Demand forecasting at Nijhof-Wassink

Implementing a forecasting procedure for transportation provider Nijhof-Wassink using company and macro economic data sources

Lex de Jager

Supervisors University of Twente
Dr. D. Prak
Dr. M.R. Machado

Supervisor Nijhof-Wassink
M. Hansté

Bachelor of Industrial Engineering and Management
University of Twente
Netherlands
3rd December 2024

Executive Summary

Nijhof-Wassink is an international transportation service provider, mainly working in the Netherlands. In the chemical logistics divisions Dry Bulk Logistics (DBL), and Liquid Bulk Logistics (LBL). The number of orders greatly influences the cost pricing of individual orders. As clients do not explicitly pay for the kilometers driven towards the pick-up points, these empty kilometers cause volatility in costs. Therefor, being able to accurately match demand with capacity reduces the driven empty kilometers which in turn lowers the cost per order. Nijhof-Wassink wants to accurately set their sales prices, which is the motivation for predicting future demand. The research question is defined as: *How can Nijhof-Wassink's chemical division utilize historical data and external factors to better understand and forecast their monthly demand?*

To answer the research question the data of Nijhof-Wassink is analyzed, external market factors are selected, and forecasting methods are made and validated. The data of Nijhof-Wassink is taken from march 2016 until April 2024. The data of LBL and DBL are aggregated to make the forecast more accurate. Linking external factors to the demand data of Nijhof-Wassink is needed before including external factors as input in the model. The Dutch Bureau of Statistics (CBS) data is used as external datasets. Two ways of linking the datasets are used; a general trend analysis called Coupling and a Correlation matrix. The external feature from the coupling method is general Turnover, while from the correlation method Turnover abroad, and LPG prices scored higher. The number of Trucks is also added as an external feature. Datetime features are added to round out the model inputs. A comparative analysis was done between four models with three approaches of taken external variables into account. Artificial Neural Network (ANN), Random Forrest (RF), Support Vector Regression (SVR), and LightGBM are used. The models considered are supervised machine learning models, which have increased considerably in popularity in forecasting literature. The three approaches are not taking external variables into account, lagging external variables, and using forecasted external variables as input. For the forecasting method for the external variables SARIMA is used, a simple linear model. The simple model is used as input for the more complex supervised machine learning model. Flat cross validation is the best way to create a model which can forecast, but introduces some bias. Flat cross validation means first estimating the hyperparameters, and after that validating the accuracy. To validate the in-sample accuracy of the models, K-Fold cross validation is used, after which gaps are defined to look how the forecast performs over longer time intervals.

The best performing model is the base SVR, closely followed by the forecasted external features RF. The feature importance analysis of the RF model showed that the external feature Turnover is used much less than the others. Thus the conclusion can be drawn that for this case the correlation method produces more useful results. The SARIMA gives errors with certain gaps and part of the fold, which makes it harder to implement. because the SVR performs best In-sample without external features, the SARIMA is substituted with the SVR model to create the best of both. The datetime feature Year creates a bad increasing trend, because for the majority of the data the number of orders is growing. The number of orders is not growing in the last 2 years, which makes it a bad indicator. With new data from may until September 2024 these models can be further validated with an out-of-sample test. This out-of-sample test is done by using the data from the in-sample test and forecasting the period of the out-of-sample test, after which the MAPE is calculated. The SVR model performs significantly worse compared to the RF model with SARIMA or SVR. The best performing model on the out-of-sample test is the RF with SVR to forecast the external features. The data is disaggregated into LBL and DBL datasets and forecasted with this model. Both the LBL and DBL MAPE is worse than the aggregated data, where the DBL MAPE is better than the LBL MAPE. Aggregated forecasts are always more accurate than disaggregated forecast, so this acceptable. Although the hyperparameters are retrained, the features are more attuned to the larger subset which is DBL.

The answer of the research question is two fold. Firstly, Nijhof-Wassink can better understand their demand by finding external factors that influence their demand with correlation. Secondly, Nijhof-Wassink can forecast their demand by using the model as described: A Random Forrest with external features forecasted with Support Vector Regression. For linking external data to the demand data of Nijhof-Wassink the method of correlation proved more useful than coupling. Nijhof-Wassink is recommended to keep the model up to date, by checking these correlation coefficients and their feature importance analysis. This model only partly solves the problem of volatility in cost. Nijhof-Wassink can improve the applicability of the by making the forecast more operational. Further academic research can be done into the methodologies of hybrid forecasting using flat cross validation.

Contents

Chapter 1

Introduction

In this chapter the company structure and the department instigating this research is introduced. After the introduction of Nijhof-Wassink, the motivation for this research is described. The problem identification looks more closely at the problem and the research design describes the way that problem will be solved.

1.1 Introduction to Nijhof-Wassink

The research laid out in this paper is conducted at Nijhof-Wassink, part of the Nijhof-Wassink Group. The Nijhof-Wassink Group is an international logistics service provider with a focus on dry and liquid bulk goods in a competitive international market. Other activities the group is also active in are warehousing, truck maintenance (NIJWA) and cleaning services. They operate inter-modally by road, water and rail from multiple different countries. The full organization of Nijhof-Wassink Group is detailed in Figure 1.1. As stated, our focus in this thesis is on the Nijhof-Wassink company, specifically the Chemical Logistics division.

The research assignment is formulated by the head of sales support. This department is a support function within Nijhof-Wassink and is responsible for calculating cost prices and assisting with tender bidding. Tenders are large scale contracts Nijhof-Wassink negotiates with clients involving multiple lanes at the same time. A lane is transportation jargon for a route from point A to point B. A tender offer involves the client giving off the specific lanes they need driven for a specified amount of time. Nijhof-Wassink then bids on a selection of the offered lanes. If the bid wins, Nijhof-Wassink will drive the included lanes for the duration of the contract. Typically, a tender has a duration of multiple years. In the tender there are indications of the demand intensity at different points in time, for example a certain number of orders per month. However, these indications are not set amounts the client will have to abide by. The way in which these tender contracts are usually structured results in little insight for Nijhof-Wassink when orders will be actually placed. This creates the possibility for increased volatility in incoming orders within a certain time period.

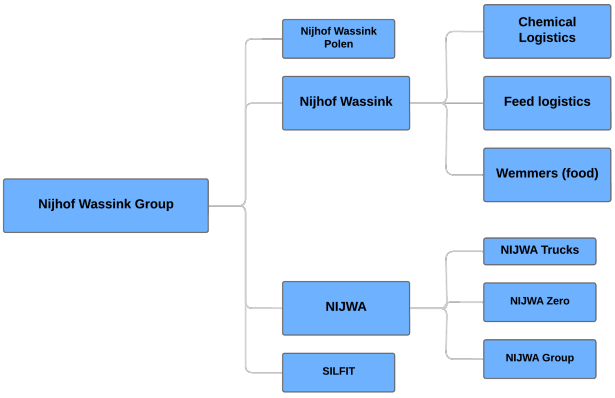

Figure 1.1: Company structure

Within the Chemical Logistics division of Nijhof-Wassink there are four business units: dry bulk transport, liquid bulk transport, fuel transport and warehousing. Each have their own individual clients and operate independently from each other. However, in practice the markets of Dry and Liquid bulk are very closely related with regards to demand fluctuations. The Liquid bulk industry is, more specialized and has fewer competitors compared to dry bulk. The reason dry and liquid are separate business units is because the trucks cannot be used interchangeably. As the business units of fuel transport (gas station inventory) and warehousing (no transport at all) are very different from dry and liquid bulk, they will be excluded from the scope of this research.

In order to determine the tactical decision to be made within the company, each month a tactical meeting is held between different departments at Nijhof-Wassink. There are two main goals to this meeting. Firstly, the meeting looks back and reflects on the actual costs incurred during the previous periods based on new calculations in order to identify which lanes have been profitable and which have not. Secondly, the meeting looks forward, estimating what trends will influence the market demand and how capacity, maintenance and pricing should be adjusted for the next period. Tactical decisions that need to be made for example are decisions on personnel capacity, truck maintenance, or sales pricing for certain lanes. A decision to incur more costs by deploying more capacity during a given period has implications for planning as this capacity needs to be used effectively in order to be profitable. These decisions have direct influences capacity, influencing the operating costs which in turn influences the sales pricing and with it the profit margins for the next period. Currently these decisions are made based on internal historical data and rough estimations of market trends based on experience.

1.1.1 Motivation for the research

As stated, Nijhof-Wassink operates in a competitive market where competitive pricing is key. In order to be more competitive in this market it is of great importance to be able to reliably calculate and predict operating costs as this is at the base of deciding sales prices which decide if tenders can be won or not. A significant part of the operating costs at Nijhof-Wassink's Chemical Logistics division are controlled by the ability to plan capacity effectively. For example, deciding the amount of capacity in such a way that tender requirements can be fulfilled while the capacity is also used efficiently, i.e. chartering out empty trucks or chartering tendered lanes during moments of low capacity, will result in an improved competitive edge within the market. The other main thing influencing the ability to plan capacity effectively comes back to the tenders and the demand. As stated the way tenders have been setup creates the possibility for volatility in the demand. Coupled with an already volatile market, makes it difficult to accurately estimate demand. Currently the demand is estimated by internal historical data and estimation of market trends based on experience. These combined tend to not give a reliable estimation of the demand. The sales support manager wants to improve on the estimations of demand based on internal historical data and market trends based on experience and move to more data driven decision making.

1.2 Problem Identification

In this section the problem is defined, from action problem to core problem via a problem cluster. Once the core problem is described the main research question is defined. The processes of Dry and Liquid bulk business units are assumed to be similar in practice, so much so that for this problem identification they will be treated as one.

1.2.1 Action Problem

As stated in the research motivation, Nijhof-Wassinks Chemical Logistics division has to operate with competitive market with volatile demand. It is difficult to accurately predict the costs per order which in turn influence the ability of the company to effectively compete within the market. The cost per order vary a lot based on how many orders are driven. The main variable influence on cost are the kilometers driven without orders. with more orders the number of kilometers driven empty can be minimized more effectively, thus overall demand is very influential for the costs per order. From this the following action problem has been defined: Uncertainty in correctly setting sales prices.

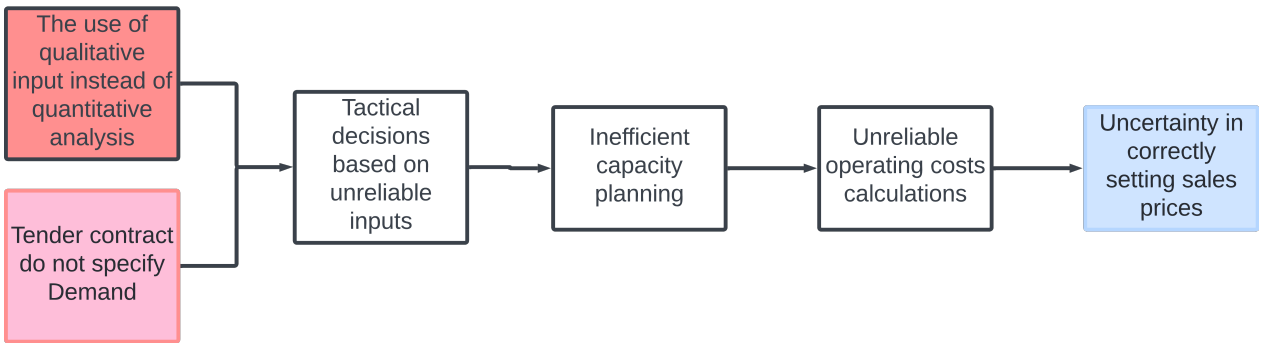

Figure 1.2: Problem Cluster

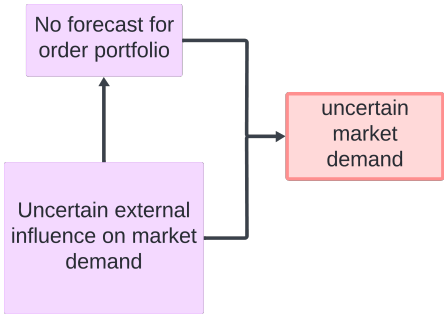

Figure 1.3: influences on the core problem

1.2.2 Problem cluster

In order to go from action problem to the actual core problems, a problem cluster was created. From the action problem the problem cluster is expended by working backwards towards the core problem. From the uncertainty in correctly setting sales prices it is found that unreliable operating cost calculations are the cause. These uncertainties are mainly caused by inefficient capacity planning of both trucks and personnel when the planned capacity is not on level with the actual demand. These inefficiencies are created by the tactical decisions which are made monthly during the tactical meetings However, these decisions are based on unreliable demand estimations. The demand estimations are unreliable as of two reasons. Firstly, they originate from qualitative experience and not from quantitative data analysis. Secondly, because of the way the tender contracts are structured, the clients have no obligation to forecast their actual order amount for a specific time period. The full problem cluster is shown in Figure 1.2.

1.2.3 Core problem

From the above shown problem cluster two core problems were identified. Firstly the tender contracts are not giving insights into actual demand in a certain period as this is not a requirement. Secondly the market demand estimations are unreliable because they use qualitative experience and not quantitative data analysis. From these two identified core problems the second problem is chosen for this research as the first problem requires changes in how tender agreements are set up, which falls outside the scope of this research. It is impossible to fully solve unreliability in demand estimation but it is possible to create an expected market demand. An expected market demand creates a quantitative basis for decision making. In order to create expected market demand two factors of the problem need to be looked at. Firstly, the relation between external factors and the orders is not defined. Secondly, the information about products is not used. In the decision process at Nijhof-Wassink external factors can be quite important, since a disruption in supply chain can immediately affect the order level. The qualitative consensus in Nijhof-Wassink is that the external factors have a strong influence on market demand. On the other side there is no forecast about expected demand whatsoever, which means the decision makers have no way to look forward and compare between expected demand and capacity. The uncertain influence of external factors also influences a potential forecast, which is visualized in Figure 1.3.

1.2.4 Research question

From the core problem we can formulate a research question to reduce the uncertainty and unreliability in demand predictions for Nijhof-Wassink's chemical division. We will try to predict the future market demand, and thus solve the chosen core problem. To that end we have formulated the following main research question:

"How can Nijhof-Wassink's chemical division utilize historical data and external factors to better understand and forecast their monthly demand?"

This questions results in the goal of creating a forecast with as input both the data of Nijhof-Wassink chemical division and certain external macro economic factors. In order for the external macro economic factors to be used, two criteria need to be met. Firstly, their usability has to be supported by analysis from within the company or literature. Secondly, there must be a trustworthy data-set available. The forecast is going to be used for tactical decision making, which means the forecast will be aggregated to a monthly forecast looking 12 months ahead. The forecast will be aggregated to one number and dis-aggregated to liquid bulk and dry bulk. If time allows other levels of aggregating will be looked into.

1.3 Research Design

To answer the research question six main questions are defined. The main questions answer different gaps in knowledge in order to answer the research questions. The main questions are explained more in the following sections, together with the necessary sub questions to answer the main questions. The sub questions are stated under the explanation of the ,ain questions referenced with the number of the question they belong to. Phases 3 to 7 of the MPSM are used as framework of this research design. (Heerkens and van Winden, 2017)

1.3.1 Problem analysis

In the problem analysis the current situation is discussed. This means investigating Nijhof-Wassink's current processes and deciding if knowledge can be gained about the markets from this data. To do this the following questions will be answered:

Main question 1: *How is forecasting used and managed in the current process?*

To that end, three topics will be analyzed. First, we analyze the current ways of working at the tactical meetings. Secondly, we analyze the data from Nijhof-Wassink. Thirdly, we create a qualitative analysis of clients via internal experts. Based on these three topics, three sub questions have been drawn up. In 1a the already existing forecasts are looked into. If there is no current forecast, the management steering tool that is used instead is investigated. In 1b the data on demand behavior of the current client base is looked at. This will give insight in which markets Nijhof-Wassink is close to via their clients. In 1c multiple interviews will be held with sales- and sales support -departments to identify which market indicators could hold predictive value. Here we can make the distinction between the dry and liquid bulk, and discuss other aggregation levels with the problem owner. Since what can be forecasted depends on what data is available.

Sub question 1a: *What is the current process of tactical decision making?*

Sub question 1b: *What demand data does Nijhof-Wassink have?*

Sub question 1c: *What type of macro economic indicators are thought to have an influence by Nijhof-Wassink?*

1.3.2 Solution generation

The Solution generation consists of two parts, the forecasting models and the external factors. For the forecasting models we only consider models that have multivariate capacity and are used in the context of transportation in theory.

Main question 2: *What are the relevant models for demand forecasting in Transportation?*

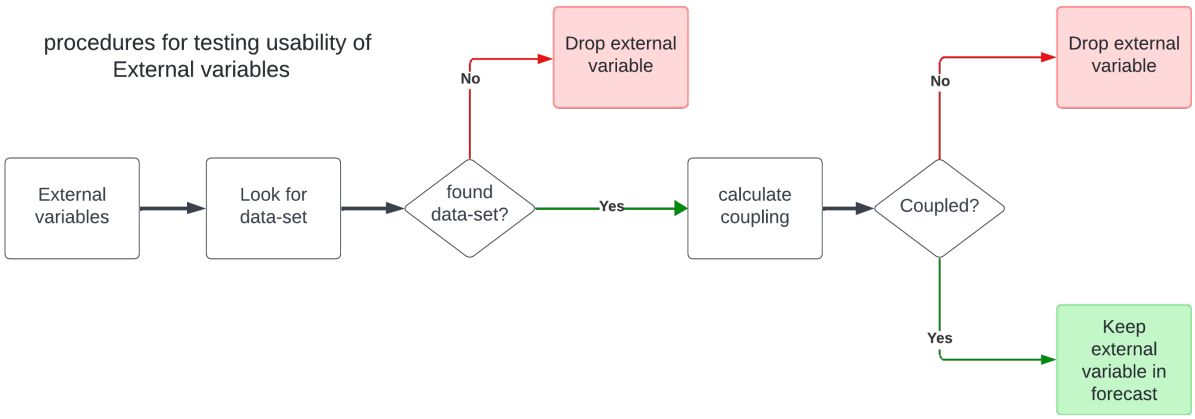

Figure 1.4: Procedure external variables

To give an answer to the main research question we will choose, train, and test a forecasting method. To make that decision we first need to expand the knowledge about forecasting models to make a sound choice, this is done with question 2a. Since data is available at Nijhof-Wassink the preferred methods are quantitative models (Profillidis and Botzoris, 2019b). A prerequisite for the models is that they need to be multivariate. In this section we will consider support vector regression, Artificial neural network, and Random Forrest as potential forecasting methods. Linear regression, fuzzy linear regression and artificial neural networks are methods described in the context of transportation by Profillidis and Botzoris (2019d);(2019e). Support vector regression and random Forrest are more novel forecasting methods and not mentioned by Profillidis and Botzoris (2019d), but in case studies these two models produce good results (Pai et al., 2010; Zhang, 2003). K-fold validation is a widely used technique to train the models, the exact workings will be described in 2b. To check the accuracy of the forecasting models different accuracy measures should be considered. To test the methods Profillidis and Botzoris (2019c) suggest using Theil's inequality coefficient or Mean absolute percentage error (MAPE), these will be considered in 2c. This theoretical basis is useful for choosing which model to recommend in 6a, since a more difficult model with the same accuracy as a simple model should not be recommended.

Sub question 2a: *Which forecasting methods are used in transportation sector?*

Sub question 2b: *How to set-up and, if applicable, train forecasting methods?*

Sub question 2c: *What evaluating methods are used for forecasting methods?*

The next sub-question covers the external factors section of the research question. We will consider two inputs for external variables, those found in literature described here and those the companies qualitative analysis gave as output in 1c. The qualitative analysis will thus be substantiated with a literature review.

Main question 3: *How can macro economic indicators that influence the transport sector in the Netherlands be used in a forecast of Nijhof-Wassink?*

To make a forecasting model with external variables data is necessary of the external variables involved. These data-sets need to be gathered from trustworthy sources, for example government sources. 3b covers this part. When no data-set can be found for a variable it will be discarded, since it cannot be used in the predictions. To check if it is useful to incorporate the external variable with the data of Nijhof-Wassink we should check if the data sets are coupled or decoupled. When the data set is not coupled the external variable should also be discarded, the procedure for external variables is visualized in Figure 1.4. 3c covers the calculations of coupling by Profillidis and Botzoris (2019a) and correlation matrix.

Sub question 3a: *Which Macro Economic indicators influence the transportation sector?*

Sub question 3b: *Do the identified macro economic indicators have trustworthy and complete data?*

Sub question 3c: *Is the external data linked to the data of Nijhof-Wassink?*

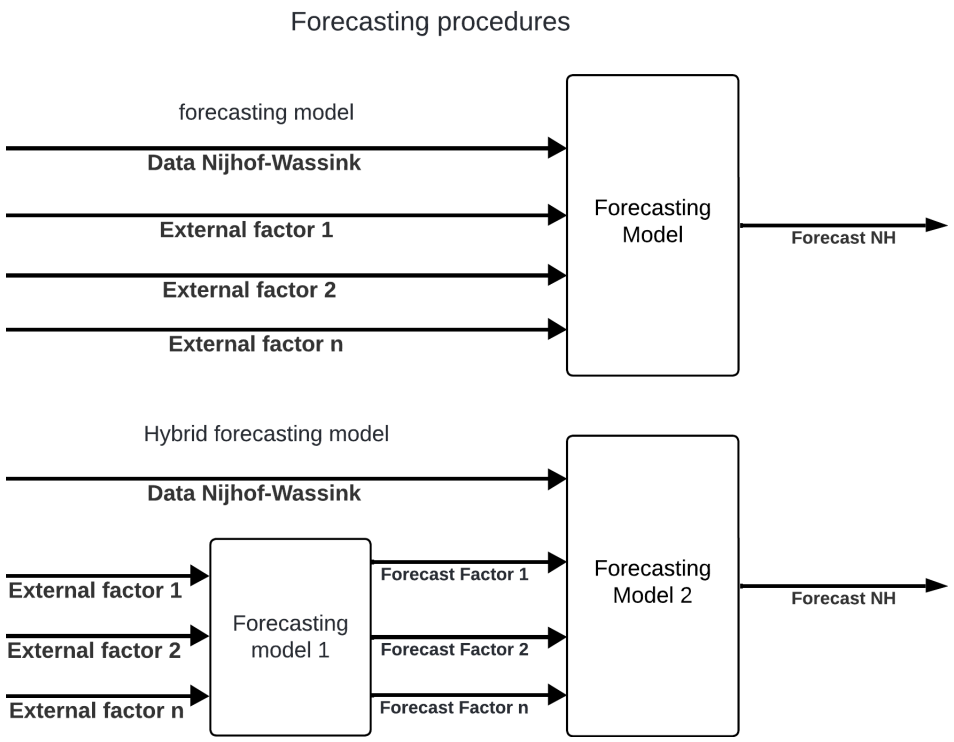

Figure 1.5: Procedure hybrid models

1.3.3 Solution choice

In this phase a solution is chosen based on the applicability to the problem and feasibility. The following question is formulated to cover the solution choice:

Main question 4: *Which models can forecast Nijhof-Wassink's demand, while taking external factors into account?*

To answer this question the information gathered in previous questions and actions is combined to chose the models which can solve the problem. We will do this by choosing different forecasting procedures: a regular forecasting model and hybrid forecasting models. In Figure 1.5 the difference is visualized. The difference is to forecast with company data and external data sets at the same time, or first forecasting the external factors and using those forecast in the model to forecast the demand of Nijhof Wassink. The external variables that come up during questions 1c and 3a need to have data and be coupled to the demand of Nijhof-Wassink. In Figure 1.4 the procedure is described when to include external variables. The different models chosen here will all be implemented to be ranked on accuracy. Before making the model it is good to choose which validation criteria will be used to assess the accuracy and who the models will be tested. The choice which model to implement and which macro variable to include, is based on availability of data and the feasibility to program, train, and evaluate the model.

Sub question 4a: *How can lagged or hybrid methods be implemented?*

Sub question 4b: *Which macro economic variables can be implemented in the model?*

Sub question 4c: *Which models are feasible an expected to give acceptable results?*

1.3.4 Solution implementation

In this phase of the MPSM the solution is implemented, which answers the question:

Main question 5: *Which forecasting model performs best for Nijhof-Wassink?*

The different models chosen will be made in python. The models chosen need to be made in Python due to the programming knowledge at Nijhof-Wassink. Different python packages can be used to ease this process, the process of creating will be described in 5a. In 5b the models need to be trained and evaluated as described in the outcomes to questions 2b and 2c. The accuracy of the different models will be described as answer on question 5c. The different models will be ranked on accuracy measures described in 2c.

Sub question 5a: *How can the chosen models be created?*

Sub question 5b: *How can the chosen models be trained and evaluated?*

Sub question 5c: *How do the chosen models perform?*

1.3.5 Solution evaluation

The implemented solution needs to be evaluated to check if it solves the problem at the company. The solution will achieve the company goal if it deliverers a Interactive forecasting tool and an advise on how to adapt the forecasting tool.

Main question 6: *How can the chosen solution be improved to enhance performance, generality, and validity?*

To solve the problem in the company the forecasting tool is the most tangible deliverable. The goal of the tool is to predict demand on a tactical level, which means a monthly forecast for 12 months ahead. An easier to understand model with the same accuracy as a more difficult model is preferable, so in 6a the different models are discussed and the best performing model for Nijhof-Wassink is chosen. The tool should also be easy to use, since it will be a reoccurring job to forecast the demand. This will be discussed in 6b. The second goal of the tool is to be easily adaptable to the different divisions in Nijhof-Wassink. This leads to questions 6c, where we will give an advise on how to adapt the forecasting tool. This advice is likely to be in the form of a manual. This manual also needs to explain how to expand the code and change the external factors for the different departments.

Sub question 6a: *Which model has the best output for Nijhof-Wassink?*

Sub question 6b: *How can the validity of the model be inproved?*

Sub question 6c: *How to adapt the forecast tool for other divisions?*

1.4 Scope of the research

It is good to reflect on the scope of the research. Due to time constraints the problem is scoped to only focus on the chemical division of Nijhof-Wassink. This means the problem is only partly solved if the main research question is answered. Another constraint is in the considered forecasting Models, the models are chosen because different sources use the models in a transportation setting. This means not all possible models are considered, which is a scoping choice. Due to time constraints it was not possible to analyze all methods before choosing. Next to that there is some evidence that the ensemble approach to forecasting improve the robustness and accuracy (Wu and Levinson, 2021). In this research this is translated to the hybrid method. Other ensemble approaches will be left out of this research. There are many different ways to forecast and due to time constraints not all novel ensemble forecasting methods can be tested.

1.5 Conclusion chapter 1

In this chapter Nijhof-Wassink is introduced and the motivation of the research is described. After which the problem identification and research design are stated. Nijhof-Wassink chemical division is an international logistics provider working with dry and liquid bulk. In the following chapters the research is conducted in the following manner: In chapter 2 the processes at Nijhof-Wassink will be further discussed, together with the data analysis of the chemical division. In chapter 3 the relevant literature for this thesis will be discussed. First the different types of forecasting methods. Secondly, the set-up necessary to create forecasts and accuracy tests. Thirdly, the evaluating methods of these tests are defined. and fourthly the datasets and connection of external variables are discussed. In chapter 4 the forecasting models are created and tested.

This is done by implementing the models and creating an in-sample test to compare the models. In chapter 5 the best performing models are further enhanced and validated by tested on new data. with the new data an out-of-sample test can be conducted. In chapter 6 the research question is answered and recommendation for further research and to Nijhof-Wassink are given.

Chapter 2

Current Situation

In this chapter the current situation at Nijhof-Wassink's Chemical logistics division is investigated in order to answer main question 1: *How is forecasting used and managed in the current process?*. This is done by following the three defined sub questions which will be reflected on in the conclusions of this chapter.

This chapter has two main parts. First, the process from order intake till tactical decision making is described in order to answer sub question 1a: *What is the current process of tactical decision making?*. Second, the available data within this process is analyzed to answer sub question 1b: *What demand data does Nijhof-Wassink have?*. In this part the sub question 1c: *What type of macro economic indicators are thought to have an influence by Nijhof-Wassink?* is also answered.

2.1 Current Process

In this section the current state of the process at Nijhof-Wassink regarding order intake, planning and tactical decision making is described in order to answer the question *How is forecasting used and managed in the current process?*. First, some more context is given on the current clients and products offered by the Nijhof-Wassink's Chemical logistics division. After this the three mentioned parts of the process are described. Lastly the currently recognized external factors are discussed.

2.1.1 Clients and Products

Nijhof-Wassink has a wide range of clients, many of which are situated within the Netherlands. Both Liquid and Dry bulk logistics have a small number of clients making up a large share of the orders. These clients are mainly situated in the chemical, pharmaceutical and petroleum industries. The total clients base produces products for a wide range of markets, for example: the plastic, petrol, construction, and manufacturing sector. The products that are transported by Nijhof-Wassink for these clients are most often semi-finished products, used to create other products. Examples of these products are plastic pellets, resins or polyols, but also the base products for medication and fuels.

2.1.2 Order Intake Process

The order intake process at Nijhof-Wassink has two main entry points, email and an online portal. Via both these channels requests come in in accordance with or separate of the tender agreements, so called SPOT orders. Orders connected to tender agreements do not need price setting but of course orders separate from these agreements need a swift cost calculation in order to set a good price point. In general, SPOT orders have a higher profit margin compared to tender orders. The requests can come in at any time of day and with varying pick up dates. For example, a client can sends a request at the end of the day for loading at the start of the next day. This is not always possible and sometimes Nijhof-Wassink sells orders to charters because they are cheaper or more flexible. Selling orders to charters reduces the usage of the trucks. On the contrary, it can also be beneficial to take on charters in the form of SPOT orders if this helps the capacity planning. Both of these options will be looked at more in the next section on planning.

2.1.3 Planning

The orders that have come in via either entry point need to be fit in the ever changing capacity planning. As said before, decisions need to be made on if the order should be executed by a company owned truck or if the order should be outsourced in the form of a charter. Sometimes using a charter is better, for example when the loading takes place far from frequently used centers. A local transportation provider can often be much cheaper than Nijhof-Wassink in certain locations. On the other end is the pursuit of the capacity planning department to minimize what is known as empty kilometers. Empty kilometers are the kilometers the trucks have to drive between orders, in which the truck doesn't carry any cargo. The clients of Nijhof-Wassink are not willing to pay extra for those empty kilometers, so it is in Nijhof-Wassink's interest to minimize these kilometers. This can be done by, as described above, chartering out the order or instead taking on SPOT orders in the area to reduce the distance driven with an empty truck. The options available to the capacity planning are also influenced by tactical decisions made by management. The specifics and impact of these decisions is detailed in the next section.

2.1.4 Tactical Decision Making

In Chapter 1 the monthly tactical meetings are briefly discussed. In these meetings, two things happen. Firstly, there is a section on looking back to certain lanes and checking if the expected margin has been made. Due to toll, empty kilometers and cleaning cost the profit margin can be fluctuate a lot to when the prices where set. Secondly, the meeting looks ahead to make tactical decisions for the upcoming period. These tactical decisions include changes to truck maintenance scheduling, personnel planning and price points. All these decisions have influence on the available capacity and the eventual demand. A delicate balance needs to be achieved in order to maximize profit margins. Currently, the decisions are mainly made based on historical order data from within the company and estimations in market changes based on experience. This means that the meetings are mainly held on the basis of historical figures and qualitative predictive information. rom a managerial point of few it is hard to steer these qualitative feelings in a productive way. We must come to the conclusion no forecast is used in the current situation. Not only no quantitative forecast, but also no methodology of qualitative forecasting is used. In the next section more will be explained about the qualitative measures in use at Nijhof-wassink.

2.1.5 External Factors

In talks with the sales team at Nijhof-wassink, the experience of the overall up and down of the market was found as a big influence on the demand level. The sales team gains these qualitative insights during talks with clients, and during the conventions in the logistic or chemical sector. The majority of their clients are producing chemical products which are used in a wide range of industries. The overall market fluctuation in the chemical sector could be used as external factors to predict demand at the Nijhof-Wassink level. Another set of indicators that came up during those conventions are the trusts of the clients in the markets they serve, if the trust is high more inventory is kept. Examples of external factors used in the chemical industry or trust indicators are producer price indexes, market trust figures or inventory levels. The idea behind these indicators is the following; if the chemical industry is doing well, more products needs to be transported, thus resulting in more orders for Nijhof-Wassink.

2.2 Data of Nijhof-Wassink

In this part of the chapter the previously mentioned historical order data of Nijhof-Wassink is analyzed. This data is gathered from the ERP system Navitrans, which has the functionality to export to Excel, in two separate data sets. This is because liquid and bulk figures are recorded separately. In this part the cleaning of the respective data sets, analysis and aggregation analysis are described.

2.2.1 Data cleaning

Before the data could be analyzed, it first needed to be cleaned. The gathered data required little cleaning as only two problems arouse during the initial screening. Firstly, a certain amount of location data was missing as a result of routes originating outside the Netherlands. This was handled by dropping the rows which had empty cells as the amount of times this occurred did not influence the size of the data sets significantly.

Secondly, the two datasets did not have the same start and end date, which was necessary in order to aggregate both together. This problem first showed in the first two months of the bulk dataset where an unusual low amount of orders were noted. Both datasets also showed orders well into the future at the time. In order to fix this problem the limits of the dataset were set to March 2016 to April 2024. In the next two sections the analysis on the two separate data sets is given. After this the data is aggregated and analyzed again. In the monthly data the number of order is called # dossiers.

2.2.2 LBL analysis

The average number of orders per month in the liquid Bulk (LBL) division is around one third of the total numbers of orders. Looking at the graph in figure 2.1 It is visible that there is no big trend to be seen in the data. Around 3% of the orders are SPOT and 24% are inter modal, the rest are orders via tenders. In this context, inter modal means international train transport. Inter modal is only used for liquid bulk because the service area of the liquid bulk division is bigger then dry bulk. With a larger service area the inter modal option makes more sense for some lanes.

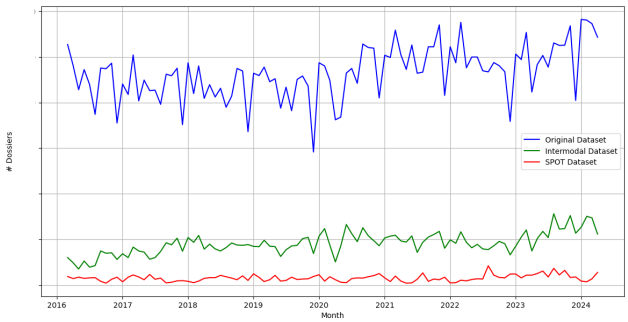

Figure 2.1: Number of orders liquid Bulk

2.2.3 DBL analysis

In the Dry Bulk division (DBL) the data looks a slightly different, as can be seen in figure 2.2 There is more variance in the data, as can be seen in the standard deviation at 236. The average number of orders per month is also more than double as in the liquid bulk. The absolute and relative number of SPOT orders in dry bulk is bigger than in liquid. The explanation for this is the fact that liquid bulk is a more specialized industry, which leads to relatively more client loyalty and less possibilities of outsourcing.

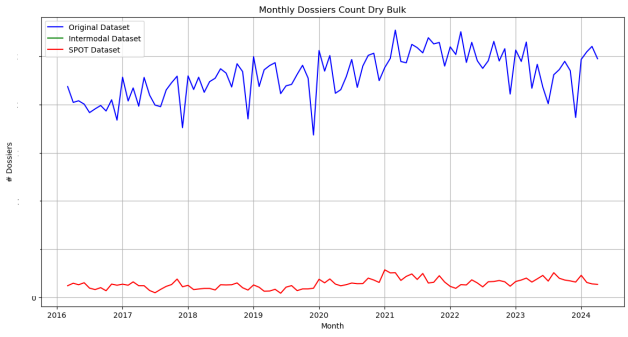

Figure 2.2: Number of orders Dry Bulk

2.2.4 Trend and seasonality

To create an accurate forecast the data of Liquid and dry bulk are aggregated to one dataset. The data from both liquid and dry bulk are found to be alike enough to be combined. The Dry bulk is around two thirds of the total orders, which induces some risk with dis-aggregating. The businesses have a lot of similarities,

which is why aggregating is deemed to be a valid option. This combined dataset is during the rest of this research.

The aggregated data can be decomposed to show the trend and seasonality using the STL method discussed by Hyndman and Athanasopoulos (2021). Decomposing into Trend, seasonal component and random factor generates picture as seen in figure 2.3. There is not one clear trend in the data, but multiple trends. A relative stable growth from 2017 until 2020, followed by an increase in growth from 2020 until 2022. This increase and decrease in growth is the same time frame as Corona. From 2022 onward a decrease in orders is noted. This could be the result of the post corona recession in Europe. The seasonal graph shows a clear dip in December, which can be explained by the company closing down between Christmas and new year.

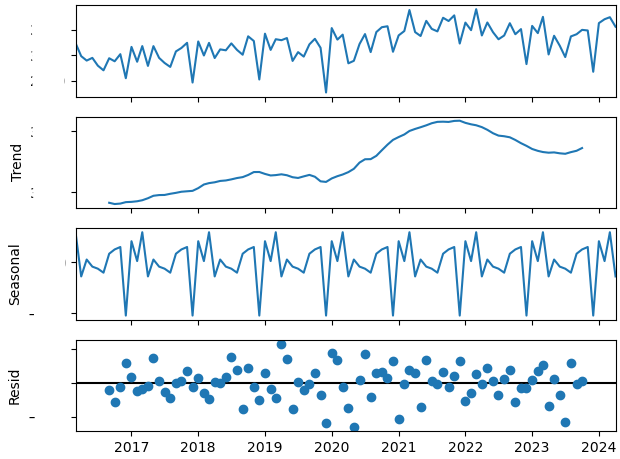

Figure 2.3: Number of orders Decomposed into trend, seasonal and random

The capacity of Nijhof-Wassink to transport orders is directly proportionate to the maximum number of orders that are possible in a month. The growth that can be observed in the number of orders is influenced by market forces and by the capacity of Nijhof-Wassink. The number of available trucks will be taken into account by using the trucks as an extra exogenous input in the forecasting procedure.

2.3 Conclusion chapter 2

At the end of this chapter the three sub questions about the current state at Nijhof-Wassink can be answered. For sub question 1a: *What is the current process of tactical decision making?* the following answer has been found. The process of tactical decision making is based on both quantitative internal historical data and qualitative external market experiences. It was noted that this process does not use quantitative data for both internal and external data which results in sub optimal estimations of demand. This sub optimal demand estimations lead to unreliable capacity planning which in turn puts pressure on profit margins.

The sub questions 1b: *What demand data does Nijhof-Wassink have?* is answered by two separate the data sets and the one aggregated datasets analyzed in the previous sections. It was found that the data was easily cleaned and gave insight in the demand per month from the years 2016 till 2024. The data shows that the demand has been increasing over past years and shows slight seasonality with the biggest drop in December, explained by company closure during that time. Multiple trends can be seen in the data, a steady increase from 2016 until 2020. From 2020 until 2022 the increase in orders becomes much bigger, flatting out and starting to decrease in December 2022.

The answer to sub question 1c: *What type of macro economic indicators are thought to have an influence by Nijhof-Wassink?* has been found to be factors used in the chemical industry. These factors are defined as producer price indexes, market trust figures and inventory levels at clients.

To answer the main question 1: *How is forecasting used and managed in the current process?* In this chapter the tactical decision making process is described, the data of Nijhof-Wassink is analyzed, and a

qualitative assessment of the macroeconomic indicators has been done. The tactical decision making is done in monthly meetings where costs are recalculated and pricing policy is determent. In these meetings only qualitative information and quantitative data of Nijhof-Wassink itself is used. No quantitative forecast is used in this monthly meeting. The Data of Nijhof-Wassink is very clean and easy to analyze. The data is aggregated and has multiple trends and relatively strong seasonality in December. External factors found by the qualitative assessment of the sales team mostly come up during meetings with clients or during conventions of the logistic sectors.

Chapter 3

Literature review

This chapter will explain the relevant literature concerning forecasting and external variables influencing demand in transportation. The chapter will answer the following two research questions: *What are the relevant models for demand forecasting in Transportation?* and *How can macro economic indicators that influence the transport sector in the Netherlands be used in a forecast of Nijhof-Wassink?*

Before diving more deeply into the different forecasting methods it is important to mention some points that are universally true about forecasting. Chopra and Meindl (2016) mention four universal characteristics of forecasts:

1. Forecasts are always inaccurate
2. Long-term forecasts are less accurate then short-term forecasts
3. Aggregate forecasts are more accurate than disaggregate forecasts
4. The farther up the supply chain a company is the greater the distortion of information it receives

When creating forecasts it is important to remember these four characteristics and ensure the minimization of their impact. The accuracy of a forecast has to be measured as it is an important metric for decision based on the data of the forecast. The different models and accuracy methods will be discussed in this chapter to answer the first main question. In the case of Nijhof-Wassink the most important part is that they are affected by global trends in industry. This awareness falls in line with what Chopra and Meindl (2016) mention as necessary for a good forecast. Namely, past demand, lead time of product replenishment, planned advertising or marketing efforts, planned price discounts, State of the economy and actions that competitors have taken. For Nijhof-Wassink, as transportation provider, the lead time, advertising, and price discounts, can be disregarded entirely. So for Nijhof-Wassink the past demand and state of the economy are most important. In the second part of this chapter more possible external indicators are sought in order to answer the second main question for this chapter.

3.1 Forecasting methods

In this part we will answer question *What are the relevant models for demand forecasting in Transportation?* In section 3.1.1 and 3.1.2 different forecasting methods are discussed. First off is the statistical methods underlying SARIMA based on Profillidis and Botzoris (2019c) as this research deals with time series data. Secondly four learning methods are chosen for this research: Artificial Neural Network (ANN), Random Forest (RF), LightGBM and Support Vector Regression (SVR).

Forecasting is a field of study that tries to predict the future as accurately as possible. There are two main types of quantitative forecasting: explanatory forecasts and time series models. Explanatory forecasts are based on predictor variables, external factors that are related to the predicted variable. Like in an example of Hyndman and Athanasopoulos (2021) in Equation 3.1, were the demand of electricity (ED) is predicted using external variables.

$$ED = f(current\ temperature,\ strength\ of\ economy,\ population,\ time\ of\ day,\ day\ of\ week,\ error) \tag{3.1}$$

Based on the data that is gathered in the previous chapter, the choice is made to focus on the time series models version of forecasting for this research. In the next section a more in depth explanation of this type of forecasting is given.

3.1.1 Time series forecasting methods

A time series refers to a number of observations with time as an implicit variable. In the example of Hyndman and Athanasopoulos (2021), the electricity demand (ED) is forecast by past electricity demand. This can be seen in equation 3.2 In the case of (S)ARIMA, the time series is a univariate time series. Univariate refers to functions with only one variable. A univariate time series analysis is appropriate for when the mechanism by which the dependent variable is affected by the independent variable is unknown or hard to quantify (Profillidis and Botzoris, 2019d).

$$ED_{t+1} = f(ED_t\ ,\ ED_{t-1},\ ED_{t-2},\ ED_{t-3},\ error) \tag{3.2}$$

SARIMA consists of multiple forecasting processes. These are in order: Seasonal (S), Auto-regressive (AR), Integrated (I), and moving average (MA). With these processes different types of forecasts can be implemented. In the following sections ARIMA and SARIMA will be discussed in more detail. Before doing so it is important to look at the AR(p) and MA(q) parts of the processes. In an autoregression model the forecast is made using a linear combination of past values of the variable. Autoregression indicates that it is a regression of the variable against itself, thus univariate. In equation 3.3 the AR(p) is described, where c is related to the mean of the forecast and ϕ the weight of the different historical data points. The c is calculated by taking the mean (μ) times $(1 - \phi_1 - \phi_2 ... \phi_p)$. Autoregression models are usually restricted to stationary data and depending on the values of μ and p the model can become equivalent to white noise. This means random walk or random walk with drift can occur (Hyndman and Athanasopoulos, 2021). A moving average model, in this case a MA(q), uses past errors to create a forecast. Of course the past errors are not observed, but calculated. This makes it not a regressions model in the usual sense as can be seen in equation 3.4 where the θ_n stands for the respective errors and the ε_t for the calculated weights.

$$AR(p) : y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + error(\varepsilon_t) \tag{3.3}$$

$$MA(q) : y_t = \mu + \varepsilon_t - \theta_1 * \varepsilon_{t-1} - \theta_2 * \varepsilon_{t-2} - ... - \theta_q * \varepsilon_{t-q} \tag{3.4}$$

To ease calculation Profillidis and Botzoris (2019d) the use of a backshift operator B is described. This operator helps with referring to specific data points. For example B^{12} means 12 data points back. In the case of monthly data B^{12} refers to the same month one year prior. The backshift operator is defined as: $B^i y_t = y_{t-i}$. Using this backshift operator the equation can be reformulated as follows according to 3.3 and 3.4:

$$AR(p) : y_t = c + \sum_{i=1}^{p} \phi_i * B^i y_t + \varepsilon_t \tag{3.5}$$

$$MA(q) : y_t = \mu - \sum_{i=1}^{q} \theta_i * B^i \varepsilon_t + \varepsilon_t \tag{3.6}$$

The auto-regressive integrated moving average process, ARIMA(p, d, q)

To create an ARIMA(p,d,q) process the AR(p) and MA(q) should be added together, leaving the μ out. The AR(p) part should be stationary, this means the stationary of the time series should be checked before the calibration of the model. Tests for checking if it is indeed stationary will be discussed in 3.2. If the AR(p) part is nonstationary, the time series must be differenced as many times as necessary to become stationary. One difference is the change of the variable from the previous period to the current one, the d-th difference is between the d-th value and the current value of the variable under forecast (Profillidis and Botzoris, 2019d). The number of differences in an ARIMA model is noted with the d. In practice d usually takes a value of 1 or 2. The backshift operator can be used to describe the d-th difference as:

$$(1-B)^d * y_t \tag{3.7}$$

Seasonal auto-regressive Integrated moving average process, SARIMA$(p,d,q)(P,D,Q)_M$

When the data is seasonal, the seasonality can cause issues with the parameters used in the ARIMA model. In the SARIMA model a different ARIMA model can be fit for high-season and low-season data. The lower case letters in $SARIMA(p,d,q)(P,D,Q)_M$ stand for the same thing as in a ARIMA (p,d,q). the M stands for the length of the seasonality, and the upper case letters stand for the parameters during the high season. SARIMA or ARIMA models are widely used in ensemble models (Zhang, 2003), and Profillidis and Botzoris (2019d) use SARIMA as statistical model in the transport sector.

3.1.2 Learning Forecasting methods

In this section the four forecasting methods are discussed that use external variables and historical data to forecast the future. When looking at the same example about electricity demand, Hyndman and Athanasopoulos (2021) it can be defined as a mixed forecasting method based on the use of both historical and external data. The example is described in equation 3.8. The models discussed below are learning models, which in this context means that the parameters are calculated by computation. The computer tries to fit the best parameters for the forecast by trying out different setups of parameters. Python packages will be used for most of the calculations, in the following sections the core concepts of each methods are discussed. Machine learning models are relatively new in the forecasting space, which was dominated by the statistical methods like SARIMA. The M series is a forecasting tournament where teams compete on the same forecasting challenge to see whose methods are best. In the first three tournaments, held between 1982 and 2000, the general consensus was that simpler models do not under perform in comparison with more complex models (Xiao et al., 2023). That changes in the 4th tournament, when hybrid models were introduced. hybrid models make use of statistical methods and supervised learning methods. Since then learning methods have taken over the tournament and most relevant literature (Maçaira et al., 2018).

$$ED_{t+1} = f(ED_t,\ current\ temperature,\ time\ of\ day,\ day\ of\ week,\ error) \tag{3.8}$$

Artificial neural network, ANN

An artificial neural network (ANN) is a method that circumvents certain drawbacks of statistical regression methods. Namely, the assumption of linearity, the problems with large amounts of data and the practice that once the relationships in statistical models have been established they remain unchanged in the forecasting process (Profillidis and Botzoris, 2019e). ANN is a machine learning method that tries to imitate the way biological neuronal systems work processing information. ANN are not based on specific rules, but are developed through a trial and error process with successive calculations. The ANN is built up by layers where each layer has a certain number of nodes or neurons, as can be seen in figure 3.1 created by Hyndman and Athanasopoulos (2021). Each ANN has one input layer, one output layer, and a certain number of hidden layers. Each layer has a different number of nodes. The input layer has the same number of inputs as number of nodes. The output layer is the predicted value. The number of nodes in the hidden layer is one of the parameters that can be set to tune the model. The ANN model that will be used in this thesis will be a feed forward model. This means that the nodes send their output value to nodes they did not directly or indirectly receive input from. The outputs of the nodes in one layer are inputs to the next layer, where the inputs to each node are combined using a weighted linear combination. In each node the data of the previous layer is collected and weighted. This creates a summation of various inputs x_i and weights w_i, as shown equation 3.9 The summation ans calculation of ξ is the first step in processing, the next step is the transfer function. The transfer modifies the input to generate an output. Hyndman and Athanasopoulos (2021) suggests using a sigmoid function as transfer function, as shown in equation 3.10. The parameters $w_{i,j}$ and b_i are estimated.

$$\xi_j = \sum_{i=1}^{n} x_i w_{i,j} \tag{3.9}$$

$$f(\xi_j) = \frac{1}{1+e^{-b_i * \xi}} \tag{3.10}$$

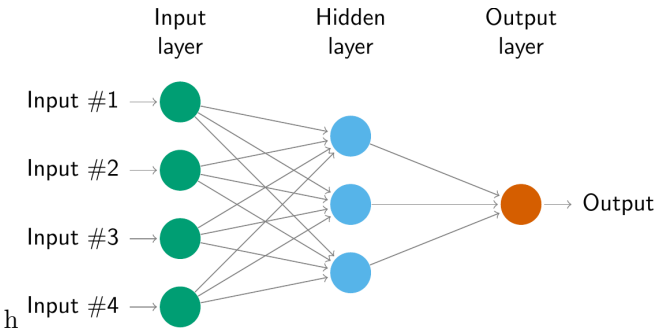

Figure 3.1: A neural network with four inputs and one hidden layer with three hidden neurons

Support vector regression, SVR

Support vector machines (SVM) is a supervised classification algorithm. It works by first graphing different points on a decision plane. It separates objects with different classes by a visible gap as much as possible between the classes. An SVM can do this for both linear and non-linear problems. Support vector regression is build on the SVM as it puts constraints on the possible outcomes and makes it an optimization problem. The idea behind Support vector regression is to create a ε insensitive loss function. Thus minimize:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_l + \xi_i^*) \tag{1}$$

Maçaira et al. (2018) describe the use of SVR in forecasting method as mainly used in the financial sector and as weather predictor. For this research the SVR model is chosen due to the robustness with little data points and the use in non-linear problem solving (Maçaira et al., 2018). A known disadvantage is the high computational power needed and the lack of transparency in the results (Khanna and Awad, 2015).

Random forest, RF

A Random Forest (RF) regression model combines multiple decision trees to create a single model. Each tree in the forest builds from a different subset of the data and makes its independent prediction. The difference between the RF methods and having multiple decision trees is the randomness included at each node of the tree. In a decision tree the split is made using the best value among all variables, in RF only a subset of the variables are used to create the split. The final prediction for input is based on the average or weighted average of all the individual trees' predictions. This is a example version of an ensemble method. This method can be improved by bootstrap aggregating, in which independent trees are constructed with a sample of the data set. RF has lower risk of over fitting due to randomness introduced by the subset of data used in bootstrapping and of the variables used to create splits (Liaw and Wiener, 2002).

LightGBM

LightGBM is a Gradient Boosting Decision Tree (GBDT), which is a Machine learning algorithm designed to lower computation time. Ke et al. (2017) discusses the improvement computation time while using LightGBM. They state that with large amounts of data, computation time can 20 times shorter when using LightGBM compared to other models. LightGBM has two novel methods that improve it over other tree growing methods: Gradient-based One-Side Sampling(GOSS) and Exclusive Feature Bundling (EFB). GOSS obtains accurate estimations of the information gain with a much smaller data size due to the exclusion of data instances with small gradient. This, leads to what Ke et al. (2017) refers to as leaf wise growth, which is visualized in 3.2. Leaf wise growth can be described as only growing the leafs of the tree that have the highest possibility of a good prediction. This is different from a normal decision tree, where all the leafs are grown to the same depth.

3.1.3 Comparative advantages of the different models

Each of these models is capable of handling multivariate time series which makes all of them applicable to for this research but each has their own strengths and weaknesses. These four models were chosen from

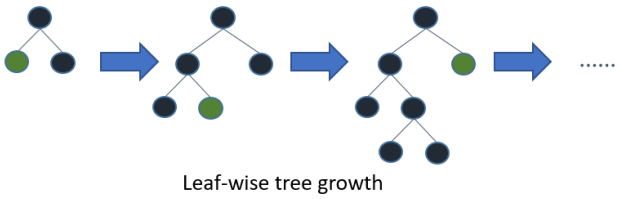

Figure 3.2: leaf wise growth

literature because of their use in industry. The literature review of Saraiva and Yoshizaki (2024) and Maçaira et al. (2018) describe the use of time series forecast, neural networks and support vector machines in the logistics sector. In the U.S. the bureau of statistics did an comparative study using machine learning in the transportation sector, where Random Forrest and Support perform well in this sector (Lim et al., 2022). The LightGBM is a more novel method, but with success in predict transport demand (Zhang et al., 2020). Each model has different strong and weak points. ANN is good in handling non-linear relations, but is not very transparent (Profillidis and Botzoris, 2019e). SVR is robust with little data points, but requires more of computational power (Khanna and Awad, 2015). RF has very little risk of overfitting, but requires lots of data (Liaw and Wiener, 2002). LightGBM is very fast, but harder to tune correctly (Ke et al., 2017). Each model had benefits and problems, which is why all these models will be tested to see which fits this dataset the best.

3.2 Model Setup and training

In this section the following question will be answered: *How to set-up and, if applicable, train forecasting methods?*. Two main topics need to be discussed to answer this question: the tuning of the models and the K-fold validation method.

3.2.1 Check for stationary data

For the autoregressive process the data needs to be stationary. To check for stationary data a statistical test can be used. Profillidis and Botzoris (2019d) advises the augmented Dickey-Fuller test. The null hypothesis H_0 for the Augmented Dickey-Fuller test is that the to be predicted variable has a unit root, which means it is nonstationary. The alternative hypothesis H_1 is the opposite, that the time series is stationary. The unit root means $\psi = 0$. The test is a student's t-test examining if the parameter $\psi = 0$, Profillidis and Botzoris (2019d) gives a calculation for the procedure and the test statistic. The calculation of the test statistic will be automated and the book of Profillidis and Botzoris (2019d) will be used to check the critical values.

$$\delta(y_t) = c + b * t + \psi * y_{t-1} + \sum_{i=1}^{p-1} \delta_i * \Delta y_{t-1} + \varepsilon_t \tag{3.11}$$

$$Augmented\ Dickey - Fuller\ test\ statistic = \frac{\hat{\psi}}{standard\ error\ \hat{\psi}} \tag{3.12}$$

3.2.2 K-fold cross validation

Cross-validation is a resampling procedure used to evaluate machine learning models and train hyperparameters. The simplest method is to define a training and test set from the data. However, if all available data is required for training as well it is better to use cross validation. Cross validation works by creating smaller training sets and test sets and evaluating the accuracy measures and averaging them out on all sub sets. Since the order of data points matters in time series it k-fold validation works by creating incremental subsets of the total data, as can be seen in figure 3.3.

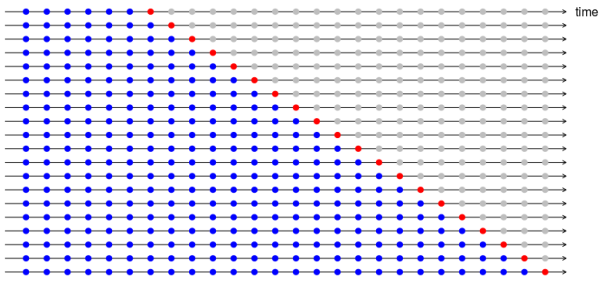

Figure 3.3: K-fold validation visualisation

3.2.3 Hyperparameter tuning method

Each machine learning model has different hyperparameters that need to be chosen before the model can run. Hyperparameters are configurations of the model, which highly influences the outcome and performance of the model. The tuning of the hyperparameter is the systematized way of looking for the configuration that works best for the Data. Wainer and Cawley (2021) describe two ways of cross validation in the tuning of the hyperparameters: nested and flat cross validation. Nested cross validation means tuning hyperparameters for each fold in the K-fold cross validation. This version leads to low bias because only the training sets are inputs for the hyperparemeter tuning algorithms. Flat cross validation does incur bias as the hyperparameters are trained on the entire dataset before the K-fold cross validation is done. According to Wainer and Cawley (2021), the bias incurred is acceptable and the loss of accuracy is minimal. A flat cross validation is necessary to use the parameters for forecasting, since a nested cross validation does not provide a singe set of hyperparameters. Picking one set of the hyperparameters would render cross validation invalid, because the results from the nested cross validation are not representative of a single set of the hyperparameters used. The Hyperparameters are tuned via Randomizedsearch as described by Bergstra and Bengio (2012). Randomizedsearch is a method of Hyperparameter optimization which is more efficient than the classical gridsearch methods. For gridsearch a multidimensional grid is created and every configuration possible is tested. With Randomizedsearch random samples are taken from the data grid to create configurations. The algorithm scores these samples and resamples the better scoring ones to find better configurations. This leads to lower computation needs without loss of accuracy when comparing with a more extensive gridsearch method.

3.3 Evaluating forecasting Methods

In this section the following question is answered: *What evaluating methods are used for forecasting methods?*. This is done by analyzing the methods discussed by Profillidis and Botzoris (2019c). The most well known measures are mean absolute deviation, mean squared error and mean absolute percentage error. These Statistics are not helpful by themselves, since they give an arbitrary number(Profillidis and Botzoris, 2019c). The statistic outcomes need to be compared between different models to chose the 'better' model. Defining a forecasting model as good is harder to do with these evaluating methods. Thus we also look at Theil's inequality coefficient, also known as Theil's U. Theil's U has a more normative outcome about if the model is a good fit for the data. Profillidis and Botzoris (2019c)

3.3.1 The Mean absolute deviation

The *mean absolute deviation* (MAD) takes the sum of the absolute difference between the actual values y_i and the predicted values $\hat{y}_i$. After which that number is divided by the number of observation n. The MAD reflects the variability of the data points around the mean, a higher mean suggest that the data points are more scattered. For forecasting that means that a lower MAD would be a better forecast. The drawback of MAD are that it cannot look for patterns or trends in data, since it treats all deviations equally.

$$MAD = \frac{1}{n} * \sum_{I=1}^{n} |y_i - \hat{y}_i| \tag{3.13}$$

3.3.2 The mean squared error

The *mean squared error* (MSE) is calculated by summing up the squared errors, and dividing them by the number of observations n. Error is calculated by taking the actual values y_i and the predicted values $\hat{y}_i$. The MSE punishes large errors more then small errors relative to the MAD, due to the squaring of the error. The *root mean squared error* (RMSE) is a variation on the MSE, where the entire formula is placed in a square root. Conceptually the difference between MSE and RMSE is like the standard deviation and the standard error. The MSE looks at the spread of predicted values around the actual values, while the RMSE looks at the error of predicted values.

$$MSE = \frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.14}$$

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.15}$$

3.3.3 The mean absolute percentage error

The Mean absolute percentage error (MAPE) is a good forecast accuracy measure if the data has significant seasonality or variability (Chopra and Meindl, 2016). The MAPE is calculated nearly in the same way as the MAD, but the measure is made a percentage by dividing by y_t and doing the entire sum times 100.

$$MAPE = \frac{1}{n} * \sum_{I=1}^{n} |\frac{y_i - \hat{y}_i}{y_i}| * 100 \tag{3.16}$$

3.3.4 Theil's inequality coefficient

Theil's inequality coefficient describes the accuracy of the model between zero and one. When the coefficient is zero the model is perfect, and when the model is one the model lacks any forecasting ability. Profillidis and Botzoris (2019c) consider values less than 0.5 to be good and values less then 0.1 as excellent.

$$Theil's \;\; U = \frac{\sqrt{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{n} * \sum_{i=1}^{n} (y_i)^2} + \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (\hat{y}_i)^2}} \tag{3.17}$$

The above described accuracy measures all have one number as output. These measures don't describe the composition of the error at all. The powerful point of Theil's U is the fact that it can be decomposed into bias U^M, variance U^V, and covariance U^C. This decomposition of the accuracy measure creates a means of analyzing each component separately. Since the bias, variance and covariance are proportions of Theil's U they add up to one: $U^M + U^S + U^C = 1$. The Bias U^M is an indicator for the systematic error, which should be minimized as much as possible. The variability proportions, U^S, indicates the ability of the model to replicate the degree of variability in the predicted variable. The non systematic error of the model, which is impossible to avoid, is the covariance proportion U^C. Since every forecast will have a non systematic error, the covariance is less worrisome than bias and variance (Profillidis and Botzoris, 2019c).

$$U^M = \frac{(y_i - \hat{y}_i)^2}{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.18}$$

$$U^S = \frac{(\sigma_Y - \sigma_{\hat{y}})^2}{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.19}$$

$$U^C = \frac{2 * (1 - r_{Y\hat{Y}}) * \sigma_Y * \sigma_{\hat{Y}}}{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.20}$$

Where

$$\sigma_{\hat{Y}} = \; Standard\ deviation\ of\ the\ fitted\ model$$

$$\sigma_Y = \; Standard\ deviation\ of\ the\ observed\ values$$

$$r_{Y\hat{Y}} = \; Pearson\ correlation\ coefficient\ between\ actual\ and\ fitted\ values.$$

3.4 External Variables

In this section the external variables discussed in chapter 2 will be validated in theory. In order to answer the question:*How can macro economic indicators that influence the transport sector in the Netherlands be used in a forecast of Nijhof-Wassink?* data is needed. Data on external variables gathered from trustworthy sources. before inputting the data into a forecasting model, the coupling or decouping should be checked, the theory behind coupling is explained here as well.

3.4.1 Data-sources for macro economic indicators

Do the identified macro economic indicators have trustworthy and complete data? To answer this question, we will look at the data available at CBS. CBS is the dutch bureau for statistics, which is a relatively reliable source for data about the Netherlands and its economics. Since the Macro economic indicators are so broadly defined we can use the open data API to quickly download and use many different datasets. The datasets that cover the content referenced in previous sections are stated in table 3.1. The rows are different tables in the CBS database. The data points in these tables are responses to questionnaires in percentages, prices, and other business cycle statistics.

The datasets were chosen to either cover part of the chemical sector, or be be an estimator of general market trends. CBS does not have oil prices in their database, so pump prices are used instead.

Table 3.1: Used CBS tables with external factors

Table CBS	Original title	Translation
85609NED	Conjuctuurenquête nederland	Business survey in the Netherlands
81234NED	Producentenvertrouwen	Producer confidence
85612NED	Ondernemersvertrouwen	Business confidence
80416NED	Pompprijzen	Pump prices
85806NED	Nijverheid	Industry
85771NED	Producentprijzen	producer prices
83133NED	Consumentprijzen index	consumer price index
83693NED	Consumentvertrouwen	Consumer confidence

3.4.2 Coupling and decoupling

(Profillidis and Botzoris, 2019a) mentions the concept of coupling and decoupling of data, focused on the question if transport is coupled with the gross domestic product. This is an macroeconomic method which we will apply to the micro economic environment of Nijhof-Wassink. Coupling measures the ratio of the rate of change of the demand and the external variable, as can be seen in equation 3.21 For two variables to be considered coupled the value of $\epsilon_{X,Y}$ has to be between 0.8 and 1.2, or -0.8 and -1.2. In this thesis we will assume that if a dataset of an external variable is coupled with the data of Nijhof-Wassink it is related enough to take the external variable into account.

$$\epsilon_{Demand,\ GDP} = \frac{\%\Delta Demand}{\%\Delta GDP} \tag{3.21}$$

3.4.3 Correlation

The correlation coefficients measure the strength of linear relationships between variables (Hyndman and Athanasopoulos, 2021). In this case the variables are the external variables and the order data. Correlation coefficients are computed as seen in equation 3.22, the coefficients are bounded between -1 and 1. If the coefficient is -1 or 1 the correlation is perfect, and a value of 0 means there is no correlation. For taking a correlated variables into account, Hyndman and Athanasopoulos (2021) warn that correlation does not mean causation. Two indicators can be very much correlated without a causal link. Without the causal link the indicator can still be useful as a predictor, however the causal link is highly preferred in an indicator.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{3.22}$$

3.5 Conclusion chapter 3

In the third chapter forecasting as a general topic is discussed, and then specified in two main parts; The description of the models, the setup of the models and the evaluating methods, as part one. The external variables that can be used as features, the dataset of these variables, and the way to link the variables with the data, as part two. In this conclusion the two main questions are answered distinctly.

To answer main question 2: *What are the relevant models for demand forecasting in Transportation?* , four topics are discussed. Fistly, the difference between statistics and learning models is described. Choosing learning models as forecasting models due to their ability in modeling nonlinear behavior and the recent success of learning models in forecasting tournaments. Secondly, four different learning models are presented to use for Nijhof-Wassink. That is, random forrest, support vector regression, artificial neural network, and LightGBM. These models were chosen because of their multivariate ability and history in the transportation sector. Thirdly, the training methods are discussed. Namely, the use of K-fold validation and hyperparameter tuning with RandomsearchCV. Fourthly, four different evaluating methods are discussed. The most interesting one of these is Theil's coefficients, which gives more insight then only a comparison between models. However, due to time constraints and finding mainly MAPE as evaluating method in literature, moving forward MAPE will be used.

To answer main question 3: *How can macro economic indicators that influence the transport sector in the Netherlands be used in a forecast of Nijhof-Wassink?*, the data sources, and linking methods were discussed. The data source for accurate and clean data for external variables will be the Dutch bureau of statistics, which can be imported by an api. Two different ways of looking at which external variable to use were discussed, coupling and correlation. Correlation looks at the small changes over time, while Coupling looks at the overall trend of the data. Both can give different insight in which external variable to use, which is why both will be used in the forecasting procedure of next chapter.

Chapter 4

Creating forecasts

> "The first 90 percent of the code accounts for the first 90 percent of the development time. The remaining 10 percent of the code accounts for the other 90 percent of the development time."
>
> Tom Cargill, Bell Labs

In this chapter the main questions *Which forecasting model performs best for Nijhof-Wassink?* and *How can macroeconomic indicators that influence the transport sector in the Netherlands be used i na forecast of Nijhof-Wassink?* will be answered. To this twelve forecasting models will be made, based on the four supervised learning methods discussed in section 3.1.2. For each of the four methods 3 models will be build: one single model without external variables, a lagged model, and a hybrid model, totaling to 12 models total. Firstly, the coupling and correlation assessments are done for the datasets described in chapter 3. Secondly, the validation of the accuracy measures are discussed in the context of external features. Thirdly, the implementation and choices of the hyper-parameter grids are discussed. Fourthly, the results of the twelve models are discussed.

4.1 Procedure Set-up

The procedure to chose on model to take as basis model is as follows: First determine the indicators that are not external. After this a dataset is created including all external variables to check for coupling. Then a correlation matrix is set up to find if the tested indicators are not to highly correlated with each other. Finally the validation of the model is done by scoring them based on the chosen metrics. From this a list of best scoring models is produced.

4.1.1 Base indicators

The non external indicators are mostly time related indicators. These indicators are used to introduce trend and seasonality into the model. To represent trend the indicator year is chosen, and to represent seasonality the indicator month is used. As discussed in chapter 2 there is a big decrease every December in the number of orders. To account for the decrease in business days in December the number of business days is also used as an indicator as # BDays.

4.1.2 Test for coupling

To test for coupling, the external data first needs to be prepared as not all datasets use the same step size in time (i.e. weekly, monthly, quarterly). This results in 2016 and 2023 being incomplete years. In order to fix this the % change will be taken from the summed years 2017 or 2018 and 2023. In the Dataset from CBS the chemical industries are represented on both an aggregate level and dis aggregate level.

4.1.3 Correlation

A correlation matrix is made to look at the correlation between the data of Nijhof-Wassink and the 44 external indicators. These indicators are found in the datasets of CBS as defined in chapter 2. An indicator is considered well correlated if the correlation result is higher than 0.5. Afterwards it should be checked if the correlation between the indicators is not to high, since that could bring unwanted bias into the forecast.

4.1.4 External feature selection

The first feature taken as external source is the number of trucks operated by Nijhof-Wassink. The further coupled datasets are Calendar Adjusted Turnover Total, season adjusted Turnover total and total turnover. These indicators are highly correlated, namely between 0.98 and 0.99. Since the total turnover is the broadest indicator and the rest of coupling is not significantly different that one is chosen. The correlation matrix gives seven indicators with a correlation higher than 0.5. These seven can be divided in two groups by correlation within the group: fuels and turnover. The turnover abroad has the highest correlation with the data from Nijhof-Wassink from the different turnover indicators, and LPG from the different fuel prices. Thus the external features are: the number of Trucks (# Trucks), Turnover in the chemical industry (totaleOmzet_4), Turnover abroad in the chemical industry (OmzetBuitenland_6), and LPG prices (LPG_3).

4.1.5 K-fold validation with gaps

The time series K-fold is used in all the different models to increase validity of the accuracy measures. This is done by creating 12 different test and train sets, also called folds. The test set becomes bigger with the length of the test set of the previous fold. The idea of the folds is to simulate forecasts and gain a measure of accuracy for the model. To do this the goal is to simulate forecasting 12 months ahead in one month intervals The length of the test set chosen for this is 6 months. 6 months as test set is a balance between the volatility 1 month test sets create and the amount of data needed for 12 month test sets. To still simulate to forecast 12 month ahead gaps are used between the train and test sets. So, to simulate forecasting 12 months ahead we will create a gap between train and test set of 12 months. An example of how K-fold cross validation with gaps looks like can be seen in figure 4.1. For each of the 12 gaps 12 folds are used. This leads to essentially 12 flat k-fold validations for each model to come to an accuracy measure.

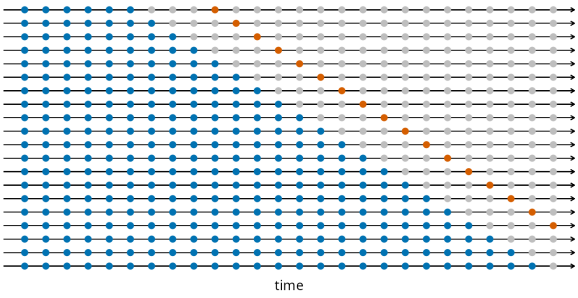

Figure 4.1: K-fold cross validation with gap 3

4.1.6 K-fold with gaps with external features

This procedure with K-fold and gaps is the base procedure without external variables, which is necessary to add external features into the model. In order to take external features into account in the validation and forecasting methods it is necessary to adapt the dataset. To create a forecast it is necessary to have data point at the future date that needs to be forecast. The base model has time related features which can be counted ahead into the future, like year and month. To do the same with external variables two variants are discussed below.

Set-up lagged models

The lagged procedure is simpler to implement, and computationally less expansive than the hybrid model. Before the train and test sets are defined the external variables are transferred ahead in time with the amount

of the test set plus the gap. So, for the data needed for gap 1 the external features are transferred 7 months, 6 for the length of the test set and 1 for the length of the gap. This transfer of the data only needs to happen once per gap, which makes it easy in coding and fast in execution.

Set-up hybrid models

SARIMA is used as forecasting method for the hybrid models. The choice for SARIMA is to keep a more linear and simpler element in the forecast. The SARIMA model used is an auto_SARIMA model, which needs data to train on as well. For every training and test split the training data is used as input for the forecast of the test set. The length of the forecasts is determined by the test set and the length of the gap. This greatly increases the computational time in relation to the lagged models, since for every fold of each gap length a new SARIMA model needs to be trained.

4.2 implementation and feature choice

In this section the implementation of the different machine learning models is discussed. The hyperparameter grids are defined by either different options or a minimum and maximum value. RandomsearchCV is a heuristic that can look through option relatively quickly, which means the size of the grid is less of a concern than with grid search. With this in mind the choice was made to expand the grid every time the randomsearchCV came in the top 10 percent of the grid. The randomsearchCV runs every time the model is used, to update the grid and new historical or external data.

ANN

For the artificial neural network the hyper-parameter grid consists of 5 different parameters. Namely, Hidden layer size, activation method, Solver Method, Alpha, and learning rate. The hidden layer size is varied from 1 to 10, where the integers represents the number of neurons in a hidden layer. Activation Method controls the function which transforms the input in the neurons. In this case the model can choose between Hyperbolic Tangent, Rectified Linear Unit and a Sigmoid. All three are not perfect for this use case, that is was left up to the model to chose the optimal of the three. Solver method determines the optimization algorithm for updating weights, which influence the learning potential. The model can choose between Adaptive Moment Estimation and a quasi-Newton optimization, both are native to the regressor package in Python. The quasi-Newton optimization is better for smaller datasets and Adaptive Moment Estimation is better for larger datasets. Alpha is the regularization term which help prevent overfitting. Usually Alphas are between 0 and 1, the model used picks between 0.0001 and 1. A low aplha means weak regularization and an alpha closer to 1 means stronger regularization. Learning rate determines the rate at which the weights are updated. This can be either constant or adaptive, which means if the model does not improve after some iteration the model stops. Adaptive can be useful to stop over fitting on the data, but could lead to worse models.

SVR

For the Support vector regressor the hyper-parameter grid consists of 3 different parameters. Namely, the C, Epsilon, and Kernel. The C is the regularization factor, where smaller numbers prioritize smoothness over fit. Higher numbers prioritize fit, which makes the model more accurate, but increases the risk off overfitting. The C grid used is between 1 and 100. The epsilon is the tolerance within no penalty is given. A smaller Epsilon makes the model more sensitive to small errors, while a larger epsilon is less sensitive to small deviations. The Epsilon chosen are between 0.1 and 1. The Kernel is the function which transforms the input to a higher-dimensional space. The options are Linear transformations, polynomial transformations, or radial basis function. Linear is for simple relationships, polynomial is useful for complex but still somewhat linear relationships. Radial basis function is used for complex non-linear relationships.

RF

For the Random Forrest the hyper-parameter grid consists of 4 different parameters. Namely, Number of estimators, max depth, minimal samples during splits, and minimal leaf samples. The number of estimators is the number of trees in the Forrest. Larger number of estimators increases the accuracy of the model, but also increases computation time. The chosen range of number of estimators is between 50 and 200. Max

depth is the size of the tree or estimator. Shallower trees are faster and prevent overfitting, while deeper trees can capture more complex relations. The grid for max depth is between 5 and 20. The minimal samples needed to split a node in the decision tree. Smaller values lead to more splits and deeper trees, while larger values restrict splitting, making the tree simpler. A grid of values between 2 and 15 is chosen for the number of indicators needed to split the node.

LightGBM

For the LightGBM the hyper-parameter grid consists of 6 different parameters. Namely, earning rate, number of leaves, number of estimators, maximum depth, L1 regularization (reg_alpha), and L2 regularization (reg_lambda). The learning rate controls how much the models changes each iteration of training. With lower values of learning rate the learning process becomes more gradual, which improves accuracy but increases processing time. The chosen grid for Learning rate is from 0.01 to 0.5. The Number of leaves is the maximum number of leaves per tree. A larger number of leaves captures more complexity but also lead to overfitting. The number of leaves is between 5 and 50. The number of estimators and the max depth is the same as with Random Forrest. Between 50 and 200, and 5 and 20 respectively. The reg_alpha and reg_lambda are both true or false. Both are penalty terms added based on either the absolute (reg_alpha) value or the squared (reg_lambda) value of the model. The regularization helps prevent overfitting.

4.3 Results of forecasting models

The average MAPE across all gaps can be seen in table 4.1. In this table the lower scores are a result of better models. Interestingly, the best performing model does not include external features. Overall the forecast features work better then lagged features and the models without external features. The forecast features and the base models are very similar in average accuracy, but the difference is which forecasting method is the best. Another measure of looking which model is better can be how much the accuracy decreases between gap 0 and gap 11. The results of this analysis can be seen in Figure 4.2. The expectation is that the average in-sample MAPE increases with the increase of the gap. The average increase in MAPE across the gaps is 25% The twelve models show quite a big spread, from a 5% increase as a minimum to a 40% increase as maximum. however, both of these extremes are in relatively bad performing models. The minimum increase is at the base model ANN, and the maximum increase is at the lagged SVR model.

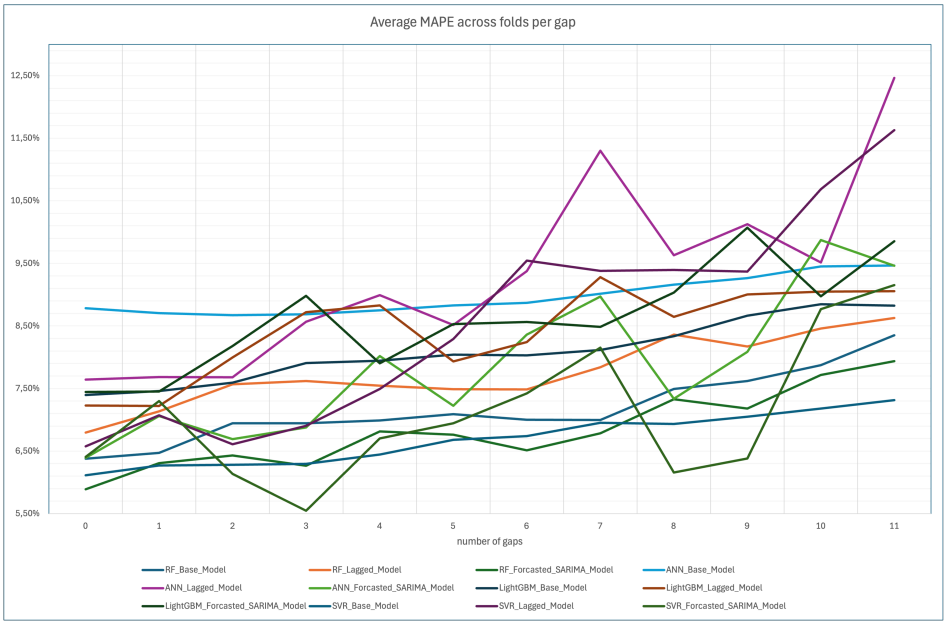

Figure 4.2: Average MAPE across folds per gap for each model

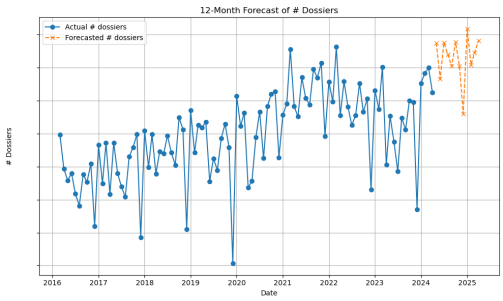

Figure 4.3: Best model, SVR without external features

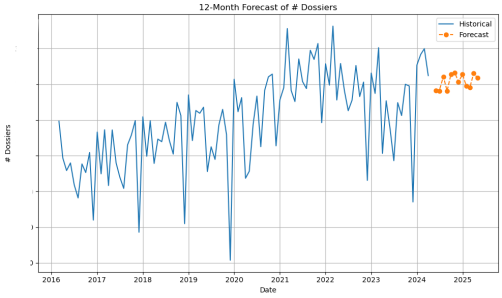

Figure 4.4: Best model with external features, Random Forrest with forecasted features

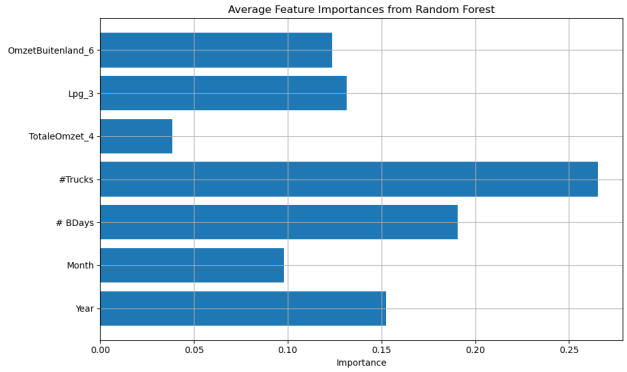

Figure 4.5: Relative feature importance of RF model with forecast external features

4.3.1 Analysis of best models

The two best models are the SVR model without the use of external features and the RF with forecast external features. The forecasts made with these models can be seen in Figures 4.3 and 4.4. The SVR model is the best scoring model, with only date features used as input. It is good to note that the SVR model scores much worse when external features are included. For the RF it is interesting to look at the feature importance, which can be seen in figure 4.5. The number of trucks is the most influential features, followed by the number of business days and year. The big difference between total turnover and the rest could be an sign that the indicator is not useful at all. Another explanation is that total turnover and turnover in foreign countries looks to similar and only one is really used in the forecast.

4.3.2 problems with forecasting using models

When trying to forecast the time period after 2024-4, the data from 2024-4 until 2024-9 can be used a validation of performance. The SVR model was forecasting significantly too high, on average 400 orders to high. To help with the analysis the time features used should be looked at in relation to the trend of the data. The RF with external features created a flat line as external forecast, which is not a good input for the model. To improve on this the way of forecasting external features should be looked at. In sample both models perform the best, but both prove ineffective when forecasting in their current state.

Table 4.1: In-sample average MAPE across gaps of the different forecasting methods

Average MAPE	Base model	Incl. Ex. features lagged	Incl. Ex. features forecast with SARIMA
RF	7.26	7.86	6.91
ANN	8.99	9.28	8.05
LightGBM	8.16	8.43	8.70
SVR	6.74	8.90	7.29

4.4 Conclusion Chapter 4

In this chapter twelve models are defined and evaluated to answer two main questions. To answer the first main question 4: *Which models can forecast Nijhof-Wassink's demand, while taking external factors into account?*, three steps are taken. firstly, discussing the three procedures of forecasting; a base model without external features, a lagged procedure, and an forecasted procedure. Secondly, by setting up the four forecasting models; ANN, SVR, RF, and LightGBM. Thirdly, to make use the models the hyperparameters need to be defined, this is done with RandomsearchCV. K-fold validation with gaps is used to create a robust accuracy measure. These steps create twelve models, eight of these models take external factors into account.

To answer the second question of this chapter question 5: *Which forecasting model performs best for Nijhof-Wassink?*, the in-sample performance in analyzed. The Lagged procedure Less accurate than the base or forecasted procedures, thus to take external factors into account it is best to use the forecasted external factors as input. The ANN and LightGBM show worse performance across the board, means the models do not work well on this dataset. The Base model SVR and the forecast external features RF model performs best, but both show problems while forecasting out of sample. In the RF model the importance of the features is very varied, from which the the not relevant features can be derived. In this case the total turnover feature is nearly not used at all, while the other external features are used. From this we can conclude that for this analysis the coupling approach is less accurate than the correlation approach. In the next chapter a more in depth analysis will be done of possibilities of solving the issues of the models, by creating a hybrid forecasting model with both models combined and checking out-of-sample performance with new data.

Chapter 5

Model Validation

In this chapter the Model described in the last chapter will be further analyzed, improved, and made more applicable for Nijhof-Wassink. This is to answer the research question *How can the chosen solution be improved to enhance performance, generality, and validity?*. A MAPE of 7% is very low, but the model does create an adequate forecast that seems logical. This can have multiple causes; like bias in the accuracy measure, noise caused by the exogenous variables, or an issue with the implementation of the forecast. The analysis is done in the following five steps. Firstly he difference between in sample and out of sample performance is discussed to analyze the performance of the current model. Secondly the forecasting method for exogenous variables will be looked at and improved, which should improve the performance of the overall model. Thirdly to account for the problem of forecasting to high on average the included features analyzed and adjusted. Fourthly an out-of-sample time series split test is done with the newly received data. Lastly the general dataset is split in Dry bulk (DBL) and Liquid bulk (LBL), and the results of these separate forecasts is analyzed.

5.1 In-sample versus out-of-sample performance

An in-sample MAPE of 7% is very good, but could be a sign of over fitting if the forecast is not good. Overfitting is a stark difference in in-sample and out-of-sample performance. In-sample accuracy shows how well the model can fit on the data provided, while out-of-sample accuracy shows how well the model can predict unseen data. It is obvious that out-of-sample performance is the more important in the context of forecasting, since the future is always unknown. Because flat cross validation was used in the creation of the model, the model trained on all data available to fit the hyperparameters. This means that it could be argued that all validation was done in-sample. When looking at the forecast of the model in figure 5.1, the forecast is higher and more packed together than expected when looking at the downward/stabilizing trend from the year 2022 onward. However, since a RandomsearchCV heuristic was used, which inherently minimizes bias, it is unlikely that over-fitting is the main issue with this forecast.

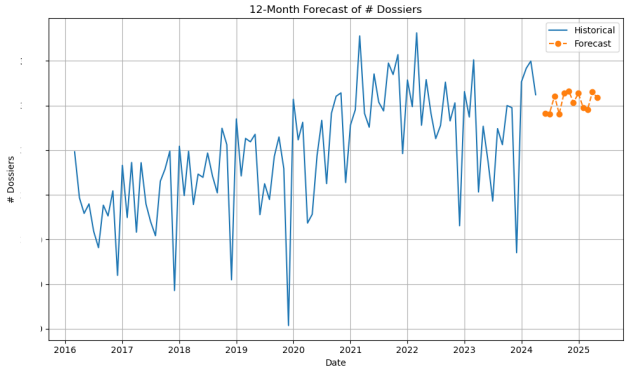

Figure 5.1: Forecast: Random Forrest with SARIMA

5.2 Forecasting method for exogenous variables

When looking at the performance of SARIMA in the forecasting model, it does not perform very well. The output for the SARIMA is stationary, which makes it a bad forecast for the external variables. In the systematic literature review of Maçaira et al. (2018), about using explanatory variables in time series analysis, the number of papers per decade using ARIMAX between 1967 and 2016 is dropping, while the use of support vector machines increase during the same time period. Since the SVR model is a support vector machine, and scored the best MAPE as a forecasting method on this dataset, it seems beneficial to swap out the SARIMA model with the SVR model. The result of the SVR as external variable forecaster can be seen in figure 5.2. However, just as with the SVR model discussed in chapter 4, the results are too optimistic. Changes in features will change the input of the model, which should lower the output.

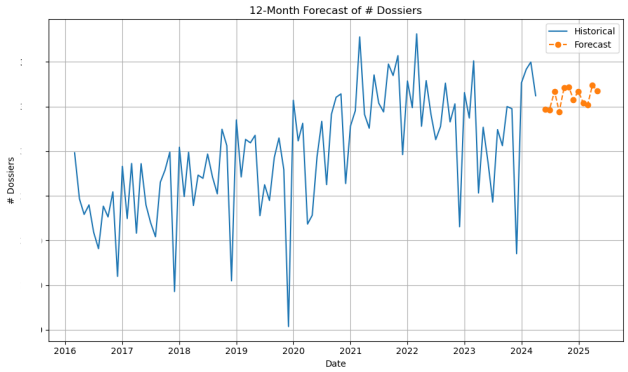

Figure 5.2: Forecast: Random Forrest with SVR

5.3 Feature analysis

Not all features are equally relevant in the model. When looking at the feature importance in figure 4.5, there is a stark difference between the relative importance of the features. The total turnover in the chemical industry is not used at all, So it can be taken out of the forecast. It can be concluded that the theory of coupling has proven less effective compared to correlation for this dataset, since lpg and foreign turnover is used significantly in the forecast. The SVR model is to optimistic when looking at the data from April 2024 onward, which is most likely caused by the significant increase in orders between 2016 and 2024. Two solutions can be implemented to loosen the relationship between years and number of orders. Either delete the earlier years which don't follow the preferred pattern, or delete Year as a feature in the model. Leaving year out as feature is the better option, since keeping more data for the model to train on increases the accuracy. When leaving out years starting from 2016 problems arise with the data point necessary to perform k-fold cross validation with gaps. The results from leaving out year and turnover can be seen in figure 5.4, and the feature importance in figure 5.3. The forecast is less optimistic, which is more realistic. The average MAPE of this model is 8.21%, which is less accurate then the SVR and RF model with SARIMA.

5.4 Out-of-sample test

The new data received is from May 2024 until September 2024. These five months of new data are used as a secondary validation test for the models. This test will be out-of-sample because nothing is changed in the Hyperparameters. The forecasts of the three models will be used to calculate a MAPE with the new actual data. This way of splitting the time series in a part were the Hyperparameters are trained on and a part that has not been seen by the time series is a simplistic way of validating a forecasting model. Generally, this way is seen as less robust and more prone to bias. However, the bias introduced by flat cross validation is not introduced in this score. The models are trained and on the data until April 2024 and tested on the data from may until September 2024. With the actual from Nijhof-Wassink an out-of-sample MAPE can be calculated for this test.

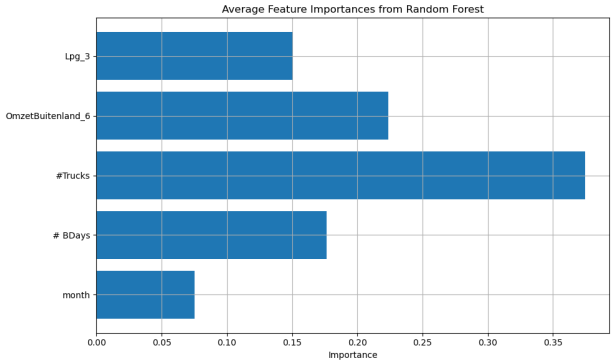

Figure 5.3: Feature importance of the Random Forrest with less features

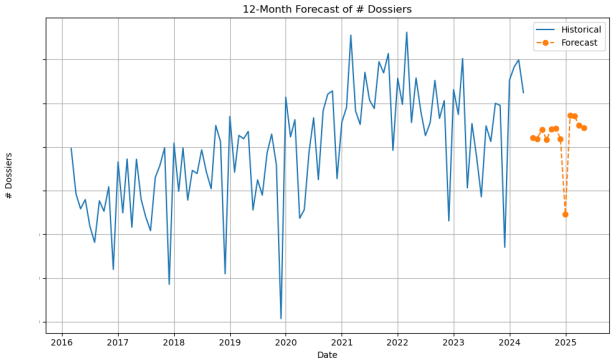

Figure 5.4: Forecast, Random Forrest with SVR, using less features

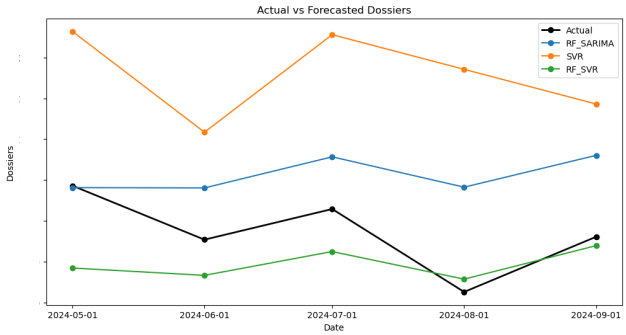

Figure 5.5: Actual vs forcasted: SVR, RF with SARIMA, RF with SVR

The results can be seen in table 5.1 and figure 5.5. The shown results are remarkable, since it shows the out-of-sample is the opposite from the In-sample MAPE. Although the sample-size of this experiment is very small, the RF models perform much better than the SVR. The big difference in performance in the SVR model can be attributed to partly over fitting on the data. Because of the risk of over fitting, for the rest of the chapter the support vector will be used to forecast external data and random forrest for the forecast of the orders of Nijhof-Wassink.

Table 5.1: In-sample and out-of-sample MAPE of the three models, SVR, RF with SARIMA, and RF with SVR.

	MAPE score	
Method	In-sample	Out-of-sample
SVR	6.74%	11.54%
RF_SARIMA	6.91%	4.30%
RF_SVR	8.21%	2.61%

5.5 Dis-aggregating the dataset

To make the forecast more insightful in the monthly meetings, the forecast needs to be made on a less aggregated level. Checking if on division level the model still has a good MAPE is also a good way to check how well the model works in different circumstances. The same forecasting model is used for both the dry bulk and liquid bulk division. The RF and support vector regressor are fitted on the new datasets, with no other changes. The number of DBL orders is roughly twice as big as the number LBL orders, which would lead to the expectation of the model being more in-tune with the DBL dataset. This is also what can be seen in the MAPE: DBL has an average of 8.7% and LBL an average of 10.6% The datasets and forecasts

can be seen in figure 5.6 and 5.7. From this it can be concluded that the model can be very well adapted to DBL but could use some improvement for LBL.

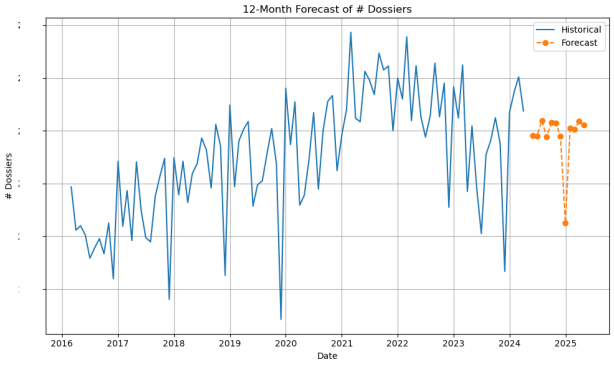

Figure 5.6: Forecast: DBL

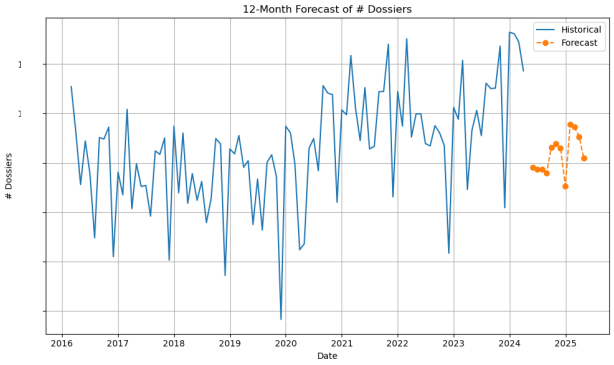

Figure 5.7: Forecast: LBL

5.6 Conclusion chapter 5

In this chapter the RF regressor and SVR chosen as the best models in chapter 4 were further analyzed and enhanced to answer the main question 6: *How can the chosen solution be improved to enhance performance, generality, and validity?*. All models discussed have lower accuracy then indicated by the MAPE score by the bias, which is inherent in the flat cross validation of hyperparameter tuning. However, using MAPE can be an easy way to state that the performance of a model is in the expected range. The SARIMA part of the model was changed to SVR, since the base SVR model was the best performing model in chapter 4. This is an interesting result, because in SARIMA is widely used in hybrid forecasting methods.

Two features were dropped in the model, namely turnover and the year input, as the first had little to no influence on the model and the second caused to much over fitting. Attributing an increase in orders every year does not reflect the reality very well. An interesting conclusion which can be drawn from the feature importance graph, is that the features chosen via a qualitative route or correlation where much more important than via coupling. This is an affirmation for Hyndman and Athanasopoulos (2021), who give warning for using only computation without qualitative assessment to find explanatory variables for forecasting.

An out-of-sample test was done with new data as extra validation, with surprising results. The RF with SVR scored very well in the out-of-sample test, while performing worse on the previously done in-sample validation method. The SVR scored much worse on the out-of-sample test than the in-sample test. It is good to continually keep testing both models, since a good in-sample MAPE does not automatically mean good forecasting performance. the out-of-sample tests indicates that the SVR model is partly over fitting, because of that the RF is used during the tests on the dis-aggregate data.

The model was lastly validated by trying it on disaggregated dataset of dry bulk and liquid bulk. For dry bulk it works quite well, which is not surprising since it is the majority of the general dataset. For liquid bulk it also works, but most likely some other features work better with this dataset.

Chapter 6

Conclusion

In this chapter the results of the research are discussed. Firstly, the conclusion is given together with an answer to the research question. Secondly, the limitations and areas of further research are discussed. Thirdly, a set of recommendation is given, directed at Nijhof-Wassink meant to be used during the implementation of the created solution.

6.1 Conclusion

To aid the decision making process, Nijhof-Wassink chemical division wanted quantification of the prediction of the future orders. The main decision making process takes place in a monthly meeting on tactical level, where decisions are made that directly influence capacity planning and profit margins. Currently no forecasting method is used to predict future demand, only qualitative statements about the markets are brought up. The wish of Nijhof-Wassinnk to quantify these rumors and create a forecast leads to the criteria that external features need to be used in the forecasting. From these criteria the research question is formulated as *"How can Nijhof-Wassink's chemical division use historical data and external factors to better understand and forecast the demand?"* Because the meetings are monthly, the time frame for the forecast is monthly as well. In the next sections the solution and answer to the research question are discussed.

6.1.1 Use of historical data

The chemical division can be split up into the dry bulk logistics (DBL) and liquid bulk logistics (LBL). The Data of dry and liquid bulk is first aggregated to make a model, this is to ease to process and create a more accurate overall model. The historical data of Nijhof-Wassink consists of the orders per month, but also of the Number of Trucks and number of working days per month. These last two came about in qualitative discussions with Nijhof-Wassink about the performance of the model. The number of trucks is a measure of capacity, if the number of trucks increases it figures the number of orders also should increase. The number of business days is a way to explain the seasonality of the data, every year there is a big dip in orders, which is explained by the closing of Nijhof-Wassink in the Christmas holiday.

6.1.2 Use of external factors

The process of looking for external factor to take into account while forecasting has multiple stages. Namely, a qualitative assessment, finding clean data, a quantitative assessment, using the chosen variables as features and finally analyzing the feature importance. The best external factors have a qualitative explanation and quantitative support. To achieve this support two methods were tried: coupling and correlation. The qualitative assessment meant narrowing down the features to macro economic features about the chemical industry instead of only using macroeconomic features about the Netherlands. This had two reasons. For one, to make looking for data manageable and to secondly already discard features which would have no explainable impact. For clean data the dutch bureau of statistics was used (CBS). The quantitative assessment of the features using coupling resulted in general turnover, while the correlation resulted in turnover in foreign country and LPG prices. From the feature importance analysis the general turnover was not used as much as the foreign turnover and LPG prices. Which leads to the conclusion that for this use case, correlation is

a better suited method compared to coupling. Next to that it is interesting to note that foreign turnover is better suited than general turnover to forecast the demand, which is supported by the data of Nijhof-Wassink that they transport many of their orders to Germany.

6.1.3 Forecasting demand

Twelve different models were created by a combination of supervised learning models and forecasting procedures to forecast demand. In order to pick the best performing model, the MAPE metric was used. The four models were random forrest (RF), support vector regression (SVR), artificial neural network (ANN) and LightGBM. The three versions of each model were to 1) not including external features as base model, 2) lagging the external features and 3) forecasting the external features using SARIMA. RandomsearchCV was used as a good heuristic for finding the hyperparameters of the model while minimizing the introduced bias. K-fold cross validation with gaps was used to create an accurate measure of the models in order to compare different settings. Random forrest was the best model including external features, while Support vector regression worked best in the base scenario. The hybrid forecast procedure was more accurate than the lagged procedure for all models. New data was introduced to further validate the models. On this new data the SVR model showed a bigger decrease in performance than the Random forrest model. To validate the model, it was again tested by doing an analysis of SARIMA and a feature analysis. SARIMA was changed to a support vector regression to forecast the external features as input for the random forrest model. This change increase the out-of-sample performance as tested with a very small dataset.

To answer the research question: *"How can Nijhof-Wassink's chemical division use historical data and external factors to better understand and forecast the demand?"*

To find external factors that are useful to understand and forecast demand, analysis of the business processes and correlation can be combined. This process of qualitative and quantitative analysis lead to better understanding of which external factors influence the demand. K-fold validation and RandomsearchCV can be used to assess the performance and optimize the Hyperparameters. The best performing model in the k-fold validation is a support vector regression. To create a forecast with external data and historical data, a hybrid model can be used. First use support vector regression to forecast the external factors, after which this is used as input for the hybrid model, together with historical data features. The forecasting model that works best with external factors is a random forrest model.

6.2 Discussion & further research

In all research projects, things do not go as planned or can certainly be improved. This is why it is good to reflect on the research design, its limitations, and how the research design was executed. Furthermore, the scientific relevance and potential avenues for further research are discussed.

6.2.1 Reflection on research design

The MPSM gave a good framework for creating a research design. Within the research design the phases of the MPSM were used from the third phase onward. Namely, the problem analysis, solution generation, solution choice, solution implementation and solution evaluation. Main questions and sub questions gave guidance to the chapters and improved the coherence of the research. The first main question focused on the current situation in the company, which is the problem analysis. The current situation had two parts, the business processes which sparked interest for the research and the analysis of the already present data within the company. This could be improved on by creating two main questions for this part, just like the literature review. The literature review also contains multiple parts, but has two main questions and multiple sub questions attached to it to reflect this. In the chapter of the literature review the solution generation phase is done. The solution choice and solution implementation is done in chapter four, for each phase a different main question is defined. In chapter 5 the solution is evaluated and improved, which is part of the solution evaluation phase, but it could be argued a new MPSM circle should have been done at that stage. Overall the MPSM was a good framework to use during this thesis, it provided clarity and was easy to use.

6.2.2 Forecasting Limitations and assumptions

In research focused on forecasting, assumptions need to be made which are limitation on the validity of the research. The biggest assumption when using historical data is that the past is representative of the future. If this turns out not to be the case, no quantitative forecast can be made. Structural breaks in data can prevent a model from working as intended. Another big assumption is that the orders of Nijhof-Wassink are a good representation of demand. Orders and market demand are of course different things, but in this thesis they are treated as if they mean the same thing. This is necessary because tracking market demand is not feasible for Nijhof-Wassink, which means there is no data to use for forecasting of demand. The supervised machine learning models used in this research give little insight how the forecast is made, which makes them harder to interpret compared to linear models.

The use of flat cross validation instead of nested cross validation introduces bias into the forecast because all the data is used to train the hyperparameters. The use of K-fold cross validation and the gaps make sure the MAPE scores achieved are a robust representation of the in-sample performance of the model. Creating an out-of-sample workflow for K-fold with nested cross validation does not produce hyperparameters as an output. The hyperparameters are needed to create a working model, thus the choice for flat cross validation was made to ensure that a forecast was possible. The out-of-sample test done in chapter 5 does not correspond with the results of the in-sample validation, which makes the results the overall results less reliable. Although a very small sample size is used, a complete turnaround in MAPE is still surprising. The out-of-sample test is not a good validation test, but it does not have the inherent bias of the flat cross validation.

In most hybrid forecasting method a simple linear model is used together with a more sophisticated machine learning model. The reason for this is to keep the benefits of both simpler linear models while using the sophisticated models to improve the non-linear relationships. In this research the linear model SARIMA was used, but SARIMA did not perform well in the k-fold. Changing the SARIMA Model to an SVR model did negate this advantage, since the SVR also uses non-linear kernels.

6.2.3 Scientific relevance and Further research

The scientific relevance is mostly as a case study. In this research supervised machine learning models are applied to a logistics forecasting problem, and external features are used to improve the forecasting of demand. The inclusion of macroeconomic variables in a forecast is not new, mostly used in the financial or energy sectors. Applying these techniques to the logistic sector is not a widely used approach.

The theory of coupling and decoupling is used successfully in the works of Profillidis and Botzoris (2019b), where it is used to conclude if Co2 emissions can be decoupled from the logistic industry. In this research this macro economic model is applied to the micro economic scale of one company. Finally this model of coupling did not work better then correlation to define which external variables are useful to forecast with this dataset.

Further academic research can be done by looking more at the methodologies of forecasting. The bias introduced by flat cross validation with hyperparameter tuning could be looked into in order to give an estimation of the difference between in-sample and out-if-sample accuracy. The benefits of randomsearchCV versus gridsearchCV in accuracy, validity and computing can be further researched, also in combination with the bias of introduced with the cross validation. Next to this, in this research Theil's inequalities are described but not used due to time constraints. Further research into the use cases of Theil's inequalities can be useful to better understand the biases in the system.

6.3 Recommendations for Nijhof-Wassink

This tool gives Nijhof-Wassink insight in the expectation of demand on a monthly level. Next to that two different ways of linking external variables to the demand of Nijhof-wassink were tested. The tool and the sight give more certainty in the demand expectation, which is part of the causes of the action problem as described in problem cluster. This solves the core problem for Nijhof-Wassink, but does not solve the action problem entirely. The uncertainty in correctly setting prices is reduced, but not solved. Nijhof-Wassink could expand on this research to solve the action problem by keeping the model up to date and adapting it

to different causes of the uncertainty.

Nijhof-Wassink can improve the presented model and improve the use case of forecasting in their business process by looking into the following areas. look at the real life performance of both the RF with SVR and SVR, to see if the in-sample validation or out-of-sample test are more valid for the Nijhof-Wassink. The RF model with SARIMA is more difficult to implement and has higher computation time, while not showing better performance. The Number of trucks feature is known for three months in advance, using the actual of this feature instead of a forecast input improves the model. Next to that look for other external features periodically to look if he correlation coefficients change.

Another area of further research is making an operational forecast on a weekly basis instead of monthly. This could improve the planning process by having an estimation of the future demand on an operational level. When the number of orders needed to be picked up from a certain location is known, an better estimation of the cost of the order can be made. The Geo location of the orders could also be used in a forecast to try to predict the number of orders from and to locations.

Chapter 7

References

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Chopra, S. and Meindl, P. (2016). *Supply Chain Management: Strategy, Planning, and Operation*. Pearson Education, 6th edition.

Heerkens, H. and van Winden, A. (2017). *Solving Managerial Problems Systematically*. Noordhoff Uitgevers. Translated into English by Jan-Willem Tjooitink.

Hyndman, R. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts, Australia, 3rd edition.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Neural Information Processing Systems*.

Khanna, R. and Awad, M. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

Lim, H., Uddin, M., Liu, Y., Chin, S.-M., and Hwang, H.-L. (2022). A Comparative Study of Machine Learning Algorithms for Industry-Specific Freight Generation Model. *Sustainability*, 14(22):1–25.

Maçaira, P. M., Tavares Thomé, A. M., Cyrino Oliveira, F. L., and Carvalho Ferrer, A. L. (2018). Time series analysis with explanatory variables: A systematic literature review. *Environmental Modelling & Software*, 107:199–209.

Pai, P.-F., Lin, K.-P., Lin, C.-S., and Chang, P.-T. (2010). Time series forecasting by a seasonal support vector regression model. *Expert Systems with Applications*, 37(6):4261–4265.

Profillidis, V. and Botzoris, G. (2019a). Chapter 1 - transport demand and factors affecting it. In Profillidis, V. and Botzoris, G., editors, *Modeling of Transport Demand*, pages 1–46. Elsevier.

Profillidis, V. and Botzoris, G. (2019b). Chapter 3 - methods of modeling transport demand. In Profillidis, V. and Botzoris, G., editors, *Modeling of Transport Demand*, pages 89–123. Elsevier.

Profillidis, V. and Botzoris, G. (2019c). Chapter 5 - statistical methods for transport demand modeling. In Profillidis, V. and Botzoris, G., editors, *Modeling of Transport Demand*, pages 163–224. Elsevier.

Profillidis, V. and Botzoris, G. (2019d). Chapter 6 - trend projection and time series methods. In Profillidis, V. and Botzoris, G., editors, *Modeling of Transport Demand*, pages 225–270. Elsevier.

Profillidis, V. and Botzoris, G. (2019e). Chapter 8 - artificial intelligence—neural network methods. In Profillidis, V. and Botzoris, G., editors, *Modeling of Transport Demand*, pages 353–382. Elsevier.

Saraiva, F. A. and Yoshizaki, H. T. Y. (2024). Logistics demand forecasting: a literature review. *Transportation Research Procedia*, 79:100–107. City Logistics 2023.

Wainer, J. and Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 182:115222.

Wu, H. and Levinson, D. (2021). The ensemble approach to forecasting: A review and synthesis. *Transportation Research Part C: Emerging Technologies*, 132:103357.

Xiao, L., Chen, S. T., Management, G. C., Tanrikulu, O., and DeWoskin, D. (2023). The m6 competition: Battle of the forecasting models.

Zhang, G. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175.

Zhang, Y., Zhu, C., and Wang, Q. (2020). Lightgbm-based model for metro passenger volume forecasting. *IET Intelligent Transport Systems*, 14:1815–1823.