Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

Simen Emil Eide Næss

Potential of Machine Learning in Demand Forecasting Based on Point of Sales Data for Food Producers

Master's thesis in Engineering & ICT
Supervisor: Anita Romsdal
June 2023

NTNU
Norwegian University of
Science and Technology

Simen Emil Eide Næss

Potential of Machine Learning in Demand Forecasting Based on Point of Sales Data for Food Producers

Master’s thesis in Engineering & ICT
Supervisor: Anita Romsdal
June 2023

Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

NTNU
Norwegian University of
Science and Technology

Abstract

This master's thesis investigates the "Potential of Machine Learning in Demand Forecasting Based on Point of Sales (POS) Data for Food Producers." The central goal was to evaluate how POS data can be efficiently utilized by food producers to improve demand forecasting.

The research involved an exhaustive comparison of various forecasting techniques, including traditional methods, machine learning (ML) models, and hybrid or ensemble approaches. Using a robust dataset of historical POS data, the study scrutinized the efficacy of these methods in generating either insightful or practically useful demand predictions for food producers.

Interestingly, while ML models did not exhibit superior performance compared to traditional methods in this instance, they revealed significant potential features. When appropriately implemented, these features could lead to enhanced forecast accuracy and facilitate the uncovering of intricate patterns in the data. Such insights could assist food producers in strategic decision-making related to inventory management and supply chain planning.

However, the research also highlighted several challenges that need further exploration. These include managing outliers, integrating unforeseen external factors, and incorporating ML models effectively into existing supply chain infrastructures.

In conclusion, the study underscores that, while ML models require further refinement for practical application in demand forecasting, they hold promise for improving forecast efficacy when harnessed appropriately. This conclusion sets the stage for further research in the field of ML-aided demand forecasting within the food production industry.

Sammendrag

Denne masteroppgaven undersøker "Potensialet for maskinlæring i etterspørselsprognoser basert på Point of Sales (POS) data for matprodusenter." Hovedmålet var å evaluere hvordan POS-data kan bli effektivt utnyttet av matprodusenter for å forbedre etterspørselsprognoser.

Forskningen involverte en grundig sammenligning av forskjellige prognoseteknikker, inkludert tradisjonelle metoder, maskinlæringsmodeller (ML), og hybrid- eller ensemblemetoder. Ved å bruke en robust datasett av historiske POS-data, studerte forskningen effektiviteten av disse metodene i å generere enten innsiktsfulle eller praktisk nyttige etterspørselsprediksjoner for matprodusenter.

Interessant nok, selv om ML-modeller ikke viste overlegen ytelse sammenlignet med tradisjonelle metoder i dette tilfellet, avslørte de betydelig potensielle egenskaper. Når disse egenskapene implementeres på riktig måte, kan de føre til forbedret prognosenøyaktighet og lette avdekkingen av intrikate mønstre i dataene. Slike innsikter kan hjelpe matprodusenter i strategiske beslutninger relatert til lagerstyring og forsyningskjedeplanlegging.

Imidlertid fremhevet forskningen også flere utfordringer som krever ytterligere utforskning. Disse inkluderer håndtering av outliers, integrering av uforutsette eksterne faktorer, og effektiv integrering av ML-modeller i eksisterende forsyningskjedeinfrastrukturer.

Til slutt understreker studien at, mens ML-modeller krever videre raffinering for praktisk bruk i etterspørselsprognoser, holder de løfte om å forbedre prognoseeffektiviteten når de blir utnyttet på riktig måte. Denne konklusjonen legger grunnlaget for videre forskning innen feltet for ML-støttet etterspørselsprognoser innen matproduksjonsindustrien.

Preface

This master's thesis was written during the final semester of the 5-year Engineering and ICT master's program, with a specialization in Production Management, at the Norwegian University of Science and Technology (NTNU) in Trondheim.

I would like to extend my gratitude to my supervisor, Anita Romsdal, for her guidance over several semesters, her valuable feedback, and assistance in finding new perspectives to consider in my research. My good friend and researcher at NTNU, Erlend Torje Berg Lundby, generously dedicated his spare time to provide his expertise and experience, which reflects his strong commitment to machine learning and his kind nature.

I would also like to express my thanks to the companies involved in this research, especially Brynhild, for their contribution to a very interesting project and for providing great insights throughout this and their previous projects.

Last but not least, I wish to express my sincere gratitude to my boss for his understanding and provision of time off, which allowed me to complete most of the thesis alongside a full-time job. I have been fortunate enough to work on projects that have enhanced my skills in relevant topics, all of which have had a substantial impact on my thesis.

Abbreviations

AI Artificial Intelligence

ANN Artificial Neural Network

ARIMA Auto Regressive Integrated Moving Average

BOM Bill of Material

CODP Customer Order Decoupling Point

ES Exponential Smoothing

FMCG Fast Moving Consumer Goods

LSTM Long Short Term Memory network

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

ML Machine Learning

MLP Multi-Layered Perceptron

MPS Master Production Schedule

MRP Material Requirements Planning

MTS Make-To-Stock

MTO Make-To-Order

NG NorgesGruppen

POS Point of Sales

PPC Production Planning and Control

RMSE Root Mean Squared Error

RNN Recurrent Neural Networks

SARIMAX Seasonal Auto Regressive Integrated Moving Average with eXogenous regressors model

WA Weighted Average

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background & Motivation

The grocery industry business landscape is highly competitive. The food supply chain's negotiation power is often dominated by wholesalers, leading to pressure on food suppliers' margins. Simultaneously, the decision to purchase food, confectionery, chocolate, or nuts is complex and influenced by a multitude of factors that further challenge a supplier's ability to anticipate demand.

Understanding and predicting what to sell and when becomes critical for suppliers to optimize their operations and minimize the negative effects of demand overestimations and underestimations. Misjudgments can lead to increased costs, waste, and customer dissatisfaction. This is particularly prevalent in the context of food waste in private homes, which often arises from the influence of wholesalers on consumers' purchasing habits. In fact, the retail sector, although only contributing to 5% of the total food waste in the supply chain, is indirectly responsible for a significant portion of the food waste in homes. This is due to frequent promotions and marketing strategies that prompt consumers to purchase beyond their actual needs.

The complexity of factors influencing food purchasing behaviors makes establishing a direct causality linking sales to these factors at a specific time challenging. Hence, historic data analysis is used to estimate future sales and identify trends. Traditional forecasting methods such as moving averages are often employed but usually consider only a few features.

However, the landscape of data analysis and forecasting is rapidly evolving with advancements in computational power and the emergence of machine learning capabilities. Machine learning, particularly deep learning, offers the ability to process and draw insights from large datasets, potentially improving traditional forecasting models. Nevertheless, the transition from information to knowledge relies on a thorough data analysis to provide insights that inform and guide the design of the forecasting model. Emphasizing preprocessing to enhance model performance is thus a vital aspect of applying machine learning in demand forecasting.

An effective forecasting model could serve as a benchmark for tackling the intricate problem of multivariate analysis and could be adopted by other players in the food supply chain. In addition to minimizing waste and enhancing operational efficiency, such a model could also provide a competitive edge in a challenging market like Brynhildgruppen's, ultimately creating a more sustainable and efficient food supply chain.

1.2 Problem Description

This thesis explores the utilization of Point of Sales (POS) data to estimate demand for Brynhild AS, a leading supplier of Fast-Moving Consumer Goods (FMCG) to Norway's largest wholesale distributors. Currently, Brynhild AS relies on historical order data for demand forecasting, a standard procedure in the industry. However, the potential of POS data for enhancing demand forecasting accuracy remains relatively untapped, particularly within the context of a food producer. At the other side of the food supply chain, POS is currently being used to forecast the aggregated demand for wholesalers. POS- data describes transactions of items for each store, and gives an in-depth understanding of how consumers buy their food. The information in this data can be mapped and utilized to gain insight in the ever-changing demands of consumers.

With the advent of Big Data and Machine Learning (ML), there is a myriad of forecasting methods available. Each comes with its unique set of advantages and disadvantages, from traditional statistical methods to more complex ML techniques. For instance, statistical methods are known for their interpretability and robustness but may fall short when dealing with non-linear and complex data patterns. On the other hand, ML methods can capture complex patterns and dependencies, but they require larger datasets, extensive prepossessing of the training data and may be more difficult to interpret.

This thesis will probe various multivariate forecasting techniques, comparing machine learning and statistical regression methods in their different applications. Some of these techniques utilize only historical sales data, while others incorporate additional features.

The primary objective is to understand how a food supplier, like Brynhild AS, can utilize these techniques to enhance the accuracy of demand forecasts based on information shared up stream in the supply chain, and in this case, POS data. Such precision in forecasting could enable Brynhild AS to align its production more closely with demand, reducing costs and waste resulting from overproduction. The research will assess the forecasting techniques' accuracy, complexity of implementation, data requirements, and interpretability, aiming to identify the most suitable method for Brynhild AS.

Ultimately, this thesis seeks to narrow the gap between POS data potential and its actual utilization in demand forecasting for food producers. The findings could offer valuable insights for the food industry, contributing to the development of more efficient and sustainable supply chain strategies.

1.3 Research Objectives & Questions

1.3.1 Objective:

The purpose of this Master's thesis is to investigate the potential of utilizing Point of Sales (POS) data to estimate demand for food production and compares various forecasting methods to develop an accurate and reliable forecast. To accomplish this, three research questions are formulated to guide the study:

Question 1: How can POS data be effectively used to estimate demand for food production?

Question 2: What are the most accurate and efficient forecasting methods for estimating food demand based on POS data?

Question 3: How do different forecasting methods impact the accuracy and reliability of demand estimates for food production?

To address the first research question, the study explores the what POS is, which potential value it presents and if there is reason to belive that there is a correlation between POS and food demand, hereby analyzing sales trends, seasonality, and other factors affecting demand.

The second research question is addressed through a comparative analysis of various forecasting methods, highlighting the role of machine learning and its potential. The comparative analysis includes time-series analysis, moving averages, exponential smoothing, and machine learning algorithms such as Holt-Winters, SARIMAX, MLP and LSTM. The accuracy and efficiency of each method are assessed using established metrics like Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE).

For the third research question, the study evaluates the impact of different forecasting methods on the overall accuracy and reliability of the demand estimates. This is achieved by analyzing the forecast error, and assessing the robustness of each method under varying market conditions and the practical implications of investing in a forecast model.

1.4 Research Scope

This research aims to explore innovative forecasting methods that can enhance forecasting accuracy and thereby aid production scheduling for a Norwegian food producer within the FMCG market. The study will focus on comparing different methods for forecast performance of retails' demand estimates to evalute the potential benefits of utilizing point-of-sale (POS) data instead of order history between producers and wholesalers. Although the primary goal of this research is to benefit Brynild, if the methods are proven effective, the findings could inspire other food producers to adopt similar approaches. Additionally, this study will investigate how the use of POS-data can add value to various stakeholders in the supply value chain beyond the producer.

2 Methodology

This thesis was developed in two distinct phases. Most of the work was done in the first phase, which lasted for 20 weeks. During this time, a comprehensive literature review was done, and a detailed case study of Brynild was undertaken. This phase blended the findings from the literature review with a hands-on programming case, which helped to confirm the theories used in the specific case.

The second phase was shorter and took place in Spring 2023. This part was focused on finishing up the necessary documentation. It also required going back to the topics from the earlier literature review to make sure all the information used was still up-to-date and in line with the latest research in the field.

In this second phase, some new insights came up. These were partly a result of the experience gained from other machine learning projects in more business-focused environments but attributed to the understanding of the cost of implementation and the sparsity of expertise on the subject. These fresh insights, gained over time and from different situations, were added into the final documentation. This ensured a thorough and reliable study of the subjects involved.

2.1 Literature study

The first part of this master's thesis was dedicated to a thorough exploration of academic literature, establishing the foundational structure for the case study. The focused study was designed to elucidate the obstacles encountered in demand forecasting by a food producer within the Norwegian food supply chain. In-depth examination was conducted on production management, specifically production planning, resulting in the identification of challenges and potential solutions such as supply chain integration. Anticipating the availability of Point-of-Sale (POS) data for the subsequent case study, deeper investigation was made into the application of POS data in higher echelons of the supply chain.

The investigation extended to various forecasting methods in current use, scrutinizing their strengths and weaknesses. This scrutiny encompassed comparisons between the theoretical constructs underpinning time series forecasting as employed in production management and those used in deep learning. The comparison offered valuable insights into how these distinct approaches address demand forecasting.

In order to visualize how state-of-the-art forecasting methods could be implemented in the food production industry, a search was conducted for the latest research on the utilization of deep learning methods across various industries. Significant contributions were found in sectors such as wind power and energy production, noted for their innovative application of technology and analytical methods.

The main goal of the literature study was finding which new prominent methods that address some of the problems with the methods that are widely used today, and should therefore be tested for demand forecasting on a food producer. This review helped direct the focus towards relevant methods and limitations to be analyzed in the case study comprising the second part of this project.

The search for relevant literature utilized keywords listed in Table **??**, using Scopus and Google Scholar as the primary search tools. The search process began with these keywords, and gradually shifted attention towards references found in the early high-relevance articles. The reliability of these articles was assessed by considering factors such as the number of citations, the reputation of the institutions involved, the track record of the authors, and of course, through critical evaluation and pre-existing knowledge on the subjects. To delve deeply into the theory of neural networks, frequent reference was made to the MIT textbook, Deep Learning (Goodfellow et al. 2016), authored by some of the top minds in AI.

Topic	**Keywords (combined with topic)**
Supply Chain	bullwhip, food production, integration, management + information, mts, neural network demand forecasting, demand + perishable
Forecasting	model selection, multivariate + demand, point of sales, comparative + holt winters, deep learning, hybrid, ensemble, global
Deep learning	multi-step, time series, ensamble, LSTM, curse of dimentionality

Table 1: Keywords used in literature study

Additionally, four master's theses in the same field of study, all undertaken for the same company featured in the case study (Olsen 2022, Khakpour 2020, Mejlholm P 2019 and Nilsen Lie 2019), were scrutinized. These theses provided insights into previous work on the problem, guided towards new literature, and flagged potential 'future work' suggestions based on the challenges encountered, but were not directly referenced as sole sources for information.

2.2 Case study

The second part of the project comprised a case study centered on a Norwegian food producer, Brynild. Brynild's existing forecasting methods were found to be less than ideal, mirroring the situation of many food producers as discovered through the literature review. The company's forecasting relies primarily on historical sales data to project anticipated future demand. Recently, Brynild obtained Point of Sales (POS) data from one of their clients, encompassing all sales transactions over a period of five years for all their stores throughout Norway. The objective was to determine whether Brynild could enhance their demand forecasts by leveraging cutting-edge methods equipped with this new data source. Both qualitative and quantitative assessments were conducted, contrasting the limitations of Brynild's current forecasting technique with newer methods grounded in deep learning or the combination of both. A blend of diverse forecasting methods was proposed as an alternative to Brynild's conventional approach, with the goal to maximize the potential of the newly acquired data.

In the course of the case study, interviews were carried out with several key figures at Brynild, including the supply chain director, the forecast manager, and a project manager. These discussions facilitated a more comprehensive understanding of the challenges Brynild faces. The interviews with the different roles at Brynild focused on giving an understading of how production planning was done and what limitations that exist with their cooperation with vendors, suppliers and distributors. It was important to hear from each of the interviewed personnel how their relationship and understanding of which information they used to make decisions regarding production. During the fall of 2021, Brynild gave a tour at their factory in Fredrikstad, which included a introduction to both the company, two interviews and a conversation with their production manager to make sure that there was a technical understanding of what they wanted to be done, as well as guided tour of their production lines.

To gain insight on potential strategies to tackle their forecasting issues, Relex Solutions, a front-runner in retail optimization software, graciously provided a technical demonstration of their demand forecasting platform and participated in a follow-up interview. Similarly, Frivind AS, another software company known for generating retail demand forecasts and specializing in enhancing historical sales data with weather conditions to yield more accurate forecasts, kindly provided some answers to relevant questions. Their insights have also helped in identifying promising methodologies for application.

3 Empirical Background of the Norwegian Food Industry

This chapter will identify and present some of the general characteristics of the Norwegian food supply chain, particularly highlighting the role of the food producer. The design of the food supply chain consists of several actors, including those particularly gaining the attention of this rapport, being producer, wholesaler, retailer, and consumer. In Norway, the wholesaler and the retailer often have the same owners. The wholesaler, therefore, sells item per item indirectly to the consumers; meanwhile, the food producers sell large batches to wholesalers. The end customer's demand is indirectly reflected in the frequency and size of orders received by the food producers.

3.1 Market Dynamics and Power Imbalance

The Norwegian grocery market is consolidated with only four main wholesalers: NorgesGruppen (with a 43.2% market share), Coop (29.3%), Rema 1000 (23.7%), and Bunnpris (3.7%) as of 2019 (Dagligvarehandelen n.d.). These wholesalers own the retail outlets, thereby making the supply chain link between wholesalers and retailers fully integrated (**WifstadK**). Certain wholesalers also serve as producers in their own supply chain, creating their own brands. Their extensive presence throughout the supply chain is dominating. Even with limited actors, the profit margins are reported to be low. The industry heavily relies on economies of scale to maintain profitability. Despite the market competition among these four wholesalers, each holds substantial bargaining power over producers. This power imbalance is evident in NorgesGruppen's reported interaction with over 1100 suppliers, both local and international, signifying fierce competition among producers.

These market dynamics significantly impact the contractual relationships that a food producer might have with these wholesalers, leading to several potential implications. While various characteristics could be mentioned, as found in the literature (Dobson and Waterson 1997, Corbett et al. 2004, Chopra and Meindl 2019), the following have been selected because they reflect Brynild's current scenario. They were mentioned or implied during the interviews conducted with Brynild's personnel. Although it's challenging to document their state, it illustrates the situation Brynild finds itself in with respect to customer relationships.

- Price Pressures: The bargaining power of wholesalers allows them to impose price reductions, leading to contracts heavily favoring them, potentially resulting in lower margins for food producers.
- Volume Commitments: Wholesalers may demand minimum volume commitments, pressurizing producers to maintain high production levels. Failure to meet these commitments can result in financial penalties or even contract termination.

- Exclusivity Clauses: Amid fierce competition, wholesalers may demand exclusivity clauses to secure a continuous supply. While this could ensure a level of security for food producers, it could also restrict their market potential and ability to diversify their customer base.
- Compliance with Retailer Requirements: Contracts may include stringent product quality, safety, packaging, and labeling requirements. Non-compliance could lead to penalties or contract termination.
- Flexible Delivery and Return Terms: Dominant wholesalers may demand flexible delivery schedules and favorable return policies. They may also require producers to manage stock levels, which could increase costs and risks for the producers.
- Data Sharing and Transparency: Contracts may involve stipulations regarding data sharing. While this could aid producers in demand forecasting and planning, it would also require them to invest in systems capable of handling such data exchanges.

In sum, the power imbalance in the Norwegian grocery market can lead to contracts skewed towards wholesalers, imposing substantial demands and risks on food producers. This necessitates effective negotiation skills and risk management strategies for producers, while fine tuning and optimizing all sub-processes in their own production to reduce costs on their side.

3.2 The chase for margins

As mentioned, competition in the food industry business and the power in negotiations held by the wholesalers, squeezes the food suppliers' margins. Today, there are conflicting interests between the different participants of the supply chain. Offering high customer service means either maintaining a high stock level or having frequent deliveries. While a food producer aims to utilize the economies of scale by producing large volumes and buffering products in stock, however the end customer does not buy in large stock, as their demand is driven by their consumption, which is fairly stable. At the same time, wholesalers strive to keep stock levels low and buy only according to their customer's demands (Strandhagen et al. 2010). The balance is difficult to maintain, as consumers have high demands when it comes to both price and service level, and act disloyally towards the wholesalers (Nielsen n.d.).

3.3 Food production characteristics

Food production exhibits unique characteristics that distinguish it from many other production industries. Among these characteristics are the limited flexibility and extended lead times brought about by the nature of the products involved. A significant challenge in the food industry is meeting the evolving demands of end consumers due to these limitations.

Raw materials and products in food production are largely perishable, losing their quality over time and consequently being subject to shelf life constraints. Once a product's quality drops below a specified level, it is deemed as food waste. This creates a delicate balance for producers as they have to navigate between production, inventory management, and timely distribution to minimize wastage.

The food industry features a high percentage of slow-moving items, but product variety is high and continues to grow, especially during promotional periods. In an effort to maximize profit and efficiency, producers often opt for large batch production to leverage the economy of scale principle. This translates to a manufacturing model that supports low variety and large volumes, coupled with extended lead times and low postponement.

In terms of cost dynamics , the potential cost of a lost sale frequently surpasses the inventory carrying cost. This cost dynamic encourages producers to manufacture towards a finished goods inventory. Furthermore, retailers typically demand high service levels, frequently placing orders that need to be delivered within a short timeframe.

The supply chain of raw materials is also subject to factors such as crop yield quality and uninterrupted shipping routes, with additional elements of seasonality, demand amplification, and economy of scale considerations coming into play. Like the food producers, suppliers face these challenges and must adapt to meet the needs of the industry.

The varying demand, exacerbated by high and increasingly frequent promotional activities, is a significant cause of the bullwhip effect in the food supply chain.

3.4 Bullwhip

The bullwhip effect is a phenomenon where order variations increase as they move up the supply chain from the retailer to the wholesaler to the manufacturer, and finally to the supplier. This variation results in inefficiencies in the supply chain, leading to increased costs and decreased service levels.

Food producers play a pivotal role in this effect. Their position in the supply chain means that they are the most affected by demand fluctuations. An increase in consumer demand often leads to retailers increasing their orders to producers. However, due to long lead times and low flexibility inherent in food production, producers may overreact to these signals and overproduce, anticipating a long-term increase in demand. This overproduction often results in excess inventory and food waste when the anticipated demand does not materialize, further contributing to the bullwhip effect.

Being the end customer, consumers pulls demand through the supply chain, buying directly from the stores owned by the wholesalers. This gives the wholesaler the advantage of direct contact with demand data. Food producers, on the other hand, only see this demand as bulk orders from the wholesalers, which leads to deviant estimations of the actual demand of the end customer. When aggregating demand for each store, supply strategies set by the wholesaler result in large orders with relatively low frequency, so-called "order batching", as compared to the constant flow of items leaving their store's shelves. The increased variability in demand amplifies as it moves upstream in the supply chain, a phenomenon referred to as the "bullwhip effect" which occurs in other supply chains as well. Previous discoveries on the topic identified four main sources to this effect, being price variations, rationing game, demand signal processing, and, as previously mentioned, order batching, as discussed in (H.L. Lee et al. 1997).

More recent research has through the use of simulations of the Beer distribution game found three factors proving to be the most statistically significant; i.e. demand forecasting updating, level of echelons, and price variations (Paik and Bagchi 2007). Demand forecasting updating suggests that the demand amplification occurs due to safety stock build-ups along the supply chain and long lead times.

3.5 Moving towards integrated supply chains

Supply chain management involves the coordination of critical flows between its actors, among them: material flow, financial flow, and information flow (Pedroso and Nakano 2009). This paper hones in on information flow, which, powered by data, allows all supply chain participants to exchange information with their immediate partners. Such an exchange can result in quicker responses, expedited delivery, and informed decision-making (Patnayakuni et al. 2006, Trapero et al. 2012). These will collectively enhance a producer ability to meet a high service level. At a macro level, as highlighted in the prior literature reference, competition has shifted from individual firms to entire supply chain networks, making information sharing a crucial driver of transformation in traditional buyer-supplier relations. To successfully derive performance benefits from these cooperative supply networks, the need for extensive information sharing across the supply chain becomes a necessity.

In the context of food producers, the supply chain demand is driven by consumer demand. The supply chain extends from a consumer purchasing food in a retail store to a wholesaler placing a bulk order from a food producer. If we disregard breakage, the overall product flow between all supply chain participants is roughly equal. Each retail transaction produces Point of Sale (POS) data, extensively used in the industry for store replenishment evaluation. This data aids in forecasting the demand for products sold.

While wholesalers commonly utilize POS data for strategic planning, it is seldom shared with manufacturers further upstream in the supply chain. The integral principle of an integrated supply chain, as conceptualized in the Industry 4.0 paradigm, is the sharing of information among different supply chain actors. This shared information is key to enabling data-driven supply chains, which utilize pull-based control principles, reducing cost levels, lead times, and stock levels (Fransoo and Wouters 2000).

As we move towards an integrated supply chain, it's critical to explore effective inventory management strategies. Vendor-Managed Inventory (VMI) and Continuous Replenishment Program (CRP) are innovative strategies that leverage shared information for better inventory control (Danese 2004; Kurtuluş et al. 2011). Under VMI, manufacturers manage retailers' inventory levels, using shared POS data and advanced forecasting to optimize replenishment. CRP, on the other hand, could utilize POS data to continually adjust inventory levels and then aligning stock with real-time demand (Raghunathan and Yeh 2001). Implementing VMI and CRP offers a significant opportunity to reduce the bullwhip effect in real-world supply chains (Disney and Towill 2003), which may enhance service levels and mitigate waste, culminating in a more sustainable food production industry.

4 Theoretical Background

This section provides a overview of theories in the fields of production planning, time series forecasting and deep learning with particular relevance to the research questions of this paper. The background theory for production planning explains the importance of forecasting. The two of them constrains the scope of time series forecasting methods utilizing statistics and deep learning that are followingly described in greater detail.

4.1 Production Planning and Control in Food Supply

Production planning and control (PPC) is a key infrastructural decision area of any production company's operations strategy and includes aspects of coordination and operation planning of a companies resources. Some of the functions of PPC include material requirements planning, demand management, capacity planning and scheduling and sequencing of jobs (Stevenson et al. 2005) and summarized as: "... the PPC system should provide an overview of the key planning and control activities and decisions within a production system and thus define the structures and information upon which managers make effective decisions." (Romsdal 2014)

In the past decades, cost-efficiency and greater speed have been the main orienting goals of Supply Chain Management, which has required PPC to adapt their strategies to minimize costs and production lead time accordingly (Romsdal 2014). As identified by Hau Lee 2004, the problem of high-speed, low-cost supply chains is that they lack the abilities to respond to unexpected changes in demand or supply. This implied that the production systems did not focus sufficiently on responsiveness. The link between responsiveness and PPC that is particularly relevant to this paper is demand anticipation. A company's ability to anticipate future demand for its products will improve its planning, making way for precisely optimizing the required resources and avoid potential lost revenues.

A common tendency in high volume industries, such as the food sector, is the production approaches make-to-order (MTO) and make-to-stock (MTS) Chetan Soman et al. 2004. These are two different planning strategies determined by the Customer Order Decoupling Point (CODP). More precisely, the CODP is the point in the value chain flow where the product is tied to a specific order from a customer; the different choices being make-to-stock, assemble-to-order, make-to-order, and engineer-to-order, illustrated in Figure 1. As a rule, the CODP coincides with the most important stock point, from where the customer order process starts (Olhager 2010). Activities that occur upstream of the CODP are typical forecast-driven; meanwhile, activities that occur downstream and include the CODP are customer-order-driven.

Companies that are classified as MTS serve their customers from a finished goods inventory. To ensure high customer service levels with short customer delivery lead times on products produced in high unit volumes with narrow product variety, MTS

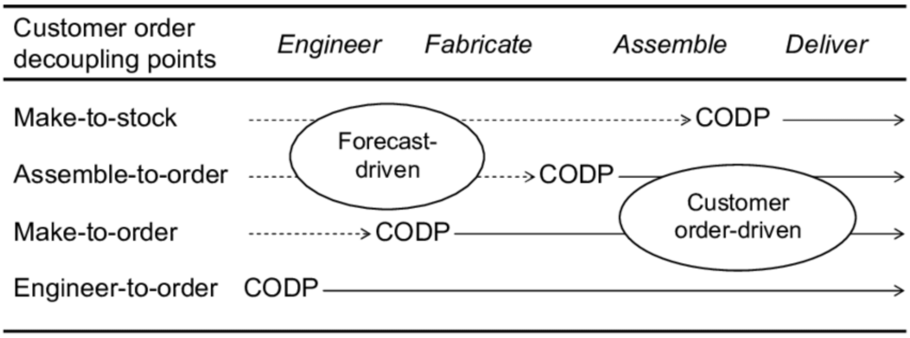

Figure 1: The four different customer order decoupling points (Olhager 2010)

in a must for many companies in the food sector (Romsdal 2014). Keeping finished goods in stock provides a buffer against variations in demand and simultaneously enables stable production levels, which have a high cost-efficiency per product (Vollmann 2005). How much to keep in stock is a trade-off between customer service level and holding costs. This balance is crucial for the company to be economically proficient. Food producers are particularly sensitive to this issue as their holding costs also will have to include that perishable products lose quality over time, or in the worst case, become obsolete (Romsdal 2014).

As for most products, they are composed of several components. For a food producer these components are ingredients. How much ingredients is needed to produce a specific amount of products are decided by the Bill of Material (BOM). Material Requirements Planning (MRP) is used by producers that follow the MTS principles as the most common requirement-initiated planning method (Romsdal 2014). To ensure that ingredients are available for production and products are available for delivery in time, MRP takes into account the demand for a product, its BOM and inventory records. The plan is based on backward scheduling, which means that jobs are started as late as possible (Brandon-Jones et al. 2013). The input to the MRP is controlled by a Master Production Schedule (MPS) that summarises the volume and timing of the demand of end products based on customer order, forecasted demand, capacity constraints on production lines and inventory levels. As for MTS production, forecasts and expectations of future demand are the main inputs to MPS (Zijm 2000).

As a result, MTS companies' production relies heavily on anticipating demand, i.e., demand forecasts, and planning to meet this demand (C. Soman 2005).

4.2 Forecasting

Suppose we have a sequence of observations from a time series from a past state until present (t), $[x_1, x_2, x_3, ..., x_t]$ and wish to speculate in how the future observation (x_{t+1}) will look like. A forecasting method will have to be used. This section will categorize different types of methods and explain the basic principles of how to predict (x_{t+1}), and potentially (x_{t+2}), (x_{t+3}), etc., by forecasting time series.

4.2.1 Fundamentals of forecasting techniques

Forecasting methods can broadly be categorized into three types: judgemental, univariate, or multivariate. Judgemental forecasts, as the name implies, hinge on human judgment, drawing on experience, expertise, or intuition. On the other hand, univariate and multivariate methods are quantitative techniques that use past and current time series observations to generate predictions.

Univariate methods use a single time series to make predictions, while multivariate methods employ additional explanatory variables to forecast future outcomes (C. Chatfield 2001). It's important to distinguish these methods from the quantitative models used in forecasting literature, which are often classified as either causal or non-causal (Cochran et al. 2011).

Here, the term "models" refers to those built using either univariate or multivariate "methods". Causal models leverage cause-and-effect relationships to explain outcomes driven by influential factors. These factors, often known as "features", are additional explanatory variables used in conjunction with the time series observed, thus deploying multivariate methods. Conversely, non-causal models generate forecasts by identifying systematic patterns in time series data, making use of univariate methods.

Implementing univariate or multivariate forecasting methods requires careful consideration of a few fundamental steps, shown in Figure 2 and originally described by Makridakis, Wheelwright et al. 1984. This framework will be used as a guideline to explore the forecasting problem outlined in this report's case study.

4.2.2 Evaluating forecast methods

Evaluation of forecasting methods is done by splitting the data set into two portions, training and test data. Training data is used to estimate the parameters of the forecasting method, while test data is used to evaluate how well the trained model is able to generalize a solution and test how it performs in terms of accuracy when tested on the second partition. The results of the test *should* provide a reliable indication of how well the model is likely to forecast on new data.

When splitting the data, determining the size of the samples does not have an exact science behind it, but as a rule of thumb, when training non-linear forecasting models, at least 20% of the original data set should be held back for testing, or in other words; for an out-of-sample forecasting evaluation (Granger 1993) which it is what it's called in many of the cited books and articles of this rapport. When comparing different methods, it's important to be aware that changing the data sample and split could affect the comparison's foundation. Different data samples could include different levels of noise, scales, or sample sizes could affect should always be used to give a fair comparison.

The error of the forecast is used to evaluate its accuracy. The error ($error_t$) is calculated as the difference between the estimated value that the model have predicted

Step 1: Problem definition
Involves developing a understanding of how the forecasts will be used and who it is for.

Step 2: Gathering information
Gathering statistical data as historical data of the items of interest. The historical data is used to construct a model which can be used for forecasting.

Step 3: Preliminary (exploratory) analysis
Graph the data for visual inspection and compute some simple descriptive statistics. If more than one series of data is available and relevant, produce scatter plots of series and related descriptive statistics (e.g., correlations).

The purpose is to get a feel for the data like if there are any consistent patterns, significant trends, if seasonality is important, any evidence of the presence of business cycles, if there are any outliers and the importance of the relationships among the variables.

Step 4: Choosing and fitting models
The preliminary analysis serves to limit the search for an appropriate forecasting and hence only pursuing one or two leading contenders for subsequent analysis. When forecasting the long-term, a less formal approach is often better. This can involve identifying and extrapolating mega trends going back in time, using analogies, and constructing scenarios to consider future possibilities.

Step 5: Using and evaluating a forecasting model
The performance of a model can only be properly evaluated after the model is used to make the forecast. It is important to be aware of how each forecasting method has performed in practice in other forecasting contexts. In addition, the accuracy of future forecasts is not the only criterion for assessing the success of a forecasting assignment. A successful forecasting assignment will usually also be a stimulus to action within the organization. In general, forecasts act as new information and management must incorporate such information into its basic objective to enhance the likelihood of a favorable outcome. Implementing forecasting is often at least as important as the forecasts themselves.

Figure 2: A framework for basic analysis to be used when building a forecasting model by Makridakis, Wheelwright et al. 1984

$(\hat{Y}_t)$ and the observed value (Y_t):
$Error_t = Y_t - \hat{Y}_t$

The goal of any forecasting model would be to minimize this error of all its predictions. Many different methods are used to summarize this error so that users can evaluate and compare their performance. Two of the methods that are widely used both in forecasting literature referred to in this rapport as well as by the company presented in the case section are the following:

Root Mean Squared Error(RMSE)

$$RMSE = \sqrt{mean(error_t^2)} \tag{1}$$

Mean Absolute Percentage Error(MAPE)

$$MAPE = mean(|\%error_t|) \tag{2}$$

RMSE, (1), is commonly used to compare different methods applied to a single time series or several different time series with the same units. Popular for it's ease of implementation and interpretation. MAPE, (2), has the advantage of being unit-free, so it is ideal for comparing forecast performance between different data sets or different units. A disadvantage is that for values $Y_t = 0$ the error will be undefined or infinite, or for values of Y_t being close to zero the error will be extremely large. Using several evaluation methods could help in getting a better picture of how the model performs.

As a note on forecast modeling, it's important to keep in mind the practical aspects when diving into a theoretical domain; hence two things should be noted in regards of evaluating forecasts (Hyndman and G. Athanasopoulos 2018):

- A model could have a perfect fit to the training data but will not necessarily forecast well in real life. There are no guarantees that the past will represent similar patterns as the future.

- Over-fitting a model to data is just as bad as failing to identify a systematic pattern in the data, even though the test errors may be close to zero.

Regarding over-fitting a model, this is the consequence of a model that corresponds too closely or exactly to a particular set of data and therefore fails to generalize a solution applicable to future scenarios. There is also the issue of under-fitting. A term used for models that are not able to obtain a sufficiently low error value when describing a particular set of data (Goodfellow et al. 2016). Both cases are illustrated in Figure 3.

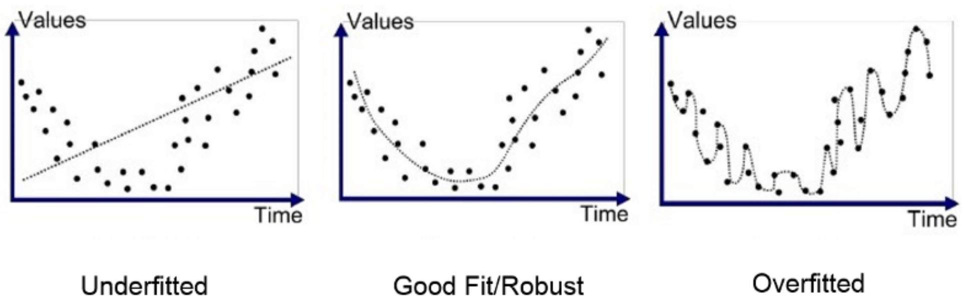

Figure 3: Illustration of over-/under-fitting a function to describe a tendency in a data set

Forecasting deals with uncertainty on many different levels. When it comes to forecasting models, this uncertainty can be categorized by the three followings (C. Chatfield 2001):
1) Uncertainty about the structure of the model
2) Uncertainty about estimates of the model parameters
3) Uncertainty about the data, including unexplained random variation in observed variables, measurements, and recording errors.

The uncertainty about the structure of the model will be the focus as we discuss different options within traditional methods of forecasting, including statistical models and modern models based on machine learning methods.

4.3 Traditional forecasting methods

This subsection will describe general characteristics of the traditional forecasting methods, namely univariate methods used in time series models, and thereby identify the advantages and disadvantages to be considered when implementing these types of methods.

4.3.1 General characteristics

Single time series are used to create univariate time series models that all have the basic underlying assumption that future patterns are similar to historical patterns. When extracting the characteristics of times series patterns, there are four properties of data that are often analyzed:

- **Trend**. Describes a tendency of growth or decline in the observed data. The trend could be either linear or non-linear; accordingly, the trend can be modeled by linear or non-linear functions.

- **Seasonality**. Describes a repeated pattern at a fixed interval, such as average air temperature.

- **Cycles**. Describes a repeated pattern, similar to seasonality, but may be of varying interval sizes. Such behavior could be found in financial times series data.

- **Randomness**. Times series data often contain both systematic patterns and random noise. The random noise creates difficulties when trying to generalize a pattern from the time series analysis, and so most times series models include a noise term to take into account the effects of randomness, such as smoothing.

A general idea of traditional forecasting methods are that they create non-causal and statistical models, and so have no other explanatory force that tendencies in previous observations. On the one hand, this makes the methods easy to implement, understand and the model easy to analyze, as well as being surprisingly effective compared to their level of simplicity Hyndman and G. Athanasopoulos 2018. On the other hand, it limits the model's ability to reflect real life and rapid response in demand changes as it would have to wait several months for previous months results to have full effect in the scenario of a shock to demand forecasting. There are well documented examples of how statistical methods fell short when Covid-19 affected sales (Herzog et al. n.d.), which demonstrates this point.

The forecasting methods are unable to show non-linear relationships, limiting their capacity to forecast exponential trends; that is, they have empirically shown to perform poorly. An important part of forecasting is estimating prediction intervals. Prediction intervals give an interval within which a future observation will lie with a specified probability (Hyndman and G. Athanasopoulos 2018). Because traditional methods are based on a strict set of rules of calculations, this interval is calculated by the standard deviation of each step forwards in time. Several simulated and empirical studies have shown that the proposed intervals made by Holt-Winters' method, one of the most established and advances traditional methods, tend to bee too narrow or too wide, in the sense that more observations than expected fall outside of the predicted interval (Koehler et al. 2001, Ord et al. 1995 and Chris Chatfield and Yar 1988).

There has been conducted several comparative studies on the many different forecasting techniques that include both trend, seasonal and cyclic effects. Some of the most widely used methods are: Winter's smoothing, Holt method, Time series regression, Decomposition and Autoregressive Integrated Moving Average (ARIMA). The findings of such studies does however not conclude that one of these methods is superior to the others. Although it is inconclusive which method outperform the others, Holt Winters' method and ARIMA are generally amongst the top performers, and therefore the most popular choices for univariate forecasting (Tirkeş et al. 2017, Markovska et al. 2016 and Veiga et al. 2014). Because of their similar performance, and because the company presented in the case section of this paper uses the same method, Holt Winters' method is later on selected as the base line model to function as a performance indicator to compared against other forecasting models.

4.3.2 Holt-Winters

Holt-Winters' method, also called triple exponential smoothing, exhibits both trend and seasonal variation, two of the most important characteristics of time series patterns. It comprises four techniques of forecasting; weighted averaging (WA), exponential smoothing (ES), double exponential smoothing, and triple exponential smoothing. Together, this creates a model where some of the noise in the data is leveled out, previous observations are weighted in regards to their relevance (time-wise), and parameters for tuning trend and seasonality (Hyndman and G. Athanasopoulos 2018). The method can describe linear trends in data and therefore do not perform well on non-linear structures. However, it there are non-linear trends in the data, it is possible to transform the input to a logarithmic scale, despite being less ideal in terms of analysis and evaluation, which will improve forecasting accuracy. The method is popular due to its low implementation costs, as it can be used as an 'out of the box'-solution. As an example of a statistical method, it produces models that are easy to analyse and understand for non-experts. Despite it's simplistic nature, it has shown to be surprisingly powerful and is widely used for *univariate*, multi-step time series forecasting (Tirkeş et al. 2017).

4.3.3 SARIMAX

SARIMAX, an acronym for Seasonal AutoRegressive Integrated Moving Average with eXogenous factors, extends the ARIMA model, incorporating a seasonal differencing term (S) and accommodating exogenous variables (X), which is used as a error component in a autoregressive model. ARIMA is an example of a autoregression model, which is a term used for models that forecast the variable of interest by using a linear combination of several past values of the variable itself (Hyndman and G. Athanasopoulos 2018). The seasonal component allows it to capture seasonal trends and patterns that are repeated over specific periods, thereby making it apt for forecasting data with pronounced seasonality.

The X in SARIMAX stands for 'exogenous' and represents the model's capacity to incorporate external variables. These exogenous factors represent independent variables that influence the forecast variable but are not influenced by it. This ability to incorporate outside factors enables SARIMAX to account for influences such as economic indicators, weather conditions, or other relevant factors that impact the data but are external to it.

As such, the multivariate nature of SARIMAX due to its capacity to incorporate exogenous variables differentiates it from univariate forecasting models. This feature, coupled with its flexibility and capacity to handle complex data structures, makes SARIMAX a compelling choice for diverse forecasting tasks, including the one at hand.

4.4 Deep learning-based forecasting methods

This sub-section will introduce key concepts that have been found relevant through a literature study in the field of machine learning in demand forecasting. The different methods will be presented in such a depth that the readers can understand some of the possibilities and challenges related to these methods, without the aspect of mathematically explaining each detail.

4.4.1 General characteristics

Within the umbrella term 'Artificial Intelligence,' the subgroup Machine Learning has proven useful in several practical forecasting tasks. In particular, deep learning and artificial neural networks' abilities within recognizing and classifying patterns have been recognized as prominent tools to predict future outcomes (Sharda 1994). The scope of ML for this rapport is limited to supervised machine learning, which in general focuses on learning how to predict the outcome of a scenario given some input about the present and previous states of different features. They are therefore trained with historical data in which both features and outcomes are described. For example, predicting whether or not someone will buy new skies the upcoming season, depending on their previous buys, their current physiology, and how much it has snowed in the preseason.

The general idea of supervised machine learning is to build a model that minimizes predictions' error by training a model based on previous states with associated outcomes. Within quantitative forecasting, forecasting methods are either based on statistics or machine learning, but they can also be combined. Multiple linear regression is based on a statistical approach but finding the optimal solution that fits the described scenario best needs to be approximated through evaluation of error. This is where machine learning can do its part and help in evaluating the accuracy of the almost endless alternatives of regression models (Kuhn and Johnson 2013).

Models of ML are not limited to linear problems, and regarding the probability that real-life generated problems often are non-linear, the use of ML in forecasting could help describe time series that include cyclic behavior and shifts between stable periods and high volatility. (C. Chatfield 2001). Actually, regardless of the problem's complexity, as long as it can be described by a function, there exists a neural network which can do the job. This universality is proven by the universality theorem, but how to construct this network is not always known, and there may not exist a technique to even do so at present time (M. A. Nielsen 2015).

The architecture of a neural network creates its most important feature: to generalize a solution by learning from experience. This is mainly the reason for why ML has been predicted a promising future within forecasting; the known relationships between the features described in the data does not need to be known a priori because the network will find them itself during training. The models "learn from experience" by using data sets of historical data, as previously mentioned, to tune the weights of the model to be adjusted to the specific task at hand. The two groups of ANNs that this study will focus on is feed-forward neural networks (also called multi-layer perceptrons (MLP)) and recurrent neural networks(RNN) as a result of their promising findings in other studies done in the field of time series forecasting, particularly when compared to statistical models such as ARIMA and Holt-Winters (Tang et al. 1991, G.Peter Zhang et al. 2001), Al-Saba and El-Amin 1999, Elkateb et al. 1998).

RNNs and MLPs are similar in how they are built. Like other neural networks, they are inspired by neuroscience, mimicking the biological structuring of a human brain, and how signals are processed back and forth between different parts of the brain. The networks consist of layers of nodes which processes an input through an activation function and generates a transformed signal that is passed to adjacent nodes, which are interconnected by links called edges. Information is feeded from the input layer, through the hidden layers, and out of the output layer. A multi-layer perceptron is illustrated in Figure 4 for easier understanding of the design of such a network. Differences between the two network's technical nuances will be explained in the following subsections.

The three types of layers are input, hidden, and output layers, where the optimal number of hidden layers may vary depending on the domain and nature of the data. The number of hidden layers gives the depth of the model, hence the name "deep

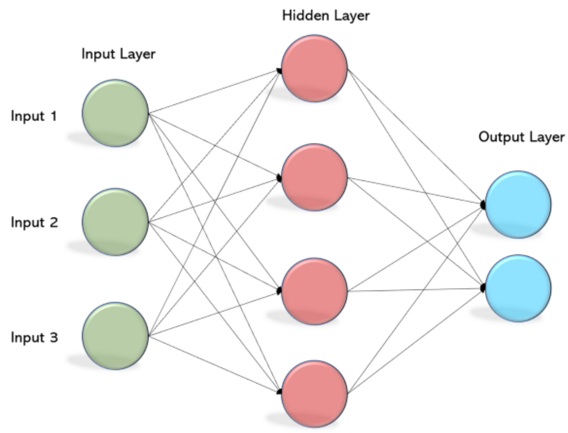

Figure 4: A conceptual illustration of a fully connected multi layer perceptron with one hidden layer, three inputs and two outputs. (Mohanty 2019)

learning", which arose from this terminology. The number of nodes in each hidden layer determines the width of the model. The width of the model enables parallel processes where some or all of the nodes of a previous layer can feed information to an adjacent node (Goodfellow et al. 2016). The adjacent node has a activation function, typically a sigmoid function (M. A. Nielsen 2015), which takes the input and transforms it with a non-linear function that it feeds forward along with its belonging weight which signalises the importance of the nodes output. The combination of many non-linear transformations across nodes and layers makes it possible to replicate highly complex non-linear relationships between the input features and describe this through a non-linear function.

4.4.2 Designing and training an ANN

Two elementary parts of building a neural network is designing and training the network. When designing a network, there is normally a manual process of tuning the hyperparameters, such as the depth of the network, learning rate, activation functions, dropout rate, and desired width of the hidden layers. These hyperparameters specify how the network is trained and the structure of the network itself (Yoo 2019, Guoqiang Zhang et al. 1998). More recent research have made efforts of tuning the hyperparameters by using automatic sequential optimizing algorithms with promising results (Bergstra et al. 2011) performed by some of the pioneers of AI, but it is constrained by available processing power.

The training of the network is done according to the hyperparameters and is an optimization process of minimizing the loss of a generalized solution by tuning parameters for each node called 'weights'. The training of the network also requires large processing powers, and can be a limiting factor in regards to developing a good model, and is one of the costs of using a neural network, as it need to continuously update it's model. The continuous input data is only a small portion of the total data sample used to train the network. Neural networks require an enormous amount of data compared to univariate time series models. Together, the relationship of the network size, data sample size, and the task at hand could lead to over-fitting, as previously discussed as a general problem in forecasting. Neural networks are particularly exposed to over-fitting, more so than statistical models, and methods of weight pruning have been developed as a potential solution when facing such problems (Weigend et al. 1991).

4.4.3 Multi Layer Perceptron

How a Multi-Layer Perceptron is built has already been explained as a concept, but will now be examined in more details to identify it's strengths and weaknesses in the aspect of forecasting. First, a disclaimer; training and tuning neural networks are not an exact science, more of an art (Similar to a quote made by Yann LeCun, one of the more acknowledged researchers in the field of AI).

Training a MLP is a procedure based on iterations of testing different versions of a network, evaluate their accuracy and error and adapt accordingly to reduce that error. Backpropagation is a learning algorithm widely used for such an optimization task (Goodfellow et al. 2016). As the first part of training, the network can be initialized with random weights. The weights, the parameters of the model, are used as input to a cost function. This cost function returns a sum of the error between each predicted outcome compared to the target outcome. By iterating over every data point in the training set, or till a benchmark accuracy is reached, the weight of each node are accordingly adjust to minimize the cost function which is fed back into the network again. This round is training is called an epoch.

The learning algorithm serves the purpose of minimizing the cost function, and update the weights accordingly for each epoch of the training (Hecht-Nielsen 1992), but as the number of epochs increase, so does the risk of over-fitting(4.2.2) and the computational costs related to training increase. To avoid this, the backpropagation algorithm introduces a way to more quickly adjust the weights, by computing the gradient of each weights contributing to the cost function, i.e. the partial derivative of each weight in regards to the cost function. The gradient of the cost function that concerns a specific set of weights describes the curvature of these particular weights in regards to the function. The sum of all curvatures for combinations of different weights and their according error makes up an error surface. By moving in the direction of where the negative derivative is the largest, the weights are updated and the function is improved. How fast to move is decided by a hyper parameter earlier introduced, namely the learning rate. A large learning rate would result in faster training but increases the chance of not finding a global minima which is where the neural network will find the set of weights with the lowest cost score, and therefore the most optimal solution to our problem.

There are no guarantees that it will find a optimal solution as several minimums could be found, as seen in figure 5. When standing in point A, there is a risk of getting a sub-optimal solution by stopping training when the local minima is reached. How to handle either saddle points or local minimums on the error surface is a part of the optimization

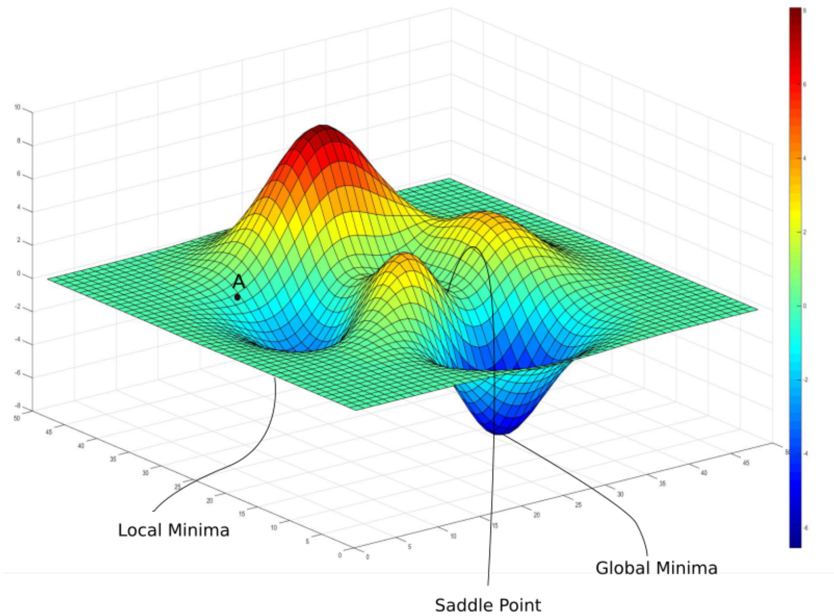

Figure 5: A 3-dimensional presentation of a cost function with only two weights. A saddle point, and local and global minima is pointed out. (Kathuria 2018)

problem, and is likely avoided by different strategies on adjusting the learning rate. Since this is in large parts an implementation problem of the network, and not of much importance to understand the basic principles of the different networks, the interested reader is referred to Phan and Hagan 2013 for an in-depth explanation on this subtopic. However, it is worth noticing that a network doesn't always find the optimal solutions, as a consequence of this problem.

An MLP is represented by a static, direct acyclic graph. This presents a problem for MLPs when it comes to utilizing sequential data, i,e. temporal data such as time series. The input layer does not allow for sequences and sees each input as an independent variable. There is, however, a possible workaround to this problem: the inputs will include a fixed number of past values, which could be chosen by a moving fixed-length window along the series.

Using a time window of a fixed size has proven to be limiting in many applications: too narrow time windows will result in leaving important information out and a too wide time window will include useless inputs that may cause distracting noise (Assaad et al. 2008). In the sense of forecasting multiple steps, each time step will have to be added as a new node in the output layer. As the output nodes are not connected, so will neither the predicted time steps be, and what has happened at time t less likely affect what happens at $(t+1)$. For this reason, MLPs are mostly used for multivariate, *one-step* time series forecasting.

MLPs does not have a feedback connection in which outputs of the model are fed back into itself. When MLPs are extended to include a feedback connection, they are called recurrent neural networks.

4.4.4 Recurrent neural network

A recurrent neural network (RNN) is very similar to a MLP, but it includes a recurring feedback from previous states of the network, and thereby specializes in treating sequential data, such as time series. An MLP would process an entire sequence at once, where it would treat each temporal instance in the sequence as a separate feature. In contrast, RNN would process sequences sequentially by iterating over each element while keeping a state in-between the elements, which can be thought of as a form of memory that remembers previously seen data. (Goodfellow et al. 2016).

The ability of sequencing enables the RNN to recognize temporal patterns and generalize across sequences (Chollet 2018). The complete network can more intuitively be thought of as consecutive copies of a network, each passing messages to a succeeding network describing important states to its successor (Olah 2015). Unlike traditional networks, the output prediction of a RNN is therefore not only based on the final input, but also on the states that are calculated based on the preceding input. The build of an RNN can be illustrated as one network with a loop to it self, but is better understood when unfolded, so that the reader gets a more intuitive perspective of how time series are handled, as shown in figure 6.

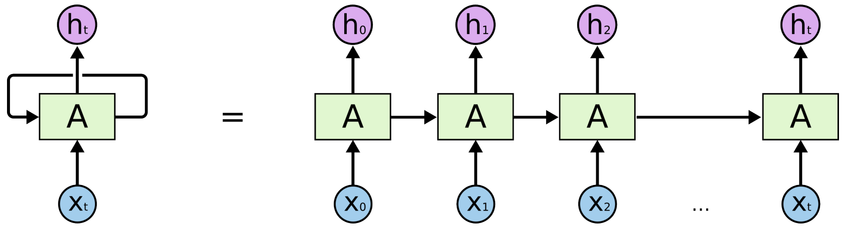

Figure 6: A unrolled Recurrent Neural Network (RNN) (Olah 2015)

In this RNN, a neural network A takes an input x_t and outputs h_t, where the different inputs and outputs are represented at their time-being, t. The unfolded network shows how it is equivalent to the repeated application of a MLP, only the previous state, h_{t-1} is updated between each iteration. For each network, the updated state will be based on calculations of the previous state for time $(t-1)$, all previous inputs of x, the input x_t for the current state and the weights of the network. This chain-like nature reveals that RNNs are intimately related to sequences and lists, and is the natural architecture of neural networks to use for such data. This makes them well suited to solve sequence-based problems such as speech recognition, language modeling, translation and time series modeling of stock prices, to mention a few.

The 'vanilla' version of RNNs are trained by applying backpropagation, just like in MLPs, but the gradient of each time step is calculated recursively, and therefore hard to train. For long-term dependencies, this results long chains of networks. In theory, this does not present a problem, but empirical use-cases of RNNs have shown that connecting information from one time step to information many steps ahead becomes increasingly difficult. The reason is known as *the vanishing gradient problem*, which is an effect that occurs when the total network becomes to deep, i.e. too many states (Chollet 2018).

Backpropagation for RNN is called Backpropagation Through Time (BPTT), and utilizes gradient-based learning, where the weights in the network are updated proportionally to the gradient of the error function, similar to backpropagation for MLP. In deep networks, such as unfolded RNNs processing long sequences, this update will become vanishing small, effectively stagnating the networks ability to learn (Goodfellow et al. 2016). For this reason, simple RNNs are almost never used as standalone models for solving real-world problems due to being too simplistic models (Chollet 2018). The vanishing gradient problem was studied in detail by Hochreiter 1997, who designed alternatives to the simple recurrent unit to help solve the problem - the Long Short-Term Memory (LSTM).

4.4.5 Long Short-Term Memory

While RNNs have a basic structure in the recurrent unit, shown as A in figure 6, usually a single neural layer, LSTMs have four neural layers interconnecting in a special way which is specifically designed to enable states to bypass their information to future states. The standard LSTM layer consist of a cell, a forget gate, an input gate and an output gate (Chollet 2018). Together, the three added gates control the flow of information that goes in and out of the cell, and effectively decide what information the network should use, or not use, for a given input (Hochreiter 1997). Consequently, this creates a fictive memory that has the ability to pick and choose which previous states to remember.

These 'memory cells' resolves the vanishing gradient problem of RNNs Olah 2015. In pratical regards of using a LSTM, this means that it now can train on long sequences of data, with several previous states, including lags of different duration. LSTM's abilities to understand time series are prominent for the utilisation of forecasting methods, but have yielded a variety of outcomes, often limited by computationally expensive training. They are, however, the most eligible choice when it comes to multivariate, multi-step time series forecasting from a performance perspective, as academic research on LSTM networks have shown promising results in demand forecasting. For example, Namini, Sima and Tavakoli (Siami Namini et al. 2018) found that LSTM outperforms classical time series and regression models in their case.

The potential of LSTM lies in its ability to model complex non-linear relationships and its proficiency in learning and remembering over long sequences, which is particularly useful for demand forecasting where long-term trends and seasonality patterns are prevalent, such as in the case of a yearly cyclic pattern spanding over several years.

However, when considering implementation costs, both MLPs and LSTM have their trade-offs. While LSTM models offer superior predictive skills for time-series data, they are more complex and require more computational resources and training time compared to MLPs (Pascanu, Mikolov, and Bengio, 2013). This might translate into higher implementation costs, particularly for large-scale applications.

Despite the higher costs associated with LSTMs, their superior predictive performance often outweighs their limitations, particularly in contexts where accurate forecasting can lead to significant cost savings or revenue increases.

4.5 Model ensembling

The concept of model ensembling adresses the problem that different methods of time series forecasting tend to highlight different nuances in the training data's relationships. The technique help reduce the generalization error in machine learning tasks. The task of ensembling a model consists of pooling the predictions of many models together in order to make a better prediction as a whole Chollet 2018. Model ensembling is a machine learning technique that involves combining the predictions from multiple models to generate a final prediction. The main idea behind model ensembling is to exploit the diverse strengths of different models and thereby increase the overall accuracy of the predictions (Opitz and Maclin 1999).

There are several ways to ensemble models, such as bagging, boosting, and stacking. Bagging, or popularly called "bootstrap aggregating", involves creating multiple subsets of the original data, training a separate model on each subset, and then aggregating the predictions. Boosting is an iterative technique that adjusts the weights of observations based on the previous model's errors. Stacking involves training a model (meta-learner or second-level model) to make the final prediction based on the predictions of multiple base-level models (Wolpert 1992).

In the context of forecasting, model ensembling can be a powerful tool. By combining different forecasting models into hybrid models, it is possible to achieve higher predictive accuracy than individual models alone. In order to generate a good ensemble, the constituent models should be good at utilizing different relations or aspects of the data to make their predictions. For instance, one could combine time series models that are good at capturing trend and seasonality, like ARIMA or Exponential Smoothing, with machine learning models that excel in capturing non-linear relationships, such as MLPs, Support Vector Machines or Random Forests (George Athanasopoulos et al. 2017).

A hybrid model might be particularly beneficial when dealing with complex forecasting problems, such as demand forecasting for food producers, where it is expected that multiple factors with non-linear relationships influence demand. For instance, if demand is influenced by price changes, promotions, competitor activity, seasonal patterns, and longer-term trends, this creates both the need for temporal sensitive. A hybrid model that combines a model capturing temporal dynamics (like an LSTM) with a model capturing cross-sectional dependencies (like a Random Forest) could potentially capture these influences more accurately (Kolassa 2019).

However, creating hybrid models comes with increased complexity and potential implementation costs. Training multiple models and combining their predictions requires more computational resources and time compared to training a single model. Also, choosing the right models to ensemble, and the right method to combine their predictions, is not a straightforward task and might require a lot of experimentation and domain expertise (Sagi and Rokach 2018).

Despite these challenges, the potential benefits of improved forecast accuracy often outweigh the associated costs and complexities. Therefore, hybrid models and model ensembling techniques are becoming an increasingly popular tool in the field of demand forecasting, given their potential to capture complex patterns and their ability to deliver superior predictive performance.

4.6 Summary

- Forecasts used for time series modelling are either univariate or multivariate, meaning their target value is only predicted with regards to one or multiple features.
- Statistical forecasting methods, such as Holt-Winters and ARIMA, are fitted to univariate time series. They are both autoregressive models that aime to describe non-stationary time series by seasonality, trend and residuals, though they are in concept best suited to describe stationary time series.
- An adoption of ARIMA called SARIMAX is a statistical forecasting method similar to ARIMA but can in particular be fitted with multivariate time series in order to predict a target variable.
- The statistical methods output a model where the predicted values are regressed on their prior values, giving a linear relationship between those values. Based on these principles, the models aforementioned struggle to model non-stationary time-series. Real world data described by time series are rarely stationary.
- Neural networks is a concept that uses a data-driven fitting to recognize patterns. They can mimic complex non-linear relationships and can be used for time series forecasting.
- Multi-Layered Perceptron(MLP) are one type of network where the input is of fixed size, and sequences in data are learned by using lagged target variables as features in the data.
- Recurrent Neural Networks(RNN) are another type where the network learns by iterating over the sequence of data, step by step. RNNs suffer from a problem of vanishing gradients when learning of long sequences. Long-Short Term Memory is an adoption of RNN that adresses this problem by introducing a memory cell and is therefor the most widely used RNN as of today.
- Different methods are sensitive to different types of bias in the data. The goal of generalizing a pattern can be done by utilizing several methods and combining those methods by ensembling them into one model. This technique is called model ensembling.

5 Case Study

The case study presented in the master thesis investigates the application of point of sales (POS) data for demand forecasting in a Norwegian fast-moving consumer goods (FMCG) food producer. This real-world example aims to provide insights into the practical aspects of implementing demand forecasting techniques using POS data and the effectiveness of various forecasting methodologies within this context is critically evaluated. This assessment draws on previously discussed empirical and theoretical frameworks and concepts discussed in the previous chapters about the empirical and theoretical background, which are then applied to this particular case for a comprehensive exploration and understanding.

The case study covers the following key aspects:

Company Background: A detailed overview of the Norwegian FMCG food producer is provided, including its position in the market, the types of products it manufactures, and its supply chain structure. This background information helps in understanding the unique challenges faced by the company and the relevance of demand forecasting for its operations, and is used to define the problem at hand, completing Step 1 in the previously referred framework for building a forecasting model (2)

Data Collection and Preparation: The study outlines the process of collecting POS data from the food producer's retail partners, as well as the data cleaning and preprocessing steps necessary to transform the raw data into a format suitable for forecasting. This includes addressing issues such as missing values, outliers, and aggregating data at different levels of granularity. This part represents the majority of Step 2 from the framework (2), which is about gathering historical data of the items of interest. This step is further expanded by gathering data not directly linked to the items, but surrounding conditions that may have had an impact on the items.

Descriptive Data Analysis and Feature Engineering: An important aspect of the case study is the identification of relevant features that can improve the accuracy of demand forecasts. This involves exploring various factors, such as seasonality, promotions, holidays, and product attributes, and incorporating them into the forecasting models as explanatory variables. This part concludes Step 3 of the framework (2) as this represents the preliminary (and exploratory) analysis of our data.

Selection of Forecasting Methods: Based on the literature review and the specific context of the Norwegian FMCG food producer, the study selects a set of appropriate forecasting methods to be tested. These include time series models, machine learning algorithms, and hybrid models, each with their own strengths and weaknesses. This helps us to limit the search for an apporpriate forecasting by pursuing the leading contenders for the following model imlementation, as in Step 4 of the framework (2).

Model Implementation and Evaluation: The selected forecasting methods are implemented using the preprocessed POS data and relevant features. The performance of each method is evaluated using a robust methodology that includes in-sample forecasting, cross-validation, and performance metrics such as MAE, MSE, and MAPE.

This concludes the final step for the framework for building a forecasting model, as given by Makridakis (2).

Results and Discussion: The case study presents a thorough analysis of the results, discussing the performance of each forecasting method in the context of the Norwegian FMCG food producer. The analysis identifies the best-performing methods and provides insights into the factors that contribute to their success, such as the ability to handle complex patterns and high volatility in the POS data.

Practical Implications and Recommendations: Based on the findings of the case study, the master thesis offers practical recommendations for the Norwegian FMCG food producer in terms of implementing demand forecasting using POS data. This includes suggestions on the most suitable forecasting methods, the importance of feature engineering, and the potential benefits of adopting a hybrid approach.

Limitations and Future Research: The case study acknowledges its limitations, such as the specific context of a Norwegian FMCG food producer and the potential biases in the POS data. It also proposes avenues for future research, including exploring additional features, testing other forecasting methods, and investigating the impact of the forecasting improvements on the company's supply chain operations.

5.1 Company Background

This section will provide a brief overview of the case company, with details of their current situation in regards to PPC and forecasting. This is used to identify their PPC related and forecasting challenges, which is used to design a solution addressing these challenges. Information about the case company has been collected from three sources: by interviews of different employees at Brynild, a guided tour at their factory in Fredrikstad and from other master theses written for the same company, described in more details in the section about methodology.

Brynild AS is one of Norway's leading food producers and distributors of Fast-Moving Consumer Goods (FMCG). Established in 1895 in Fredrikstad, where they continue to house their production plant and headquarters, the company boasts an annual turnover of approximately 900 million NOK and a staff of over 200 employees. Brynild's product range spans candy, nuts, chocolate, dried fruits, and hygiene products. All products, barring the hygiene line, are produced by Brynild. The hygiene products are exclusively sold and distributed on behalf of a prominent German company, Beiersdorf. The company markets products under various brands such as Minde Sjokolade, Dent, Nivea, Den Lille Nøttefabrikken, and St. Michael, comprising more than 100 standard products. They distribute their products throughout all Scandinavian countries, including Finland, with the largest market share originating from the Norwegian market. In the Norwegian retail market, chocolates account for 3%, nuts for 32%, candy for 18%, and mints for 17%.

Brynild's primary clientele are three major Norwegian wholesalers, who collectively own approximately 40 distribution warehouses across the country. These wholesalers, in turn, supply Brynild's products to around 4,000 retail stores and 2,000 other sales locations such as kiosks, petrol stations, and various convenience stores. As mentioned previously in this paper's empirical background, a merger between retail and wholesalers has occurred in the Norwegian market over the past few decades. This led to the consolidation of retail brands and wholesalers under the same companies, specifically; Coop, Reitan Gruppen, and NorgesGruppen/ASKO. These three Brynild customers maintain similar, but not identical, buy-in strategies. However, the POS data received for this case study is exclusively to NorgesGruppen/ASKO's retail stores.

5.2 PPC and demand forecasting at Brynild

The confectionery goods production market is dominated by large international firms such as Mondelez, Cloetta, Nestlé, and Orkla. Consequently, Brynhild grapples with stiff competition, and their customers, such as Norgesgruppen, possess numerous alternative sources for their FMCG product supply. The total food sales market is colossal, generating substantial turnover of more than 200 billion NOK only in Norway (Statista 2023). This results in a market characterized by minimal margins, compelling all vendors to minimize costs to achieve break-even. Given such narrow margins, these companies rely on vast volume and the principles of Economy of Scale to generate revenue (**Scale**). This principle involves purchasing and producing in large quantities to minimize per-unit cost. Brynhild's production layout is structured for large batch productions with extended setup times, indicating low flexibility in the production lines. While some of their products, like Nivea, are imported, the majority of their turnover stems from products manufactured at their Fredrikstad factory. Most of these products are produced for a finished goods warehouse, classifying Brynhild as a Make-To-Stock production company.

Brynhild's customers, the wholesalers, expect a service level of over 98% from them, which represents the percentage of on-time delivered orders out of the total orders placed. Given the lengthy production lead times, this necessitates considerable stock holding, making the company susceptible to overproduction to meet their orders. According to their production manager, the opportunity cost of not selling a product is greater than the cost of overproducing one unit. Some of their products are perishable goods, deteriorating in quality over time. If Brynhild consistently delivers low-quality products or fails to meet the agreed service level, they risk finding themselves in a situation where their customers reduce their usual orders or completely shift to competitors.

Owing to Brynhild's production strategy, they are unable to produce and deliver on short notice. They have therefore established three deadlines for their sales solutions, categorized as follows:

- Seasonal articles, which are products that are only in season for certain public holidays
- Campaigns, including both standard products and specialized campaign products

- Standard replenishment for distribution storages

The orders need to be received in advance of delivery for Brynild to be able to deliver on time. Not all products can be taken directly from the finished goods inventory. A Master Production Schedule (MPS) is used to schedule production capacities, work staff, and raw material ordering. It describes all orders and production activities within the next 8 months. Raw material orders need to be placed months ahead of the delivery dates to their customers. Production is set for each week, while production amounts are estimated for each day. Following is an production plan which runs 4-6 weeks in advance of delivery to estimate the approximate production levels of different products. The granularity of the production plan goes further down in details to organize production lines and staffing when approaching the delivery date. As the level of details of the plan increases: the level of flexibility decreases accordingly. Close to delivery dates, the flexibility of the MPS is low, and at approximately 3 week ahead, there is little room for urgent orders to squeeze into the plan.

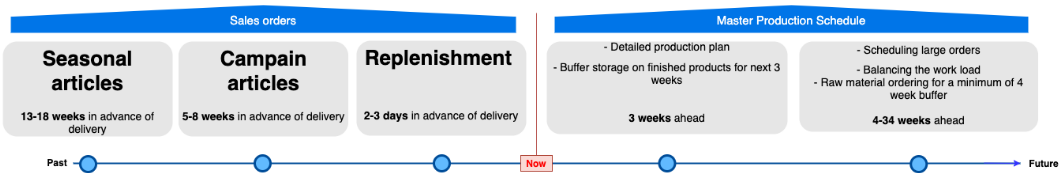

Figure 7: Timeline of orders and MPS for Brynild

The timeline in 7 illustrates how sales are placed in advance of delivery at Brynild, and how the MPS is structured accordingly. The three different types of sales orders have different dependencies, and therefor different time constraints in order for Brynild to be able to successfully handle them. Despite these constraints, some orders are still difficult to fulfill. Different reasons for this are supply shortcomings of materials, the lack of flexibility on production lines due to long change-over times for different products or too low stock either in raw materials or buffer storage.

Brynild's PPC currently faces a set of challenges and opportunities that influence their operational performance, environmental footprint, and bottom-line results. Here is a structured explanation of these issues.

5.2.1 Sales Volumes and Contract Structure

Sales volumes between Brynild and their customers (wholesalers) do not consist of a constant stable flow of small product batches. Instead, transactions are typically organized in large bulks. This approach confers certain advantages, including quantity discounts from Brynild's suppliers and reduced per-product costs leveraged through the principles of economy of scale. This cost reduction extends to areas like transportation, where fixed costs are distributed across a larger number of products.

Brynild's sales contracts with their wholesalers operate on a buy-back principle. If wholesalers are unable to sell the full quantity of purchased goods, Brynild agrees to buy back the remaining products. This principle leads to significant revenue loss for Brynild, both in economic terms - due to the cost of buying back unsold products - and in environmental terms due to the waste generated by perishable goods that reach their expiration date.

5.2.2 Forecasting Model and MPS Setting

Demand forecasts form as one of the inputs for setting Brynild's Master Production Schedule (MPS). These forecasts currently used are generated by Brynild's Enterprise Resource Planning system, SAP, which employs the Holt Winters forecasting method (4.3.2) .

The Holt Winters method uses previous years' sales records to estimate future demand. These estimates are then reviewed by Brynild's forecast manager, who decides whether to adjust them based on subjective assessments. Once finalized, these forecasts are handed over to the production planning team and used as input to the MPS.

While the current forecasting model generally provides reasonably accurate estimates, according to the forecast manager, it falls short in predicting significant events such as holidays, and the production manager have faith that new data sources with additional information about both products and locations will create an addition to the current forecast. Although Holt Winters incorporates a parameter for cyclical behavior, it doesn't account for holidays that do not fall on the same date each year, like Easter. The model also doesn't utilize features like sales prices and locations in its forecasts.

5.2.3 Service Level, Ordering Strategies, and Variabilities

Brynild's customers expect a high service level and base their ordering strategies on this expectation. However, these strategies can vary between wholesalers, leading to differing order types and predictability levels.

Moreover, multiple variables can impact sales, such as seasonality, promotional campaigns, new product launches, and other events. Additionally, external factors like changes in the Norwegian sugar and chocolate tax can have a significant impact on Brynild's profit margins. These taxes effect the prices which is not included in the forecasts, so forecasts have no way of comparing one year against the other with regards to a fluctuant sugar tax.

5.2.4 Addressing Challenges and Improving PPC with POS-data Forecasting

The master's thesis proposes a potential solution to these challenges in the form of improved supply chain transparency and precise coordination of production and distribution activities, underpinned by sharing actual demand information. This approach, if implemented effectively, is expected to mitigate the bullwhip effect, characterized by demand forecast distortion as one moves further up the supply chain.

The first step toward this solution is to understand how to forecast using Point-of-Sale (POS) data. Then, an appropriate model needs to be found that can integrate this forecasting method into Brynild's existing forecasting system, thus enhancing their current MPS setting process.

However, to evaluate the true impact of including POS forecasting, one would need to create a replica of Brynild's current forecast model. This will allow a comparative analysis to determine whether POS forecasting improves performance and adds value to Brynild's production planning. This case study focused on the initial two steps, leaving the final step as a potential area for future research.

5.3 Data Collection and Preparation

The study outlines the process of collecting POS data from the food producer's retail partners, as well as the data cleaning and preprocessing steps necessary to transform the raw data into a format suitable for forecasting. This includes addressing issues such as missing values, outliers, and aggregating data at different levels of granularity. This part represents the majority of Step 2 from the framework (2), which is about gathering historical data of the items of interest. This step is further expanded by gathering data not directly linked to the items, but surrounding conditions that may have had an impact on the items.

The entirety of the technical project was done as a Python project, where all data was processed, combined, plots created and models where built. The different packages and libraries used for both processing and plot creation are specified in the code found in 8.2.

The data collected was as following:

- POS-data from NorgesGruppen
- Storage data from NorgesGruppen
- Product data from Brynild

Our main dataset covers a period of 5 years, from fall 2015 to fall 2020, of Point-Of-Sales data. The raw data consists of multiple time series grouped by day per product per store of Brynild's products sold from NorgesGruppen's stores. It also has a datapoint for turnover for aggregated sale, indirectly giving some insights to sales price. The original dataset consists of 49 million rows with a average of 2 products per transaction; hence a total of 98million sales over the period of 5 years. Considering the previously written theses on the same topic for the producer, this is an huge increase in data, and thereby addressing some of their suggested improvements as part of their chapter for 'further work'. The datatypes found in the POS-data makes it both geo-spatial and temporal, meaning 'spanding over both geographic space and time', which is considered the most challenging data type to work with.

The vast size of the POS-data resulted that the data dump from NG was in the form of 5 CSV-files- one for each year. The storage data had an identical form and postfix, while the three remaining data sources where single files and easier to handle in regards to size and in a relational, non-temporal data format.

File Names	**Source**	**Description**	**Data Format**	**Size**	**Key Variables**
NGPosdata	NG	POS data from all NG retails on daily basis	Temporal and spatial	3.4 GB	Datetime, Product ID, Store ID, Store brand, Turnover, Number of Sales
NG5year_beh	NG	Stock level in all NG retails on weekly basis	Temporal and spatial	1.1 GB	Datetime, Product ID, Store ID, Store brand, Number of SKU
Dim_Vare	Brynild	Description of each Product	Tabular	5.4 MB	Product ID, Product Group and Sub-Groups
Dim_Kunde	NG	Retailer data, added Brynild's own store ID	Tabular	5.4 MB	Store ID, Location, Size of store

Table 2: Overview of Data Files

The Practicalities of handling Big Data

Managing and manipulating Big Data can be a challenging task, particularly when working with large datasets, such as the one used in this thesis for demand forecasting. The Point of Sales dataset was more than 3 GB in size and contained data about numerous products from various stores across the country. A significant portion of time was spent simply loading and editing this data, as minor adjustments took a considerable amount of time due to the dataset's size.

Several strategies were employed to tackle these challenges. One of them involved the use of Pandas, an open source library, providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. This was employed in Jupyter Notebook, a web-based interactive computing environment that allows the creation and sharing of documents containing live code, equations, visualizations, and narrative text. This tool was particularly useful for managing large datasets as it facilitated the retention of data frames in memory, removing the need to reload the entire dataset each time an operation was performed.

However, handling Big Data often requires substantial computational resources, and even using Jupyter Notebook didn't entirely mitigate the challenges. There were frequent instances where the Python kernel, the program that runs and manages the user's computations, crashed due to insufficient resources, necessitating a restart. The hardware used was a Apple Macbook Pro with an ARM-processor called M2 Max with 64 GB RAM, one of the most powerful personal computers on market as of spring 2023.

To increase processing speed and power, a cloud-based service called Paperspace were used. Paperspace is a cloud computing platform that allows users to rent GPU space on a virtual machine (VM) for high-performance tasks. Even though this platform provided additional computational resources that helped improve the speed of data handling, it was also susceptible to crashes, presumably due to the significant demands imposed by the large dataset.

These experiences underline the practical challenges of dealing with Big Data, here in the form of static csv-files. Optimizing data management, ensuring computational efficiency, and securing sufficient resources are essential to effectively handle large datasets. Moreover, it highlights the necessity of robust hardware and software infrastructure to sustain the intense computational demands of Big Data, particularly in the field of demand forecasting.

5.4 Descriptive Data Analysis and Feature Engineering

An important aspect of the case study was the identification of relevant features that could improve the accuracy of demand forecasts. This involved exploring various factors, such as seasonality, promotions, holidays, and product attributes, and incorporating them into the forecasting models as explanatory variables. This part concludes Step 3 of the framework (2) as this represents the preliminary (and exploratory) analysis of our data.

As seen in table 2 , the POS and stock level data received from Brynild was originally provided to them by NorgesGruppen, one of their biggest customers. There had already been performed a data cleaning on the data sets, along with aggregating each Point of Sale, being similar to data received for each receipt of purchase, to a daily basis sum of all purchases from that day.

To investigate the quality of the overall data, several analysis was conducted. This was done so that it would come clear which features should be removed, which had potential to be combined into new features, and which ones that created bias due to corrupt data.

The data analysis included creating tables of description of statistical distribution of data and visual inspections of graphs, which helped to identify the following in the POS data:

- Redundancies

- Negative transactions
- Missing values
- Negative turnover despite positive sales numbers
- Outliers and deviations

These findings showed that the dataset was not properly cleaned, even though the data had previously been inspected by both NorgesGruppen and Brynild. By assumption, these deviations, in example the outliers, may be noise in terms of generalization, and will therefor throw our model off by adopting this noise as a part of the pattern that it trains to recognize. For that reason, all data identified with these errors was removed from the master data, summarizing to consist of approximately 3 million rows.

The analysis also helped identify cyclic behavoir and understand Brynild's top selling products better (19, 18).

Another part of the exploration involved mapping all stores on a map for visual inspection. This helped in understanding where different brands were located, a small indicator of how much they sold, and the ratio of how much of Brynild's products they sold compared to the stores total turnover. 23, (24)

Instead of always testing all 200 different products, there was taken samples, often of 5-8 products to test for different characteristics, such as stationarity, which was not present.

5.4.1 Granularity

Brynild's production plans are, as previously mentioned, rarely flexible enough to move orders up the que with a days notice. The last week until delivery, there is little room for change in the production scheduling, and their scope is usually three-four weeks ahead. It is therefor not of their interest to estimate forecasts on a daily notice, but preferably on a weekly notice.

By aggregation all POS-data on a weekly basis, where the number of products are summerised along with turnover for each store and each product, the level of detail are adequate for the desired outcome- namely to use the forecast for the purpose of production planning, including staff planning, line balancing, raw material ordering, etc.

This greatly improves the computational restraints that comes with the size of data generated by POS.

5.4.2 Feature selection:

According to interviews with Brynild's personal, the sales prices have a great effect on sales, and some products may be more sensitive to price altercations than others, depending on category and competing products. As an example: A lollipop which costs 7 NOK would not be greatly affected by a price reduction of 30 percent, meanwhile a bag of cashew nuts which costs 90 NOK would experience a huge increase in sales.

Alot of Brynild's products are seasonal, and only sold for certain periods of time. There are several different periods of time who generate a boost in sales, but those who generate most sales are Christmas and Easter (18) with a lot of specialized products aimed for the holiday season. While Christmas is always on the same date, this is not the case for Easter. Therefore, an additional dataset was created with all holidays from 2015-2020 and joined with the master data to enrich it with the understanding of when a holiday have had an effect on the sales. The variable was simply used as a flag.

The 'Inventory'-variable is mainly a control variable as it is a result of products recieved versus products sold. When estimating products sold, inventory is redundant as a feature in a forecasting model because it only depends on production and the result of the forecast. Zero inventory would reflect in zero sales, and large inventories indicate a large buy-in, which would be likely before a sales promotion, or other sort of campaigns that would boost sales numbers.

POS is already aggregated by the data owner, our "third party"- wholesaler, to describe the total volume of each product sold for each day in each store, thereby making the turnover a sum of the products sold. The estimated price for each product is therefor not guaranteed to be precise, but will, somewhat assisted by the law of large numbers, be a good implication of the actual price, thanks to the sheer amount of stores and sales for each day. This was used to create another feature, namely "Sales Promotion".

The variable was designed to flag a promotion if the following scenario occured: Grouped by product, store and week, when the price of the same product from the same store brand is 35% less at both time t-3 and time t+3 than it was at time t, then it should be tagged as a sales promotion. How well this correlated with some of the products can be seen in (22) in the appendix, but was overall a strong indicator. An assumption was that the promotions was always often only for different store brands, based on how prices fluctuated differently for the different brands, as seen in 21, where three top selling products where chosen for three different brands.

Another feature added was on population density. Each store has a unique id which is available for services such as OpenStreetMap. OpenStreetMaps API called Overpass Turbo helped to extract a GPS-coordinate in lat/long-form for each store that used to extract and calculate all house adresses in a fixed radius from each store, giving a indicator on how populated the surrounding area of each store is.

The POS-data has been enriched with features that is assumed to add value to insight to the sales mechanisms behind Brynilds products. The size of the data created challenges in regards to computer processing and memory space as the files in the format of comma-separated-values had to be joined in the RAM of the computer. After multiple joins, the master data file had the following features:

- Date of sale
- Store ID
- Product ID
- Number of Sold Products (target value for forecasting)
- Turnover
- Number of the same product in stock
- Product Category
- Store Brand
- Store Size
- Day of the year
- Week of the year
- Month

Repackaging of products

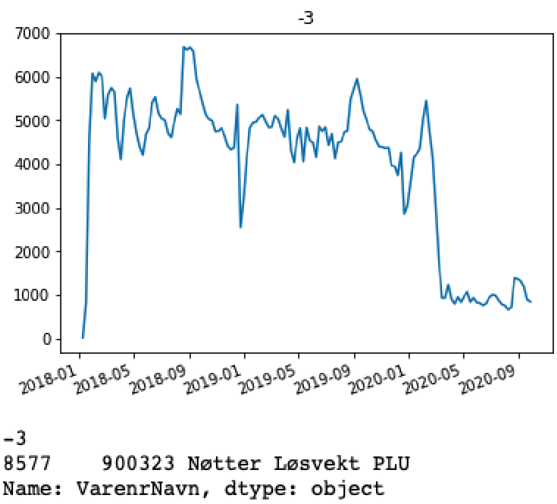

Figure 8: Total sales, separated by brand

Using POS does not provide a complete estimate of Brynilds sold products, because one of NG's number one top selling product, in terms of turnover, is not sold in the same form of packaging. Due to the design of pick & mix-sales installments for nuts, the bags of nuts that are sold are a mixture of the different nuts, chosen by the consumer, where some types sell out faster than others, and therefore creates more orders to the producer. The mixed bag is labeled as 'Pick & Mix Nuts' and does therefor not give any information of what kind is actually sold. Also, this is a example of a product that was partially discountinued when Covid-19 affected consumers freedom of movement, and the fear of infection was high. Nuts had to be repackaged in order to be sold as cups of sampled nuts, which clearly affected sales volumes, as seen in (8). This will create bias in our data. The effect of Covid-19 is not easy to pinpoint for all products, but may be present in other parts of the data.

Resource shortages

Other types of data that are not clearly mentioned, and creates noise for a learning algorithm, is supply shortages. There was a trouble with supply of paranuts, also called Brazil nuts kernels, due to local weather condicions which made the year crops minimal. There is no data for this, just a explanatory negative shift in sales. In other words, the forecasting models are just told that there weren't sold any nuts, and tries to understand why. (9). This is a clear indicator of bias.

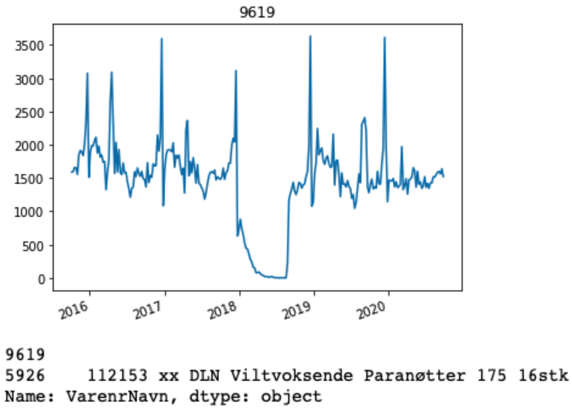

Figure 9: Total sales volumes of wild Brazil nut kernels. Notice the dip in 2018

The key outputs believed to enrich the original data is as follows:

- Number of nearby houses to each store within a 3km radius.
- Weekly updated inventory for each store.
- Holidays. Christmas, Easter and other dynamical holidays in the Norwegian calendar that are defined according to Easter.
- Store specific details, such as size, brand, subcategory.
- Price per Product. An average of each product prize per store brand was estimated by assessing total turnover and total sales volumes per week.
- Sales promotion

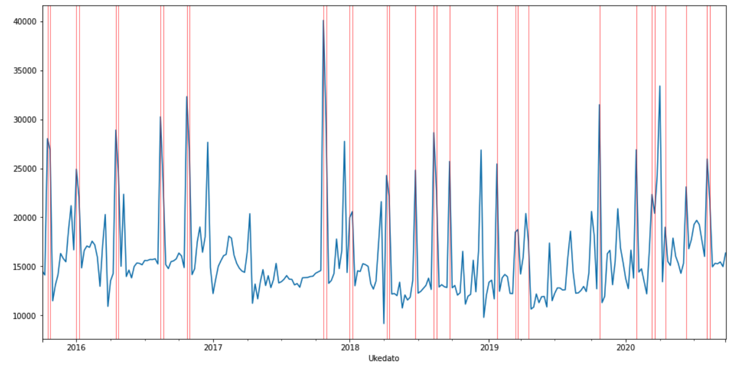

Figure 10: Sales volumes of a selected top selling product. The red vertical line indicates that there was a sales promotion for the product current, and is flagged as a sale in feature for sales promotion is

- Lagged number of sales for previous weeks. An important feature for fixed-window time series forecasting methods to understand sequences.

5.4.3 Horizon selection

When forecasting in time ahead, two different forecasting horizons where chosen. The short-term horizon forecast was 3 weeks ahead, appropriate considering the low flexibility in production scheduling and the desired window for buffer storage at Brynild. With 3 weeks notice, the production line can vary different plans, such as staffing, line balancing and express orders of stock item(sjekk dette med anita, finner ikke kilde). The long-term horizon used is 12 weeks in advance, to meet allow for them to adapt their production planning accordingly, and have the flexibility to handle deviations from what their original plan was. Their flexibility allows the to make adaptions to staffing, line balancing, orders of certain products or ingredients, ect.

5.5 Selection of Forecasting Methods

Based on the literature review and the specific context of the Norwegian FMCG food producer, the study selects a set of appropriate forecasting methods to be tested. These include time series models, machine learning algorithms, and hybrid models, each with their own strengths and weaknesses. This helps us to limit the search for an apporpriate forecasting by pursuing the leading contenders for the following model imlementation, as in Step 4 of the framework (2).

5.6 Model Implementation and Evaluation

The selected forecasting methods are implemented using the preprocessed POS data and relevant features. The performance of each method is evaluated using a robust methodology that includes in-sample forecasting, cross-validation, and performance metrics such as MAE, MSE, and MAPE. This concludes the final step for the framework for building a forecasting model, as given by Makridakis (2).

All models were all compared on the same split between training and test data. The split was adjusted with the selection of the forecasting horizon, being either 3 weeks (13th of August 2020- 28th of September 2020) or 12 weeks (15th of June 2020 - 28th of September 2020). In other words, as the test and training data did not overlap, the models outputted an out of sample- prediction.

5.7 Results

The case study presents a thorough analysis of the results, discussing the performance of each forecasting method in the context of the Norwegian FMCG food producer. The analysis identifies the best-performing methods and provides insights into the factors that contribute to their success, such as the ability to handle complex patterns and high volatility in the POS data.

This subsection will give a overview of the different results from forecasting using Holt-Winters, SARIMAX, MLPs, LSTMSs and an ensemble model which combines, amongst others, exponential smoothing models and machine learning models and bootstrap those best performing for that particular product.

Holt-Winters and SARIMAX where hyperparamater optimized to fit the test data set. MLP and LSTM where fitted to generalize at best by splitting their training data to validation and training, and then estimate how well their algorithm would describe the test data. Hyperparamater tuning of the neural networks where to computationaly heavy to fully automate, due to each model having to build a new network for each product, so hyperparamaters where chosen by optimizing on a very limited set of parameters. This resulted in both networks having a architecture of 3-4 layers depth, approx 14000 nodes, batch sizes of 256 and 20-40 epochs. All specifications of the models can be found in the code uploaded at the github found in the appendix.

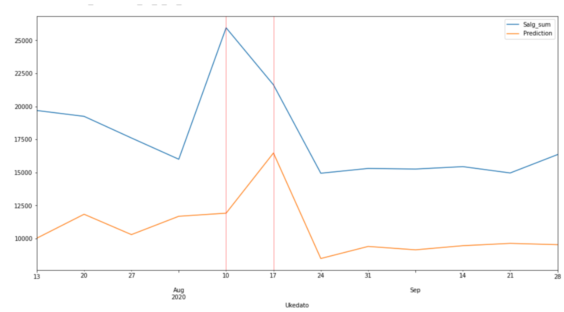

Figure 11: Test set of a spesified product, with a SARIMA forecast over the test set. The red vertical lines represent a indicator of sales promotion during the following week

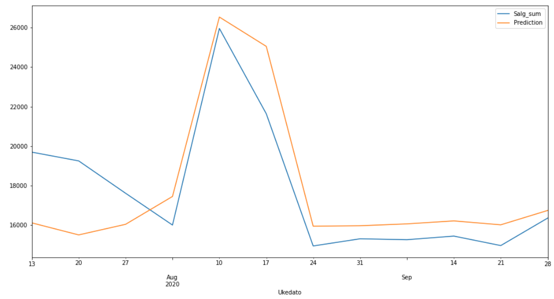

Figure 12: Test set of a spesified product, with a SARIMAX forecast over the test set. Feature for sales promotion was used as exogenous variable, simulating a forecast where sales promotions are available in advance.

Figures 11 and 12 show an example for a forecast of one of the products after adapting SARIMA to handle exogenous variables, in this case the feature 'Sales Promotion'. The top selling products had in general a lot more indicators for sales promotions than the mid and low selling products. This is an example of a forecast of a univariable model compared to a multivariable model, and shows that for this product, the advantage of enriching a model with several features give a great increase in forecasting accuracy.

Holt-Winters takes a fraction of the time of the other models both to set up and to train. The results show a naive forecast, and shows a trend towards averaging recent sales as its best ability to predict new sales. Holt-Winters is the only of the model that fits well when forecasting products that are tested for a low-selling ressesion. This results in a much better MAPS score all over the portoflio as a relative estimation of 100 products when the target is 2 is a much bigger MAPS than estimate of 10000 when target is 8000. HW is used as the baseline model to compare against the other alternatives.

SARIMAX uses a multivariate forecast which makes it a lot better on some products, but as seen in the results table, the general performance over all the products gives results similar to HW. The Sarimax took about 68 hours to train. The forecast was inputed historic sales aggregated per week, and the average amount of sales promotions for that week. The main metric used to evaluate the models performance is RMSE, due to MAPS being infinite when compared against values of zero, and RMSE punishes large deviations rather than many small deviations. By visually inspecting some of the product forecasts, this seems to perform better than HW when not comparing products that have been tested for periods with small sales volumes.

MLP and LSTM have no understanding of what they are prediciting, but generalize according to the data they are fed. Their results are highly variant, scaling from best to worst forecasts for the different groups.

To understand if different products should be forecasted with different models, and to understand if ensemble forecasts could be a prominent way to pursue, Amazon's automated machine learning pipeline called Gluon Time Series was used to see if there were any 'winning' algorithms and compare forecast accuracy on the same products as the other models. The pipeline ensembles a huge variety of models and compare their performance to asses if they should be combined or selected singularly for the best forecast. The results were inconclusive, but pointed in the direction of Exponential Smoothing Models to be selected for the products with low sales volumes, with such long training times that it was not possible to reach final results for full comparison with the other models. However, some of the products that had interesting results were plotted to be evaluated separately (26, 27, 28, 29, 30).

Few to none of the products have a perfectly syncronized seasonality which makes them inadequate to predict in the same model.

Model	**Holt-Winters**		**SARIMAX**	
Error metric	RMSE	MAPE	RMSE	MAPE
Forecast Horizon	3 weeks	3 weeks	3 weeks	3 weeks
Top Selling Products	6 513	43	3 586	31
Mid Selling Products	1 527	45	2 550	65
Low Selling Products	431	Inf	453	Inf
Seasonally Dependent Products	5	Inf	147	Inf
All Products	10 326	Inf	7 554	Inf

Model	**Holt-Winters**		**SARIMAX**	
Error metric	RMSE	MAPE	RMSE	MAPE
Forecast Horizon	12 weeks	12 weeks	12 weeks	12 weeks
Top Selling Products	18 624	147	17 546	112
Mid Selling Products	5 303	109	7 948	132
Low Selling Products	1 202	Inf	826	Inf
Seasonally Dependent Products	64	Inf	2 280	Inf
All Products	32 607	Inf	33 848	Inf

Model	**MLP**		**LSTM**	
Error metric	RMSE	MAPE	RMSE	MAPE
Forecast Horizon	3 weeks	3 weeks	3 weeks	3 weeks
Top Selling Products	5 183	35	11 924	158
Mid Selling Products	2 299	96	2 502	107
Low Selling Products	626	Inf	561	Inf
Seasonally Dependent Products	4 315	Inf	7 755	Inf
All Products	13 210	Inf	25 775	Inf

Model	MLP		LSTM	
Error metric	RMSE	MAPE	RMSE	MAPE
Forecast Horizon	12 weeks	12 weeks	12 weeks	12 weeks
Top Selling Products	17 502	105	26 422	139
Mid Selling Products	6 996	129	6 360	133
Low Selling Products	1 603	Inf	1 422	Inf
Seasonally Dependent Products	5 917	Inf	12 124	Inf
All Products	39 849	Inf	53 165	Inf

```
-*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*-
Vare_Id:  14536
Mean Absolute Error: 146.265
Mean Squared Error: 1411.725
Root Mean Squared Error: 188.8
Mean absolute percentage error: 9.765
Scaled Mean absolute percentage error: 10.519
Mean forecast error: 1256.798
Normalised mean squared error: 0.702
Theil_u_statistic: 0.0
-*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*-
```

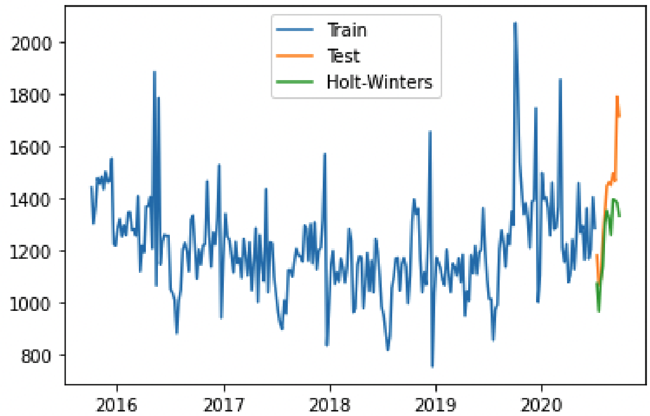

Figure 13: Forecast generated from Holt-Winters with 12 week forecasting horizon of product 14536

6 Discussion

In this discussion chapter, we will address the various aspects of utilizing point of sales (POS) data for demand forecasting in food production, focusing on the challenges and potential solutions that can be derived from the different forecasting models.

The relationship between the POS data potential and its actual utilization in demand forecasting for food producers is so tightly bonded that the research objectives are discussed as a whole in the following sections, to more clearly understand how the data and the methods have a bi-directional relationship.

Product variability:

```
-*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*-
Vare_Id:  14536
Mean Absolute Error: 82.097
Mean Squared Error: 1411.725
Root Mean Squared Error: 96.96
Mean absolute percentage error: 6.335
Scaled Mean absolute percentage error: 6.119
Mean forecast error: 1457.923
Normalised mean squared error: 0.185
Theil_u_statistic: 0.0
-*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*--*-
```

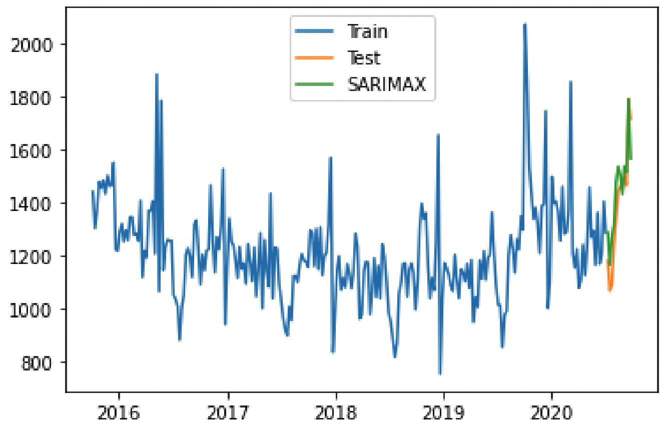

Figure 14: Forecast generated from Sarimax with 12 week forecasting horizon of product 14536

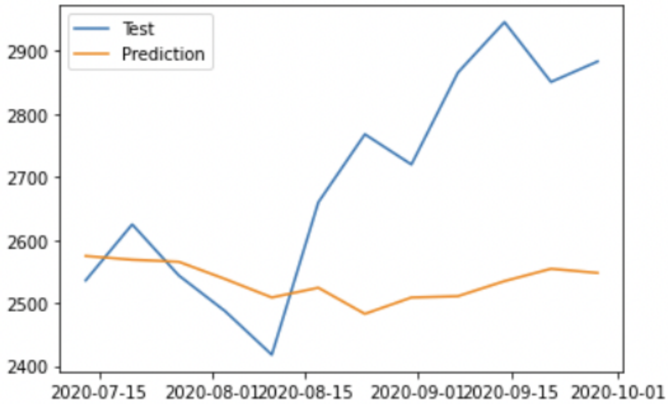

Figure 15: Forecast generated from LSTM-network with 12 week forecasting horizon of product 14536

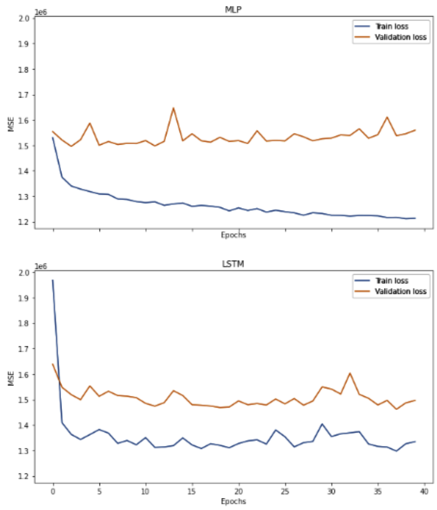

Figure 16: The total loss of the MLP and LSTM models show that it struggles to generalize models that fit with the validation data.

The challenges posed by the variability in product characteristics became significantly evident in the case study, as exemplified by figures 9 , 8 and in products that were produced for short seasons or events. These diverse products display distinct seasonal patterns and encounter unique challenges that can impact the performance of forecasting models. This variety underscores the difficulties encountered when applying a universal forecasting model.

To address this issue in a practical sense, forecasts could be created taking into consideration each product's inherent seasonality and unique characteristics. Alternatively, products can be grouped and forecasted collectively, based on shared features. While such an approach of aggregation may not contribute directly to the task of Brynild's production manager in managing the Master Production Schedule (MPS) at a granular product level, it can be instrumental in identifying trends. Moreover, it can serve as a crucial parameter for individual product forecasts.

The concept of aggregation and its associated benefits are examined in detail in Dekker et al. 2004. The research indicates promising advancements in improving forecast accuracy by elevating the level of aggregation and suggests effective methods for implementing such an approach. The insights derived from this study can provide valuable guidance in shaping the future strategies of demand forecasting while using POS data, in particular when the product portfolio is of a wide range of products with a few shared characteristics. As seen in 17, the majority of sales is of only 10% of the products. This means that there is not much data from the remaining 90% of the product portfolio.

The concept of aggregation also addresses the challenge of having products with too little data, as was handled in the case by changing the granularity of time from days to weeks. In the scope of using our models to understand demand for a food producer with low flexibility in production lines, there is less need to understand what happens on a monday or a thursday if there is at most one weekly shipment to the retailers, independently. At the same time, there is less need of preprocessing the data, as some models are unfit to handle timesteps in the POS without sales, and the general performance of models tend to improve with larger numbers.

Addressing deviations and variations:

Variations in the POS data could arise from a multitude of factors such as fluctuations in product availability, effects of marketing campaigns, degrees of product visibility, and many others. To highlight the discussion of those just mentioned, here is a discussion on the source and possibly how to address these variations.

Product availability might fluctuate due to miscalculations in inventory, restrictive orders from wholesalers to retailers, or even supply shortages from the producer's side.

Even though the current dataset does not include marketing campaign data, it could be integrated from an alternative source if initiated by the retailer. Presently, there are instances when wholesalers launch promotional events of which Brynild is not entirely informed, leading to a less competitive stance for its supply chain network (Patnayakuni et al. 2006). This observation is supported by tests showing an impact on forecasts when sales promotions are not included. Therefore, sharing such information upstream could significantly enhance Norgesgruppen's supply chain performance, particularly with respect to competitive positioning.

In the absence of promotions or campaign data, identifying bias sources in the data becomes challenging, causing what may appear as unexplainable variances or 'noise'.

As for the matter of product visibility, although it isn't as straight forward, existing models can quantify visibility levels, as suggested by Lu and Seo 2015. These could be individually applied to each product. Yet, this approach conflicts with the aim of consolidating forecasts at the product group level. Therefore, to maintain a clear research focus, this thesis should either delve further into the potential advantages of multivariate forecasting — incorporating features such as visibility — or aim at pinpointing aggregation levels that offer more accurate forecasts. This could significantly enhance long-term forecasting from a producer's viewpoint.

Understanding these distinct types of deviations is instrumental in pinpointing the origins of forecasting errors. This knowledge can then inform the refinement of forecasting models, thereby enhancing their accuracy and overall performance.

Deviations resulting from poor data quality don't inherently characterize the data, but rather may manifest as a pattern due to the substantial volume of data and potential mismanagement on the retailer's side. Such issues could result in noisy or incomplete data, outliers, and errors, potentially leading to inaccurate forecasts. This was indeed the situation observed with several of the products found in the Point of Sales (POS) data received from NorgesGruppen.

Challenges with limited observations and numerous variables:

Datasets with few observations but many variables can lead to a large hypothesis space and increased complexity in model training. One approach to addressing this issue is to involve production managers with domain knowledge about product raw materials, availability, shelf placement, campaigns, sales patterns, and other factors that can help in feature selection and model development. Another is to limit the number of variables as suggested in the sub chapter about Further work. While it is unclear wether or not that was the reason for under-/overfitting the machine learning models of this thesis, it is a possible root to the problem.

Practical considerations in model selection: Understanding the impact of deviations

The impact of deviations can fluctuate depending on several factors. These include the sales volume of the product (top-seller vs. low-seller), the direction of the deviation (negative vs. positive), and the confidence level of the prediction interval (such as a 95% confidence interval). A thorough understanding of these elements can facilitate the interpretation of results, thereby guiding informed decision-making, as in this particular case; the choosing of the best performing method.

To illustrate, consider a scenario where Brynild is deciding between two forecasting models. One model has a Root Mean Square Error (RMSE) that is half the value of the other for two products used for model evaluation. On surface, the first model appears superior. However, if the error distribution across the two products is examined, a different picture might emerge. It may reveal that the seemingly inferior model actually performs slightly better on a top-selling product. In contrast, the other product, which contributes less to overall turnover and returns, may skew the error distribution in the first model's favor. In such a case, choosing the "better" model according to RMSE could lead to an inefficient allocation of resources, thus presenting a cost problem. This implies that the models and split of training and test data should not be done without domain knowledge, which leads back to how much resources should be prioritized to build a new forecast model.

Comparing Time Investment: Building Machine Learning versus Statistical Methods

Investing time in building machine learning (ML) and autoregressive (AR) pipelines can require a diverse range of resources, often posing significant challenges for businesses. Especially when considering the implementation and maintenance of complex ML models, the practical feasibility and benefits of employing such models for demand forecasting become vital considerations.

In hindsight, it could be suggested that this thesis may have attempted to cover too broad a scope, potentially overcomplicating the core issues. Future efforts might be better directed towards exploring the most effective ways of implementing a single innovative method, such as Long Short-Term Memory (LSTM). The emphasis should be placed on refining the selected model and experimenting with diverse data preprocessing techniques to optimize the chosen approach.

Model selection could be effectively conducted via competitions based on similar data. Such competitions could facilitate the process of identifying the optimal model for the task at hand. This approach would leave more room to focus on preprocessing and potentially stacking smaller networks, rather than attempting to apply comparable data across a range of different models, as suggested by Makridakis (Makridakis, Spiliotis et al. 2018).

Tuning ML models is an intricate art form 4.4.3, a fact that should not be underestimated. This task can be remarkably labor-intensive, and given its lack of guaranteed success in creating a well-functioning forecasting model, it is not to be undertaken lightly. Therefore, food producers contemplating the implementation of ML forecasts should carefully evaluate the cost-effectiveness of such a project. They may find more value in addressing simpler, more immediate issues - or "low-hanging fruit" - especially if resources are limited. A way of mean may be to implement a multivariate regressional method, such as Sarimax presented in this case study.

Another focus area should be how models that are dynamical should be updated during the lifetime in a production ML pipeline. The machine learning methods are computationally expensive to train and run, and requires infrastructure to handle the continuous data flow that would be received of POS. At the same time, the implementation of the ML models showed how prone they were to break, and should be monitored by a data engineer to ensure that it remains operational.

From a cost perspective, the resulting question as a producer is if the gain of improving the company's demand forecast gives enough operational benefits of utilizing ML models in it forecast, compared with the size of capital and operational expenditures required to build infrastructure, build and maintain ML forecasting models. As a remark, there are also an alternative cost in

Comparison of model performance:

The performance of various models, such as univariate vs. multivariate and statistical regression vs. deep learning, should be evaluated using error metrics that quantify the deviation between predicted and actual sales. It is important to consider the aggregate performance of the models, as well as the performance of individual sub-models, to ensure that the overall forecast accuracy is not skewed by a few well-performing models. As there were no given parameters for seasonality, trend and residuals used to fit the statistical regression models, a hyper parameter optimization was conducted.

Dues, introduces a new problem not found during the literatur study: Comparing a Holt-Winters or SARIMAX forecasting model with an MLP (Multilayer Perceptron) or LSTM (Long Short-Term Memory) model after all have undergone hyperparameter optimization can lead to an inequitable comparison. The discrepancy arises primarily from the distinct nature of these model types and the standard practices used in their application.

Hyperparameter optimization is a common practice in machine learning. For models like MLP or LSTM, this process involves tuning various parameters like the number of layers, the number of neurons in each layer, learning rate, etc. These adjustments help to better adapt the model to the data, potentially improving its predictive performance. This process of optimization is integral to the design and training of machine learning models.

Conversely, statistical models like Holt-Winters and SARIMAX generally have few parameters that can be 'tuned' in the same sense. These models typically depend on inherent characteristics of the data, like trend, seasonality, and autoregressive or moving average components, rather than complex, adaptable structures. When we apply the same hyperparameter optimization process to these statistical models that we use for machine learning models, we create an artificial scenario that strays from their intended use.

Essentially, by optimizing statistical models, we may force them to behave more like machine learning models, pushing them beyond their design specifications. This adjustment may lead to overfitting, where the model fits the data too closely and may not generalize well to new, unseen data. Thus, although both types of models are optimized, the comparison becomes unfair because we are, in effect, comparing a machine learning model with another model that's been coerced into mimicking a machine learning model, rather than its inherent statistical model structure.

Another challenge when comparing statistical models like Holt Winters (HW) to machine learning models lies in their inherent data handling strategies. The Holt-Winters model, by design, predominantly focuses on the latter part of a data sequence to predict future outcomes. As such, while it can be tested on a different segment of the data - say 2018 - the underlying requirement is to remove all training data post-2018 and utilize only the 2015-2017 dataset. In contrast, machine learning models can utilize all available data, and keep out for the specific test set year, such as 2018. This discrepancy creates a fairness issue in the comparison, as these models cannot be directly evaluated against each other. Instead, they can only be tested on the most recent part of their respective temporal data.

It's not beneficial to use data from 2018-2021 to test the prediction performance for November 2017, as the models are inherently designed to forecast future trends. However, a Multilayer Perceptron (MLP) could potentially perform such a task, although this might not be feasible with LSTM (Long Short-Term Memory).

Testing the models accurately becomes even more challenging when we aim to forecast three or twelve weeks ahead, not necessarily confined to 2020, as we selected in the test data. Various unpredictable factors, including market trends, competition, income level drops (for instance, due to COVID-19), and other factors that influence purchasing behavior, can cause significant fluctuations for other years' test data. Therefore, there's a possibility that 2021 cannot be accurately predicted based on the information gleaned from 2015-2020, further complicating model comparison and assessment.

The third major challenge when it comes to building out-of-sample forecasting is the 'black box' issue regarding machine learning methods, as they in a scenario where they are aggressivly overfitted, produce estimates with exponential long term growth or decline, as shown in the example of 27. By simply visually inspecting several of the models, it comes clear that the Ensemble model from Amazon has issues with restraint 28, cycles and seasons 29, 30 and is sometimes not able to adapt at all, without giving any good explanation as to why a regression model without trend or seasons was the best possible fit 26. The poor performance of some of the models is a proof to the

Evaluation metric and under estimations:

The importance of accuracy metrics and evaluation methods for forecast models has been emphasized repeatedly throughout this thesis. To achieve optimal performance, a variety of evaluation methods have been employed, providing a comprehensive and fair comparison between different models. However, one crucial element appears to have been overlooked.

The ultimate objective of forecasting is to perfectly identify future demand, and while this has been the primary focus of the efforts put into the case study, a perfect forecast-to-future-value alignment is, realistically, unattainable. This raises an important question: when perfection is not feasible, how would an underestimation versus an overestimation impact the outcome of a given forecast? This reflection is a result of the discussions by Kolassa 2019.

As discussed in the case study, as a food producer, an improvement of the demand forecast could directly influence costs by reducing waste, reducing stockouts and meeting the requirements in the SLA with customers. Hence, when training a forecast model based on machine learning, the evaluation metric which pinpoints what goal our forecasts aims for should also represent a point with a minimum of cost. In the example of Brynild, this will have to be at a minimum of 98% of an order fullfilment, effectively requireing a forecast of POS to always aim to only have positive residuals, meaning never underestimate.

This thesis addresses this challenge by proposing a novel evaluation metric: the Weighted Mean Squared Error (WMSE), or more specifically, the "Underestimate-Weighted Mean Squared Error" method. This method is of particular relevance due to the following reasons:

Service Level Agreement (SLA) Compliance: The WMSE metric heavily penalizes underestimations in forecasted demand, aligning well with the company's SLA that requires at least 98% fulfillment of order volumes. Consequently, a model trained using this loss function would be incentivized to minimize underestimations, thereby aiding in SLA compliance.

Cost Optimization: Overestimations may result in excess inventory and elevated carrying costs. However, these are often less than the costs associated with underestimations, such as stockouts and lost sales. Thus, the WMSE, by emphasizing underestimations, optimizes the balance between carrying and stockout costs.

The tailored forecasting accuracy metric, the "Underestimate-Weighted Mean Squared Error," has been created, considers these points. This function calculates the Mean Squared Error, sharing characteristics with RMSE such as handling days without sales in the training data. Additionally, it imposes a penalty on underestimations.

In the provided Python implementation, the underestimation penalty parameter controls the penalty level for underestimations. This parameter is adjustable to achieve the desired level of penalty and align with the service level rate of 98%.

Always remember, forecasting models should be developed with a specific use case in mind. It's not merely about identifying patterns in historical data; it's also about understanding how predictions will inform decision-making. Considering the substantial cost associated with underestimations in the case of a potential termination of sales contracts, models should be trained to minimize them.

Training models without considering their intended application can lead to suboptimal or even detrimental outcomes. A model that equally penalizes underestimation and overestimation might perform well in testing scenarios but could cause frequent SLA breaches and associated penalties in real-world applications.

Therefore, it is vital to train forecasting models to optimize for specific objectives aligning with their intended use. The proposed WMSE algorithm accomplishes this by training the model in a manner most beneficial for the company's business objectives.

Combining machine learning and statistical methods:

Recognizing the individual limitations of machine learning and statistical methods, there lies an opportunity in a hybrid approach, blending the strengths of both. This fusion aims to harness the respective strengths of different methods, thereby enhancing the overall efficacy of demand forecasting. This synergy could offer a more robust solution, capable of adjusting to the dynamic nature of demand, ultimately contributing to the company's strategic and operational objectives. While Amazons ensamble model appealed as as a promising, quick-fix solution, the reliability of 'out of the box' methods for machine learning can often be called into question. Despite Amazon's renowned expertise in machine learning, their ensemble model failed to yield accurate forecasts in the given context, although it compbined the powers of both statistical regression and machine learning techniques.

The reasons for its subpar performance can be attributed to a myriad of factors, but a crucial one lies in the inherent limitations of 'out of the box' solutions. Machine learning models necessitate careful calibration to the unique characteristics and demands of the dataset at hand. A model that is not sufficiently customized might lack the necessary precision and adaptability to capture the intricate nuances and complex patterns of the data.

In essence, machine learning is not a one-size-fits-all domain; success heavily relies on the thoughtful customization of the model, which accounts for the specificities of the problem and data in question. As such, the 'out of the box' ensemble model's inability to generate satisfactory forecasts serves as a testament to the necessity for tailor-made machine learning solutions over generalist, pre-packaged models.

6.1 Further Work

Different neural networks have again and again shown that they are difficult to replicate because of the lack of standarization when it comes to building successfull models. This is a problem that should be addressed in order for the work of Brynhild to contribute to improve forecasting among other peers in the Norwegian food industry.

Because different models have different advantages and disadvantages, there has been made suggestions of combining methods with promising results (Ebrahimpour et al. 2011). For example by exploiting the less costly training of an MLP to solve a classification problem, instead of a numerical problem, to either decide if the forecasted event will expect to experience a decrease or an increase in sales, and then add this as a input in a LSTM network that already predicts on the entire POS data as input. The MLP network could in the case of Brynild use a different dataset about the store location, size, nearby competitors, competitors size, previous weather anomalies, etc. Most of these features (except weather) does not need to be updated, as opposed to POS-data, as often. This could also potentially utilize the strengths of different methods and should be tested.

The primary goal will be to increase the accuracy of the model, but there's also a need for the use of models that are easier to interpret for non-experts. And so, another alternative is to combine a machine learning methods with a statistical method, such as given in the example of G.Peter Zhang 2003 where a ARIMA model is combined with an ANN tested on Wolf's sunspot data, a standard data set used widely to compare results to other methods, among with other data sets.

Another promising finding was the use of ensemble models. The concept is to generalize a prediction by combining several neural networks, because different networks tend to be better at different patterns in the data. MLPs are less computationaly costly to train, compared to LSTMs, but are not suitable for multi-step forecasting of long sequences. In the literature study, there was found examples where a MLP was used on forecasting three step ahead. In the case of Brynild, this would equal three weeks, which is not sufficient as an horizon. LSTMs does however have the ability of handling sequences to predict enough time steps and could be trained on different parts of the data, such as POS and historic sales.

During the second term at writing this thesis, a interesting article was found during the update of the literature study; it investigated a dataset including the sales history of furniture in a retail store and applied classical time-series forecasting techniques such as ARIMA and compared it to models such as Prophet, LSTM, and CNN, while measuring the accuracy with RMSE and MAPE, similar to this project. The results showed the superiority of the Stacked LSTM method over the other methods. In addition, the results indicate the good performances of the Prophet and CNN models. This study was done on univariate time series, which underlines how comprehensive the task of comparing all the different methods, and different forms of stacking with both uni- and multivariate time series on both long and short term forecasting. All in within the task of a master thesis. While there were no clear indications that a stacked LSTM would outperform the other methods in this case study, this and other studies mentioned such as Siami Namini et al. 2018, G.Peter Zhang 2003 and Alon et al. 2001 are all convincing. The outcome is likely a result of over- or underfitting the models due to improper training.

One of the underlying motivations for doing this project was to see if POS-data, in particular, would improve the forecasts when combined with methods that could make use of a multivariate forecasting. A prominent results could make way for motivating wholesalers to improve communication and sharing of information in the supply chain. It does however not mean that their buy-in strategies will change, and so the same bullwhip effect as seen today may occur. An important notice is therefor to include historic sales, even when using POS data.

The problem of interpretability and explainability in machine learning models make them difficult to use for forecasting controllers/managers that does not have the skill set, and therefor would potentially lead to the need of new employments.

Covid effect

The impact on deep learning on occurance of special events. Some events being game changers, in terms of great shifts or simply a change in trend.

Such examples where shown in the test data, such as Covid-19's impact on nuts 8. The production manager also had a theory that the change of sugar taxes also impacted their sales. Such events will have long term effects. There are also events that would which will change trends, and some creating short term spikes in sales volumes or prices, such as shortages on raw material, boycott of competitors or strikes. Such events could in some regards be implemented in a forecasting model, as such McKinsey's report on rapid demand forecasting sugested Becdach et al. 2023. They proposed a solution where Covid infection levels could be used as a parameter to input in a forecast. McKinsey reported a ~35% increase in sales volumes over three weeks in 2019 in the US within the categories 'Snacks' and 'Personal Care Products'. In the regards to Brynild; such an event would create a substatual pull from the retailers as it is reasonabel to belive that stockoutages would result in large orders on Brynild's end.

Short term spike and boycott

Recent days also brought the attention to the event of boycotts of competing companies. The result of a blacklist for companies claimed to be 'sponsors of war' made several norwegian vendors of FMCG products boycott Mondelez, despite that there were clear recommendations on not to boycott companies within critical sectors such as food industry. Some critics refer to the boycott as 'naming and shaming' by companies that want to position them self to highlight their own ethics, which indicates the power of the medias effect on PR and sales. This boycott resulted in a marketing campain by one of Brynild and Mondelez competitors; Nidar. As the event is still ongoing, the outcomes as not final, so the effect this has had on sales volumes are unclear, but the fact that a disruptive PR could shift the market is interesting, and close to impossible to include in a forecast on Brynilds part.

Dimentionality and PCA

Another challenge with the information gathered for the forecast building is what may seems to be tendencies to over or underfit. A source for this is potentially what is referred to as 'The curse of dimensionality', which refers to challenges faced when dealing with high-dimensional data, such as datasets combining geo-spatial and tabular data similar to the one used in the case. In machine learning, these high-dimensional datasets can lead to overfitting, where a model learns from noise in the training data and consequently performs poorly on unseen data. They can also lead to underfitting, where the model is too simplistic to capture the complexity in the data (Jolliffe 2002).

Principal Component Analysis (PCA) is a solution to these issues. PCA is a technique used to reduce the dimensionality of a dataset, by creating new variables, called principal components, that capture as much of the original data's variance as possible. These principal components captures the variance of the original data through a process that identifies the axes (or directions) in the feature space along which the original data varies the most (Jolliffe 2002).

By reducing the dimensionality of the dataset, PCA can mitigate overfitting and underfitting. It reduces the complexity of the model and improves computational efficiency. However, it does not guarantee better performance for all tasks, and the results might be harder to interpret since principal components don't correspond to original variables, so there is no way of knowing for a business controller, that wants to decompose which variable that corresponded to a increase or decrease of the forecast in a certain point of time.

7 Conclusion

The literature study found POS data to be highly relevant for a food producer as it is an example of how increased information flow could bring the industry a step closer to integrated supply chains. Different methods found in literature within forecasting, in particular highlighting the potential in utilizing multivariate forecasting, as apposed to univariate, opened for the possibility to regard machine learning techniques that could operate on both types of data. The downside of operation with a 'Black box'-forecast found in deep learning techniques in machine learning could be outweighed by a increase in accuracy. With the introduction of ensemble models we mitigate some of the risks of using machine learning techniques who operate clear of statistics and in a nonlinear space.

The case study showed that the real POS-data inquired had characteristics which maked it difficult to fit predictive models accurately. The large varience between products, and large quantities of data points gathered from sources where input can be affected by factors such as seasonality, promotions, stock-outs and other means that no "straight out of the box" predictive modeling technique can be applied if one expects promising results. Along with the huge amount of data from POS comes the challenge of handeling such large amounts, and the domain expertise to effectively select how to preprocess the data. It sets high expetations towards a producers willingnes to invest large resources to build, deploy and maintain a forecasting model feasting on streamed POS data. At the same time, the huge amount of data still was not enough to forecast all products, as the majority of products are not frequent sellers. The result was that there is need to aggregate the data on different levels to stay clear of "The Curse of Dimensionality".

In regards to integration of supply chains and the case of information sharing besides POS, the value of making marketing plans available to the food producer showed evident, as it resulted in more accurate forecasts when taking into use in multivariate methods. Although the multivariate forecasts were generated with limited training, they showed promising potential in capturing and generalizing patterns within the data. With proper refinement and optimization, these methods have the capability to achieve remarkable outcomes.

Bibliography

Alon, Ilan, Min Qi and Robert J. Sadowski (2001). 'Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods'. In: *Journal of Retailing and Consumer Services* 8.3, pp. 147–156. ISSN: 0969-6989. DOI: https://doi.org/10.1016/S0969-6989(00)00011-4. URL: https://www.sciencedirect.com/science/article/pii/S0969698900000114.

Assaad, Mohammad, Romuald Boné and Hubert Cardot (2008). 'A new boosting algorithm for improved time-series forecasting with recurrent neural networks'. In: *Information Fusion* 9.1. Special Issue on Applications of Ensemble Methods, pp. 41–55. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2006.10.009. URL: https://www.sciencedirect.com/science/article/pii/S1566253506000820.

Athanasopoulos, George et al. (Feb. 2017). 'Forecasting with Temporal Hierarchies'. In: *European Journal of Operational Research* 262. DOI: 10.1016/j.ejor.2017.02.046.

Becdach, Camilo et al. (2023). *Rapidly forecasting demand and adapting commercial plans in a pandemic*. URL: https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/rapidly-forecasting-demand-and-adapting-commercial-plans-in-a-pandemic#/ (visited on 15th June 2023).

Bergstra, J. et al. (2011). 'Algorithms for Hyper-Parameter Optimization'. In: *NIPS*.

Brandon-Jones, Alistair, Nigel Slack and Robert Johnson (June 2013). *Operations Management*. English. 7th. United Kingdom: Pearson Prentice Hall.

Chatfield, C. (2001). *Time-Series Forecasting*.

Chatfield, Chris and Mohammad Yar (Jan. 1988). 'Holt-Winters Forecasting: Some Practical Issues'. In: *The Statistician* 37. DOI: 10.2307/2348687.

Chollet, F. (2018). 'Deep Learning with Python'. In: *Manning Publications*.

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

Cochran, J.J. et al. (Feb. 2011). *Wiley Encyclopedia of Operations Research and Management Science, 8 Volume Set*. Chap. Evaluating and Comparing Forecasting Models. DOI: 10.1002/9780470400531.eorms0307.

Corbett, Charles, Deming Zhou and Christopher Tang (Apr. 2004). 'Designing Supply Contracts: Contract Type and Information Asymmetry'. In: *Management Science* 50, pp. 550–559. DOI: 10.1287/mnsc.1030.0173.

Dagligvarehandelen (n.d.). *Dagligvarefasiten av Nielsen*. Accessed: 09-12-2020.

Danese, Pamela (Dec. 2004). 'Beyond Vendor Managed Inventory: the Glaxosmithkline Case'. In: *Supply Chain Forum: an International Journal* 5, pp. 32–40. DOI: 10.1080/16258312.2004.11517131.

Dekker, Mark, Karel van Donselaar and Pim Ouwehand (2004). 'How to use aggregation and combined forecasting to improve seasonal demand forecasts'. In: *International Journal of Production Economics* 90.2. Production Planning and Control, pp. 151–167. ISSN: 0925-5273. DOI: https://doi.org/10.1016/j.ijpe.2004.02.004. URL: https://www.sciencedirect.com/science/article/pii/S0925527304000398.

Disney, Stephen and D.R. Towill (June 2003). 'Vendor-Managed Inventory and Bullwhip Reduction in a Two-Level Supply Chain'. In: *International Journal of Operations Production Management* 23, pp. 625–651. DOI: 10.1108/01443570310476654.

Dobson, Paul and Michael Waterson (1997). 'Countervailing Power and Consumer Prices'. In: *Economic Journal* 107.441, pp. 418–30. URL: https://EconPapers.repec.org/RePEc:ecj:econjl:v:107:y:1997:i:441:p:418-30.

Ebrahimpour, Reza et al. (July 2011). 'Mixture of MLP-experts for trend forecasting of time series: A case study of the Tehran stock exchange'. In: *International Journal of Forecasting* 27, pp. 804–816. DOI: 10.1016/j.ijforecast.2010.02.015.

Elkateb, M.M., K. Solaiman and Y. Al-Turki (1998). 'A comparative study of medium-weather-dependent load forecasting using enhanced artificial/fuzzy neural network and statistical techniques'. In: *Neurocomputing* 23.1-3, pp. 3–13. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0032495104&doi=10.1016%5C%2fS0925-2312%5C%2898%5C%2900076-9&partnerID=40&md5=b58acb31d95c7028df471237a26ec7d1.

Fransoo, J.C. and M.J.F. Wouters (2000). 'Measuring the bullwhip effect in the supply chain'. In: *Supply Chain Management* 5.2, pp. 78–89. URL: https://www.scopus.com/inward/record.uri?eid=2%5C-s2.0%5C-0038519919&doi=10.1108%5C%2f13598540010319993&partnerID=40&md5=38222c1cfcc6926f4038531841d0687f.

Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Granger, C.W.J. (1993). 'Strategies for Modelling Nonlinear Time-Series Relationships'. In: *Economic Record* 69.3, pp. 233–238. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84984416383%5C&doi=10.1111%5C%2fj.1475-4932.1993.tb02103.x%5C&partnerID=40%5C&md5=a991ddae2a7bd98bd153db7bb5164567.

Hecht-Nielsen, R. (1992). 'III.3 - Theory of the Backpropagation Neural Network'. In: ed. by Harry Wechsler, pp. 65–93. DOI: https://doi.org/10.1016/B978-0-12-741252-8.50010-8. URL: http://www.sciencedirect.com/science/article/pii/B9780127412528500108.

Herzog, c., F. Jost and N. Greiner (n.d.). 'Machine learning has revolutionized forecasting during covid-19'. In: (). URL: https://www.oliverwyman.com/our-expertise/insights/2020/nov/machine-learning-will-revolutionize-forecasting.html.

Hochreiter S. Schmidhuber, J. (1997). *Long Short-term Memory*. Neural Computation. Chap. 9(8).

Hyndman, R.J. and G. Athanasopoulos (2018). *Forecasting: principles and practice, 2nd edition*.

Jolliffe, Ian T (2002). *Principal component analysis*. Springer.

Kathuria, A. (2018). 'Intro to optimization in deep learning: Gradient Descent'. In: URL: https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/.

Khakpour, A. (2020). 'Using Machine Learning and Big Data in sales forecasting for production and retail'. In: *Master's thesis, Østfold University College*.

Koehler, Anne B., Ralph D. Snyder and J.Keith Ord (2001). 'Forecasting models and prediction intervals for the multiplicative Holt–Winters method'. In: *International Journal of Forecasting* 17.2, pp. 269–286. ISSN: 0169-2070. DOI: https://doi.org/10.1016/S0169-2070(01)00081-4. URL: http://www.sciencedirect.com/science/article/pii/S0169207001000814.

Kolassa, Stephan (July 2019). 'Why the "best" point forecast depends on the error or accuracy measure'. In: *International Journal of Forecasting* 36. DOI: 10.1016/j.ijforecast.2019.02.017.

Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. Chap. 6.
Kurtuluş, Mümin, Sezer Ülkü and Beril Toktay (May 2011). 'The Value of Collaborative Forecasting in Supply Chains'. In: *Manufacturing Service Operations Management* 14. DOI: 10.2139/ssrn.1841183.
Lee, H.L., V. Padmanabhan and S. Whang (1997). 'Information distortion in a supply chain: The bullwhip effect'. In: *Management Science* 43.4, pp. 546–558. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0031108822&doi=10.1287%5C%2fmnsc.43.4.546&partnerID=40%5C&md5=df71e1efd92918d97406f1d87598d928.
Lee, Hau (Nov. 2004). 'The Triple-A Supply Chain'. In: *Harvard business review* 82, pp. 102–120, 157.
Lu, Yi and Hyun Seo (Feb. 2015). 'Developing Visibility Analysis for a Retail Store: A Pilot Study in a Bookstore'. In: *Environment and Planning B: Planning and Design* 0, pp. 0-0. DOI: 10.1068/b130016p.
Makridakis, Spyros, Evangelos Spiliotis and Vassilios Assimakopoulos (2018). 'Statistical and Machine Learning forecasting methods: Concerns and ways forward'. In: *PLoS ONE* 13.
Makridakis, Spyros, S. Wheelwright and Rob Hyndman (Jan. 1984). 'Forecasting: Methods and Applications'. In: vol. 35. DOI: 10.2307/2581936.
Markovska, Marija, Aneta Bučkovska and Dimitar Taskovski (Sept. 2016). 'Comparative study of ARIMA and Holt-Winters statistical models for prediction of energy consumption'. In.
Mejlholm P T. Christiansen G, H.C. (2019). 'Planning of exceptional demand'. In: *Aalborg University*.
Mohanty, A. (2019). 'Multilayer Percetptron (MLP) Models on Real World Banking Data'. In: URL: https://becominghuman.ai/multi-layer-perceptron-mlp-models-on-real-world-banking-data-f6dd3d7e998f.
Nielsen (n.d.). *Norwegians judgements about the grocery chains are ready.* Accessed: 12-12-2020.
Nielsen, Michael A. (2015). *Neural Networks and Deep Learning.*
Nilsen Lie, E. (2019). 'Exploring factors effecting sales using machine learning'. In: *Master's thesis, Østfold University College.*
Olah, C. (2015). 'Understanding LSTM networks'. In: URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.
Olhager, Jan (Dec. 2010). 'The role of the customer order decoupling point in production and supply chain management'. In: *Computers in Industry* 61, pp. 863–868. DOI: 10.1016/j.compind.2010.07.011.
Olsen, Sindre E. (2022). 'Potential benefits of information sharing for a producer in the Norwegian food supply chain'. In: *Master thesis, NTNU.*
Opitz, David and Richard Maclin (Dec. 1999). 'Popular Ensemble Methods: An Empirical Study'. In: 11. DOI: 10.1613/jair.614.
Ord, Keith, Anne Koehler and R.D. Snyder (Feb. 1995). 'Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models'. In: *Journal of the American Statistical Association* 92. DOI: 10.2307/2965433.
Paik, S.-K. and P.K. Bagchi (2007). 'Understanding the causes of the bullwhip effect in a supply chain'. In: *International Journal of Retail and Distribution Management* 35.4, pp. 308–324. URL: https://www.scopus.com/inward/record.uri?eid=2%5C-

s2.0%5C-33847301406%5C&doi=10.1108%5C%2f09590550710736229%5C&partnerID=40%5C&md5=06b5360c87e2435606fd01c11964d23d.

Patnayakuni, Ravi, Arun Rai and Nainika Seth (2006). 'Relational Antecedents of Information Flow Integration for Supply Chain Coordination'. In: *Journal of Management Information Systems* 23.1, pp. 13–49. ISSN: 07421222. URL: http://www.jstor.org/stable/40398826 (visited on 15th June 2023).

Pedroso, Marcelo Caldeira and Davi Nakano (2009). 'Knowledge and information flows in supply chains: A study on pharmaceutical companies'. In: *International Journal of Production Economics* 122.1. Transport Logistics and Physical Distribution Interlocking of Information Systems for International Supply and Demand Chains Management ICPR19, pp. 376–384. ISSN: 0925-5273. DOI: https://doi.org/10.1016/j.ijpe.2009.06.012. URL: https://www.sciencedirect.com/science/article/pii/S0925527309002011.

Phan, M. C. and M. T. Hagan (2013). 'Error Surface of Recurrent Neural Networks'. In: *IEEE Transactions on Neural Networks and Learning Systems* 24.11, pp. 1709–1721. DOI: 10.1109/TNNLS.2013.2258470.

Raghunathan, Srinivasan and Arthur Yeh (Dec. 2001). 'Beyond EDI: Impact of Continuous Replenishment Program (CRP) Between a Manufacturer and Its Retailers'. In: *Information Systems Research* 12, pp. 406–419. DOI: 10.1287/isre.12.4.406.9701.

Romsdal, Anita (Jan. 2014). 'Differentiated production planning and control in food supply chains'. PhD thesis.

Al-Saba, T. and I. El-Amin (1999). 'Artificial neural networks as applied to long-term demand forecasting'. In: *Artificial Intelligence in Engineering* 13.2, pp. 189–197. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0033116849&doi=10.1016%5C%2fS0954-1810%5C%2898%5C%2900018-1&partnerID=40&md5=f9765b46c7813f75702745c1c0c4b206.

Sagi, Omer and Lior Rokach (Feb. 2018). 'Ensemble learning: A survey'. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, p. 1249. DOI: 10.1002/widm.1249.

Sharda, R. (1994). *Neural networks for the MS/OR analyst: An application bibliography.*

Siami Namini, Sima, Neda Tavakoli and Akbar Siami Namin (Dec. 2018). 'A Comparison of ARIMA and LSTM in Forecasting Time Series'. In: pp. 1394–1401. DOI: 10.1109/ICMLA.2018.00227.

Soman, C. (2005). 'Make-to-order and make-to-stock in food processing industries'. In.

Soman, Chetan, Dirk Pieter Van Donk and Gjc Gaalman (July 2004). 'Combined make-to-order and make-to-stock in a food Production System'. In: *International Journal of Production Economics* 90, pp. 223–235. DOI: 10.1016/S0925-5273(02)00376-6.

Statista (2023). *Consumer goods and FMCG- Grocery retail industry in Norway - statistics facts.* URL: https://www.statista.com/topics/10739/grocery-retail-industry-in-norway/#topicOverview (visited on 13th June 2023).

Stevenson, M, L Hendry and B G Kingsman (2005). 'A review of production planning and control: the applicability of key concepts to the make-to-order industry'. In: *International Journal of Production Research* 43.5, pp. 869–898.

Strandhagen, J.O., H.C. Dreyer and A. Romsdal (2010). *Control model for intelligent and demand-driven supply chains.* IGI Global, pp. 49–70. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84899284247%5C&doi=10.4018%5C%2f978-1-61692-862-9.ch003%5C&partnerID=40%5C&md5=37db882fc697bc03aae25912511f1bd7.

Tang, Z., C. de Almeida and P.A. Fishwick (1991). 'Time series forecasting using neural networks vs. Box-Jenkins methodology'. In: *Simulation* 57.5, pp. 303–310. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0026258339&doi=10.1177%5C%2f003754979105700508&partnerID=40&md5=9bf070f262a12ab178f5aa070212fc79.

Tirkeş, Güzin, Cenk Güray and Nese Celebi (Sept. 2017). 'Demand Forecasting: A Comparison Between The Holt-Winters, Trend Analysis and Decomposition Models'. In: *Tehnicki Vjesnik* 24, pp. 503–509.

Trapero, Juan, Nikolaos Kourentzes and Robert Fildes (Dec. 2012). 'Impact of information exchange on supplier forecasting performance'. In: *OMEGA the International Journal of Management Science* 40, pp. 738–747. DOI: 10.1016/j.omega.2011.08.009.

Veiga, Claudimar Pereira da et al. (2014). 'Demand forecasting in food retail : a comparison between the Holt-Winters and ARIMA models'. In.

Vollmann, Thomas E (2005). *Manufacturing planning and control for supply chain management.*

Weigend, Andreas S, David E Rumelhart and Bernardo A Huberman (1991). 'Generalization by weight-elimination with application to forecasting'. In: *Advances in neural information processing systems*, pp. 875–882.

Wolpert, David (Dec. 1992). 'Stacked Generalization'. In: *Neural Networks* 5, pp. 241–259. DOI: 10.1016/S0893-6080(05)80023-1.

Yoo, YoungJun (2019). 'Hyperparameter optimization of deep neural network using univariate dynamic encoding algorithm for searches'. In: *Knowledge-Based Systems* 178, pp. 74–83. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2019.04.019. URL: http://www.sciencedirect.com/science/article/pii/S0950705119301923.

Zhang, G.Peter (2003). 'Time series forecasting using a hybrid ARIMA and neural network model'. In: *Neurocomputing* 50, pp. 159–175. ISSN: 0925-2312. DOI: https://doi.org/10.1016/S0925-2312(01)00702-0. URL: http://www.sciencedirect.com/science/article/pii/S0925231201007020.

Zhang, G.Peter, B.Eddy Patuwo and M.Y. Hu (2001). 'A simulation study of artificial neural networks for nonlinear time-series forecasting'. In: *Computers and Operations Research* 28.4, pp. 381–396. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035314205&doi=10.1016%5C%2fS0305-0548%5C%2899%5C%2900123-9&partnerID=40&md5=299f3f62c7cf828f5dc67ee8fe256b79.

Zhang, Guoqiang, B. Eddy Patuwo and Michael Y. Hu (1998). 'Forecasting with artificial neural networks:: The state of the art'. In: *International Journal of Forecasting* 14.1, pp. 35–62. ISSN: 0169-2070. DOI: https://doi.org/10.1016/S0169-2070(97)00044-7. URL: http://www.sciencedirect.com/science/article/pii/S0169207097000447.

Zijm, W.H.M. (Aug. 2000). 'Towards intelligent manufacturing planning and control systems'. In: *OR Spektr* 22, pp. 313–345. DOI: 10.1007/s002919900032.

8 Code

8.1 Training Metric

```python
import numpy as np

def underestimate_weighted_mse(y_true, y_pred, weight=2.0):
    """
    This function computes the underestimate-weighted MSE.

    Parameters:
    y_true : ndarray
        True values.
    y_pred : ndarray
        Estimated values.
    weight: float
        Weight for underestimation. Empirically estimated.
        Defaults to 2.0

    Returns:
    float
        Underestimate-weighted MSE
    """

    if len(y_true) != len(y_pred):
        raise ValueError("The lengths of the input arrays do 
not match.")

    # Calculate residuals
    residuals = y_true - y_pred

    # Apply weights
    weighted_residuals = np.where(residuals > 0, residuals *
weight, residuals)

    # Calculate the underestimate-weighted MSE
    uwmse = np.mean(weighted_residuals**2)

    return uwmse
```

```
#Define variables and print examples

y_true = np.array([5, 10, 15, 20])
# Two arrays with similar absolute error, but the first
    underestimates and the second overestimates
y_pred_low = np.array([3, 10, 13, 17])
y_pred_high = np.array([7, 10, 17, 23])

uw_mse_low = underestimate_weighted_mse(y_true, y_pred_low)
uw_mse_high = underestimate_weighted_mse(y_true, y_pred_high)
print("Underestimate-weigthed MSE, low estimations:",
    uw_mse_low)
print("Underestimate-weigthed MSE, high estimations:",
    uw_mse_high)
```

8.2 Github

All code from the data cleaning and model building has been uploaded to a project on github assigned for the purpose of documenting what has been done as a part of the technical project for this thesis:

https://github.com/BGforecastDSproject/BG_complete_project.git

Appendix

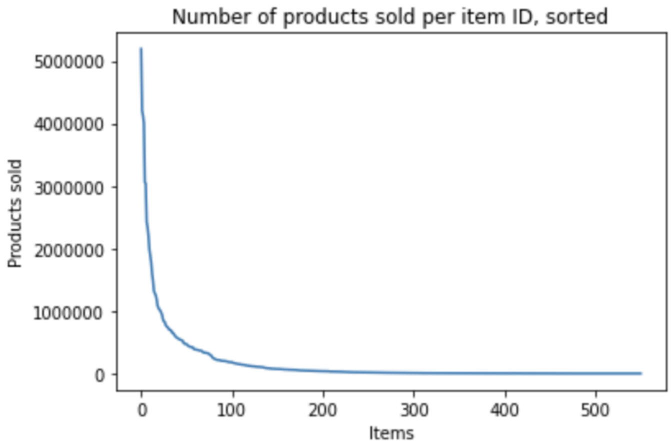

Figure 17: Product portfolio sorted from most selling to least selling products. The curve show a Pareto distribution where 10% of the products makes up for almost 90% of the sales volumes

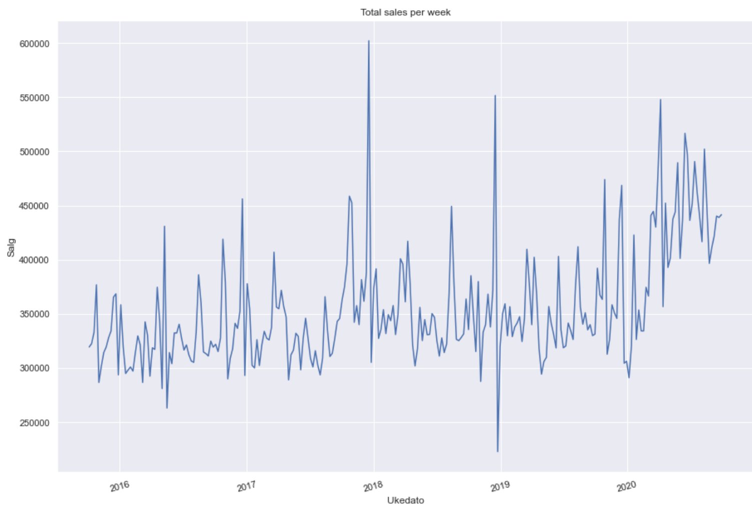

Figure 18: Total number of sales for all products in all stores, aggregated sum of each week, with time represented on the X-axis

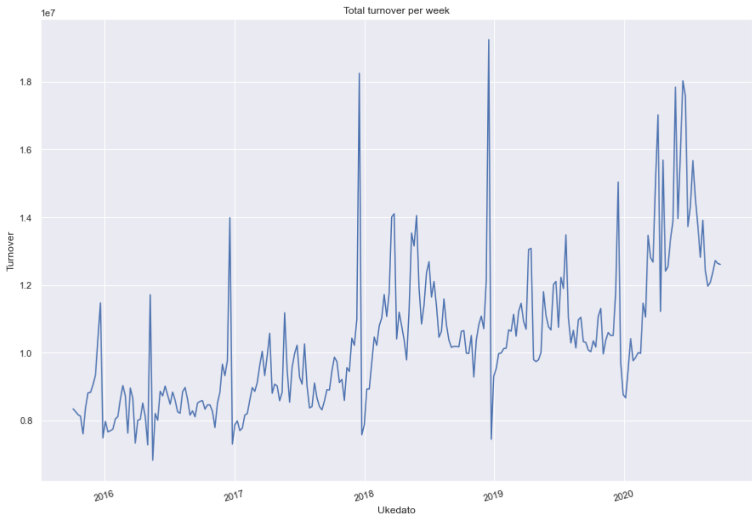

Figure 19: Total turnover for all products in all stores, aggregated sum of each week, with time represented on the X-axis

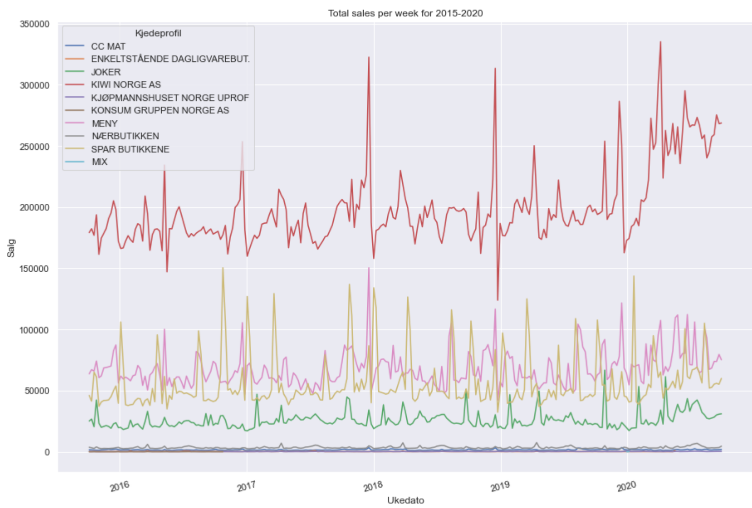

Figure 20: Total sales volumes for all products, separated by all the store brands included in the POS-data

Top Selling Products:	Mid Selling Products:	Low Selling Products:	Seasonally Dependent Products
Brynild Badeball kjpp 11g 120stk	Minde Nøttekos Peanøtter 180g 18stk	NIVEA Creme 75ml	DLN Brente Mandler 190g
Brynild Gomp Fruktpastiller 25g 30stk	DLN Cashew 280g 15stk	NIVEA Men Shower Sensitive, 250 ml	Brynild Juleskum 190g 18stk
Dent Eukalyptus 24g 30stk	Brynild Gul Kamfer 170g 15stk	NIVEA Q10 Anti-Wrinkle Night Care	Brynild Julemix 280g 18stk
B.Vepsebol Jordbær 42g 50stk	NIVEA Soft Body Face cream, boks 200 m	DLN Aprikoser 190g 16stk	Minde Tennis Mørk 120g 22stk
Brynild Chokiss kjpp 11g 120stk	B.Vepsebol Hot Lakris 42g 50stk	Minde Tennis Mørk 120g 22stk	
DLN Nøtti Frutti 350g 15stk	DLN Pistasj 60g 10stk	Brynild Pulverpadder Sur 70g 15stk	
Dent Trio 24g 30stk	B.Supermix Original 240g 20stk	SM Chip Nuts Barbeque 110g 10stk	
DLN Spesial 270g 15stk	Brynild Seigmenn NG 300g 18 stk	NIVEA Sun Prot. Sensiti Lotion SPF 3	
Minde Risbrød 290g 14stk	DLN Pinje 60g 10stk	Minde Knas 160g 20stk	
	DLN Nøtt Bær 60g 26stk		
	Minde Nøttekos Nøtteblanding 170g 18stk		
	Minde Jubileum 250g 16stk		
	NIVEA Double Effect Eye Make-up Remov		
	DLN Brente Mandler 190g		
	NIVEA Q10 Anti-Wrinkle Day Care		
	NIVEA Sun Lotion SPF20, 200ml		
	NIVEA Smooth Caring Body Lotion, 250 m		
	Brynild Kongen av Danmark 60g 15stk		
	Brynild Juleskum 190g 18stk		
	Brynild Julemix 280g 18stk		

Table 3: Names of the products selected to represent each category. The names are fetched directly from Brynild product catalogue used for production planning, which is why each product include the size of a production sized unit, while POS data is only per singular product, which can be misleading

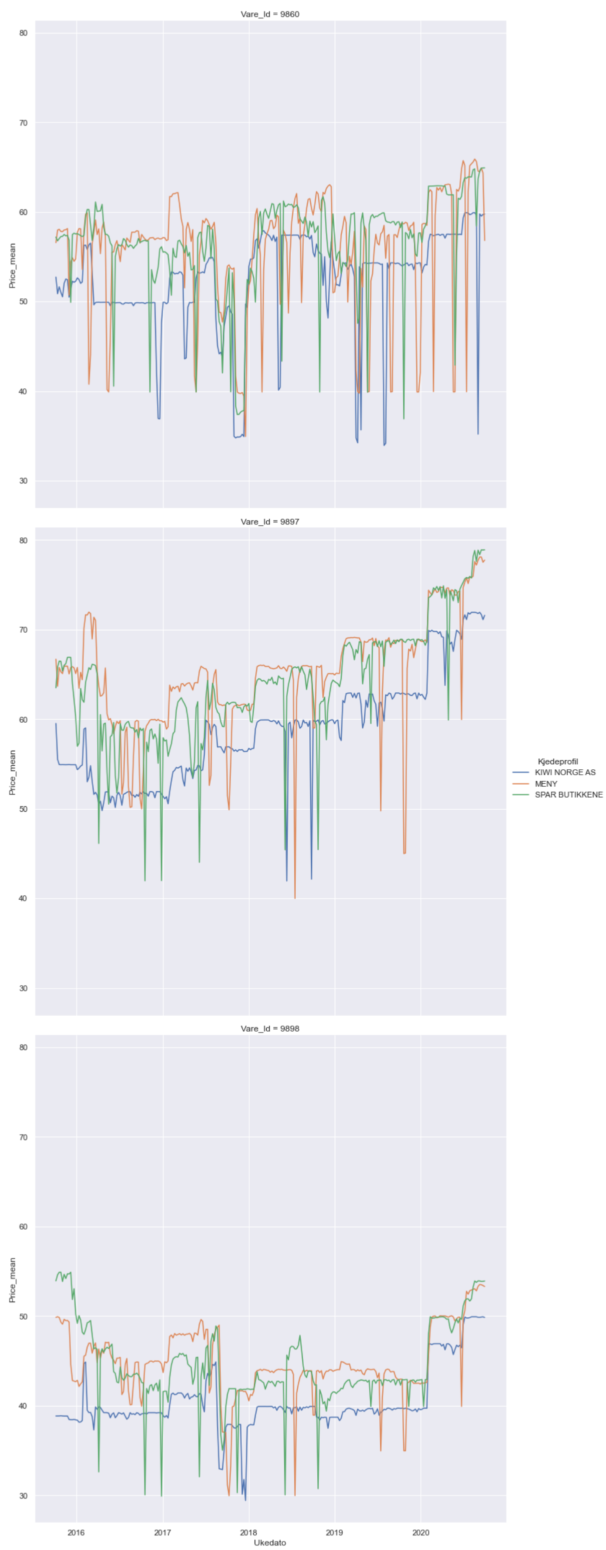

Figure 21: The mean prices of three top selling products seperated by 3 store brands. The big dips in prices indicate that there has been a sales promotion on the product. The dips rarely occur on the same date, meaning that sales are not global between all brands within Norgesgruppen, which means that sales campaigns are not a binary value to be used with the POS-data to estimate sales volumes

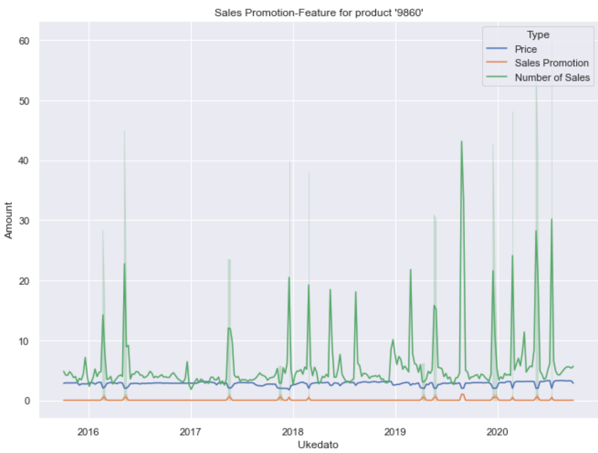

Figure 22: Total sales for product 9860 for one of the store brands. The scale on the green line for 'Number of Sales' are log-scaled, for more easily to combine the three lines in one plot. The new feature is present on all minor spikes on the orange line, indicating that there is believed to be a sales promotion. This plot show the concurrency of deviations in price, which is examined to have a effect on peaks in sales volumes. The feature appears to be a strong indicator, from a visual inspection.

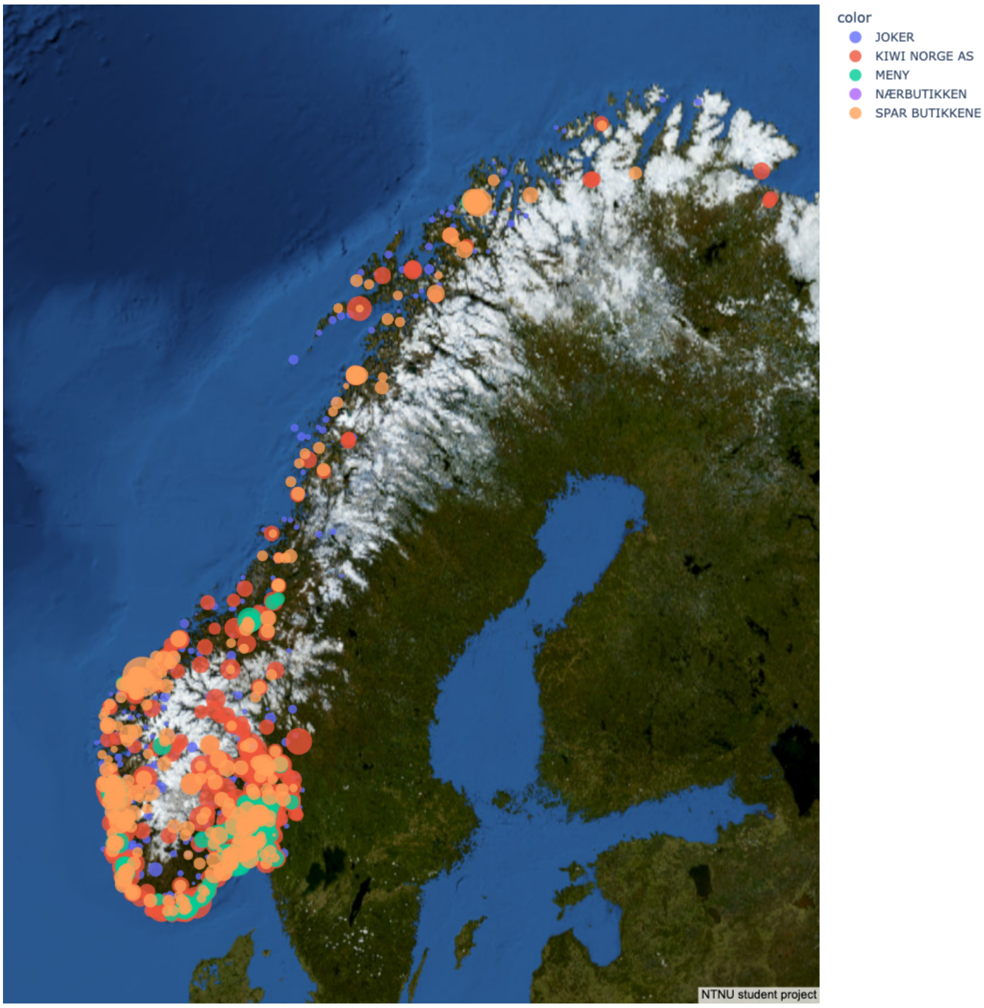

Figure 23: A map of Norway containing a scatter of markers for all stores in the POS dataset. The hue indicates the different brands and the radius of each circle is proportionate with each store's number of total sales of Brynild's products during the time period presented in our POS-data.

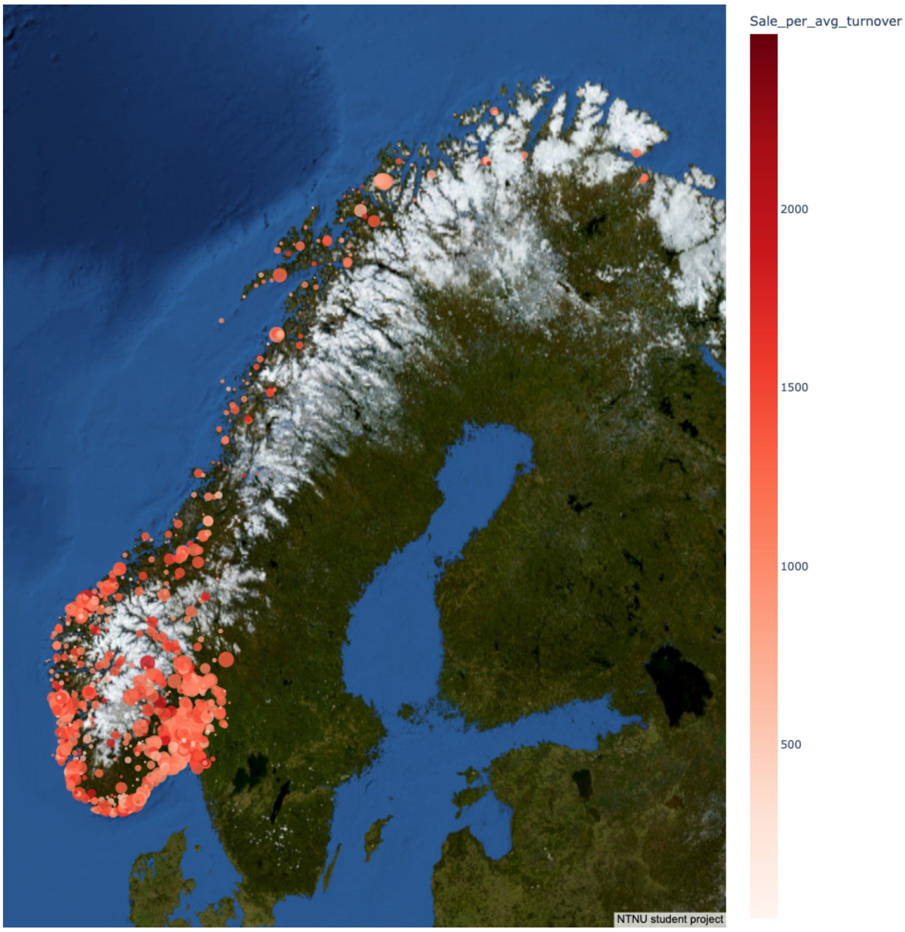

Figure 24: A scatter of Norgesgruppens stores selling Brynilds products. The radius of each circle indicates a relative volume of number of products sold for the store, and the hue of the circle indicates the relative size of number of sales of Brynild products compared to the total turnover for all products in the store. This was done to visually look for geographical correlations or implications that could lead to valuable feature engineering.

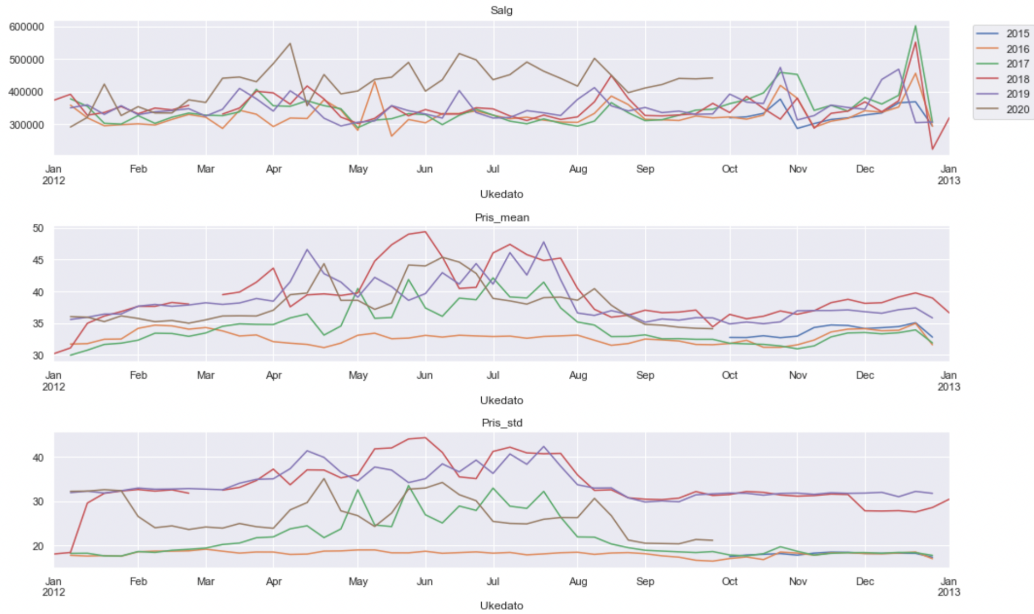

Figure 25: Total sales of Brynilds portifolio seperated by year to find deviations that could lead to new features. The x-axis is wrongfully indicating a year chosen, but is mearly used as a reference calender due to limitations in the python package chosen to illustrate the time lines. In 2018, the sugar taxes target many of Brynilds products was at peak high, which left no clear marks in the prices of their sold products from the wholesalers part. The variance in prices is believed to be due to Brynilds product portfolio, consisting of seasonally dependent products, which gives explanation to why there are few differences between years, but only seasons.

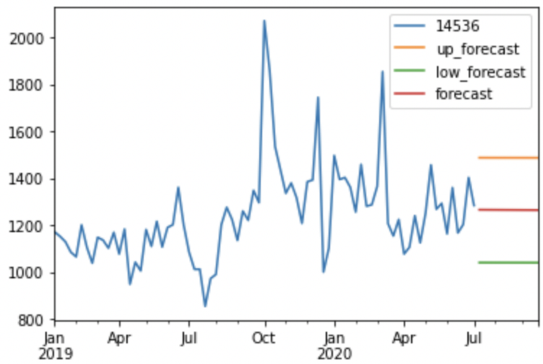

Figure 26: Amazons Ensemble model visualising a forecast, as well as upper and lower boundary of a 95% confidential interval. It shows a 12 week horizon, with a out-of-sample forecast. The best ensamble was made with only a regressor model without trend or cycles.

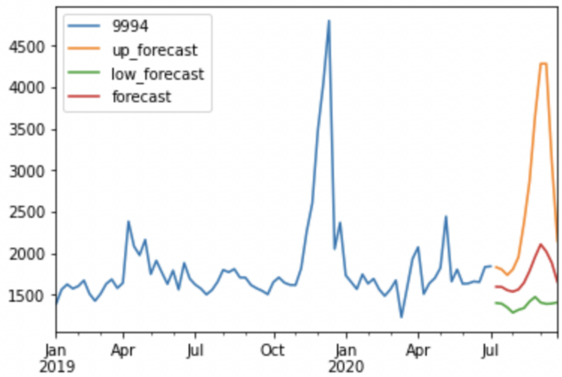

Figure 27: Amazons Ensemble model visualising a forecast, as well as upper and lower boundary of a 95% confidential interval. It shows a 12 week horizon, with a out-of-sample forecast. Notice the large devaiation between lower and upper boundary.

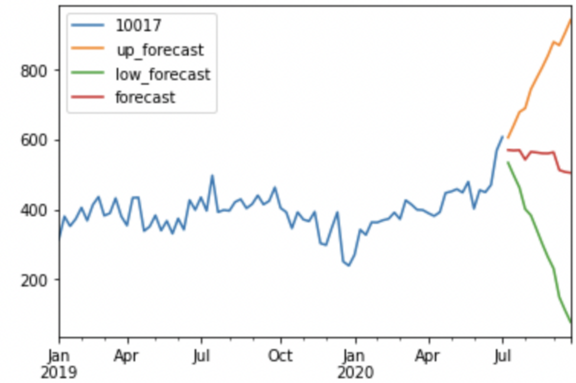

Figure 28: Amazons Ensemble model visualising a forecast, as well as upper and lower boundary of a 95% confidential interval. It shows a 12 week horizon, with a out-of-sample forecast. Notice different trends and large deviation.

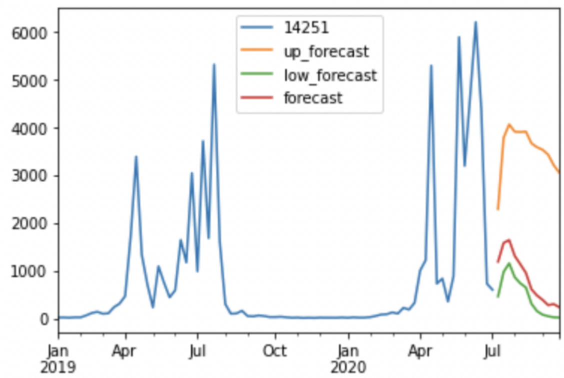

Figure 29: Amazons Ensemble model visualising a forecast, as well as upper and lower boundary of a 95% confidential interval. It shows a 12 week horizon, with a out-of-sample forecast.

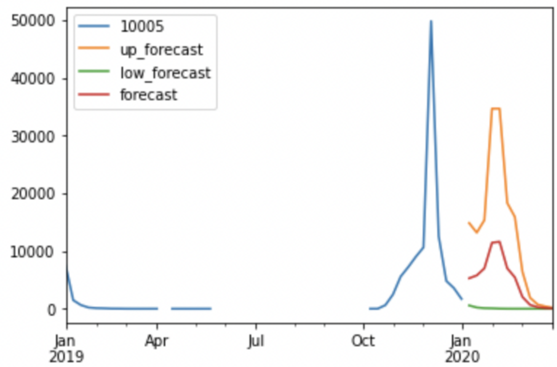

Figure 30: Amazons Ensemble model visualising a forecast, as well as upper and lower boundary of a 95% confidential interval. It shows a 12 week horizon, with a out-of-sample forecast. An attempt to forecast a seasonal product which is mostly sold during the christmas season.

NTNU

Norwegian University of
Science and Technology