 2015; Shiwakoti et al., 2023; Tirkeş et al., 2017). Additive seasonality is used when the magnitude of seasonality stays approximately constant throughout the time series (Tirkeş et al., 2017).

Multiplicative seasonality is used when seasonal fluctuations of the data change proportionally to the average level of the observations. To assess this we start by simply visualizing the graphed demand in Figure 2.2 in Section 2.2.3. We do not identify the fluctuation to be larger as the average observation increases. This would imply the use of additive seasonality.

Next, we compute the Coefficient of Variation (CV) to further analyze which seasonality seems to be more suitable. CV can be calculated from the following formula:

$$CV = \frac{\sigma}{\mu} * 100\% \tag{20}$$

We obtain a CV of 23.3% for the training set. This suggests that variability is relatively low compared to the mean, further confirming the choice of additive seasonality.

As our data contains only a slight trend, we test Holt-Winters with and without the trend component. This allows us to test whether trend is a significant factor in the data. We also choose to use additive trend as in Section 2.2.3 from Figure 2.2 we concluded that there may be a subtle linear upwards trend.

For developing the model, we continue by initializing the parameters. We need initial estimates of the level, trend, and seasonal components (Chopra & Meindl, 2016). The availability of data makes it more difficult to optimize parameters. We use the *ExponentialSmoothing* function from the *statsmodels.tsa.holtwinters* library in Python, suggested by Huang and Petukhina (2022). For using the implemented parameter optimization method it requires two full seasonal cycles of data, which we do not have. Thus, we must use other common methods for initializing the three components of the model. For the level, we use the value of the first data point. For the trend, we calculate the initial value as the average of the differences between consecutive data points. For the initial seasonal value, we calculate the ratio of all data points to the average of the observations during the same seasonal cycle (time of the year in our case) in the other years we have data from. The optimal smoothing parameters resulting from the model are presented in Appendix E2.

The meaning of the smoothing parameters is explained in Section 3.2.2. The smoothing parameter of the level component of about 0.23 (with a possible range from 0 to 1) implies that the model is able to respond to some fluctuation in the data, while also remaining relatively stable, keeping a balance between recent and historical observations. The smoothing parameter of the trend is very subtly positive, which is why it is not surprising that not incorporating trends can return more accurate results. It indicates that there is most likely a smooth longer term trend and that the data is more dependent on past observations instead of very recent ones. This further suggests the event-driven seasonality, with quite insignificant trends. The smoothing parameter of the seasonal component of about 0.74 implies that the model is well able to account for seasonal changes, putting emphasis more on the very recent observations.

Table 4.1 indicates the resulting measures of accuracy of Holt-Winters models. The results show that excluding the trend significantly improves the results, however, the difference between the forecasted values and the actual demand is still about 200 units. The bias tells us that on average the forecast tends to overestimate the demand. The accuracy may be affected by having less than two full seasonal cycles of data, since the Holt-Winters method expects seasonal patterns to stably repeat each cycle. One observation for a time period may not be enough to capture the seasonal fluctuation and thus, finding the optimal parameters is more difficult. However, for the ex-ante forecast we have close to two years of data, which can increase the accuracy. On the other hand, making a forecast for a longer horizon may balance out the improvements caused by more data.

From the graph in Figure 4.2 we can see, as indicated by the bias, that the forecast with trend tends to significantly overestimate the observed demand. We know that the training set had a slight positive trend, however, in the beginning of 2024 (where the testing set begins) the supermarkets decreased their price for Heineken. This is possibly the cause of a decreased demand in the first months of 2024, causing the inclusion of trend to predict inaccurately. By removing the trend we see much better results with an accuracy of about 85.5% and much lower bias, depicted also in Figure 4.2.

KPIs	Seasonal Naive	H-W With Trend	H-W Without Trend
MAPE	14.0%	21.1%	14.5%
RMSE	198.3	278.6	200.5
bias	93.8	221.4	100.6

Table 4.1 Results of the seasonal naive and Holt-Winters methods.

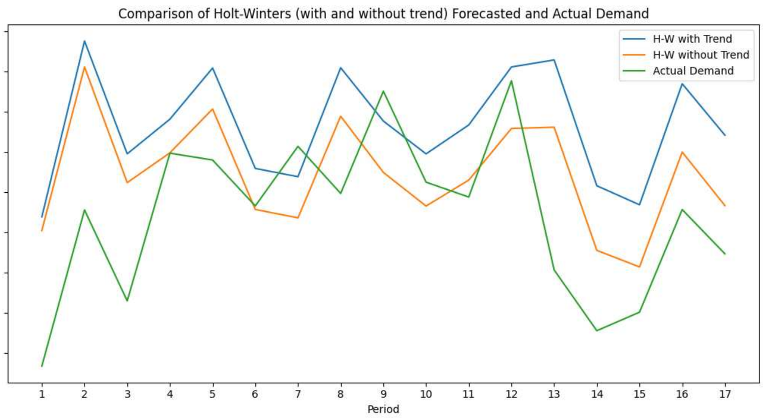


Figure 4.2 Comparison graph of the Holt-Winters forecasts and the actual demand (values hidden for confidentiality).

4.3 (S)ARIMA

In this section, we test the autoregressive integrated moving average (ARIMA) method with an extension of seasonality (SARIMA). The method has an assumption that the data is stationary, simply indicating that it does not contain trends or seasonality. In Appendix D, we analyze the data in detail by fitting a trend line and plotting the (partial) autocorrelation functions. Furthermore, we use a method that can be utilized to check whether a unit root is present in the data, called the Augmented Dickey-Fuller test. We conclude the data to be slightly non-stationary and difference the data to make sure the assumption is met. For the detailed process, refer to Appendix D.

4.3.1 ARIMA

To develop the ARIMA(p, d, q) model, we start by parameter selection. According to Huang and Petukhina (2022), there are two essential measures to take for determining the parameters p and q of ARIMA. The first one is computing and analyzing the ACF and PACF plots. The second aspect is using certain information criteria such as the Akaike Information Criterion and Bayesian Information Criterion. Furthermore, the source explains that the amount of times the data should be differenced to make it stationary (parameter d) should be determined before p and q, as models with different d values cannot be compared with AIC and BIC. In Appendix D, we determined the data to be stationary after one differencing, indicating d=1.

We compute the correlogram and PACF plot for the differenced data, depicted in Figure 4.3.

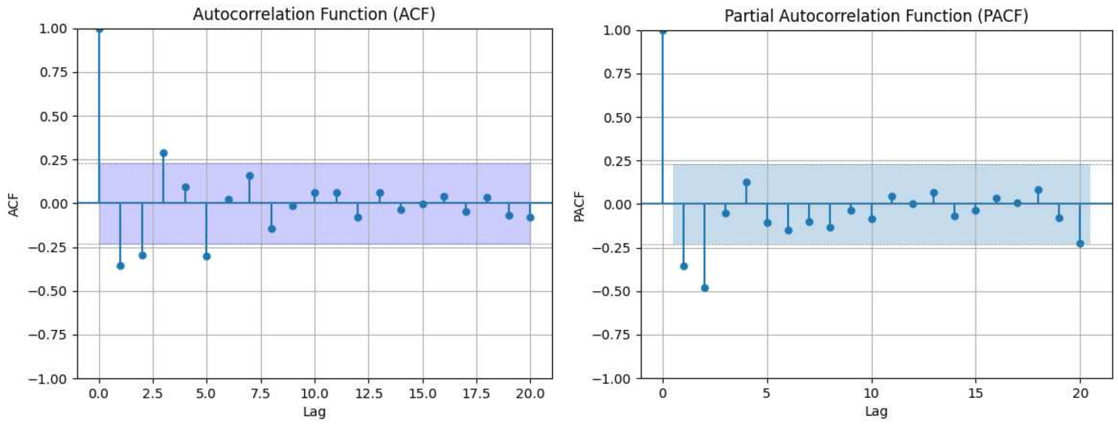


Figure 4.3 ACF and PACF plots of the differenced training set.

From the PACF plot we can see that there are significant values observed at the first two lags, before they move inside the significance threshold. This suggests that two autoregressive terms (AR(2)) for the model may potentially be suitable, meaning that the model accounts for the effect of the two previous weeks for each forecasted point.

Making a conclusion from the ACF plot is not as clear. There are several significant values, suggesting that there might be an MA process present indicating the usefulness of an MA or ARIMA model instead of only an AR process. Values outside the significance threshold can be seen at lags 1, 2, 3, and 5. It may suggest some additional patterns in the data, but as they are not at equal intervals, it is difficult to say whether it is a seasonal pattern. We must further examine the parameter selection for the use of ARIMA. Thus, we use AIC and BIC to help with model selection.

AIC and BIC are information criteria that evaluate the quality and fit of different models for comparison. We compute AIC and BIC for all combinations of p and q between 0 and 3 to find the model with the best performance, taking into account the goodness of fit as well as complexity of the model. AIC may often overfit the model by preferring models that have too many parameters, while BIC places a larger penalty for additional parameters (Huang & Petukhina, 2022; Onder & Wei, 2022). The models that return the lowest values of AIC and BIC should be used. For specific formulas for AIC and BIC, refer to Huang and Petukhina (2022, p. 115).

Both AIC and BIC receive low values for the ARIMA(2, 1, 0) model, which is also suggested by the PACF plot. In addition, AIC suggests ARIMA(3, 1, 3) and BIC ARIMA(0, 1, 1). As mentioned, BIC penalizes extra parameters more, which is why it makes sense that the suggested model of AIC contains more parameters. We further analyze these models to determine their suitability.

Huang and Petukhina (2022), and Singh et al. (2024) introduce the maximum likelihood estimation (MLE) method for estimating the values of the AR and MA parameters, which finds the values with the highest chance of returning the observed data. MLE assumes that the data points are independent and identically distributed, which is why we analyze diagnostics of the resulting models to approximate whether we can make these assumptions. Based on the Q-Q plot of the residuals of the models, they are approximately normally distributed. In addition, the correlograms do not show lags with significant values. However, again the lack of data makes it more difficult for the MLE to determine optimal parameter values, which may decrease the accuracy of the model. The models result in the KPIs presented in Table 4.2, with parameter values presented in Appendix E4.

Comparing the values of Table 4.2, it is not straightforward to conclude whether ARIMA(3, 1, 3) or ARIMA(0, 1, 1) is the best model for the data. We use k-fold cross-validation, which helps to further analyze each model to get a better idea of their performance on different subsets of the data. The results of the cross-validation confirm that ARIMA(2, 1, 0) is less suitable as it results in a significantly higher average MSE. The value for ARIMA(0, 1, 1) is slightly lower than that of ARIMA(3, 1, 3) and thus, we can conclude it may be a better fit for the data. However, looking at comparison graphs of the forecasts and the actual demand we notice that the models do not account for seasonal fluctuations well. For reference, Figure 4.4 depicts the graph for ARIMA(3, 1, 3).

4.3.2 SARIMA

The basic ARIMA model does not explicitly account for seasonality, which is why we want to extend the model, knowing the data contains yearly seasonality. The SARIMA model adds seasonal order parameters to the basic model, taking the form SARIMA(p, d, q)(P, D, Q, m), where m corresponds to the length of the seasonal cycle.

We test the SARIMA model to see whether it results in higher accuracy than the basic models. Again, we use optimizations methods such as AIC to determine p and q as well as the seasonal orders P and Q, and MLE to determine the seasonal and non-seasonal AR and MA parameters. The resulting optimal model, with weekly seasonality (m=52), is the SARIMA(2, 1, 1)(1, 0, 0, 52). Note that since we have less than two full seasonal cycles of data, the seasonal differencing D is 0. Looking at Table 4.2, the MAPE is slightly higher compared to some of the basic ARIMA models, however, as can be seen from the graph depicted in Figure 4.4, the model significantly better accounts for the seasonal fluctuations. Additionally, the model results in a lower RMSE indicating smaller differences between the forecast and actual demand.

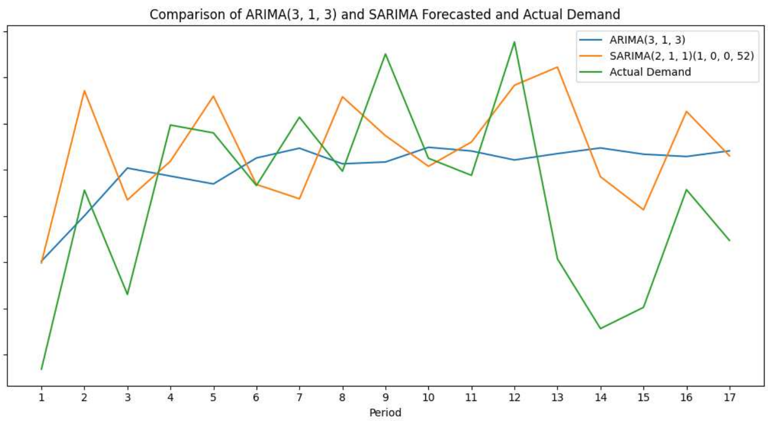


Figure 4.4 Comparison of the ARIMA(3, 1, 3) and SARIMA(2, 1, 1)(1, 0, 0, 52) models and the actual demand (values hidden for confidentiality).

KPIs	ARIMA(2, 1, 0)	ARIMA(3, 1, 3)	ARIMA(0, 1, 1)	SARIMA(2, 1, 1)(1, 0, 0, 52)
MAPE	15.8%	14.1%	14.0%	14.5%
RMSE	233.2	195.1	202.7	194.8
bias	-116.4	61.4	56.2	101.9

Table 4.2 Results of the (S)ARIMA models.

4.4 Multiple Regression

We continue to test a causal model for forecasting, multiple linear regression. We include two exogenous variables, the exams and events that take place in the four delivery cities, expected to affect the demand. Turff has developed a year calendar for September 2023 to August 2024 containing the exam and resit periods as well as some significant events. We use this to develop the values of the exogenous variables. However, as the company has not made a similar calendar for August 2022 - August 2023 (the remainder of the training set), we make an assumption that the exam periods and events took place in the same weeks during that time period.

The first exogenous variable is Exams, including resits, incorporated as a binary variable. It takes the value 1 if there are exam or resit periods in any of the cities during that week, and 0 otherwise. The exam periods are expected to have a negative impact on the demand of Heineken, as the customer base consists of student houses.

The second exogenous variable is Events, expected to positively influence the demand. A few of the main events per city are included, as well as events that affect all the cities (i.e. King's Day). The independent variable is scored on a scale from 0 to 8, depending on the number and significance of the events happening that week. For more details of the scoring, refer to Appendix E5. Unlike the exams, the variable Events is shifted one week prior as the company often experiences peak demand the preceding week.

Next, we consider the assumptions of the multiple linear regression model, including that the residuals should be random and normally distributed, and that the independent variables do not have multicollinearity (F. et al., 2021). In addition, there should be a linear relationship between the dependent variable (demand) and the two independent variables.

Before developing the model, we can test whether the independent variables have multicollinearity. F. et al. (2021, p. 2) proposes to use Variance Inflation Factor (VIF) to test whether the assumption is met, describing it in the following way: "VIF measures how much the variance of the estimated regression coefficient is inflated if the independent variables are correlated." A VIF of 1 indicates no correlation between the variables. Furthermore, a value between 5 and 10 shows some correlation possibly causing problems for the use of the variables. We obtain a result of 1.00022 for both of the independent variables, indicating no significant collinearity.

Next, we use an F test to evaluate if a linear relationship exists between the demand and the independent variables (F. et al., 2021). The null hypothesis of the model is that the model is not suitable for predicting the dependent variable, and if it is rejected it indicates that at least one of the suggested independent variables affects the demand. We use the analysis of variance to conduct the test and utilize R^2 as a measure of how well the model can explain the data, with R^2 values ranging between 0 and 1 (F. et al., 2021). An R^2 value close to 1 means that the model fits the data well as the variability is small. We use a significance level of 0.05. We fail to reject the null hypothesis, indicating that the model cannot be used for forecasting the dependent variable. In addition, we obtain an R^2 value of 0.024, confirming a poor fit as the model only explains 2.4% of the variables' relationship.

We further test the model by taking logarithms of both the dependent and independent variables, following Winston (2004). This is done when the relationship between the variables is expected to be non-linear. Because our data contains zeros, we must adjust the data by using $ln(x + 1)$ for each value, since the logarithm of 0 is undefined. Despite the changes, we again fail to reject the null-hypothesis of the F-test and the R^2 value remains low of about 0.022.

We can conclude that multiple (linear) regression is not suitable for the data and thus, move on to test SARIMAX, a model that combines time series forecasting with exogenous variables, and random forest, a more advanced machine learning technique.

4.5 SARIMAX

The SARIMAX model adds regression components to the SARIMA model used in Section 4.3.2, which is why it can be described as a combination of time series methods and causal methods. We have reason to believe that adding exogenous variables will further improve the model as we are dealing with event-driven seasonality. It allows us to implement the effect of external factors and estimate the correlation between them and the demand.

The two independent variables are the same as explained in Section 4.4 for multiple regression and details of them can be found in Appendix E5. We also test a third independent variable, a categorical variable taking into account the month the demand occurred in. However, this does not improve the model and thus, is left out. The model assumes the last observed values of the independent variables will continue on to the future, as this decrease in complexity was found to improve the results.

We can expect the SARIMAX model to result in a more precise forecast compared to the multiple linear regression model as it better accounts for seasonal factors. We have concluded that the event-driven seasonality is a significant aspect of the demand data. However, the SARIMAX model does bring additional complexity to the model.

By adding the exogenous variables we are able to increase the forecast's precision, shown in Table 4.3. The MAPE indicates the accuracy of the forecast to be about 87.9%. As mentioned in Section 1.3.6, explained by Da Veiga et al. (2014), a MAPE below 20% indicates a potentially good forecast. The RMSE tells us that on average the difference between the forecasted values and the observed demand is 165.6 units. The bias of 35.3 indicates that on average the forecast tends to overestimate the real demand to a small degree. In addition, from Figure 4.5 we can see that the forecasted values are closer to the actual demand compared to the previously tested models.

KPIs	SARIMAX(2, 1, 1)(1, 0, 0, 52)
MAPE	12.1%
RMSE	165.6
bias	35.3

Table 4.3 Results of the SARIMAX(2, 1, 1)(1, 0, 0, 52) model with two exogenous variables.

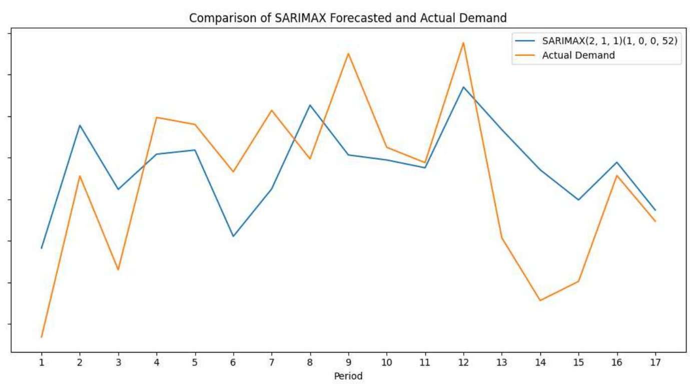


Figure 4.5 Comparison graph of the SARIMAX(2, 1, 1)(1, 0, 0, 52) forecast and the actual demand (values hidden for confidentiality).

The correlation coefficients of the SARIMAX model vary between -1 and +1 depending on their impact on demand, and are shown in Appendix E6. The value for exams is just below zero, indicating a minor negative effect on the demand during these periods. The coefficient of events, on the other hand, is positive showing that more demand is observed in the week before events. This further indicates that the model fits, as these relationships have been observed by the company in the past.

4.6 Random Forest

Last, we test random forest from the ML techniques. We can classify ML as supervised, unsupervised, and semi-supervised depending on the used data. Supervised learning utilizes labeled data, which in our case is the real historical demand data. Supervised learning was also used for multiple regression. Unsupervised learning utilizes only unlabeled data, while semi-supervised learning uses labeled and unlabeled data. We will test supervised and semi-supervised learning to compare results. We obtain unlabeled data by using the other developed forecasting models to generate more data points for the dependent variable. This allows the algorithm to learn with more data, and thus we expect it to provide better results.

We begin by developing the random forest model using only real historical observations, classified as supervised learning. The first step is to define the independent variables. We test the following options:

- The events as a categorical variable or numerical variable with the scale of 0-8 (as defined for regression and SARIMAX)
- The month or the week the demand occurred in, as a categorical variable
- The exams as a binary variable

After testing several combinations, we again choose to use Exams and Events with the same numerical values as in SARIMAX and regression, due to them providing the most accurate results. We standardize the variables to have a mean of 0 and standard deviation of 1, commonly done in machine learning to adjust the independent variables to the same scale to ensure one of them does not have a more significant effect. Furthermore, we add a categorical variable Month, simply indicating the month the demand occurs in. The model automatically divides the data into a training and testing set. We choose to reserve 20% of the data for testing, as that is approximately what was done for the other forecasting methods.

We use grid search to optimize parameters of the model. It is a technique that tests several combinations of parameter values to determine the most optimal ones for the data. It is based on creating a “grid” of all the possible combinations of parameter values based on set limitations, and uses cross-validation to train and evaluate them. It then uses the R^2 values of the combinations to choose the best parameters. For details of the performed grid search and the best values resulting from the tested models, refer to Appendix E7.

Table 4.4 indicates the results of random forest. We use R^2 as a third indicator instead of the bias, as it gives us more indication on how well the model captures the patterns. We can see that the resulting values of MAPE and RMSE for the supervised model are clearly worse compared to SARIMAX. Additionally, only 25% of the fluctuation of the demand is captured by the model based on the R^2 value. This is significantly better than that of multiple regression but still does not well reflect the historical demand.

KPIs	Random Forest, Supervised	Random Forest, Semi-supervised
MAPE	16.6%	10.3%
RMSE	248.7	174.2
R^2	0.25	0.16

Table 4.4 Results of the supervised and semi-supervised random forest techniques.

With the aim of improving the accuracy and robustness of the forecast, we move on to test semi-supervised learning with random forest where additional data is used to allow the model to train on a larger set. We generate additional data points to have about 3.5 times the amount of the historical data. We do not change anything else in the model. This results in an accuracy of almost 90%, which is an improvement from SARIMAX. However, the resulting RMSE is higher. We can observe that the more data we feed for the model the better it seems to perform. Furthermore, the R^2 value is lower than for the supervised learning model, indicating that the model does not well capture the patterns of the demand.

The feature importance diagram in Figure 4.6 shows how important each of the independent variables are for the model for predicting demand, specifically for the semi-supervised model. It shows the relative importance of all the months used as a categorical variable, as well as Exams (remainder_x0) and Events (remainder_x1). It is computed based on how much the variable reduces uncertainty in each decision tree. In other words, how important the variable is for making an accurate prediction.

We see that for the semi-supervised model, the months where demand is higher and varies more have higher importance, especially September, August, and July are important for predicting the dependent variable. Comparing to the feature importance diagram for the supervised model depicted in Figure E7.1 in Appendix E7, we can notice that the importance of the months with higher demand variability becomes higher with the use of more data. The additional data may improve the model's ability to capture these patterns and relationships. The differences in the importances of the months for both models indicate significant seasonality.

We can also see that Exams and Events are important features. However, a question arises of why Events is more important, especially when the scales are standardized. We test to adjust Events to a binary variable to match Exams for the semi-supervised model to see how the feature importance values are affected. It results in the importance of Events dropping significantly, from about 0.22 to about 0.15. However, this is still clearly higher than the importance of Exams of close to 0.05 and thus, must also be caused by other factors. The absolute value of the correlation of Demand and Events is higher than that of Demand and Exams, as depicted in the heatmap in Appendix E7 Figure E7.3, which might be one reason for the discrepancy in importance. Additionally, the occurrence of events is more frequent than exams. We also know that the demand is driven especially by events, which is why the importance of events being higher is no surprise. For further analysis on the independent variables' effects on demand for the supervised and semi-supervised models, refer to Appendix E7.

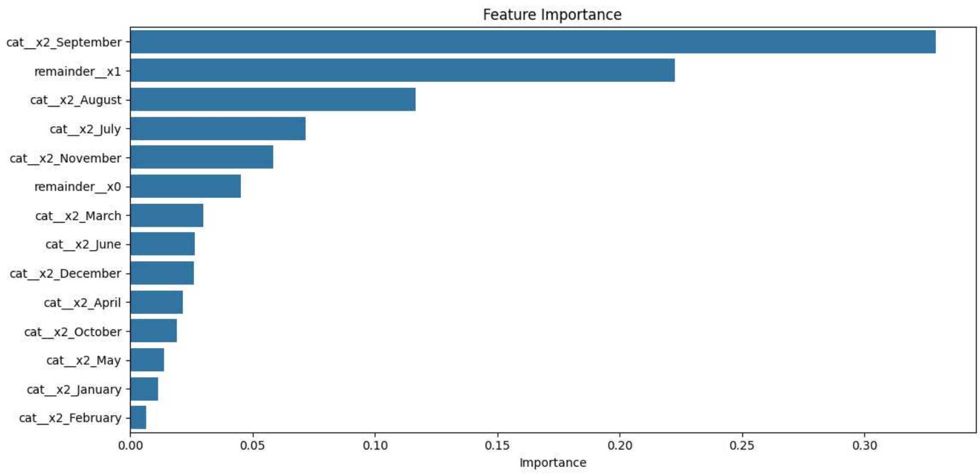


Figure 4.6 Feature importance of the independent variables used for semi-supervised random forest.

4.7 Model Selection and Ex-ante Forecast

In this section, we conclude the best forecasting method by comparing the KPIs. After choosing the best method and analyzing its results, we develop the final ex-ante forecast.

Analyzing the performance indicators of the different forecasting models, we can conclude that SARIMAX(2, 1, 1)(1, 0, 0, 52) and the semi-supervised random forest outperform the other models. SARIMAX results in a MAPE of 12.1% and an RMSE of 165.6, while semi-supervised random forest in the same KPIs of 10.3% and 174.2, respectively. The lower RMSE indicates that the forecast errors are smaller on average, measuring in the same unit as demand. RMSE penalized larger errors more, which is a suitable metric for the company, as large errors can cause more issues in some weeks. The lower MAPE of random forest indicates that the percentage accuracy of the forecast is slightly better, however, reducing larger prediction inaccuracies is more important for developing a more optimal inventory policy. Large errors can lead to unreasonable amounts of excess stock or stockouts. Furthermore, the low R^2 value of the semi-supervised random forest of about 0.16 indicates that the model does not capture the patterns of the data well. Last, the reduced complexity of the model further backs up the selection of SARIMAX for the ex-ante and future forecasting. From the results of the forecasts we can observe that adding complexity does not always improve results. Adding a seasonal component to ARIMA does not significantly increase accuracy, as well as the selection of SARIMAX over random forest.

Although the bias does not straightforwardly tell us about the model performance, comparing the absolute value of the biases of the time series models, SARIMAX also has the lowest value. The reason why bias cannot be used for making a conclusion of the precision of the forecast is because it takes the average of the differences of the forecast and the actual demand. For instance, if biases of -100 and 100 are observed it will result in a bias of 0, incorrectly describing the accuracy of the forecast. From the bias of 35.3 for the SARIMAX model we can conclude that the forecast tends to slightly overestimate the demand.

The last method we want to test is taking a simple and a weighted average of two forecasting models with the aim of further increasing accuracy, as explained by Chopra and Meindl (2016), and Onder and Wei (2022). For a weighted average, we base the weights on the resulting RMSE of the models. We choose the seasonal naive as the second most optimal method for this, as it is more feasible in the future. Choosing random forest would make future forecasting extremely complex, as it would require the use of both SARIMAX and semi-supervised random forest, which needs additional data generated by additional methods. Regardless, the combined forecasts (simple and weighted) result in higher values for MAPE and RMSE, which is why we continue with SARIMAX.

We graph the obtained MAPE and RMSE values of the chosen model, depicted in Figure 4.7. We can see that there is high variability ranging between about 1% and 33% for MAPE and between 12 and 314 for RMSE. This should be taken into account in inventory decisions to ensure the probability of stockouts is kept to a minimum. Possible approaches are to increase service level (leading to a higher safety stock) or implement a dynamic safety stock based on the expected error or standard deviation.

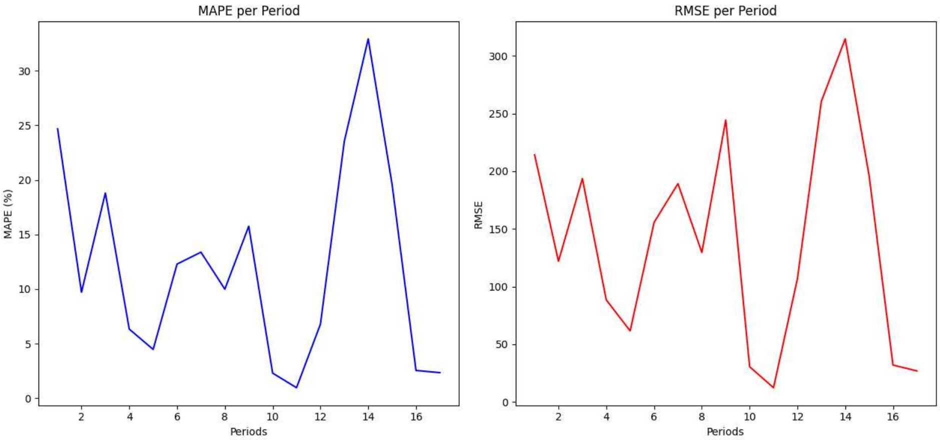


Figure 4.7 Obtained MAPE and RMSE values per week of the testing set for the SARIMAX(2, 1, 1)(1, 0, 0, 52) model.

For the result of the ex-ante forecast we can expect accuracy to slightly decrease due to a longer forecast horizon. On the other hand, we add the data from the testing set as well as new observed data, which can increase accuracy. Additionally, only one year's worth of data was available for the ex-post forecasts for the period of the testing set (weeks 1-17). This led to almost all time series models to clearly overestimate the demand, especially in weeks 13-17 when the observed demand of 2023 was significantly higher than that of 2024. For the ex-ante forecast, we have two years worth of historical observations for most of the horizon excluding about a two month period, which can make the forecast more robust. After similar parameter optimization as done for the ex-post forecast, we conclude to use SARIMAX(1, 0, 1)(0, 0, 1, 52). We cannot calculate the error of the ex-ante forecast, however, when we graph the forecast, depicted in Figure 4.8, we can compare it to the graph of the available historical data in Section 2.2.3 Figure 2.2 and conclude that the model well reflects the event-driven seasonal patterns into the future. We also plot the ACF of the residuals of the ex-ante forecast, and see that there are no lags showing crucial correlation, indicating a good model fit. The parameters of the ex-ante forecast are presented in Appendix F.

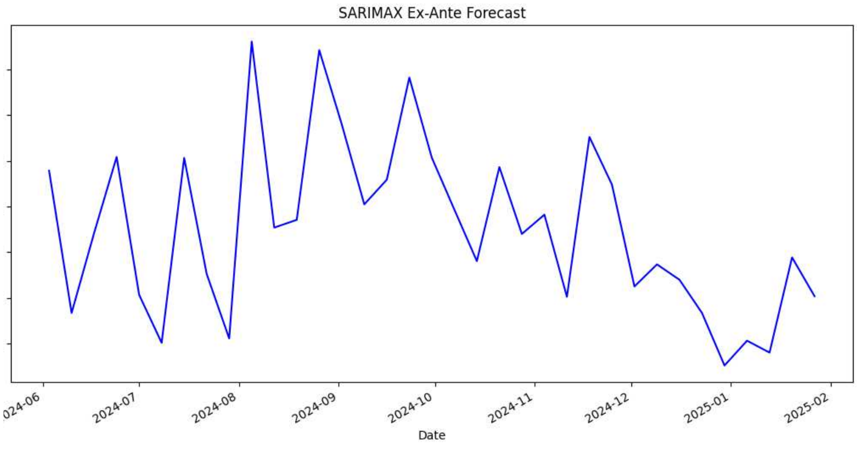


Figure 4.8 Final ex-ante forecast using the SARIMAX(1, 0, 1)(0, 0, 1, 52) model (values hidden for confidentiality).

In this chapter, we have answered two of the sub-research questions: *How can the chosen forecasting methods be applied to the case of Turff?* and *How can we assess the accuracy of the chosen methods?* We have done this by testing various forecasting methods with a training and testing set (ex-post) and analyzing the results with KPIs. Furthermore, we computed the ex-ante forecast until the end of January 2025. Developing a forecast gives us valuable information on the expected fluctuation of the demand, however, we need to take one step further to improve decision making by analyzing optimal order quantities. Thus, the ex-ante forecast will be utilized in Chapter 5 for developing an ordering policy.

Chapter 5 - Ordering Policy

In this chapter we want to answer the sub-research question: *How can we optimize inventory levels by developing an ordering policy for Heineken?* We will use the demand forecast from Chapter 4 to estimate optimal order quantities using the (*R, S*) inventory policy. We develop a dynamic policy to ensure it fits the fluctuating demand.

5.1 Selection of Approach

In this section, we choose the type of policy that best fits the way of replenishment ordering at Turff. First, options include a continuous and periodic review policy. In the former, a new order is placed every time the inventory level drops below a predetermined level, while in the latter, the inventory is reviewed in equal time intervals and an order is placed to bring the on-order inventory up to a certain level.

We choose an (*R, S*) periodic review policy, since Turff has a contract with their supplier for weekly replenishment orders. It works in a way that every time period *R* we place an order to bring the on-order inventory up to *S*. A periodic review policy is easier to manage, however, can lead to higher holding costs as more safety stock is required. Orders are usually placed on Friday and delivered Monday, resulting in a constant lead time of three days. Thus, the lead time plus review period (L+R) equals 10 days. Lead time is defined as the time between order placement and delivery for replenishment orders (Chopra & Meindl, 2016). Based on data from 2024 (the testing set), approximately 39% of weekly demand occurs during lead time (Friday-Sunday).

The second step is to choose the suitable approach. In Section 3.7.2, we introduced two different approaches for the (*R, S*) policy. The first one is based on cost minimization while the second focuses on a service level approach. After analyzing the current method of ordering as well as consulting the company supervisor, we have concluded that the service level approach is more suitable for Turff. The current ordering is based on ensuring that all demand can be met, as the cost of lost sales is high compared to overstock. The company prefers to keep a similar approach in the future. Furthermore, the lack of specific information on different costs, such as holding costs, means that rough estimates may not bring the company desired results.

The ordering policy will be based on the developed forecast, making it dynamic. This means that there is no constant value *S*, but it is calculated per each review period to account for seasonal fluctuation of the demand. Thus, equation (18) is altered to base the order up to level S_i on time *i*:

$$S_i = E(D_{L_i+R_i}) + ss_i \tag{21}$$

Equation (21) indicates that we first need to estimate expected demand for each week as well as the safety stock. $E(D_{L+R})$ is the forecasted demand for the review period and the lead time, varying per week. Thus, for example, calculating the expected demand for a replenishment order made on Friday in week 1 for week 2 sums up the forecasted demand of week 2 and 39% of the forecasted demand of week 1. Figure 5.1 visualizes how the (*R, S*) policy works, including varying demand and a dynamic safety stock based on the standard deviation of the period.

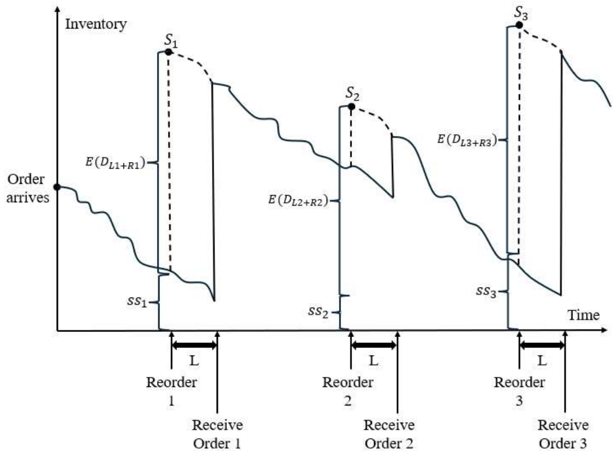


Figure 5.1 Visualization of the periodic (*R, S*) policy with dynamic order up to levels S_i.

5.2 Safety Stock Determination

Safety stock is inventory used to meet demand exceeding the forecasted orders (Chopra & Meindl, 2016). It grows larger the higher the desired availability of the product is for the company (service level). Contrary to continuous review policies where safety stock is affected only by the variability of the lead time, for the periodic policy we must account for the period of $L+R$ (Silver et al., 2017). The dynamic safety stock is calculated per week, utilizing a chosen CSL and standard deviation of demand during the lead time and review period. Again, we add time to equation (19) to make safety stock a dynamic variable:

$$ss_i = F_s^{-1}(CSL) * \sigma_{D_{L_i+R_i}} \tag{22}$$

A suitable level of safety stock is affected by the following aspects: the unpredictability of demand and supply, the desired level of availability of the product, and the used replenishment policy (Chopra, 2019). The uncertainty of demand for Heineken is relatively high, as the demand patterns have significant fluctuation and there is limited data available for comparison and forecasting. Supply uncertainty, on the other hand, is not significant due to the reliability of the supplier. The company wants to have almost full availability of the product, as the cost of a stockout exceeds that of overstock, further increasing the required level of safety inventory. Last, there is little flexibility on placing the replenishment orders, as the contract is fixed to once a week orders (and possibly a second order if it is communicated early with the supplier). A continuous review policy would allow for a lower safety stock for the same level of availability (Chopra, 2019). Overall, these factors indicate that the safety stock for Heineken will be relatively high.

To calculate the safety stocks, we start by determining the desired CSL. In Chapter 4 we showed that the MAPE was on average 12.1%, although it fluctuates among weeks. Turff wants to have a high service level to satisfy most demand. We do a simple sensitivity analysis on service levels between 92% and 99%, comparing the resulting fill rate and number of periods without stockouts with the testing set. Fill rate is the individual units of demand that can be met from stock. Based on the experiments and expert opinion, we select a CSL of 96% to balance between high availability and excess inventory. The high service level can also account for the known fluctuation of the MAPE that brings uncertainty to the ordering policy. Additionally, a high service level is preferred when the cost of lost sales exceeds that of excess inventory.

Historical data is typically used to calculate the standard deviation used for safety stock calculation, however, having only few data points for each week may result in an unreliable value. We exclude data from 2022, as it is significantly lower compared to the other years and would result in a less accurate standard deviation. Furthermore, from weeks 23 to 52 we have only one observation (from 2023). Thus, we aggregate the standard deviation per month, and include values from additional forecasting models developed in Chapter 4 with the aim of obtaining a more reliable estimate.

To account for deviation in the lead time, we sum the variances from 39% of the week where L occurs and the variance from the week of the review period, and take the square root to obtain the standard deviation. Based on a visual inspection of the histogram and Q-Q plot as well as a Shapiro-Wilk test for normality, we know that the full data set approximately follows a normal distribution. For simplicity, we assume that each monthly data set is also normally distributed, with an individual mean and standard deviation.

5.3 Development and Analysis of the Policy

We test this policy on the testing set also used for forecasting, as that is the period we have real data and forecasted data available. Comparing the suggested S and the actual observed demand from $L+R$, we notice that it results in a stockout of a small amount of units for two periods. This results in a CSL of 87.5% (=14/16). The individual units of demand that can be delivered from stock is 99.6% (fill rate), which is a better measure for a short testing set of 16 periods. The main challenge with the policy is that the resulting overstock from the other 14 periods fluctuates significantly, ranging from about 10 units up to about 580 units. This amount of excess inventory can lead from the fluctuating accuracy of the forecast, as seen from Figure 4.6. In addition, as explained, the nature of the product for Turff is such that it requires a relatively high safety stock. We test another approach with the aim of obtaining a more robust solution.

The second option we test is utilizing RMSE as a measure of standard deviation, which results in a constant safety stock. RMSE can be estimated weekly by comparing historical forecasts and actual demand to allow for a dynamic safety inventory, however, we do not have enough data to use this approach. We use RMSE in the same way as the standard deviation in the previous approach but with a constant value of 165.6 obtained for SARIMAX. For the testing set, we obtain a service level of 100%, with overstock fluctuating between 60 and 630 units.

From the two approaches tested we can conclude that the first method with a dynamic safety stock is more suitable for Turff and thus, we continue with this policy. The first approach results in close to all demand met as well as lower fluctuation of overstock than the second policy. Figure 5.2 depicts the fluctuation of MAPE for the testing set and the excess inventory that results from the ordering policy for the same duration. We can see that the periods with less accurate forecasts often result in stockouts or higher overstock, showing that the excess inventory is necessary to cover for the unexpected fluctuation, as well as indicating the importance of a precise forecast as it directly reflects on the ordering policy.

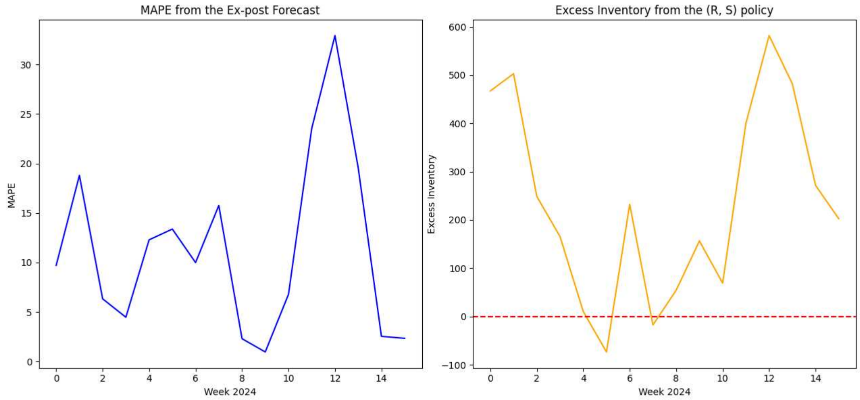


Figure 5.2 Comparison of MAPE and resulting overstock for the testing set.

Additionally, Turff offers the following option for customers: while receiving a delivery, the customer has the possibility to return six empty beer crates in exchange for a free crate of Heineken. This must be accounted for in the order up to levels. We use an estimate of the average number of free crates given out per *L+R* in 2023-24, which is 57 crates. Consequently, 57 crates are added to each order up to level.

In Figure 5.3 we compare the final resulting order up to levels with the demand that occurred on the same week in the year before. It shows that the dynamic policy matches the expected fluctuation of the demand. The gap is caused by the difference in the forecast and historical demand, the added safety stock, as well as accounting for the demand in *L+R* instead of only for one week. Looking at the individual weeks more specifically, we see that for example in the middle of November, the historical demand decreases but *S* increases. As the forecast for that week decreases as well, the discrepancy tails from the safety stock. This can happen due to the monthly aggregation of standard deviations, which was done due to the availability of few data points for each week. Furthermore, as Heineken has a short lead time of 3 days it allows us to see peaks in the order up to levels in the same weeks as the demand peaks. For longer lead times it is common to see lagged demand peaks since orders are placed in advance to anticipate for this.

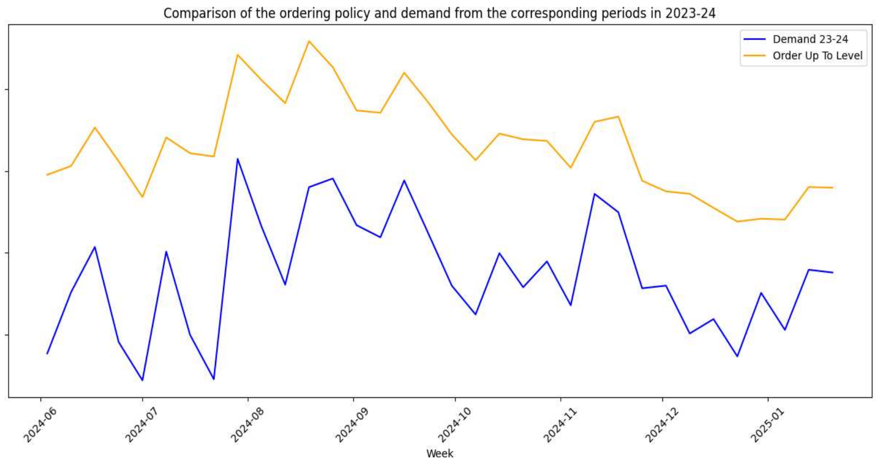


Figure 5.3 Comparison of the order up to levels and the demand from the year before (values hidden for confidentiality).

The periodic (*R, S*) policy is used to calculate the order up to level *S*, which is why it is vital that the current inventory level is subtracted for replenishment decisions. Thus, the order quantity is calculated as $Q = S - I$.

Implementation of the policy requires that orders are consistently placed on Fridays. This is currently the case in most weeks, however, deviation from it makes the policy less reliable. Furthermore, in the current policy we used additional values generated by some forecasting models to acquire more data points per week as well as calculated standard deviations with monthly aggregation. When more data is available in the future, the standard deviation can be calculated only using historical figures for a more reliable estimate and possibly per week with the aim of increased precision.

Chapter 6 - Conclusion and Outlook

In this chapter, we conclude how we have answered our main research question and make recommendations on how the processes can be enhanced by gathering additional data, automation, and further research. By doing that we aim to answer the final sub-research question: *How can the forecast be improved and applied in the future?*

6.1 Conclusions

Following the sub-research questions based on the CRISP-DM methodology, we have answered the main research question: *How can inventory management be improved at Turff for their best selling delivery product, Heineken, in order to increase forecasting accuracy from about 75-78% to 85-90% and optimize inventory levels?* Accurate demand forecasting is vital for companies to optimize inventory decisions. The findings indicate that the company can benefit from a more scientific demand forecast and inventory policy. We are able to increase the accuracy of the forecast from about 75-78% to 87.9%. The used scientific methods allow for easier further automation of inventory processes as well as optimizing inventory levels. The more automated process also reduces dependency on one employee.

Our aim was to develop a forecast by testing multiple methods to find one with suitable characteristics for the demand data of Heineken. We started by analyzing the current method of decision making as well as the historical data available. Currently, replenishment decisions are made by experience based calculations by comparing the percentage increases of the previous periods as well as from the year before. We analyzed the data to have yearly seasonality, highly affected by outside factors, which limited our choice of forecasting methods. We explored suitable options for the time series data by conducting a literature search to get an understanding of what methods match the characteristics of the data. We chose to test the seasonal naive due to the yearly seasonal patterns. The method is simple but fits the data well and provided accurate results. We tested the Holt-Winters method with and without trend, which led us to the conclusion that trend is not a significant component of the data, as the model without trend returned significantly better results. Furthermore, we tested ARIMA with its extensions of seasonality and exogenous variables. In this case, increasing the complexity of the model improved its performance. Finally, we tested multiple (linear) regression and random forest from the introduced causal methods. Multiple regression did not capture the fluctuation of the data well with the independent variables tested, while we found random forest to perform better.

Testing a wide range of models provided us with further insights on the characteristics of the data and allowed us to compare what the ideal model is to use in the future as well. The selection of the best method was made based on the KPI values of each model for the testing set. We chose to use SARIMAX, which is a time series model that incorporates exogenous regressor components. The model is well suited for time series data that is not only affected by its past observations but also independent factors that affect the demand. These characteristics fit the demand of Turff. Another forecasting technique that well competed with SARIMAX was random forest with semi-supervised learning. The model resulted in a higher percentage accuracy of 89.7%, however, the worse RMSE value and the complexity of the model led to the selection of SARIMAX. Semi-supervised learning requires the generation of additional data points, which is why using the method is not only more complex itself but also requires the use of other forecasting models. Random forest can be a great option for the future when more representative data is available, however, for now SARIMAX is more suitable and better at capturing the fluctuating demand patterns.

Developing the forecast allowed us to make an inventory policy based on a service level approach to directly assist in decision making for the product. Although we were able to improve the forecasting accuracy, the fluctuating error still brings uncertainty and can result in non-optimal inventory levels in some weeks. On the other hand, the higher safety stocks on the weeks of demand with high variability are necessary to avoid possible stockouts. In Section 6.2.1 we make recommendations on how the matter can be improved upon in the future.

The resulting MAPE of 12.1% of the ex-post forecast tails from having only one historical observation of the same time period and thus, the forecast directly relies on the fluctuations of the previous year. The forecast and the observations of 2023 for the same time period are only about 6.6% apart, likely caused by the effect of the independent variables and the trends and patterns captured from the rest of the data. Starting from week 31, we have 2 years worth of data per week and thus, expect the forecast to be more robust. The more data we gather the more we can expect the forecast to improve. For now, it was of high importance to develop a model with higher accuracy and more automation than the current method.

The validity and reliability of the forecast can be considered quite high from the perspective of the consistent tracking of the historical demand data. However, the validity is decreased by the use of interpolation to account for the few weeks when stockouts occurred as well as having a shorter testing set than the forecast horizon. Furthermore, using the results of multiple forecasts to acquire more data points for the determination of standard deviations for the safety stocks was done with the aim of increasing accuracy. However, using only historical demand values, in case more is available, will improve the validity. The performance of the forecast and the inventory policy is evaluated with KPIs, giving confirmation on the reliability of the results.

6.2 Future Considerations

The section consists of discussing various recommendations for the company and the assumptions made in the research that limit the performance of the models as well as how they can be lifted with further research.

6.2.1 Recommendations

In this section, we discuss how the forecast accuracy can be improved in the future by gathering additional data for further analysis, as well as discuss how automation of the inventory process can improve results.

First, we recommend Turff to keep more detailed track on the occurrence of stockouts. This includes the amount of time the stockout lasts, regardless of whether any emergency solutions are used to meet demand as this does not reflect the demand met from ordered stock. We used interpolation to estimate the lost demand, however, more information helps make the estimations more valid and the forecast more accurate. Second, tracking promotions of Heineken and other products of the same category can provide valuable information on their dependency with the demand. Last, price changes from Turff and from competitors can significantly impact demand, which is why it is important to gather details on the effects. The promotions and price changes can be added as exogenous variables to the model when more data is acquired. The times and the extent of promotions and price changes should be noted in order to be able to inspect the demand patterns afterwards to analyze whether the demand was affected and how. Furthermore, the two currently used exogenous variables, Exams and Events, should be carefully monitored when the model is applied again. The current values of the variables represent the occurrences in 2023-24 but must be checked whether they land on the same weeks in the future. More analysis can also be done on their relationship with the demand with the aim of adjusting the values and scales of the variables for improved forecast accuracy.

We advise Turff to set up KPIs (MAPE, RMSE, bias) to monitor the performance of the forecast by frequently comparing the forecast to the observed demand of the period utilizing the KPI equations found in Appendix B. This will further allow the company to estimate how much trust can be placed on the forecast for decision making. If the forecast starts to stabilize with the use of more data, the company can estimate whether they wish to lower the service level to lower safety inventory and thus, reduce holding costs. This is dependent on whether the levels of excess inventory are perceived as too high when demand uncertainty is reduced.

We recommend the forecast to be regenerated more frequently than the eight-month horizon used in this research. Shorter term forecasts often perform better compared to longer term ones (Chopra & Meindl, 2016). The eight-month forecast horizon was justified based on the wishes of the company, as well as allowing us to compare whether the seasonal fluctuations are well reflected on to the future by the model. However, as the method is now developed and tested, generating a forecast with new observations more often (i.e. quarterly) should be preferred. This would also better allow the testing set to be as long as the forecast horizon. Additionally, as we have seen from Figure 4.7 in Section 4.7, the error of the forecast fluctuates weekly, further highlighting the possible disadvantage of a longer horizon.

Second, we want to discuss improvement opportunities regarding the automation of inventory management. These are considered as longer term recommendations that can further help to improve the forecasting and ordering decisions. The company could benefit from Enterprise Resource Planning (ERP) that would utilize all information live to make replenishment decisions. This would shorten the forecast horizon, incorporating all available information, and this way increase its precision. Currently, the drivers manually calculate inventory levels three times a week, and the values are matched with the order quantities and sales. Additionally, the number of driving shifts per day is determined manually. The ERP system can help to automate stock count when paired with a barcode beeper or such, and the forecast could be connected to estimate the need of drivers to be hired per day. This could decrease costs by reducing shift lengths, however, require initial investments.

A more automated system would give the company more possibility to implement a continuous review ordering policy instead of the current periodic approach. A periodic review policy is often used when there is no sophisticated computer control (Silver et al., 2017). Periodic review means the stock and demand forecast is inspected in equal time intervals to make a replenishment order. In continuous review, a new order is placed every time the inventory drops below a predetermined level. This would require negotiation from the suppliers for a different contract but could reduce costs as it allows for optimization of reordering by, for example, ordering less frequently on consecutive weeks with lower demand. Other benefits of a continuous review policy include the need for a lower safety stock, as the safety stock would only need to provide protection for the lead time instead of L+R. A periodic review policy may be easier for a company to manage, however, often also results in higher holding costs than when continuous review is applied.

Other ways that help to reduce the level of required safety stock (without reducing the CSL) include reducing the lead time from the supplier and reducing demand uncertainty (Chopra, 2019). Reducing the lead time requires negotiation from the supplier, as it requires more effort from their side. Furthermore, the source explains that reducing demand uncertainty can be done by utilizing more sophisticated forecasting methods as well as gathering more detailed data on independent variables affecting the demand.

The automation opportunities become more important as the company grows and thus, are considered longer term opportunities. Of course, some investment is required for using an ERP system as well as the implementation, including automating all the suggestions in an optimal manner. For further research, a cost benefit analysis can be made to determine profitability.

6.2.2 Limitations

Further research can be done to lift some of the assumptions made in this research that can limit the models. In this section, we discuss what assumptions were made, why, and how this can be improved upon in the future.

For forecasting methods, we emphasized the benefits of simplicity over complexity to ensure that the model can be smoothly utilized in future forecasting. Machine learning techniques have been developing at a fast pace and are becoming more and more common for forecasting due to their improved ability to handle complex and big data sets. We tested two machine learning methods: regression and random forest. We concluded that the random forest with semi-supervised learning provided nearly similar results to SARIMAX, and can also be more accurate on a longer horizon as more data is used, however, significantly adds complexity. Investigating additional machine learning techniques, such as Long Short-Term Memory (LSTM) networks, gives an opportunity to further increase accuracy of the forecast. Furthermore, a selection of time series analysis methods was made to limit the scope and thus, gives a possibility to research other commonly used models, such as Prophet.

An assumption was made that the independent variables Exams and Events used in multiple regression, SARIMAX, and random forest occurred in the same weeks in the year before. The assumption was made for simplicity, as the company has tracked the occurrences only for a year from September 2023 to August 2024. The assumption reflects reality for the majority of the events impacting all delivery cities. For the exam periods and smaller events, as well as the common events the assumption was not met for, further research can be done to determine and utilize the exact timings. This can improve the accuracy of the forecast and give more information about the relationship between the demand and the independent variables.

The aggregated monthly datasets used for determining standard deviations for the safety stock were assumed to each follow a normal distribution with an individual mean and variance. Separately determining the distributions of the subsets may help the precision of the model, although adds complexity. This assumption was made for simplicity and due to the full data set following a normal distribution. To gain information on the distribution of each monthly data set, statistical methods and tests (also used in this research), such as histograms, Q-Q plots, and the Shapiro-Wilk test, can be utilized to test for normality. In case a subset does not follow a normal distribution, additional goodness-of-fit tests can be performed to determine the correct distribution. Furthermore, when more data is gathered, the data generated with other forecasting methods should not be used for determining the standard deviations for the safety stock calculations. The historical observations can be aggregated monthly to still ensure a large enough sample.

The lead time was assumed constant in this research for simplicity and because it closely reflects the current situation. However, if orders are not always placed at the same point, the determined order quantity becomes less accurate. Thus, orders should be aimed to be placed on Friday morning each week or arrangements made to switch to a continuous review policy, as explained in Section 6.2.1. Furthermore regarding the ordering policy, we assumed that the company always places orders once per week. This is true most weeks, however, in some cases a second replenishment order is utilized, if communicated with the supplier in advance or in emergency situations to meet demand. With the optimized ordering policy, we aim to reduce the need for the use of the emergency solutions.

In conclusion, further research can help to increase precision of the models and in that way possibly improve accuracy. The assumptions were made for simplicity and as they often closely reflected the reality but can be seen as limitations of the research.

References

Chambers, J. C., Mullick, S. K., & Smith, D. D. (1971, July). How to Choose the Right Forecasting Technique. *Harvard Business Review*.

Chopra, S. (2019). *Supply Chain Management : Strategy, Planning, and Operation* (7th ed.). Pearson.

Chopra, S., & Meindl, P. (2016). *Supply chain management : Strategy, Planning, and Operation* (6th ed.). Pearson.

Cooper, D. R., & Schindler, P. S. (2014). *Business Research Methods* (12th ed.). McGraw-Hill/Irwin.

Da Veiga, C. P., Da Veiga, C. R. P., Catapan, A., Tortato, U., & Da Silva, W. V. (2014). Demand forecasting in food retail: A comparison between the Holt-Winters and ARIMA models. *WSEAS Transactions on Business and Economics*, *11*(1), 608–614.

F., F., Dachyar, M., Taurina, Z., & Qaradhawi, Y. (2021). Disclosing fast moving consumer goods demand forecasting predictor using multi linear regression. *Engineering and Applied Science Research*, *48*(5), 627–636.

Heerkens, H., & van Winden, A. (2017). *Solving Managerial Problems Systematically*. Noordhoff Uitgevers.

Hotz, N. (2023, January 19). *What is CRISP DM?* Data Science Project Management. https://www.datascience-pm.com/crisp-dm-2/

Huang, C., & Petukhina, A. (2022). *Applied Time Series Analysis and Forecasting with Python*. Springer.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.

Ivanov, D., Tsipoulanidis, A., & Schönberger, J. (2019). *Global Supply Chain and Operations Management : A Decision-Oriented Introduction to the Creation of Value* (pp. 319–333). Cham Springer International Publishing.

Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M. O., Atak Bulbul, B., & Ekmis, M. A. (2019). An Improved Demand Forecasting Model Using Deep Learning Approach and Proposed Decision Integration Strategy for Supply Chain. *Hindawi Complexity*, *2019*. https://doi.org/10.1155/2019/9067367

Kim, J.-A., & Jeong, J. (2018). Simulation Evaluation for Efficient Inventory Management Based on Demand Forecast. In O. Gervasi, B. Murgante, S. Misra, E. Stankova, C. M. Torre, A. M. A. C.

Rocha, D. Taniar, B. O. Apduhan, E. Tarantino, & Y. Ryu (Eds.), *Computational Science and Its Applications – ICCSA 2018*. Springer International Publishing.

Matsumoto, M., & Komatsu, S. (2015). Demand forecasting for production planning in remanufacturing. *The International Journal of Advanced Manufacturing Technology*, *79*(1-4), 161–175. https://doi.org/10.1007/s00170-015-6787-x

Onder, I., & Wei, W. (2022). Time Series Analysis : Forecasting Tourism Demand with Time Series Analysis. *Applied Data Science in Tourism*, 467–480. https://doi.org/10.1007/978-3-030-88389-8_22

Profillidis, V. A., & Botzoris, G. N. (2019). Modeling of transport demand: Analyzing, calculating, and forecasting transport demand. *Modeling of Transport Demand*, 225–270. https://doi.org/10.1016/b978-0-12-811513-8.00006-6

Shiwakoti, R. K., Charoenlarpnopparut, C., & Chapagain, K. (2023). Time Series Analysis of Electricity Demand Forecasting Using Seasonal ARIMA and an Exponential Smoothing Model. *2023 International Conference on Power and Renewable Energy Engineering, PREE 2023*, 131–137. https://doi.org/10.1109/pree57903.2023.10370319

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2017). *Inventory and Production Management in Supply Chains*. Taylor & Francis Group, Broken Sound Parkway Nw, Suite 300, Boca Raton, Fl 7- Crc Press.

Singh, V. P., Singh, R., Paul, P. K., Bisht, D. S., & Gaur, S. (2024). *Hydrological Processes Modelling and Data Analysis* (Vol. 127). Springer Nature.

Taparia, V., Mishra, P., Gupta, N., & Kumar, D. (2023). Improved Demand Forecasting of a Retail Store Using a Hybrid Machine Learning Model. *Journal of Graphic Era University*. https://doi.org/10.13052/jgeu0975-1416.1212

Tirkeş, G., Güray, C., & Çelebi, N. (2017). Demand forecasting: a comparison between the Holt-Winters, trend analysis and decomposition models. *Tehnicki Vjesnik - Technical Gazette*, *24*, 503–509. https://doi.org/10.17559/tv-20160615204011

Winston, W. L. (2004). *Operations Research : Applications and Algorithms* (4th ed.). Brooks/Cole—Thomson Learning.

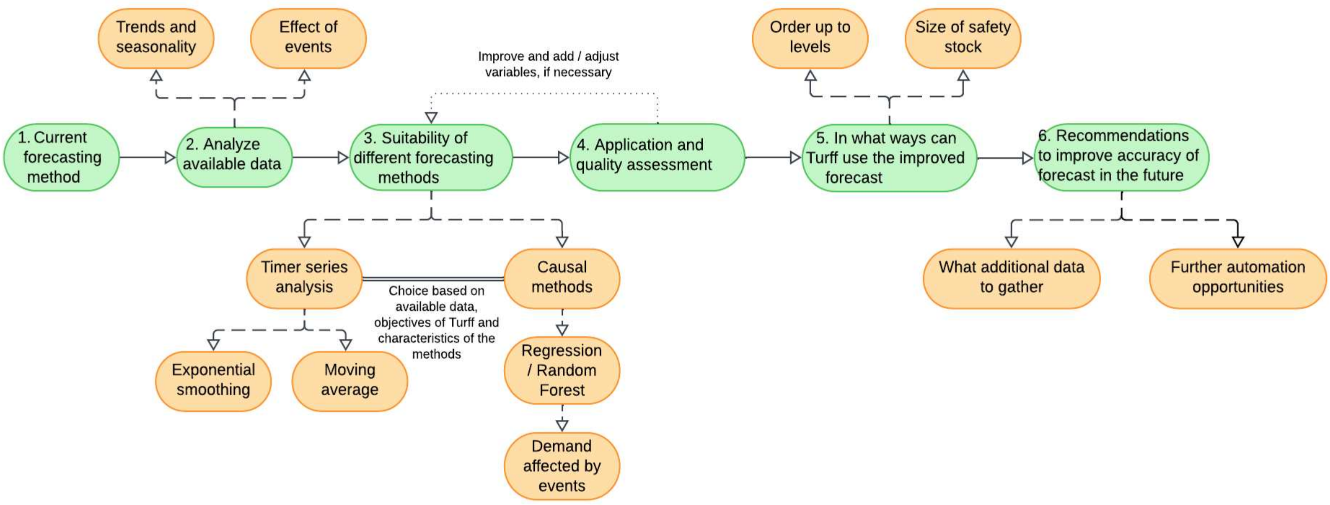
Appendices

Appendix A - Designing the Research

A1 Initial List of Problems

Initial list of problems
Number of drivers needed per day calculated manually
Inventory tracked manually, no live count
Forecasting done manually for each week, based on experience
Manual order placement
Orders canceled manually
No ERP system (or similar) implemented
Broken crates during delivery drives

A2 Visualization of Research Steps



A3 Research Design per Sub-question

Research question	CRISP-DM phase	Research type	Research population	Data gathering method	Research Strategy	Activity plan
1. What is the current forecasting method?	Business Understanding	Exploratory	Company supervisor	Conversations	Qualitative	Talk with the supply chain manager of the company
2. What historical data is available and what characteristics does it possess?	Data Understanding, Data Preparation	Exploratory	Historical demand data	Conversations, data provided from the company, literature	Quantitative	Access historical data, analyze data, identify possibilities
3. What are the most relevant forecasting methods available in literature?	Modeling	Descriptive	Literature, forecasting methods	Literature study	Qualitative	Literature review
4. How can the chosen forecasting methods be applied to the case of Turff?	Modeling	Explanatory	Literature, company supervisor, forecasting methods	Literature study, data	Qualitative, Quantitative	Data analysis, definition of objectives, method decision, application
5. How can we assess the accuracy of the chosen methods?	Evaluation	Evaluative	Literature, forecast	Literature, calculations based on findings	Quantitative	Model comparison, assessment metrics, analysis of metrics
6. How can we optimize inventory levels by developing an ordering policy for Heineken?	Evaluation	Exploratory	Company	Literature, developed forecast	Qualitative, Quantitative	Exploring various ordering policies that fit the forecasted data, choice and application based on company preference
7. How can the forecast be improved and applied in the future?	Evaluation	Descriptive	Company	Findings	Qualitative	Analysis of findings, identification of missing information, reflection on improvement options

Appendix B - KPI Formulas

We provide the formulas for the KPIs used for assessing forecast accuracy. For further information on the methods, refer to Section 1.3.6.

$$MAPE = \frac{1}{n} * \sum_{i=1}^{n} \left| \frac{Actual_i - Forecast_i}{Actual_i} \right| ,$$

where n = total number of periods

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Actual_i - Forecast_i)^2}$$

$$bias = \frac{1}{n} \sum_{i=1}^{n} (Forecast_i - Actual_i)$$

Next, we introduce the formulas for the KPIs used to evaluate the performance of the inventory policy:

$$CSL = \frac{\#\ of\ periods\ without\ stockout}{n} * 100$$

$$fill\ rate = \frac{Demand\ met\ from\ inventory}{Total\ demand} * 100$$

Appendix C - Comparison of Forecasting Methods

Red indicates that the forecasting method does not meet the criteria, yellow indicates that it meets the criteria adequately, and green indicates that the criteria is sufficiently met.

Criteria	SES	Holt's	Holt-Winters	ARIMA	SARIMA(X)	Multiple Regression	Random Forest
Medium-term	Red	Yellow	Yellow	Yellow	Green	Green	Green
Trends	Red	Green	Green	Green	Green	Yellow	Yellow
Seasonality	Red	Red	Green	Yellow	Green	Yellow	Yellow
Limited amount of historical data	Yellow	Yellow	Yellow	Yellow	Yellow	Green	Yellow
Applicable in 1-2 weeks	Green	Green	Green	Green	Green	Green	Green
Easily understandable	Green	Green	Green	Yellow	Yellow	Yellow	Red

Appendix D - Stationarity of Data for ARIMA Models

Some time series forecasting methods, such as the ARIMA models, include an assumption that the data is stationary. If a change in time does not affect the data points of the time series, as well as its mean and variance, it is considered stationary (Onder & Wei, 2022). This means that the data does not contain trend or seasonality. Based on a visual inspection, we could already identify some possibility of trend and seasonality. We use additional methods to confirm our observations. We analyze the training set specifically, since that will be first used for fitting the forecasts to assess their accuracy.

To confirm whether the training set contains trends, we fit a trend line to the data. From Figure D.1 we see a slight upwards trend.

Figure D.1 Training data with a fitted trend line (values hidden for confidentiality).

Next, we compute the Autocorrelation Function (ACF) as well as the Partial Autocorrelation Function (PACF) to further analyze stationarity. The ACF shows whether there is a correlation between the time series and its lagged values. The PACF is an extension of ACF, accounting for the influence of the intermediate lags (Singh et al., 2024). In other words, ACF shows how much the previous observations affect the current one, i.e. lag 3 tells us how much the observation 3 periods ago affects the current observation. PACF describes the same except ignoring the effect of the observations in between. The correlation at lag 0 is always 1, as it shows the correlation of the time series with itself. A large positive ACF value at a certain lag k indicates a significant correlation of the observations k time periods apart (Singh et al., 2024).

Correlogram is a graph that shows the autocorrelations plotted against the lags (Huang & Petukhina, 2022). Figure 4.4 depicts the correlogram of the training set as well as the plotted PACF. There are several rules of thumb for determining the number of lags to use, one of which states that if $n \geq 50$ then number of lags $\leq n/4$ (Box et al., 2016, as cited in Huang & Petukhina, 2022). The number of lags was chosen as 20 based on this principle. A 95% confidence interval is added to each graph with a blue shade.

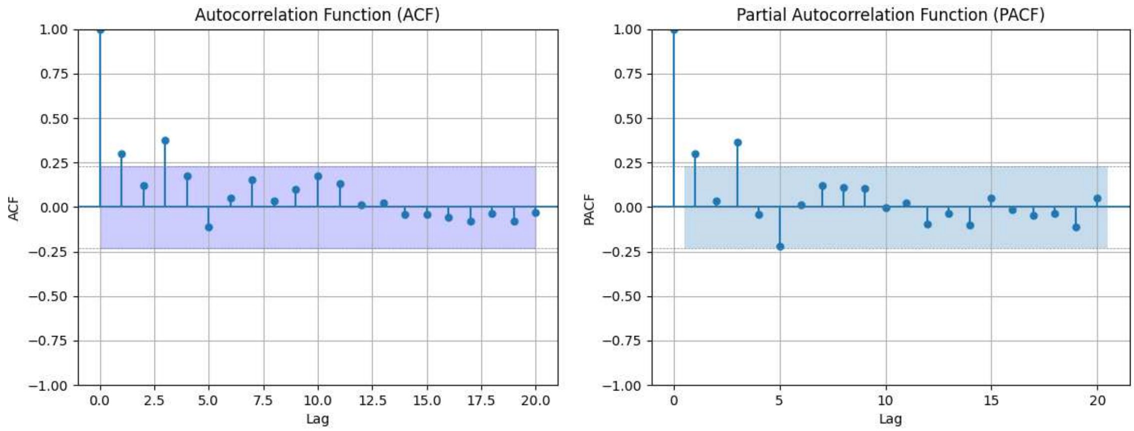


Figure D.2 The correlogram and PACF plot of the training data.

If the ACF slowly decreases as we move further from lag 0, it indicates that the time series is non-stationary (Huang & Petukhina, 2022). In general, the values of ACF for small lags are often strongly positive and slowly decreasing when the data contains trends (Hyndman & Athanasopoulos, 2021). This is because in that case, nearby observations are often also nearby in value. Furthermore, for seasonal data the values of ACF should be larger at each multiple of the seasonal period. As depicted in the correlogram in Figure D.2, the ACF rapidly decreases, opposing the indication of non-stationarity. As we are dealing with yearly event-driven seasonality, it cannot be seen directly from the ACF plot in peaks at constant intervals, but having multiple significant lags may suggest seasonal aspects.

Another sign of non-stationarity is that the ACF value of lag 1 is positive and highly significant, which can be seen to be true from the correlogram. Thus, we can conclude that the data may be non-stationary. Last, we can state that autocorrelation values that lie outside the confidence interval are considered statistically significant. They indicate correlation and non-randomness of the data.

From the PACF plot we can see that there are two significant values (outside the confidence interval). This may suggest some complex patterns of correlation in the data. We do not observe clear seasonal patterns. The ACF and PACF plots are also used later to determine parameters for ARIMA models.

Last, we conduct the Augmented Dickey-Fuller (ADF) statistical test to check whether its results indicate non-stationarity of the data. The ADF checks whether a time series is stationary based on whether a unit root is present in the data (Onder & Wei, 2022; Singh et al., 2024). The null-hypothesis of ADF is non-stationarity.

The results of the ADF test indicate that for a significance level of 0.05 we can reject the null hypothesis but for a significance level of 0.01 we fail to reject it. This means the data can be slightly non-stationary, however, significant trends or seasonality do not seem to be present. The test statistic (t-statistic) is also larger than the critical value at a 1% significance level, further indicating possible non-stationarity.

We can conclude that the training data appears to be slightly non-stationary. Due to the requirement of the ARIMA models, we difference the training data and conduct the ADF again to assess whether the differencing transformed the data into fully stationary. Differencing of data means calculating the difference between each consecutive data point. Sometimes, only one differencing does not remove the trend and/or seasonality from the data. The differenced data is depicted in Figure D.3.

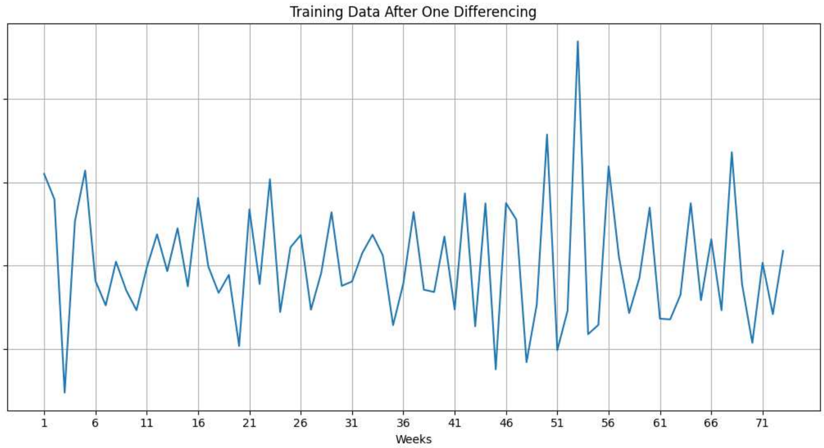


Figure D.3 Training data per week after one differencing (values hidden for confidentiality).

After a visual inspection of the differenced data, it does not seem to contain any clear trend or seasonality. To further prove our claim of stationary, we use the ADF test. We can clearly reject the null hypothesis because the obtained p-value is lower than the significance level 0.01. In addition, the test statistic (t-statistic) is smaller than the critical value at 1%, 5%, and 10% significance levels. We can conclude the data to be stationary after one differencing.

Appendix E - Ex-post Forecasting

E1 Seasonal Naive and Holt-Winters Results for Testing Set

Week / 2024	S. naive	H-W	H-W, no trend
1	1443	1858.23	1807.30
2	2220	2513.39	2416.86
3	1802	2093.01	1985.76
4	1965	2222.17	2096.79
5	2240	2412.81	2260.14
6	1842	2038.61	1885.92
7	1776	2007.82	1854.48
8	2256	2413.79	2233.06
9	2073	2215.21	2023.80
10	1929	2092.91	1898.55
11	2042	2200.61	1995.10
12	2319	2416.73	2187.53
13	2411	2442.94	2192.26
14	1874	1974.07	1733.35
15	1712	1903.22	1671.61
16	2193	2354.17	2099.62
17	1976	2162.62	1900.25

Note that the forecast values are multiplied by an anonymous factor for confidentiality of the company.

E2 Holt-Winters Smoothing Parameters

Smoothing parameters	With trend	Without trend
α (level)	0.2171	0.2643
β (trend)	0.0001	-
γ (seasonal)	0.7829	0.7357

E3 (S)ARIMA(X) Results for Testing Set

Week / 2024	ARIMA(0, 1, 1)	ARIMA(2, 1, 0)	ARIMA(3, 1, 3)	SARIMA(2, 1, 1)(1, 0, 0, 52)	SARIMAX(2, 1, 1)(1, 0, 0, 52)
1	1947.70	1797.35	1653.27	1647.48	1623.46
2	1947.70	1643.50	1800.70	2207.18	2066.94
3	1947.70	1649.80	1955.73	1852.23	1835.54
4	1947.70	1725.99	1929.62	1978.09	1962.75
5	1947.70	1680.26	1904.16	2189.40	1977.56
6	1947.70	1666.27	1988.50	1902.48	1665.59
7	1947.70	1697.76	2020.27	1855.74	1837.11
8	1947.70	1687.46	1969.53	2187.41	2139.92
9	1947.70	1676.89	1975.42	2060.95	1959.82
10	1947.70	1688.12	2023.16	1961.49	1941.69
11	1947.70	1687.34	2011.16	2040.42	1913.47
12	1947.70	1681.95	1981.98	2224.83	2205.23
13	1947.70	1685.36	2002.65	2283.87	2051.39
14	1947.70	1686.25	2021.01	1927.57	1906.20
15	1947.70	1683.99	2000.77	1820.15	1797.31
16	1947.70	1684.79	1993.26	2139.86	1933.48
17	1947.70	1685.51	2011.66	1995.29	1760.83

Note that the forecast values are multiplied by an anonymous factor for confidentiality of the company.

E4 Parameter Values for ARIMA Models

ARIMA(0, 1, 1)	Value
MA1	-0.7852
ARIMA(2, 1, 0)	
AR1	-0.5573
AR2	-0.5181
ARIMA(3, 1, 3)	
AR1	0.3566
AR2	-0.5837
AR3	0.5043
MA1	-1.0631

MA2	0.7107
MA3	-0.5686
SARIMA(2, 1, 1)(1, 0, 0, 52)	
AR1	-0.4937
AR2	-0.4848
MA1	-0.1857
SARIMAX(2, 1, 1)(1, 0, 0, 52)	
AR1	-0.5347
AR2	-0.4619
MA1	-0.1523

E5 Data of Exogenous Variables: Exams and Events

The independent variable Events is scored on a scale from 0 to 8 based on the number and significance of the events per week. 0 is given for a week if no events are happening in any of the cities or nationally, if an event is happening in 1 city the week is given a 1, continuing incrementally until 4 points. Additionally, if there is a common event, the week receives an additional score of 4 as it affects all of the four cities. The points range from 0 to 8, as there happens to only be up to two different national events happening in the same week. Table E5.1 shows an overview of the scoring and the values above 0 of the exogenous variables.

# of Cities with Events	Score
0	0
1	1
2	2
3	3
4	4
Common Events	+4

Table E5.1 Overview of the scoring of events.

Table E5.2 contains the week numbers where there are exams in any of the delivery cities, week numbers where there are significant events in any of the delivery cities, and the score indicating how wide range of events there are that week. The values are based on occurrences in the time span of September 2023 to August 2024.

Weeks Exams = 1	Weeks Events > 0	Value of Events
5	2	4
9	5	4
11	6	8
16	8	1
17	11	1
21	13	8
26	16	4

27	17	5
28	21	1
29	22	4
30	25	1
40	26	1
43	27	2
45	28	1
50	32	8
51	33	8
	34	8
	36	1
	37	3
	40	4
	45	1
	46	4
	48	4

Table E5.2 Data of exogenous variables > 0.

E6 Correlation Coefficients of SARIMAX Exogenous Variables

Exogenous Variable	Correlation Coefficient
Exams	-0.0615
Events	0.1999

E7 Further Analysis of Random Forest

We first explain the results of the used grid search of the supervised and semi-supervised random forest and move on to further analyze the results of the models with regards to the independent variables utilized.

The grid search was performed by utilizing the *GridSearchCV* function from the *sklearn.model_selection* library in Python for choosing the best values. Two to three different parameter values were tested to aim to optimize the number of trees, the maximum depth of the trees, the minimum number of samples in a leaf, and the minimum number of samples required for splitting. The grid search for both the supervised and the semi-supervised learning random forests resulted in the same values from the tested options:

- Number of trees: 100
- Max. depth of trees: None
- Min. number of samples in a leaf: 4
- Min. number of samples for splitting: 2

Figure E7.1 depicts the feature importance diagram for the supervised random forest. Compared to the feature importance of the semi-supervised model presented in Figure 4.6 in Section 4.6, we can notice significant differences. Events (remainder_x1) and Exams (remainder_x0) are more important for the supervised model. This can be caused by the supervised model not having enough data to capture the effect of each individual month in as much detail. When generating more data, the effect of the months with more fluctuation and higher demand seem to become more important, as months September, August, and July are of higher importance for the semi-supervised random forest. For both learning techniques, we see differences in the importance of each month, indicating clear seasonality of the data.

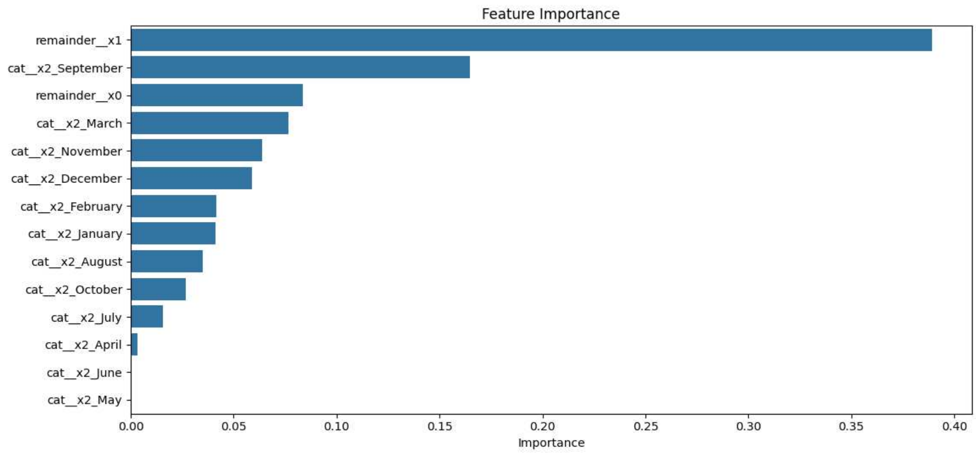


Figure E7.1 Feature importance of the independent variables used for supervised random forest.

The violin plot, depicted in Figure E7.2, indicates the distribution of the demand data of each month, giving us further information on the relationship of the demand with each month. It shows that the variability of demand, especially in August but also July and September to December, is high, while from January to June the demand is more consistent. For the semi-supervised random forest, the shapes of the violins also strongly seem to relate to the feature importance, as the months from January to June have low importance. However, the supervised model does not seem to capture this pattern as strongly, possibly due to less data.

The wider the violin for the month, the more data points land in the corresponding demand values. Thus, we can see that from about December to May the variability is low as the observations are more densely close to the mean, while June to November the fluctuation is higher. This can also clearly be observed from the graphed demand in Chapter 2.2.3 Figure 2.2. The differences of the shapes of the violins suggest seasonality, just as concluded from the feature importance values. When the violin is more spread out (especially for August) it also indicates that the demand is more difficult to predict. The white line in the interquartile range depicts the median observation of each month. It stays relatively stable throughout the months, except a clear peak in September. The violin plot of the semi-supervised random forest has similar patterns among the months, however, the violins are generally narrower and more spread out, as more data is added the variability also increases.

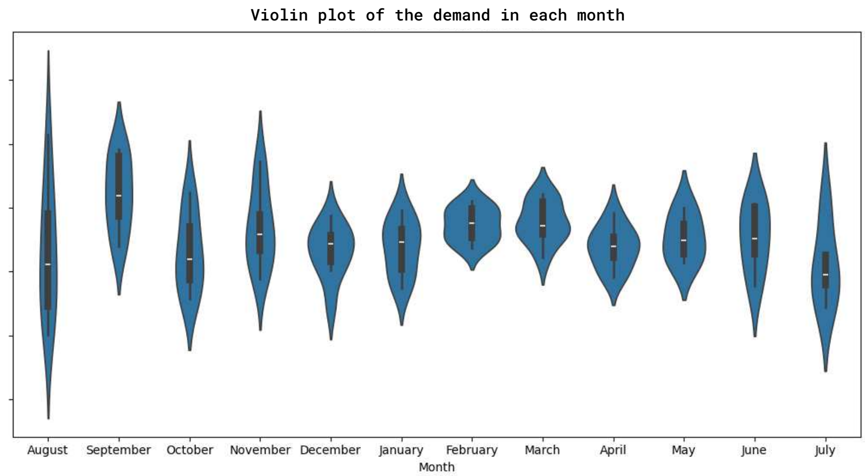


Figure E7.2 Violin plot of the supervised random forest for each month (values hidden for confidentiality).

The correlation heatmap for the supervised random forest, shown in Figure E7.3, indicates that the correlation between Exams and the demand is negative, meaning that the occurrence of exams causes the demand to go down. The correlation of demand and Events, on the other hand, is shown to have the opposite effect. Both of these relationships were similarly observed with SARIMAX. A similar correlation heatmap for the semi-supervised learning shows a value of -0.059 for demand and exams, and 0.17 for demand and events. Thus, it results in stronger relationships between the variables, possibly because there is more data to train on.

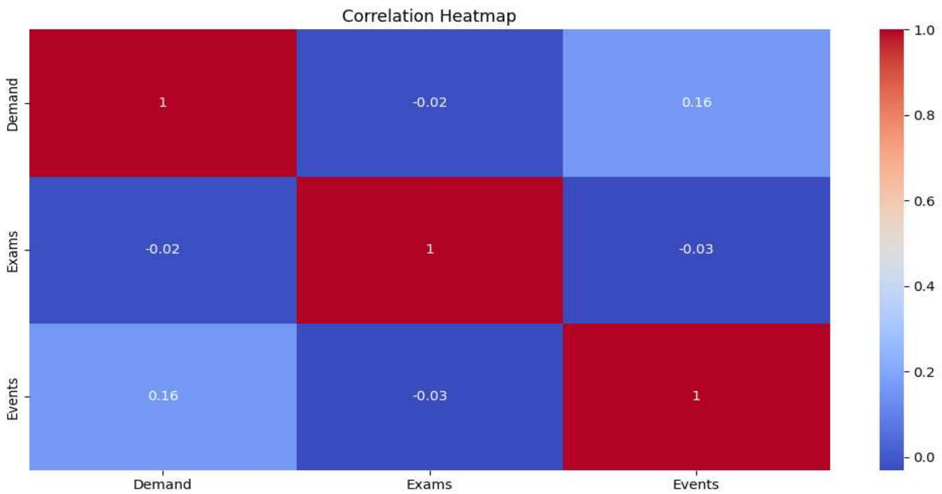


Figure E7.3 Correlation heatmap of the supervised random forest's numerical variables.

Appendix F - Ex-ante Forecast with SARIMAX(1, 0, 1)(0, 0, 1, 52)

Start Date	Forecast value
03/06/2024	2067.91
10/06/2024	1600.64
17/06/2024	1863.31
24/06/2024	2113.13
01/07/2024	1660.31
08/07/2024	1502.59
15/07/2024	2109.87
22/07/2024	1728.85
29/07/2024	1517.03
05/08/2024	2492.05
12/08/2024	1881.04
19/08/2024	1906.52
26/08/2024	2463.96
02/09/2024	2220.07
09/09/2024	1957.31
16/09/2024	2038.41

23/09/2024	2373.89
30/09/2024	2110.46
07/10/2024	1940.35
14/10/2024	1770.84
21/10/2024	2079.88
28/10/2024	1860.30
04/11/2024	1923.48
11/11/2024	1653.79
18/11/2024	2178.75
25/11/2024	2022.25
02/12/2024	1687.71
09/12/2024	1760.26
16/12/2024	1709.74
23/12/2024	1600.66
30/12/2024	1428.14
06/01/2025	1509.65
13/01/2025	1470.61
20/01/2025	1783.09
27/01/2025	1655.66
Parameter	Value
AR1	1.0000
MA1	-0.7987
Exogenous Variables	Correlation Coefficients
Exams	-0.0433
Events	0.1527

Note that the forecast values are multiplied by an anonymous factor for confidentiality of the company.

Appendix G - Ordering Policy

The week number corresponds to the **week where the replenishment order is made**. Thus, the Friday when the order is placed for the following week. Note that to obtain the final order quantity, the **current inventory level must be subtracted from the order up to level**. Finally, we **round the order up to levels up** to obtain integer values.

Week (2024-25)	Safety Stock (CSL = 96%)	Order Up to Level (S)
23	472.34	2966
24	472.34	3046
25	472.34	3398
26	518.83	3089
27	525.54	2762

28	525.54	3307
29	525.54	3163
30	855.00	3132
31	894.54	4064
32	894.54	3834
33	894.54	3621
34	894.54	4188
35	683.24	3950
36	645.07	3554
37	645.07	3533
38	645.07	3900
39	509.57	3632
40	485.66	3335
41	485.66	3099
42	485.66	3342
43	533.23	3291
44	540.10	3275
45	540.10	3030
46	540.10	3450
47	540.10	3498
48	348.45	2911
49	309.06	2814
50	309.06	2791
51	309.06	2663
52	398.37	2537
1	410.26	2563
2	410.26	2556
3	410.26	2853
4	410.26	2847

Note that the values of the ordering policy are multiplied by an anonymous factor for confidentiality of the company.

Appendix H - Use of Python Libraries

For developing the forecasts, we used several libraries in Python to help automate the calculation and better optimize parameters. First, we used the following general libraries for data analysis: *Numpy*, *Pandas*, *Scipy*, *Matplotlib*, and *Statsmodels*. For further details on what each of the libraries are commonly used for, refer to Huang and Petukhina (2022, p. 12).

There are also specific libraries for various forecasting models. For ARIMA and its extensions we used the *auto_arima* function from the *pmdarima* library. The function was used to optimize parameters of the models. However, we encountered a notification stating that Maximum Likelihood Estimation failed to converge, for models with higher numbers of autoregressive and moving average parameters. MLE is used to optimize the values of these parameters. After analyzing the notification, we conclude that it is most likely due to the limited amount of data (of less than two seasonal cycles). The issue is further addressed by analyzing the accuracy of the resulting model as well as whether there is correlation in its residuals. We find that the parameters return precise results regardless. In addition, the ex-ante forecast's fluctuations are compared to that of previous years', and we can see that even for a medium term horizon it performs quite robustly and accurately portrays past seasonality and fluctuations.

Furthermore, as suggested by Huang and Petukhina (2022, p. 265), we utilized the *SARIMAX* function from the *statsmodels.tsa.statespace.sarimax*. This was used to fit the optimal model and develop the forecast. For Holt-Winters, we used the *ExponentialSmoothing* function from the *statsmodels.tsa.holtwinters* library in Python, also suggested by Huang and Petukhina (2022, p. 67). For automated initialization of parameters the function requires two full seasonal periods worth of data, which is why this had to be done manually.

For random forest we used several additional libraries, from which we will mention the main ones. To adjust categorical variables and standardize data, we worked with the *OneHotEncoder* and *StandardScaler* functions from the *sklearn.preprocessing* library. To perform the grid search we used the *GridSearchCV* function from the *sklearn.model_selection* library. Last, we utilized the *RandomForestRegressor* from the *sklearn.ensemble* library.

Appendix I - Use of AI

The main used tools that (may) utilize Artificial Intelligence (AI):

- ChatGPT OpenAI
- MyBib Reference Manager
- Google Docs

AI in the form of OpenAI (ChatGPT) is used in this work for minor assistance for programming in Python, understanding concepts, and improving the writing and readability of small parts of the text. The used MyBib reference manager may use AI for generating references, and it was used for the purpose of easier and more accurate tracking and citing of sources. The spell check and writing suggestions on Google Docs were used to eliminate spelling errors that can affect and disrupt the readability of the report.

All AI based content was carefully reviewed and edited as needed. We take responsibility for the content of the work.