National Technical University of Athens

School of Mechanical Engineering

Sector of Industrial Management & Operational research

A Machine Learning based Decision Support Tool for Supporting the Spare Parts Bulk Ordering Process in the Shipping Industry

Diploma Thesis of Fiorentia Zoi Anglou

Supervisor: Stavros Ponis

Athens, 2019

Acknowledgments

First, I would like to express my sincere gratitude to my thesis supervisor, Professor Stavros Ponis. His help and trust allowed me to gain valuable experience and knowledge and his advice contributed to the completion of this thesis.

I would also like to sincerely thank the ship management company where the case study has been carried out, for providing me with the unique opportunity to complete my thesis in a real business environment. Specifically, I would like to thank Mr. Konstantinos Petrocheilos for guiding me, Dr. Athanasios Spanos for providing me with valuable insight and help, Mrs. Nancy Saridou and Mr. Paris Perlegkas for supporting me throughout the entire process.

Lastly, I would like to express my gratitude to my family and friends that showed me unconditional support during my studies.

Summary in Greek

Η παρούσα διπλωματική εργασία εντάσσεται στην επιστημονική περιοχή της διοίκησης της εφοδιαστικής αλυσίδας και επικεντρώνει το ενδιαφέρον και τις εργασίες της στη διαδικασία της προμήθειας (purchasing) των επιχειρήσεων που δραστηριοποιούνται στον κλάδο της ναυτιλίας. Η διαδικασία προμήθειας στον κλάδο της Ναυτιλίας είναι μια πολύπλοκη διαδικασία, πολύ κοστοβόρα και με μεγάλα περιθώρια βελτίωσης. Στόχος της παρούσας διπλωματικής εργασίας είναι να δημιουργήσει ένα εργαλείο υποστήριξης των αγοραστικών αποφάσεων του τμήματος προμηθειών μιας ναυτιλιακής εταιρίας το οποίο βασίζεται σε έναν συνδυασμό αλγοριθμικών τεχνικών ανάλυσης δεδομένων και είναι σε θέση να προβλέψει τις αναγκαίες ποσότητες για την κάλυψη των αναγκών σε ανταλλακτικά για το σύνολο του στόλου της, με το ελάχιστο δυνατό κόστος και σεβόμενο τις απαιτήσεις της τελικής ζήτησης σε ορίζοντα ενός έτους.

Η εργασία χωρίζεται σε δύο μέρη. Στο πρώτο μέρος της εργασίας γίνεται μια συνοπτική περιγραφή του λειτουργικού μοντέλου του κλάδου της Ναυτιλίας με έμφαση στη διαδικασία προμήθειας η οποία και μελετάται λεπτομερώς στο πλαίσιο της παρούσης εργασίας. Στη συνέχεια ακολουθεί μια βιβλιογραφική επισκόπηση των βασικών εννοιών της ανάλυσης δεδομένων και των διαθέσιμων αλγοριθμικών τεχνικών και των χαρακτηριστικών τους. Η έρευνα αυτή, οδηγεί στον προσδιορισμό των αλγορίθμων τεχνητής νοημοσύνης που θα χρησιμοποιηθούν στο δεύτερο μέρος, στο οποίο παρουσιάζονται οι εργασίες υλοποίησης του εργαλείου υποστήριξης αποφάσεων που αποτελεί και το κύριο παραγόμενο προϊόν της εργασίας.

Το δεύτερο μέρος της εργασίας ακολουθεί τρία βασικά μεθοδολογικά βήματα υλοποίησης. Στο πρώτο προσδιορίζονται συγκεκριμένοι κωδικοί πάνω στους οποίους θα επικεντρωθεί η ανάλυση του κόστους και η πρόγνωση των αναγκαίων ποσοτήτων για την κάλυψη των αναγκών του στόλου. Χρησιμοποιούνται αλγόριθμοι ομαδοποίησης (clustering) προκειμένου να καταταχθούν οι κωδικοί με βάση μια σειρά κριτηρίων που αφορούν στοιχεία της ζήτησης τους και τελικά προσδιορίζονται αυτοί που αποτελούν το μεγαλύτερο κομμάτι του συνόλου των εξόδων για το σύνολο του στόλου.

Στη συνέχεια, για τους κωδικούς που εντοπίστηκαν στο προηγούμενο βήμα, εκπονείται ανάλυση της ζήτησης τους με στόχο τον καθορισμό των συνολικών ετήσιων αναγκών για αυτούς τους κωδικούς σε κάθε πλοίο. Πιο συγκεκριμένα, η ανάλυση περιλαμβάνει τη δημιουργία μοντέλων πρόβλεψης των ονομαστικών αναγκών κάθε πλοίου για το έτος αναφοράς και τον καθορισμό του τρόπου με τον οποίο οι αγοραστικές αποφάσεις επηρεάζουν τις τελικές ανάγκες των πλοίων. Ο απώτερος επιχειρηματικός στόχος των αναλύσεων αυτού του βήματος είναι η εκλογίκευση της διαδικασίας προμήθειας μέσα από τη δημιουργία στοχευμένων παραγγελιών σε μεγάλες ποσότητες με όσο το δυνατόν μικρότερο σφάλμα, που θα οδηγήσει σε σημαντική μείωση του κόστους προμήθειας (λιγότερα stock outs, μείωση του κόστους αποθεματοποίησης, επίτευξη καλύτερων μέσων τιμών προμήθειας ανά κατηγορία ανταλλακτικών και συνολικά).

Στο τρίτο και τελευταίο μεθοδολογικό βήμα της εργασίας, αναπτύσσεται ένα ρυθμιστικό (prescriptive) μοντέλο προκειμένου να υποστηρίξει τις βέλτιστες αποφάσεις για την αγορά ανταλλακτικών, χρησιμοποιώντας σαν βάση τα αποτελέσματα του προγνωστικού μοντέλου του προηγούμενου βήματος. Για το σκοπό αυτό δημιουργήθηκε μια σύνθετη συνάρτηση κόστους με διαφορετικές συνιστώσες, όπως το κόστος κτήσης, το διαχειριστικό κόστος, το μεταφορικό κόστος, το κόστος εργασίας (διαχείρισης) και το κόστος αποθέματος. Ως αποτέλεσμα, το μοντέλο που προκύπτει είναι σε θέση να προτείνει στα στελέχη εφοδιαστικής και προμηθειών μιας ναυτιλιακής

εταιρίας την ανάθεση εντολών προμήθειας (ανταλλακτικά ανά προμηθευτή) η οποία δίνει το χαμηλότερο δυνατό συνολικό κόστος και ταυτόχρονα ικανοποιεί τις απαιτήσεις της ζήτησης.

Summary in English

This diploma thesis is part of the scientific field of supply chain management and focuses on the purchasing-related tasks of a shipping company. Purchasing is a complex process, bears high cost and exhibits great improvement margins. The main objective of this thesis is to create a decision support tool that is based on data analytics and machine learning algorithms and can forecast the quantities needed to cover the needs of the fleet for spare parts, that concurrently minimizes the cost and respects the demand requirements.

The thesis is divided in two parts. The first part briefly describes the shipping and maritime industry focusing on the supply chain specific aspects that are examined in detail in the main body of the thesis. Furthermore, it contains the theoretical research and the literature review of the data analytics concepts such as descriptive, predictive, prescriptive and machine learning algorithms as well as an in-depth analysis of the ones that will be used in the case study. The theoretical research aims to define the machine learning algorithms that will be used in the second part which constitutes the main product of the thesis.

The second part of the thesis follows three main steps. The first step focuses on the definition of the product codes upon which predictive and prescriptive models will be applied. Clustering is used to classify the product codes based on a series of different criteria referring to demand in order to define the product codes that drive the cost of the spares for the whole fleet.

Furthermore, for the product codes determined in the previous step, analyses are performed in order to determine the total needs of the fleet for the following year for each vessel. More precisely, machine learning models are developed and forecasting of nominal needs is attempted as a function of vessel, demand characteristics and decisions regarding the source of purchase. The business reasoning behind these analyses is the rationalization of the purchasing process by placing targeted orders in high quantities with minimum error possible that will lead to significant decrease of purchasing-related costs (less stock outs, decrease of stock out cost, better average price of purchase per spare type and in total).

Lastly, in the third step, a prescriptive model is developed to support cost optimal decisions in terms of spare parts procurement using as a basis the outcome of the predictive model. For this purpose, a complex cost function is created that includes the acquisition cost, the logistic/forwarding cost, the administrative cost and the inventory cost. As a result, the model can advise the executives of the purchasing department of a shipping company the allocation of spares to vendors that minimize the total cost while respecting the level of demand.

Contents

Figures

Tables

1 Problem Statement

1.1 The shipping industry and the case company

Seaborne trade allows for the bulk transportation of raw materials and the import/export of affordable provisions and manufactured goods. Seaborne trade accounts for the carriage of approximately 80 percent of global trade and more than 70 percent of its value is carried on board and handled by seaports worldwide (UNCTAD, 2018). Over 50,000 merchant ships exist (International Chamber of Shipping, 2019) trading internationally and transporting a large variety of cargoes. Greece continues to be the largest ship owning country in terms of cargo-carrying capacity (dwt), followed by Japan, China, Germany and Singapore (UNCTAD, 2018). It should be noted that the total cargo carrying capacity for these countries accounts for around 50 percent of the globally existing dwt, as shown in the table below.

Table 1.1-1: Ownership of the world fleet, regarding ocean-going vessels of 1.000 gross tons and above (Source: **(UNCTAD, 2018)**)

Country of Ownership	DWT [thousands of tons]
Greece	330,176
Japan	223,615
China	183,094
Germany	107,119
Singapore	103,583
China	97,806
Korea	77,277
USA	68,932
Norway	59,380
Bermuda	54,252
World	1,910,012

The main types of vessels that will be discussed below are:

- Tankers, which are used for the transportation of crude oil, oil products, chemicals and gas.
- Dry bulk carriers, which are used for the transportation of several dry cargoes.
- Container ships and multipurpose ships, which are used for the transportation of general cargo.

In the shipping business, usually, each vessel is owned by a company, which is called the ship-owning company. The companies operating the vessels, not necessarily owning the assets themselves, are called ship management companies. The case company described in this Thesis is a ship management company operating worldwide. The company operates in the spot market, which means that it does not undertake long contracts but rather fixes its vessels for smaller voyages. The company operates 86 vessels (tankers, containerships, dry bulk carriers), most of them oil tankers with an average age of the vessels of the company is 9.97 years. The case company has 67 oil tankers with an average age of 10.6 years and 21 dry carriers with average age of 7.9 years.

Table 1.1-2: Vessels sizes and types operated and average age of vessels by the case company

Vessel Size	Vessel Type	Number of Vessels	Average Age [years]
CAPESIZE	Dry	5	6.1
CONTAINER	Dry	3	6.7
KAMSARMAX	Dry	4	10.0
PANAMAX DRY	Dry	1	13.0
SUPRAMAX	Dry	4	11.9
ULTRAMAX	Dry	4	4.1

AFRAMAX	Tanker	33	10.7
MR1	Tanker	7	16.7
MR2	Tanker	10	9.7
SUEZMAX	Tanker	8	9.8
VLCC	Tanker	7	6.5

For reference, the sizes of the vessels are presented below.

Table 1.1-3: Average dead weight tonnage of vessels per size and type

Vessel Size	Vessel Type	DWT [tons]
CAPESIZE	Dry	180,200
CONTAINER	Dry	49,600
KAMSARMAX	Dry	82,200
PANAMAX DRY	Dry	75,600
SUPRAMAX	Dry	56,200
ULTRAMAX	Dry	62,700
AFRAMAX	Tanker	110,900
MR1	Tanker	39,700
MR2	Tanker	49,100
SUEZMAX	Tanker	161,100
VLCC	Tanker	311,900

In 2018 the average age of oil tankers across the worldwide fleet was 29.2 years (UNCTAD, 2018) and for dry bulk carriers 42.5 years (UNCTAD, 2018)

Table 1.1-4: World fleet statistics (Source: **(UNCTAD, 2018)**)

		Years				
		0-4	**5-9**	**10-14**	**15-19**	**20+**
Oil tankers	Percentage of total ships	14.97	21.89	17.04	8.46	37.64
	Percentage of dead-weight tonnage	21.7	33.86	24.6	14.3	5.55
	Average vessel size (dwt)	78543	84016	78643	93525	8303
Dry Bulk Carriers	Percentage of total ships	27.83	41.32	12.9	8.72	9.24
	Percentage of dead-weight tonnage	29.99	43.04	12.93	7.22	6.82
	Average vessel size (dwt)	79281	76618	73750	60907	54304

The technical condition of a vessel directly affects the performance of the ship and the economic outcome it produces, as well as, the well-being of the men on board and the protection of the environment. A vessel's operating capabilities, its fuel consumption and its resale price, heavily depend on the vessel's technical condition. As a result, maintenance and repair activities are required to ensure that the vessel and its equipment strictly meet current standards for safe and efficient operation. The maintenance of a vessel is divided in two main categories, i.e. planned and unplanned maintenance. Unplanned (or corrective) maintenance is maintenance which is carried out after unexpected failure detection and is aimed at restoring an asset to a condition in which it can perform its intended function. Planned maintenance includes the overhauls and the dry-docks. The overhauls take place at specific time intervals for specific type of machinery on board the vessel during the voyage or a port call. The dry docking procedure is an extensive maintenance process that takes place at specific time intervals (longer than routine overhauling) and requires the vessel to stay at a shipyard for some time.

The maintenance costs of a vessel heavily depend on its age mainly because of two factors:

– Because of malfunctions by the aged machinery of the vessel and

- Because of the increased mandatory maintenance procedure that a vessel above 15 years needs to undertake (dry docking takes places every 2.5 years when a vessel reaches 15 years instead of 5)

1.2 Bulk ordering concepts

The main scope of the bulk orders in general is to aggregate demand and exploit the advantages of high volume. The main advantages of bulk orders are:

- Low administrative costs: Order aggregation creates economies of scale reducing the administrative cost per order ratio.
- High negotiating power: when the demand of several needs is accumulated the volume increases and the customer can press for lower prices, higher discounts and better contractual terms.
- Low logistics costs: The decisions regarding time and place of delivery can be optimized ensuring lower forwarding costs
- Uniformity: the lower number of suppliers is a step towards guarantee of a same level of quality of purchased goods across a company.
- Traceability: The uniformity of quality across the company makes it easier to identify problems and malfunctions if need arises.

Undoubtedly, the maintenance of the machinery onboard a vessel is a critical task. The scheduled and organized maintenance of a vessel can make a considerable difference on the operating expenses of the vessel. Therefore, ship management companies establish full proof and robust planned maintenance frameworks and systems, whilst taking the planned maintenance of the vessels very seriously and undertaking cross departmental projects to ensure timely delivery, high quality of spare parts with the lowest total cost of ownership possible.

The planned maintenance of each vessel is a timely task that demands overhauls at specific time intervals and/or equipment running hours. Depending on the age of the vessel and the type of machinery those needs may vary but the overhaul needs when accumulated for the whole fleet may amount to a considerable expense for the company.

The bulk ordering process in the shipping industry has the below characteristics:

- The suppliers that can provide the necessary parts in the necessary volumes for overhauling processes are a few and are concentrated in two geographic regions (Europe and Asia).
- The number of distinct items ordered each year is considerably high amounting to several thousand different spare parts
- The delivery locations are not constant and are subject to the vessel's movements.

The main idea behind the bulk orders of the case company is that the individual vessel's needs for overhauls and general planned maintenance are accumulated across the fleet and then grouped into four categories: compressors, purifiers, main engine and diesel generators, which are the main machinery components requiring overhauls every one or two years. Then, smaller groups are created based on characteristics of the underlying vessels and items are grouped into rfqs. Then the rfqs are sent to suppliers making it easier to negotiate prices and terms as the one-off revenue for suppliers increases. Finally, after the supplier selection process ends, the grouped queries are again broken down to individual vessels, the purchase orders are released and the items are then delivered to the vessels accordingly. The main steps of the bulk orders in the case company are as follows:

- **A. Requisitions from the vessels are created**: The Vessel's Chief Engineer (C/Eng) updates the quantity requirements for each needed item (VCE is a member of the vessel's crew and the person responsible for all technical issues safeguarding the smooth operation of the vessel). In his decision, C/Eng considers the Superintended Engineer's (VSE) insight and the vessel's stock. VSE shares the same responsibilities with C/Eng but is based on shore. This task generally starts in late March to early April and ends late May.
- **B. Queries creation:** The requirements of the vessels are aggregated per component and manufacturer to create the rfqs.
- **C. Price collection:** The suppliers revert with prices and after negotiations the winner is selected.
- **D. Purchase order finalization:** Purchase orders are created and the final quantities are determined by the technical department after reviewing the updated needs of the vessel. This step generally takes place in November, so it is highly probable that the vessel's needs have changed. The finalized purchase orders are then sent out to corresponding suppliers and the parts are delivered to the vessels accordingly. This process takes place in the start of the year.

The bulk order process is also presented in a form of a flow chart below:

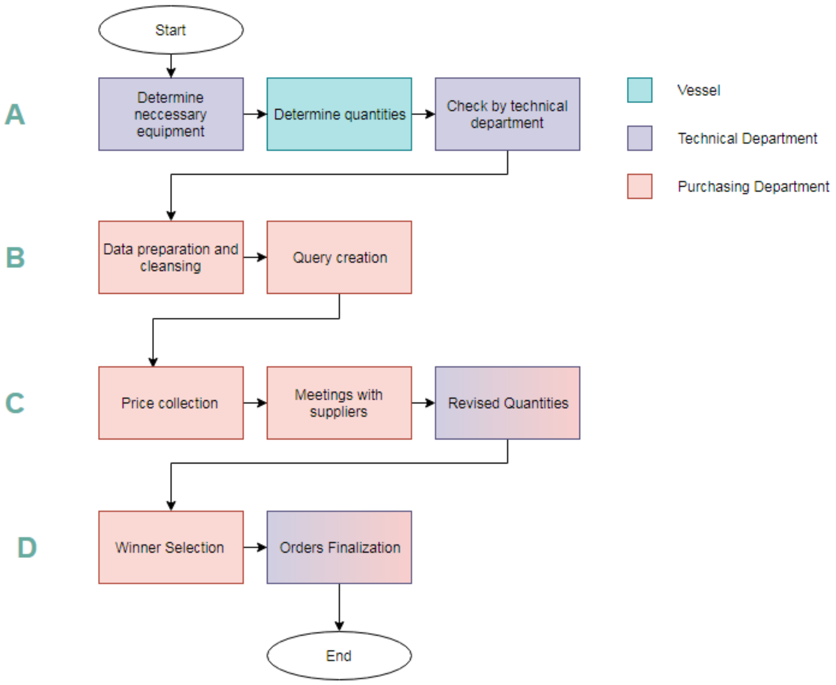


Figure 1.2-1: Bulk order process flow diagram

To provide a deeper understanding of the process an indicative time plan is presented below:

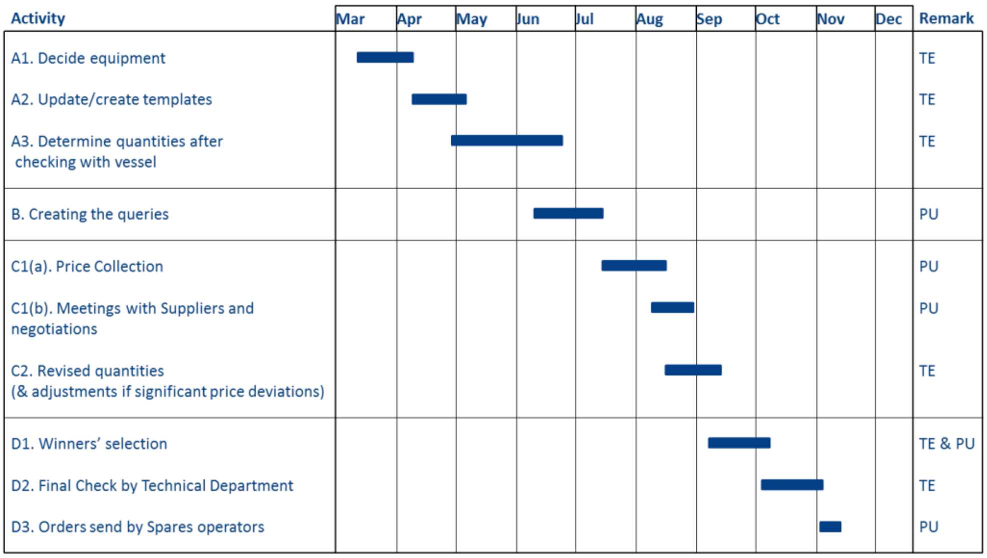


Figure 1.2-2: Indicative timeline of the bulk ordering process

In the figure above the remark section indicates the responsible department for completion of relevant activity, 'TE' for technical department and 'PU' for purchasing department.

As can one easily understand from the time plan above, the bulk order process takes a lot of time and requires the attention of several departments and individuals. The main challenges that the case company encounters in this process are the following:

- The volume: the bulk orders refer to more than 50 vessels and more than 4,000 items every year making it very time-consuming to negotiate with the implicated suppliers and conclude the selection process.
- The administrative workload: a high number of interconnected parties and stakeholders participate in the process, which makes the process very unwieldy and slow-moving. By approximation, 1.7 FTEs throughout the year are needed for the smooth completion of the process.

The above challenges have triggered the case company in scouring for ways to optimally address them and unlock further value of the bulk order process. The case company looked to machine learning due to its current strong standing and high maturity profile in deploying advanced analytics to increase effectiveness and boost efficiency in supply chain areas such as general consumables forecasting, crew scheduling and strategic network design. This gave rise to the topic of this diploma thesis which will aim to address the aforementioned key challenges by taming a very sizeable and overly complex dataset, providing ways to extract useful information and insights from historical data, facilitating the ability to forecast the needs of the fleet, reducing administrative workload and support the decision-making process by generating indicative solutions.

In the following sections, the applicability of machine learning in dealing with similar business issues will be examined (in Chapter 0) so as to formulate bulk orders analytics framework (in Chapter 3) that will enable the design of an integrated tool that aims to tackle challenges throughout the process of the bulk orders. More specifically, clustering and forecasting of the quantities needed by the vessels will take place to provide a laser focused and current view of the critical needs of the vessels by integrating exogenous factors which in their way orchestrate demand, e.g. vessel age and in the process eliminate the back and forth between the technical department and the vessel which presents the main hurdle in step (A), (see Figure 1.2-1). Furthermore, to further reduce the administrative workload and generate cost optimal scenaria in steps (C) and (D) (see Figure 1.2-1), blending of analytics with traditional operations research, i.e. prescriptive analytics, will be examined and to drive winner selection will take place on the basis of minimum total cost of ownership.

2 Theoretical research

2.1 Basic Concepts

2.1.1 Data Analytics Concepts

Most recent advances in artificial intelligence (AI) have been achieved by applying machine learning to very large data sets (Russel & Norvig, 2009)Machine learning algorithms detect patterns and learn how to make predictions and recommendations by processing data and experiences, rather than by receiving explicit programming instruction. The algorithms also adapt in response to new data and experiences to improve efficacy over time. Data analytics, that are a major part of machine learning algorithms, can be divided into three major categories with increasing complexity.

- Descriptive analytics focus on models that try to describe what happened and are deployed by most industries as they give valuable insight in the past.
- Predictive analytics use statistical models and forecasts techniques to understand the future and are used to answer the question ‘what could happen?’ Predictive analytics are employed in data-driven organizations as a key source of insight.
- Prescriptive analytics mainly employ optimization and simulation algorithms to provide recommendation on what to do to achieve specific goals.

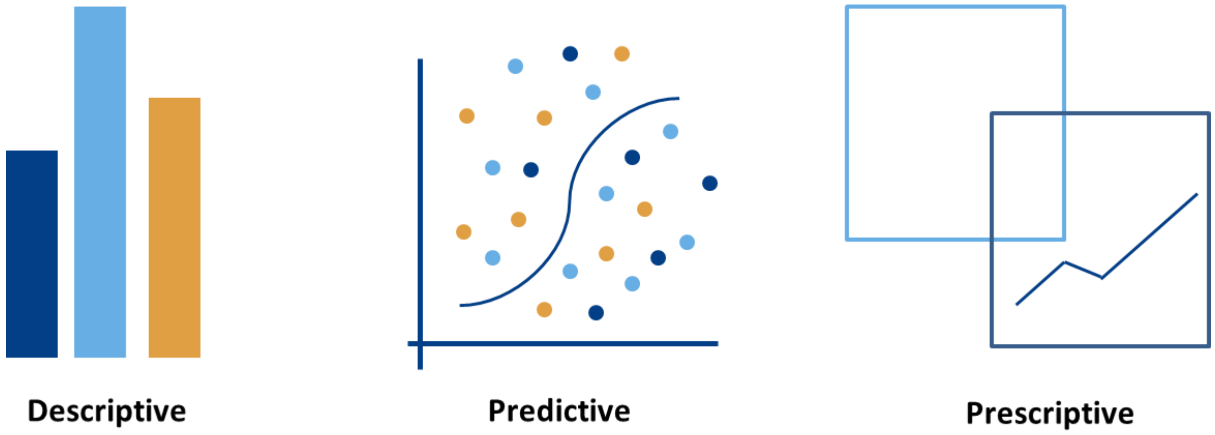


Figure 2.1-1: Types of data analytics (Source: McKinsey Analytics)

2.1.2 Machine learning Concepts

Machine learning focuses on the last two types of data analytics, predictive and prescriptive. The main idea of the research behind artificial intelligence (AI) is that ‘every aspect of learning or any other feature of intelligence can, in principle, be so precisely described that a machine can be made to simulate it’ (McCarthy, Minsky, Rochester, & Shannon, 1955)Generally artificial or computational intelligence is the study of intelligent agents, described as entities that act in an environment (Poole, Mackworth, & Goevel, 1998). As the research in the field grew, researchers defined AI as the ‘effort to make computers think’ and to create ‘machine with minds in the full and literal sense’ (Haugeland, 1985).

All in all, artificial intelligence is typically defined as the ability of a machine to perform cognitive functions we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, and even exercising creativity. Examples of technologies that enable AI to solve business problems are robotics and autonomous vehicles, computer vision, language and text processing, virtual agents, and machine learning.

Artificial intelligence can be divided into three main categories explained in detail in the next sections:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

The main goal of machine learning algorithms is to effectively build a mathematical model based on a subset of the available data, most commonly called the training set, to make predictions about the future and take actions without being explicitly instructed the way to perform the task (Bishop, 2006)

2.2 Algorithms and use cases

2.2.1 Descriptive models

Descriptive analytics are used to describe previous data and situations and extract value from them. These models make extensive use of statistical tools that can quantitatively describe and summarize features of a dataset. For each use case, the tools used are different.

In the specific use case, the bulk ordering process of a ship management company, for the time being the descriptive analytics are used mainly to compare strategies and results of each year to enable the company to negotiate better in the future.

2.2.1.1 Unsupervised learning algorithms

Unsupervised learning algorithms can be used to construct descriptive models and derive underlying relationships and give insight to past data (Hinton & Sejnowski, 1999). Unsupervised learning algorithms are machine learning algorithms that ‘learn’ from the test data that have not been labeled, classified or categorized. In contrary to the supervised algorithms where a human gives feedback, in the process of unsupervised learning the algorithm detects common elements in the data and reacts based on the presence or the absence of such common elements. Therefore, the unsupervised learning algorithms are typically used when one does not know how to classify the data and they want the algorithm to find patterns and categorize the data for them. Most common use cases of the unsupervised learning algorithms are:

- Segment employees, suppliers, customers and generally business partners into categories based on their performance
- Use clusters for behavior prediction to identify the important data necessary for making a recommendation

One of the most important aspects of unsupervised learning algorithms is clustering. Clustering is the process that includes grouping a set of objects in the same group, in a way that group’s items are more alike than others belonging in a different group. There are many data clustering algorithms present in literature, since clustering is one of the most common tasks in machine learning. Some clustering algorithms are:

- k-means: the k-means algorithm groups items into a predefined number of clusters aiming to minimize the distance of data points from the mean of each cluster. This algorithm will be explained in detail below.
- DBSCAN: dbscan is a density-based algorithm that groups items based on how closely they are packed. This algorithm will be explained in detail below.
- Kohonen neural network: a neural network algorithm that is mostly used for dimensionality reduction. The algorithm is trained using unlabeled data to produce a low-dimensional representation of the input space (Kohonen, 1982).

Another aspect of unsupervised learning is anomaly detection, the process where the data points that differ significantly from the whole of the dataset are identified and labelled as outliers. Typically, the outliers refer to problems in the smooth operation of an organization. The unsupervised anomaly detection search for outliers in an unstructured dataset using the assumption that most of the data-points can be considered normal. Therefore, the algorithms detect instances (data points) that fit the least to the remainder of data in the dataset. The algorithms that perform outlier detection can be as simple as the creation of box plots or more complex using clustering (Zimek & Filzmoser, 2018)

2.2.2 Predictive models

Predictive models make use of statistical techniques from data mining and machine learning. They are used to analyze current and historical data to make predictions about future events (Nyce, 2007). They capture the relationship between the specific performance of a unit in a sample and one or more attributes and features of the units. The objective is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. Generally, data can be divided into two major categories. The first is the training sample, or training set, that encompasses data that have known attributes and performance and upon which the models are applied and tested. The other contains data that have known attributes but unknown performance. Some indicative use cases that predictive models are broadly used:

- Customer/ Supplier relationship management: An approach that is used to manage a company’s interaction with current and potential customers. It involves the construction of a holistic view of the relationship with customers throughout the lifecycle of the relationship. It is used to predict customer’s buying habits and to promote relevant product codes.
- Project risk management: The process of managing an uncertain event that if it occurs it has either positive or negative impact on the objectives of the project.
- Demand forecasting: The process of trying to understand and predict customer demand to optimize supply decisions using supply chain and business management
- Trade promotions optimization: Tools helping companies achieve profitable growth from their trade promotions that are optimized to generate more sales and profitability.

The predictive models mainly make use of supervised learning, a sub-category of machine learning, which uses algorithms to learn a function that maps an input to an output based on input-output pairs (Russel & Norvig, 2009). The supervised algorithms use labeled training data and feedback from humans to learn the relationship of given inputs to a given output. (Mohri, Rostamizadeh, & Talwalkar, 2012)In supervised learning algorithms, each data set can be described as a pair that consists of an input object (most of the times, a vector) and an output (a value). The algorithm, then, analyses the data and produces a function that connects the input variables and the output. Once the training of the algorithm is complete and the algorithm is sufficiently accurate, it is applied on new data.

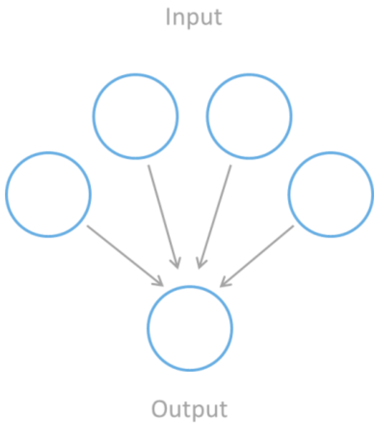


Figure 2.2-1: Supervised learning algorithms main logic (Source: McKinsey Analytics)

Supervised learning can be divided into two main categories: the classification and the regression algorithms. Classification algorithms recognize patterns form the input data and then use them to classify the new observation. On the other hand, regression algorithms forecast a continuous numeric variable using input data and reveal the underlying function.

Generally, most algorithms can work both for classification and regression problems. Depending on the specifics of the case at hand, relevant changes are made so that the algorithm can be used.

The main steps to applying a supervised learning algorithm are:

- Gathering of the training set: the training set needs to be a cross section sample that respects the requirements of the chosen algorithm and concurrently has a satisfactory number of data points relation to the variables to successfully complete the training phase.
- Treating of the gathered data: the input data of the algorithm are represented in a way that will increase the accuracy. The user needs to be cautious of the number of variables that will be used to avoid the 'curse of dimensionality' (Bellman, 1957) meaning the increase of various phenomena detrimental to the accuracy of the algorithms that arise when analyzing high dimensional spaces (e.g. overfitting, underfitting etc.).
- Deciding the algorithm that will be used to train the data and determine control parameters.
- Evaluating the accuracy of the algorithm by applying the algorithm on the test data.

Below some major algorithms are presented:

Table 2.2-1: Machine learning algorithms used in predictive analytics

Algorithm	Main Logic	Application Example
Linear regression	Linear approach to modelling the relationship between a dependent variable and one or more independent variables to predict future values of output variables.	Prediction and understanding of economic elements such as consumption spending, fixed investment spending and economic drivers such as competition, distribution, marketing initiatives, etc.

Logistic regression	A model that resembles linear regression although the outcome is a binary variable. It is now extended to include categorical outputs with more than two values.	Classification of people for business purposes based on how likely it is for them to spend a certain amount of money, to repay a loan etc. Can also be used in medicine and other scientific fields.
ARIMA model	The ARIMA (autoregressive integrated moving average) models can be applied to time series data either to better understand the data or to predict future points in the series.	Prediction of sales for the next months based on previous year’s sales to better plan production and accuracy of sales targets.
SARIMA models	An extension of the previously mentioned model, the SARIMA models can also account for the seasonality in the time series.	Prediction of retail sales to account for seasonal peaks within the year, e.g. Christmas holidays etc.
Naive Bayes	Classifier that makes use of Bayes' theorem with strong 'naïve' assumptions between the features. It allows the probability of an event to be calculated based on knowledge of factors that might affect that event.	It is mainly used to analyze sentiment to assess a product's perception in the marker Can also be used to classify several people based on measured characteristics.
Random forest	Classification or regression model that fairly improves the accuracy of a single decision tree by generating multiple decision trees and taking a majority vote of them to predict the output.	Can be used to predict the customers that will repay their debts in time, to predict a stock's behavior and whether a customer will buy a product or not.
Neural networks	Model in which artificial neurons (software-based calculators) make up an input layer, one or more hidden ones where calculations take place and an output layer.	Due to their ability to model nonlinear processes they have vast applications in system identification, medical diagnosis and decision making.
Deep learning	Based on neural networks, deep learning methods were inspired by the processing of information in biological systems.	They are broadly used for voice, text and character recognition.
Decision tree	Uses decision trees to go from observations about an item to conclusions about the item's target value. It can be either a classification tree, where the target variable can only have discreet values or regression tree if it is continuous.	A decision tree can be used to provide a defined decision framework eg. it can be used to understand product attributes that make a product more likely to be bought.
Support vector machine	It represents the examples as points in space in way that the example of the separate categories is divided by a gap that is as wide as possible. When generalized it can be used for regression.	It is widely applied in biological and other sciences. It has also been used to classify images and in text and hypertext categorization.

Boosting trees	Generates sequential decision trees where each decision tree focuses on correcting the errors coming from the previous tree model. The final output is a combination of results from all decision trees.	Forecasting of product demand and inventory levels.

As several algorithms have been developed, it is important that the best – performing algorithm for each use case is selected. In the selection process, the following aspects need to be considered:

- Bias-variance tradeoff: Errors in machine learning algorithms can be divided into two major categories: the bias error, which is the error that derives from faulty assumptions in the learning algorithm, and the variance error, which derives from sensitivity to small fluctuations in the training set (James, 2003). Thus, the first major issue that one must consider is the tradeoff between bias and variance (Geman, Bienenstock, & Doursat, 1992).The tradeoff is the conflict that tries to simultaneously minimize these two sources of error that prevent the supervised learning algorithms from generalizing beyond their training set.
- Function complexity and amount of training data: The second issue that arises is the amount of training data that will be used. As expected, the higher the complexity of the 'true' function the higher the amount of data needed to extract the relationship between the variables. However, if the function is too complex the algorithm can be prone to overfitting, meaning that the results of the algorithm respond more closely to the train set but fail to fit additional data (i.e. the test set).
- Noise in the output values: If the desired output variables exhibit high levels of noise, meaning these variables are often incorrect, then the algorithm should not attempt to fit the data. Attempting to fit misleading data could lead to overfitting or to incorrect definition of the underlying function. To avoid this issue, it is common to apply techniques that remove noisy training examples prior to the training (e.g. outlier elimination) or try to alleviate noise in the output (e.g. early stopping) (Brodley & Friedl, 1996).

It is important to consider the above when selecting the algorithm for each use case and to experiment between different algorithms to determine the best algorithm for each application (Geisser, 1993).

2.2.2.1 Reinforcement Learning

Reinforcement learning is an area of machine learning that is broadly used in prescriptive models. In reinforcement learning the algorithm learns how to perform a specific task in an environment. The algorithm receives rewards when performing correctly and penalties when performing incorrectly. Thus, the algorithm learns without any intervention from humans by trying to maximize the rewards and minimize the penalties. One of the main issues to be addressed when using reinforcement learning is the exploration – exploitation trade off. The exploration can be defined as the 'random' search of the possible solutions without searching in a specific area. This allows the algorithm to explore the solution space and not trap it-self to a local optimum. On the other hand, the premise of exploitation is searching thoroughly promising solution neighborhoods identified during the exploration phase. The main disadvantage of the exploration is that is time-consuming, and the main disadvantage of the exploitation is that the algorithm can be easily trapped to a local optimum. To overcome these shortcomings, efficient and effective neighborhood operators need to be constructed to account for the intricacies of the feasible solution space, like the "Big Valley" phenomenon, i.e. the clustering of very strong local optima around the global one, whilst not sacrificing utilization of the "Backbone", e.g. the commonly shared orientations amongst promising permutations.

2.2.3 Prescriptive analytics

Prescriptive analytics is a recently introduced concept of data analytics as it encompasses techniques and results from descriptive and predictive analytics. Prescriptive analytics uses optimization methods to identify the best alternatives to minimize or maximize some objective (Evans & Linder, 2012) Prescriptive analytics suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and show the implications of each decision option. The algorithms may consider new data that become available to re-assess the decisions and the risks. Generally, prescriptive analytics models consider both structured data (numbers, labelled data) and unstructured data and business rules.

One of the simplest methods used to search for the optimum solution in a problem is the brute force search or exhaustive enumeration. Brute-force search is a general problem-solving technique that consists of systematic enumeration of all possible candidates for the solution and checking the performance of each candidate on the objective function and on the constraints of the problem. Its application though is dependent on the underlying optimization problem, complexity and size of solution space. Should the former be NP-Hard and the latter expansive, usage of brute force might be prohibited due to the large computational overhead generated. In this case, alternate approximation approaches might be considered like generalized metaheuristics, e.g. greedy adaptive randomized search, particle swarm optimization, and/ or heuristics encompassing problem specific knowledge in hybrid algorithmic frameworks.

2.3 In-depth analysis of machine learning algorithms to be used

2.3.1 Unsupervised learning – Clustering Algorithms

Clustering is a major part of unsupervised machine learning. As previously discussed (in 2.2.1.1), clustering makes use of historical data to create classes (often also called groups) based on certain criteria. Clustering is performed when it is believed that the data have undisclosed relationships with one another that can be unveiled with the underlying cluster labels. The underlying clustering labels and classes may help uncover useful information about the data and the groupings can be made based on several dimensions in a structured way that will help choose better actions for each group.

2.3.1.1 k-means

K-means is an algorithm that organizes data into groups (k) that each contains data with similar characteristics. K-means groups n observations into k clusters based on the distance between observations of each cluster from a centroid. The algorithm tries to minimize this distance between observations from same cluster.

More specifically, it puts N data points of an I-dimensional space into K clusters. Each cluster can be parameterized by a vector $\mathrm{m_k}$ called its mean. The data points are denoted by vector $\mathrm{x^{(n)}}$ where n runs from 1 to N (where N is the number of data points). The vector x has I components $\mathrm{x_i}$. We can compute the distance as:

$$\mathrm{d\,(x,m)} = \sum_{\mathrm{i=1}}^{\mathrm{i=I}} (\mathrm{x_i} - \mathrm{m_i})^2$$

The first step of k-means is the initialization of the centroids (the means). This can be performed using a number of different methods which will be discussed further below. Then, the two main steps of the algorithm are performed in iteration. The first step is called the assignment step and the second the update step.

In the assignment step, each data point is assigned to the nearest mean. The guess for the cluster $\mathrm{k}^{(\mathrm{n})}$ that the point $\mathrm{x}^{(\mathrm{n})}$ belongs is denoted by $\mathrm{k}^{\wedge\,(\mathrm{n})}$:

$$\mathrm{k}^{\wedge\,(\mathrm{n})} = \mathrm{argmin}_{\mathrm{k}}\{\mathrm{d}(\mathrm{m}^{\mathrm{k}}, \mathrm{x}^{\mathrm{n}})\,\}$$

, where argmin is the function that attains the k for which the distance as defined above is the minimum.

Then, $\mathrm{r}_{\mathrm{k}}^{(\mathrm{n})}$ is set to 1 if m_{k} is the closest mean to data point $\mathrm{x}^{(\mathrm{n})}$, otherwise it is set to 0.

$$\mathrm{r}_{\mathrm{k}}^{(\mathrm{n})} = \begin{cases} 1, & \text{if } \mathrm{k}^{\wedge\,(\mathrm{n})} = \mathrm{k} \\ 0, & \text{if } \mathrm{k}^{\wedge\,(\mathrm{n})} \neq \mathrm{k} \end{cases}$$

To summarize the two steps are presented below:

- In the first step, the algorithm computes the distance between the mean and each cluster. The k cluster for which the mean has the minimum distance, is the cluster that the data point will be part of.
- In the second step the $\mathrm{r}_{\mathrm{k}}^{(\mathrm{n})}$ takes the value 1 if the data point belongs in cluster k and 0 otherwise.

In the update step the means (m_{k}) that have been initialized (in the first iteration) or have been computed in the previous iteration are updated.

$$\mathrm{m}_{\mathrm{k}} = \frac{\sum_{\mathrm{n}} \mathrm{r}_{\mathrm{k}}^{(\mathrm{n})} \mathrm{x}^{\mathrm{n}}}{\mathrm{R}^{(\mathrm{k})}}$$

where,

$$\mathrm{R}^{(\mathrm{k})} = \sum_{\mathrm{n}} \mathrm{r}_{\mathrm{k}}^{(\mathrm{n})}$$

The steps are repeated until the assignments do not change. Alternatively, it can be said that the algorithm stops when the means, the parameters of the model, stabilize.

There are two cases that k-means can’t handle. The first refers to the distances between a data point and two (or more) centers of clusters. If said distances are equal, then the algorithm cannot decide where to assign the data point. However, this is easily solved by assigning the data point to the smallest k. The second case occurs when a cluster has no data points assigned to it. If this is the case, then $\mathrm{R}^{(\mathrm{k})} = 0$ and m_{k} cannot be updated. If this happens then no changes are required to be made to m_{k}.The initialization of the algorithm influences the clustering result as sometimes the k-means is trapped in a local optimum. Furthermore, the initialization also affects the total number of iterations of the algorithm and therefore the complexity of the problem.

The first method for the initialization is the random method. According to this data points are assigned randomly to clusters and then the mean is calculated.

The second is the Forgy method. This is one of the most commonly used methods where k data points are randomly chosen and are used as initial means. (Hamerly & Elkan, 2002)

The Forgy method tends to spread centers out in the data, while the Random Partition method tends to place the centers in a small area near the middle of the dataset. Random Partition was found to be a preferable initialization method for its simplicity (Pena, Lozano, & Larranaga, 1999)However, for standard k-means algorithms the Forgy method of initialization is preferable (Zhang, 2003). The algorithm does not guarantee convergence to the global optimum. The result may depend on the initial clusters. As the algorithm is usually fast, it is common to run it multiple times with different starting conditions.

One of the major characteristics of k-means is the fact that it uses the Euclidian distance as a metric and variance as a measure of cluster scatter. Another main characteristic is that the number of output clusters is a pre-defined parameter by the user making the algorithm subject to the user's perception of the dataset. Poor choice of the parameter k (number of clusters) may yield poor results.

Another key limitation of k-means is the cluster model is the main concept of the algorithm. The main concept is based on spherical clusters which may fail to uncover underlying relationships between the data.

2.3.1.2 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm that given a set of points in some space, classifies in the same category points that are closely packed together. At the same time, it marks as outliers points that lie in low-density regions. The key idea is that for each point of a cluster, the neighborhood of a given radius must contain at least a minimum number of points, i.e. the density in the neighborhood must exceed some threshold. As k-means, DBSCAN usually uses the Euclidean distance to measure how close data points are from each other (Ester, Kriegel , Sander, & Xu, 1996). Before applying DBSCAN, there two parameters that need to be defined:

- eps: its value specifies the minimum distance between two points for them to be considered neighbors. If eps value is too small, a large part of the data will not be clustered. On the other hand, if the value is too high then most of the data points will be put in the same cluster.
- minPts: which is the parameter that specifies how many neighbors a point should have to be included into a cluster. Generally, the higher the value of minPts the more significant the clusters that will be created will be

The first parameter that must be defined is minPts. (Ester, Kriegel , Sander, & Xu, 1996) who wrote the first paper on DBSCAN suggest to setting minPts to 4, for two-dimensional data but in a next paper (Sander, Ester, Kriegel, & Xu, 1998) it is suggested that the minPts is set to twice the dataset dimension. Generally, minPts needs to satisfy the relationship

$$\mathrm{minPts} \geq \mathrm{D} + 1$$

, where D is the number of dimensions of the problem. For datasets that have a lot of noise, that are very large, that are high dimensional, or that have many duplicates it may improve results to increase minPts. (Schubert et al, 2017).

The value of eps is usually calculated using the k-distance graph, plotting the distance to the k=minPts-1 nearest neighbor sorted from largest to smallest value. The value of eps can then be decided based on the point that the graph shows an elbow. (Schubert, Sander, Ester, & Kriegel, 2017) (Sander, Ester, Kriegel, & Xu, 1998) (Ester, Kriegel , Sander, & Xu, 1996). For the purposes of the algorithm data points are classified as core points, border points or outliers:

- A data point is considered a core point if at least minPts are within distance eps of it (including the original point it self)
- A data point is considered a border point when it is within distance eps from a core point but is not a core point itself (therefore it does not meet the minPts criterion)
- A data point is considered a noise point if it does not belong in any of the aforementioned categories. Those points represent outliers in the data set that do not belong to any cluster

Two points are considered 'directly density-reachable' if one of the points is a core point and the other point is within its eps radius. If we considered three data points denoted as p, m, q and p is directly density reachable from m, which is directly density-reachable from q. The set of points within the eps radius of p -> m -> q form one cluster.

The algorithm chooses a point p arbitrarily. Then, it retrieves all points directly density reachable from p with respect to the minimum distance eps. If p is a core point, then a cluster is formed. Then, it recursively finds all its density connected points and assign them to the same cluster as p. If p is not a core point, then the algorithm iterates through the remaining unvisited points in the dataset. The process is terminated when the algorithm has gone through all the points.

More explicitly, the algorithm begins by picking an arbitrary point from the data set. If there are more than minPts data points within distance eps from that point (including itself), therefore if the data point is a core point, a cluster is formed. Then the algorithm checks all the points that were included in the cluster to determine if they too have more than minPts points within a distance eps. If they do the cluster grows and this process continues. If the above constraint is not satisfied, then the algorithm starts the process again by choosing randomly another data point that has not yet been assigned to a cluster. If the data point chosen happens to be a noise point, then the algorithm picks a new point. The main characteristics of DBSCAN are summarized below:

- The algorithm optimizes the number of clusters without using feedback from the human thus increasing its efficiency. The user does not need to perform a sensitivity analysis as per the number of clusters.
- DBSCAN can find any shape of cluster, as opposed to the k-means algorithm that finds only circle-shaped clusters.
- The algorithm self-adjusts for outlying data points.
- It is not entirely deterministic, meaning that border points that are reachable from more than one clusters can be part of any of those clusters
- It cannot cluster datasets with large difference in densities, since the minPts-eps combination cannot be appropriately chosen for all clusters (Kriegel, Kroger, Sander, & Zimer, 2011)

2.3.2 Supervised learning – Forecasting Algorithms

2.3.2.1 Random Forest

Random forest is a supervised learning algorithm: meaning that the input and output variables are pre-defined by the user and is commonly used in machine learning.

To better understand the algorithm at hand, the building block of random forest, the decision tree will be explained below. Decision tree learning uses decision trees to go from observations about an item to forecast about an item's target value. The main advantages of the decision trees are:

- Decision trees can handle both categorical and numerical data (Gareth, Witten, Hastie, & Tibshirani, 2015)
- Decision trees do not require advanced data handling. Many algorithms in machine learning, also k-means and dbscan described above, require data normalization and indirect creation of dummy variables. However, decision trees do not require such actions.
- The decision trees use a white box model in contrast with other machine learning algorithms, such as neural networks that make use of black box models.

However, decision trees exhibit certain limitations as well:

- Decision trees can be very robust and small changes in the training set could result in large changes in the outcome of the tree (Gareth, Witten, Hastie, & Tibshirani, 2015)
- Decision trees are prone to overfitting (Bramer, 2007)This happens when an over-complex tree is created that cannot generalize well from the training data. (Hothorn, Hornik, & Zeileis, 2006)
- Lastly, decision trees that have more categorical variables with different number of levels may be biased towards attributes with more levels (Deng, Runger, & Tuv, 2011).This can be easily avoided by a two-stage approach (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013)

Random decision forests were firstly introduced as an attractive method for classification due to their high execution speed (Ho T, 1995)In random forest tree predictors are combined so that each tree depends on the values of a random vector sampled independently (Breiman , 2001)To overcome limitations on accuracy exhibited on single trees, several decision trees in different subspaces are combined to form a forest thus increasing the validity of the results (Tin Kam Ho, 1995). In other words, random forests are a way of averaging multiple deep single decision trees that may be in risk of overfitting and have been trained on different parts of the training set to reduce error metrics (Hastie, Tibshirani, & Friedman, 2008).

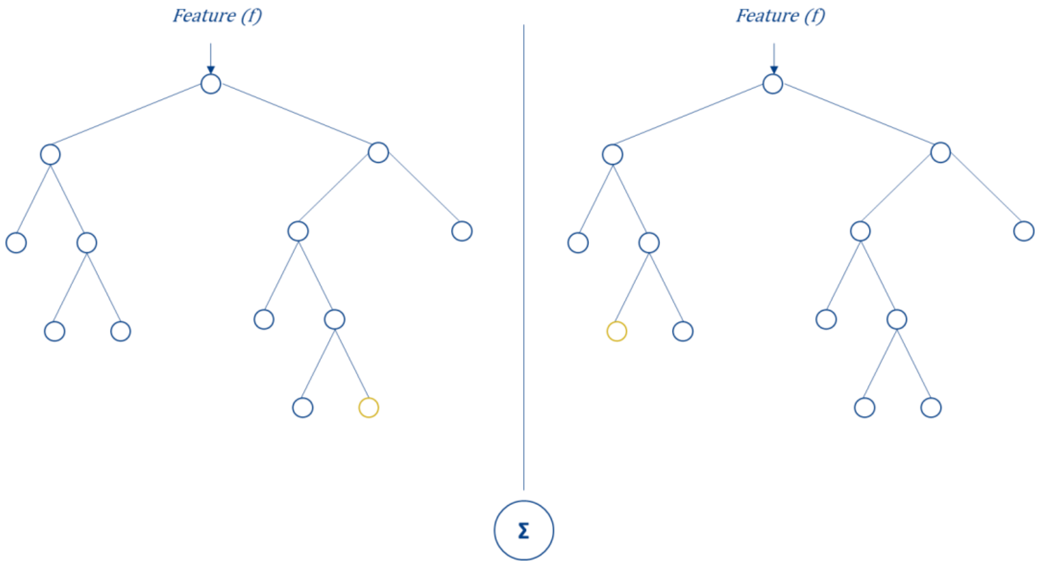


Figure 2.3-1: Random forest with two trees visualized

The random forest algorithm makes use of the general technique of bootstrap aggregating (bagging) to combine the results of the single trees (Breiman , 2001).

Given a training set $X = x_1, x_2, \ldots\ldots, x_n$ with responses $Y = y_1, y_2, \ldots\ldots\ldots, y_n$ bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, \ldots\ldots, B$

- Sample with replacement n training examples from X , Y (denoted as X_b, Y_b)
- Train regression (or classification) tree f_b on X_b, Y_b

After training, predictions for x' can be made by averaging all the predictions from previously trained trees f_b using

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Therefore, random forest is one of the algorithms that will be tested in the next sections mainly because of its ability to avoid overfitting (Hastie et al, 2008) and its superior efficiency (Tin Kam Ho, 1995).

2.3.2.2 Generalized Linear Model

The generalized linear model is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution (Nelder & Wedderburn, 1972)

In a general linear model the dependent variable y_i, *i* = 1,….,n is modelled by a linear function of explanatory variables x_j, *j*=1,….,p plus an error term as follows:

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

In a simple linear model, the independent variable is only one variable.

A generalized linear model is made up of a linear predictor

$$n_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

and two functions

- A link function that describes how the mean $E(Yi) = \mu_i$, depends on the linear predictor $g(\mu_i)=n_i$
- A variance function that describes how the variance $var(Yi)$ depends on the mean $var(Yi)=\varphi V(\mu)$ where the dispersion parameter ɸ is a constant

One of the advantages of the generalized linear model is that it can account both for categorical and numerical variables. At the same time, the generalized linear model for regression in this case is superior to the simple linear regression as it does not pose limitations on the error distributions of the variables (Nelder & Wedderburn, 1972).

2.3.2.3 Principal Component Regression

In statistics, principal component regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA). Typically, it considers regressing the outcome (also known as the response or the dependent variable) on a set of covariates (also known as predictors, or explanatory variables, or independent variables) based on a standard linear regression model but uses PCA for estimating the unknown regression coefficients in the model. **(Bair , Hastie, & Debashis , 2005).**

The main structure of principal component regression can be divided into three main steps:

- Perform principal component analysis on the independent variables using statistical methods
- Perform regression on the vector of results of the previous step using simple regression methods such as the ordinary least squares method
- Transform the vector back to the scale of actual covariates to get the final principal component regression estimator.

The main concept of the principal component analysis (which is the stepping stone of the PCR algorithm), is to make use of statistical procedures to convert a set correlated variables into a set of linearly uncorrelated variables which are called principal components. Principal component analysis firstly introduced by Pearson in 1901, (Pearson , 1901) makes use of the principal axis theorem in mechanics and creates a vector of uncorrelated orthogonal basis set.

The two main advantages of the PCR are:

- The algorithm can be performed when the number of variables is high in relation to the number of available data. (Jackson, 1991)
- PCR can perform regression even when the explanatory variables are highly correlated to each other.

The two advantages above make the principal component regression one of the algorithms that will be used as the dataset has correlated independent variables and a large` number of data is not available. These are the two main reasons that this algorithm will be used in the forecasting of the demand.

2.3.3 Prescriptive model – Brute Force Analysis

In computer science, brute-force search or exhaustive search is a very general problem-solving technique and algorithmic paradigm that consists of systematically enumerating all possible candidates for the solution, checking whether each candidate satisfies the problem's statement and assessing their performance.

While a brute-force search is simple to implement, and will always find a solution if it exists, its cost is proportional to the number of candidate solutions – which in many practical problems tends to grow very quickly as the size of the problem increases (combinatorial explosion) (Coursera, 2018). Therefore, brute-force search is typically used when the problem size is limited, or when there are problem-specific heuristics that can be used to reduce the set of candidate solutions to a manageable size. The method is also used when the simplicity of implementation is more important than speed. At the same time there is merit in exploring the performance of brute force analysis as the computing power of the IT systems has increased considerably in the last year therefore, decreasing significantly the computational overhead of the implementation of such methods.

The most efficient way to speed up a brute force algorithm is to reduce the search space efficiently by applying business or other rules. This analysis may reduce the candidates to the set of all valid solutions; thus yielding an algorithm that directly enumerates all the desired solutions without wasting time with tests and the generation of invalid candidates.

2.4 Conclusions

Having discussed in detail the applicable machine learning algorithms and advanced analytics concepts that can be applied in this project, the bulk orders analytics framework is formulated. The bulk orders analytics framework encompasses the algorithms and computation steps that will lead to the completion of the three methodological steps of the practical part of the thesis:

- Identification of high interest items: in this methodological step of the thesis clustering will be performed so as to decrease the size of the dataset. This step encompasses the concepts of descriptive analytics and relevant algorithms discussed (see section: 2.3.1) will be applied.
- Forecasting of demand: in this methodological step of the thesis machine learning algorithms will be applied so as to decrease the amount of administrative workload. This step encompasses the concepts of predictive analytics and relevant algorithms discussed (see section: 2.3.2) will be applied.
- Allocation of items: in this methodological step of the thesis a prescriptive model will be created so as to minimize total cost incurred throughout the bulk orders process. This step encompasses the concepts of predictive analytics and relevant methodologies discussed (see section: 2.3.3) will be applied.

3 Bulk Orders Analytics Framework

3.1 Overview

Below the bulk orders analytics framework is presented. This framework aims to tackle challenges as presented in section 1.2, with the data analytics concepts thoroughly discussed in chapter 0. As described in the summary the main scope of the analytics section of the thesis is to provide a comprehensive decision support tool for market decisions for the use of the case company in the bulk ordering process. The main steps of the bulk orders are presented below (see also: Figure 1.2-1)

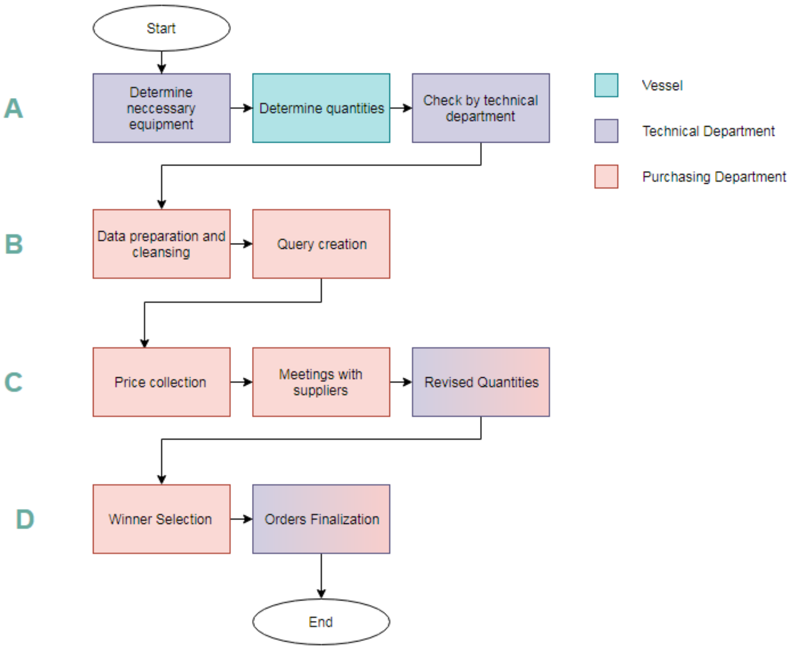


Figure 3.1-1: Bulk order process flow diagram [duplicate of Figure 1.2 1)]

The main challenges ,(see also section 1.2), of the bulk orders are focused on the large volume of the ordered items and the continuous back and forth between the vessel and the departments that generates increased amount of administrative workload. Therefore, the practical part of this thesis aims to tackle those challenges by making use of data analytics and machine learning by decreasing the number of items, by providing forecasted total needs of each vessel for each spare part for the next year and by allocating each spare part to vendors. More precisely:

- The decrease of total items aims to facilitate steps C and D where the price collection, the negotiations and the winner selection will focus only on a fraction on items and thus completion time for these steps will be smaller.
- The forecast of the total needs of the vessel's aims to facilitate step A by providing an insight to the technical department about the actual vessel's needs and how they are shaped depending on the decisions regarding the source of purchase
- The cost-based allocation aims to facilitate step D. 'Winner Selection' and rationalize the whole process by providing indicative allocation of spares to vendor category (i.e. maker or non-maker) on the basis on minimum cost incurrence.

To achieve the above aims of the project an extensive bulk order analytics framework is created which is presented in the following figure:

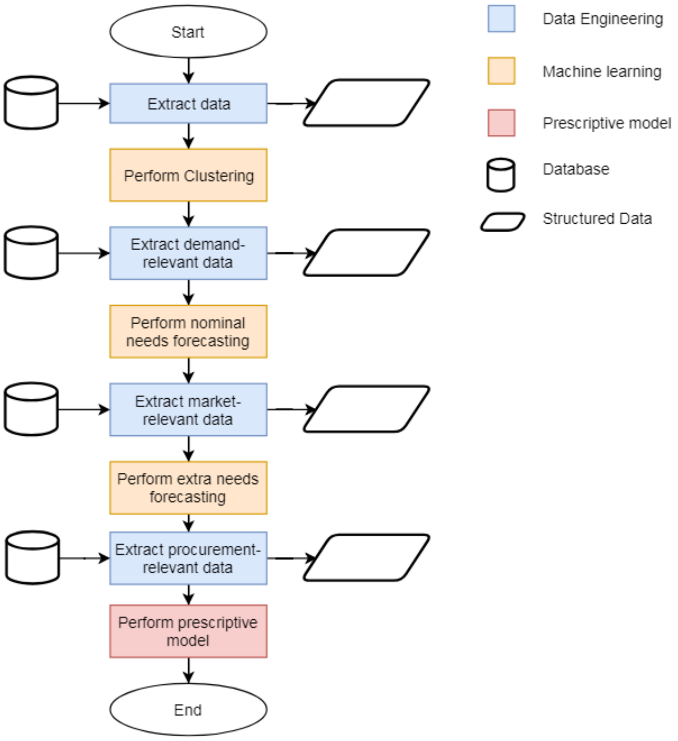

Start
Extract data
Perform Clustering
Extract demand-relevant data
Perform nominal needs forecasting
Extract market-relevant data
Perform extra needs forecasting
Extract procurement-relevant data
Perform prescriptive model
End
Data Engineering
Machine learning
Prescriptive model
Database
Structured Data

Figure 3.1-2: Overview of bulk orders analytics framework

The practical part of the thesis is divided in three main methodological steps, described below:

- Identification of high-interest items: where from all the items of the bulk orders the ones that with certain criteria can be classified as high-interest are identified.
- Forecasting of demand: where the nominal needs of the fleet based on vessel characteristics and the extra needs of each vessel based on market- related decisions are defined
- Prescriptive model: where a complex cost function is created to determine the optimum allocation of vendor to items so as to minimize total cost, while respecting demand requirements.

Detailed flow diagrams for each methodological step have been created and are presented below:

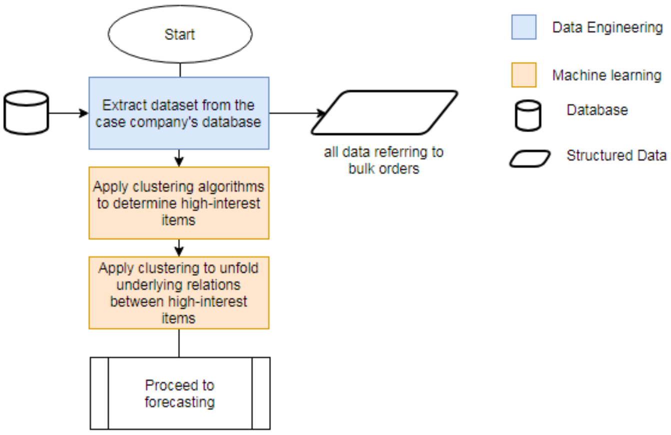

Start
Extract dataset from the case company's database
all data referring to bulk orders
Apply clustering algorithms to determine high-interest items
Apply clustering to unfold underlying relations between high-interest items
Proceed to forecasting
Data Engineering
Machine learning
Database
Structured Data

Figure 3.1-3: Flow diagram for the first part of the thesis

The first methodological step (presented above in Figure 3.1-3) focuses on the identification of the high interest items and aims to reduce the total administrative cost.

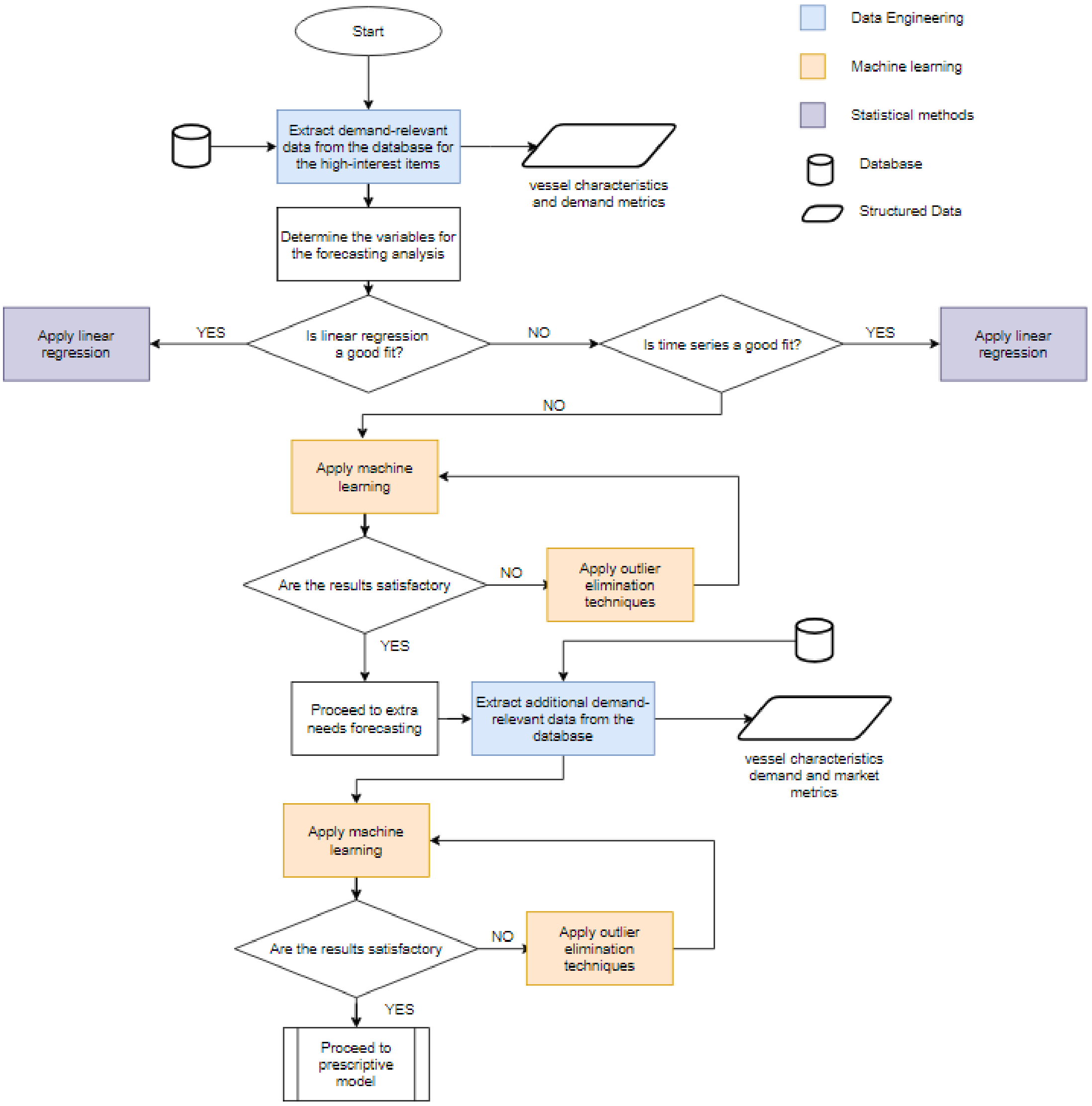
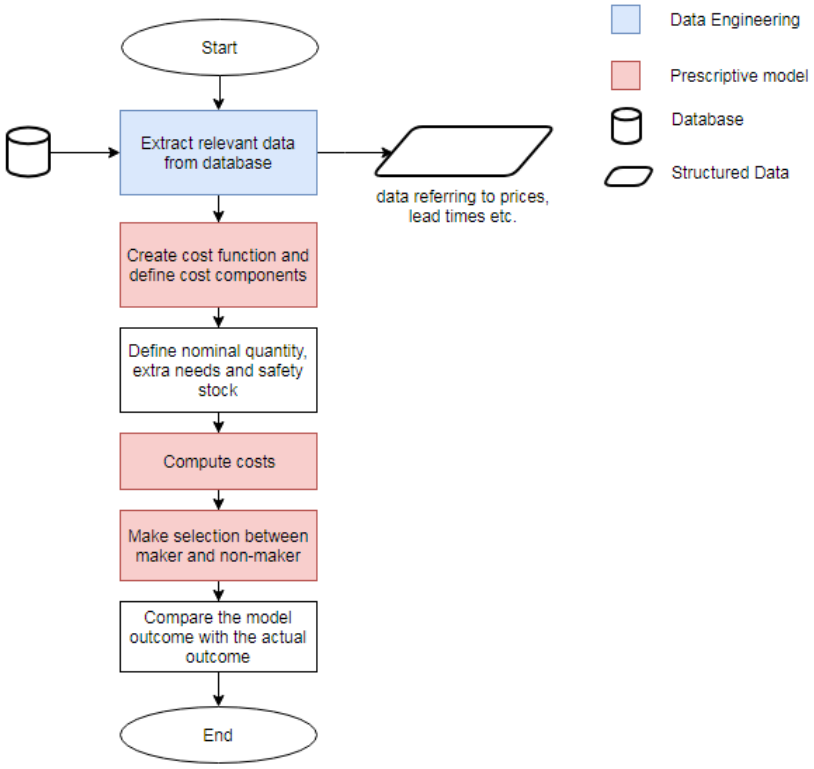

Start
Extract demand-relevant data from the database for the high-interest items
vessel characteristics and demand metrics
Determine the variables for the forecasting analysis
Is linear regression a good fit?
YES
Apply linear regression
NO
Is time series a good fit?
YES
Apply linear regression
NO
Apply machine learning
Are the results satisfactory
NO
Apply outlier elimination techniques
YES
Proceed to extra needs forecasting
Extract additional demand-relevant data from the database
vessel characteristics demand and market metrics
Apply machine learning
Are the results satisfactory
NO
Apply outlier elimination techniques
YES
Proceed to prescriptive model
Data Engineering
Machine learning
Statistical methods
Database
Structured Data

Figure 3.1-4: Flow diagram for the second part of the thesis

The second methodological step (presented above in Figure 3.1-4) focuses on the prediction of the nominal and additional needs of the vessels.



Start
Extract relevant data from database
data referring to prices, lead times etc.
Create cost function and define cost components
Define nominal quantity, extra needs and safety stock
Compute costs
Make selection between maker and non-maker
Compare the model outcome with the actual outcome
End
Data Engineering
Prescriptive model
Database
Structured Data

Figure 3.1-5: Flow diagram for the third part of the thesis

Lastly, the final methodological step includes the creation of the complex cost function.

In the next sections each of the above sub-problems are discussed in detail.

3.2 Data Engineering

A large part of the thesis is focused on the data engineering part of the problem. As all the aspects of the practical part of the thesis entail sub-problems where historical data need to be used either to extract specific values or whole datasets to train models.

The data engineering part of the thesis was completed using the case company's data warehouse. The data warehouse is a structured database that consolidates raw data from disparate sources and heterogeneous systems within the company, in a hub-and-spoke architecture. More specifically, it houses information stemming from the below integrated systems:

- the SEASOFT, which houses information regarding the position of the vessel
- the SAP ERP, which houses information regarding invoice checking and payments
- the AMOS PMS (Planned Maintenance System), which houses the basic vessel information and detailed relationship between vessel's components while also provides spares, supplies and lubricants procurement support and entails all relevant information
- the dedicated SAP Forwarding Tool, which houses information regarding supply chain related costs as well as information regarding the stock in the company's warehouses.

A large number of data in the database, which was used for the analysis, are produced from the planned maintenance system, AMOS, which also supports the entire procure-to-pay lifecycle of the spare parts demand: from the requisition part (where the vessel raises the need) to the delivery part (where the purchased goods are sent to the vessel). This cycle also entails the quotations phase, where several suppliers quote prices and lead times for specific parts as well as the procurement order phase where, after vendor selection, the official purchase order is sent to the selected vendor and the parts are purchased.

For the case study in question, each year more than 16,000 spare parts orders are made referring to more than 30,000 different maker references (which is a unique code referring to one item). In the bulk order process, which is the scope of the thesis, more than 4,000 items are procured grouped in 600 orders.

The variables that are of interest to the case study amount to more than 30 including historical prices, lead times, vessel characteristics, historical data about dry-docking etc.

To create the datasets extensive use of SQL was made creating more than 35 temporary tables to create 4 final datasets. Below the dimensions of the initial tables:

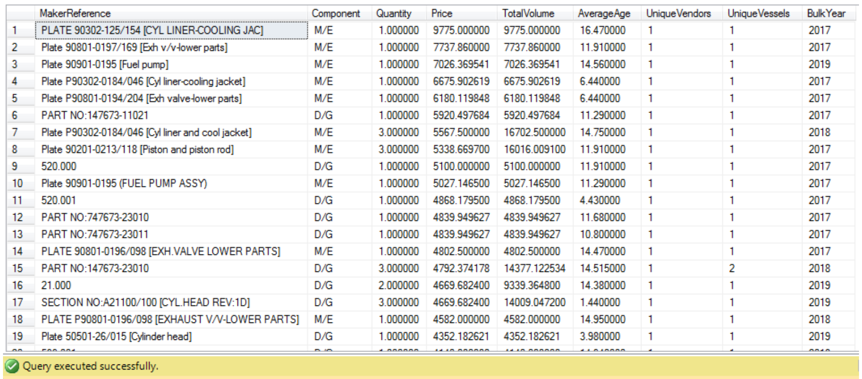
- First dataset: including all bulk orders items for all the years: 13,967 rows x 6 columns
- Second dataset: including all vessel characteristics and demand metrics for forecasting of nominal needs of the fleet: 3,625 rows x 18 columns
- Third dataset: including all vessel characteristics, demand metrics and market details for forecasting of extra needs of the fleet: 4,508 rows x 20 columns
- Fourth dataset: including important variables from second dataset and historical data about demand, prices, weight and lead times used for prescriptive model: 3,625 rows x 18 columns

Below a small part of the SQL code used:

```
USE [ADW_Analytics]
GO
/****** Object:  StoredProcedure [pu].[uspT_MakerVSnonMaker]    Script Date: 21/06/2019
16:02:07 ******/
SET ANSI_NULLS ON GO
SET QUOTED_IDENTIFIER ON GO
ALTER procedure [pu].[uspT_MakerVSnonMaker] AS
SET NOCOUNT ON;
IF OBJECT_ID('tempdb..#itemsdetailsinitial') IS NOT NULL DROP TABLE #itemsdetailsinitial
IF OBJECT_ID('tempdb..#final') IS NOT NULL DROP TABLE #final

.......... [more lines that have not been included]

--drop table #final
select distinct f.component, , f.Vessel_Code, f.MakerReference, f.BulkYear
, sum(f.Avg_pv)/ count (f.Vessel_Code) as Quantity, avg (f.VesselAge) as Avg_Age, avg
(f.AvgPrice) as Avg_Price
        , f.Market, f1.VesselSize_Desc, f1.BoughtStatus, f1.NationalityCountry_Code,
v.VesselSegment_Desc
        , v.VesselType_Desc as VslType, dwt.Avg_DWT, f.NextYearDD,v.CargoOilType_Desc,
v.ClassificationClass_Desc
        , year (v.Bought_DateTime) as YearBought, v.BuildYear, v.HullBreadthMolded,
v.HullTonnageGT
        , v.HullSpeedAbsMax, v.InsuredValue, f.Level3_Desc
        , case when e.Avg_pv_NB is null then 0 else e.Avg_pv_NB    end as 'ExtraNB'
        , casewhen (ga.GeographicArea_Desc is null and v.Yard_Name like '%korea%') OR
Yard_Name like '%hyundai%' OR Yard_Name like '%samho%' OR Yard_Name like '%DAEWOO%' or
v.Vessel_Code = 'h8' then 'SOUTH KOREA'
                      when v.Yard_Name like '%imabari%' then 'JAPAN'
                      when v.Yard_Name like '%hudong%' then 'CHINA'
                      else GA.GeographicArea_Desc
          end as 'VslOrigin'
into #final
from #finalv2 f
```

```
left join #finalv2 f1 on f.Component = f1.Component and f.Level3_Desc = f1.Level3_Desc and
f.BulkYear = f1.BulkYear and f.Vessel_Code = f1.Vessel_Code
left join #extraorders e on f.Component = e.Component and f.Level3_Desc = e.Level3_Desc and
f.BulkYear = e.OrderYear and f.Vessel_Code = e.Vessel_Code
left join adw.team.Vessels v on v.Vessel_Code = f.Vessel_Code
left join adw.team.VesselsNextDryDocking dd on dd.Vessel_Code = f.Vessel_Code
left join adw.team.GeographicAreas ga on ga.GeographicAreas_Key =
v.YardCountry_GeographicAreas_FK
left join #avgDWT dwt on dwt.VesselSize_Desc = f.VesselSize_Desc
where f1.Maker_Code is not null
group by f.Component, f.Level3_Desc, f.BulkYear, f.Vessel_Code      , f1.Maker_Code,
f1.Maker_Name, f1.VesselSize_Desc, f1.BoughtStatus
          , f1.NationalityCountry_Code, f.VesselAge, v.VesselSegment_Desc,
v.Fleet_ShortDesc, v.CF_DWT_Max, f.NextYearDD, ga.GeographicArea_Desc
          , v.Yard_Name, v.Vessel_Code, v.VesselType_Desc, v.CargoOilType_Desc,
v.ClassificationClass_Desc, year (v.Bought_DateTime)
          , v.BuildYear, v.HullBreadthMolded, v.HullTonnageGT, v.HullSpeedAbsMax,
v.InsuredValue, e.Avg_pv_NB, f.Level3_Desc, dwt.Avg_DWT
          ,f.MakerReference,f.Market

.......... [more lines that have not been included]

GO
```

Additionally, some snapshots of the extracted data:

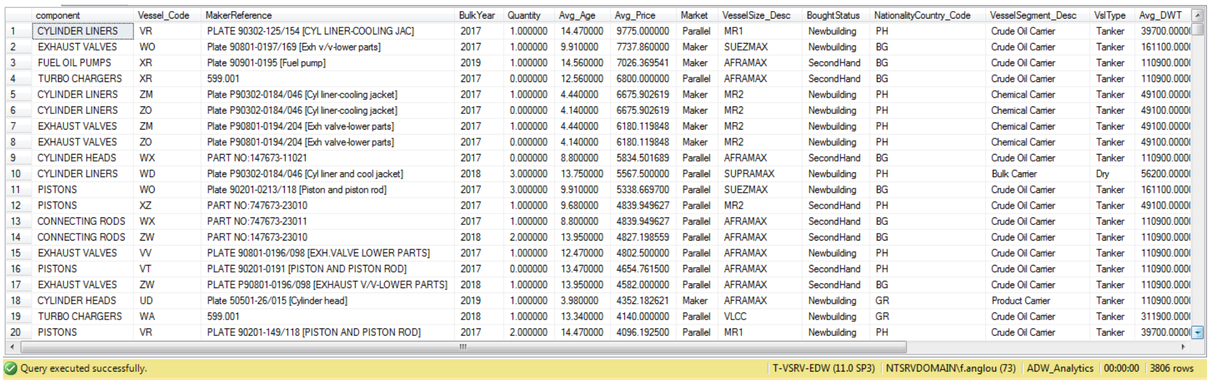


	MakerReference	Component	Quantity	Price	TotalVolume	AverageAge	UniqueVendors	UniqueVessels	BulkYear
1	PLATE 90302-125/154 [CYL LINER-COOLING JAC]	M/E	1.000000	9775.000000	9775.000000	16.470000	1	1	2017
2	Plate 90801-0197/169 [Exh v/v-lower parts]	M/E	1.000000	7737.860000	7737.860000	11.910000	1	1	2017
3	Plate 90901-0195 [Fuel pump]	M/E	1.000000	7026.369541	7026.369541	14.560000	1	1	2019
4	Plate P90302-0184/046 [Cyl liner-cooling jacket]	M/E	1.000000	6675.902619	6675.902619	6.440000	1	1	2017
5	Plate P90801-0194/204 [Exh valve-lower parts]	M/E	1.000000	6180.119848	6180.119848	6.440000	1	1	2017
6	PART NO:147673-11021	D/G	1.000000	5920.497684	5920.497684	11.290000	1	1	2017
7	Plate P90302-0184/046 [Cyl liner and cool jacket]	M/E	3.000000	5567.500000	16702.500000	14.750000	1	1	2018
8	Plate 90201-0213/118 [Piston and piston rod]	M/E	3.000000	5338.669700	16016.009100	11.910000	1	1	2017
9	520.000	D/G	1.000000	5100.000000	5100.000000	11.910000	1	1	2017
10	Plate 90901-0195 (FUEL PUMP ASSY)	M/E	1.000000	5027.146500	5027.146500	11.290000	1	1	2017
11	520.001	D/G	1.000000	4868.179500	4868.179500	4.430000	1	1	2017
12	PART NO:747673-23010	D/G	1.000000	4839.949627	4839.949627	11.680000	1	1	2017
13	PART NO:747673-23011	D/G	1.000000	4839.949627	4839.949627	10.800000	1	1	2017
14	PLATE 90801-0196/098 [EXH.VALVE LOWER PARTS]	M/E	1.000000	4802.500000	4802.500000	14.470000	1	1	2017
15	PART NO:147673-23010	D/G	3.000000	4792.374178	14377.122534	14.515000	1	2	2018
16	21.000	D/G	2.000000	4669.682400	9339.364800	14.380000	1	1	2019
17	SECTION NO:A21100/100 [CYL.HEAD REV:1D]	D/G	3.000000	4669.682400	14009.047200	1.440000	1	1	2019
18	PLATE P90801-0196/098 [EXHAUST V/V-LOWER PARTS]	M/E	1.000000	4582.000000	4582.000000	14.950000	1	1	2018
19	Plate 50501-26/015 [Cylinder head]	D/G	1.000000	4352.182621	4352.182621	3.980000	1	1	2019

Query executed successfully.

Figure 3.2-1: Snapshot of SQL code's results (1/2)

Above snapshot (Figure 3.2-2) is produced by SQL code used for the first methodological step of the practical part of the thesis.



	component	Vessel_Code	MakerReference	BulkYear	Quantity	Avg_Age	Avg_Price	Market	VesselSize_Desc	BoughtStatus	NationalityCountry_Code	VesselSegment_Desc	VslType	Avg_DWT
1	CYLINDER LINERS	VR	PLATE 90302-125/154 [CYL LINER-COOLING JAC]	2017	1.000000	14.470000	9775.000000	Parallel	MR1	Newbuilding	PH	Crude Oil Carrier	Tanker	39700.0000
2	EXHAUST VALVES	WO	Plate 90801-0197/169 [Exh v/v-lower parts]	2017	1.000000	9.910000	7737.860000	Maker	SUEZMAX	Newbuilding	BG	Crude Oil Carrier	Tanker	161100.000
3	FUEL OIL PUMPS	XR	Plate 90901-0195 [Fuel pump]	2019	1.000000	14.560000	7026.369541	Maker	AFRAMAX	SecondHand	BG	Crude Oil Carrier	Tanker	110900.000
4	TURBO CHARGERS	XR	599.001	2017	0.000000	12.560000	6800.000000	Parallel	AFRAMAX	SecondHand	BG	Crude Oil Carrier	Tanker	110900.000
5	CYLINDER LINERS	ZM	Plate P90302-0184/046 [Cyl liner-cooling jacket]	2017	1.000000	4.440000	6675.902619	Maker	MR2	Newbuilding	PH	Chemical Carrier	Tanker	49100.0000
6	CYLINDER LINERS	ZO	Plate P90302-0184/046 [Cyl liner-cooling jacket]	2017	0.000000	4.140000	6675.902619	Maker	MR2	Newbuilding	PH	Chemical Carrier	Tanker	49100.0000
7	EXHAUST VALVES	ZM	Plate P90801-0194/204 [Exh valve-lower parts]	2017	1.000000	4.440000	6180.119848	Maker	MR2	Newbuilding	PH	Chemical Carrier	Tanker	49100.0000
8	EXHAUST VALVES	ZO	Plate P90801-0194/204 [Exh valve-lower parts]	2017	0.000000	4.140000	6180.119848	Maker	MR2	Newbuilding	PH	Chemical Carrier	Tanker	49100.0000
9	CYLINDER HEADS	WX	PART NO:147673-11021	2017	0.000000	8.800000	5834.501689	Parallel	AFRAMAX	SecondHand	BG	Crude Oil Carrier	Tanker	110900.000
10	CYLINDER LINERS	WD	Plate P90302-0184/046 [Cyl liner and cool jacket]	2018	3.000000	13.750000	5567.500000	Parallel	SUPRAMAX	Newbuilding	PH	Bulk Carrier	Dry	56200.0000
11	PISTONS	WO	Plate 90201-0213/118 [Piston and piston rod]	2017	3.000000	9.910000	5338.669700	Parallel	SUEZMAX	Newbuilding	BG	Crude Oil Carrier	Tanker	161100.000
12	PISTONS	XZ	PART NO:747673-23010	2017	1.000000	9.680000	4839.949627	Parallel	MR2	SecondHand	PH	Crude Oil Carrier	Tanker	49100.0000
13	CONNECTING RODS	WX	PART NO:747673-23011	2017	1.000000	8.800000	4839.949627	Parallel	AFRAMAX	SecondHand	BG	Crude Oil Carrier	Tanker	110900.000
14	CONNECTING RODS	ZW	PART NO:147673-23010	2018	2.000000	13.950000	4827.198559	Parallel	AFRAMAX	SecondHand	BG	Crude Oil Carrier	Tanker	110900.000
15	EXHAUST VALVES	VV	PLATE 90801-0196/098 [EXH.VALVE LOWER PARTS]	2017	1.000000	12.470000	4802.500000	Parallel	AFRAMAX	Newbuilding	PH	Crude Oil Carrier	Tanker	110900.000
16	PISTONS	VT	PLATE 90201-0191 [PISTON AND PISTON ROD]	2017	0.000000	13.470000	4654.761500	Parallel	AFRAMAX	SecondHand	PH	Crude Oil Carrier	Tanker	110900.000
17	EXHAUST VALVES	ZW	PLATE P90801-0196/098 [EXHAUST V/V-LOWER PARTS]	2018	1.000000	13.950000	4582.000000	Parallel	AFRAMAX	SecondHand	BG	Crude Oil Carrier	Tanker	110900.000
18	CYLINDER HEADS	UD	Plate 50501-26/015 [Cylinder head]	2019	1.000000	3.980000	4352.182621	Maker	AFRAMAX	Newbuilding	GR	Product Carrier	Tanker	110900.000
19	TURBO CHARGERS	WA	599.001	2018	1.000000	13.340000	4140.000000	Parallel	VLCC	Newbuilding	GR	Crude Oil Carrier	Tanker	311900.000
20	PISTONS	VR	PLATE 90201-149/118 [PISTON AND PISTON ROD]	2017	2.000000	14.470000	4096.192500	Parallel	MR1	Newbuilding	PH	Crude Oil Carrier	Tanker	39700.0000

Query executed successfully. | T-VSRV-EDW (11.0 SP3) | NTSRVDOMAIN\f.anglou (73) | ADW_Analytics | 00:00:00 | 3806 rows

Figure 3.2-2: Snapshot of SQL code's results (2/2)

Above snapshot (Figure 3.2-2) is produced by SQL code used for the second methodological step of the practical part of the thesis.

In the next sections, the analytics framework provided in this chapter will be used to perform the three methodological steps (namely: the identification of high interest items, the forecasting of demand and the implementation of the prescriptive model) of the practical part of the thesis to create the integrated decision support tool for the bulk order process.

This page has been intentionally left blank

4 Identification of high-interest items

The bulk ordering process, as described (see section 1.2), is a time-consuming project and this is in part due to the very high number of line items that comprise a bulk order. In the interest of industry-wide standardization, each item in the spare parts industry can be referred to using a unique number called a maker reference. In each bulk order, there are around 4,000 distinct maker references making it very time consuming to compare the items and even to insert the prices of each supplier in a data base or/ and an ERP system.

Therefore, it seems important to be able to narrow down the high-interest items for each bulk order to facilitate and expedite the process. This way the analysis can be focused only on items that have been identified as high-interest and therefore the volume of administrative workload for the departments will be smaller.

4.1 Identification of analysis criteria

The identification variables that will be the input in the unsupervised learning algorithm is of great importance as the relationship between those will determine the items upon which the forecasting will be performed. The dataset that is used for clustering the items consists of data from all the bulk order projects (in total 3 bulk orders, one for the needs of 2017, one for the needs of 2018 and one for those of 2019) that the company has undertook. Therefore, the below variables were identified.

- Price: this variable indicates the acquisition price of the item (also accounting for discounts-if any apply)
- Quantity: this variable indicates the number of times this item was purchased in all the three bulks
- Total Volume: this variable indicates the product of the price and the quantity and is meant to increase the importance of items that have a medium price and were ordered a considerable amount of times thus making the total volume of those items quite large.
- Number of Unique Vessels: this variable indicates the number of different vessels that the item was installed on. This variable was inserted to increase the importance of an item, even if it doesn't have a considerable volume, price or quantity, if it is installed on many vessels and therefore has an increased influence in the uniformity and possible problems across several vessels.
- Average Age: this variable indicates the average age of the vessels said item is installed on.

4.2 Descriptive modelling and clustering analysis

4.2.1 The dataset

The dataset has been extracted from the case study company's data warehouse using SQL. Below a sample of the dataset is presented.

The below notation will be used here on after for the identification of components:

- Main Engine: M/E
- Diesel Generator: DG
- Purifiers: PUR (not present in below sample dataset)
- Compressors: COMP (not present in below sample dataset)

Table 4.2-1: Sample of dataset used for clustering

Maker Reference	Category	Quantity	Price	Total Volume	Average Age	Unique Vessels
[redacted]	M/E	2	10480.16	20960.32	14.01	1
[redacted]	M/E	1	9775	9775	16.14	1
[redacted]	M/E	1	7737.86	7737.86	11.58	1
[redacted]	M/E	1	7026.37	7026.37	14.23	1
[redacted]	M/E	1	6675.903	6675.903	6.11	1
[redacted]	M/E	1	6180.12	6180.12	6.11	1
[redacted]	DG	1	5920.498	5920.498	10.96	1
[redacted]	M/E	3	5338.67	16016.01	11.58	1
[redacted]	DG	1	5100	5100	11.58	1
[redacted]	M/E	1	5027.147	5027.147	10.96	1
[redacted]	DG	1	4868.18	4868.18	4.1	1
[redacted]	DG	1	4839.95	4839.95	10.47	1

4.2.2 DBSCAN Clustering

The scope of the clustering exercise is to determine the items that have an abnormally high price, quantity, combination of both or/ and are installed on several vessels. Therefore, what needs to be performed is a clustering that will identify the 'outliers' of the dataset thus labelling the items that have the characteristics described above. The ideal algorithm for this exercise is the dbscan algorithm also described in the previous chapter of the thesis (see sections 2.3.1.2) as it automatically creates a cluster containing the outliers.

Having decided the variables of the analysis, as per literature review (see sections 2.3.1.2) k is determined as:

$$\mathrm{k} = 2\,\mathrm{dim} = 10$$

, where dim is the number of dimensions (or variables) of the problem (5 variables as seen in the table above; Quantity, Price, Total Volume, Average Age, Unique Vessels).

The next step will be to determine the value of the parameter eps. As previously stated, the k-Nearest Neighbor (kNN, 10-NN) is created and is shown below.

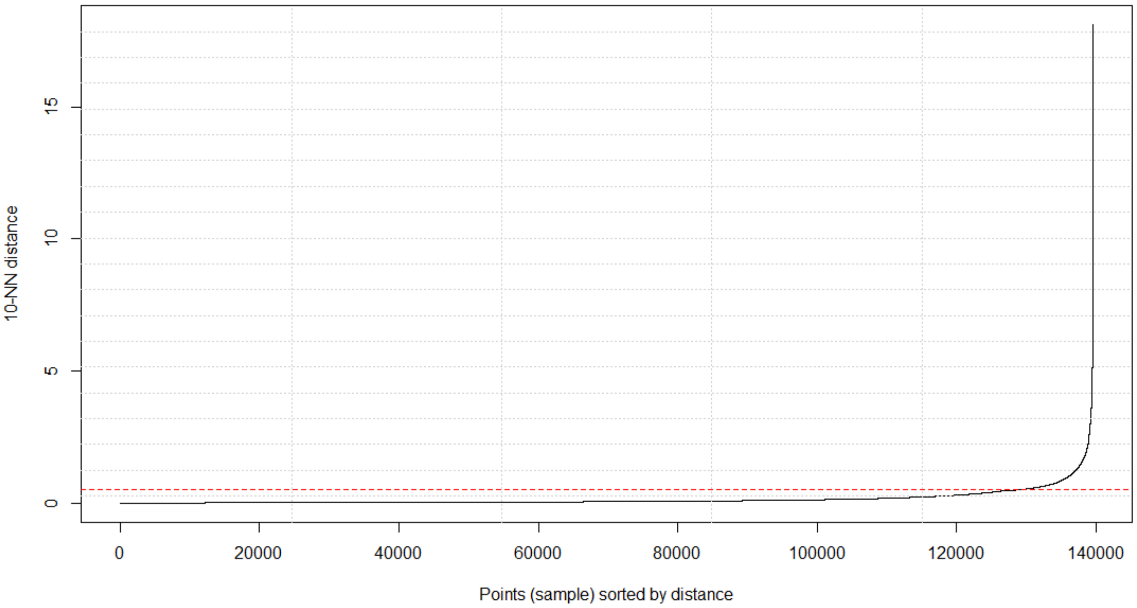


Figure 4.2-1: k-NN plot

As discussed in the literature review (see section 2.3.1.2) the value of eps is determined by the elbow of the graph (in the above graph shown by the red dotted line). The value of the eps parameter is determined at eps=1.8

Table 4.2-2: Cluster Means of Variables[1]

Cluster	Data Points	Quantity	Price	Total Volume	Average Age	Unique Vessels
0	376	110	841.05	4581.38	10.47	5.62
1	12254	11.54	56.407	221.28	10.68	2.05
2	9	168.6	1.3091	219.89	3.26	3.89

As can be seen from the table above three clusters have been created. The first cluster (cluster 0), here on after the outlying cluster, contains the outliers of the analysis. The mean quantity of the outlying class is considerably higher than the one of the second cluster (cluster 1), which contains the clear majority of the data, here on after the average class. The same can be said for the price of the outlying class as compared to the price of the average class. Evidently, the total volume, which is computed the product of the aforementioned characteristics (price and quantity), is also considerably higher. Finally, the unique vessels, that as mentioned before describes the number of different vessels that the specific item is installed on, is also considerably higher in the outlying class. However, the average age of the vessels is virtually the same for the two clusters.

All in all, the items of the outlying class represent the items that have considerable higher volume, price and quantity than the average meaning that they represent the cost drivers.

The third cluster (cluster 2) contains a small fraction of the total items that have a large quantity and are installed on several young vessels.

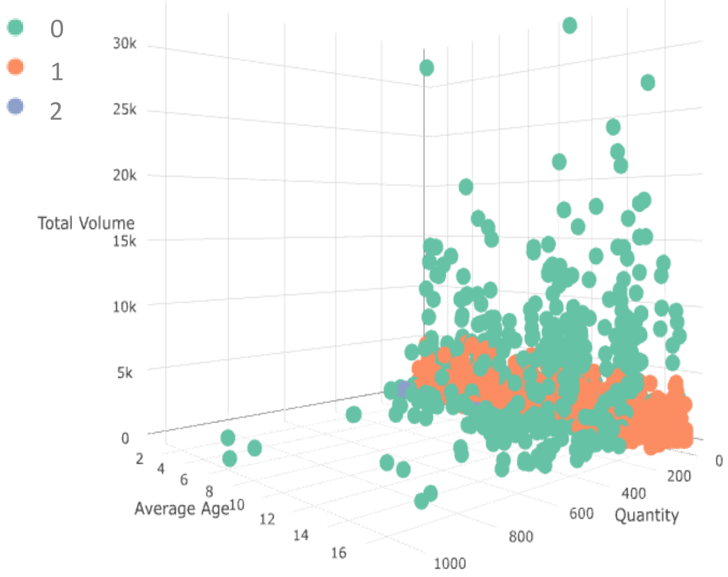


Figure 4.2-2: Results of clustering [2]

[1] Made using (Hahsler & Piekenbrock , 2018) Library: dbscan

[2] Made using (Sievert, 2018) Library: plotly for R

The items identified as cost drivers using clustering represent 3.06% of the items and 42.54% of the total cost of the bulk orders. Therefore, the cost drivers are the items that will be used further in the project to create the decision support tool.

4.2.3 Further analysis using k-means

To further analyse the data k-means clustering on the previously identified as high-interest items is performed. Identifying the optimum number of clusters is one way to limit the main disadvantage of the k-means. Therefore, for each cluster the total Euclidean distance, which is the total within-cluster sum of squares, is computed.

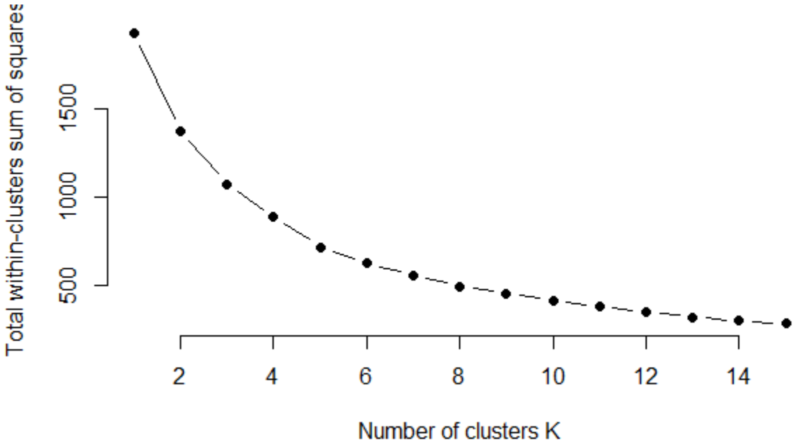


Figure 4.2-3: Optimum number of clusters

The figure above shows that the total distance between the data points and the centroid of the cluster that these data points belong in, decreases as the number of clusters increases. In the figure above there is no clear change of slope, except in cluster 7. However, considering the small number of data the number of clusters was chosen to be three. It is noted that any number between 3 and 5 could have been chosen as there is no significant change in the slope of the curve. The smallest was chosen to group the items in a smaller number of clusters thus creating broader business categories.

Table 4.2-3: Means of clusters produced by k-means algorithm

Cluster	Data Points	Total Volume	Price	Unique Vessels	Average Age
1	174	4172.0	1799	1.40	10.9
2	96	3169.0	1383	1.51	9.26
3	105	2781.7	584	1.70	11.06

As can be seen from the table above the means of the clusters cannot provide any more indications about the data in them. Therefore, the below graph is created.

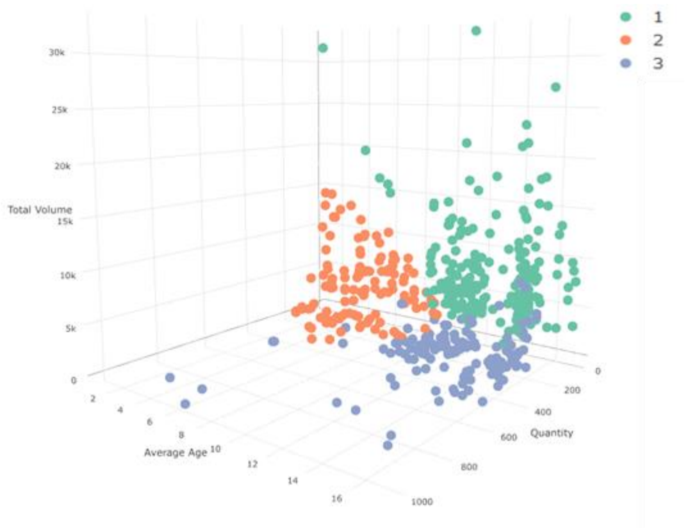


Figure 4.2-4: Kmeans 3D plot with 3 of the variables[3]

Making use of the graph above the description of each cluster is easier to be defined. Clusters 1 and 2 consist of high-volume items of low quantities. The differentiating variable here is the age where cluster 2 consists mainly of young vessels and cluster 1 of older vessels. Cluster 3 consists mainly of low volume and high quantity items for older vessels.

Table 4.2-4: Summarized results for 2nd clustering

Cluster	Description	Data Points	Total Volume	Percentage
1	High volume – young vessels	217	$1,617,972.90	73.27%
2	High volume – old vessels	126	$478,920.90	21.68%
3	Low volume	139	$111,329.80	5.04%

The cluster of k-means was further used as an independent variable in the forecasting analysis of the next chapter.

The items identified as cost drivers using clustering will be used as a basis for the bulk order price collection and winner selection. Having identified around 3% of the total items that represent around 50% of the total cost, the purchasing department will focus only on the pre-identified items to collect prices, assess the quotations, negotiate the prices and select the winner. Therefore, the administrative workload will decrease considerably in these steps of the bulk order process (see also Figure 3.1-1 and section 3.1). Concurrently, the items identified as cost drivers will provide the basis of analysis in the following steps of the practical part of the thesis. These items will be included in the forecasting part of the thesis (see section 5) and for the final step of the thesis which is the creation of the decision support tool (see section 6).

[3] Made using (Sievert, 2018) Library: plotly for R

This page has been intentionally left blank

5 Predictive forecasting model for spare parts

After having defined the items, identified by a unique number (maker reference), the predictive analysis is implemented.

This phase is divided in two parts. The first part refers to the forecasting of nominal bulk quantities based mainly on the characteristics of each vessel and the second phase refers to the forecasting of the extra quantities, if any, that will be needed during the next year based mainly on the buying patterns of the previous years.

5.1 Nominal Bulk Quantities

5.1.1 Identification of parameters

The first part of the forecasting process is to create the dataset. Using clustering a number of maker references has been identified as the cost and volume drivers of the spare parts bulk ordering process (see section 0). In the forecasting process, the machine learning algorithms will try to find a pattern between the data points in the dataset. The level on which the algorithms will try to find the relationships is described by the granularity level. Therefore, if we try to predict the nominal bulk quantities for each maker reference and for each vessel, we will need to create a small dataset for each maker reference. This way, the maximum number of data points that each dataset will have is 3, equal to the number of bulk years, thus making the forecasting process very difficult.

Therefore, the maker references have been grouped based on their relationship with certain components that can be found in the following table.

Table 5.1-1: Components and Number of Maker References

Component	Number of Maker References
AIR COMPRESSORS	19
ASSEMBLY	3
CAMSHAFT	1
CONNECTING RODS	10
CONROD (BIG END) BEARINGS	1
CYLINDER HEADS	72
CYLINDER LINERS	17
DIESEL GENERATOR	21
DRIVE SECTION	1
EXHAUST VALVES	13
FUEL INJECTION VALVES	38
FUEL OIL PUMPS	31
FUEL OIL PURIFIERS	22
FUEL OIL SYSTEM	4
LO SYSTEM	2
LUB OIL PURIFIERS	19
LUBRICATING SYSTEM	3
MAIN BEARINGS	2
MAIN DIESEL GENERATORS	26
MAIN ENGINE	7
MECHANICAL SYSTEM	12
PISTONS	49
SHAFT ASSEMBLY	5
TURBO CHARGERS	6

Therefore, the maker references have now been grouped into 24 components[4]. Then the dataset is created making extensive use of SQL code as described in the previous section (see section 3.2)

5.1.1.1 Defining parameters for one component

The final product of the forecast will be the nominal quantity that a vessel needs for the following for each vessel and for each maker reference. The variables that will be used for the forecasting of the final quantity are:

- Average Age: The age of the vessel is one of the most important vessel characteristics and as described in the first chapter the maintenance of the vessel and thus the quantities of the item that will be ordered are highly correlated.
- Average Price: The price of an item is one of the most important demand characteristics and in the sections below its relationship with the final quantity will be examined.
- DWT: This variable describes the dead-weight tonnage of the vessel and therefore is an indicator to the size of the vessel and to its needs
- C/Eng Nationality: This variable refers to the nationality of the chief engineer of the vessel and aims to unveil influences of education and culture that can be associated with the chief engineer's nationality. It is a categorical variable of three levels: GR for Greece, BG for Bulgaria and PH for Philippines.
- Type: The combination of this variable with the DWT declares the size of the vessel (eg. Aframax, Suezmax etc.) Categorical variable of two levels: Tanker and Dry
- Origin: This variable indicates the country of construction of the vessel. It is a categorical variable of three levels: SOUTH KOREA, JAPAN and CHINA and aims to unveil correlations between the shipyard and the quality of the vessel.
- k-means cluster: as described in the previous section (see section 4.2.3) the cluster of the k-means groups the items based on k-means algorithm creating 3 clusters.

Table 5.1-2: Sample dataset for forecasting of nominal needs for component AIR COMPRESSORS

Maker Reference	Quantity	Average Age	Average Price	C/Eng Nationality	Type	DWT	Origin	K-means cluster
[redacted]	1	8.3	$ 1,139.21	PH	Tanker	110900	CHINA	1
[redacted]	1	7.87	$ 1,221.59	PH	Tanker	110900	CHINA	1
[redacted]	1	8.3	$ 1,461.78	PH	Tanker	110900	CHINA	1
[redacted]	26	0.79	$ 0.35	GR	Tanker	161100	KOREA	1
[redacted]	10	1.17	$ 0.15	PH	Tanker	110900	JAPAN	2
[redacted]	10	1.65	$ 0.15	PH	Tanker	110900	JAPAN	2
[redacted]	5	1.17	$ 0.26	PH	Tanker	110900	JAPAN	1
[redacted]	5	1.65	$ 0.26	PH	Tanker	110900	JAPAN	1
[redacted]	1	6.33	$ 0.15	PH	Tanker	49100	KOREA	1

5.1.1.2 Exploring regression techniques

In order to understand the variables of the dataset and the effect that they have on the nominal quantity of the bulk order, a number of statistical methods are applied. The analysis below was performed on the component AIR COMPRESSORS.

[4] There are 11 maker references that have not been grouped in one of the components of Table 5.1-1. This happens because in the database no component was registered for said maker reference.

5.1.1.2.1 Linear Regression

The first method explored was linear regression models. Having determined the variables of the analysis a correlation matrix is created and is visualized below.

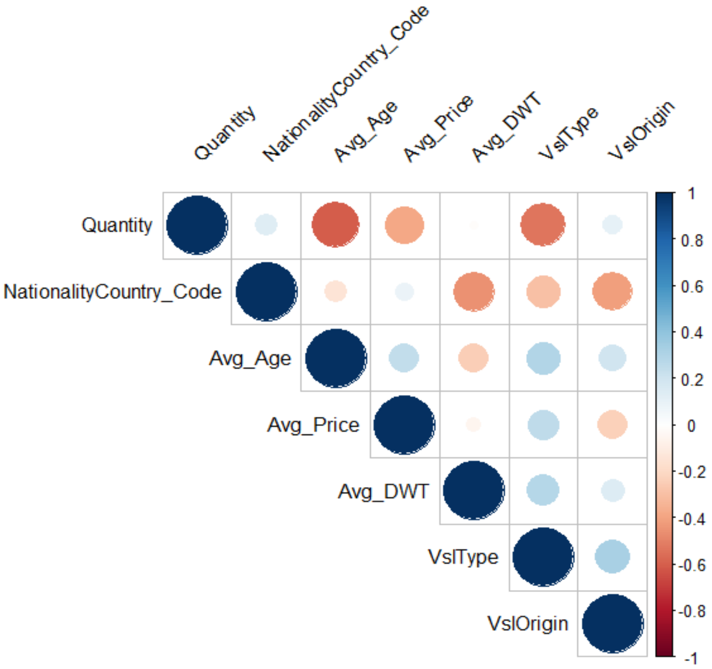


Figure 5.1-1: Correlation matrix for air compressors

As can be seen from figure above there is no significant correlation between any of the variables and the final quantity. This is assuming that a significant (strong) correlation (either positive or negative) is described by a correlation coefficient absolute value of higher than 0.6 (Evans J. , 1996)The two variables that have the highest significance for the final nominal need for air compressors are the age of the vessel (as Avg_Age) and the type of the vessel (as VslType).

Since the correlation between age of the vessel and the quantity is negative it means that as the age of the vessel increases the quantity decreases. There are two factors that can explain this finding. First, since the dry-docking procedures are happening in shorter time intervals the quantities needed to be procured each year in the bulk orders are lower. At the same time, for commercial reasons sometimes maintenance activities are not completed with the same intensity as for younger vessels for cost containment purposes since the vessel is coming towards its end of life. For example, if the company looks to sell an older vessel will not spend a considerate amount in overhauls and dry-docks. Instead, it will keep costs to the lowest possible levels.

As far as the vessel type is concerned, the correlation coefficient value is -0.5 which can be considered moderate relationship. For calculation purposes tanker- type was denoted as 0 and dry-type was denoted as 1. The quantities are therefore higher for tanker vessels in relation to the dry bulk carriers.

Finally, there is a negative correlation between the price and the quantity. The value of the correlation coefficient is -0.2 which indicates a weak negative relationship. This is expected as it is normal when the quantity increases for the price to drop.

Having checked the correlations between several variables and the dependent variable (quantity) linear models are created. Linear models were created for the numeric variables (the age of the vessels denoted as avg_age and the price denoted as avg_price) and for the categorical variable with the highest correlation with the dependent variable: the type of the vessel.

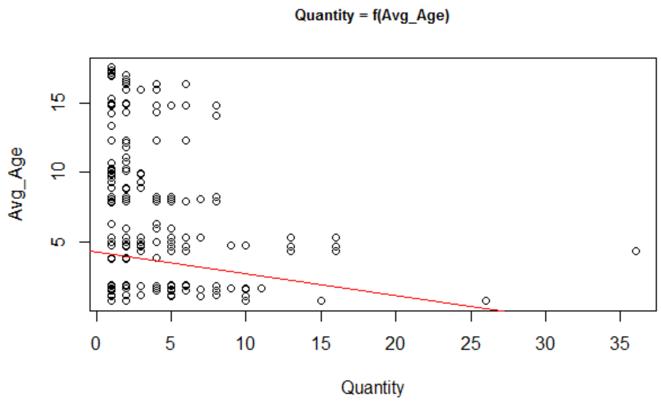


Figure 5.1-2: Linear regression for AIR COMPRESSORS

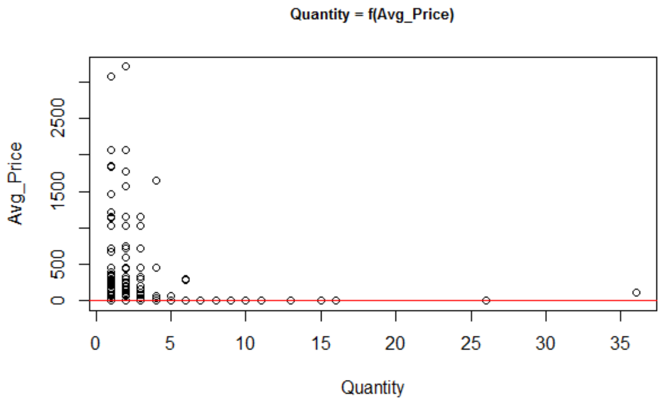


Figure 5.1-3: Linear regression for AIR COMPRESSORS

As is observed by the graphs above, it is very difficult to approximate the relationship between the ordered quantity and the different aspects of the vessel or/ and the component that influence it.

Table 5.1-3: Linear Regression and results

Number of Variables	Variable	X	Coefficient	R^2
1	Average Age	-0.158	4.34	4.1%
1	Average Price	-0.002	3.88	6.4%
1	Vessel Type	-1.701	2.762	3.8%

The analysis performed with statistical models, has shown that there is no profound relationship and correlation between the variables, thus making the forecasting with conventional methods rather difficult.

Linear regression with two variables was also explored. However, the results are not satisfactory. Even when combining the variables Average Age and Average Price the coefficient of determination, R^2, is 8.31% which is very low. However, since it is not justified to dismiss linear regression solely on the coefficient of determination below graphs, and their ideal shape is also included.

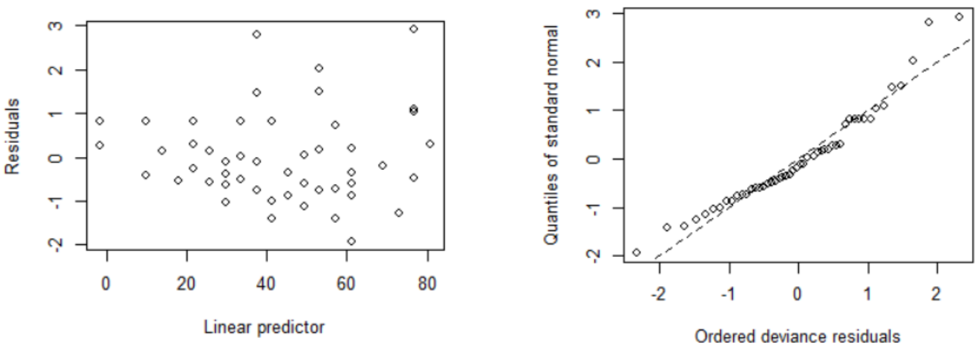


Figure 5.1-4: Ideal Shape of residuals vs fitted (left) and QQ plot (right)

The residuals vs fitted graph tests whether the relationship between the variables is linear (i.e. linearity) and whether there is equal variance along the regression line (i.e. homoscedasticity). The ideal residuals vs fitted plot should be relative shapeless (as shown in the figure above) and be generally symmetrically distributed around the 0 line.

The QQ plot helps determine if the dependent variable is normally distributed by plotting quantiles from the dataset's distribution against a theoretical distribution. If the data is normally distributed it will be plotted in a generally straight line (as shown in the figure above)

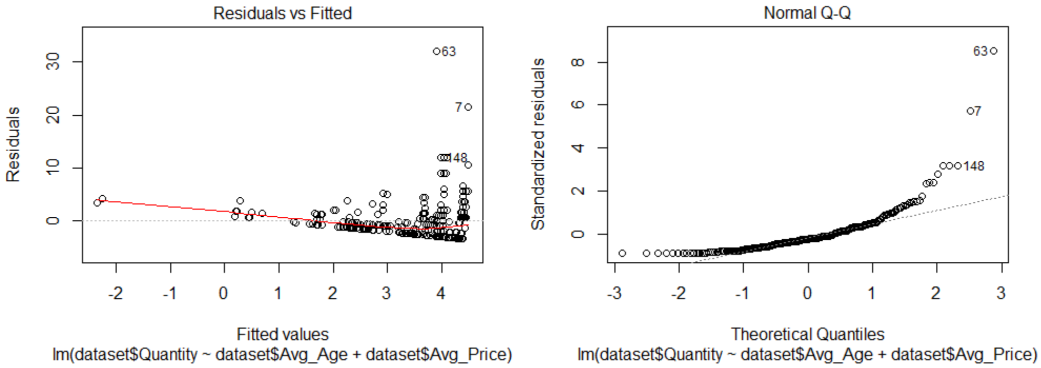


Figure 5.1-5: Residuals vs fitted (left) and QQ plot (right) after implementing linear model with age and price

As can be seen from the graphs above and from the comparison with the ideal shapes neither the residuals vs fitted plot nor the normal QQ plot are satisfactory.

The above along with the low value of R^2 are enough to dismiss normal linear regression as a forecasting method.

5.1.1.2.2 Time series analysis

Another approach that could be used for forecasting is the time series analysis. The bulk ordering is a project that is undertaken every year and the demand of the spare parts is influenced by several factors. Initially, someone could argue that the best way to forecast demand would be by using time series analysis. However, in this case the time series analysis is not performed due to the reasons below:

- The size of the dataset for the desired granularity is extremely small. As previously described (5.1.1), the forecasting will take place for distinct maker references and for a series of vessel characteristics that practically define a unique vessel. Therefore, the dataset of each quantity that needs to be forecasted has at most 3 data points, each data point representing one of the bulks already carried out by the case company. As is cited in literature review (Hanke & Wichern , 2009) minimum of the sample size to capture patterns in time series needs to be at least 50. This makes the time series analysis for the problem a poor choice.

- Additionally, demand of spare parts is heavily influenced by exogenous factors such as the vessel characteristics. As it was observed by the correlation matrix above (5.1.1.2.1) the vessel characteristics heavily influence the final demand of the vessel. Therefore, the time series forecasting which is used to predict values based on previously observed values (Imdadullah, 2014)cannot easily account for the needed exogenous factors.
- Using above analyses of linear regression, it is noted that the above dataset exhibits signs of multicollinearity, heteroscedasticity and non-stationarity and therefore certain transformations need to be performed (Deviant , 2012)This can be proven quite difficult considering the two points above.

The inapplicability of time series analysis was validated by utilization of the auto.arima package within R. The end goal was to identify a data set transformation viable enough to generate reliable results using the optimum ARIMA parametrization. For most of the components of the dataset auto.arima identified as optimum the set-up ARIMA (0,0,0), i.e. approximating the time series as “white noise”, which means that the dataset can be characterized as a sequence of random numbers and cannot be predicted if no further actions are taken. Below an example of forecasted quantities is presented:

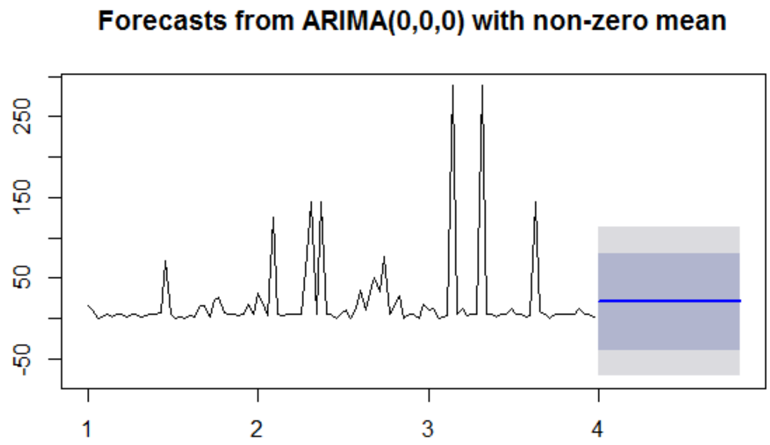


Figure 5.1-6: Results of ARIMA forecasts for the majority of the components[5]

The approximation of the series as white noise results in a constant non-zero value as the forecasted quantity which does not capture the desired result.

5.1.2 Application of machine learning

Having considered the above analyses as well as the small size of the dataset, the applicability of machine learning, and advanced statistics will be explored to see if there is indeed ground to

[5] Made using (Hyndman, et al., 2019) Library: foreast

formulate a predictive model in this problem domain. The algorithms described in the previous chapter (see section 2.3.2) of the thesis will be tested and evaluated to this end.

Smaller datasets for each component including all the distinct maker reference numbers that fall under said component along with the extracted aforementioned variables from the database are created. A loop is then created, which initially checks if there are more than 20 entries in said sub-dataset for the component to ensure a large number of data to perform the next task. Then the data are divided into two sets: the training set and the test set with a random 80-20 data partition, as for most of the sub-datasets the number of data points is not sufficient enough to also create a validation set. Finally, the training of the models is performed using the trainsets.

The following pseudo-code explains the process described above:

```
data <- get dataset from SQL[6]
for all components [i=1,. . .,N]
    component_dataset [[i]] <- data (where component[i] = component)
    if nrow (component_dataset [[i]]) >20
        data_partition <- 80/20
        trainset <- 0.8*data
        testset <- -(trainset)
        model_prediction 1<- train random forest
        model_prediction 2<- train generalized linear regression
        model_prediction 3<- train principal component regression
    end
end
```

Having trained the models, the testing process is then performed. For each model the test set is inserted as input and the output is then compared to the actual quantity. For each entry the forecasting error is computed and for each component the mean average percentage error is also computed.

```
for all components [i=1,. . .,N]
    if nrow (component_dataset [[i]]) >20
        model_results 1<- test random forest
        model_results 2<- test generalized linear regression
        model_results 3<- test principal component regression
        mape[i] <- compute mape for component i
    end
end
```

5.1.2.1 Forecasting results

For each component the mean absolute percentage error is computed as per below formula:

$$MAPE = \frac{|forecast - actual|}{actual}$$

The results are presented in the table below. For some components forecasting did not take place as the entries were not enough to properly train and test the algorithms. Here on after the three algorithms used for forecasting are denoted as follows:

- Random Forest: RF
- Generalized Linear Model: GLM
- Principal Component Regression: PCR

[6] Made using (Ripley & Lapsley, 2017) Library: RODBC

Table 5.1-4: Mean absolute errors for each component [7]

Component	RF	GLM	PCR	Number of Points
AIR COMPRESSORS	34%	119%	112%	268
ASSEMBLY	-	-	-	6
CAMSHAFT	64%	61%	58%	22
CONNECTING RODS	77%	111%	272%	31
CONROD (BIG END) BEARINGS	-	-	-	9
CYLINDER HEADS	140%	220%	221%	893
CYLINDER LINERS	111%	311%	301%	155
DIESEL GENERATOR	125%	297%	146%	114
DRIVE SECTION	39%	46%	27%	22
EXHAUST VALVES	-	-	-	17
FUEL INJECTION VALVES	98%	109%	128%	283
FUEL OIL PUMPS	103%	104%	103%	153
FUEL OIL PURIFIERS	77%	100%	94%	427
FUEL OIL SYSTEM	22%	45%	34%	26
LO SYSTEM	-	-	-	16
LUB OIL PURIFIERS	52%	99%	103%	519
LUBRICATING SYSTEM	-	-	-	4
MAIN BEARINGS	-	-	-	13
MAIN DIESEL GENERATORS	122%	78%	88%	50
MAIN ENGINE	-	-	-	10
MECHANICAL SYSTEM	137%	172%	166%	194
PISTONS	64%	256%	259%	294
SHAFT ASSEMBLY	64%	70%	85%	80
TURBO CHARGERS	48%	70%	48%	54

As can be seen from the table above the forecasting error is, in some cases, considerable and in some cases, it can be characterized as rather satisfactory (those below 40%).

To better visualize the performance of the algorithms below indicative results for each component are presented.

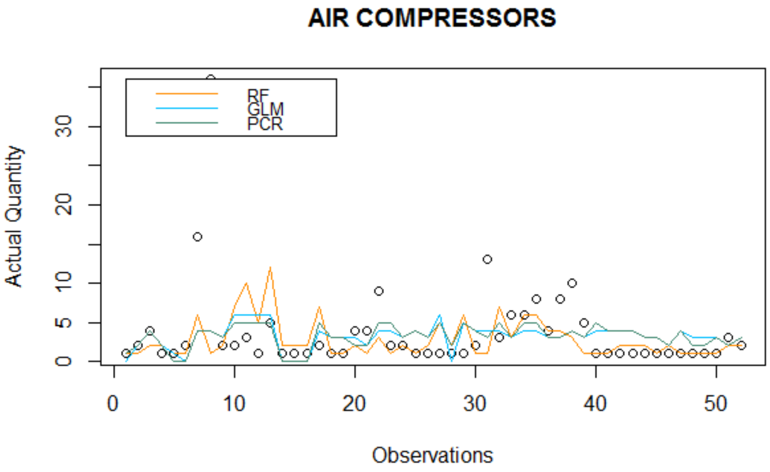


Figure 5.1-7: Forecasting results for the component AIR COMPRESSORS

[7] Made using (Kuhn, et al., 2018) Library: caret.
All the following forecasting results were found using caret library.

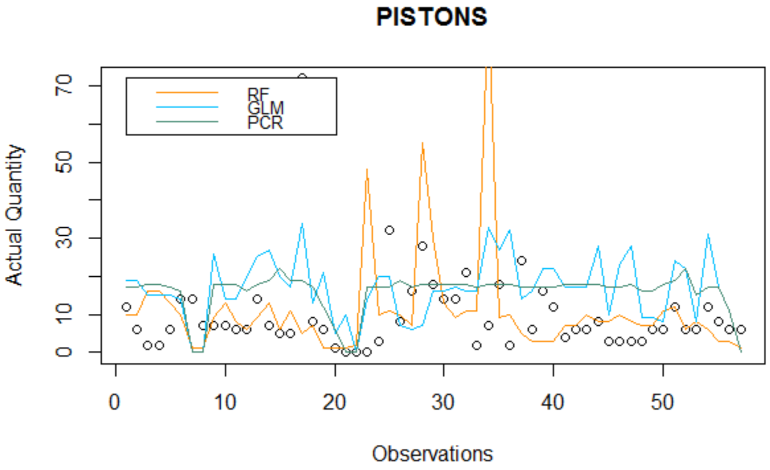


Figure 5.1-8: Forecasting results for the component PISTONS

What needs to be noted is that x axis in the graph above only denotes the observation and does not have any affiliation with time.

5.1.2.2 Handling of data and forecasting results

For some components, making use of the graphs it was observed that even though the fitting of the model seemed rather satisfactory the error was extremely high. An indicative case is the below:

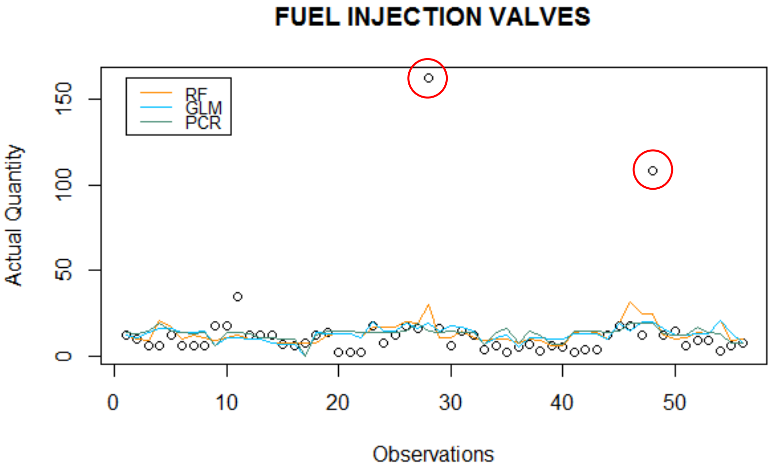


Figure 5.1-9: Forecasting results for the component FUEL INJECTION VALVES

For the component fuel injection valves, the error of the random forest algorithm is 98% (which is the minimum of the three algorithms).

However, as can be observed from the graph the fitting seems to be much better than the error. The two data points circled in red in above graph that have extremely high quantity as compared to the rest of the dataset drive the error to higher levels.

Therefore, a data cleansing method is used to determine those data points and eliminate them from the training and evaluating sets of the algorithms. In the specific dataset it is common to come across data points that can be considered as outliers. In a business sense, this can be explained by a superintended engineer calculating the values mistakenly or a vessel having abnormally high or low needs for a year.

Therefore, dbscan (see section 2.3.1.2) is used to determine the outliers and exclude those data points from the analysis (see section 2.2.1.1).

While performing the above data cleansing method the below parameters were used

- minPts = 6: this parameter was chosen as dimensions of the problem multiplied by two (see 2.3.1.2). The dimensions for which dbscan is applied are only the numeric ones (age of the vessel, DWT, price of the item)
- eps = 0.5: this parameter was chosen as dictated by literature review (see 2.3.1.2). However, since every component has a different optimum eps value the average was assumed as eps value for the parameters of DBSCAN performed below.

In the next section (see 0) the assumption of the eps value is tested to determine the optimum eps for the components.

The table below presents the number of outliers for each component and the relative relation with the total number of observations for each component.

Table 5.1-5: Number of outliers per category

Component	Outliers	Number of Points	Percentage
AIR COMPRESSORS	27	268	10.1%
ASSEMBLY	-	6	-
CAMSHAFT	14	22	63.6%
CONNECTING RODS	20	31	64.5%
CONROD (BIG END) BEARINGS	-	9	-
CYLINDER HEADS	34	893	3.8%
CYLINDER LINERS	20	155	12.9%
DIESEL GENERATOR	12	114	10.5%
DRIVE SECTION	22	22	100.0%
EXHAUST VALVES	17	17	100.0%
FUEL INJECTION VALVES	37	283	13.1%
FUEL OIL PUMPS	46	153	30.1%
FUEL OIL PURIFIERS	22	427	5.2%
FUEL OIL SYSTEM	8	26	30.8%
LO SYSTEM	-	16	-
LUB OIL PURIFIERS	51	519	9.8%
LUBRICATING SYSTEM	0	4	0.0%
MAIN BEARINGS	-	13	-
MAIN DIESEL GENERATORS	6	50	12.0%
MAIN ENGINE	-	10	-
MECHANICAL SYSTEM	20	194	10.3%
PISTONS	42	294	14.3%
SHAFT ASSEMBLY	26	80	32.5%
TURBO CHARGERS	22	54	40.7%

For the components EXHAUST VALVES and DRIVE SECTION the outliers are 100% of the dataset. This means that dbscan cannot classify the items and labels them as outliers and therefore, the components have been excluded from the implementation of the algorithms below.

Having handled the dataset in the way that was described above, the three algorithms are applied again and are trained and tested on the ‘cleaned’ dataset. The pseudo code of the analysis and its results are presented below. In the following pseudo code suffix ‘_no’ is used to emphasize the absence of the outliers.

```
eps <- 0.5
```

```
minPts <- 5
for all components [i=1,. . .,N]
    db[[i]] <- dbscan on component_dataset [[i]]
    dataset_no [[i]] <- db[[i]] [where db$cluster <> 0]
    if nrow (component_dataset [[i]]) >15
        data_partition <- 80/20
        trainset_no [[i]] <- 0.8*dataset_no [[i]]
        testset_no [[i]] <- - (dataset_no[[i]])
        model_prediction_no 1<- train random forest
        model_prediction_no 2<- train generalized linear regression
        model_prediction_no 3<- train principal component regression
    end
end
```

and hereafter the second loop used to extract the results:

```
for all components [i=1,. . .,N]
    if nrow (component_dataset [[i]]) >15
        model_results_no 1<- test random forest
        model_results_no 2<- test generalized linear regression
        model_results_no 3<- test principal component regression
        mape_no[i] <- compute mape for component i
    end
end
```

The table below presents the results of the implementation of the algorithms already discussed (see section 5.1.2) after performing data handling methods.

Table 5.1-6: Mean absolute errors for each component after outlier elimination

Component	RF	GLM	PCR	Number of Points
AIR COMPRESSORS	31%	115%	94%	241
ASSEMBLY	-	-	-	6
CAMSHAFT	-	-	-	8
CONNECTING RODS	0%	0%	0%	11
CONROD (BIG END) BEARINGS	-	-	-	9
CYLINDER HEADS	125%	153%	158%	859
CYLINDER LINERS	122%	242%	149%	135
DIESEL GENERATOR	83%	117%	93%	102
DRIVE SECTION	-	-	-	0
EXHAUST VALVES	-	-	-	0
FUEL INJECTION VALVES	44%	63%	56%	246
FUEL OIL PUMPS	68%	100%	101%	107
FUEL OIL PURIFIERS	46%	89%	91%	405
FUEL OIL SYSTEM	10%	20%	30%	18
LO SYSTEM	17%	17%	17%	16
LUB OIL PURIFIERS	53%	93%	92%	468
LUBRICATING SYSTEM	-	-	-	4
MAIN BEARINGS	-	-	-	0
MAIN DIESEL GENERATORS	162%	198%	141%	44
MAIN ENGINE	-	-	-	10
MECHANICAL SYSTEM	71%	72%	76%	174
PISTONS	53%	107%	125%	252
SHAFT ASSEMBLY	29%	46%	55%	54
TURBO CHARGERS	0%	17%	0%	32

To better visualize the performance of the algorithms below indicative results for each component are presented.

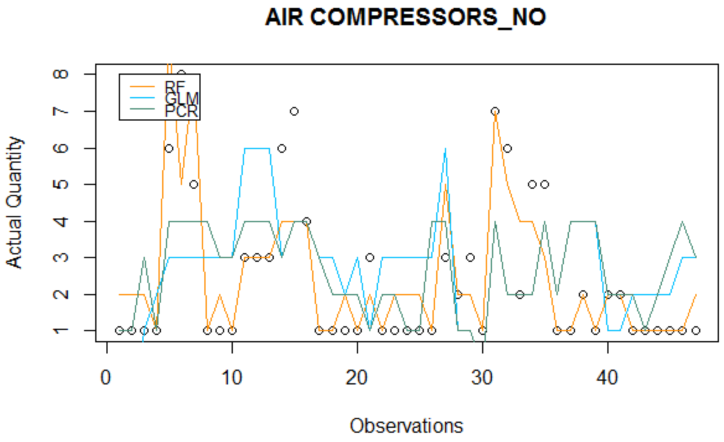


Figure 5.1-10: Forecasting results for the component AIR COMPRESSORS after outlier elimination

For the component AIR COMPRESSORS total error was improved by 3% (see Figure 5.1-7 and Table 5.1-4). Another example is provided below:

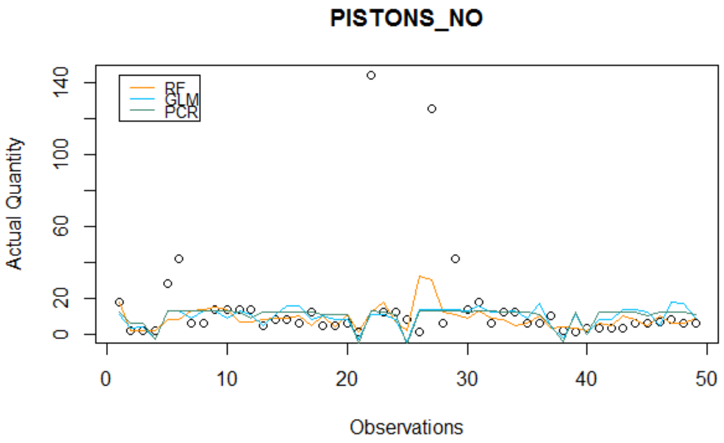


Figure 5.1-11: Forecasting results for the component PISTONS after outlier elimination

For the component AIR COMPRESSORS total error was improved by 3% (see Figure 5.1-8 and Table 5.1 4).

By comparing the two tables that summarize the results of the two methods (see Table 5.1-4 and Table 5.1-6), the forecasting after outlier elimination is a process that yields better results than simple forecasting. As an indication the total average error of the forecasting without handling of outliers is 73% and after outlier elimination it decreases by 17% to 56%.

5.1.2.3 Optimization of data cleansing algorithm and forecasting results

As previously discussed, the results of the chosen outlier handling method (dbscan) are heavily influenced by its two parameters (see section 2.3.1.2). The first parameter (k) is defined as per literature review. As the forecasting of the different components is governed by the same number of dimensions, this parameter can be considered as constant across the different components.

However, the second parameter, eps, is defined by interpreting a graph. This does not allow the algorithm to be fully automated as the user needs to interpret the produced graph and then to determine the value of the eps parameter. To avoid this, a loop is performed to determine the best value for this parameter.

However, as eps decreases, meaning that dbscan will consider data points to be in the same cluster only if they have distance smaller than eps, the number of points that are labelled as outliers increases. This way, the forecasting error decreases but the probability the model over fits the data increases. Therefore, it is very difficult to determine the ideal eps for each component.

As can be seen from the table of the previous section (5.1.2.2) there are some components that the outlier handling could not decrease the error to satisfactory levels. Namely those components are:

- Cylinder Heads
- Cylinder Liners
- Main Diesel Generators

For those components below the analysis of the optimum eps value is presented.

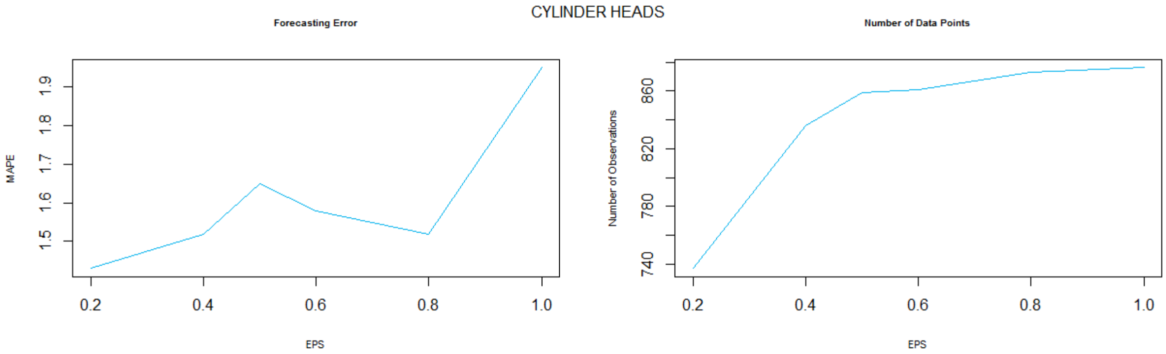


Figure 5.1-12: Change of MAPE (left) and number of data points (right) as eps parameter increases for component CYLINDER HEADS

As described above, when the eps parameter decreases the number of data points included in the analysis is increased (the outliers are decreasing). At the same time the forecasting error decreases as well. As can be seen from the graph the optimum error (without simultaneous elimination of a considerable amount of data points) is at eps=0.8.

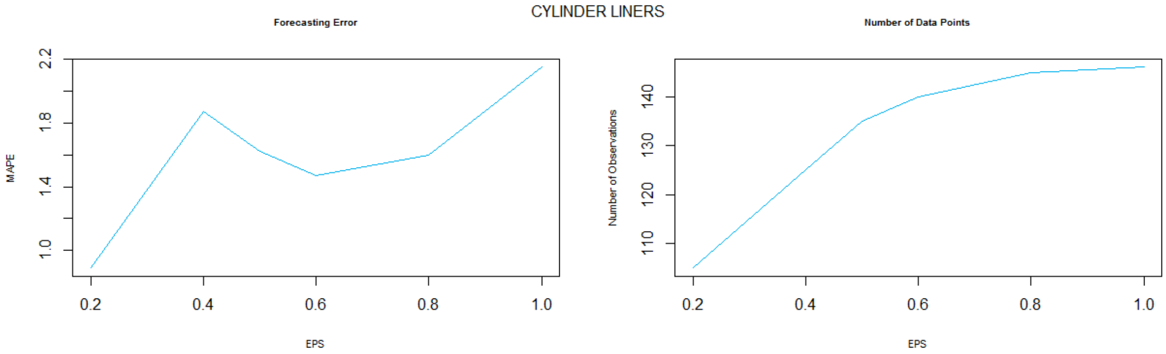


Figure 5.1-13: Change of MAPE (left) and number of data points (right) as eps parameter increases for CYLINDER LINERS

Another example is given in the figure above referring to component main diesel generators. The behavior here is similar to the component described above where the optimum point can be found at eps=0.8.

The analysis above can be considered an area of further research into optimizing the model and especially the data handling section.

5.1.3 Synthesis of results

To conclude, the forecasting of the nominal quantities the six models produced by the three different methods (RF, PCR, GLM) are evaluated and the results are presented on the table below. For the next steps of the thesis the algorithms presented below are considered the best performing algorithms and are used for further steps.

Table 5.1-7: Least mean absolute error for each component for all methods

Component	Minimum MAPE	Method
AIR COMPRESSORS	31%	Random Forest No Outliers
ASSEMBLY	-	-
CAMSHAFT	58%	Principal Component Regression
CONNECTING RODS	0%	Random Forest No Outliers
CONROD (BIG END) BEARINGS	-	-
CYLINDER HEADS	125%	Random Forest No Outliers
CYLINDER LINERS	111%	Random Forest
DIESEL GENERATOR	83%	Random Forest No Outliers
DRIVE SECTION	27%	Principal Component Regression
EXHAUST VALVES	-	-
FUEL INJECTION VALVES	44%	Random Forest No Outliers
FUEL OIL PUMPS	68%	Random Forest No Outliers
FUEL OIL PURIFIERS	46%	Random Forest No Outliers
FUEL OIL SYSTEM	10%	Random Forest No Outliers
LO SYSTEM	17%	Random Forest No Outliers
LUB OIL PURIFIERS	52%	Random Forest
LUBRICATING SYSTEM	-	-
MAIN BEARINGS	-	-
MAIN DIESEL GENERATORS	78%	Generalized Linear Model
MAIN ENGINE	-	-
MECHANICAL SYSTEM	71%	Random Forest No Outliers
PISTONS	53%	Random Forest
SHAFT ASSEMBLY	29%	Random Forest No Outliers
TURBO CHARGERS	0%	Random Forest No Outliers

As can be easily observed from the table above the best performing method for the vast majority of the components is the random forest algorithm. This result was expected as the random forest algorithm (as previously discussed in 2.3.2.1) best handles the exogenous factors that influence the outcome in a stochastic manner that makes it impervious to over/ under fitting.

5.2 Additional quantities during the year

As previously described, the quantities of the bulk ordering process are meant to cover the fleet's needs for the next year, yet on occasion those items are reordered during the year. This is either due to miscalculation of the vessel's needs or failure of previously bought equipment. To determine those additional quantities a forecasting process is used.

First, the dataset is created following the same actions as for the nominal quantities. The same vessel characteristics are extracted from the database, but an additional column is created.

- Bulk Market: This variable is a categorical variable that has two values: maker and parallel. This variable describes the source of purchase of each item purchased in the bulk order process. These variables will be used to explore possible correlation between the source of purchase and the additional quantities

What needs to be predicted is the variable here on after denoted as “NonBulk”, which is the extra quantity that was purchased during the year.

Table 5.2-1: Sample dataset for forecasting of extra needs for component FUEL INJECTION VALVE

Maker Reference	Category	Bulk Market	Non-Bulk	Quantity	Type	DWT	Average Price	Age	Origin	C/Eng Nationality
[redacted]	DG	Maker	0	2	Dry	49600	$ 448.24	3.75	KOREA	PH
[redacted]	DG	Maker	0	3	Tanker	161100	$ 429.03	0.79	KOREA	GR
[redacted]	DG	Maker	9	7	Tanker	161100	$ 177.70	0.79	KOREA	GR
[redacted]	DG	Maker	56	112	Tanker	161100	$ 1.35	0.79	KOREA	GR
[redacted]	DG	Parallel	0	0	Tanker	110900	$ 729.63	8.67	KOREA	BG
[redacted]	DG	Parallel	6	1	Tanker	311900	$ 72.25	12.35	KOREA	GR
[redacted]	DG	Parallel	0	2	Dry	56200	$ 717.22	12.63	CHINA	PH
[redacted]	ME	Parallel	4	3	Tanker	110900	$ 465.48	13.35	KOREA	PH
[redacted]	ME	Parallel	0	4	Tanker	110900	$ 232.74	7.87	CHINA	PH
[redacted]	ME	Parallel	0	6	Tanker	161100	$ 503.81	9.79	KOREA	BG
[redacted]	ME	Parallel	0	6	Tanker	39700	$ 558.57	15.55	KOREA	PH

In the training of the algorithms for the forecasting of the extra quantities there is the difficulty that most of the data (used for training and testing) are zero, meaning that an item was purchased in the bulk order but was not re-purchased during the year.

Because of the intermittent nature of demand in these cases it is proven difficult to train the algorithms in some of the sub-datasets. It was decided that if any sub-dataset had a mean of non-bulk quantities of less than 0.1 it would not be used for forecasting, but, rather the extra quantities would be assumed to be all zeroes. This threshold was decided as it was the minimum variation in the target variable required for efficient training of the dataset. For the forecasting of the extra quantities, the same procedure as previously is used. The pseudocode is presented below:

```
for all components [i=1,. . .,N]
    extra_dataset [[i]] <- all from data where component[i] = component
    if nrow (extra_dataset [[i]]) >20 && mean (extra_dataset [[i]]) >0.1
        data_partition <- 80/20
        trainset <- 0.8*data
        testset <- -(trainset)
        model_prediction 1<- run random forest
        model_prediction 2<- run generalized linear regression
        model_prediction 3<- run support vector regression
    end
end

for all components [i=1,. . .,N]
    if nrow (extra_dataset [[i]]) >20 && mean (extra_dataset [[i]]) >0.1
        model_results 1<- test random forest
        model_results 2<- test generalized linear regression
        model_results 3<- test principal component regression
        mape[i] <- compute mape for component i
    end
end
```

The models are trained and tested according to the pseudocodes above thus producing the results of the following section.

5.2.1 Forecasting results

For each component the mean absolute percentage error is presented in the table below. For some components an accurate forecast could not be generated as the entries were not enough to properly train and test the algorithms.

Table 5.2-2: Mean absolute errors for extra needs for each component

Component	RF	GLM	PCR	Number of Points
AIR COMPRESSORS	-	-	-	326
ASSEMBLY	-	-	-	6
CAMSHAFT	672%	1361%	530%	24
CONNECTING RODS	14%	114%	97%	35
CONROD (BIG END) BEARINGS	-	-	-	10
CYLINDER HEADS	161%	267%	317%	1061
CYLINDER LINERS	43%	37%	17%	173
DIESEL GENERATOR	193%	126%	158%	149
DRIVE SECTION	0%	43%	43%	32
EXHAUST VALVES	0%	0%	0%	18
FUEL INJECTION VALVES	304%	319%	345%	329
FUEL OIL PUMPS	257%	324%	194%	183
FUEL OIL PURIFIERS	28%	57%	69%	505
FUEL OIL SYSTEM	200%	357%	86%	31
LO SYSTEM	-	-	-	17
LUB OIL PURIFIERS	2%	8%	12%	729
LUBRICATING SYSTEM	-	-	-	11
MAIN BEARINGS	-	-	-	15
MAIN DIESEL GENERATORS	-	-	-	50
MAIN ENGINE	-	-	-	10
MECHANICAL SYSTEM	5%	10%	8%	325
PISTONS	121%	146%	168%	314
SHAFT ASSEMBLY	-	-	-	99
TURBO CHARGERS	0%	17%	0%	56

As can be seen from the table above, the forecasting results for some components is not satisfactory and the error metric is considerably higher than that in the forecasting for the nominal quantities. To better visualize the results of the forecasting process, indicative figures for each component are presented below.

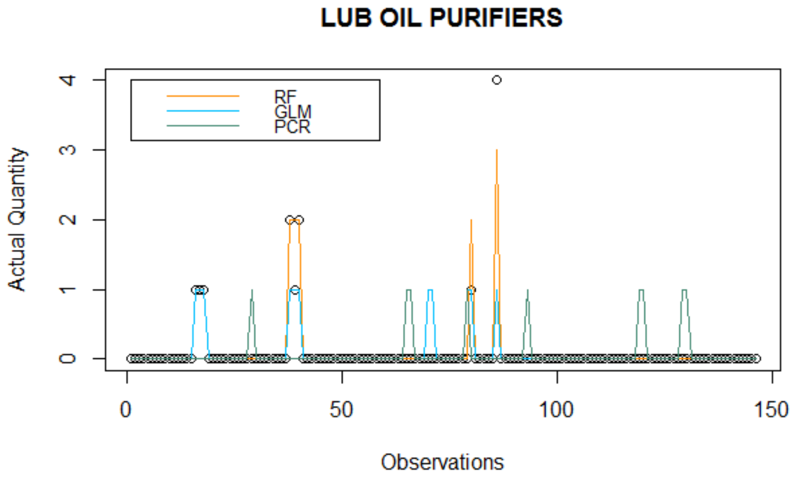


Figure 5.2-1: Forecasting results for the component LUB OIL PURIFIERS

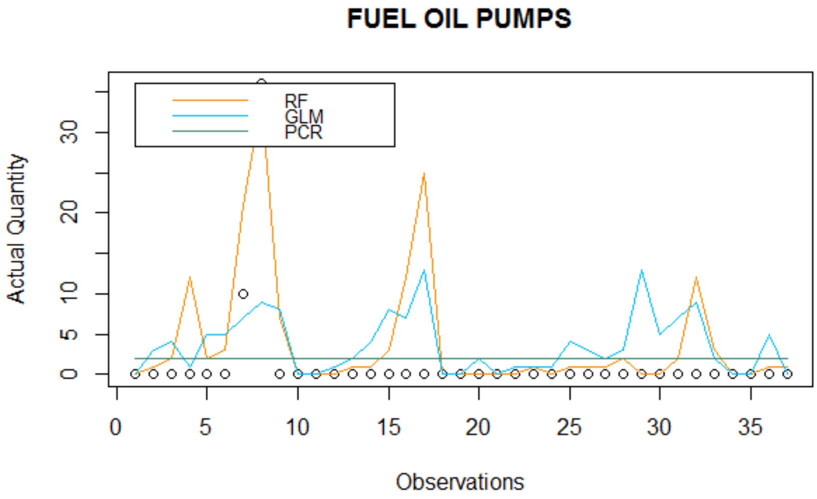


Figure 5.2-2: Forecasting results for the component FUEL OIL PUMPS

From the figures above, the need to better handle the dataset to produce better results is becoming clear. As previously discussed (see section 2.3.3), this part of the forecasting exercise will be used in the next section (see section 6) where the prescriptive model is constructed.

The main scope of the forecasting of the extra needs is to determine whether extra needs for the vessels are influenced by the initial source of purchase (maker or non-maker)

5.2.2 Data handling and forecasting results

The same behavior observed before is also observed in this section of the forecasting exercise.

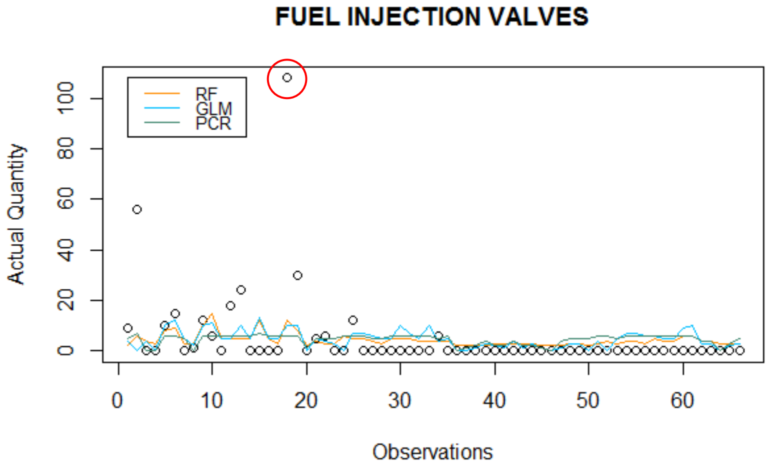


Figure 5.2-3: Forecasting results for the component FUEL INJECTION VALVES

What is interesting to be observed here, is the scale of the difference of the circled point. The quantity of all the points is below 40 and there is only one that is more than 100. This data point not only increases the error but also could be influencing the training of the algorithms and the rest of the results. The mean percentage absolute error for fuel injection valves is more than 300%.

Therefore, the same process as previously is employed, and handling of the outliers is performed.

Table 5.2-3: Mean absolute errors for extra needs for each component after outlier elimination

Component	RF	GLM	PCR	Number of Points
AIR COMPRESSORS	-	-	-	295
ASSEMBLY	-	-	-	6
CAMSHAFT	-	-	-	8
CONNECTING RODS	-	-	-	15
CONROD (BIG END) BEARINGS	-	-	-	10
CYLINDER HEADS	70%	181%	183%	971
CYLINDER LINERS	7%	13%	0%	153
DIESEL GENERATOR	6%	44%	48%	121
DRIVE SECTION	-	-	-	16
EXHAUST VALVES	-	-	-	0
FUEL INJECTION VALVES	88%	122%	127%	285
FUEL OIL PUMPS	14%	19%	17%	145
FUEL OIL PURIFIERS	16%	18%	6%	472
FUEL OIL SYSTEM	-	-	-	20
LO SYSTEM	-	-	-	17
LUB OIL PURIFIERS	0%	3%	6%	687
LUBRICATING SYSTEM	-	-	-	11
MAIN BEARINGS	-	-	-	6
MAIN DIESEL GENERATORS	-	-	-	50
MAIN ENGINE	-	-	-	10
MECHANICAL SYSTEM	3%	5%	2%	308
PISTONS	13%	17%	9%	293
SHAFT ASSEMBLY	-	-	-	99
TURBO CHARGERS	-	-	-	33

As per usual, below forecasting results are presented in figures to better visualize the results.

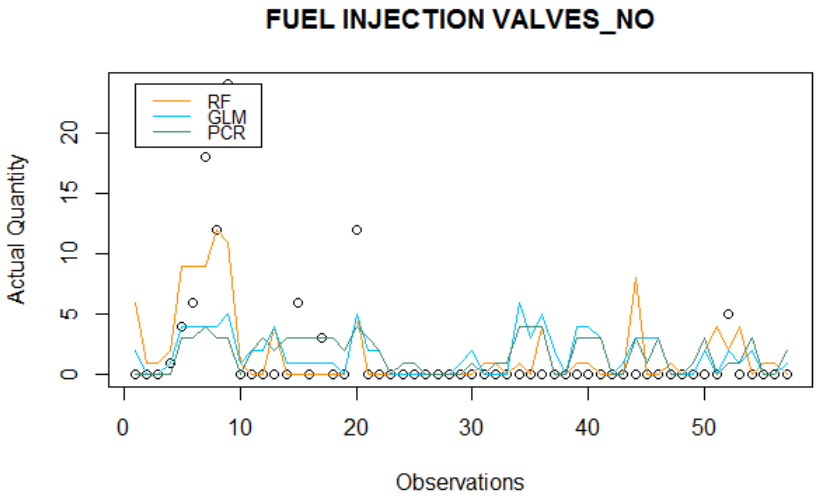


Figure 5.2-4: Results for forecasting of extra needs for component FUEL INJECTION VALVES after outlier handling

What needs to be noted here is that by eliminating in total 44 data points (both from the training and the test sets) the error decreases from 304% to 84%, which for the purposes of this business case can be considered acceptable. Some more indicative results are presented below:

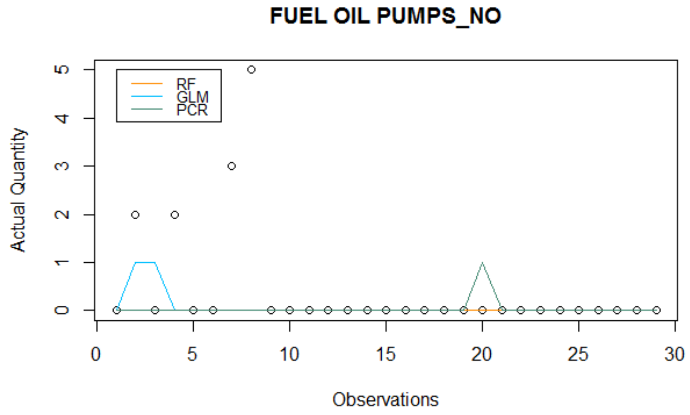


Figure 5.2-5: Results for forecasting of extra needs for component FUEL OIL PUMPS after outlier handling

For fuel oil pumps error was reduced from 257% to 14% with the elimination of 38 outliers.

By comparing the two tables that summarize the results of the two methods (see Table 5.2-2 and Table 5.2-3), the forecasting after outlier elimination seems to be a process that yields better results than simple forecasting. Indicatively the total average MAPE for the forecasting without outlier elimination is 88%[8] and after outlier elimination it decreases to 24%.

The forecasting of the extra quantities is performed only on a small number of components as the rest do not exhibit variations in the data and thus the completion of the training is very difficult. Especially after outlier handling the number of forecastable components drops even further, from 13 to 8. This section of the thesis is one of the areas that future work could be focused on.

5.3 Application on newly-released data

The bulk ordering process of the case company starts in April each year (see section 1.2), therefore the first phase of the process, where the requisitions are created, is an ideal case for the algorithms to be tested. For each component the qualifying algorithm out of the six previously mentioned models can be found below.

Table 5.3-1: Forecasting results for real case of the best performing algorithm

Component	MAPE	Method Used
AIR COMPRESSORS	36%	Random Forest No Outliers
ASSEMBLY	-	-
CAMSHAFT	35%	Principal Component Regression
CONNECTING RODS	62%	Random Forest No Outliers
CONROD (BIG END) BEARINGS	-	-
CYLINDER HEADS	105%	Random Forest No Outliers
CYLINDER LINERS	110%	Random Forest
DIESEL GENERATOR	66%	Random Forest No Outliers
DRIVE SECTION	78%	Principal Component Regression
EXHAUST VALVES	-	-
FUEL INJECTION VALVES	66%	Random Forest No Outliers
FUEL OIL PUMPS	106%	Random Forest No Outliers
FUEL OIL PURIFIERS	65%	Random Forest No Outliers
FUEL OIL SYSTEM	16%	Random Forest No Outliers

[8] This was computed excluding the forecast of the component CAMSHAFT as the error is extremely high and the results of the above will not be used further.

LO SYSTEM	-	-
LUB OIL PURIFIERS	111%	Random Forest
LUBRICATING SYSTEM	-	-
MAIN BEARINGS	-	-
MAIN DIESEL GENERATORS	78%	Generalized Linear Model
MAIN ENGINE	-	-
MECHANICAL SYSTEM	99%	Random Forest No Outliers
PISTONS	55%	Random Forest No Outliers
SHAFT ASSEMBLY	113%	Random Forest No Outliers
TURBO CHARGERS	5%	Random Forest No Outliers

For indicative components forecasting results are presented using graphs to better visualize the performance of the chosen algorithm.

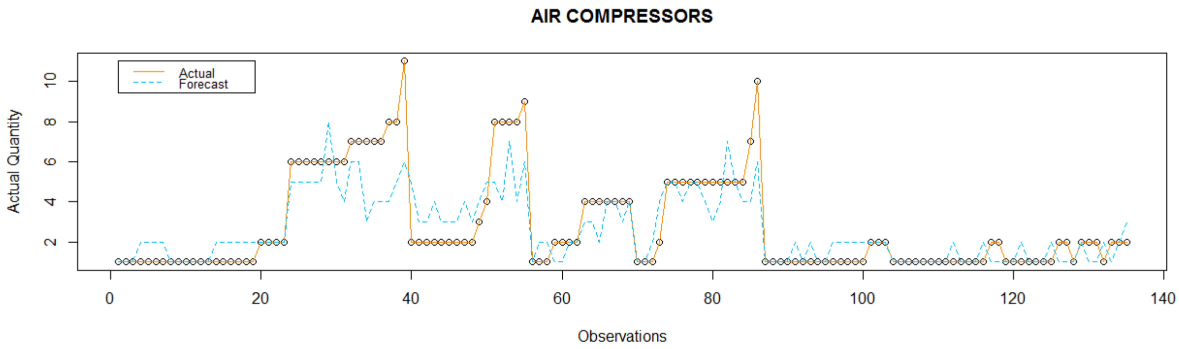


Figure 5.3-1: Forecasting of RQ quantities for BO 2020 for component AIR COMPRESSORS

For the forecasting of component air compressors, random forest with outlier handling was used. The error in the training was 31% and the actual error was 36%.

In general, the average error on the newly released data does not deviate significantly from the average error in the training set. However, it should be kept in mind that the quantities that are used to test the performance of the algorithms on the newly released data are the initial quantities of the bulk order process. As it was explained in previous sections (see section 1.2) the initial quantities are reduced or increased based on the company's decisions regarding necessary equipment.

As further analysis, all six algorithms were applied to the data to evaluate if the best algorithm is indeed the best fit. The results are presented below. On the table below, the minimums have been marked with bold.

Table 5.3-2: Results of all methods on real data

Component	RF	GLM	PCR	RF_NO	GLM_NO	PCR_N0
AIR COMPRESSORS	**36%**	113%	109%	36%	108%	94%
ASSEMBLY	-	-	-	-	-	-
CAMSHAFT	49%	75%	**35%**	-	-	-
CONNECTING RODS	**54%**	179%	175%	62%	89%	71%
CONROD (BIG END) BEARINGS	-	-	-	-	-	-
CYLINDER HEADS	116%	184%	195%	**105%**	298%	299%
CYLINDER LINERS	110%	200%	228%	**70%**	224%	187%
DIESEL GENERATOR	100%	151%	168%	**66%**	86%	109%
DRIVE SECTION	**64%**	189%	78%	-	-	-
EXHAUST VALVES	-	-	-	-	-	-
FUEL INJECTION VALVES	89%	102%	110%	**66%**	66%	73%
FUEL OIL PUMPS	168%	174%	174%	**106%**	132%	130%
FUEL OIL PURIFIERS	83%	122%	128%	**65%**	99%	104%

FUEL OIL SYSTEM	24%	65%	29%	**16%**	90%	29%
LO SYSTEM	-	-	-	-	-	-
LUB OIL PURIFIERS	111%	156%	148%	**78%**	144%	144%
LUBRICATING SYSTEM	-	-	-	-	-	-
MAIN BEARINGS	-	-	-	-	-	-
MAIN DIESEL GENERATORS	30%	78%	47%	**26%**	67%	41%
MAIN ENGINE	-	-	-	-	-	-
MECHANICAL SYSTEM	153%	197%	175%	**99%**	150%	123%
PISTONS	112%	304%	198%	**55%**	167%	248%
SHAFT ASSEMBLY	**70%**	163%	165%	113%	8492%	9130%
TURBO CHARGERS	47%	57%	48%	**5%**	71%	62%

By comparing the two tables, it is observed that for most components the best performing algorithm on the test data is also the best performing for the newly released data. However, as can be seen by comparing Table 5.3-2 with Table 5.3-1 there are some components that the algorithm identified as the best performing is not. As previously stated the quantities of the newly released data are not the finalized purchased quantities, therefore explaining some of the deviations in the results. At the same time, it needs to be kept in mind that for most components the data available are not sufficient enough to capture precisely the complex nature of the problem therefore it is expected, especially in the first years of implementation that such deviations arise. However, it is expected that as the data set grows and additional dimensions (i.e. variables) are added the accuracy of the results will improve. The components that have different best performing algorithms for the train set and for the newly released date can be seen below:

Table 5.3-3: Components for which the test and the real set have different behavior

Component	**Error in train**	**Best Performing on train**	**Error in new data**	**Best Performing on new data**
CONNECTING RODS	62%	Random Forest No Outliers	54%	Random Forest
CYLINDER LINERS	110%	Random Forest	70%	Random Forest No Outliers
DRIVE SECTION	78%	Principal Component Regression	64%	Random Forest
LUB OIL PURIFIERS	111%	Random Forest	78%	Random Forest No Outliers
SHAFT ASSEMBLY	113%	Random Forest No Outliers	70%	Random Forest

For the components of the table above, in the following sections both algorithms were applied, and the average was taken as the final forecasted quantity.

The forecasting of the nominal needs of the vessels exhibits satisfactory results (average MAPE 53%) and could, in the future when the training samples increase, become more and more accurate. For some specific components that show increased accuracy, e.g. fuel oil system (MAPE = 10%) the tool can be used to expedite the process while decreasing the workload both for the vessel and for the shore- based engineers. However, the forecasting of the extra needs does not yield such results. The average MAPE is increased compared to the forecasting of the nominal quantities while the number of components upon which forecasting is applied decreases.

The next methodological step of the thesis encompasses the results of the above forecasting exercises to create the final objective function that leads to optimum allocation of items to vendors so as to minimize total cost of the bulk orders.

This page has been intentionally left blank

6 Prescriptive cost optimization model

The final step of the analysis is the creation of the main product of the thesis, the prescriptive model. The prescriptive model ties in the entire bulk order analytics framework and gears it in the decision support domain by serving as a guideline on the optimal cost basis of spare parts procurement The model will allow the case company to determine whether each spare part should be ordered more times than the nominal need of the vessel and whether it should be bought from maker or from the parallel market.

For the prescriptive model data from the dataset used for the forecasting of nominal bulk quantities for the year 2018 will be used.

6.1 Identification of cost analysis parameters

The cost function will be the objective function of this optimization problem. Therefore, the components of this function are of great importance to the result.

The components of the cost function have been identified as of below

- Acquisition cost: it represents the cost of purchase for each item. It depends on the total quantities and on the acquisition price of each item. What needs to be noted here is that for the two main categories of suppliers, makers and non-makers, the acquisition price changes considerably.

  $$Acquisition\ Cost = (Nominal\ Needs + Extra\ Quantities\ + Safety\ Stock) * Acquisition\ Price$$

  Safety stock: depending on the desired service level (SL) the level of the safety stock will be determined. The safety stock will also be added to the acquisition and forwarding cost as it is assumed that both the target inventory and the safety stock are bought together, considering that price fluctuations in the spare parts are not high. The safety stock follows the formula below.

  $$\text{Safety Stock} = \text{Z} * \sqrt{\text{Avg LT} * (\text{st dev of demand})^2 + \text{Avg demand} * (\text{st dev LT})^2}$$

  where LT is the lead time, Z is the inverse distribution function of a standard normal distribution with cumulative probability of the underlying service level and demand refers to the historical demand of the relevant item. Both for the lead time and for the average demand there are more than 30 observations therefore by the central limit theorem it can be said that these variables satisfy the underlying assumptions (i.e. normal distribution) of the above formula.
- Forwarding cost: this cost component represents the cost of the transportation of each item on board the vessel. This cost depends on several parameters such as the location of the supplier, the trading route of the vessel, any specific requirements for clearance etc. For the purposes of this analysis it is assumed that the forwarding cost depends mainly on the lead time. As of current situation in the market, there are two locations that supply ship spare parts and can cover the needs of the overhauls which are accumulated in the bulk orders: Europe and Korea. Around 40% of makers are in Europe and transportation costs to Rotterdam, the main logistic hub of the company, for the spare parts located in Europe is

assumed to be zero, as the European makers use groupage trucks on their account to transfer the spare parts. For the rest of the items, we will follow the function below.

$$\text{forwarding cost} = \begin{cases} \dfrac{\text{weight of order}}{\text{total weight in container}} * \text{container rental rate}, & \text{lead time} < 30 \\ \text{weight of order} * \text{airfreight rate} \quad , & \text{lead time} \geq 30 \end{cases}$$

Therefore, according to the above the final formula for the forwarding cost is the below:

$$\text{final forwarding cost} = \begin{cases} (1 - 0.4) * \text{forwarding cost}, & \text{market} = \text{Maker} \\ \text{forwarding cost}, & \text{market} = \text{Parallel} \end{cases}$$

- Inventory Cost: this cost component represents the costs that are incurred because of the inventory held on the vessel. The inventory cost follows the simple formula below

$$Inventory\ Cost = \left(SS + \frac{TI}{2}\right) * Acquisition\ Price * WACC$$

where: SS is the safety stock and TI is the target inventory where

$$Target\ Inventory = Nominal\ Needs + Extra\ Quantities$$

- Stock out cost: this cost component represents the costs that are incurred when an item that should have been on board the vessel is not. It is computed using the formula below:

$$Stock\ out\ cost = \frac{100 - SL}{2} * Additional\ Cost$$

where

$$Additional\ Cost = Acquisition\,' + Forwarding\,' + Administrative'$$

What needs to be noted here is that the components of the additional cost are significantly higher than the respective costs during a normal/ routine ordering process. Therefore, the three components of the additional cost function will be increased by a factor.

6.2 Analysis

6.2.1 Acquisition Cost

The acquisition cost is the first cost component that will be discussed. The main components of the acquisition cost are the total quantities and the purchase price. Both of those components will be analyzed below.

6.2.1.1 Total Quantities

To determine the level of the nominal needs for each maker reference the models of the previous section (see section 5) are being used. The dataset is structured in the same way as the dataset that was used for training and testing the models of the previous chapter (see section 5). The best performing model (the one having the smallest error) was used (see section 5.1.3)

Table 6.2-1: Components that each algorithm is applied on

Algorithm	Components
Random Forest No Outliers	AIR COMPRESSORS, CONNECTING RODS, CYLINDER HEADS, DIESEL GENERATOR, FUEL INJECTION VALVES, FUEL OIL PURIFIERS, FUEL OIL PUMPS, FUEL OIL SYSTEM, LO SYSTEM, MECHANICAL SYSTEM, TURBO CHARGERS
Random Forest	CYLINDER LINERS, LUB OIL PURIFIERS, PISTONS
Principal Component Regression	CAMSHAFT, DRIVE SECTION
Generalized Linear Model	MAIN DIESEL GENERATORS

To determine the extra quantities for the fleet, the datasets (also including the forecasted data for nominal needs) are duplicated. The first dataset will have the bulk market equal to Maker thus making the indirect assumption that all the items were purchased though the original market and the second will be created assuming that all items were purchased through the parallel market.

This was performed to determine and highlight the differences in the quantities that need to be purchased as influenced by the source of purchase.

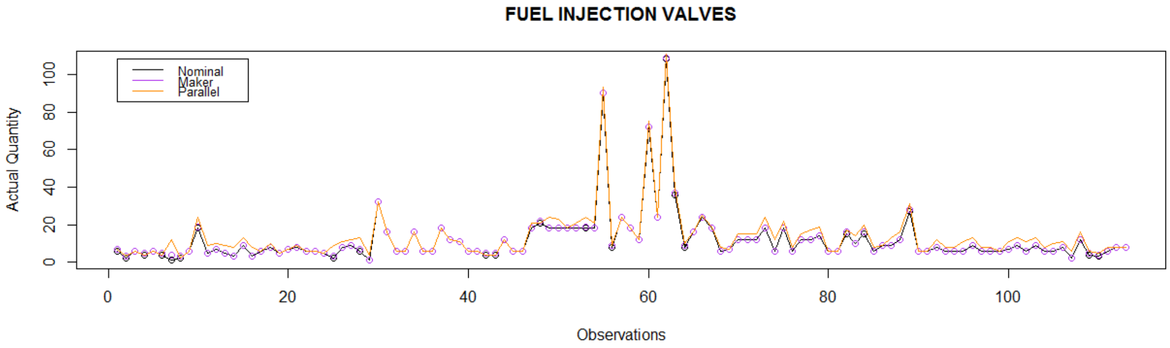


Figure 6.2-1: Extra needs depending on the market for FUEL INJECTION VALVES

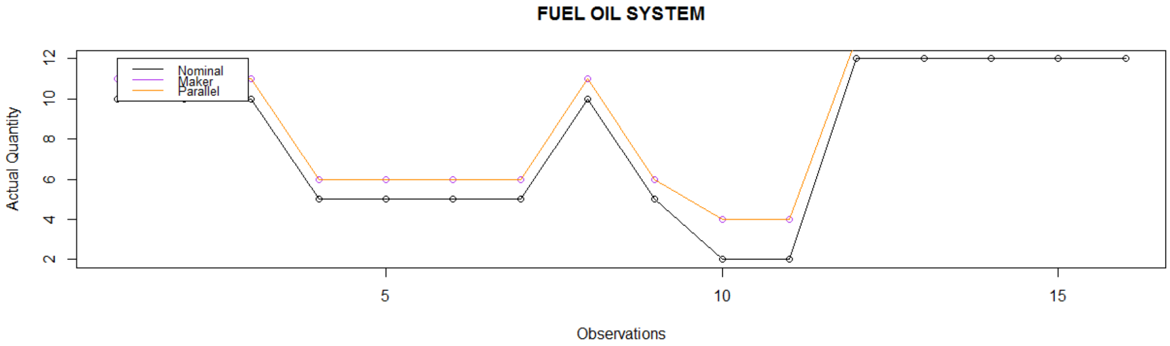


Figure 6.2-2: Extra needs depending on the market for FUEL OIL SYSTEM

The next step was to compute the safety stock for each item. As per previous section the formula of the safety stock is the below

$$\text{Safety Stock} = Z_{SL} * \sqrt{\text{Avg LT} * (\text{st dev of demand})^2 + \text{Avg demand} * (\text{st dev LT})^2}$$

where LT is the lead time computed in the following pages and Z_{SL} is the inverse distribution function of a standard normal distribution with cumulative probability of the service level (SL).

The average demand and the standard deviation of demand were computed regardless of the market using past data.

The lead time was derived from past data for each category for makers and for parallel market. To capture the necessary time for delivery on board the production lead time was increased by 5 days.

Table 6.2-2: Lead times per component for each market category

Component	Parallel Lead Time	Maker Lead Time
AIR COMPRESSORS	30	8
ASSEMBLY	21	5
CAMSHAFT	30	15
CONNECTING RODS	30	7
CONROD (BIG END) BEARINGS	30	15
CYLINDER HEADS	24	13
CYLINDER LINERS	16	45
DIESEL GENERATOR	27	22
DRIVE SECTION	0	5
EXHAUST VALVES	20	35
FUEL INJECTION VALVES	15	16
FUEL OIL PUMPS	21	15
FUEL OIL PURIFIERS	1	5
FUEL OIL SYSTEM	18	33
LO SYSTEM	0	5
LUB OIL PURIFIERS	0	5
LUBRICATING SYSTEM	30	15
MAIN BEARINGS	30	15
MAIN DIESEL GENERATORS	23	13
MAIN ENGINE	30	41
MECHANICAL SYSTEM	0	15
PISTONS	26	29
SHAFT ASSEMBLY	0	5
TURBO CHARGERS	30	0

Lastly, to account for the forecast errors of the previous models, that sometimes are significant, the following procedure is used.

- For each component the forecast bias is computed, and it is determined whether there is an over-forecasting or an under-forecasting bias
- If there is an over-forecasting bias, then the safety stock computed is multiplied by the accuracy of the forecast of nominal quantities

The forecast bias is computed using the following formula

$$forecast\ bias = \frac{forecast - actual}{forecast + actual}$$

The above, is sometimes called the normalized forecast metric and is broadly used to compute the bias. As can be seen, the metric ∈ [-1, +1] where 0 indicates the absence of forecast bias. Negative values show a tendency to under-forecast and positive values to over-forecast.

In a business sense, the safety stock is needed to cover needs that cannot be covered by the nominal demand. However, if the demand has been forecasted with a method that indicates over -forecast bias the final quantity that will be purchased will be unnecessarily high. This reasoning explains the final formula of the safety stock:

$$\text{Safety Stock} = \begin{cases} \text{safety stock} * \text{accuracy}, & \text{forecast bias} > 0 \\ \text{safety stock}, & \text{forecast bias} \leq 0 \end{cases}$$

To visualise the above thinking the diagram below is presented.

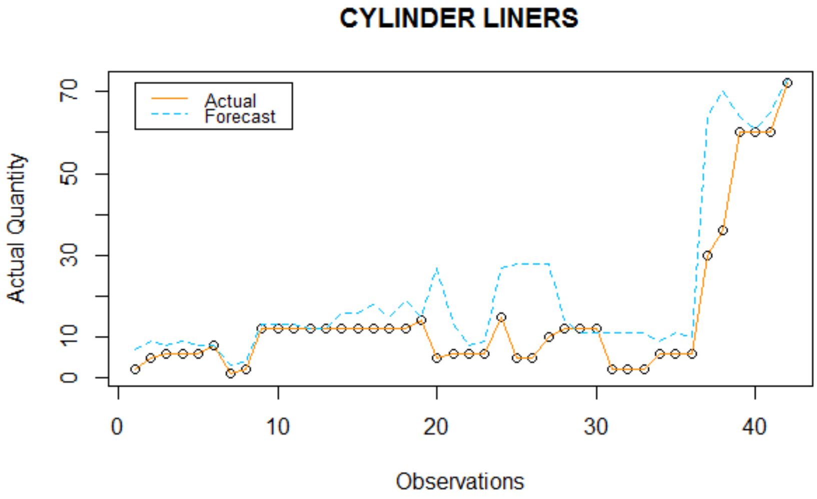


Figure 6.2-3: Example of positive forecast bias in the component CYLINDER LINERS

The forecast for some components (as the one illustrated in the figure above) tends to overestimate the quantities that will be needed. Therefore, the over forecasted quantities can be used as safety stock. This will avoid over-stocking the vessels with unnecessarily high quantities of items that have been forecasted with methods that exhibit high positive forecast bias.

To determine the optimum service level of each time and market exhaustive enumeration was used. Random service levels were used to compute the total cost of the items and the service level having the minimum total cost was identified as the optimum service level and was used in the final step of the prescriptive model. The random service levels were chosen in the range of 95% to 99.9%. As the items ordered in the bulk process are critical for the smooth running of machinery, this range was chosen mainly for business and technical reasons.

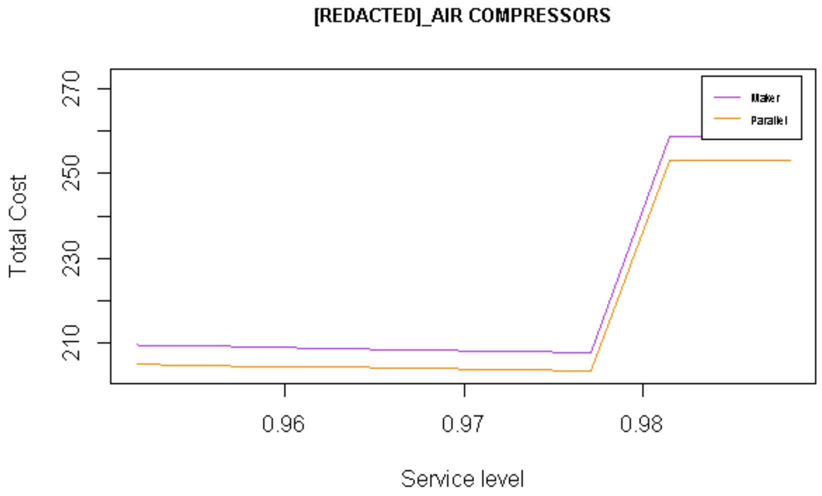


Figure 6.2-4: Cost of item of item [REDACTED] of component AIR COMPRESSORS as a function of the service level

As can be observed from the graph, the minimum cost is achieved for the same level 97.71% for both maker and non-maker. In the figure above, for SL ∈ [95%, 97.71%] the total cost decreases slightly and for SL ∈ [97.71%, 99.9%] the cost is increased while the slope remains the same. To understand this behavior below graph (total quantity to service level) is created. As one can see from the graph above there is a high fluctuation of the total cost (around 30%) as the service level changes, highlighting the need to determine the optimum service level.

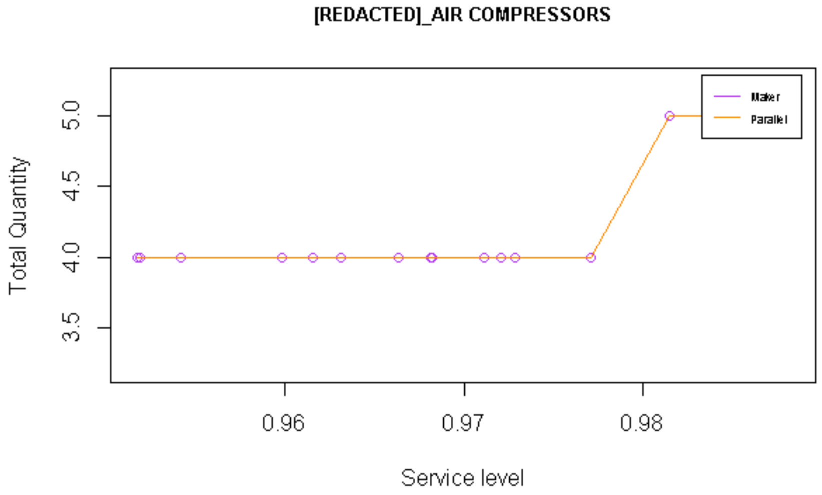


Figure 6.2-5: Quantity of item [REDACTED] of component AIR COMPRESSORS as a function of the service level

The figure above visualizes the change of the total quantity purchased as the service level increases. As can be observed from the total quantities for maker and non-maker are the same. Using this graph, the change of the total cost can be interpreted more easily. The total quantity drives the increase as the service level dictates purchase of one additional quantity therefore driving the acquisition, forwarding and inventory costs up. Below another example is given:

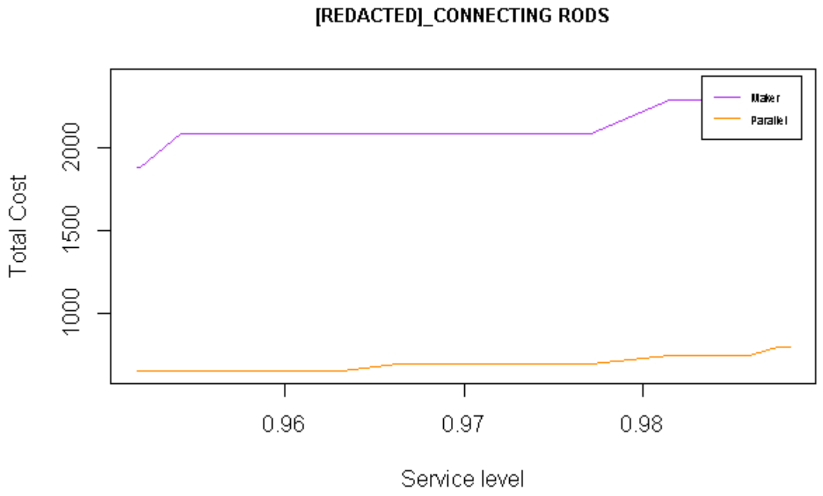


Figure 6.2-6: Cost of item [REDACTED] of category CONNECTING RODS as a function of the service level

For this item the optimum service level is different for the maker (maker optimum service level = 95.18%) and for the non-maker (parallel optimum service level = 96.31%). Below the quantities of this item as a function of the service level are also disclosed.

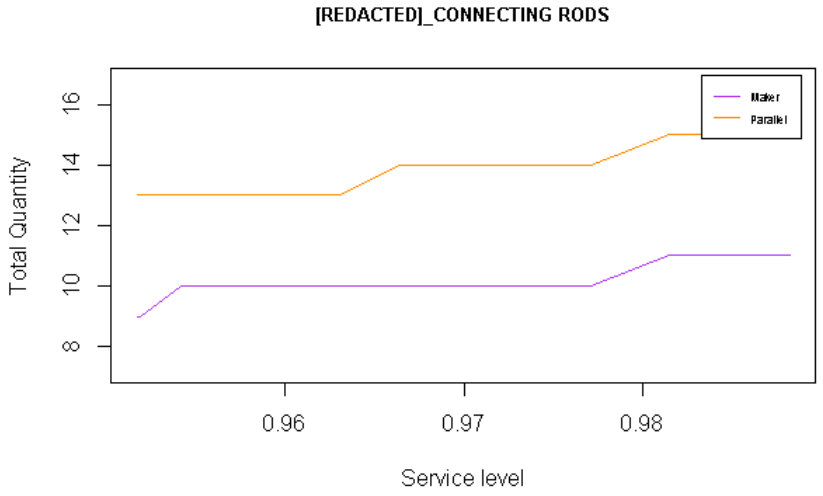


Figure 6.2-7: Total quantity of item [REDACTED] of category CONNECTING RODS as a function of the service level

6.2.2 Acquisition Price

To calculate the acquisition cost of each item the prices of the items depending on the market need to be determined. To simplify the procedure the comparable items (same items purchased in both bulks) of bulk orders 2018 and 2019 were extracted and for them, the prices were compared for each category (in the dataset labelled as category).

Figure 6.2-8: Acquisition cost per category and per market

Market_19	Market_18	Category	AVG_(Price19/Price18)	MAX_(Price19/Price18)	MIN_(Price19/Price18)
Maker	Parallel	D/G	4.39	52.08	0.23
Maker	Parallel	M/E	3.61	17.24	0.71
Maker	Parallel	Compressor	0.92	4.83	0.16
Maker	Parallel	Purifier	1.12	5.49	0.26

Using the above table, the price on record is multiplied or divided by the relevant entry in the column labeled 'AVG_ (Price19/Price18)' that practically expresses the difference of prices between makers and non-makers. Therefore, the prices depending on the market are extracted. To accommodate for randomness and for changes in prices the market prices of the items were then multiplied by a factor following the continuous distribution with a minimum of 0.8 and a maximum of 1.2.

The elimination of assumptions regarding the prices of the two market categories can be achieved by making the prescriptive model vendor specific. This would improve accuracy in the forecasting part and efficiency in the decision-making part as sometimes the difference in pricing of vendors, even if they belong in the same category, can be significant. At the same, discrepancies between average historical prices and average market prices and/ or lead times can be omitted, thus making the model more efficient.

6.2.3 Forwarding Cost

The forwarding cost is the next cost component that will be examined. As previously discussed (6.1), to determine the forwarding cost approximates the weights of the orders. The weight of the items depends on the category that they belong. Those are presented in the table below.

Table 6.2-3: Components and Categories

Component	Category [1]	Category [2]
AIR COMPRESSORS	AIR COMPRESSORS	-
ASSEMBLY	MAIN ENGINES	MAIN DIESEL GENERATORS
CAMSHAFT	MAIN DIESEL GENERATORS	-
CONNECTING RODS	MAIN DIESEL GENERATORS	-
CONROD (BIG END) BEARINGS	MAIN DIESEL GENERATORS	-
CYLINDER HEADS	MAIN DIESEL GENERATORS	-
CYLINDER LINERS	MAIN ENGINES	MAIN DIESEL GENERATORS
DIESEL GENERATOR	MAIN DIESEL GENERATORS	-
DRIVE SECTION	LUB OIL PURIFIERS	-
DRIVE SECTION	FUEL OIL PURIFIERS	-
EXHAUST VALVES	MAIN ENGINES	-
FUEL INJECTION VALVES	MAIN ENGINES	MAIN DIESEL GENERATORS
FUEL OIL PUMPS	MAIN ENGINES	MAIN DIESEL GENERATORS
FUEL OIL PURIFIERS	FUEL OIL PURIFIERS	-
FUEL OIL SYSTEM	MAIN DIESEL GENERATORS	-
LO SYSTEM	LUB OIL PURIFIERS	-
LUB OIL PURIFIERS	LUB OIL PURIFIERS	-
LUBRICATING SYSTEM	MAIN ENGINES	MAIN DIESEL GENERATORS
MAIN BEARINGS	MAIN DIESEL GENERATORS	-
MAIN DIESEL GENERATORS	MAIN DIESEL GENERATORS	-
MAIN ENGINE	MAIN ENGINES	-
MECHANICAL SYSTEM	FUEL OIL PURIFIERS	LUB OIL PURIFIERS
MECHANICAL SYSTEM	MAIN ENGINES	-
PISTONS	MAIN DIESEL GENERATORS	MAIN ENGINES
SHAFT ASSEMBLY	FUEL OIL PURIFIERS	LUB OIL PURIFIERS
TURBO CHARGERS	MAIN ENGINES	MAIN DIESEL GENERATORS

For each category the average weight is shown below

Table 6.2-4: Average weight of items per category

Category	Weight per Item [kg]
MAIN DIESEL GENERATORS	2.8
MAIN ENGINES	8.9
LUB OIL PURIFIERS, FUEL OIL PURIFIERS	3.7

For the component air compressors, there is no available data and therefore the weight of the items needs to be approximated using other methods. After conducting interviews with superintendent engineers and experienced spare part operators of the case company, average weight is assumed around 3 kgs per item.

If a component belongs to more than one category, then the average of the two is taken as the weight. Therefore, the forwarding cost follow the following formula:

$$\text{forwarding cost} = \begin{cases} \dfrac{\text{weight per item} * \text{quantity}}{\text{total weight in container}} * \text{container rental rate}\,, & \text{lead time } < 30 \\ \text{weight per item} * \text{quantity} * \text{airfreight rate} \qquad , & \text{lead time} \geq 30 \end{cases}$$

The forwarding cost must be separated into two categories:

- Forwarding cost for routine orders. The routine last mile cost is low because in a routine shipment there is usually more than 10 orders being shipped and therefore the costs are being allocated to a high number of orders. Additionally, when routine shipment is arranged the port index, the corresponding price index of a port, is considered.
- Forwarding cost for unplanned orders. In this case, the two factors above are not being considered as the spare part needs to reach the vessel regardless of its location and the port index is not considered.

Further analysis for the forwarding cost for unplanned orders will take place in the stock out cost analysis section (see section 6.2.5).

6.2.4 Inventory Cost

As previously discussed the inventory cost follows the equation below

$$Inventory\ Cost = \left(SS + \frac{TI}{2}\right) * Acquisition\ Price * \ WACC$$

, where WACC is the weighted average cost of capital which is the average interest rate of the case company which for privacy reasons is redacted.

As further analysis for this parameter a sensitivity analysis will take place at the next section. For the purposes of this analysis the WACC ∈ [3%, 8%]. As reference, the WACC of the maritime industry is 7.05% (NYU Stern, 2019).

6.2.5 Stock- out Cost

This cost component, as previously discussed, represents the cost of re-supplying the ship with the item on an urgent level if the existing stock of the vessel runs out.

As one can easily understand, this cost will be increased compared to the previous cost mainly for the reasons below:

- When a requisition is made on an urgent basis there is no time to receive quotations from several vendors or to make price negotiations with them as can be done in the bulk ordering process.
- When a requisition is made on an urgent basis, the selection is mainly made considering the lead time and not the price, therefore there is a possibility that the price is higher than usual.
- When a requisition is made on an urgent basis, the forwarding cost of the shipment can be extremely higher both because of inconvenient delivery port and because of absence of other orders (meaning that the fixed costs of the shipment are not being allocated to many items).

Therefore, the formula of the stock out cost is as per below

$$Stock\ out\ cost = (Probability\ of\ Stockout) * \left(forwarding\ cost^{'} + acquisition\ cost^{'} + administrative\right)$$

where,

$$Acquisition\ Cost\ ' = \frac{a * Price_{maker} + (1-a) * Price_{parallel}}{2} * \ Stock\ out\ order\ quantity$$

where a is the percentage of the times that on a spot basis the maker is chosen. As per analysis of the case company's buying patterns a= 40% and,

$$Stock\ out\ order\ quantity = (Safety\ Stock_{service\ level\ 99.9\%} - Safety\ Stock_{of\ selected\ SL}\ )$$

and,

$$Probability\ of\ Stockout = (0.999 - Selected\ SL)$$

where SL is the service level for each item.

$$Forwarding\ Cost' = \ Stock\ out\ order\ quantity * avg\_weight * airfreight\ cost * urgency$$

where the urgency factor of the equation above models the increased forwarding cost of an unplanned event.

For calculating the urgency factor the following simple equations were used.

$$\begin{cases} Routine\ Cases_1 * Routine\ Cost + Unplanned\ Cases_1 * Unplanned\ Cost = Cost_1 \\ Routine\ Cases_2 * Routine\ Cost + Unplanned\ Cases_2 * Unplanned\ Cost = Cost_2 \end{cases}$$

To compute Cost$_1$ and Cost$_2$ the budgeting tool of the company was used which will not be disclosed for privacy reasons.

Therefore, the urgency can be defined as

$$urgency = \frac{Unplanned\ Cost}{Routine\ Cost}$$

For the fleet total it was found that urgency ∈ [1.2, 2.5]

Since a more precise way to define the urgency would be out of scope of this thesis, a sensitivity analysis will take place in the next section to explore the way that the urgency coefficient influences the model.

Finally, to compute the administrative cost of an unplanned case it is assumed that each unplanned case is an order. To compute the administrative cost of an order, interviews with spare parts operators of the case company were conducted. At the same time, by retrieving data regarding the number of orders completed each year by each operator, it was found that the total administrative cost of an order is 19USD.

6.3 Synthesis of results

Using the default values for the two variable parameters of the problem as below

- WACC = [redacted]
- Urgency Coefficient = 2

, the below table is produced.

The column maker percentage presents the percentage of entries for which the total cost of maker is less than the total cost of the parallel market, and thus the maker is chosen as the source of purchase.

Table 6.3-1: For the default values of the parameters maker percentage for each component

Component	Maker Percentage	Total Acquisition for Makers	Total Acquisition for Parallel
AIR COMPRESSORS	83%	$ 82,836.16	$ 35,874.97
ASSEMBLY	-	-	-
CAMSHAFT	0%	$ -	$ 62.82
CONNECTING RODS	0%	$ -	$ 41,602.46
CONROD (BIG END) BEARINGS	-	-	-
CYLINDER HEADS	39%	$ 4,586.61	$ 112,763.44
CYLINDER LINERS	4%	$ -	$ 42,856.87
DIESEL GENERATOR	11%	$ 43.75	$ 10,068.39
DRIVE SECTION	91%	$ 496.69	$ 77.83
EXHAUST VALVES	-	-	-
FUEL INJECTION VALVES	1%	$ 19.17	$ 97,703.79
FUEL OIL PUMPS	0%	$ -	$ 68,889.19
FUEL OIL PURIFIERS	57%	$ 15,952.30	$ 2,779.68
FUEL OIL SYSTEM	0%	$ -	$ 30.70
LO SYSTEM	14%	$ 1.56	$ 7.16
LUB OIL PURIFIERS	58%	$ 6,966.83	$ 44.04
LUBRICATING SYSTEM	-	-	-
MAIN BEARINGS	-	-	-
MAIN DIESEL GENERATORS	33%	$ 128.81	$ 1,105.06
MAIN ENGINE	-	-	-
MECHANICAL SYSTEM	2%	$ 709.56	$ 38,875.61
PISTONS	0%	$ -	$ 226,474.99
SHAFT ASSEMBLY	44%	$ 111.75	$ 5,137.98
TURBO CHARGERS	0%	$ -	$ 36,175.98

A sensitivity analysis is then performed for both the WACC and for the urgency coefficient.

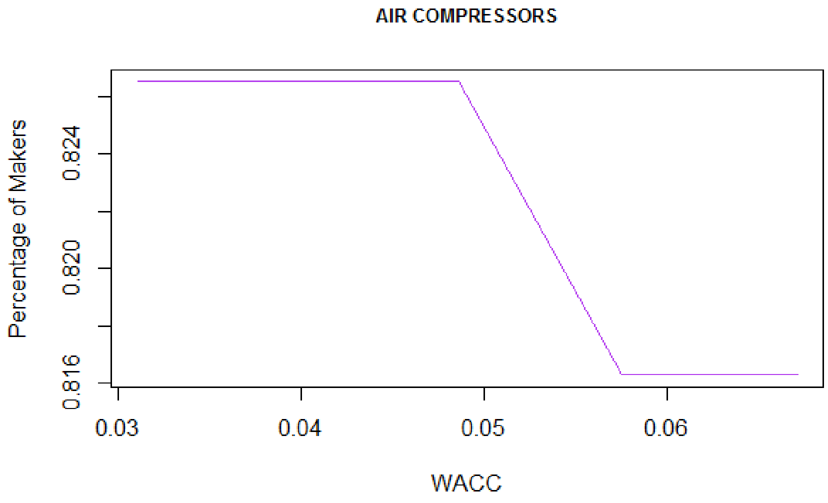


Figure 6.3-1: Results of sensitivity analysis for WACC variable for component AIR COMPRESSORS

The graphs above have been produced while keeping the urgency coefficient constant at 2. The graph visualizes the influence of WACC on the percentage of entries that the maker is selected as a source of purchase.

As can be seen the WACC does not influence heavily the selection process. The total change in the selection is 1% (total size of the dataset: 98) which means that the selection switches from maker to non-maker for only one item when the WACC increases. This derives from the increased inventory cost for said item due to high price of maker as compared to that of the non-maker.

The change in total gross revenue from makers is small and is depicted in the graph below.

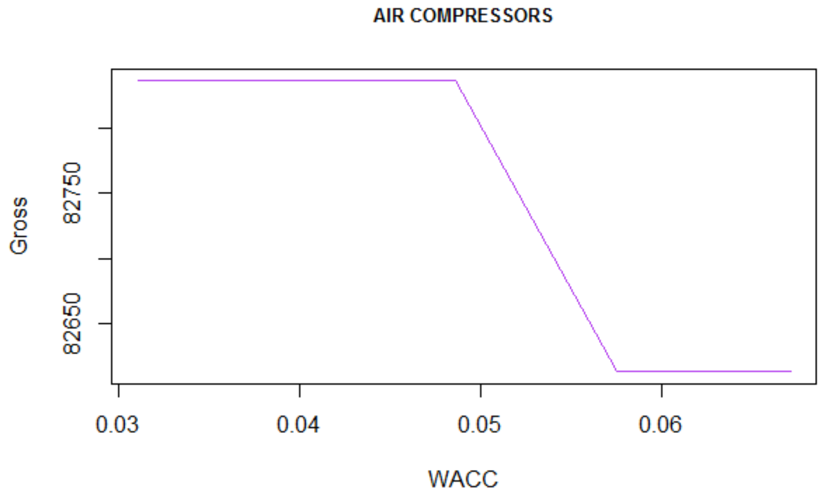


Figure 6.3-2: Change in total gross revenue of makers as a function of WACC

As can be seen from the graph above the total gross revenue for makers is decreased by less $200 (less than 0.5% of total gross revenue). Therefore, it is safe to conclude that for the component air compressors the WACC does not heavily influence the selection process.

A further analysis on the influence of the WACC on important components of the cost function is conducted and the results can be seen in the graph below.

The graph includes:

- Change of the total quantity purchased
- Change of the total stock out cost
- Change of the total acquisition cost
- Change of the total cost

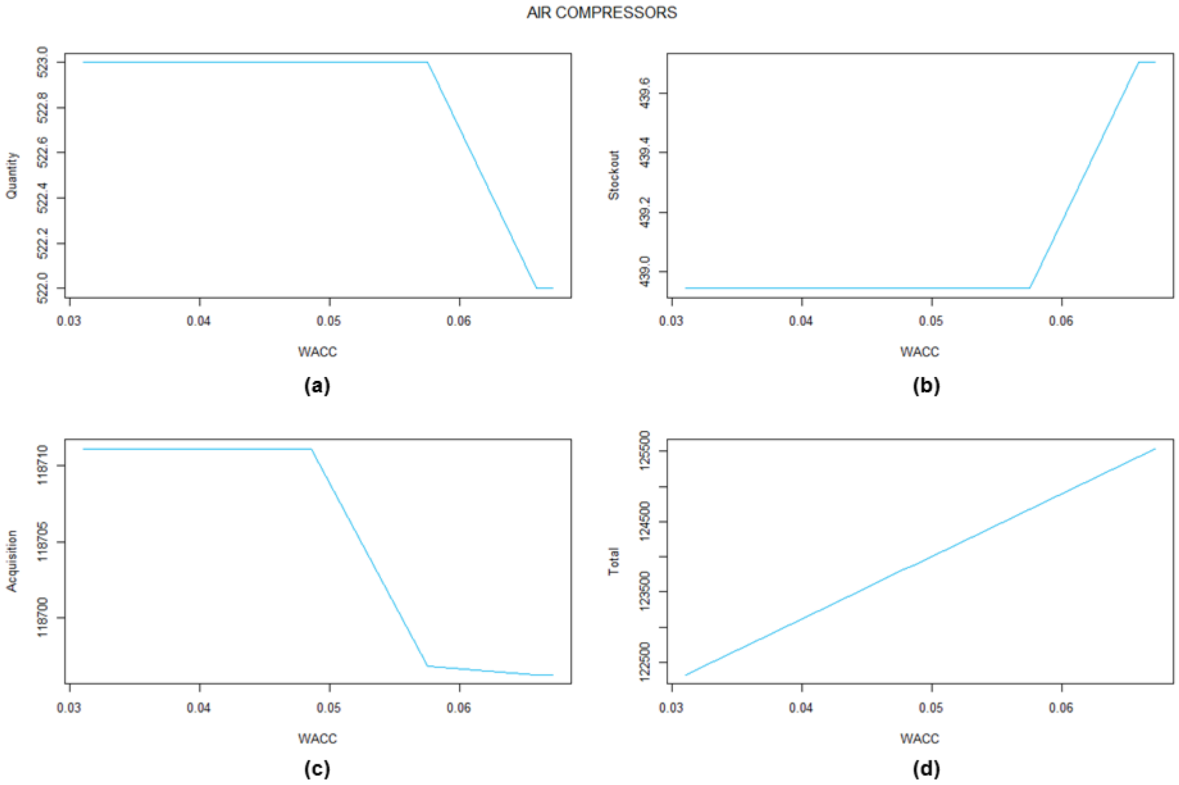


Figure 6.3-3: Results of sensitivity analysis for important cost components

It is important to be noted that the above graph was produced while keeping the urgency coefficient constant at 2.

In the above graph (a) shows the change of the quantity due to the increased WACC. As can be seen the total quantity procured decreases thus also decreasing the acquisition cost (c) and increasing the stock out (b). In the last graph (d), the increase of the total cost is increased due to the increased inventory cost.

In the next section the influence of the urgency coefficient is shown.

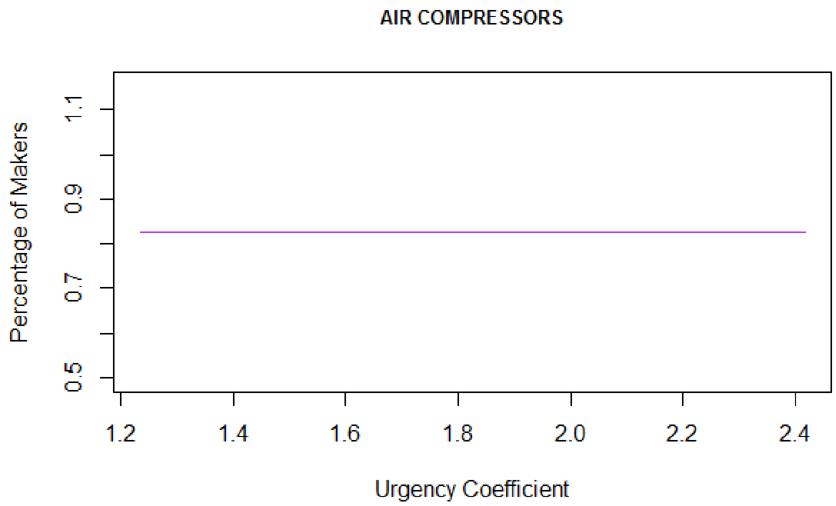


Figure 6.3-4: Total acquisition cost to makers for component AIR COMPRESSORS as a function of the urgency coefficient

As is seen in the graph above, the urgency coefficient does not influence the selection process for the component air compressors.

However, in the graph below it is observed that there is a slight increase in the total gross revenue by makers, that since no percentage change is exhibited means that the quantities procured are increased.

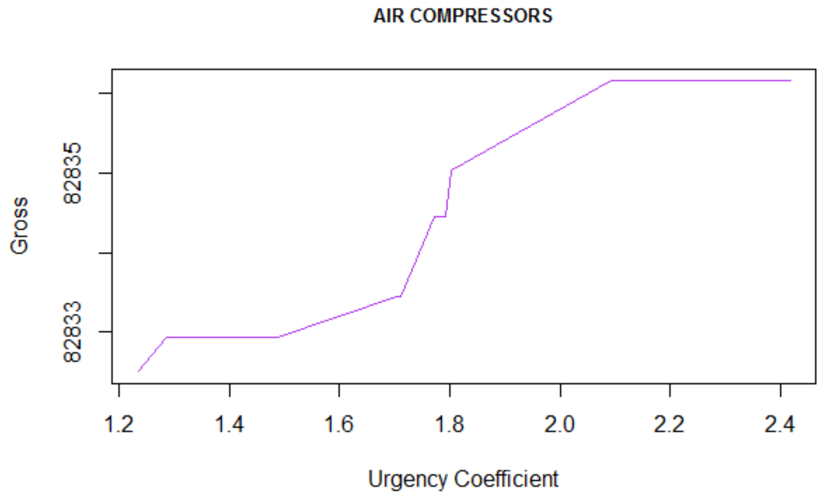


Figure 6.3-5: Change in total gross revenue of makers as a function of the urgency coefficient

A further analysis on the influence of the urgency coefficient on important components of the cost function is performed as in the previous section and the results can be seen in the graph below.

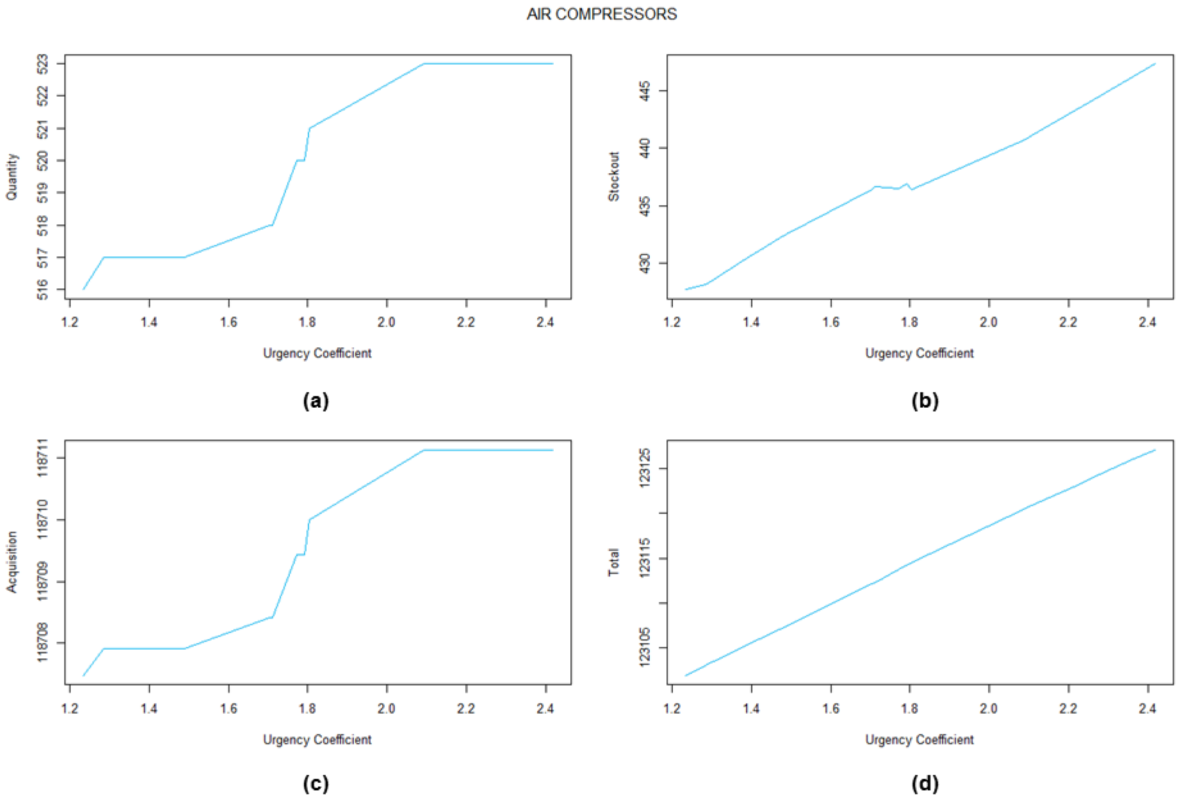


Figure 6.3-6: Results of sensitivity analysis for important cost components

As can be seen in the quantity – urgency graph (a) as the urgency coefficient increases the quantity increases as well. This is expected as if the costs for unplanned cases are higher, it makes sense to increase the stock kept on board a vessel. At the same time the acquisition cost increases (b) with the same slope as the quantity. The behavior of the stock out cost, which can be seen at (c), is the most interesting one. At first, it was expected that with the increase of the urgency coefficient the stock out cost would increase considerably, however the simultaneous increase of the quantity procured slows down the effect of the urgency factor and even if the stock out cost increases it does not increase with the same intensity.

6.4 Comparison with real-case results

To compare the performance of the model with the real case results, the bulk ordering process of 2018 of the case company is used. Here on after, the real case bulk is denoted with the suffix [actual] and the selections and costs produced by the model with the suffix [model].

Table 6.4-1: Comparison of percentage of maker selection and quantities procured for each component

Component	Percentage [model]	Percentage[actual]	Quantity [model]	Quantity[actual]
AIR COMPRESSORS	83%	36%	523	355
ASSEMBLY	-	-	-	-
CAMSHAFT	0%	0%	357	228
CONNECTING RODS	0%	30%	57	46
CONROD (BIG END) BEARINGS	-	-	-	-
CYLINDER HEADS	38%	2%	6519	7387
CYLINDER LINERS	0%	0%	928	950
DIESEL GENERATOR	11%	22%	471	410
DRIVE SECTION	91%	0%	62	55
EXHAUST VALVES	-	-	-	-
FUEL INJECTION VALVES	1%	52%	1382	1153

FUEL OIL PUMPS	0%	61%	287	237
FUEL OIL PURIFIERS	77%	0%	515	407
FUEL OIL SYSTEM	0%	86%	117	75
LO SYSTEM	14%	0%	35	16
LUB OIL PURIFIERS	58%	0%	554	399
LUBRICATING SYSTEM	-	-	-	-
MAIN BEARINGS	-	-	-	-
MAIN DIESEL GENERATORS	33%	0%	101	92
MAIN ENGINE	-	-	-	-
MECHANICAL SYSTEM	1%	0%	724	809
PISTONS	0%	3%	2156	2128
SHAFT ASSEMBLY	44%	0%	284	190
TURBO CHARGERS	0%	4%	85	47
Total			**15,157**	**14,984**

As can be seen from the table above, the percentage of the makers is increased considerably especially in some categories (e.g. the air compressors, the drive section and others).

This can be explained mostly because of the decreased forwarding and stock out costs and will be further explored below.

What needs to be noted here are the assumptions below:

- Stock out cost of the actual case has been computed in the same way that the stock out cost for the purposes of the thesis has been calculated.
- No safety stock was assumed for the actual bulk orders therefore, the probability of stock out is 50%. In the following pages, a sensitivity analysis regarding this probability is carried out.
- Urgency coefficient for the previous bulk was assumed 1.2 (the lower bound of the urgency coefficient bounds).

Table 6.4-2: Comparison of total cost of acquisition and forwarding in each component

Component	Acquisition [model]	Acquisition[actual]	Forwarding [model]	Forwarding[actual]
AIR COMPRESSORS	$ 118,711.13	$ 67,478.15	$ 397.07	$ 1,551.73
ASSEMBLY	-	-	-	-
CAMSHAFT	$ 65.77	$ 42.87	$ 1,428.00	$ 1,428.00
CONNECTING RODS	$ 41,602.46	$ 28,448.57	$ 228.00	$ 179.47
CONROD (BIG END) BEARINGS	-	-	-	-
CYLINDER HEADS	$ 128,783.77	$ 65,694.61	$ 12,453.07	$ 26,606.13
CYLINDER LINERS	$ 59,627.99	$ 38,336.30	$ 3,712.00	$ 3,712.00
DIESEL GENERATOR	$ 10,192.18	$ 4,668.98	$ 1,884.00	$ 1,892.00
DRIVE SECTION	$ 574.52	$ 531.71	$ 16.53	$ 19.20
EXHAUST VALVES	-	-	-	-
FUEL INJECTION VALVES	$ 97,791.30	$ 141,458.83	$ 5,528.00	$ 5,072.00
FUEL OIL PUMPS	$ 85,144.19	$ 148,031.18	$ 1,148.00	$ 1,040.00
FUEL OIL PURIFIERS	$ 18,746.07	$ 11,263.45	$ 137.33	$ 157.07
FUEL OIL SYSTEM	$ 30.70	$ 62.54	$ 468.00	$ 460.00
LO SYSTEM	$ 8.72	$ 4.14	$ 9.33	$ 9.33
LUB OIL PURIFIERS	$ 9,953.03	$ 5,255.58	$ 147.73	$ 166.13
LUBRICATING SYSTEM	-	-	-	-
MAIN BEARINGS	-	-	-	-
MAIN DIESEL GENERATORS	$ 1,236.90	$ 840.21	$ 277.07	$ 408.00
MAIN ENGINE	-	-	-	-
MECHANICAL SYSTEM	$ 46,248.38	$ 6,942.63	$ 211.73	$ 193.33
PISTONS	$ 268,400.23	$ 135,354.90	$ 8,624.00	$ 8,620.00

SHAFT ASSEMBLY	$ 5,249.72	$ 787.33	$ 75.73	$ 75.20
TURBO CHARGERS	$ 36,175.98	$ 19,471.70	$ 340.00	$ 328.80
Total	**$ 928,543.05**	**$ 674,673.67**	**$ 37,085.60**	**$ 51,918.40**

Table 6.4-3: Comparison of total cost of inventory and stock out cost in each component

Component	Inventory [model]	Inventory[actual]	Stock out [model]	Stock out[actual]
AIR COMPRESSORS	$ 3,571.59	$ 1,349.56	$ 438.94	$ 37,860.21
ASSEMBLY	-	-	-	-
CAMSHAFT	$ 1.46	$ 0.86	$ 15.97	$ 614.81
CONNECTING RODS	$ 1,309.17	$ 568.97	$ 362.11	$ 28,329.28
CONROD (BIG END) BEARINGS	-	-	-	-
CYLINDER HEADS	$ 2,737.88	$ 1,313.89	$ 2,696.39	$ 47,976.15
CYLINDER LINERS	$ 1,508.30	$ 766.73	$ 841.09	$ 38,659.21
DIESEL GENERATOR	$ 235.69	$ 93.38	$ 200.41	$ 6,093.52
DRIVE SECTION	$ 13.72	$ 10.63	$ 23.86	$ 608.89
EXHAUST VALVES	-	-	-	-
FUEL INJECTION VALVES	$ 2,354.66	$ 2,829.18	$ 506.90	$ 29,224.29
FUEL OIL PUMPS	$ 2,366.53	$ 2,960.62	$ 276.23	$ 38,526.97
FUEL OIL PURIFIERS	$ 519.57	$ 225.27	$ 186.15	$ 10,855.38
FUEL OIL SYSTEM	$ 0.78	$ 1.25	$ 8.64	$ 408.81
LO SYSTEM	$ 0.28	$ 0.08	$ 1.27	$ 313.07
LUB OIL PURIFIERS	$ 258.52	$ 105.11	$ 120.83	$ 8,385.28
LUBRICATING SYSTEM	-	-	-	-
MAIN BEARINGS	-	-	-	-
MAIN DIESEL GENERATORS	$ 26.22	$ 16.80	$ 14.51	$ 401.65
MAIN ENGINE	-	-	-	-
MECHANICAL SYSTEM	$ 1,190.03	$ 138.85	$ 137.48	$ 20,087.71
PISTONS	$ 7,162.21	$ 2,707.10	$ 2,585.96	$ 124,924.43
SHAFT ASSEMBLY	$ 124.25	$ 15.75	$ 24.34	$ 2,356.84
TURBO CHARGERS	$ 1,174.75	$ 389.43	$ 324.46	$ 28,468.20
Total	**$ 24,555.61**	**$ 13,493.47**	**$ 8,765.52**	**$ 424,094.72**

Table 6.4-4: Final comparison of total costs for actual and model

Component	Percentage [model]	Percentage[actual]	Total [model]	Total[actual]
AIR COMPRESSORS	83%	36%	$ 123,118.73	$ 108,239.66
ASSEMBLY	-	-	-	-
CAMSHAFT	0%	0%	$ 1,511.20	$ 2,086.54
CONNECTING RODS	0%	30%	$ 43,501.73	$ 57,526.29
CONROD (BIG END) BEARINGS	-	-	-	-
CYLINDER HEADS	38%	2%	$ 146,671.11	$ 141,590.79
CYLINDER LINERS	0%	0%	$ 65,689.38	$ 81,474.24
DIESEL GENERATOR	11%	22%	$ 12,512.29	$ 12,747.87
DRIVE SECTION	91%	0%	$ 628.63	$ 1,170.43
EXHAUST VALVES	-	-	-	-
FUEL INJECTION VALVES	1%	52%	$ 106,180.86	$ 178,584.29
FUEL OIL PUMPS	0%	61%	$ 88,934.95	$ 190,558.78
FUEL OIL PURIFIERS	77%	0%	$ 19,589.12	$ 22,501.17
FUEL OIL SYSTEM	0%	86%	$ 508.12	$ 932.61
LO SYSTEM	14%	0%	$ 19.60	$ 326.62
LUB OIL PURIFIERS	58%	0%	$ 10,480.12	$ 13,912.11
LUBRICATING SYSTEM	-	-	-	-
MAIN BEARINGS	-	-	-	-
MAIN DIESEL GENERATORS	33%	0%	$ 1,554.70	$ 1,666.67
MAIN ENGINE	-	-	-	-
MECHANICAL SYSTEM	1%	0%	$ 47,787.62	$ 27,362.53
PISTONS	0%	3%	$ 286,772.41	$ 271,606.42
SHAFT ASSEMBLY	44%	0%	$ 5,474.05	$ 3,235.11

TURBO CHARGERS	0%	4%	$ 38,015.18	$ 48,658.13
Total			**$ 998,949.79**	**$ 1,164,180.27**

As can be derived from the previous tables there is quite a difference between the actual decisions and the model decisions. The two drivers of the cost are the acquisition and the stock out cost and are the two factors that differentiate the model decisions with the actual ones.

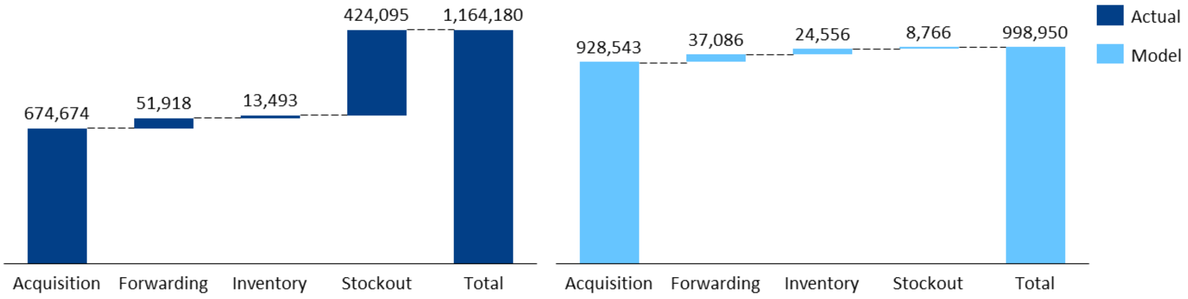


Figure 6.4-1: Final total costs of previously realized bulk and as proposed by the model

The figure above visualizes the difference between the total actual costs and the total model costs. As can be seen the stock out cost of the actual case is considerably high unlike the low stock out cost derived from the model.

The cost components that drive the difference between the actual and the model are the acquisition cost and the stock out cost. The acquisition cost is increased because of two changes. The first refers to increased quantities due to market selection and safety stock and the second refers to the different choices made by the model that increase the acquisition cost because of the difference of price between makers and non-makers.

However, the increased acquisition cost is covered by the decreased stock out cost due to the safety stocks. This is the reason that the stock out cost of the actual case is considerably higher. As previously said the stock out cost of the actual case was computed assuming that the actual quantities did not account for safety stocks. The graph below visualizes the effect that the probability of stock out.

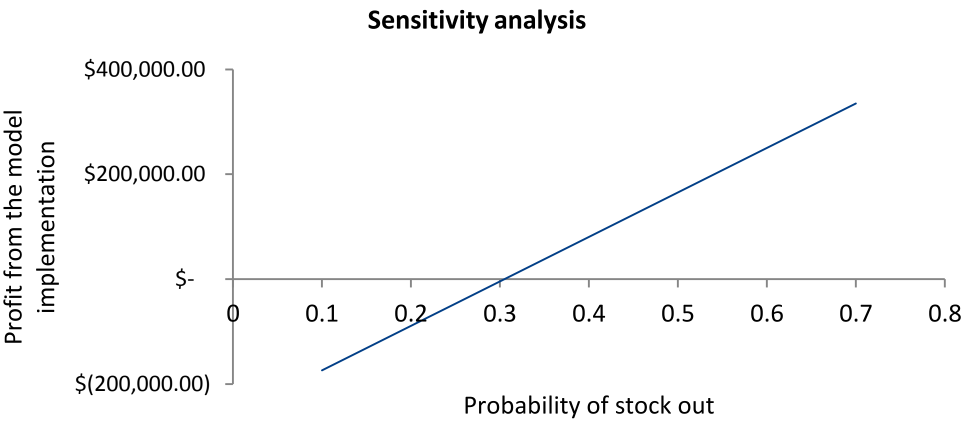


Figure 6.4-2: Sensitivity analysis for stock out probability

As is seen by the graph above, the implementation of the model is profitable when the probability of stock out for the actual case is more than 30%. However, one of the assumptions for the computation of stock out cost (6.4) in the actual case is that the urgency is at the lowest possible level since when the probability is high the more probable is to deliver in a more convenient port. Therefore, as the stock out probability decreases the urgency coefficient would need to be increased. This analysis is outside the scope of the current thesis and the results presented here refer to a decreased urgency coefficient of 1.2 and therefore the conclusions can be considered conservative.

What needs to be noted here is that the purchase of safety stock can be treated as an investment. The safety stock purchased in the first year of implementation of the model will not be replenished in total and in the next year the quantities bought will not be as increased.

Therefore, it is also important to note that if the selection suggestions of the model are used, without purchasing extra quantities (accounting for market and safety stock), the total costs would be as in the graph below.

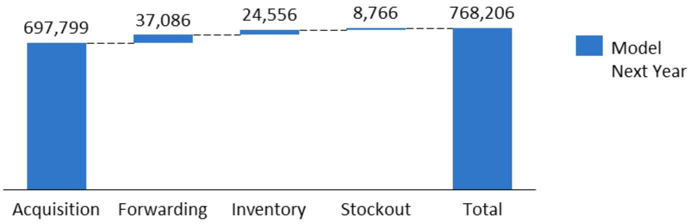


Figure 6.4-3: Final total costs incurred by the model in the 2nd year

As previously said the increased quantities purchased can be considered as an investment that, keeping all other factors stable, would be paid back in full in the 2nd year of implementation of the model regardless of the probability of stock out.

In conclusion, the model supports that with the purchase of safety stock the total cost incurred in the bulk ordering process will lead to decreased costs. The total indirect profit of the implementation of the model as computed for a fraction of the total bulk order process amounts to more than $150,000, which translates to 14.8% reduction of the total costs incurred. The safety stock demands an initial investment of almost one quarter of the total acquisition cost that, according to the model, will decrease the stock out probability, thus decreasing the total costs of the process.

This page has been intentionally left blank

7 Summary of key findings and further research

The original research objective as set out in the summary was to create a comprehensive decision support tool that would help to facilitate the process of the bulk orders and optimize the purchasing decisions.

This would be achieved by firstly reducing the base of analysis by identifying the high interest items of the bulk order. This part focuses mainly on decreasing the workload of the departments involved and on the creation of a targeted subset for further analyses. The next step would be to create a forecasting tool for estimating the expected needs of the fleet regarding the previously identified items and to test whether the needed quantity is influenced by the source of purchase. Lastly, a cost-related decision support tool is created to allocate in the most cost-effective way the items to a group of vendors.

The main conclusions of the above are:

- The identified as high- interest items represent less than 5% of the total items but more than 40% of the total cost. This means that if the bulk ordering process is only focused on these items the administrative workload would decrease considerably both for the departments involved internally, decreasing by around 0.5 FTE, but also for the business partners of the case company
- The forecasting of the nominal needs of the vessels exhibits satisfactory results (average MAPE 53%) and could in the future, when the training samples increase, become more and more accurate. For some specific components that show increased accuracy, e.g. fuel oil system (MAPE = 10%) the tool can be used to expedite the process while decreasing the workload both for the vessel and for the shore- based engineers.
- The forecasting models of the extra needs based on market characteristics are not performing well when it comes to accuracy (average MAPE 165%). What is observed is that the source of purchase in the bulk orders does not heavily influence the extra needs that need to be covered during the year after the bulk. However, a further analysis including more independent variables, related to the urgency of purchases, the types of machinery and their nominal running hours etc. may be performed in the future. This analysis was not performed as said data were not easily accessible in a structured format.
- Lastly, the prescriptive model supports that increased quantities would lead to decreased total costs by 14.8%, as one major component of the cost function is the stock out cost- which represents the increased cost to deliver an item on board with heightened urgency. Even if the previous exercise showed no influence of the source of purchase on the extra quantities, the prescriptive model supports an increased allocation of items to makers. This, however, mainly stems from the fact the makers have smaller production times, thus driving the safety stock down. All in all, the model supports an initial investment on increased safety stock that would be paid back in full (all other factors constant) after the 2nd year of implementation.

In addition, throughout the course of this thesis the below areas were identified as the ones most deserving of further research:

- Re-evaluation of the entire algorithmic framework when more data and bulk order cycles have been amassed. More specifically, with a sizeable enough critical mass of data, deep learning algorithms could be performed and assessed to see if they would further improve the MAPE.

- Add new dimensions (e.g.: urgency of purchase, related maintenance work orders etc.) in order to tackle the increased complexity of the extra needs forecasting. Including these dimensions may lead to more accurate predictions regarding the items re-purchased during the year after machinery failure-and not because of miscalculations in the forecasts
- Further improve accuracy by making the predictive and prescriptive models vendor specific. Given the augmented variance noted in terms of prices, lead-times and total cost of ownership amongst suppliers- and even amongst original makers- any effort to increase granularity across this dimension would most likely yield improved decision making on cost. Furthermore, trade routes of the vessels could also be added as a way to increase accuracy in the computation of the forwarding cost.
- Finally, there is need to reevaluate for these critical items their P-F curves, particularly if their origin is from the parallel market. Retrieving and consolidating this information in a structured format from the engineering crew onboard in between overhauls with the bulk order spare parts would subsequently pave the way for a more holistic predictive maintenance model. The latter should gauge this improved visibility into the spare part reliability when predicting demand or prescribing outcomes for optimum total cost of ownership.

8 Conclusions

The findings of this thesis confirm that there is merit in applying advanced analytics concepts and machine learning algorithms in attempt to rationalize spare parts purchasing and tackle present challenges as discussed in Chapter 0. However, and given the initial investment needed, the application of such concepts will yield a higher ROI for companies that manage a fleet size potentially exceeding 30 vessels. A sizeable fleet is needed to generate a large enough pool of spare part needs to drive the economies of scale that lie at the heart of bulk order execution.

Concurrently, the concepts presented in Chapter 0require an all-encompassing supply chain footprint of commonly adhered to procure-to-pay processes enabled by robust and scalable systems, e.g. AMOS, SAP, and Data Warehouse. The latter, acting as a holistic supply chain system of record, provide the primary source of a raw data set with the necessary breadth and depth to make concepts such as machine learning attractive, meaningful and effective.

Of course the data set alone is not enough to unlock value, so any ship management company looking to go down this road will need to invest in developing internally its analytics capabilities by building a pool of dedicated data scientists coupled with analytics translators, who are capable of bridging the business need with the algorithmic capability necessary. The latter will link the business need with the analytics capability necessary the former will employ so as to drive value in any such initiative. It is worth noting that supply chain systems of record and organically grown analytics capabilities remain elusive concepts in the maritime industry which has traditionally proven tardy in following the digitalization traits, minus some nominal exceptions such as the case company. Yet even in this case which fulfills the aforementioned requirements to a large extent, limitations and obstacles were faced when working towards the completion of this thesis. The primary one was data availability and depth: advanced analytics are data hungry to the extent that three years of bulk order data, in the time dimension, and over 90 vessels in the space dimension were not enough to enough to even entertain the thought of deploying deep learning algorithms. The ones that remained were hand-picked to ensure their structure was not prone to overfitting due to lack of extensive data to train them on, e.g. Random Forest.

Additionally, such exercises as the one tackled in this thesis would need to also leverage the vantage point of core operations owners such as the Technical department since the Bulk Order analytics framework created could steadily evolved into a predictive maintenance tool with cross-functional ownership and applicability. In practice this proves difficult to ascertain and manage given the different pace, systems and tools the alternate departments follow, the Chinese walls raised in some cases. These could be handled under the umbrella of an end-to-end Maintenance Transformation Programme, complete with a Change Management methodology, provided all stakeholders involved see and realize the untapped value advanced analytics can generate.

The McKinsey Global Institute (McKinsey & Company, 2018) estimates the net effect potential of Artificial Intelligence on the world economy to be an incremental 13 trillion USD of economic activity by 2030, or a 16% higher cumulative GDP than today. It will be hard to imagine such a tremendous impact leaving the maritime industry unaffected and ship management companies already start veering towards AI adoption at a faster pace than in the past in an effort to keep up with the rest of the world, much like the case company is doing. However, since there does not appear to be a silver bullet for the adoption of such concepts it makes sense to focus on tangible and ripe use-cases- e.g. forecasting for bulk order spares- generate a monetary and workload related benefit, communicate it accordingly cross-departmentally and build the remaining use-cases from that point onwards. That was the main contribution of this diploma thesis to the case company.

9 Bibliography

Bair , E., Hastie, T., & Debashis , P. (2005). Prediction by Supervised Principal Compoents.

Bellman, R. (1957). *Dynamic Programming.* Princeton University Press.

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.

Bramer, M. (2007). *Principles of Data Mining.* Springer.

Brandmaier, A., von Oertzen, T., McArdle, J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*.

Breiman , L. (2001). Random Forests. *Machine Learning*.

Brodley, C., & Friedl, M. (1996). Identifying and Eliminating Mislabeled Training Instances.

Coursera. (2018). *Machine Learning: Clustering & Retrieval*. Retrieved from Complexity of brute force search: https://www.coursera.org/lecture/ml-clustering-and-retrieval/complexity-of-brute-force-search-5R6q3

Deng, H., Runger, G., & Tuv, E. (2011). *Bias of importance measures for multi-valued attributes and solutions.* Springer.

Deviant , S. (2012). *Statistics Handbook .* CreateSpace Independent Publishing Platform.

Ester, M., Kriegel , H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* Portland, Oregon.

Evans, J. (1996). *Straightforward statistics for the behavioral sciences.* Pacific Grove.

Evans, J., & Linder, C. (2012). Business Analytics: The Next Frontier for Decision Sciences. *Decision Line*.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning.* Springer.

Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall/CRC.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*.

Hahsler, M., & Piekenbrock , M. (2018). *dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and related algorithms.*

Hamerly , G., & Elkan, C. (2002). Alternatives to the k-means algorithm that find better clusterings. *ACM CIKM International Conference on Information and Knowledge Management.* McLean.

Hanke, J., & Wichern , D. (2009). *Business Forecasting.* Pearson.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning.* Springer.

Haugeland, J. (1985). Artificial Intelligence: The Very Idea . *MIT Press*.

Hinton, G., & Sejnowski, T. (1999). Unsupervised learning: Foundation Computation. *MIT Press*.

Ho T. (1995). Random Decisions Forests. *ICDAR '95 Proceedings of the Third International Conference on Document Analysis and Recognition.*

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres , G., Chay, L., O'Hara- Wild, M., . . . Yasmeen , F. (2019). *_forecast: Forecasting function for time series and linear models_*.

Imdadullah. (2014). Time Series Analysis. *Basic Statistics and Data Analysis*.

International Chamber of Shipping. (2018). *http://www.ics-shipping.org/*.

International Chamber of Shipping. (2019). *http://www.ics-shipping.org/*.

Jackson, E. (1991). *A User's Guide To Principal Components.* John Wiley & Sons.

James, G. (2003). Variance and Bias for General Loss Functions. *Download PDF*.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*.

Kriegel, H., Kroger, P., Sander, J., & Zimer, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Kuhn, M., Wing , J., Weston , J., Williams , A., Keefer, C., Engelhardt , A., . . . Hunt , T. (2018). *caret: Classification and Regression Training.*

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*.

McKinsey & Company. (2018). *Notes from the AI frontier: Modeling the impact of AI on the world economy.*

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning.* MIT Press.

Nelder , J., & Wedderburn, W. (1972). *Generalized Linear Models.* Wiley for the Royal Statistical Society.

Nyce, C. (2007). *Predictive Analytics White Paper.*

NYU Stern. (2019). *Cost of Capital per Sector.*

Pearson , K. (1901). On Lines and Planes of Closest Fit to Points in Space.

Pena, J., Lozano, A., & Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*.

Poole, D., Mackworth, A., & Goevel, R. (1998). *Computational Intelligence: A Logical Approach.* Oxford University Press.

Ripley , B., & Lapsley, M. (2017). *RODBC: ODBC Database Access.*

Russel, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach.* Pearson.

Sander, J., Ester, M., Kriegel, H., & Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*.

Schubert, E., Sander, J., Ester, M., & Kriegel, H. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*.

Sievert, C. (2018). *plotly for R.*

UNCTAD. (2018). *Review of Maritime Transport.*

Zhang, B. (2003). Comparison of the performance of center-based clustering algorithms. *PAKDD '03 Proceedings of the 7th Pacific-Asia conference on Advances in knowledge discovery and data mining.*

Zimek, A., & Filzmoser, P. (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.