SCANIA

Developing a shipment forecast for carriers, by incorporating uncertainty factors through the utilization of machine learning methods

Research for creating a shipment volume forecast for the overseas carriers and freight forwarders

July 20th 2023

Fabian Bosscher

Student Industrial Engineering & Management

SCANIA

Author
Fabian Bosscher

Education
Industrial Engineering & Management
Specialization
Production & Logistic Management

Educational institution
University of Twente
Drienerlolaan 5
7522 NB Enschede

Company
Scania Logistics Netherlands B.V
Blaloweg 22
8041 BA Zwolle
038 497 8023

Company supervisor
Thanos Pavlakis
Material Supply Engineer

University supervisors
Dr. M.C. van der Heijden
Dr. L. Xie

Date of publication
July 20th 2023

Version 2.1

Preface

Before you is my thesis, which marks the end of my master program in Industrial Engineering and Management at the University of Twente, Enschede. Embarking on this master's journey almost three years ago has been quite an adventure. Throughout this period, I have acquired valuable knowledge and experience that will stay with me for life. While this thesis signifies the end of my student life, it does not mark the end of my learning journey. To quote the words of Steve Jobs, "Stay hungry, stay foolish."

I would like to take this opportunity to express my gratitude to everyone who has supported me during my research. Firstly, my first supervisor from the University, Matthieu van der Heijden, for all the time, support and feedback during this research. Matthieu has been involved since the beginning and his guidance has been consistently helpful, providing me with crucial insights. I would also like to extend my thanks to my second supervisor, Lin Xie, for her expertise in machine learning algorithms and the valuable insights she contributed towards the end of the research. Although relatively new to the University, Lin's knowledge significantly contributed to the overall study.

Secondly, I want to express my gratitude to Scania Logistics Netherlands for providing the research opportunity. Specifically, I am grateful to my supervisor, Thanos Pavlakis, for guiding me through the organization, providing relevant information, and offering valuable feedback. With Thanos' assistance, we were able to develop an initial version of the carrier forecast using PowerBI. Additionally, I would like to thank all the other members of the MSE team for making my internship a pleasant experience.

Lastly, I would like to acknowledge my fellow students who undertook their own graduate projects alongside me, as well as my friends and family who supported me throughout this journey. It has been quite a challenging path, and not everything went smoothly, but I am immensely proud of the final result. I could not have accomplished it without their support.

Warm regards,

Fabian Bosscher

Management summary

This research is performed for Scania Logistics Netherlands (SLN), which is responsible for the entire supply network of inbound supplies for Scania in Central Europe. Scania sources materials, parts, and resources from approximately 1,000 direct and 10,000 indirect suppliers worldwide. This research focusses on the North Bound Flow (NBF), which is an internal supply to Scania production units in Europe. From here parts from suppliers outside Europe are stored, custom cleared, repacked, and held at the facility in Hasselt. Since NBF suppliers are located overseas, shipments from these suppliers are transported via air or sea freight and Scania uses the services from carriers to ship these shipments in containers.

Scania and the carrier prefer to have advance knowledge of the shipment volume that will be received at the port of loading where the overseas suppliers are assigned to. Such that the carrier can reserve sufficient space on vessels. The present situation lacks insight for SLN and the contracted carriers regarding the expected volume to receive from suppliers and the specific transportation requirements for Scania's facilities. This volume holds significant importance as it directly affects the required vessel capacity. If the volume exceeds the reserved space on the vessel, the additional volume will be subject to spot market pricing, which are often higher.

The goal of this research is to develop a high-quality carrier shipment forecast, by utilizing the order information of suppliers and the expected packaging information. This forecast determines the total volume (expressed in cubic meter, which can be converted to the number of containers) that carriers need to transport. The time interval for the forecast is set at 10 weeks, meaning a forecast is produced for one week ahead up to and including 10 weeks ahead. The forecast is used to provide insights and pre-book the required transport capacity at the carriers in advance, such that the freight shipped on spot market rates is minimized.

The literature shows that there are various methods that are intriguing to employ. Although most of these methods are suitable for specific elements within the forecast, none of them are suitable for the forecast as a whole. Based on the different methods in the literature, we make a combination of multiple models to create one flow of forecasting. This forecasting flow consists of 10 steps and makes a distinction between the deterministic and stochastic part of the forecast. We assessed the lead time variability by comparing the differences between the planned and actual delivery dates for each supplier. To calculate the shipment volumes accurately, we multiplied them by the expected proportions for each week of shipment. To account for the frozen period in the forecast, we multiply the shipment volumes by the proportions in which week the shipment is expected. The ML methods are relevant for addressing packaging uncertainty. By testing four ML methods to predict the shipment volumes for the non-Scania packaging suppliers, the LightGBM exhibits the highest accuracy among the methods. After selecting the best ML method, we conduct the final forecasting with three test orders.

For a comparison between the old forecast, the new forecast, and the actual shipment volumes, we are limited to data availability. We were only able to conduct an analysis using the data from the final 10 weeks of 2022. The existing forecasting approach yielded a MSE of 59.7 TEUs, indicating an average deviation of 59.7 TEUs over the course of the 10 weeks. However, with the introduction of the new forecasting method, the MSE does have a significantly lower value of 1.6 TEUs. The variance tends to increase when the number of TEUs is relatively high, which aligns with logical expectations. As the shipment volume for a given week becomes larger, the likelihood of fluctuations in lead time and packaging volume also tends to rise. This variance serves as an indicator for the additional costs incurred when shipping extra containers. By decreasing the

variance, we can also lower the costs associated with shipping these additional containers. Applying the principles derived from the newsvendor problem, we calculate a CSL of 66.7%. Consequently, our objective is to identify the optimal number of containers that allows us to achieve this CSL. As a general observation, we find that overbooking with one TEU tends to result in meeting the desired CSL.

The final forecast is now implemented in a first prototype within PowerBI, incorporating the three data sources. The dashboard includes filters for port of loading, supplier location, and Incoterms, allowing users to customize their view. It displays information such as backlog shipments, geographical view of ports and suppliers, shipment volumes, and TEUs in tables and diagrams. Future steps to be taken to this protype are the application of the ML method and the lead-time variability. These improvements contribute to a more accurate and comprehensive forecast for shipment volumes within SLN.

The last step of the implementation of the forecast, are the activities to maintain the quality of the forecast. It is recommended to regularly analyse the performance of different ML methods using new datasets to ensure the highest possible accuracy. We also emphasize the importance of updating supplier information regularly. Examples of situations that require updates include the introduction of new suppliers, supplier relocations, tender implementations, and changes in lead-time variability. The ML method update should be conducted once a year. Supplier information and packaging information should be updated three times a year. Regular updates of the ML method, supplier information, and packaging information are essential for maintaining accurate and reliable forecasts. By adhering to the recommended updating procedures, SLN can ensure that the forecasting process remains up-to-date.

Further research can be done by presenting it to the carriers itself, receiving feedback from them would be highly valuable and appreciated. Also, it might be interesting to incorporate the expected weight of shipments in the forecast. Although weight is less crucial than volume, it can provide insights into the required number of TEUs. Lastly, Scania should explore the potential opportunities of utilizing the vendor evaluation module of the non-Scania packaging analysis which is currently be done.

Table of Contents

List of abbreviations

The following abbreviations are frequently used in this research:

NBF	North Bound Flow
SBF	South Bound Flow
SLA	Scania Latin America
LCL	Less than Container Load
FCL	Full Container Load
SLN	Scania Logistics Netherlands
LACE	Landed Cost Estimation
TMS	Transport Management System
ML	Machine Learning
NN(s)	Neural Networks
CDF	Cumulative Distribution Function
PDF	Probability Distribution Function
TEU	Twenty-foot equivalent unit
MSE	Material Supply Engineering
SIS	Shipment information sharing
CSL	Cycle service level

1 Introduction

This assignment is created by Scania Logistics Netherlands, further referred to in this thesis as SLN. SLN is a division of the larger organization Scania, where SLN is responsible for the entire supply network of inbound supplies for Scania in Central Europe. The mother organization, Scania, is a Swedish manufacturer of trucks, buses, and power solutions. In 2021, its market share for trucks of Scania was in Europe 49%, where it sold 85.930 trucks worldwide. Scania is operating in more than 100 countries with 54.000 employees, where some of the warehousing and one of its production facilities is in the Netherlands (Scania, 2021).

1.1 Background Information

1.1.1 The supply chain

Scania sources materials, parts, and resources from approximately 1,000 direct and 10,000 indirect suppliers worldwide (Scania, 2021). Given the global nature of the suppliers and the internal flow of parts between Scania facilities, the supply chain of SLN involves multiple material flows, which can be categorized as follows:

1. **EURONET** :encompassing flows from European suppliers to Scania facilities within Europe, as well as between different Scania facilities in Europe.
2. **North Bound Flow**: (NBF) internal supply to Scania production units in Europe, where parts from suppliers outside Europe are stored, custom cleared, repacked, and held at the facility in Hasselt. The NBF involves over 125 suppliers from more than 10 countries, covering more than 700 unique parts.
3. **South Bound Flow**: (SBF) can be seen as the flow of parts towards the production facilities in Latin America, also known as Scania Latin America (SLA). So it is importing of goods to the Scania facility in São Paulo (Brazil) where the parts are customs cleared and stored. For example, the flow of materials from the warehousing facility in Hasselt to the production facility in São Paulo is determined as the SBF. When a supplier is located close to the production facility, SLN always considers if a direct shipment is an option.

Figure 1: Facilities Scania worldwide (Scania, 2021)

Figure 1 (included in Appendix I) provides an overview of Scania facilities worldwide. This research specifically focuses on the flow of materials from non-European suppliers to Europe, thus emphasizing the NBF.

Freight forwarders and carriers

Since NBF suppliers are located overseas, shipments from these suppliers are transported via air or sea freight. Therefore, SLN collaborates with carriers and freight forwarders. Carriers are companies that possess transportation assets, such as vessels and containers. Freight forwarders, on the other hand, have agreements with carriers and utilize multiple carriers within their network to transport containers (Shang & Lu, 2012). Engaging a freight forwarder offers flexibility, as alternative carriers can be utilized if any issues arise with a particular carrier's vessel. For simplicity, we use the term "carrier" to refer to both carriers and the services provided by freight forwarders.

The selection of containers for shipping is based on two options: 20-feet or 40-feet containers, differing in length. Furthermore, there are two types of shipments: Less than Container Load (LCL) and Full Container Load (FCL). LCL applies when cargo from multiple owners is combined in one container, while FCL reserves an entire container for a single owner, even if not all the space is utilized (Hsu, Tai, Wang, & Chou, 2021).

Each carrier covers a specific geographical region and the corresponding port of origin. This implies that the location of the supplier is an important argument for selecting the appropriate carrier. For overseas suppliers, there are 27 ports covered by seven carriers. From the port of origin, cargo is transported by vessel or airplane to the destination port, typically Rotterdam or Amsterdam (Schiphol). For FCL shipments, a contracted carrier ensures transportation to the logistics center in Hasselt via barge or truck. In the case of LCL or air shipments, the forwarder manages the transportation to Hasselt. From there, the NBF distributes the cargo to production facilities in Europe. Approximately 90% of the NBF volume is shipped by sea, as it is the most cost-effective method for these parts.

Volume of the cargo

The packaging that is used is an important element in this research since this determines the volume of the cargo. Key packaging information includes the packaging type, dimensions (volume), and the number of parts per package. The packaging type varies among suppliers, with two distinct groups: those using their own packaging and those using Scania packaging.

The order information from November 2022 up to November 2023 (which consists of 16.000 orders) shows that approximately 60% of the NBF are coming from Brazil, Argentina, and Turkish suppliers. For the Brazilian, Argentinian, and some Turkish suppliers, Scania packaging is utilized. This is because in some of these countries, they exchange Scania packaging, or a consolidation point is located, which collects all the deliveries from the suppliers in that country and repacks them according to the Scania packaging. From the Scania packaging, we know exactly the packaging type, what the dimensions are, and how many parts fit within one packaging type. However, information about packaging type and dimensions is unavailable for the remaining 40% of suppliers who use their own packaging.

Confirmation of the shipment volume is typically sent by the supplier to the carrier approximately two weeks before arrival. This confirmation allows the carrier to determine the exact volume to be shipped and allocate capacity on the vessel accordingly. The carrier could use a 20-feet or 40-feet container, where the difference is in the length of the container. There is also a distinction between Less than Container Load (LCL) and Full Container Load (FCL).

1.2 Research motivation

Until approximately one and half year ago, SLN used to book one hundred percent of their freight on the *spot market*. However, there has been a shift towards booking based on contracts since then. SLN now seeks to establish contractual agreements with carriers because it provides them with a guaranteed amount of transport at a fixed price. These contracts involve SLN and the carrier agreeing on the transportation locations, freight rates, and expected total volume (in terms of FCL/LCL containers) on an annual basis. Currently, only suppliers located in a particular continent remain uncontracted.

The expected volume in these contracts depends on whether it is the initial contract with the carrier or an extension of a previous one. In the former case, the expected volume is based on data provided by SLN, while in the latter case, it is derived from the total volume of the previous year. The historical data serves as a baseline for forecasting the volume. It is the responsibility of the carrier to ensure that there is sufficient capacity to accommodate the expected total volume. If the actual volume falls short of the expected volume, the carrier can sell the available capacity on the spot market. Failure to meet the reserved volumes may result in financial penalties for Scania.

However, there are situations where no contract exists or the total volume exceeds the agreed amount of transport. For instance, while the carrier typically transports one standard FCL per week, there may be cases where the supplier suddenly requests the transportation of six FCLs. These additional five FCLs cannot be shipped at the contract price but must be booked at spot market rates. Usually, finding additional capacity on the spot market incurs higher costs compared to the fixed price agreed upon in the contract. Historical spot market freight rates and contracted freight rates from SLN indicate an average of 20% higher expenses for booking containers on the spot market.

SLN and the carrier prefer to have advance knowledge of the shipment volume to reserve sufficient space on vessels. This practice, known as pre-booking, involves the carrier searching for a suitable shipping line that meets Scania's requirements and provides adequate space for the intended cargo. By pre-booking, the volume to be transported for SLN is already allocated on that specific shipping line.

Currently, SLN utilizes an ERP system, which contains order information such as the number of parts ordered, the supplier details, and the expected shipment arrival date. However, there is currently no suitable transport management system (TMS) that integrates this information with the supplier's packaging type and translates it into a volume shipment forecast for each carrier, considering the specific port of origin covered by each forwarder (as a single forwarder may cover multiple ports). Sharing this forecast with the carrier in advance enables pre-booking to take place. Figure 2 provides a schematic overview of the processes involved in FCL container shipments, including estimated lead times.

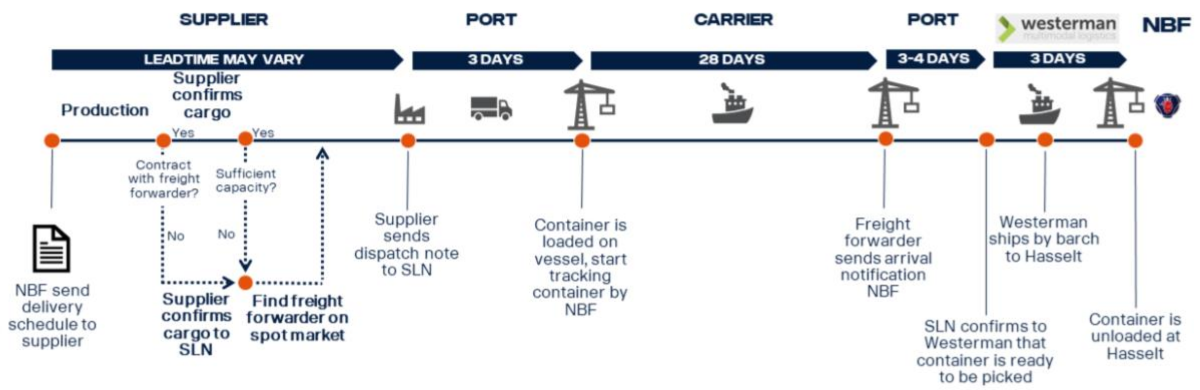


Figure 2: Physical and information flow, from the supplier to the arrival of parts in Hasselt

1.3 Problem identification

The present situation lacks insight for SLN and the contracted carriers regarding the expected volume to receive from suppliers and the specific transportation requirements for Scania's facilities. This volume holds significant importance as it directly affects the required vessel capacity. If the volume exceeds the reserved space on the vessel, the additional volume will be subject to spot market pricing. In cases where there is an urgent need for a specific part in production (referred to as a rush order or speed transport), SLN has the option to utilize air freight, which involves considerably higher cost compared to sea freight. When considering the long-term consequences, opting for shipping at spot market prices is consistently more cost-effective than experiencing production losses due to delayed parts.

Certain aspects of this research necessitate further investigation, with some involving a level of uncertainty. Firstly, there is a *frozen period* for each supplier, which denotes the period before the delivery date during which SLN is prohibited from making changes to the order, such as modifying specifications or quantities. All suppliers adhere to a consistent duration for this frozen period, which amounts to two weeks. Secondly, the *lead time variability* of suppliers introduces uncertainty regarding when the goods are ready to be transported by the carrier. For instance, if suppliers consistently deliver parts late, it significantly impacts the total volume that needs to be shipped. Furthermore, for suppliers who do not utilize Scania packaging, there is a lack of available *information* regarding their *packaging*. Based on SLN's experience, it is known that non-Scania packaging suppliers employ different shapes and volumes. However, SLN lacks knowledge regarding the quantity of parts that can be accommodated in each packaging type, as well as the dimensions and corresponding volume of these packages.

Regarding packaging types, as mentioned in Section 1.1.1, approximately 60% of the NBF will employ Scania packaging. We have precise information about the dimensions of Scania packaging and the number of parts it can accommodate. The total volume from countries using Scania packaging can be calculated deterministically, in addition to accounting for other uncertainties, in contrast to deliveries from suppliers in other countries. If the estimated volume to be transported by the carrier is known a few weeks in advance, pre-booking activities can be initiated to secure sufficient capacity, ensuring any additional volume can be shipped at the appropriate price without resorting to the spot market. Apart from reducing costs, having this information also leads to a more stable workload for the material planning teams at SLN. In summary, the core problem can be outlined as follows:

Currently, there is no comprehensive overview of the volumes that must be transported by the carrier, leading to a substantial portion of the volumes needing to be reserved through spot market bookings.

1.4 Research goal

The objective is to integrate order information from the ERP system with the anticipated packaging volume used by the supplier. This integration enables the creation of a standardized shipment volume forecast for each carrier, which can be filtered based on the ports they serve. By incorporating the order and packaging information, along with other elements of uncertainty, the forecast can be improved for greater accuracy. A more accurate forecast allows for early pre-booking activities, providing financial benefits to SLN by avoiding the need to enter the spot market.

The forecast should offer a visibility of 10 weeks (prior to the expected arrival at the loading port) regarding the total volume that needs to be transported. This allows carriers to reserve adequate capacity on a vessel. Providing a forecast with less than 10 weeks of visibility does not allow carriers sufficient time to adjust their capacity, while a forecast exceeding 10 weeks becomes less accurate due to changes in production planning. Pre-booking activities can be carried out based on the total volume, involving reserving capacity within a single container (LCL) or assigning multiple containers (FCL).

Each carrier should be able to receive a dynamic report which lists all Scania suppliers that fall under the geographical region they cover, that predicts the total shipping capacity needed to cover Scania's volumes. The data used to create these reports are on daily basis and should be translated to a weekly-based report because the capacity of a carrier is considered on a weekly basis. If for example, a delivery from a supplier to the carrier, arrives one or two days late in that same week, this does not have a high influence on the weekly-based report since there is some flexibility between the arrival of the cargo at the port of loading and departure of the vessel.

The report should provide an approximate prediction, considering the inherent uncertainties, and cannot guarantee 100% accuracy. Factors such as variability in lead time and differences in packaging types used by suppliers contribute to the unpredictability. It is important for carriers to understand that the forecast serves as a volume indication and not a booking request. Regular updates to the forecast, preferably on a daily basis, are necessary. The report can be shared through automated email or a web-based tool with appropriate access rights. This statement can be summarized in the following research goal:

*Develop a **high-quality carrier shipment forecast**, by utilizing the **order information** of suppliers and the **expected packaging information**. This forecast determines the **total volume** (expressed in cubic meter, which can be converted to the number of containers) that carriers need to transport. The time interval for the forecast is set at **10 weeks**, meaning a forecast is produced for one week ahead up to and including 10 weeks ahead. The forecast is used to provide insights and **pre-book** the **required transport capacity** at the carriers in advance, such that the freight shipped on **spot market rates is minimized**.*

1.5 Research approach

To achieve this research goal, the main research question and several sub-questions have been formulated. The DMAIC (Define, Measure, Analyze, Improve & Control) method is used as the process model throughout the research. This method, consisting of five phases, facilitates process improvement and aligns with the Lean Six Sigma managerial approach (Theisens, 2016). These phases guide the research, dividing it into distinct chapters. By answering these questions, we address the main research question, ultimately solving the core problem.

1.5.1 Research questions

The main research question is described as follows:

Main research question
In which way can the order information of suppliers be translated to provide a 10-week ahead carrier shipment forecast, that provides insight into the required transport capacity of the carriers?

The DMAIC process model serves as a framework for this research, allowing for the formulation of corresponding research questions specific to each phase. The research approach outlines the actions undertaken to address these questions and complete each phase of the DMAIC method. Below, the main questions, along with their corresponding sub-questions and approaches, are provided:

Define

What is the current way of forecasting and its performance?

1. *What information is currently available to create a high-quality carrier shipment forecast?*
2. *What are the requirements for the forecasting model?*

SLN currently utilizes an ERP system to store pertinent information regarding parts ordered from suppliers, as discussed in the Research motivation. The initial step involves creating an overview of the available information and assessing its relevance for further research. This phase also entails a detailed analysis of the elements of uncertainty. The existing forecasting approach is defined, analyzed, and the requirements for the model are determined.

Measure

Which possibilities are there to improve the carrier forecast?

3. *Which methodologies are most appropriate to forecast the capacity, expressed in cubic meters, for the carriers?*
4. *How should the elements (packaging information, lead time variability & frozen period) be taken into account in this carrier forecast?*

To identify the most suitable methodology for developing the carrier forecast, a comprehensive literature study is conducted. This study focuses on finding methods that effectively combine order information with relevant uncertainty elements for creating the carrier forecast. Additionally, the methodology addresses strategies for managing uncertainty elements in the future.

Analyse

What are the improvements for SLN by using this suggested methodology?

5. *How can this suggested methodology be transformed to the situation of SLN?*
6. *What is the performance of this suggested methodology?*

Based on the insights gained from the literature study, the selected methodology must be adapted to fit the specific context of SLN. This adaptation is achieved through the development of a forecasting design that outlines where the different models are applied. Following the necessary data preparations, the performance of the proposed methodology is evaluated in the subsequent chapter. The most effective methods are then employed in the final forecasting design, which undergoes test runs.

Improve

In what way can this supposed carrier shipment forecast successfully be implemented?

7. *How should the process look that creates a carrier shipment forecast, as automated as possible?*
8. *The first prototype; how does it work?*

Upon identifying the optimal forecasting model during the analysis phase, a process is devised to outline how information should be generated, combined, and updated to create the carrier shipment forecast. This process also addresses the transformation of information to produce the forecast. In the latter stages of the improvement phase the operation of the initial prototype is discussed, along with the subsequent steps required for its integration into SLN.

Control

Which actions have to take place to maintain the high-quality carrier shipment forecast?

9. *What processes should be designed and implemented within SLN?*
10. *At which interval does this carrier shipment forecast have to be reviewed again?*

This phase of the research focuses on defining the actions necessary to monitor and anchor the forecasting process. Once the anticipated carrier shipment forecast has been successfully implemented, measures are implemented to ensure its continued reliability and accuracy in the future. Additionally, decisions need to be made regarding the frequency at which the forecast should be reviewed.

1.5.2 Scope of the research

To ensure the feasibility of this research, a well-defined scope is necessary. The research project is expected to span approximately 20 weeks, with each week consisting of 42 working hours, totaling 840 hours. However, the actual duration may be affected by holiday periods and public holidays. The objective of this research is to develop a carrier shipment volume forecast, with the prototype version being build.

As mentioned in the Background Information Section, there are various methods for transporting parts from overseas countries to the warehouse facility in the Netherlands. Since sea freight represents the majority of shipments and air freight is limited to specific suppliers or used for rush orders, this research solely focuses on sea freight. The carrier shipment forecast should encompass a time span ranging from one week ahead up to and including 10 weeks ahead. The volumes within this carrier shipment forecast are expressed in cubic meters (further referred to in this thesis as m^3) on a weekly basis, allowing for translation into container space in terms of twenty-foot equivalent units (TEU).

Material Supply Engineering (MSE) is the name of the team where this research takes place. The key stakeholders benefiting from this research are primarily the carrier and the MSE and transport planning teams within SLN. By having advance knowledge of the shipment volume to be transported, both SLN and the carriers can engage in pre-booking activities, enabling them to reserve container space in TEUs in advance. This, in turn, leads to a more stable workload internally (SLN) and externally (carriers).
Scania operates multiple warehousing facilities in the Netherlands, with the majority located in Hasselt, approximately 18 km away from the production facility in Zwolle. Parts are transported from this facility to production facilities within and outside the Netherlands. For the purpose of this research, the focus is solely on the flow of parts from overseas to the warehouse facility in Hasselt (NBF/import).

2 Current system analysis & future requirements

In this chapter, we examine the current system implemented in SLN. Initially, in Section 0, we define the current available information for generating the carrier shipment forecast. We consolidate this information in Section 0, presenting it as a table of relationships that serves as an input data for the forecast. Moving on to Section 2.2, we explain the current forecasting and performance of SLN. Finally, in Section 2.3, we discuss the model requirements. The conclusion of this chapter, presented in Section 2.4, answers the following sub-questions:

Define

What is the current way of forecasting and its performance?

1. *What information is currently available to create a high-quality carrier shipment forecast?*
2. *What are the requirements for the forecasting model?*

2.1 Available information

To develop a high-quality carrier shipment forecast, we utilize the currently accessible data. SLN employs an ERP system which stores relevant information about parts on order from suppliers. These orders are placed based on the production schedule, which is planned up to a year in advance. However, there is currently no TMS designed to support SLN's logistics operations and match the order information with the associated uncertainties.

2.1.1 Order information

In this study, our aim is to create a carrier forecast that offers insights into the required transport capacity of the carriers. The most important data to gather is information about the parts currently on order from suppliers. This information resides in the ERP system and can be accessed through queries. We can assume the reliability of this data as it is based on parts scheduled for production. However, when importing the data, it is essential to verify its correctness and completeness. Otherwise, the forecast's accuracy may be compromised, or it may not be possible to create the forecast altogether. To ensure automation of the processes to the greatest extent, it is vital to document the steps involved in cleaning the data. The data needs to be modified to suit our requirements, such as generating a weekly forecast by considering the weeks as a time series. Each line of the order information query corresponds to a specific part of an order (multiple lines may be from the same order). Each line comprises 27 columns of information, with Table 1 summarizing the most pertinent details.

Table 1: Important variables within the ERP system regarding the parts on order

Column name	***Description of the column***
Pick-up date	*Planned shipment date by the supplier*
Planned delivery date	*Planned delivery date at Scania (Hasselt)*
Quantity on order	*The quantity on order for that specific part*
Status of the order	*Scheduled line = order information is sent to the supplier* *Shipment Notification = confirmation that the supplier has sent the parts*
Supplier ID	*ID number of the supplier that is used internally*
Supplier name	*Full name of the supplier (should not be displayed in the final forecast)*
Country of supplier	*Country where the supplier production facility is located*

2.1.2 Table of relations

It is important to customize the forecast for each carrier by specifying the assigned supplier and country of origin. Typically, a specific carrier serves all the suppliers in the same country. This relationship (assigning a carrier to specific supplier) is considered as a categorical variable as it indicates the category to which a specific supplier belongs. To establish this table of relations for each supplier, historical data has been examined and certain assumptions have been made. In terms of historical data, we analyzed all the deliveries from the previous year (2022) and compiled an overview of the suppliers' port usage, assigned forwarder, and the quantities of FCL and LCL shipments. Since a carrier can serve multiple ports, this table of relations provides insights into the suppliers from which they receive shipments.

The following assumptions were made in creating this table of relations:

- For suppliers without available historical data but located in a specific country, a classification for port and forwarder was determined based on other suppliers in the same country.
- Some suppliers/countries do not have a primary port and/or forwarder and are classified as spot market. As shipments in the spot market can originate from any port, these suppliers/countries are combined in a separate report.

2.1.3 Packaging information

One of the main uncertainties in this research pertains to the packaging used for the NBF. Fortunately, for the majority of flows (60%), we have information about the packaging because parts from these countries/suppliers are packed in Scania packaging. In such cases, we know the packaging type, the number of parts accommodated in one package, and the package volume. For instance, for part X, we know that 10 pieces can fit on a Euro pallet with four collars (raised edges). This Euro pallet has a volume of $0.475\ \mathrm{m}^3$. If we need to ship 100 pieces of this part, we can expect 10 Euro pallets with a total volume of $10 \times 0.475 = 4.75\ m^3$. The previously discussed table of relations also includes information about which suppliers the Scania packaging is applicable to.

However, we also need to consider the suppliers that do not use Scania packaging. Unfortunately, there is currently no available data on the packaging used by these suppliers, as Scania does not record how they receive parts from suppliers. Requesting this data from suppliers would be excessively time-consuming for this research, as it involves approximately 750 different parts from 65 suppliers. Therefore, an alternative method is devised to provide an indication of the packaging volumes from a specific supplier as accurately as possible. The data used to develop this method is sourced from selected carriers and covers the period from the middle of 2019 to 2022. This dataset includes all the shipment information handled by these carriers, including shipment volumes. The utilization of this data within the research are discussed in Section 4.3.

2.1.4 Lead time variability

The next uncertainty element for this research is supplier reliability. “Supplier reliability is simply defined as the ability of a company to consistently supply an acceptable product at the required time” (National Research Council, 1995). Based on the experience of Material Supply Engineers at SLN, there are no major discrepancies between scheduled and actual deliveries, as evident from the deviations in delivery dates. This variability in lead time is crucial for the forecast because it strongly influences the shipment volume that needs to be transported and reserved on the vessel.

To assess the variability in lead time, historical shipment data was analyzed. This involved comparing the planned shipment date with the actual shipment date for various suppliers. The time differences were converted into weeks, allowing us to determine the percentage of shipments that were sent before, during, or after the planned week. However, it is not possible to perform a comparable analysis for deviations in the quantity of parts.

During the course of this research (starting from January 2023), SLN was implementing a new module called the ACS Vendor Evaluation Module, which automatically assesses the "score" of a supplier. This score is based on the supplier's on-time delivery performance and adherence to the ordered quantities. Both criteria carry equal weight and assign a score between 0 and 100 to the supplier. However, this module does not utilize historical data, and since only data from January (upon its release) is available, it cannot be used for this research. Nevertheless, it holds potential for future forecast improvements.

2.1.5 Frozen period

The frozen period at the supplier, which is the duration before the delivery date when SLN cannot make any changes to the order, such as modifying specifications or quantities, must also be taken into account for the carrier shipment forecast. All suppliers adhere to a consistent duration for this frozen period, which amounts to two weeks, and should be recorded in the table of relations. More about the frozen period and the application of this factor to the forecast, are discussed in Section 4.2. It can be asserted that the forecast accuracy improves after the frozen period has elapsed.

2.2 Current forecasting & performance

Currently, SLN does not perform extensive forecasting activities for carriers. The existing forecast used by the carriers to determine vessel capacity is based on historical shipment volumes from previous years. In the case of a new contract, SLN provides the total number of FCL and/or LCL shipments from the previous year. The carrier then calculates the weekly capacity reservation by dividing the annual number of FCL and/or LCL shipments by the number of weeks. In the case of contract extensions, the carrier uses historical data from the previous year to determine volumes for the upcoming year, with the previous year's container count serving as the forecast for the current year.

To evaluate the performance of the current forecast, the available data is limited. An analysis can only be conducted for a specific flow (from port to port) where the agreed-upon number of FCL containers annually is compared to the actual number of containers shipped for that flow. The performance measures for this specific flow's current forecasting approach are summarized in Table 2. The performance metrics spanning from 2019 to 2022 demonstrate fluctuating performance throughout the years. The differences between the number of shipped containers and the contracted amount have increased (MSE) in 2020 and 2021, indicating greater variability in measurements and a higher percentage error compared to the actual container count (MAPE). The rise in the number of containers can be attributed to the growing emphasis on maintaining higher stock levels. Based on these findings, it is not possible to definitively determine whether all the additional containers were booked at spot-market prices. The available information does not provide details on whether SLN paid the full spot-market price, received a financial discount, or adhered to the contracted price.

Table 2: Difference between contracted and actual containers

Year	% Difference between actual vs contracted # of containers	MSE	MAD	MAPE
2019	11%	33	4,8785	60%
2020	125%	229	8,0432	148%
2021	106%	181	6,2722	126%
2022	35%	254	9,1903	111%
Average	**69%**	**174**	**7,0960**	**111%**

The existing forecast is based on an annual timeframe, with the carrier disaggregates this forecast over the weeks. However, this forecast overlooks crucial factors such as Scania's production levels on a weekly basis, packaging volumes, lead time variability, and the frozen period specific to each supplier. Consequently, relying on historical volumes from previous years to determine shipment capacity results in significant deviations between the expected and actual number of container shipments. As a result, the current forecast lacks accuracy, making it an unviable approach for the future. This underscores the importance of shipment volume forecasting, highlighting the significance of this research. A forecast that incorporates order information and effectively accounts for the associated uncertainties is expected to yield substantially improved performance.

2.3 Model requirements

The model's requirements are closely aligned with the research scope and forecast needs. The forecast should be **weekly based**, with a time interval of 10 weeks. This entails generating forecasts for one week ahead up to and including 10 weeks ahead. The process should be highly **automated** so that any changes in orders (excluding the frozen period) are reflected in the forecast. Given the potential daily occurrence of these changes, the forecast should be **updated at least once a week**, ideally on a daily basis. Accuracy is expected to improve as the forecast approaches the actual shipment arrival or when the frozen period has expired.

The forecast should provide aggregated information at the **carrier** and **supplier levels**. If no specific carrier is assigned, these should be consolidated into a single forecast. The forecast should indicate the total volume, in m^3, for each port of loading and supplier, encompassing all parts expected to be received from a particular supplier in a given week. This volume indication should also be **translated** into the **number** of **TEUs**. Considering that a carrier may serve multiple ports, understanding the expected shipment volumes at different ports is essential due to varying capacity requirements across regions.

A requirement for the forecast is thorough verification of input and output data. Adhering to the principle of "garbage in is garbage out," it is crucial that the input data and variables are accurate and correctly formatted to ensure the model's performance. Similarly, the accuracy of the model's output should be tested and validated to ensure reliability.

The forecast results need to be incorporated into a **weekly** report for internal use within SLN and to be shared with assigned carriers. These reports serve as the basis for decision-making, enabling carriers to reserve adequate vessel space. Microsoft Power BI is utilized to develop and share these reports, leveraging information from the ERP system and relevant variables for forecast creation. The report creation and sharing process should be as automated as possible, minimizing manual effort. Given the confidentiality of most information in these reports, careful consideration must be given to ensure secure sharing. Additionally, to maintain anonymity, Scania prefers not to include the full supplier names in the forecast, suggesting the use of suppliers IDs or postal codes to anonymize the data. In cases where manual updating procedures are necessary, knowledge sharing on the process should be provided, and the report should be user-friendly, potentially including instructions. A summary of the requirements is provided in Table 3.

Table 3: Summary of the model requirements

Model requirements
✓ *Weekly-based, time interval of 10 weeks*
✓ *As automated as possible*
✓ *Updated at least each week, but preferably daily*
✓ *Aggregated on port and supplier level*
✓ *Input and output are verified*
✓ *Possible to share with interesting parties in a secure way*
✓ *Forecast should be user-friendly and if needed contain instructions*

2.4 Conclusion

In conclusion, the development of a high-quality carrier shipment forecast for SLN requires careful consideration of available information, current forecasting practices, and specific model requirements. The utilization of the ERP system's order information, including parts on order from suppliers, forms the basis of the forecast. Customizing the forecast for each carrier involves establishing a table of relations that categorizes suppliers based on historical data and their assigned ports and forwarders. Packaging information is another important factor, although the availability of data is limited to parts packed in Scania packaging. To accurately estimate packaging volumes for suppliers using alternative packaging, we are devising an alternative method based on selected carriers' data.

The current forecasting approach based on historical shipment volumes has shown a decline in performance over the years. It overlooks crucial factors such as production levels, packaging volumes, lead time variability, and the frozen period. The model requirements emphasize a weekly forecast with a 10-week time interval, automated processes for reflecting order changes, and regular updates to ensure accuracy. The forecast should provide aggregated information at the carrier and supplier levels, indicating total volumes in m^3 and translated into TEUs. Thorough verification of input and output data is essential, and the forecast results should be incorporated into automated weekly reports shared within SLN and with carriers. Secure sharing of confidential information, anonymization of supplier names, and user-friendly report generation are additional requirements. Overall, by addressing these requirements and incorporating the necessary information and processes, the developed carrier shipment forecast has the potential to greatly enhance SLN's logistics operations and decision-making processes.

3 Literature study

In this chapter, we explore relevant theories that are pertinent to our research. In Section 3.1, we delve into the literature requirements for the necessary models. Subsequently, in Section 3.2, we provide a comprehensive overview of freight shipment and highlight the advantages of information sharing within supply chains. Moving forward to Section 3.3, we examine applicable forecasting methods for our research, including those that have demonstrated success in predicting shipment volumes. Lastly, in Section 3.4, we analyze how the elements (of uncertainty) can be effectively incorporated into the forecasting methods. In 3.6 we introduce the Newsvendor problem. The conclusion of this literature study, presented in Section 3.6, answers the following subquestions:

Measure

Which possibilities are there to improve the carrier forecast?

3. *Which methodologies are most appropriate to forecast the capacity, expressed in cubic meters, for the carriers?*
4. *How should the elements (packaging information, lead time variability & frozen period) be taken into account in this carrier forecast?*

By addressing these sub-questions, we aim to consolidate our understanding of the existing literature, establish a foundation for our research, and lay the groundwork for further exploration in subsequent chapters.

3.1 Literature model requirements

Prior to embarking on the literature research, it is important to define the specific objectives and criteria for the literature search. As outlined in Section 2.3, our focus is to determine weekly shipment volumes aggregated at the carrier and port levels. To accomplish this, we require information on the types and quantities of parts, as well as the corresponding packaging volumes expected to arrive at specific ports. Given the presence of stochastic elements in this study, arising from unknown packaging quantities from certain suppliers and lead-time variability, the literature review should identify methods capable of predicting these variables, enabling their integration into a unified carrier forecast.

3.2 General

Despite the impact of the Covid-19 pandemic, which resulted in decreased demand for sea freight, the past decade has witnessed a significant growth in sea freight transportation. This growth can be attributed to global trade and its interconnection with supply chains, which exhibit a strong correlation with economic expansion (MacroMicro, 2022). As noted by Gorman et al. (2014), sea freight operations are resource-intensive, involving assets, labor, and fuel. The management of variable and fixed costs within a network-based operational structure presents considerable complexities, necessitating close coordination among supply chain partners to optimize overall performance and distribute realized returns among the stakeholders (Gorman, et al., 2014).

Information sharing in supply chains

A basic enabler for tight coordination is information sharing, which has been greatly facilitated by the advances in information technology. Lee and Whang (2000) describes the types of information that can be shared within supply chains and underscore their relevance, including inventory data, sales information, demand forecasts, order statuses, and production schedules. Additionally, Zang et al. (2006) assess the value of shipment information sharing (SIS), also known as advanced shipping notification, in which a stage within the supply chain communicates shipment quantity information to its downstream customers. This practice is particularly valuable as suppliers may not always be able to fulfill consumer orders precisely on time due to service imperfections, thus underscoring the importance of sharing such information with customers.

One specific aspect of information sharing pertains to shipment quantities, where the actual quantities received by the customer after a given lead time may differ from their expectations. The prior knowledge of this information allows the customer sufficient time to adapt and address this uncertainty by adjusting their future ordering decisions. Zang et al. (2006) affirm that sharing shipment quantity information proves effective in mitigating this uncertainty within the supply chain.

According to Lee and Whang (2000), a producer can leverage its supplier's delivery schedule to enhance its own production schedule. However, this does not apply to SLN as they establish their own production schedule, assuming that suppliers are capable of meeting the demand. In order to assist participants in making informed production and inventory decisions, we describe SIS as follows: within a supply chain network, one stage shares its shipment information (time and quantity) with its immediate downstream stage. Zang et al. (2006) conducted a case study comparing a wholesaler that implemented SIS with its downstream customers against a scenario without information sharing (NSIS). The results demonstrate that SIS enables downstream customers to effectively adapt and address shipment uncertainty, whereas the adjustment process is significantly slower in the NSIS environment. The more uncertain the shipments, the greater the potential benefit an organization can derive from implementing SIS.

3.3 Forecasting methods

When it comes to determining shipment volumes using forecasting methods, there are various approaches available. These methodologies are not exclusively designed for sea freight but have been applied to other modes of transportation (such as train or air freight) or different sectors (like retail). In this discussion, we provide a brief overview of these methods and their past applications. However, it is important to note that none of these methods have been employed specifically for actual orders, meaning they have not been used to determine volumes based on scheduled orders. This study aims to differentiate itself from previous research by focusing on forecasting methods specifically applicable to shipments for which the packaging dimensions are unknown.

3.3.1 Traditional forecasting techniques

Time series models

A time series consists of observations collected at different points in time (Brockwell & Davis, 2016). Time series models analyze forecasted variables based on historical data patterns. Among the various methods within time series forecasting, one of the simplest approaches is the moving average. It involves summing up all observations in a collection and dividing the sum by the total number of observations. As new data is added, the oldest observation is removed from the average calculation. This method is relatively basic as it does not consider weights, trends, or cyclical information (Archer, 1980)

Exponential smoothing is another well-known method for time series forecasting, which takes into account elements such as trends and seasonality. Over time, these techniques have evolved to incorporate more sophisticated patterns. There are a total of 15 different models, including popular ones like simple exponential smoothing (SES) without trend or seasonality and Holt-Winters' additive method with additive trend and multiplicative seasonality factors. Each model calculates a stable demand level and, depending on the method, incorporates trend and seasonality. A trend indicates a consistent increase in demand over time, while seasonality accounts for variations in demand during specific periods, such as every January. The selection of a model depends on the current demand pattern's trend and seasonality (Hyndman, et al., 2002).

ARIMA
The Auto Regressive Integrated Moving Average (ARIMA) is a method that is based on past values (auto regression) and forecast errors (moving average). The model is characterized by three terms:

- p is the number of auto regression terms;
- d is the number of differences required to make this time series stationary; and
- q is the number of moving averages.

Stationary time series is necessary for ARIMA models because they use linear regression, which performs best when predictors are not correlated. A stepwise method can be employed to search through various permutations of the input parameters (p, d, and q) to select the model that performs the best based on a chosen accuracy measure (Fattah, et al., 2018). Seasonality can also be incorporated into the ARIMA model by including additional seasonal terms (P, D, Q) that involve a backshift of the seasonal period (Ramshorst, 2022)

Causal models
While time series and ARIMA models focus on the variables being forecasted, causal models analyze the relationship between an explanatory variable (e.g., part number and supplier) and the variable of interest (e.g., packaging volume), also known as the response variable (Archer, 1980). Causal models predict the response variable based on these explanatory variables. Linear regression is a simple method for predicting a response using one or more predictor variables (Ramshorst, 2022). It finds the best linear relationship through the least squares approach, which can then be utilized for prediction. Many statistical learning approaches can be considered as generalizations or extensions of linear regression (James G., et al., 2021).

The M-competition
The objective of forecasting competitions is to empirically assess the effectiveness of both new and existing forecasting methodologies, allowing for experimentation similar to that in the hard sciences (Hyndman R. , 2020) (Makridakis, et al., 2020). The M competition is the most influential and widely referenced competition in the field of forecasting (Makridakis, et al., 1982-2020), with the most recent being the M5 competition held from March to June in 2020 (Makridakis, et al., 2022). Compared to previous M competitions, M5 extends the results of the previous M competitions by: (1) significantly increasing the range of participating methods, particularly those falling under the Machine Learning (ML) category, (2) evaluating the performance of uncertainty distribution as well as point forecast precision, (3) incorporating time series data along with exogenous/explanatory variables, (4) utilizing correlated, grouped time series, and (5) focusing on series demonstrating intermittency (Makridakis, et al., 2022).

In (Hyndman, et al., 2002), an algorithm is proposed to automatically select the most suitable model from the exponential smoothing family of models for each series at the product-store level. The M5 competition utilizes this algorithm as a baseline for comparing other forecasting methods, and this algorithm has already performed exceptionally well; 92.5% of the participating teams in the M5 competition failed to surpass this algorithm, particularly when producing forecasts at the product or product-store level (Makridakis, et al., 2022). Consequently, the baseline against which additional forecasting techniques must compete is already fairly high.

The M5 competition demonstrated the superiority of ML techniques, especially LightGBM, based on the aggregated demand. The top 50 methods outperformed the most accurate statistical benchmark by more than 14 percent, while the top five methods outperformed it by over 20 percent. Additionally, the M5 competition, like previous M competitions, revealed that combining models enhances forecasting accuracy. There is also a shift in the performance of methods, as simple statistical methods were found to be more accurate than complex and sophisticated methods in M1, M2, and M3. In M4, only two sophisticated methods were more accurate than simple statistical methods, whereas in the M5 competition, all 50 top-performing methods were based on ML (Makridakis, et al., 2022).

3.3.2 Machine learning techniques

As discussed in Section 3.3.1 within the M5 competition it can be seen that nowadays ML is becoming a more important forecasting method. ML is not only becoming famous for forecasting approaches, ML is nowadays widely used in domains such as business, healthcare, industries and military (Aggarwal, et al., 2022). With ML you can develop a system that automatically learns from (new) data (Domingos, 2012). Where you normally need to create manually a system or method, an ML is trained with an existing data set. It can be seen that the an ML model learns from itself. When the ML model has been finalized, it can perform learned tasks on new data. A set of relevant input variables and example instances are needed to enable learning from data. These input variables are called features (Koch, 2018). For ML to be effective, the amount of the data needed to train the model must be sufficient. As a rule of thumb regarding machine learning is that you need at least ten times as many rows (data points) as there are features (columns) in your dataset (Smolic, 2022).

After the development of the ML method, there are various tasks that can be done by the ML method where we only focus on one of them, *supervised learning*. The supervised learning approach is based on the premise that a teacher or supervisor is available who organizes the training instances into classes and makes use of the data on the class membership of each training instance (Sathya & Abraham, 2013). The most common application of ML in the industry is *Data Mining* (DM). DM describes the applied discovery of knowledge within databases and includes the process of data understanding, data preparation, modeling, evaluation, and implementation (Knolla, Prüglmeierb, & Reinharta, 2016).

An example of a paper where ML has been used as a predictive method is the paper of (Knolla, Prüglmeierb, & Reinharta, 2016). In this study, a method for predictive inbound logistics planning is presented. ML can be used to extract general knowledge from logistical procedures and utilize that knowledge to forecast future events. Such as, providing insight into the packaging dimensions, and then the shipment volume can be determined. Also (Koch, 2018) applied a similar approach of ML in the air cargo supply chain using *Random Forest*. This ensemble learning approach is utilized for both regression and classification. It constructs various decision trees that combine to produce a solution.

According to the M5 competition, several ML methods excelled in this challenge. For this research we chose the top three performing ML methods and explain the functioning behind it. Those three ML methods are LightGBM, XGBoost and N-Beats, where LightGBM and XGBoost are both based on Gradient Boosting. Gradient Boosting is a machine learning algorithm which builds 'ensembles' by sequentially training weak models, combining them together and improving loss functions to achieve strong predictive power. It uses a combination of weak learners (usually decision trees) and combines them to create a strong one. A weak learner is a machine learning model that is only slightly better than random guessing. However, when weak learners are combined, they can form a strong learner that is much more accurate. This can be used for both regression and classification problems (Saha, 2023). The distinction between boosting, where each new predictor learns from the mistakes of the preceding forecasters, and bagging, where samples are obtained with replacement, is depicted in Figure 3.

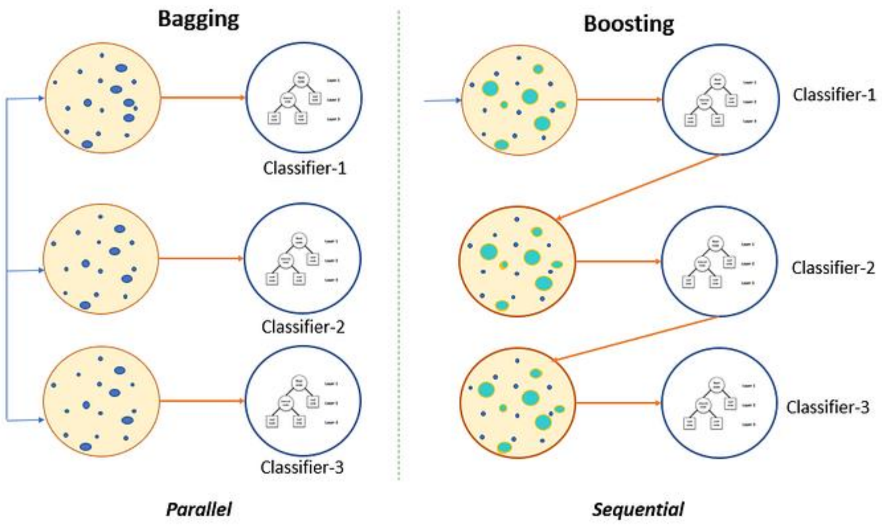


Figure 3: Bagging (independent predictors) vs. Boosting (sequential predictors) (Pal, 2020)

Tree-based methods

Two of the selected ML methods (XGBoost and LightGBM) are tree-based methods, where each of them do have a specific specialty. Tree-based methods are a type of supervised ML methods used for both classification and regression tasks. These algorithms create decision trees or ensembles of decision trees to make predictions based on the input features. The basic idea behind tree-based methods is to recursively split the data based on the most informative feature until a stopping criterion is reached. Each node in the decision tree represents a splitting point based on a specific feature, and the decision tree is constructed by choosing the feature that provides the highest information gain or the lowest impurity at each node, which is a is a measure of how "mixed" the classes or target variable values are in a set of data (James G. et al., 2021).

In regression tasks, tree-based methods construct decision trees that predict a continuous target variable. The final prediction is made by traversing the decision tree based on the values of the input features until a leaf node is reached, which provides the predicted output. Tree-based methods have several advantages, including their ability to handle both numerical and categorical data, their interpretability, and their ability to capture complex non-linear relationships between the input features and the target variable. However, they may suffer from overfitting if not properly regularized and may not perform well when the data is noisy or contains missing values (James G. et al., 2021).

XGBoost

Extreme Gradient Boosting, or XGBoost, is a machine learning technique that prioritizes model performance and computing speed. It has been created to operate with significant and complex datasets. In order to create a powerful learner, XGBoost combines a number of weak learners. Training a variety of decision trees is how XGBoost operates. A subset of the data is used to train each tree, and the predictions from each tree are then combined to get the final prediction.

In this Section, we describe the steps for the XGBoost technique in more detail. The **first step** (1) of the algorithm is make an initial prediction and calculate the residuals between the observations and the prediction. This prediction can be anything, but for now we assume our initial prediction is the average value of the variables we want to predict, which is 70. The **second step** (2) is to build an XGBoost tree, where each tree start with a single leaf and all the residuals (observation – initial prediction) go into that leaf. For this leaf, we need to calculate somtheing that is called a Similarity Score (see Equation 1), where λ (lambda) is a regularization parameter, which means that it is intended to reduce the predictions sensitivity to individual observations. Also known as preventing to overfit the training data. The default value of λ is 1.

Equation 1

$$Similarity\ Score = \frac{(\sum Residuals)^2}{Number\ of\ Residuals + \lambda}$$

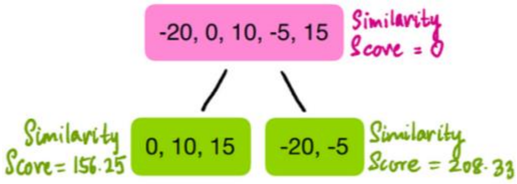


Figure 4: Example of splitting and calculating similarity scores (Rao, 2021). For example, the node on the bottom left: (0+10+15)^2 / 3+1 = 156.25

After all the residuals has been placed into that leaf, it is time to cluster the residuals if we split them into two groups using the features. In this research that can be for example the supplier or the number of parts. Splitting the residuals basically means that we are adding branches to our tree. We then calculate the Similarity Scores for the left and right leaves of the above split, see Figure 4 for an example. Now we do have the Similarity Scores of the top leave and the left and right leaves of the split, the next step it to quantify how much better the leaves cluster similar residuals than the root does. We can do this by calculating the gain (Equation 2) of splitting the residuals into two groups. If the gain is positive, then it is a good idea to split, otherwise, it is not.

Equation 2

$$Gain = Left_{Similariy} + Right_{Similarity} - Root_{Similariy}$$

We compare this gain to those of the splits of the other features. In the case this other features is a continuous variable, the process to find the different splits is a little more complicated. First, we arrange the rows of our dataset according to the ascending order. Then we calculate the average values of the adjacent values. Now we split the residuals using the averages as thresholds and calculate gain for each of the splits. We continue this process up to all features has been split and determine the split with the greatest gain value, which we use as our initial split. Now we can add more branches to the tree by splitting our initial split again using the same process. Only now, we use the initial leave as our root node and try splitting them by the greatest gain value that is greater than 0. Here, it is important to note that observations are only added to a node when they land in that node.

Pruning the tree, removing of branches, is the **third step** (3) of this ML method, where we avoid overfitting the data. We do this by starting from the bottom of the tree with a given treshold γ (gamma) and work our way up to see if the gain is sufficient. The default value of γ is 0, so only positive gains are kept in the tree. The **fourth step** (4) is to calculate the output values of the leaves, because a leaf can now contain multiple residuals. This is similar to the formula to calculate the Similarity Score except we are not squaring the residuals. Doing this fourth step where the final tree is ready, we can continue to the **fifth step** (5) and make a prediction. For this prediction we use Equation 3, where the learning rate ϵ (epsilon) is importance of the constructed tree for the new prediction which is by default 0.3, because then we do consider the previous tree.

Equation 3

$$Prediction = Initial\ prediction + \ \varepsilon\ \times Final\ tree$$

In the **sixt step** (6) we calculate the residual values for our new predictions. You probably see that the residuals are smaller than the residuals from step 1. As we repeat this process (from step 2 to

6), our residuals get smaller and smaller indicating that our predicted values are getting closer to the observed values. We do this until the residuals are super small or we reached the maximum number of iterations we set for our algorithm. If the tree we built at each iteration is indicated by T_i, where i is the current iteration, then the formula to calculate predictions can be found in Equation 4 (Starmer, 2020) (Rao, 2021).

Equation 4

$$Initial\ prediction + \varepsilon T_1 + \varepsilon T_2 + \varepsilon T_3 + \cdots + \varepsilon T_i$$

LightGBM

At the M5 competition, where it was embraced by all top 50 contestants, LightGBM emerged as the undisputed winner. The M5 competition demonstrates how retail businesses may make better use of this technology to enhance the accuracy of their everyday activities and sales projections. Although the model operates similarly to the XGBoost model, there is a significant difference.

When building the trees, the XGBoost model splits the tree depth wise or level wise. Which means that in the steps described of the XGBoost model, the model jumps directly from the left leaf in a split to the right leaf. Whereas, the LightGBM model splits the tree leaf wise with the best fit, so it continues first in the left leaf until no further splits are applicable or it has reached its maximum depth of the tree. The biggest advantage of this way of working is that it no longer invests in unnecessary nodes and leaves that do not contribute to a better tree. The LightGBM structure continues to grow with the most promising branches and leaves (nodes with the biggest gain), holding the number of decision leaves constant. Moreover, it moves extremely quickly, hence the name "Light" (Khandelwal, Analytics Vidhya, 2020). Figure 5 shows the level-wise tree growth in XGBoost and Figure 6 the leaf-wise tree growth in Light GBM.

The lightGBM model also uses historical binning of continuous features, which is a technique for reducing the cardinality of continuous data (Oracle, 2023). The use of binning numerical values decreases significantly the number of split points to consider in decision trees, and they remove the need to use sorting algorithms, which are always computation-heavy. Therefore, LightGBM is generally speaking faster than traditional gradient boosting, such as XGBoost (Tuychiev, 2021).

Figure 5: Level-wise tree growth in XGBoost (Khandelwal, 2020)

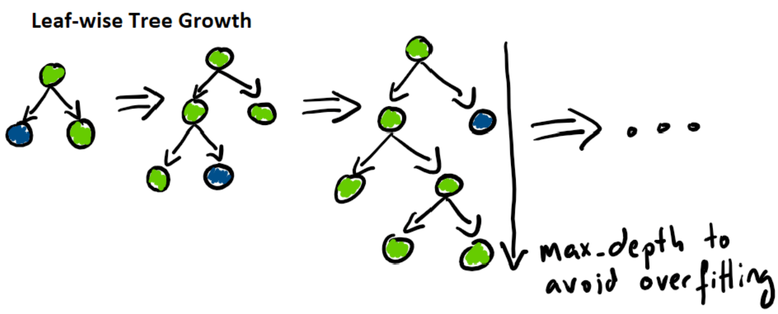


Figure 6: Leaf-wise tree growth in Light GBM (Khandelwal, 2020)

N-BEATS

A method that also performed quite well, was the one from N-BEATS, which is a deep-learning neural networks (NNs) method for time series forecasting (Onnen, 2022). N-BEATS is a specialized architecture designed for interpretable time series forecasting, while a neural network is a broader term referring to various architectures used in machine learning. N-BEATS emphasizes interpretability through the use of basis functions, while neural networks offer more flexibility for different tasks but may lack inherent interpretability. A neural network consists of multiple neurons, which can usually be read from left to right, where the input variables are added on the left cite of the network and the possible output variables are on the right side of the network. The term "deep learning" comes from neural networks that contains several hidden layers, also called "deep neural networks" .

The operations done by a neuron are pretty simple to understand. First, it adds up the variables from the previous neuron (layer) where it is connected to. For example in Figure 7, we see that there are 3 inputs (x_1, x_2, x_3) coming to the neuron. These input variables are multiplied by another variable called the weight (w_1, w_2, w_3), which determines which controls how the two neurons are connected. Each neuronal connection has a unique weight, and only those numbers change as a result of the learning process. The calculated total number may also include a bias value. Although it is not a value that originates from a particular neuron and is selected prior to the learning phase, it might be useful for the network to include it. Finally, after performing all those summations, the neuron applies a function known as the "activation function" to the value it had obtained.

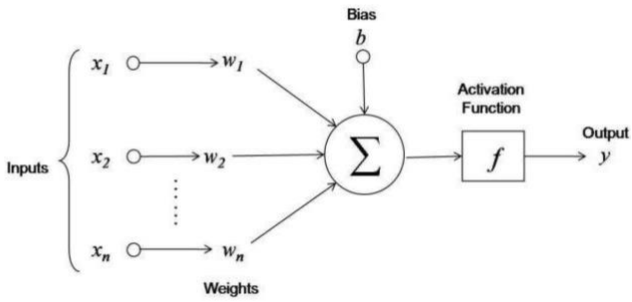


Figure 7: Operations done by a neuron (Arnx, 2019)

The activitation function usually translates the calculated value by the neurons to a number between 0 and 1, the sigmoid function (Equation 5) can be used for this. There are other functions that may alter the boundaries of our function while maintaining the same goal of restricting the value. In principle, that is all what a neuron does. It takes all the values from the connected neuron multiplied by their respective weight, add them, and apply an activation function. The neuron is then ready to communicate its updated value to other neurons.

Equation 5

$$sig(t) = \frac{1}{1+e^{-t}}$$

To be able to make some predictions, we need to learn the neural network how to respond to the input variables. First off, keep in mind that a neural network produces an outcome after receiving an input. Since it is unable to produce the desired result on its first attempt (barring extreme chance), during the learning phase each input comes with a label that indicates which possible result the neural network should have predicted. If the network predicted the correct label, the actual parameters are kept and the following input is given. If the output obtained fails to match the label, weights are altered. Only those parameters are possible to be altered during the learning phase. The learning rate, that we already briefly introduced for XGBoost, is the rate that controls the way the neural network learns. This value influences how quickly the neural network learns, or more specifically, whether it modifies a weight little by little or by bigger steps. A reasonable value for the learning rate is typically 1 (Arnx, 2019).

3.4 Incorporating elements

In the previous Section, we presented the general forecasting method that is relevant to this research. These methods have been selected based on the situation of SLN. As introduced in 1.4 and further explained from Section 2.1.3 up to 2.1.5, three elements in this research need some extra attention for applying to the final forecast. How to deal with them, is discussed in this Section.

3.4.1 Packaging uncertainty

As stated in Section 0 the packaging volume of approximately 60% of the NBF flows can be determined deterministically. For the remaining flows, we are dealing with stochastic packaging volumes and we have to find a method how to cope with this uncertainty. For the packaging uncertainty, we have looked again at the research from (Koch, 2018) and (Knolla, Prüglmeierb, & Reinharta, 2016). In both types of research, they are also dealing with packaging uncertainty to forecast the total shipment volume and in both, they used ML methods (random forest and data mining, which includes the process of data understanding, data preparation, modelling, evaluation and implementation) . Historical data regarding the packaging volumes should be analyzed, such that a prediction model can be created and can be applied to determine the packaging volumes from a specific supplier, part, and port of origin.

3.4.2 Lead time variability

As introduced in this chapter, there has been an increase the last decade in the number of firms that source parts from overseas. Although this has reduced procurements costs, it has increased supply chain risk; procurement lead times are longer and are often unreliable (Wang & Tomlin, 2009). It is therefore important to incorporate lead time variability in your forecast, which is the variability in the time (a batch of) parts has actual been shipped by the supplier. Wang & Tomlin (2009) describes a cumulative distribution function (CDF) and a probability distribution function (PDF) of the delay. Where L denotes the standard lead time and ω denote a stochastic, non-negative delay. Then, with probability θ, there is a delay and the lead time is $L + \omega$; with probability $1 - \theta$, there is no delay and the lead time is simply L. The model collapses to a constant lead time case when $\theta = 0$ and a pure stochastic lead time case when $\theta = 1$. Hereafter, we refer to θ as the delay probability and ω as the delay. Two commonly used measures are the standard deviation and the coefficient of variation (CV). The standard deviation provides a measure of the dispersion of lead times, while the CV takes into account the average lead time, allowing for better comparison between suppliers or product lines.

3.4.3 Frozen period

The frozen period is the period before the delivery date when SLN is not allowed to apply any changes to the order (from production planning to the supplier), such as changing specifications or the quantity of an order. “The frozen schedule provides some stability in the short term, as any short-term changes in the demand forecast get accumulated and then deferred until beyond the frozen period.” (Graves, 2011). A frozen period policy was implemented by many organizations to limit changes to later periods while allowing changes to occur earlier. By applying a frozen period, you limit the schedule nervousness in a manufacturing system (Pujawan, 2001).

Stadtler & Kilger (2000) states that organizations often plan according on a rolling horizon basis, which is that production/shipments etc. are actually set into practice when they fall inside the frozen planning period. For example, at the beginning op January a plan is made that covers

January to December. But only the first period, the so-called frozen planning period, is actually put into practice. Stadtler & Kilger suggests that organizations should work according an event-based planning method, where a new plan is not drawn up in regular intervals but in case of an important event, for example, unexpected sales, major changes in customer orders, breakdown of a machine, etc. This procedure requires that all data which are necessary for planning are updated continuously so they are available at any arbitrary event time.

3.5 Newsvendor problem

During the elaboration of this research we see overlap with the newsvendor problem. The newsvendor problem is a classic inventory optimization problem in operations research. It gets its name from the analogy of a news vendor who must decide how many newspapers to order for sale the next day, with uncertain demand. The newsvendor problem is characterized by a single-period decision-making scenario in which a seller must determine the optimal order quantity for a perishable product, such as newspapers, that has uncertain demand (Congzheng, et al., 2022).

The objective of the newsvendor problem is to find the order quantity that maximizes the expected profit or minimizes the expected cost. The challenge lies in balancing the trade-off between the potential profit from meeting demand and the cost of leftover inventory. To solve the newsvendor problem, various factors must be taken into account, such as the cost of ordering, the price of the product, the salvage value of unsold items, and the probability distribution of demand. Techniques like the cycle service level (CSL) are commonly used to determine the optimal order quantity that minimizes the expected cost or maximizes the expected profit (Congzheng, et al., 2022).

The newsvendor problem is widely applicable in various industries, beyond just newspapers, including retail, manufacturing, and supply chain management, where perishable or seasonal products are involved and demand is uncertain. The newsvendor problem is also applicable to the research conducted at SLN, as it involves the need to determine the quantity of containers to pre-book, despite the uncertainty surrounding the exact demand.

3.6 Conclusion

The objective of this chapter was to identify suitable methods for constructing a carrier forecast. The literature review revealed several forecasting techniques, but none of them were based on actual orders, and only a few specialized in predicting shipment volumes. A summarized overview in Table 4 highlights the industries in which these different methods have been applied. Recent research, particularly the M5 competition, demonstrated that ML methods are currently the most promising for forecasting. Notably, LightGBM, XGBOOST, and N-beats exhibited relatively low forecasting errors. It would be worthwhile to explore how these methods can be implemented in the context of SLN.

Table 4: Industries in which these different methods have been applied

	Shipment information sharing (SIS)	***Time series models***	***ARIMA***	***The M5-competition***	***ML***	***Random forest***
Wholesaler	*Zang et al. (2006)*					
Finance		*(Hyndman, et al., 2002)*				
Transport			(Ramshorst, 2022)			
Retailer				*(Makridakis, et al., 2022)*		
Manufacturing industry					*(Knolla, et al., 2016)*	
Air cargo						(Koch, 2018)

Based on the literature review, it has become evident that incorporating certain elements, namely packaging uncertainty, lead time variability, and frozen period, is crucial. ML methods, such as the aforementioned ones, are relevant for addressing packaging uncertainty. Lead time variability can be incorporated by considering the average variability in lead time measured in days. Changes related to the frozen period element should only be applied when the frozen period has been utilized.

There is a noticeable overlap with the newsvendor problem in our research, as we aim to determine the optimal order quantity despite lacking precise knowledge of the demand. Exploring the potential integration of this theory into our study would be interesting to further investigate. While various methods are intriguing to employ, they are most suitable for specific elements within the forecast, such as predicting volumes for a non-Scania packaging supplier. Therefore, it is recommended to construct the final forecast by combining different parts, where each part utilizes a specific method.

4 Forecasting design

Up to now we have made an analysis of the current forecasting method used within SLN and made an overview of the requirements for the new forecasting model. Based on these requirements, we have conducted a literature review to find methods that can be usefull for this forecast. The overall conclusion of this literature study is that there is not one straightforward model that can be used for this forecast. Since we are dealing with different elements, deterministic and stochastic, we should make a combination of multiple models to create one flow of forecasting.

In this chapter, we discuss how the different methodologies found in chapter 3 can be translated and combined to this research. We start in 0 by constructing the flow of forecasting, where we also address where the application of the different models takes place. In 4.2 we focus on each part of the forecast design and describe the steps for each part in more detail. The data prepation for the ML methods and the ML methods itself is discussed in 4.3. The conclusion of this chapter, presented in Section 4.4, answers the first subquestion of the analyse phase:

Analyse

What are the improvements for SLN by using this suggested methodology?

5. *How can this suggested methodology be transformed to the situation of SLN?*

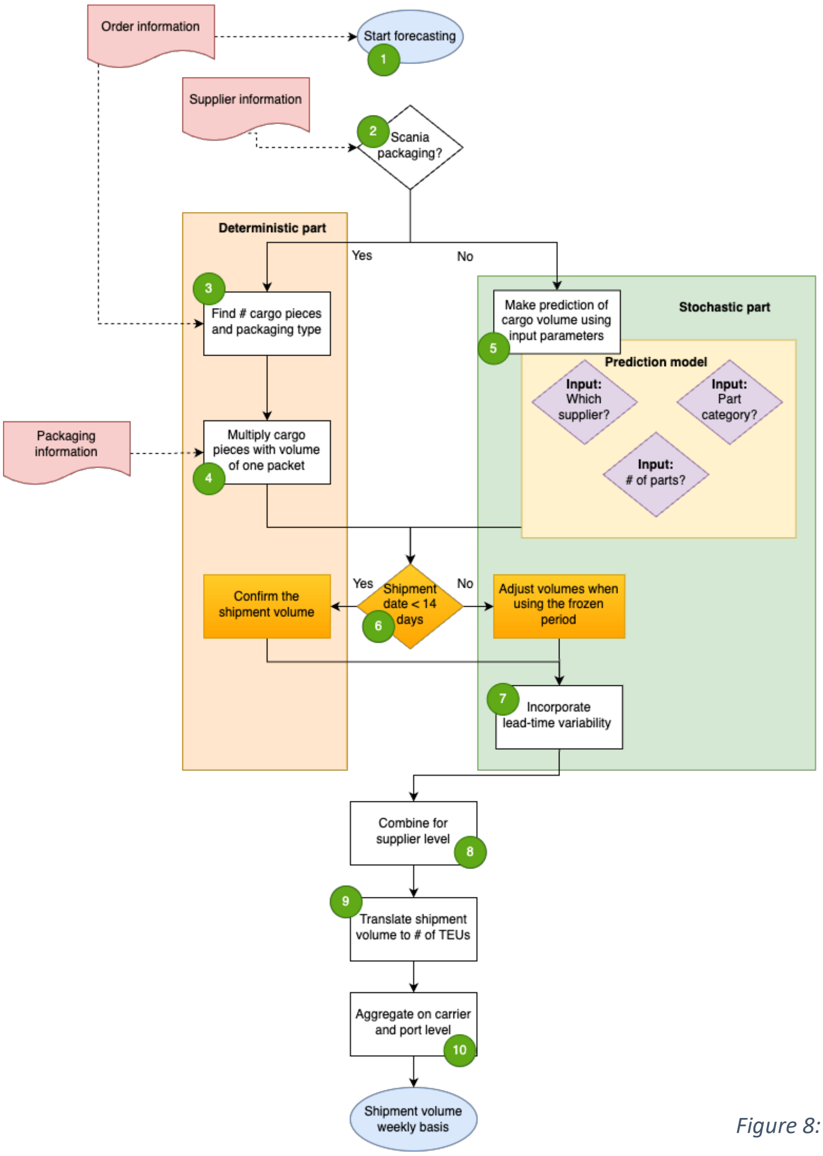


Figure 8: Flow of forecasting

4.1 Flow of forecasting

Since the methodologies found in the Literature study should be applied to a specific element in the forecast, it is important to sketch how the flow of forecasting does look like. The reason behind this is that there are multiple elements that should be incorporated in order to determine the shipment volumes. The forecasting flow is shown in Figure 8, starting with the scheduled order and ending with the final shipment volume on a weekly basis that should be received at a particular port. This flowchart also illustrates which portions of the flow are deterministic (Scania packaging is applicable) and which portions are stochastic (supplier does not use Scania packaging and the lead-time variability).

4.2 The forecast: step-by-step

This Section focuses on each part of the forecast design outlined in Figure 8 and provides a detailed description of the steps involved.

Step 1: Start forecasting flow

The initiation of the forecasting process starts at the *Start forecasting (1)* leaf, taking into account a scheduled orderline from the *Order information* data as its starting point of reference. If the data lacks a scheduled orderline (unavailability), the forecasting flow will not start.

Step 2: Check for Scania packaging

Each row in the order information data represents a part ordered with a specific quantity (see 2.1.1 for more information about this source). The goal is to determine shipment volumes for each row and aggregate them on a weekly basis. In the *Scania packaging (2)* leaf, the supplier name from the order is checked against the *Supplier information* data, specifically the *Scania packaging* column. If the Scania packaging column for this supplier contains the boolean value TRUE, the forecast proceeds to the deterministic part (orange) shown in Figure 8. Otherwise, it goes to the stochastic part (green). Table 5 provides a visualization of these two tables.

Table 5: Table on the left is the order information and table on the right is the supplier information

Partnumber	*Quantity*	*Supplier name*
1234	*500*	*X*
5678	*750*	*Y*
9101	*350*	*Z*

Supplier name	*Scania packaging*
X	*True*
Y	*False*
Z	*True*

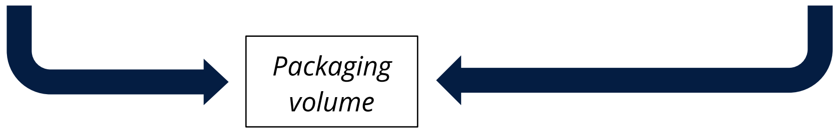
Steps 3 to 4: Scania packaging → TRUE

If the supplier for the specific order line uses Scania packaging, the volume can be determined deterministically. This involves considering the *number of cargo pieces* (where the quanity ordered is divided by the number of parts per cargo piece and rounded up) and *packaging type (3)*, which are available in the data order information. The volume of the specific packaging type is then obtained from the *Packaging information* table by multiplying the number of cargo pieces with the packaging volume (4). This provides the packaging volume for the order line if the supplier uses Scania packaging. Table 6 provides a visualization of these two tables.

Table 6: Table on the left is the order information and table on the right is the packaging information, with a multiplication of the number of pieces by the corresponding packaging volume. This results in the packaging volume for this orderline.

Partnumber	***Number of cargo pieces***	***Packaging type***
1234	*9*	*D2*
5678	*4*	*18*
9101	*13*	*A4*

Packaging type	***Packaging volume***
A4	$1.81\ m^3$
D2	$0.75\ m^3$
18	$2.42\ m^3$



Step 5: Scania packaging → FALSE

If the column Scania packaging in the supplier information data contains the boolean value FALSE for the supplier, the stochastic part of the forecast is used. For suppliers not using Scania packaging, a ML method is employed to determine the expected packaging volume. The ML method is specifically trained on historical data and *predicts (5) the packaging volume* using three input parameters: supplier (number or name), part number, and order quantity for the scheduled line. Consequently, each ML method is trained using all three features. Figure 9 illustrates an example of how this ML method works with some example parameters. The development of these methods is discussed in 4.3.

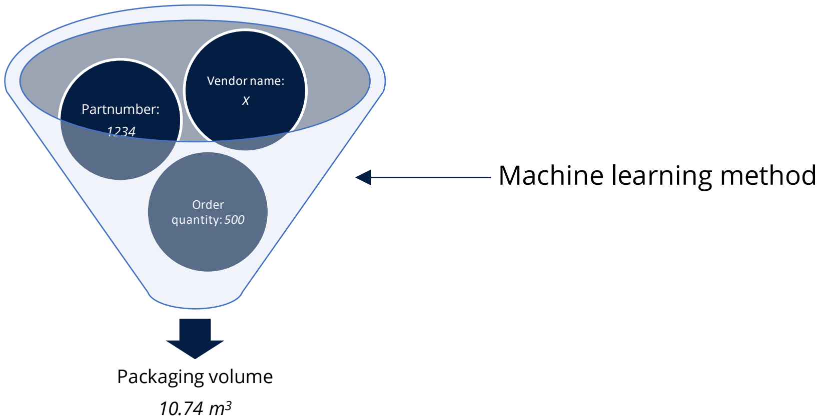


Figure 9: Visualization of the ML method, that uses three input parameters and predicts the packaging volume

Step 6: frozen period

The frozen period refers to the period before the delivery date when no changes can be made to an order. There are situations where the quantity of an order is changed by SLN or an entire order is dropped. The frozen period is part of the agreement between SLN and the supplier, and it is recorded in the ERP system. In our case, all overseas suppliers have a fixed frozen period of 14 days. This means that no changes can be made to the orders within 14 days before the planned delivery date, and shipments can be confirmed to the carrier during this period. Visual representation of this time window can be seen in Figure 10.

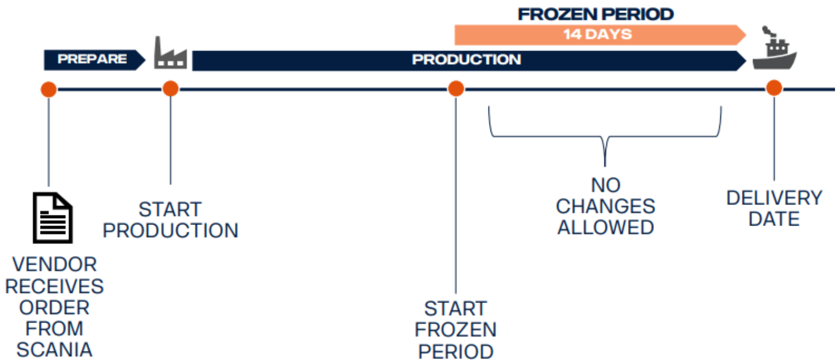


Figure 10: Visualization of the time window with the frozen period

To *incorporate the frozen period (6)* in our forecast, we divide it into the deterministic and stochastic components. The deterministic part pertains to orders falling within the frozen period, which is 14 days before the planned shipment date for all overseas suppliers (indicated by the orange part in Figure 10). During this period, no changes are allowed to the order, and the shipment load is confirmed. If the planned shipment date is more than 14 days in the future, we move on to the stochastic part of the forecast. To account for stochasticity, we conducted an analysis based on order information from December 2022 to May 2023. We compared scheduled order lines across these reports and recorded any instances where a scheduled line was removed, considering it as an order cancellation.

Table 7: The number of scheduled orders and cancellations for each supplier, and the cancellation rate is calculated as the ratio of cancellations to scheduled orders.

Supplier name	***Number of scheduled orders***	***Number of cancellations***	***Cancellation rate***
x	*100*	*10*	*10%*
y	*80*	*12*	*15%*
z	*160*	*32*	*20%*

This analysis was performed for all suppliers, comparing the number of cancellations with the number of scheduled orders. The results are presented in Table 7, which includes the suppliers along with their corresponding number of scheduled orders and the recorded number of cancellations. We refer to Table 8 to explain how we incorporate this information into our forecast. The shipment volume for a specific scheduled order is determined either deterministically or stochastically using the ML method. Since the planned shipment date is more than 14 days away from the current day (5-6-2023), we need to account for the cancellation rate specific to the supplier. To do this, we reduce the shipment volume by multiplying it by one minus the cancellation rate. If the cancellation rate for a specific supplier is unknown, we assume it to be 0, and the shipment volume remains unchanged. If the planned shipment date is less than 14 days away, we assume that the shipment will take place, and the only remaining uncertainty is the arrival time. With this information, we proceed to the next step of the forecast.

Table 8: Input, process, and output for incorporating the frozen period (step 6).

Input		***Process***	***Output***	
Supplier name	x	*Shipment volume* $\times$ *(*$1-$ *Cancellation rate)*	*Supplier name*	x
Cancellation rate	*10%*		*Cancellation rate*	*10%*
Shipment volume	*10.3* m^3		*Shipment volume*	*9.27* m^3
Planned shipment date	*1-7-2023*	$10.3\ m^3 \times (1-10\%) = 9.27\ m^3$	*Planned shipment date*	*1-7-2023*

Step 7: Lead time variability

The forecast incorporates *lead time variability (7)* for situations where the supplier uses or does not use Scania packaging. Lead time variability refers to the analysis of the average number of days a supplier is classified as late or early. Historical data is used to analyze the differences between planned delivery time and actual delivery time for each supplier. To incorporate lead time variability into the forecast, the average variance between actual and planned delivery dates is measured for each supplier. Due to the limited amount of data available, we are unable to conduct this analysis at the part level. Therefore, we make the assumption that the same level of lead time variability applies to all deliveries from this supplier. The variance, represented by σ^2, is calculated as follows:

Equation 6

$$\sigma_x^2 = variance\ for\ supplier\ x = \frac{\sum_x Variances}{Total\ \#\ of\ observations}$$

For example, to determine the lead time variability for supplier x, the variances for ten orders are summed to a total of 93 days. Therefore, the variance for supplier x is $\sigma_x^2 = \frac{93}{10} = 9.3\ \approx 9\ days$. The application of this approach is not representative when we have only one observation. In such cases, we exclude any consideration of lead time variability. The final forecast is being aggregated on a weekly level, so variances below 7 days will not be reflected in the final forecast. However, our example of 9 days should be visible. Based on the planned shipment date, the shipment can take place 9 days before and after the planned shipment date.

To aggregate shipment volumes on a weekly basis, we utilize proportions. We examine all the measured variances and categorize them into bins based on their range. The number of variances falling within each range is divided by the total number of observations, yielding our proportions. This aggregation is done at a weekly level. An example of this analysis for one supplier is provided in Table 9.

Table 9: Lead time variability analysis supplier x.

Variance measured	***Count***	***Proportion (count / 50)***	***Proportion weekly level***
-12	*2*	*0.04*	*0.08*
-11	*1*	*0.02*	
-7	*1*	*0.02*	
-5	*9*	*0.18*	*0.7*
-4	*23*	*0.46*	
-3	*3*	*0.06*	
7	*7*	*0.14*	*0.18*
11	*2*	*0.04*	
17	*1*	*0.02*	*0.04*
44	*1*	*0.02*	
37	*50*	*1*	*1*

We refer to Table 10 to explain how we incorporate this information into our forecast. Building upon the previous steps, we use the same supplier, shipment volume, and planned shipment date, while also integrating the lead time variability from Table 9. To aggregate the shipment volume over the weeks, we multiply the shipment volumes by the proportions derived from the lead time variability. For instance, if the planned shipment date for a scheduled order is 1-7-2023, corresponding to week 26, the expected volume for this week would be 9.27 multiplied by 0.7, resulting in 6.489 m³. We repeat this process for the other weeks as well.

Table 10: Input, process, and output for incorporating the lead time variability (step 7).

Input			Process	Output
Supplier name	x		*Shipment volume × proportion lead time analysis*	
Shipment volume	*9.27 m^3*		W25 $9.27 \times 0.08 = 0.742\, m^3$	
Planned shipment date	*1-7-2023*		W26 $9.27 \times 0.7 = 6.489\, m^3$	
Lead time variability	$t-1$	*0.08*	W27 $9.27 \times 0.18 = 1.669\, m^3$	
	t	*0.7*	W28 $9.27 \times 0.04 = 0.371\, m^3$	
	$t+1$	*0.18*		
	$t+2$	*0.04*		

Week	*25*	*26*	*27*	28
Volume	*0.742*	*6.489*	1.669	0.371

Step 8: Combine for supplier level

Up until now, we have calculated the shipment volumes in m^3 for each order row and incorporated lead time variability by aggregating the volumes over the weeks. However, this has been done for each scheduled order, which pertains to a specific part and its corresponding quantity. Now, we need to combine these volumes from different scheduled orders for this supplier. This enables us to obtain the total volumes expected to be received from this supplier on a weekly basis. To understand how we integrate this information into our forecast, we refer to Table 11. We gather all the scheduled orders for this supplier and add up the volumes if the shipments are scheduled for the same week. Ultimately, we have the total volume in m^3 that can be anticipated from supplier x.

Table 11: Input, process, and output for incorporating combining for supplier level (step 8).

Input

Supplier name x

	Week	*25*	*26*	*27*	*28*
#1	*Volume*	*0.742*	*6.489*	1.669	0.371

	Week	*24*	*25*	*26*	*27*
#2	*Volume*	*0.993*	*8.687*	2.234	0.496

	Week	*27*	*28*	*29*	*30*
#3	*Volume*	*1.461*	*12.782*	3.287	0.730

Process

	Week	*24*	*25*	*26*	*27*	*28*	*29*	*30*
Order	*#1*		*0.74*	*6.49*	*1.67*	*0.37*		
	#2	*5,77*	*50.5*	*13.0*	*2.88*			
	#3				*1.46*	*12.78*	*3.29*	*0.73*
	Sum	*5.77*	*51.1*	*19.5*	*6.01*	*13.15*	*3.29*	*0.73*

Output

Week	Total shipment volume in m^3
24	5.77
25	51.1
26	19.5
27	6.01
28	13.15
29	3.29
30	0.73

Step 9: Translate to TEUs

The carriers are primarily interested in the number of containers required for transportation from the origin port to the destination port. Therefore, we also convert the shipment volumes from m^3 to the number of TEUs. The volume of one TEU is equivalent to 38.51 m^3, and a forty-foot container counts as 2 TEUs (CBS, 2023). Considering SLN's experience, we assume a filling rate of 85%, meaning that 32. m^3 can be allocated for cargo. We refer to Table 12 to explain how we incorporate this conversion into our forecast. Using the input parameters, we translate the shipment volume to the number of TEUs by dividing it by 32.73 m^3 and round this number up to whole figures. This results in the number of TEUs to expect from supplier x in week t.

Table 12: Input, process, and output for translating the shipment volume to the number of TEUs (step 9).

Input		Process	Output	
Supplier name	x	$\frac{Cargo\ volume}{32.73} = \#of\ TEUs$	*Supplier name*	x
Week	$t = 25$		*Week*	$t = 25$
Shipment volume	*51.21 m^3*	$\frac{51.21}{32.73} = 1.56\, TEU = 2\, TEUs$	*Shipment volume*	*51.21 m^3*
Net volume one TEU	*32.73 m^3*		*Number of TEUs*	*2*

Step 10: Aggregate

The final step involves aggregating the volumes on the supplier and port levels, as well as on a weekly basis. This can be achieved by filtering the data accordingly. Supplier-level aggregation is straightforward since we already have the supplier names in the order information table. For the port level, we need to determine the port associated with each supplier, which can be found in the supplier details data. We refer to Table 13 to explain how we incorporate this conversion into our forecast. Currently, we are including additional input parameters (country, postal code, and incoterms), which are specifically significant for the end user of the forecast. This enables the end-user of the forecast to utilize a filter that takes into account these parameters.

Table 13: Input, process, and output for the aggregation (step 10).

<table>
<tr><th colspan="2">Input</th><th>Process</th><th colspan="2">Output</th></tr>
<tr><td>Supplier name</td><td>x</td><td rowspan="9">Sum <u>all the volumes</u> and <u># of TEUs</u> from <u>all suppliers</u> that are assigned to that specific <u>port of loading</u> on a <u>weekly level</u>.</td><td>Assigned port of loading</td><td>Santos</td></tr>
<tr><td>Shipment volume</td><td>51.21 m^3</td><td>Country</td><td>Brazil</td></tr>
<tr><td>Number of TEUs</td><td>2</td><td>Carrier</td><td>C</td></tr>
<tr><td>Week</td><td>$t = 25$</td><td>Supplier names</td><td>x, y, z</td></tr>
<tr><td>Assigned port of loading</td><td>Santos</td><td colspan="2" rowspan="5"><table><tr><th>Week</th><th>Shipment volume in m³</th><th>TEUs</th></tr><tr><td>24</td><td>175.66</td><td>6</td></tr><tr><td>25</td><td>478.77</td><td>16</td></tr><tr><td>26</td><td>45.60</td><td>2</td></tr><tr><td>27</td><td>442.53</td><td>15</td></tr><tr><td>28</td><td>488.83</td><td>17</td></tr><tr><td>29</td><td>599.20</td><td>20</td></tr><tr><td>30</td><td>547.85</td><td>19</td></tr></table></td></tr>
<tr><td>Country</td><td>Brazil</td></tr>
<tr><td>Postal code</td><td>BR 25846</td></tr>
<tr><td>Incoterms</td><td>Free carrier</td></tr>
<tr><td>Carrier</td><td>C</td></tr>
</table>

4.3 ML methods

We lack knowledge of the packaging volumes for 40% of the shipments, as these suppliers employ non-Scania packaging. Therefore, it is imperative to discover an approach that enables us to anticipate these volumes. To accomplish this, we utilize the ML methods (as introduced in step 5). Given the significant influence of these shipments on the majority of deliveries, we delve into a thorough discussion of these methods. This Section discusses both the data preparation procedures for ML methods and the methods themselves. It is essential to thoroughly execute the data preparation processes prior to employing ML methods. Upon completing this step, the process can be replicated for all the ML methods.

4.3.1 Data preparation

The dataset used for the prediction models is based on historical shipment volumes from mid-2019 through 2022. It specifically consists of information about containers shipped by a particular carrier, including the volume of each container. Although the data was collected during the Covid-19 period when shipment volumes were lower, it remains usable because the parts do not differ significantly. By utilizing the container number and data from the ERP system, we can determine the specific parts and quantities included in each shipment. However, the supplier names in the carriers dataset do not match the names used internally by SLN. Consequently, we manually replaced all the supplier names in the dataset with the correct ones.

The features *supplier number* and *supplier name* essentially contain the same unique data, so we drop the *supplier number* column. Some prediction models require categorical features, for example *part number* and *supplier name,* to be translated into dummy variables. However, this approach is not preferred in our case because it would result in a total of 1,272 dummy variables in the current dataset. During the development of the first model, we observed that having such a large number of dummy variables negatively impacted the model's performance.

To address this issue, we decided to reduce the number of categories for the part number feature by replacing it with the corresponding category name called *part category*. The part category is a feature that indicates the category to which a specific part belongs. The number of unique values in this feature is considerably lower than the part number feature. As each part corresponds to a specific category, this feature can be leveraged for predicting the shipment volume.This step reduced the number of dummy variables to 614, which is still relatively high. Unfortunately, we do not have any additional information available to further reduce this number. After removing redundant columns, our dataset consists of three features, one target value, and 23,383 observations. Please refer to Table 14 for a summary of this final dataset.

Table 14: Summary dataset, where the shipment volume is the target value.

Column name	***Description***	***Feature / target***	***Catergorical / numeric***	***Number of observations***
Part category	*Description of the category where this material belongs to.*	*Feature*	*Catergorical*	*496 (unique)*
Quantity ordered	*The number of units on the order for this specific part.*	*Feature*	*Numeric*	*23,383*
Supplier name	*The name of the supplier, which is a specific category.*	*Feature*	*Catergorical*	*118 (unique)*
Shipment volume	*The volume for that order that need to be shipped from port X in week t.*	*Target*	*Numeric*	*23,383*

We use R-studio to prepare the data and apply the different methods. Upon importing the data into R-studio, we manually transform the categorical features into factors. We randomly split the 23,383 observations into two sets for the validation of the models, with the training set containing 80% of the data points and the test set containing the remaining 20%. To reduce the likelihood that a product is not in the training or test set (because a part can be introduced later in the dataset), sampling is done without replacement and at random.

4.3.2 Over-fit

For the ML methods, it is important that we do not *over-fit* our training-set. Over-fitting is when the model relies too much on randomness/noise in the training set to make its classifications. As a result, it will probably not extend well to a new dataset (Tatman, 2018). There are several ways to overcome overfitting in machine learning, where we focus on the application of the following three:

1. **Use cross-validation**: Cross-validation is a technique that can be used to estimate the performance of a model on unseen data. By splitting the data into multiple folds and training the model on different subsets of the data, cross-validation can help to identify whether the model is overfitting.
2. **Add more data augmentation**: If the training data is limited, data augmentation techniques such as image rotation, flipping, and zooming can be applied to artificially increase the size of the dataset and help the model learn more generalized patterns.
3. **Early stopping**: This is a technique where the training process is stopped early when the model's performance on a validation dataset stops improving. This can help prevent the model from continuing to learn the noise in the training data.

4.3.3 Multiple linear regression

As a baseline for comparison with the ML methods, we utilize the multiple linear regression model. We opted for multiple linear regression because it allows us to establish the relationship between the target value (shipment volume) and one or more features (Kanade, 2022). Implementing the multiple linear regression model is relatively straightforward compared to the ML methods. There are several ways to build a multiple linear regression model, depending on the specific requirements and available resources. In this research we use a ML library, which is called MASS and is built-in R-studio. This library offer a pre-implemented algorithms and functionalities for fitting multiple linear regression models, including options for regularization techniques like Ridge or Lasso regression.

We built the multiple linear regression model using all the features and three additional feature combinations: [Part category & Quantity shipped], [Part category & Supplier name], and [Quantity shipped & Supplier name]. When applying the multiple linear regression function to our training set, the model automatically generates a set of dummy variables for our categorical features (Supplier name and Part category). To avoid overfitting, we employ cross-validation. The multiple linear regression model produces coefficients for all its features, which are used in the prediction function. The predict function utilizes the test set to predict shipment volume based on the features in the test set. After making the predictions, we assess the model's performance by comparing it to the actual shipment volumes in the test set. For the complete code used in R-studio, we refer to Appendix II.

4.3.4 XGBoost

The first gradient boosting method we employ on our dataset is XGBoost, aiming to determine if it surpasses the baseline model. XGBoost operates differently, requiring specific data preparations to be compatible with R-studio.Since XGBoost cannot handle factors and the categorical features are currently factors, we manually convert them into dummy variables. We achieve this by integrating these factors into a matrix and assigning them as dummy variables. The categorical feature matrix and the numerical features are then combined. After applying the training and test split, we convert these sets into a specific matrix to enable the XGBoost model to function.

To prevent overfitting, we need to fine-tune the model's parameters. Two adjustable parameters are the maximum depth of the tree and the learning rate. We define the parameters we want to test and repeatedly run the model with different parameter values. We run the model using the values 2, 4 and 6 for the maximum depth and 0.01, 0.2 and 0.4 for the learning rate. The tuning loop stores the best result and the corresponding parameters. The full code for the XGBoost method, including this tuning loop, can be found in Appendix III.

4.3.5 LightGBM

The next ML method we employ for predicting shipment volume for suppliers not using Scania packaging is LightGBM. Many models require thorough variable preprocessing to achieve accurate predictions. Unlike XGBoost, LightGBM excels at handling categorical variables (factors), so there is no need to convert variables into dummies (one-hot encode). In fact, it is not recommended to do so, as it can slow down the process and potentially result in worse performance. After applying the training and test split, we convert these sets into a specific matrix to enable the LightGBM model to work.

To address overfitting, we utilize the same loop of parameters as in the XGBoost model. This loop stores the parameters that yield the best performance. Once the loop completes, we employ these parameters in the LightGBM model, resulting in the final LightGBM model with optimal performance. The complete code for the LightGBM model, including this tuning loop, can be found in Appendix IV.

4.3.6 N-Beats

The final method we employed was the N-Beats method, which utilizes a neural network. The steps to implement the N-Beats model are similar to those of the baseline model, so no additional data preparation steps are required. We built the N-Beats method using all the features and three additional feature combinations: [Part category & Quantity shipped], [Part category & Supplier name], and [Quantity shipped & Supplier name]. The complete code for the N-Beats model can be found in Appendix V.

In the next chapter, we evaluate the performance of the baseline model and the other ML methods in predicting the packaging volumes for suppliers that do not use Scania packaging. The best performing prediction method is applied to the final forecasting method, represented by the yellow part in Figure 8.

4.4 Conclusion

In this chapter, we discussed the flow of forecasting and provided a step-by-step guide to designing the forecast. The forecasting flowchart illustrated the deterministic and stochastic parts of the forecast, depending on whether the supplier used Scania packaging or not, the frozen period and the lead time variability. For suppliers using Scania packaging, the volume could be determined deterministically based on cargo pieces and packaging type. However, for suppliers not using Scania packaging, ML methods were employed to predict the expected packaging volume. The lead time variability was measured as the variances between planned and actual delivery dates for each supplier, and we multiply the shipment volumes by the proportions in which week the shipment is expected. The frozen period was incorporated into the forecast by reducing it with the cancellation rate when we are outside the frozen period. The shipment volumes in m^3 were translated into TEUs to represent the number of containers required for transportation.

Additionally, we introduced ML methods as a solution for predicting packaging volumes for suppliers not using Scania packaging. The data preparation procedures for ML methods were described. The steps for building and evaluating these models were explained, emphasizing the importance of parameter tuning to avoid overfitting. Overall, this chapter provided a comprehensive overview of the forecasting flow, step-by-step procedures, and the application of ML methods for predicting packaging volumes, laying the foundation for accurate and reliable forecasts in the subsequent chapters.

5 Results

In this chapter, we present the performance of the designed forecasting flow. In 5.1 we discuss the performance of the baseline method and the additional ML methods to determine the shipment volumes. We compare all these methods in 5.2 and select the best performing ML method as our final method to predict the volumes in case Scania packaging is not applicable. In 5.3 we have our final forecasting model and we run some test-run with it and check its performance. Following that, we conduct a thorough comparison in 5.4 between the current forecast, the proposed forecast developed in this research and the actual number of containers. In 5.5, we discuss the parallels with the newsvendor problem, highlighting their similarities and how this theory can be used as a decision supporting tool. The conclusion of this chapter, presented in Section 5.6, answers the last subquestion of the analyse phase and therefore also the main question from the analyse phase:

Analyse

What are the improvements for SLN by using this suggested methodology?

6. *What is the performance of this suggested methodology?*

5.1 Performance ML methods

As introduced in Section 4.3, for the prediction model we apply three different ML methods. In this Section we elaborate on all three, from running the methods, tuning the parameters, and comparing the performance of the three methods. For the comparison of the methods, we also apply a multiple linear regression model, which serves as a baseline. The ML methods were executed in R-Studio (version 12.0) on a MacBook Pro 14" 2022 equipped with an M1 processor, 16GB RAM, and a 512GB flash storage. We utilize the ML method that demonstrates the highest accuracy, in step 5 of the forecasting flow.

5.1.1 Multiple linear regression

We start with our initial model and identify the most effective model for the training set, which included the features of part category and supplier name as shown in Table 15. The model with all features does not perform well, and the combinations of [Part category & Quantity shipped] and [Quantiy shipped & Supplier name] have the lowest accuracy. For a summary of performance results for all feature combinations, we refer to Appendix VI.

Table 15: Performance Multiple Linear Regression

Performance measure	Result	Unit of measurement
MSE	*152.2419*	*m^3*
MAE	*9.889811*	*m^3*
RMSE	*12.33863*	*m^3*
Bias	*0.06761485*	*m^3*
Running time	*114.8*	*seconds*

Table 15 presents the description of the top-performing multiple linear regression model. Considering the extensive set of dummy variables, we only focus on those with substantial estimators, indicating significant contributions to the outcome. Our examination of residuals, i.e., the variation between predicted values and actual values, reveals that the median is precisely 0. This implies that the distribution of residuals is somewhat symmetrical, indicating that our model is performing well for both low and high-value predictions.

Moreover, the standard error of the coefficient provides an approximation of the coefficient's standard deviation, revealing the degree of uncertainty associated with it. The t-statistic is determined by dividing the estimate by the standard error, and high t-statistics are desirable as they suggest a proportionally smaller standard error compared to the coefficient. Upon examining Table 16, we can only observe a positive t-value for PartCategory - C. We use the p-value in conjunction with the t-statistic to determine the significance of the coefficient within the model, usually considering any p-value less than 0.05 as significant. Referencing Table 16 again, PartCategory - A is the sole statistically meaningful coefficient.

Table 16: Feature importance Multiple linear regression

Residuals	***Min***	***1Q***	***Median***	***3Q***	***Max***
	-30.533	*-8.259*	*0.00*	*7.935*	*38.875*

Feature	**Parameters estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
PartCategory - A	-19.48797	5.16694	-3.772	0.000163 *
PartCategory - B	-0.64107	6.15095	-0.104	0.916994
SupplierName - A	-18.06172	12.60694	-1.433	0.151966
PartCategory - C	8.383e-01	1.401e+00	0.599	0.549509

Our initial model has yielded promising results. The RMSE for this model is 12.33, indicating that when SLN books a shipment of one FCL (approximately 67 m^3), the shipment volume prediction is off by 12.33 m^3, or approximately 16% of an FCL. Additionally, there is minimal bias, indicating that the target function's prediction varies slightly (in m^3) with changes in the training dataset. With this as our baseline, we compare the other ML methods.

5.1.2 XGBoost

XGBoost is the first gradient boosting method that we apply. The XGBoost model runs with the standard parameters, namely a 0.3 learning rate and a maximum depth of 6, performing 10,000 rounds, but the early stopping round is set to 1,000, meaning the process stops if the test RMSE fails to improve after 1,000 rounds. We evaluate the model's performance using the predict function and Table 17 highlights the validation results.

Table 17: Performance initial XGBoost

Performance measure	***Result***	***Unit of measurement***
MSE	*154.3603*	m^3
MAE	*9.955656*	m^3
RMSE	*12.42418*	m^3
Bias	*0.08584038*	m^3
Running time	*63.3*	*seconds*

Upon comparing the initial outcomes with the baseline model, there was no noteworthy progress. In order to combat the issue of overfitting, we must adjust the parameters in our model. We establish a collection of parameters that we wish to evaluate and proceed to test them using a loop function. The loop function retains record of the optimal score and its corresponding parameters. Once this loop is complete, we use the best parameters found (refer to Table 18) and check the performance of the XGBoost model using these best parameters.

Table 18: Best parameters found XGBoost model

Parameters	***Best***
Maximum depth	*2*
Learning rate	*0.2*

Table 19 reveals that parameter tuning results in enhanced performance improvements. The performance of the XGBoost model has shows differences from the baseline model, and this is attributable to the gradient boosting method employed by XGBoost. This method involves using prior training data samples to forecast future models, and it has been frequently observed that it outperforms the bagging method in literature. Although there has been a rise in bias, it remains relatively insignificant.

Table 19: Performance final XGBoost

Performance measure	Result	Unit of measurement	Difference in % from baseline
MSE	151.2342	m^3	-0,662%
MAE	9.899847	m^3	0,101%
RMSE	12.29773	m^3	-0,323%
Bias	0.09409155	m^3	39,158%
Running time	73.3	seconds	-36,150%

5.1.3 LightGBM

We use LightGBM as another ML method to forecast shipment volume for suppliers who do not employ Scania packaging. Initially, we conduct a LightGBM model run with standard parameters such as a maximum depth of 2, a learning rate of 0.1, a maximum of 7 leaves per tree, and a minimum of 1 data point per leaf. We are training the model for 10.000 rounds, with early stopping set at 1.000 rounds. This means the model continuously trains until the test RMSE does not improve for 1.000 rounds. Subsequently, we verify the model using the predict function, and its performance is summarized in Table 20.

Table 20: Performance inital LightGBM

Performance measure	Result	Unit of measurement
MSE	153.4476	m^3
MAE	9.927458	m^3
RMSE	12.3874	m^3
Bias	0.1021023	m^3
Running time	38.5	seconds

To address the problem of overfitting, adjustments must be made to the model parameters. To achieve this, we employ the same parameter tuning loop utilized in the XGBoost model, with the optimal parameters given in Table 21.

Table 21: Best parameters found LightGBM model

Parameters	Best
Maximum depth	2
Learning rate	0.3
Maximum number of leaves	4
Minimum number of data in one leaf	1

Examining Table 22 reveals that the LightGBM model slightly outperforms the baseline and XGBoost models, potentially due to differences in tree construction. While the bias has increased slightly in the LightGBM model, it remains comparatively small. Thus far, the LightGBM model is the most effective machine learning method for predicting shipment volume for non-Scania packaging suppliers.

Table 22: Performance final LightGBM

Performance measure	Result	Unit of measurement	Difference in % from baseline
MSE	150.8456	m^3	-0,917%
MAE	9.857454	m^3	-0,327%
RMSE	12.28192	m^3	-0,452%
Bias	0.07623363	m^3	12,747%
Running time	54.9	seconds	-52,178%

5.1.4 N-Beats

The N-Beats method, a neural network-based approach, is the final ML method employed. As described in 4.3.6, we construct the model using all the features and also explore three other feature combinations. However, we do not observe any differences in the network architecture or accuracy performance across these feature combinations. Figure 11 provides a visualization of the neural network structure. Our model consists of a hidden layer with 2 neurons. The black lines represent the connections with their respective weights, while the blue line represents the bias term.

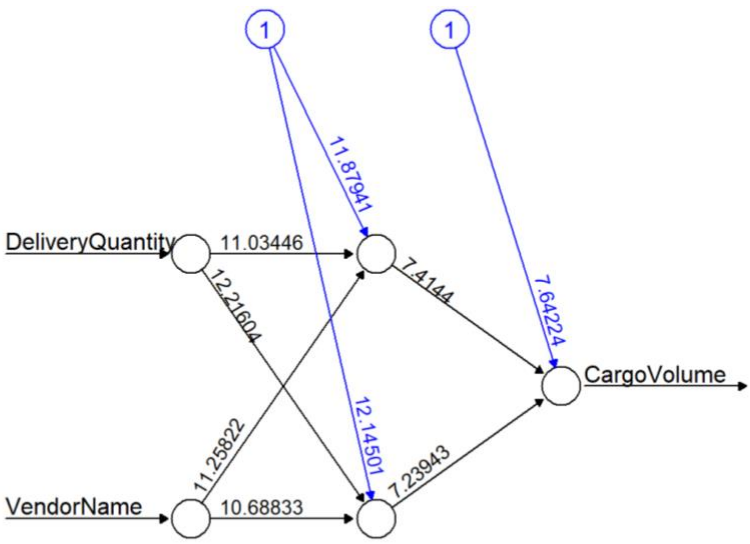


Figure 11: Layout neural network

The performance of all these combinations are the same and are summarized in Table 23.

Table 23: Performance final N-Beats

Performance measure	Result	Unit of measurement	Difference in % from baseline
MSE	200.287	m^3	31,558%
MAE	11.59909	m^3	17,283%
RMSE	14.15228	m^3	14,708%
Bias	0.03984227	m^3	-41,075%
Running time	134.8	seconds	17,422%

After analyzing the performance measures presented in Table 23, we determine that the N-Beats method does not surpass the performance of other methods, including the baseline. Previous studies have consistently shown that tree-based methods tend to outperform neural networks (Sharma, 2022). These tree-based methods are generally simplified versions of neural networks, which could potentially explain their superior performance compared to neural networks.

5.2 Comparison ML methods

Table 24 presents the consolidated performance measures for all the ML methods executed thus far. From this table, we deduce that LightGBM emerges as the most effective method for SLN in predicting shipment volume for suppliers who do not employ Scania packaging. This method offers a relatively straightforward implementation and understanding, while also delivering the highest level of accuracy. Therefore, it is strongly recommended for SLN to adopt this prediction method for suppliers who do not use Scania packaging.

Table 24: Comparison ML methods

Performance measure	Multiple linear regression	XGBoost	LightGBM	N-Beats
MSE	154.3603	151.2342	150.8456	200.287
MAE	9.955656	9.899847	9.857454	11.59909
RMSE	12.42418	12.29773	12.28192	14.15228
Bias	0.08584038	0.09409155	0.07623363	0.03984227
Running time	114.8	63.3	54.9	134.8

5.3 Final forecasting model: test-run

After selecting the LightGBM method as our ultimate prediction model for suppliers who do not utilize Scania packaging and determining the cancellation rate for the stochastic frozen period and the variability of supplier lead time, we have now achieved our finalized forecasting model. In this Section, we conduct three test runs using the forecasting model, each based on unique historical orders. These orders' characteristics is used to forecast the expected shipment volume and the corresponding week of arrival at the port of loading. Table 25 presents the specific details of these three orders, including parameters such as lead time variability and Scania packaging information which is stored in the supplier data.

Table 25: Order details for test-run

Order	Part category	Quantity	Planned delivery date	Week-number	Supplier	Lead time variability σ	Scania packaging?	Frozen period – cancellation rate
#1	A	1,792	26/06/2022	26	x	$t-1$: 0.08 t: 0.7 $t+1$: 0.18 $t+2$: 0.04	False	10%
#2	B	3,366	11/09/2022	38	y	$t-1$: 0.06 t: 0.16 $t+1$: 0.26 $t+2$: 0.52	False	15%
#3	C	624	11/08/2022	33	z	$t+1$: 0.55 $t+2$: 0.28 $t+3$: 0.05 $t+4$: 0.13	False	20%

Using the order details, we proceed with making forecasts following the design outlined in 4.2, iterating through the necessary steps. Table 26 illustrates each step alongside its corresponding input and output.

Table 26: Output test-run

<table>
<tr><th>Step of the forecast</th><th>Description</th><th>Order #1</th><th>Order #2</th><th>Order #3</th></tr>
<tr><td>Steps 1 to 2: check for Scania packaging</td><td>In the supplier information data we can see that all the suppliers (x,y and z) do not use Scania packaging, and therefore we <u>continue to step 5</u> of the forecast.</td><td>False</td><td>False</td><td>False</td></tr>
<tr><td>Step 5: Scania packaging → FALSE</td><td>We use the LightGBM model to make a prediction of the expected packaging volume for this specific order, using the part category, quantity and supplier.</td><td>24.32 m^3</td><td>22.96 m^3</td><td>26.88 m^3</td></tr>
<tr><td>Step 6: frozen period</td><td>The planned shipment date is more than 14 days in the future, so we lower the expected volumes with the cancellation rate of the supplier.</td><td>23.83 m^3</td><td>19.52 m^3</td><td>21.50 m^3</td></tr>
<tr><td>Step 7: Lead time variability</td><td>We aggregate the shipment volumes over the weeks by multiplying it by the proportions of the lead time variability. Now we do have the expected shipment volumes on a weekly level.</td>
<td><table><tr><th>Week</th><th>Prop.</th><th>Vol</th></tr><tr><td>25</td><td>0.08</td><td>1.91</td></tr><tr><td>26</td><td>0.7</td><td>16.7</td></tr><tr><td>27</td><td>0.18</td><td>4.29</td></tr><tr><td>28</td><td>0.04</td><td>0.95</td></tr></table></td>
<td><table><tr><th>Week</th><th>Prop.</th><th>Vol</th></tr><tr><td>37</td><td>0.06</td><td>1.17</td></tr><tr><td>38</td><td>0.16</td><td>3.12</td></tr><tr><td>39</td><td>0.26</td><td>5.08</td></tr><tr><td>40</td><td>0.52</td><td>10.15</td></tr></table></td>
<td><table><tr><th>Week</th><th>Prop.</th><th>Vol</th></tr><tr><td>34</td><td>0.55</td><td>11.83</td></tr><tr><td>35</td><td>0.28</td><td>5.91</td></tr><tr><td>36</td><td>0.05</td><td>1.08</td></tr><tr><td>37</td><td>0.13</td><td>2.69</td></tr></table></td></tr>
<tr><td>Step 8: combine for supplier level</td><td>We combine these volumes from different scheduled orders for this supplier. This enables us to obtain the total volumes expected to be received from this supplier on a weekly basis.</td>
<td><table><tr><th>Week</th><th>Vol</th></tr><tr><td>25</td><td>51.1</td></tr><tr><td>26</td><td>19.5</td></tr><tr><td>27</td><td>6.01</td></tr><tr><td>28</td><td>13.2</td></tr></table></td>
<td><table><tr><th>Week</th><th>Vol</th></tr><tr><td>37</td><td>5.8</td></tr><tr><td>38</td><td>50.5</td></tr><tr><td>39</td><td>27.8</td></tr><tr><td>40</td><td>3.3</td></tr></table></td>
<td><table><tr><th>Week</th><th>Vol</th></tr><tr><td>34</td><td>26.9</td></tr><tr><td>35</td><td>42.2</td></tr><tr><td>36</td><td>36.4</td></tr><tr><td>37</td><td>9.6</td></tr></table></td></tr>
<tr><td>Step 9: translate to TEUs</td><td>Assuming a filling rate of 85% of a TEU, we divide the packaging by 32.73 m^3 such that we can determine the # of TEUs.</td>
<td><table><tr><th>Week</th><th>Vol</th><th>TEUs</th></tr><tr><td>25</td><td>51.1</td><td>2</td></tr><tr><td>26</td><td>19.5</td><td>1</td></tr><tr><td>27</td><td>6.01</td><td>1</td></tr><tr><td>28</td><td>13.2</td><td>1</td></tr></table></td>
<td><table><tr><th>Week</th><th>Vol</th><th>TEUs</th></tr><tr><td>37</td><td>5.8</td><td>1</td></tr><tr><td>38</td><td>50.5</td><td>2</td></tr><tr><td>39</td><td>27.8</td><td>1</td></tr><tr><td>40</td><td>3.3</td><td>1</td></tr></table></td>
<td><table><tr><th>Week</th><th>Vol</th><th>TEUs</th></tr><tr><td>34</td><td>26.9</td><td>1</td></tr><tr><td>35</td><td>42.2</td><td>2</td></tr><tr><td>36</td><td>36.4</td><td>2</td></tr><tr><td>37</td><td>9.6</td><td>1</td></tr></table></td></tr>
<tr><td>Step 10: aggregate</td><td>The final step is to aggregate the volumes on supplier, port level (and therefore also on forwarder level since the port is covered by a specific forwarder) and on a weekly basis.</td><td colspan="3">The results are in principle the same as step 9, but now also assigned to a specific port and carrier.</td></tr>
</table>

Upon completing this forecast design, we obtain the anticipated shipment volumes measured in TEUs and aggregate them by the expected arrival weeks at the port of loading. We compare these predictions with the actual status of the scheduled lines, as depicted in Table 27.

Table 27: Comparison between the prediction and actual orders

Order	Weeknumber	Predicted volumes and TEUs		Actual volumes and TEUs	
		Volume	TEUs	Volume	TEUs
#1	25	51.1	2	59.3	2
	26	19.5	1	21.1	1
	27	6.01	1	3.9	1
	28	13.2	1	14.3	1
#2	37	5.8	1	7.9	1
	38	50.5	2	54.1	2
	39	27.8	1	33.8	2
	40	3.3	1	4.6	1
#3	34	26.9	1	29.3	1
	35	42.2	2	44.5	2
	36	36.4	2	31.2	1
	37	9.6	1	12.7	1
Total		292.3	16	298	15

Table 27 reveals that overall, the accuracy of shipment volume prediction is satisfactory. Although occasional discrepancies between predicted and actual volumes are observed, the total number of TEUs remains consistent, except for order #2 in week 39 and #3 in week 36. However, regarding the designated week, we do not perceive any transfer of volumes to different weeks. Additionally, it is crucial for the carrier that the arrival occurs within the designated week.

5.4 Comparison – old forecast vs new forecast

Now we compare the current forecasting approach with the forecasting model developed in this research and the actual number of containers needed. Due to limited available data, our comparison is somewhat constrained. For the current forecasting method, we examine the shipments from the previous year (2022) for a specific contracted flow from port to port. According to the contract, 457 containers are scheduled to be shipped at contracted prices for this flow. The forwarder aggregates this information on a weekly basis and reserves approximately 9 containers per week on the vessel. However, the actual number of containers for this flow averages around 12 containers per week. Shipping these additional 3 containers incurs extra costs for SLN. To minimize fluctuations in the container count, it would be beneficial to have this information in advance.

In order to evaluate the variability using the new forecast, we inputted the available limited historical data from 2022 into the model. Figure 12 offers an overview of the comparison between the predicted number of containers (using the old and new forecast) and the actual number of containers for specific weeks. For all the numeric values behind Figure 12 and determining the MSE, we refer to Appendix VII. Notably, during the final 10 weeks of 2022, the forecast demonstrates a satisfactory performance. Previously, the old forecasting method consistently predicted 9 containers for all weeks, which led to a significant discrepancy compared to the actual number of containers. This results in a high variance and a MSE of 59.7 TEUs. However, with the implementation of the new forecasting method, we have successfully reduced the MSE to a significantly lower value of 1.6 TEUs. We observe that the variance between the predicted and actual number of TEUs using the new forecasting method is minimal. However, the variance tends to increase when the number of TEUs is relatively high, which aligns with logical expectations. As the shipment volume for a given week becomes larger, the likelihood of fluctuations in lead time and packaging volume also tends to rise.

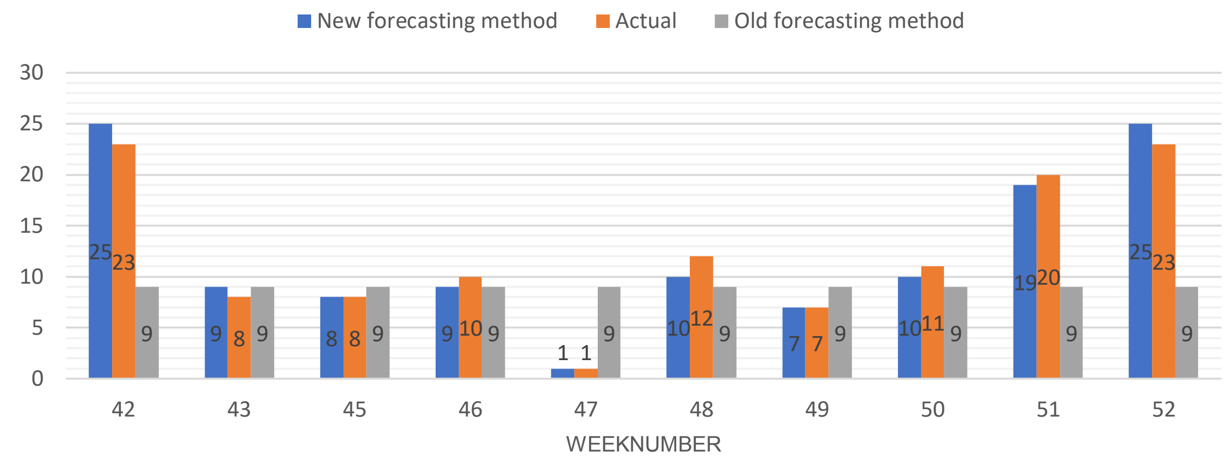


Figure 12: Comparison between predicted and actual number of TEUs

Based on the limited data at our disposal, we observe a reduction in the variance of container numbers between the current forecasting approach and the proposed new forecasting approach. This variance serves as an indicator for the additional costs incurred when shipping extra containers. As indicated in Section 1.2, we find that expenses for booking containers on the spot market are, on average, 20% higher. By decreasing the variance, we can also lower the costs associated with shipping these additional containers.

5.5 Newsvendor problem

SLN can utilize the principles of the newsvendor problem to guide their decision-making process when reserving TEUs on vessels. This approach requires a probability distribution of the expected number of TEUs arriving in a given week. For example, there is a 60% probability that the number of TEUs in a specific week is 2. In order to approximate this analysis, we assessed the variability of the forecasting error in the comparison of the previous Section. The error observed for the new forecasting method in this analysis spans from -2 (indicating an underestimation of 2 TEUs compared to the actual number) to +2 (indicating an overestimation of 2 TEUs compared to the actual number). We determined the frequency of each error by counting the number of occurrences and dividing it by the total number of observations. The findings of this analysis are presented in Table 28. We can incorporate these findings as follows: for instance, if the forecast predicts that the number of TEUs for a specific week is 5, there is a 30% chance that the actual number of TEUs is indeed 5 and a 10% chance that it is 7. The expected number of TEUs for this example is $3 * 0.20 + 4 * 0.1 + 5 * 0.3 + 6 * 0.3 + 7 * 0.10 = 5.0$.

Table 28: Probability distribution number of TEUs

Error	***Example***	***Probability***
-2	*3*	*20%*
-1	*4*	*10%*
0	*5*	*30%*
1	*6*	*30%*
2	*7*	*10%*

For the newsvendor problem, we have to incorporate the cost of understocking (Cu) and the cost of overstocking (Co). In case SLN chooses not to pre-book the required TEUs, the shipping costs for these additional units is 20% higher. Assuming that the cost of shipping one TEU from a specific port is €2,000,-, this translates to an extra cost of €400,- per TEU when opting for the spot market. Therefore, the Cu is €400,-. We assume that the Co is €200,-. Considering the disparity between Cu

and Co, it is prudent to consistently pre-book a higher number of TEUs than the expected quantity. This approach is favored as it is significantly more disadvantageous to encounter a shortage. Consequently, we advise rounding up the expected number of TEUs. We use the CSL to determine the optimal number of TEUs (O^*) to pre-book. Our objective is to identify the minimum number of TEUs where the probability of the actual demand (D) being lower than our optimal quantity, is at least equal to the CSL. We can express this concept using the following equations: $P(D \leq O^*) \geq CSL$, where the cycle service level (CSL) can be determined in the following way:

$$CSL^* = \frac{C_u}{C_u + C_o} = \frac{400}{400 + 200} = 0.667 \approx 66.7\%$$

We aim to meet a CSL of 66.7% and assess the O^* that satisfies this threshold. To accomplish this, we test various values of $O^*$and calculate the probability of the demand being lower or equal to O^*. To illustrate this process, we utilize the example provided in Table 28, where we determine the probabilities for different TEU quantities, including the actual number being 5 or higher. For instance, if we choose O^* as 3, the probability $P(D \leq 3)$ is 10%, which does not exceed the CSL requirement. We repeat this analysis for all O^* values listed in Table 28. Upon examination, we find that a value of O^* equal to 6 satisfies the CSL criterion. Therefore, we recommend SLN to pre-book 6 TEUs based on this forecast. If the probability distribution for this particular newsvendor problem remains consistent and the CSL threshold remains at 66.7%, the conclusion remains constant: always pre-book an additional TEU beyond the forecasted amount to meet the CSL requirement.

Table 29: Find the number of TEUs that meets the CSL criterion.

$P(D \leq 4) = 30\%$	*Lower than our CSL, so increase O^**
$P(D \leq 5) = 60\%$	*Lower than our CSL, so increase O^**
$P(D \leq 6) = 90\%$	*Higher than our CSL, so stop. The O^* = 6*

In summary, SLN can benefit from implementing the principles of the newsvendor problem to enhance their decision-making process. By utilizing the approach outlined in this Section, SLN can determine the probability-based wisdom of pre-booking a particular number of TEUs at a specific port of loading on a weekly basis. By adopting this approach, we increase the likelihood of consistently having sufficient space on the vessel to accommodate the containers and achieve our CSL. Additionally, when we ensure that we have reserved enough capacity in advance, we can reduce the cost of understocking Cu.

5.6 Conclusion

The best-performing ML method, for predicting volumes in the case of non-Scania packaging, is LightGBM. It slightly surpassed the baseline and XGBoost models in terms of accuracy and it is relatively simple to understand and implement. Additionally, we present the final forecasting model and conducted test runs, using unique historical orders, to evaluate its performance. The model successfully predict the expected shipment volumes and corresponding week of arrival at the port of loading for these orders. The final forecasting model based on LightGBM can help SLN make more accurate volume predictions and enhance their logistics operations for non-Scania packaging suppliers.

We run the final forecasting model with three unique orders at three different suppliers. We compare the outcomes of these runs (the number of TEUs) with the actual number of TEUs and observe that only for two orders there was a deviation of 1 TEU between the predicted an actual number of TEUs. Therefore, the accuracy of forecasting method is satisfactory. The performance comparison between the current forecasting method and the new forecasting method shows that the current way of forecasting resulted in a MSE of 59.7 TEUs. With the implementation of the new forecasting method, we successfully reduce the MSE to a significantly lower value of 1.6 TEUs. This reduction in variance is indicative of lower costs associated with shipping extra containers, as booking containers on the spot market incurs higher expenses.

By applying principles from the newsvendor problem, SLN can make proactive decisions to ensure sufficient equipment reservation without excessive booking. Using the Cu and Co we determined a CSL of 66.7%, and we determine the optimal number of containers that we need to pre-book in order to meet this CSL. If the probability distribution and the CSL remains the same, the conclusion for SLN is to always pre-book an additional TEU beyond the forecasted amount. By minimizing fluctuations and improving forecast accuracy, SLN can effectively decrease costs and optimize their container reservation strategy. Further research and evaluation with additional data would provide a more comprehensive understanding of the benefits and potential of the proposed forecasting model.

6 Implementation

In Chapter 5, we determined the optimal forecasting model for SLN, and now it is important to implement it within SLN. This chapter discusses the implementation process. We begin in Section 6.1 by outlining the necessary actions to ensure the accuracy and usability of the data sources required for the forecast. In the subsequent Section 6.2, we provide a detailed description of how this forecast is constructed. Section 6.3 covers the evaluation of the initial prototype and the subsequent steps needed for its integration into SLN. The conclusion of this chapter, presented in Section 6.4, answers the following subquestions.

Improve

In what way can this supposed carrier shipment forecast successfully be implemented?

7. *How should the process look that creates a carrier shipment forecast, as automated as possible?*
8. *The first prototype; how does it work?*

6.1 Data preparations

Previously, we conducted data preparations specifically tailored to historical shipments to train the ML methods. In this Section, we examine the essential steps that SLN needs to take for each data source to prepare them for the final forecast design. We also address any anticipated barriers in implementing these steps and propose actions to overcome them.

6.1.1 Order information

The data used in this forecast is sourced from the ERP system. The preparation of this data primarily involves ensuring its correct formatting (dates, numeric values) and eliminating unnecessary columns. The identified barrier for this particular source is the requirement for SLN to maintain these manual steps over an extended period. To address this obstacle, we have introduced automation for data preparation within the PowerBI dashboard itself, streamlining the process for SLN. They only need to ensure that an export from the ERP system is generated and placed in the designated directory. Consequently, we obtain a dataset that consists of the summarized columns listed in Table 30.

Table 30: Data - Order information

Part number	***Part category***	***Delivery date***	***Supplier name***	***Quantity***	***Packaging type***	***Number of shipping units***
1234	*X*	*1-6-2023*	*X*	*500*	*D2*	*20*

6.1.2 Supplier information

This dataset contains all the shipping information for each supplier. Within SLN, it is essential to consolidate all the independent sources into a single table that encompasses all the information specified in Table 31. The provided information is static, and SLN should implement any necessary changes if they arise. The barrier encountered with this specific source is the infrequency of changes to the associated file, which may result in SLN overlooking the need for regular updates. To ensure the information in this data remains up-to-date, it is imperative for SLN to establish a routine and maintain a procedure in updating it consistently, even if changes are infrequent and the data remains unchanged for extended periods.

Table 31: Data – Supplier information

Supplier name	***Primary forwarder***	***Port of loading***	***Scania packaging?***	***Frozen period***
X	*Forwarder name*	*Rotterdam*	*Yes/No*	*14 days*

6.1.3 Packaging information

This data should include the volumes of different packaging types used within SLN. If the "Scania packaging" column in the Supplier information dataset states "Yes" for a particular supplier, we can calculate the shipment volume by multiplying the number of packages by the volume of one packaging. If the column returns the Boolean value "No," we utilize the ML method to predict the shipment volume for that supplier line. In addition, the packaging information is characterized by its static nature, with infrequent changes occurring over time. It is expected that this information is provided by a different team within the organization. Similar to the barrier observed with supplier information, the challenge with this specific source lies in the limited frequency of updates, leading to a potential oversight of the need for regular updates. To address this, SLN should establish a systematic procedure for updating this data, ensuring that it remains accurate and up-to-date. An example of this data can be seen in Table 32.

Table 32: Data – Packaging information

Packaging type	***Volume of one package***
D2	*2.5 m^3*

6.2 Creating the forecast

After completing the data preparations described in Section 6.1, we integrate all the datasets into the forecast. We have optimized the forecast to operate automatically and require minimal effort from the team MSE. In this Section, we provide a detailed description of the process for creating and updating the forecast.

To facilitate the forecast, a Cloud environment has been established, which serves as the backdrop for the forecast and stores all the data files mentioned in Section 6.1. This Cloud environment interfaces with the forecast and acts as the input parameters. The MSE team, responsible for the forecast, needs to ensure that the data in these files remains up-to-date. Every workday at 9:30 AM, the forecast reloads the data from these sources and recalculates the volumes and TEUs, while also performing aggregations on a port and supplier level. We have adjusted this time to ensure there is enough time for manual updates to the sources.

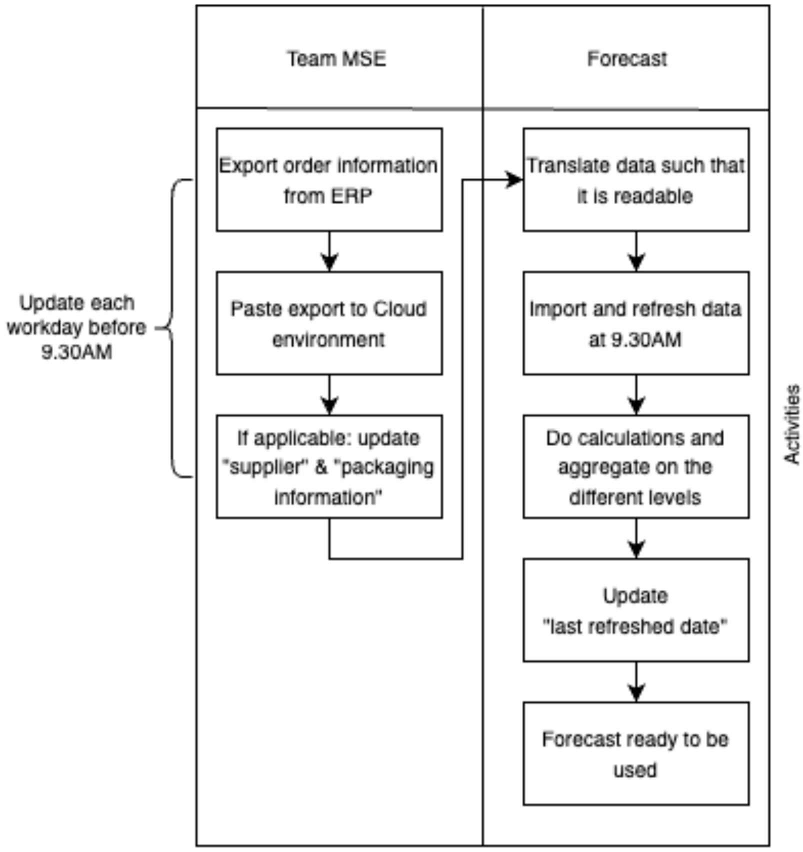


Figure 13: Activities needed to get forecast

The order information file is the data source that is subject to the most significant changes. A daily export of this data from the ERP system must be transferred to a designated file in the Cloud environment to ensure compatibility with the forecast. As for the other two data files, changes do not occur on a daily basis and should be reviewed regularly. Further details regarding this is discussed in Section 7.1. Figure 13 provides an overview of the activities required by each role to obtain the forecast.

6.3 First prototype

Throughout our research and thesis writing, we have simultaneously worked on developing an initial prototype of the forecast. This prototype is built using the PowerBI dashboard builder. It utilizes the three discussed data sources mentioned in Section 6.1 and incorporates the steps outlined in Section 4.2. Please note that the forecast displayed in this prototype is an example and not the actual forecast due to the confidential nature of the data.

6.3.1 How does it work?

The prototype comprises two pages. The first page (refer to Figure 14) provides users with background information on how the forecast operates, while the second page serves as the main dashboard for the forecast (see Figure 15). At the top, there is a gray area containing four filters. Users can apply filters based on the port of loading, the country and postal code of the supplier, as well as the desired incoterms. Additionally, a button allows switching between port-level and supplier-level aggregation (Figure 16), and an erase filter button has been included.

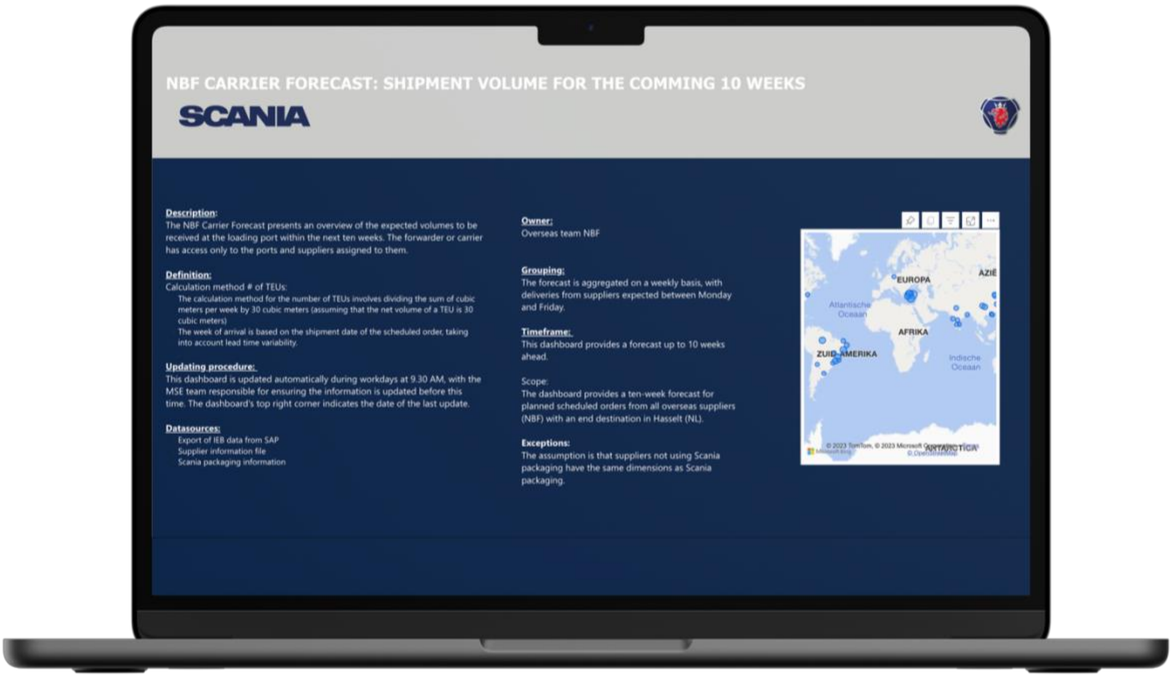


Figure 14: First page PowerBI dashboard

The table in the top left corner displays the backlog of shipments that have not been confirmed as shipped and whose delivery date is 10 weeks in the past. These shipments have the potential to arrive on short notice, so it is crucial for the forwarder to be aware of them. The geographical view illustrates the locations of ports or suppliers, with bubble size indicating the shipment volumes for the upcoming 10 weeks. The table in the top right corner presents the volumes and number of TEUs for each port or supplier for the next 10 weeks, while the figure in the bottom right corner visualizes these numbers in a bar chart.

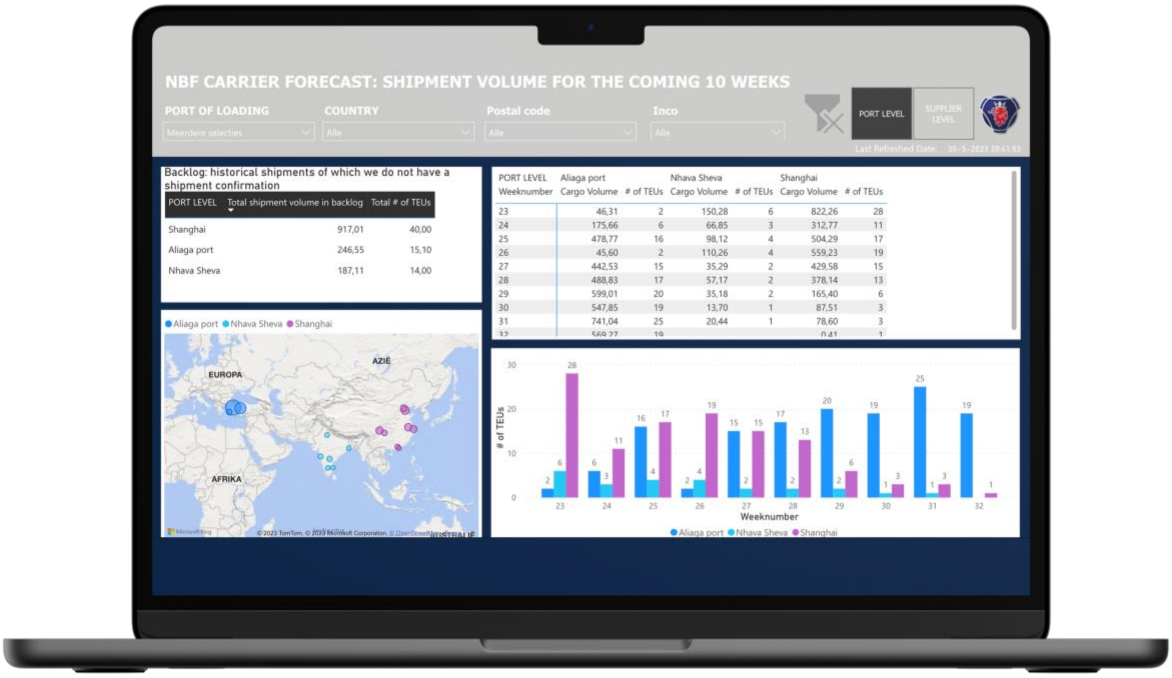


Figure 15: Second page PowerBI dashboard – Port level

Users can apply filters to the dashboard using the filters in the gray area, or they can utilize the tables and diagrams in the blue area. By clicking on a specific week, for instance, all other visualizations adjust accordingly to reflect that filter. The same applies to selecting a particular geographical location or supplier. This PowerBI dashboard utilizes row-level security, allowing data access to be restricted for specific users. For example, forwarders can only view the ports and suppliers assigned to them.

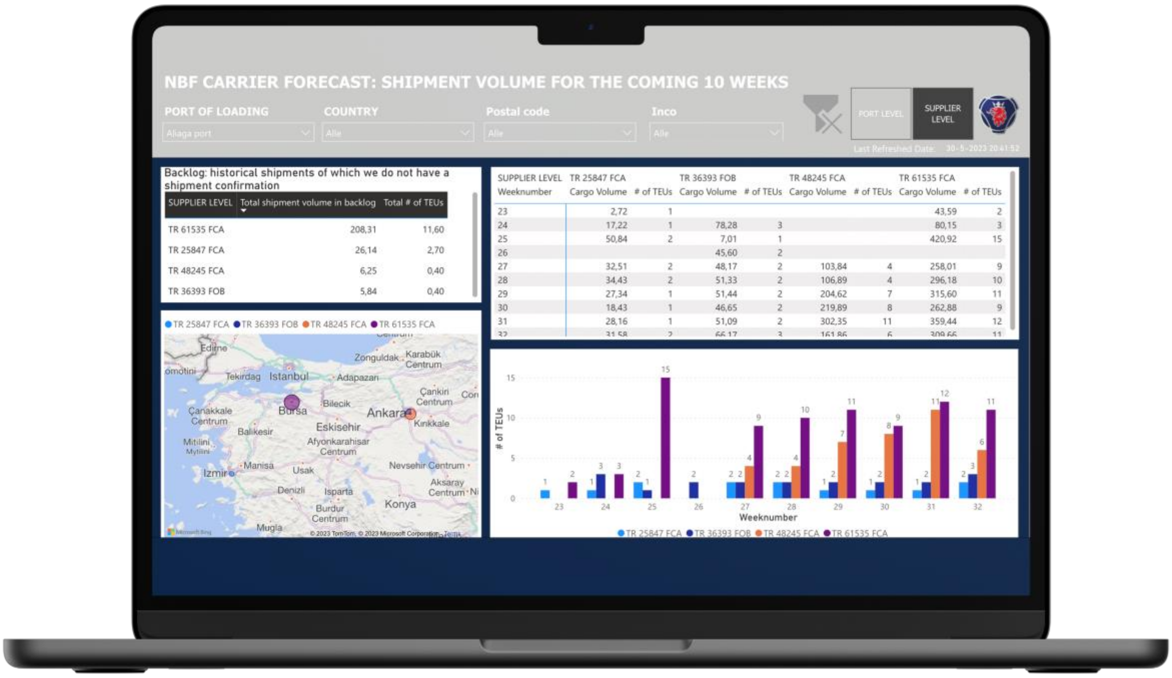


Figure 16: Second page PowerBI dashboard – Supplier level

6.3.2 Future steps

Due to time constraints during the research, we were unable to fully fine-tune the prototype forecast to create the definitive forecast for deployment within and outside the organization. Some elements are not yet implemented in this prototype, requiring additional steps.

ML method

Currently, the prototype does not employ the ML method to predict the shipment volume for suppliers that do not use Scania packaging. It assumes that Scania packaging is applicable for all scheduled lines and can be determined deterministically. The LightGBM method we developed within R-studio can be exported and imported as a separate module within the PowerBI dashboard. Documentation is available on exporting this model and importing it into PowerBI, although implementing it can be relatively time-consuming.

Lead-time variability

The prototype does not currently incorporate lead-time variability analysis, meaning the week of the delivery date is assumed to be the arrival date at the port of loading. To incorporate lead-time variability, SLN can utilize a similar model from R-studio as used for the ML method. There is documentation available that describes multiple methods for implementing R-studio scripts within PowerBI. Implementation of this process is complex due to the requirement of developing a script that consolidates the determined shipment volume for a scheduled order across multiple weeks, taking into account the proportions derived from the lead-time variability analysis.

6.4 Conclusion

In conclusion, the text discusses the data preparations and the first prototype of a forecast design for shipment volumes. The data preparations involve ensuring the correct format and necessary columns for each data source, including the worksheet forecast file, supplier information, and packaging information. A Cloud environment has been established as the foundation, interacting with the forecast and storing the relevant data files. The MSE team must ensure the data in these files remains current. Notably, the order information file undergoes the most significant changes and requires a daily export from the ERP system to maintain compatibility with the forecast.

The first prototype is built using PowerBI and incorporates the three data sources. The dashboard includes filters for port of loading, supplier location, and Incoterms, allowing users to customize their view. It displays information such as backlog shipments, geographical view of ports and suppliers, shipment volumes, and TEUs in tables and diagrams. Although the prototype provides valuable insights, there are future steps to be taken. One aspect to address is the implementation of the ML method for predicting shipment volume for suppliers not using Scania packaging. Another aspect to consider is lead-time variability analysis, which is not currently incorporated in the prototype. In summary, while the prototype provides a foundation for the forecast design, further development is necessary to fully utilize the ML method and incorporate lead-time variability analysis. These improvements will contribute to a more accurate and comprehensive forecast for shipment volumes within SLN.

7 Controlling the forecast

Up to now we have developed a protype forecast and described the steps how to construct this forecast. However, in order to maintain the quality of the forecast, it is important to plan some activities. In 7.1 we describe the updating processes for the different sources that we use in this research. Lastly, in 7.2 we plan these activities in a planning horizon of one year. The conclusion of this chapter, presented in Section 7.3, answers the following subquestions:

Control

Which actions have to take place to maintain the high-quality carrier shipment forecast?

9. *What processes should be designed and implemented within SLN?*
10. *At which interval does this carrier shipment forecast have to be reviewed again?*

7.1 Updating processes

Apart from the updating procedure of the forecast itself in 6.2, it is also important to pay attention to the sources on which this forecast is based. Therefore we describe in this Section the elements that need to be updated on a regular basis.

7.1.1 ML method

The LightGBM ML method is used to give a prediction of the shipment volumes for the suppliers that do not use Scania packaging. Based on the historical dataset of one specific carriers, we have seen that the LightGBM method was the ML method with the highest accuracy. There is however the possibility that if we use a new dataset from a new year for example, we can possibly see that a different ML method performs better. Therefore, it is important to do this analysis on a regular basis in order to make sure we are using a method that offers us the highest accuracy possible. We would recommend to use the same approach as described in 4.3 and use their corresponding code which can be found in the appendix. By using this approach, you have a good comparison with the analysis of this research and future analysis.

Non-Scania packaging analysis

By the time of writing, SLN has started a project to find out for each supplier that does not use Scania packaging, how each part is shipped and what are the dimensions of the packaging they use. This is a project that has started recently and would take approximately a year to finalize this. Depending how accurate this analysis is, this information can be used instead of the ML method. In principle, the lead-time variability is then the only stochastic part in the forecast.

7.1.2 Supplier information

The supplier information is also a source that is subject to changes and therefore should be reviewed on a regular basis. Below there are some example situations where updates are necessary to the supplier information file.

Introduction of a new supplier

Now and then SLN is going to source a part from a new supplier. In principle if a new supplier is introduced, a new row with this supplier can be added to data. For this new supplier we have to collect and determine all the data that is necessary for the carrier forecast. We sum the most important ones:

- Postalcode and City;
- Incoterms;
- Port of Loading that is used (usually determined based on the geographical location of the supplier);

- Primary carrier;
- Scania packaging or non-scania packaging.

Relocation
If the supplier relocates to a different location, this should also be updated in this table. Especially when a different port of loading or carrier changes.

Tender implementation
If SLN implements a new tender, this means that a specific carrier is assigned to transport goods from the supplier to a specific destination. This updating procedure results that the carrier in question should (only) see the suppliers that are assigned to this carrier. From the moment that these changes are applied to the data, these changes are considered the next time the forecast automatically updates. So, this should be done only when this new tender actually goes live.

Lead-time variability
The lead analysis we performed in this research, should also be reviewed on a regular basis. It is however a temporary analysis, since SLN can use the vendor evaluation module, discussed in 2.1.4, in the future.

7.1.3 Packaging information

The data for the different packaging types with the corresponding dimensions and volumes that SLN uses, should also be updated on a regular basis. For example if changes to existing packaging takes place or the introduction of a new packaging type. This information is not available in a direct system and should be requested from a specific department.

7.2 Time window

Since now we have determined the sources that need an update on a regular basis, it is now important to plan these activities in a planning horizon of one year. In Table 33 you see when the updating of these sources should take place.

Table 33: Time window when the different elements need to be updated

	Jan	***Feb***	***Mar***	***Apr***	***May***	***Jun***	***Jul***	***Aug***	***Sep***	***Okt***	***Nov***	***Dec***
ML method	X											
Supplier information				X				X				X
Packaging information				X				X				X

ML method
The data on which the ML method is based on yearly basis, so SLN does receive this report from the carrier in January. Therefore, it has no affect do to this analysis multiple times a year and should therefore only be done in January. The updating procedure, and the use of the ML method, can be removed when SLN has finished the analysis for non-Scania packaging suppliers.

Supplier information
The updating procedure of the supplier information should be done 3 times a year. This is in principle taking a new export from the ERP system, checking if the information for the forwarder and port of loading are still up to date and doing the lead-time analysis again. The lead-time analysis can be dropped if the vendor evaluation module is ready to use.

Packaging information
The packaging information should also be updated 3 times a year. This in principle requesting this data (again) from the specific departments, make sure the format is the same as the previous ones and placing it in the Cloud-environment.

7.3 Conclusion

In this chapter, we discussed the importance of updating various elements in the forecasting process on a regular basis. These updates are necessary to ensure accurate and reliable forecasts. Firstly, we highlighted the need to periodically assess the performance of the ML method used for predicting shipment volumes of suppliers not using Scania packaging. It is recommended to regularly analyze the performance of different ML methods using new datasets to ensure the highest possible accuracy. Furthermore, we discussed the ongoing project of analyzing non-Scania packaging for each supplier. This project aims to determine how each part is shipped and the dimensions of the packaging used. Depending on the accuracy of this analysis, the information gathered from it can be utilized instead of the ML method.

We also emphasized the importance of updating supplier information regularly. Examples of situations that require updates include the introduction of new suppliers, supplier relocations, tender implementations, and changes in lead-time variability. Additionally, the packaging information, including dimensions and volumes, should be updated periodically. To establish a time window for these updating processes, we recommended specific intervals. The ML method update should be conducted once a year. Supplier information and packaging information should be updated three times a year. In conclusion, regular updates of the ML method, supplier information, and packaging information are essential for maintaining accurate and reliable forecasts. By adhering to the recommended updating procedures, SLN can ensure that the forecasting process remains up-to-date.

8 Conclusion, discussion and recommendations

This chapter gives the conclusion, discussion and recommendations of this research that is created by SLN. The main research question was: *In which way can the order information of suppliers be translated to provide a 10-week ahead carrier shipment forecast, that provides insight into the required transport capacity of the carriers?* This research was structured according to the DMAIC process model, with each phase addressing specific subquestions related to the main research question. By following this approach, the core problem was effectively addressed.

8.1 Conclusion

The development of a high-quality carrier shipment forecast for SLN requires careful consideration of available information, current forecasting practices, and specific model requirements. The utilization of the ERP system's order information forms the basis of the forecast, while customizing it for each carriers involves establishing a table of relations based on historical data. To construct a carrier forecast, suitable methods were identified based on a literature review. ML methods, such as LightGBM, XGBoost, and N-beats, showed promising results for predicting shipment volumes. It is recommended to combine different methods to address specific elements within the forecast, such as non-Scania packaging volumes. Incorporating packaging uncertainty, lead time variability, and the frozen period is crucial for accurate forecasts.

The flow of forecasting was discussed, providing a step-by-step guide for designing the forecast. ML methods were introduced as a solution for predicting packaging volumes for non-Scania packaging suppliers. The best-performing ML method, LightGBM, demonstrated accuracy in predicting volumes for non-Scania packaging suppliers. The final forecasting model based on LightGBM can enhance SLN's logistics operations. Proactive decisions, based on principles from the newsvendor problem, can ensure sufficient equipment reservation without excessive booking. The CSL indicated that it is prudent to consistently overbook by one TEU. The proposed forecast approach showed a reduction in variance and can lead to cost optimization by minimizing fluctuations and improving forecast accuracy.

The prototype incorporated three data sources and provided valuable insights, but further development is necessary. Implementing the ML method and incorporating lead-time variability analysis are important steps for a more accurate and comprehensive forecast. Updating processes were emphasized to ensure accurate and reliable forecasts. Regular assessment of the ML method, analysis of non-Scania packaging, updates to supplier information, and packaging information are necessary. Specific intervals for these updates were recommended. Regular updates of the ML method, supplier information, and packaging information are essential for maintaining accuracy. By incorporating the recommended updates and addressing the identified requirements, the developed carrier shipment forecast has the potential to greatly enhance SLN's logistics operations and decision-making processes. Continual improvement and further research contributes to a more accurate and comprehensive forecast for shipment volumes within SLN.

8.2 Discussion

One of the major challenges encountered in this research was the selection of suitable models and methods to test. Despite conducting a comprehensive literature study, there was no single model that directly addressed the specific requirements of this research. The development of a shipment volume carrier forecast involved considering multiple sources of information and various elements, necessitating the creation of a combined forecasting flow using multiple methods. The chosen methods were based on their promising performance in the literature study, but it is acknowledged that there are still numerous other models that could be tested to potentially improve accuracy. However, due to time constraints, the research was limited to the selected models and methods.

Another aspect of discussion is the limited availability of data for conducting this research. Efforts were made from the beginning to quantify the research and establish a benchmark by comparing the current situation with the proposed new forecast method. However, quantifying this benchmark proved challenging. Ultimately, the focus shifted to comparing the number of pre-booked TEUs with the actual number of shipped TEUs. While it cannot be confirmed with 100% certainty that SLN paid spot-rates for the additional TEUs, it is understood that having this information in advance would contribute to a more stable supply chain. Data limitations were not only encountered in defining the benchmark but also in training the ML method, analyzing contracted versus spot rates, and gathering packaging information.

Regarding the implementation of the prototype, the aim was to have a final model that could operate automatically to a significant extent. However, due to time constraints, setting up the necessary data structure took more time than anticipated and could not be fully completed. Nevertheless, the steps for further implementation within SLN are well-defined, and the initial prototype already serves as a valuable tool.

8.3 Recommendations

We propose some recommendations as well as possible future research possibilities:

- **Data Collection**: as discussed in the previous Section, it is advisable for SLN to gather and store additional data from historical events. This expanded dataset can serve as a valuable resource for future analyses and research.
- **Present to carriers**: while the current prototype has been developed and shared internally, it is recommended to present the prototype to one or more carriers. Since they will be working directly with the forecast, receiving feedback from them would be highly valuable and appreciated.
- **Integration of weight Information**: During the final phase of this research, SLN expressed the desire to incorporate the expected weight of shipments in the forecast. Although weight is less crucial than volume, it can provide insights into the required number of TEUs. Including this information in the forecast is recommended.
- **Updating procedures**: it is advised that SLN follows the updating procedures discussed in Section 7.1. Regularly updating the ML method, supplier information, and packaging data is essential to maintain accurate and reliable forecasts.
- **Vendor Evaluation Module**: in Section 2.1.4, we introduced the vendor evaluation module, which could be highly beneficial for future applications, including its integration into the carrier forecast. SLN should explore the potential opportunities of utilizing this module and investigate whether supplier reliability, as measured by lead time variability, can be incorporated into the carrier forecast.
- **Non-Scania Packaging Analysis**: SLN has recently initiated an analysis of the packaging utilized by non-Scania packaging suppliers. Once this research is completed, the findings can be used as an input source for the carrier forecast. However, SLN should establish a procedure for regularly updating this information to ensure its accuracy and relevance.

9 Bibliography

Aggarwal, K., Mijwil, M., Sonia, Al-Mistarehi, A.-H., Alomari, S., Gök, M., . . . Abdulrhman, S. (2022). Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning. *Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning*.

Archer, B. (1980). Forecasting demand: Quantitative and intuitive techniques. *International Journal of Tourism Management*, 5-12.

Archer, B. (1980). *International Journal of Tourism Management*, 5-12.

Arnx, A. (2019). First neural network for beginners explained (with code). *Towards Data Science*.

Atwater, J., & Pittman, P. (2022). *Supply Chain Dictionary* (17 ed.). ASCM.

Brockwell, P., & Davis, R. (2016). *Introduction to Time Series and Forecasting.* Springer.

CBS. (2023). *Twenty Foot Equivalent Unit (TEU)*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 78-87.

Embassy Freight Services. (2018, 1 18). *Varklaring FCL en LCL*. Retrieved from https://www.embassyfreight.nl/nieuws/verklaring-fcl-en-lcl/

Fattah, J., Ezzine, L., Aman, Z., Moussami, H., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Bussiness Management*, 1-9.

Garrido, R. (2007). Procurement of transportation services in spot markets under a double-auction scheme with elastic demand. *Transportation Research Part B 41*, 1067–1078.

Gorman, M., Clarke, J.-P., Gharehgozli, A., Hewitt, M., Koster, R., & Roy, D. (2014). State of the Practice: A Review of the Application of OR/MS in Freight Transportation . *Informs*, 535-554.

Graves, S. (2011). Uncertainty and Production Planning.

Hsu, H.-P., Tai, H.-H., Wang, C.-N., & Chou, C.-C. (2021). Scheduling of collaborative operations of yard cranes and yard trucks for export containers using hybrid approaches. *Advanced Engineering Informatics*, 2.

Hyndman, R. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, 7-14.

Hyndman, R., Koehler, A., Snyder, R., & Grose, S. (2002). A state space frame- work for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 439-454.

James , G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning.* (2nd, Ed.) Springer.

Kanade, V. (2022, 12 14). *What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022*. Retrieved from Spiceworks: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/

Khandelwal, E. (2020, 03 27). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

Khandelwal, E. (2020, 3 27). *Which algorithm takes the crown: Light GBM vs XGBOOST?* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

Knolla, D., Prüglmeierb, M., & Reinharta, G. (2016). Predicting Future Inbound Logistics Processes using Machine Learning. *Procedia CIRP 52*, 145-150.

Koch, M. (2018). An Optimisation and Forecast Framework for ULD Packing in the Air Cargo Supply Chain.

Lee, H., & Whang, S. (2000). Information sharing in a supply chain. *International Journal of Technology Management*, 373-387.

Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, 451-476.

Makridakis, S., & Spiliotis, E. (2022). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 1325-1336.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., . . . Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 111-153.

Makridakis, S., Chatfield, C., Hibon, M., & Lawrence, M. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 5-22.

Makridakis, S., Fry, C., Petropoulos, F., & Spiliotis, E. (2020). The Future of Forecasting Competitions: Design Attributes and Principles. *INFORMS Journal on Data Science*, 96-113.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 54-74.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 1346-1364.
National Research Council. (1995). *High-Purity Chromium Metal: Supply Issues for Gas-Turbine Superalloys.* Retrieved from https://doi.org/10.17226/9248.
Onnen, H. (2022, 1 10). *N-BEATS Unleashed: Deep Forecasting Using Neural Basis Expansion Analysis in Python*. Retrieved from Towards Data Science: https://towardsdatascience.com/n-beats-unleashed-deep-forecasting-using-neural-basis-expansion-analysis-in-python-343dd6307010
Oracle. (2023, 3 27). Data Mining User's Guide.
Pal, A. (2020, 10 1). *Gradient Boosting Trees for Classification: A Beginner's Guide*. Retrieved from Medium: https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea
Pujawan, N. (2001). Schedule nervousness in a manufacturing system: A case study. *Production Planning & Control*, 515-524.
Ramshorst, H. (2022). *Forecasting Transportation Volumes at Farm Trans.*
Rao, S. (2021, 8 22). *XGBoost Regression: Explain It To Me Like I'm 10*. Retrieved from Towards data science: https://towardsdatascience.com/xgboost-regression-explain-it-to-me-like-im-10-2cf324b0bbdb
Reinau, K., Østergaard, C., & Karam, A. (2021). Horizontal collaboration in the freight transport sector: barrier and decision-making frameworks. *European Transport Research Review*, 13-53.
Saha, S. (2023, 1 25). *XGBoost vs LightGBM: How Are They Different*. Retrieved from Neptune.ai: https://neptune.ai/blog/xgboost-vs-lightgbm
Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*.
Scania. (2021). *Annual and Sustainability Report.* Södertälje, Sweden.
Shang , K.-C., & Lu, C.-S. (2012). Customer relationship management and firm performance: An empirical study of freight forwarders services. *Journal of Marine Science and Technology*, 67-72.
Smolic, H. (2022, 12 15). *How Much Data Is Needed For Machine Learning?* Retrieved from Graphite: https://graphite-note.com/how-much-data-is-needed-for-machine-learning#:~:text=Generally%20speaking%2C%20the%20rule%20of,100%20rows%20for%20optimal%20results.
Stadtler, H., & Kilger, C. (2000). *Supply Chain Management and Advanced Planning: Concepts, Models, Software and Case Studies.*
Starmer, J. (2020). XGBoost Part 1 (of 4): Regression.
Tatman, R. (2018). *Machine Learning with XGBoost (in R)* . Retrieved from Kaggle: https://www.kaggle.com/code/rtatman/machine-learning-with-xgboost-in-r#Tuning-our-model
Theisens, H. (2016). In *Climbing the Mountain* (3 ed.). Amstelveen: Lean Six Sigma Academy (LSSA).
Tuychiev, B. (2021). You Are Missing Out on LightGBM. It Crushes XGBoost in Every Aspect. *Towards Data Science*.
Wang, Y., & Tomlin, B. (2009). To Wait or Not to Wait: Optimal Ordering Under Lead Time Uncertainty and Forecast Updating. *Wiley InterScience*.
Zang, C., Tan, G.-W., Robb, D., & Zheng, X. (2006). Sharing shipment quantity information in the supply chain. *Omega*, 427-438.

Appendix I


Luleå
Lahti
St. Petersburg
Södertälje
Oskarshamn
Słupsk
Meppel
Zwolle
Angers
Beijing
Rugao (2022)
Taipei
Bengaluru
Kuala Lumpur
São Paulo
Tucumán
Johannesburg
Regional product centres
Production units
Research and development
Sales and services
Power solutions service network

Appendix II

Code for Multiple Linear Regression in R-studio

```
# libraries we're going to use
library(readxl)
library(MASS)
library(dplyr)
library(caTools)
library(Metrics)
library(glmnet)
library(caret)
library(modelr)
library(purrr)

#Clean Global Environment
rm(list = ls())

#Importing the dataset
df = read_excel("Database historical shipment volumes.xlsx")
#Drop columns that we do not need for the model
df = select(df, -Delivery, -VendorID, -Orderline, -CargoPieces, -POL, -PartNumber)
#Remove NAs
df <- na.omit(df)

#Translate categorical values into factors
df$VendorName = as.factor(df$VendorName)
df$MaterialDescription = as.factor(df$MaterialDescription)

# set a random seed
set.seed(1234)

attach(df)
train_test_split = sample.split(CargoVolume, SplitRatio = 0.8)

#The actual train and test set
train_X = subset(df, train_test_split == TRUE)
train_y = train_X$CargoVolume
test_X = subset(df, train_test_split == FALSE)
test_y = test_X$CargoVolume

#Building the model with the required predictors
lr <- lm(CargoVolume ~., data = df)
summary(lr)

#Start measuring running time
start_time <- Sys.time()

# Define training control
train.control <- trainControl(method = "cv", number = 10)
# Train the model
model <- train(CargoVolume ~ MaterialDescription + VendorName, data = train_X, method = "lm",
        trControl = train.control)
# Summarize the results
print(model)

#Gathering the predicted class labels and the actual class labels
```

```
pred_tt = predict(model, test_X)

test_mse = Metrics::mse(test_y, pred_tt)
test_mae = Metrics::mae(test_y, pred_tt)
test_rmse = Metrics::rmse(test_y, pred_tt)
test_bias = Metrics::bias(test_y, pred_tt)

cat("MSE: ",test_mse, "\nMAE: ", test_mae,"\nRMSE: ", test_rmse, "\nBias: ", test_bias)

#Stop measuring running time and print running time
end_time <- Sys.time()
end_time - start_time
```

Appendix III

Code for XGBoost in R-studio

```
# libraries we're going to use
library(xgboost)
library(readxl)
library(dplyr)
library(tidyverse) # general utility functions
library(DiagrammeR)
library(caret)
library(microbenchmark)
library(Metrics)

#Clean Global Environment
rm(list = ls())

#Importing the dataset
df = read_excel("Database historical shipment volumes.xlsx")
#Drop columns that we do not need for the model
df = select(df, -Delivery, -VendorID, -Orderline, -CargoPieces, -POL, -PartNumber)
#Remove NAs
df <- na.omit(df)

#Translate categorical values into factors
df$VendorName = as.factor(df$VendorName)
df$MaterialDescription = as.factor(df$MaterialDescription)

# set a random seed
set.seed(1234)

attach(df)
train_test_split = sample.split(CargoVolume, SplitRatio = 0.8)

#The actual train and test set
train_X = subset(df, train_test_split == TRUE)
train_y = train_X$CargoVolume
test_X = subset(df, train_test_split == FALSE)
test_y = test_X$CargoVolume

###up to here it is the same as previous models

# get the vector of training labels
CargoVolume = df$CargoVolume

xgb.data.train = xgb.DMatrix(data.matrix(train_X[, colnames(train_X) != "CargoVolume"]), label = train_X$CargoVolume)
xgb.data.test = xgb.DMatrix(data.matrix(test_X[, colnames(test_X) != "CargoVolume"]), label = test_X$CargoVolume)

#Start measuring running time
start_time <- Sys.time()

xgb.bench = microbenchmark(
 xgb.model <- xgb.train(data = xgb.data.train
                        , params = list(objective = "reg:squarederror"
                                        , eta = 0.1
```

```
                                   , max.depth = 7 #deze aangepast...deeper XGBoost model to compare accuruacy
                                   , min_child_weight = 100
                                   , subsample = 1
                                   , colsample_bytree = 1
                                   , nthread = 3
                                   , eval_metric = "rmse"
                                   , tree_method = "hist"
                                   , grow_policy = "lossguide"
                          )
                          , watchlist = list(test = xgb.data.test)
                          , nrounds = 10000
                          , early_stopping_rounds = 1000
                          , print_every_n = 20
 )
 , times = 5L
)

print(xgb.bench)
print(xgb.model$best_score)

# Make predictions on test set
xgb.test.pred = predict(xgb.model
                      , newdata = data.matrix(test_X[, colnames(test_X) != "CargoVolume"])
                      , ntreelimit = as.double(xgb.model$best_ntreelimit)
                      )

test_mse = Metrics::mse(test_y, xgb.test.pred)
test_mae = Metrics::mae(test_y, xgb.test.pred)
test_rmse = Metrics::rmse(test_y, xgb.test.pred)
test_bias = Metrics::bias(test_y, xgb.test.pred)

cat("MSE: ",test_mse, "\nMAE: ", test_mae,"\nRMSE: ", test_rmse, "\nBias: ", test_bias)

#Stop measuring running time and print running time
end_time <- Sys.time()
end_time - start_time

# Check for the features that are the most important
xgb.feature.imp = xgb.importance(model = xgb.model)
print(xgb.feature.imp)

##########################################################################

#Parameter tuning: grid Search to find the best hyperparameter combinations
max.depths = c(2,4,6)       #depth of the tree, default is 3
etas = c(0.01, 0.2, 0.4)     #learning rates of the tree, , default is 0.1

best_params = 0
best_score = 0

count = 1
for( depth in max.depths ){
 for( num in etas){

   bst_grid <- xgb.train(data = xgb.data.train
                          , params = list(objective = "reg:squarederror"
                                   , eta=num #learning rates of the tree
                                   , max.depth = depth #depth of the tree
                                   , min_child_weight = 100
```

```
                              , subsample = 1
                              , colsample_bytree = 1
                              , nthread = 3
                              , eval_metric = "rmse"
                      )
                      , watchlist = list(test = xgb.data.test)
                      , nrounds = 10000
                      , early_stopping_rounds = 1000
                      , print_every_n = 20
  )

  if(count == 1){
   best_params = bst_grid$params
   best_score = bst_grid$best_score
   count = count + 1
  }
  else if( bst_grid$best_score < best_score){
   best_params = bst_grid$params
   best_score = bst_grid$best_score
  }
 }
}
best_params
best_score

#Start measuring running time
start_time <- Sys.time()

# Run the algorithm again with the best learning rate and max.depth of the tree
xgb.bench.tuned = microbenchmark(
 xgb.model.tuned <- xgb.train(data = xgb.data.train
                     , params = list(objective = "reg:squarederror"
                             , eta = best_params$eta
                             , max.depth = best_params$max_depth
                             , min_child_weight = 100
                             , subsample = 1
                             , colsample_bytree = 1
                             , nthread = 3
                             , eval_metric = "rmse"
                             , tree_method = "hist"
                             , grow_policy = "lossguide"
                     )
                     , watchlist = list(test = xgb.data.test)
                     , nrounds = 10000
                     , early_stopping_rounds = 1000
                     , print_every_n = 20
 )
 , times = 5L
)

print(xgb.bench.tuned)
print(xgb.model.tuned$best_score)

# Make predictions on test set
xgb.test.pred.tuned = predict(xgb.model.tuned
                   , newdata = data.matrix(test_X[, colnames(test_X) != "CargoVolume"])
                   , ntreelimit = as.double(xgb.model.tuned$best_ntreelimit)
)
```

```
test_mse = Metrics::mse(test_y, xgb.test.pred.tuned)
test_mae = Metrics::mae(test_y, xgb.test.pred.tuned)
test_rmse = Metrics::rmse(test_y, xgb.test.pred.tuned)
test_bias = Metrics::bias(test_y, xgb.test.pred.tuned)

cat("MSE: ",test_mse, "\nMAE: ", test_mae,"\nRMSE: ", test_rmse, "\nBias: ", test_bias)

#Stop measuring running time and print running time
end_time <- Sys.time()
end_time - start_time

# Check for the features that are the most important
xgb.feature.imp = xgb.importance(model = xgb.model.tuned)
print(xgb.feature.imp)
```

Appendix IV

Code for LightGBM in R-studio

```
# libraries we're going to use
library(pROC)
library(lightgbm)
library(microbenchmark)
library(Metrics)

#Clean Global Environment
rm(list = ls())

#Importing the dataset
df = read_excel("Database historical shipment volumes.xlsx")
#Drop columns that we do not need for the model
df = select(df, -Delivery, -VendorID, -Orderline, -CargoPieces, -POL, -PartNumber)
#Remove NAs
df <- na.omit(df)

#Translate categorical values into factors
df$VendorName = as.factor(df$VendorName)
df$MaterialDescription = as.factor(df$MaterialDescription)

# set a random seed
set.seed(1234)

attach(df)
train_test_split = sample.split(CargoVolume, SplitRatio = 0.8)

#The actual train and test set
train_X = subset(df, train_test_split == TRUE)
train_y = train_X$CargoVolume
test_X = subset(df, train_test_split == FALSE)
test_y = test_X$CargoVolume

###up to here it is the same as previous models

lgb.train = lgb.Dataset(data.matrix(train_X[, colnames(train_X) != "CargoVolume"]), label =
train_X$CargoVolume)
lgb.test = lgb.Dataset(data.matrix(test_X[, colnames(test_X) != "CargoVolume"]), label =
test_X$CargoVolume)

#Define the parameters for the LightGBM technique
params.lgb = list(
 objective = "regression",
 metric = "rmse",
 min_data_in_leaf = 1,
 min_sum_hessian_in_leaf = 100,
 feature_fraction = 1,
 bagging_fraction = 1,
 bagging_freq = 0,
 learning_rate = 0.1,
 num_leaves = 7,
 num_threads = 2
)

#Start measuring running time
```

```
start_time <- Sys.time()

lgb.bench = microbenchmark(
 lgb.model <- lgb.train(
  params = params.lgb
  , data = lgb.train
  , valids = list(test = lgb.test)
  , nrounds = 10000
  , early_stopping_rounds = 1000
  , eval_freq = 20
 )
 , times = 5L
)

print(lgb.bench)

# Make predictions on test set
lgb.model.test = predict(lgb.model, data.matrix(test_X[, colnames(test_X) != "CargoVolume"]),
             params = list(predict_disable_shape_check = TRUE))

test_mse = Metrics::mse(test_y, lgb.model.test )
test_mae = Metrics::mae(test_y, lgb.model.test )
test_rmse = Metrics::rmse(test_y, lgb.model.test )
test_bias = Metrics::bias(test_y, lgb.model.test )

cat("MSE: ",test_mse, "\nMAE: ", test_mae,"\nRMSE: ", test_rmse, "\nBias: ", test_bias)

#Stop measuring running time and print running time
end_time <- Sys.time()
end_time - start_time


# Check for the features that are the most important
lgb.feature.imp = lgb.importance(model = lgb.model)
print(lgb.feature.imp)

########################################################################

#Parameter tuning: grid Search to find the best hyperparameter combinations

max.depths = c(1,2,3)       #depth of the tree, default is 3
etas = c(0.3, 0.4, 0.6)     #learning rates of the tree, , default is 0.1
num.leaves = c(3,4,5)   #maximum number of leaves in one tree
min_data.leaf = c(1, 2, 3) #minimum number of data in one leaf

best_params = 0
best_score = 0

count = 1
for( depth in max.depths ){
 for( num in etas){
  for( leaf in num.leaves){
   for( min_data in min_data.leaf){

     #Get the time to train the lightGBM model
     bst_grid <- lgb.train(data = lgb.train,
                params = list(
                 objective = "regression",
                 metric = "rmse",
```

```
                        min_data_in_leaf = min_data, #minimum number of data in one leaf
                        min_sum_hessian_in_leaf = 100,
                        feature_fraction = 1,
                        bagging_fraction = 1,
                        bagging_freq = 0,
                        max_depth = depth,
                        learning_rate = num,  #learning rates of the tree
                        num_leaves = leaf, #maximum number of leaves in one tree
                        num_threads = 2), #depth of the tree
                      valids = list(test = lgb.test),
                      nrounds = 10000,
                      early_stopping_rounds = 1000,
                      eval_freq = 20)
      if(count == 1){
        best_params = bst_grid$params
        best_score = bst_grid$best_score
        count = count + 1
      }

      else if( bst_grid$best_score < best_score){
        best_params = bst_grid$params
        best_score = bst_grid$best_score
      }
    }
  }
 }
}
best_params
best_score

#Start measuring running time
start_time <- Sys.time()

lgb.bench.tuned = microbenchmark(
 lgb.model.tuned <- lgb.train(data = lgb.train,
                      params = list(
                        objective = "regression",
                        metric = "rmse",
                        min_data_in_leaf = best_params$min_data_in_leaf, #minimum number of data in one leaf
                        min_sum_hessian_in_leaf = 100,
                        feature_fraction = 1,
                        bagging_fraction = 1,
                        bagging_freq = 0,
                        max_depth = best_params$max_depth,#depth of the tree
                        learning_rate = best_params$learning_rate, #learning rates of the tree
                        num_leaves = best_params$num_leaves, #maximum number of leaves in one tree
                        num_threads = 2),
                      valids = list(test = lgb.test),
                      nrounds = 10000,
                      early_stopping_rounds = 1000,
                      eval_freq = 20)
 , times = 5L
)

print(lgb.bench.tuned)

# Make predictions on test set
lgb.test.pred.tuned = predict(lgb.model.tuned, data.matrix(test_X[, colnames(test_X) != "CargoVolume"]),
                 params = list(predict_disable_shape_check = TRUE))
```

```
test_mse = Metrics::mse(test_y, lgb.test.pred.tuned )
test_mae = Metrics::mae(test_y, lgb.test.pred.tuned )
test_rmse = Metrics::rmse(test_y, lgb.test.pred.tuned )
test_bias = Metrics::bias(test_y, lgb.test.pred.tuned )

cat("MSE: ",test_mse, "\nMAE: ", test_mae,"\nRMSE: ", test_rmse, "\nBias: ", test_bias)

#Stop measuring running time and print running time
end_time <- Sys.time()
end_time - start_time


# Check for the features that are the most important
lgb.feature.imp = lgb.importance(model = lgb.model.tuned)
print(lgb.feature.imp)


print(best_params)
```

Appendix V

Code for N-Beats in R-studio

```
# libraries we're going to use
library(readxl)
library(MASS)
library(dplyr)
library(caTools)
library(Metrics)
library(glmnet)
library(caret)
library(modelr)
library(purrr)
library(neuralnet)

#Clean Global Environment
rm(list = ls())

#Importing the dataset
df = read_excel("Database historical shipment volumes.xlsx")
#Drop columns that we do not need for the model
df = select(df, -Delivery, -VendorID, -Orderline, -CargoPieces, -POL, -PartNumber)
#Remove NAs
df <- na.omit(df)

#Translate categorical values into factors
df$VendorName = as.factor(df$VendorName)
df$MaterialDescription = as.factor(df$MaterialDescription)

# set a random seed
set.seed(1234)

attach(df)
train_test_split = sample.split(CargoVolume, SplitRatio = 0.8)

#The actual train and test set
train_X = subset(df, train_test_split == TRUE)
train_y = train_X$CargoVolume
train_X = data.matrix(train_X)
test_X = subset(df, train_test_split == FALSE)
test_y = test_X$CargoVolume
test_X = data.matrix(test_X)

#Start measuring running time
start_time <- Sys.time()


#Building the model with the required predictors
nn_model <- neuralnet(CargoVolume ~ DeliveryQuantity + VendorName, data = train_X, hidden =2,
err.fct = "sse",
                linear.output = TRUE)

nn_model$result.matrix
plot(nn_model)

#Gathering the predicted class labels and the actual class labels
pred_tt = predict(nn_model, test_X)
```

```
test_mse = Metrics::mse(test_y, pred_tt)
test_mae = Metrics::mae(test_y, pred_tt)
test_rmse = Metrics::rmse(test_y, pred_tt)
test_bias = Metrics::bias(test_y, pred_tt)

cat("MSE: ",test_mse, "\nMAE: ", test_mae,"\nRMSE: ", test_rmse, "\nBias: ", test_bias)

xgb.feature.imp = nn.importance(model = xgb.model.speed)
print(xgb.feature.imp)

#Stop measuring running time and print running time
end_time <- Sys.time()
end_time - start_time
```

Appendix VI

Results <u>Multiple Linear Regression</u>

	All features	PartCategory + Delivery quantity	PartCategory + SupplierName	Delivery Quantity + SupplierName
MSE	152.2526	154.1344	152.2419	154.1344
MAE	0.890467	10.0446	0.889811	10.0446
RMSE	12.33907	12.41509	12.33863	12.41509
Bias	0.06879498	0.04554565	0.06761485	0.04554565

Appendix VII

Results between different forecasting methods

Table 34: Comparison between predicted number of TEUs (using old and new method) and the actual number of TEUs.

Weeknumber	Actual	Old forecasting method			New forecasting method		
		Forecast	Error	^2	Forecast	Error	^2
42	*23*	*9*	*14*	*196*	*25*	*-2*	*4*
43	*8*	*9*	*-1*	*1*	*9*	*-1*	*1*
45	*8*	*9*	*-1*	*1*	*8*	*0*	*0*
46	*10*	*9*	*1*	*1*	*9*	*1*	*1*
47	*1*	*9*	*-8*	*64*	*1*	*0*	*0*
48	*12*	*9*	*3*	*9*	*10*	*2*	*4*
49	*7*	*9*	*-2*	*4*	*7*	*0*	*0*
50	*11*	*9*	*2*	*4*	*10*	*1*	*1*
51	*20*	*9*	*11*	*121*	*19*	*1*	*1*
52	*23*	*9*	*14*	*196*	*25*	*-2*	*4*
MSE				***59,7***			***1,6***