Norwegian School of Economics
Bergen, Autumn 2023

AIS-data & Machine Learning

A Quantitative Approach to Predicting Freight Rates

Ole Fredrik Odfjell & Magnus Haugland

Supervisor: Jonas Andersson

Master thesis, Economics and Business Administration

Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible – through the approval of this thesis – for the theories and methods used, or results and conclusions drawn in this work.

Acknowledgements

This thesis is written as a part of our Master of Science degree in Economics and Business Administration within our major in Business Analytics at Norwegian School of Economics. This thesis summarizes our five years attending Norwegian School of Economics, and we are grateful for the excellent opportunity to meet remarkable people and learn from some of the best professors in the confines of the economic field. During these five years, we have gained insight and aspiration through past courses to explore the shipping industry further and thus contribute to research embedded in this field.

We sincerely thank our supervisor, Professor Jonas Andersson, for encouraging us to write within the shipping field and for his guidance and constructive feedback throughout the semester. His advice has given us valuable insight and motivation for a seamless realization of this thesis.

Additionally, we thank Assistant Professor Gabriel Moises Fuentes from the Department of Business and Management Science for providing the AIS data used in this thesis.

Norwegian School of Economics

Bergen, December 2023

Ole Fredrik Odfjell

Magnus Haugland

Abstract

The emerging availability of data and the development of real-time tracking systems, also known as AIS, have engaged a new field of study within the shipping segment. AIS data has a pivotal role in enhancing safety at sea. Moreover, the accessibility of real-time data over the majority of merchant vessels around the world has instigated researchers to investigate how to adopt this information to create further value in the decision-making process.

Together with machine learning methodologies and data processing capability, this thesis aspires to contribute to further investigate the deployment of AIS-derived data. More specifically, we will examine the predictive ability of AIS data on a route-specific freight rate. In addition to AIS variables, we have included other data expected to influence freight rate, and the results from a series of machine learning models have been thoroughly examined. Our results indicate that AIS-derived data offer some additional value when predicting the freight rate. However, in this exact case, the additional contributory value is negligible.

Keywords – Business Analytics, NHH, Machine Learning, Shipping, AIS, Master thesis, Maritime Economics

Contents

List of Figures

List of Tables

1 Maritime Economics and Tanker Shipping

The dynamics of sea-borne transportation pricing are influenced by the delicate balance of supply and demand. Ship owners and charterers negotiate freight rates that are mutually acceptable in the market. Supply, as outlined by (Lun et al., 2010), is determined by factors such as overall cargo capacity and fleet operational efficiency. Wijnolst and Wergeland (1996) identify load factor, sailing speed, and waiting days per trip as crucial elements of fleet operational efficiency.

The demand side of the tanker market is inherently linked to the volume of crude oil transported, which is further connected to the global economic activity and overall oil production consumption (Stopford, 2009). The distance from which the cargo is transported plays a pivotal role. The demand for tanker freight is quantified on a tonne-mile basis, multiplying the metric tonnes shipped by the distance traveled (Adland et al., 2017). For instance, the demand for shipping a tonne of crude oil from the Arabian Gulf (AG) to the US Gulf differs from shipping to Europe.

1.1 Maritime Economics

Transporting goods and raw materials by sea has been essential for global trade and economic growth for centuries. Shipping enables development by bringing together capital goods, intermediaries, and consumers through global transport networks, promoting global and regional economic integration. Today's society depends on the imports and exports provided by the worldwide shipping industry. The shipping industry accounts for over 80% of world trade and has been ever-increasing in volume. Increasing industrialization, economic liberalization, and technological advances have and will support continued growth. This chapter will explain the key fundamentals of shipping before focusing on the tanker market. This will serve as both essential information and the basis for understanding the selection of variables and models in further analysis.

1.1.1 Shipping Segments

The global merchant fleet consists of more than 50,000 vessels (> 1,000 DWT), transporting all types of cargo placed among three vessel types. These are dry bulk carriers, oil tankers, and container ships. The two first categories are traditionally tramp shipping, an on-demand service, while container ships are referred to as liner shipping, with predefined routes and scheduled services. The oil tanker segment includes gas carriers (LNG & LPG) and specialized and chemical vessels.

1.1.2 Four Shipping Markets

Within shipping, there are four primary markets to consider: New-building, Sale & Purchase, Demolition, and the Freight market (Stopford, 2009). The cash flows between these markets are driving the overall shipping market. The freight market is the shipowners' primary driver of positive cash flow. However, this depends on the current market situation. In specific periods, sale and purchase or demolition could account for the most positive cash flow, yet the freight market is the main driver in the long run. Figure 1.1 sums up the different cash flows.

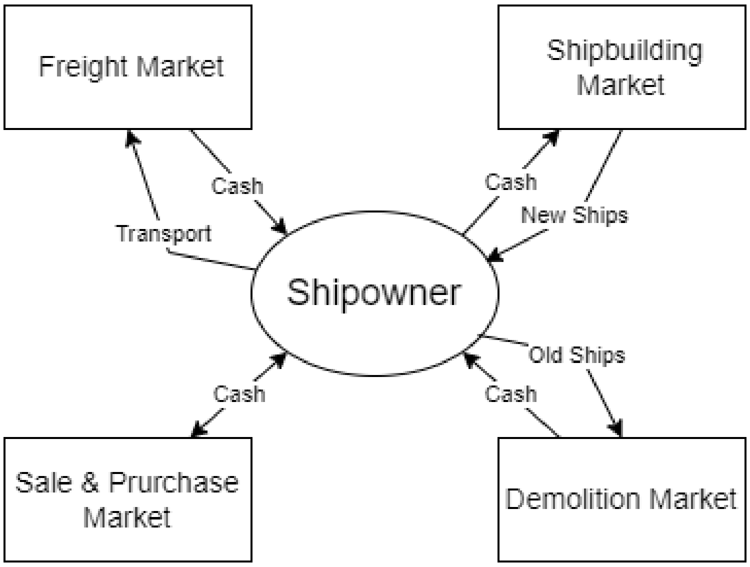


Figure 1.1: Four Shipping Markets

1.1.3 Stakeholders

Filling the four different shipping markets are its stakeholders. The key stakeholders are Shipowners, Charterers, Shipbrokers, Shipyards, and Capital markets (Stopford, 2009). Shipowners (often referred to as just "owners") are the companies owning and operating ships. Some owners have fleets working across several shipping segments, while others specialize in one. Charterers are the companies that hire ships to transport their cargo. The shipbroker is the connection between shipowners and charterers, and their task is to find a suitable and available ship for a given cargo/voyage. Shipbrokers are also involved in negotiating contractual terms and are hired by both owners and charterers. Shipyards are responsible for new-building and offering repairs and modifications (e.g., retrofitting). Lastly, the capital markets allow the stakeholders to raise capital, restructure, and liquidate.

1.1.4 The Shipping Market Model

The shipping market model highlights the mechanisms that control the supply and demand on a macroeconomic level. As Figure 1.2 illustrates, the freight market is characterized by the interplay of supply and demand, which ultimately affects freight rates. Owners in the tanker market generate income by transporting bulk liquid from one port to another, and the freight rate determines the revenue generated. As mentioned, the freight market is the main driver of positive cash flow for the owners, and freight rates are the link between supply and demand (Stopford, 2009).

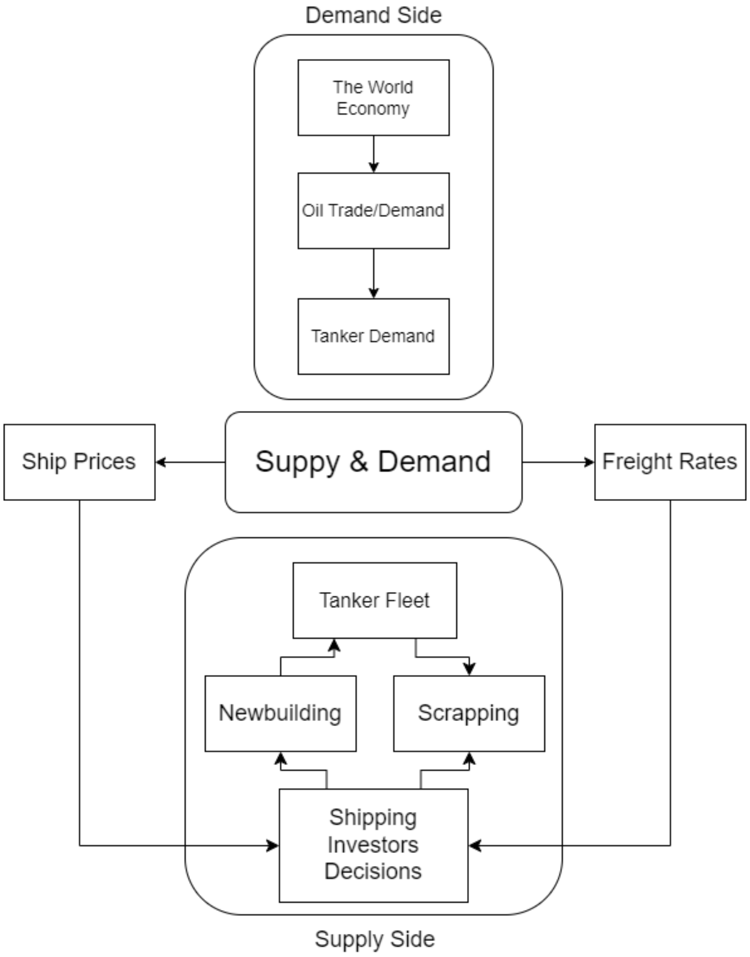


Figure 1.2: Simplification of (Stopford, 2009) Shipping Market Model

1.1.5 Freight Rate Mechanism

The freight rate mechanism links the theory of supply and demand with the theory of shipping cycles (Stopford, 2009). A combination of volatile demand and significant time lag for supply sets the frame for the shipping cycle(s). Charterers and shipowners negotiate to set freight rates that reflect the balance of available ships and cargo in the market. This balance is continuously adjusting, and so are the freight rates. In times when supply is low, the freight rates will increase. As a consequence, more suppliers and shipowners are willing to enter the market, either by repositioning in the short-term or by buying new ships, thus increasing the overall global supply. With the arrival of new-buildings, the fleet size increases, and in some cases, causes an overcapacity. This dynamic pressures freight rates, and shipowners are forced to underbid each other to secure cargo. This is, by many, considered an example of perfect competition, and to analyze it further, we turn to supply and demand functions.

1.1.6 Supply and Demand Curves

Supply and demand curves in shipping are modeled as in Figure 1.3, where the demand and supply curve intersect are theoretical freight rate equilibrium's. The supply curve indicates the amount of transportation, measured in tonne-miles, that the fleet is willing to supply. The curve is composed of each individual ships supply curve, and new, modern ships contribute to the lower part, as they are able to operate more cost efficient in comparison to older ships. The supply curve is elastic up to a certain point, at which the fleet sails at maximum capacity. To increase capacity in the short-term, ships increase their sailing speed, thus increasing the tonne-miles supplied. At some point, the fleet cannot react to short-term increases in demand, and thus, the supply curve becomes inelastic. As shown, the demand curve is almost vertical. This shape is supported by the lack of competing transport modes, freight representing a small cost for charterers, and owners need for cargo. In addition low rates do not lead charterers to ship extra cargo (Stopford, 2009).

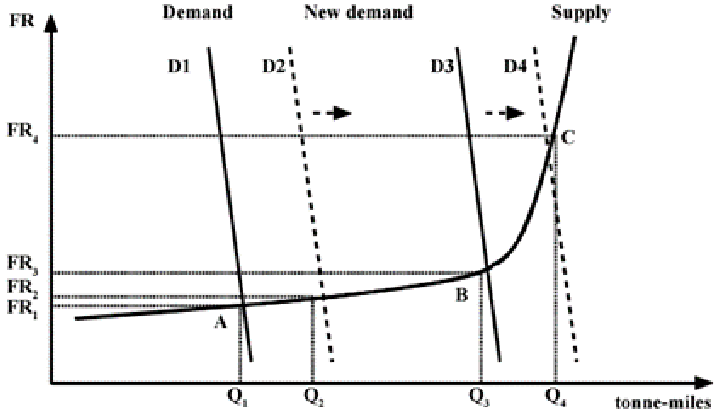


Figure 1.3: Supply & Demand

At the intersection of the supply and demand curves is the equilibrium price. The point at which shippers and shipowners have found an acceptable freight rate. In finding the equilibrium, the time frame is essential, and there are three time periods to consider: momentary, short-term, and long-term. The momentary is the freight rate negotiated for readily available ships and cargoes. In the world of shipping, this happens in "local" markets, such as the Arabian Gulf (AG) and US Gulf (UG). This can lead to local peaks and troughs, and as an owner, it is essential to anticipate when choosing between fixing cargo or ballasting to other loading areas. Moving to the short-term, there is more time to respond. This can be done for owners by moving ships in and out of lay-up or relocating ships. As demand moves upward, ships are being taken out from lay-up, and as the supply curve changes, demand has smaller effects on the freight rate. If demand growth persists, all ships are eventually taken out from lay-up. The next step to increase fleet productivity is to increase sailing speed.

In this case, a similar alteration in demand as earlier can cause significant changes in the freight rate, and when all ships are running at full speed, there is essentially no more supply available, and rates can go to any level. This often leads to investments by owners and charterers. Lastly, we consider the long run where new-building or scrapping can adjust the fleet size. When rates peak, owners want to add to their fleet by bidding up the price of secondhand ships until it is more reasonable to acquire new ships. This is subject to time lag as new-building usually takes at least a year to build. Eventually, it increases the supply which causes a downward pressure on the freight rates. In the opposite case, when freight rates are low, the profitability of the ships decreases, and eventually, it reaches the scrap price for the least efficient ships, which diminishes the supply. Another supply-decreasing effect is the conversion of vessels to enter other markets and the use of ships as floating storage. Again, this is something that owners will consider when freight rates reach critically low levels.

The dynamics between supply and demand are the causes of shipping cycles, characterized by fluctuating periods of boom and bust, with regard to attainable freight rates. Most commonly, a cycle is divided into four periods: peak, collapse, through, and recovery, see Figure 1.4. The peak is when the demand-to-supply ratio is at its max (i.e., high freight rates), while the ratio is at its minimum in the through. The collapse and recovery are the stepping stones between these two points. While the existence of cycles is agreed upon, they often differ in length and severity. Although typically lasting for a couple of years. On the demand side, major influences are global economic conditions, trade volumes, and geopolitical factors. Meanwhile, shipbuilding, regulatory changes, and technological developments influence supply.

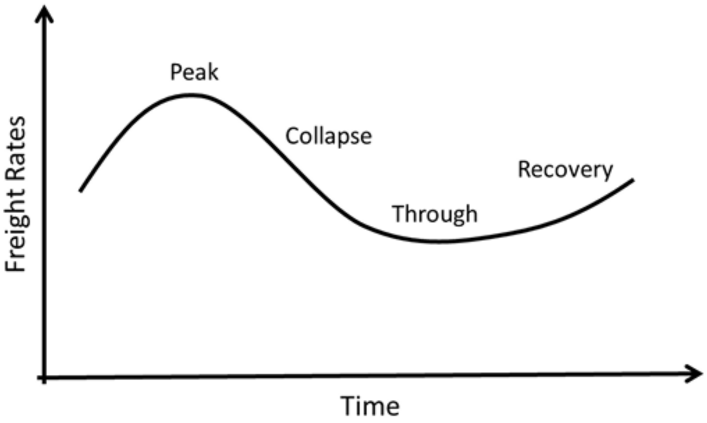


Figure 1.4: Shipping cycle stages

1.2 Global Tanker Trade

Global tanker trade involves transporting liquid cargo, primarily crude oil, and petroleum products, and it is essential for meeting worldwide energy needs. Historically, there have been major changes in the size of the vessels and their technical standards. The first Very Large Crude Carrier (VLCC) was built in 1966, to the 850 VLCCs of today. The introduction was due to economies of scale and the increased oil trade from the Middle East during the mid-20th century. Development in technical standards has been chiefly due to regulatory changes.

While long-term contracts dominated the tanker trade for an extended period, there has been a shift towards the spot market. Today, the spot market dominates the trade, allowing for more short-term changes in the freight rate.

1.2.1 Tanker Supply

The supply of tankers is represented by the cargo carrying capacity of the fleet. In early 2018, the cargo capacity was 561 million tonnes (million DWT), and the total number of tankers surpassed 6100. This accounted for around 40% of the total world shipping fleet. Within the tanker fleet, we have ships of different sizes, showcased in Table 1.1. As the names suggest, the Suezmax and Panamax sizes can transit in their respective canals. An Aframax could also, for instance, transit the Suez Canal as it is smaller than a Suezmax. The table also reveals that the transport of crude oil is the most essential liquid bulk cargo, with more than 2200 ships and a capacity of over 400 million DWT, the VLCC fleet is the most significant contributor.

In the long-term, tanker supply depends on the fleet size. It is regulated by scrapping and new-building. In a shorter time horizon, the fleet size also depends on the number of ships in lay-up. This governs the size of the active fleet and is often the first response to increased demand. When the whole fleet is operating, fleet productivity is the main driver of supply changes. Fleet productivity refers to the operational effectiveness of the fleet. This can increase (decrease) supply in the short-term by increasing (decreasing) sailing speed, thus increasing (decreasing) the tonne-miles supplied. Increasing cargo utilization by carrying the maximum amount of cargo also increases tonne-miles.

Tanker type	**No. tankers**	**Tanker size (k.DWT)**	**Fleet size (mill.DWT)**	**Main cargo**
U/VLCC	850	200+	262.15	Crude oil
Suezmax	622	125-200	95.96	Crude oil
Aframax	1,073	85-125	115.59	Crude oil & oil products
Panamax	459	55-85	33.50	Oil products
Handysize	2,434	10-55	101.78	Oil products
Tanker Fleet	**5,438**		**608.98**	

Table 1.1: Tanker Fleet (2022)

1.2.2 Tanker Demand

The tanker shipping market is the largest shipping sector in terms of trading volume and weight. To derive this factor and quantify the demand, one must account for the distance ("haul") the cargo is transported. This is done by multiplying the amount of cargo (tonnes) transported by the average distance of the trade. Thus giving a measure of demand in tonne-miles.

On a macroeconomic level, tanker demand depends on the international oil trade, which in turn depends on overall world economic activity (Alizadeh and Talley, 2011). Stopford (2009) mentions GDP as a significant factor for predicting future demand. He also adds transport costs and political events. In the case of tankers, the oil price is influential, as the oil price, which is relevant to coal and gas, directly affects the global energy mix. Consequently affecting the demand for the transportation of oil.

Table 1.2: Crude Oil Export by Loading Zone/Region

	Volume	
Load Zone	Volume (KT)	Barrels
Middle East	814,069.924	5,954,116,804
North America	243,008.311	1,807,278,813
Russia, FSU	215,272.723	1,566,774,037
Other Africa	186,444.732	1,360,868,881
Mediterranean, North Africa	154,234.615	1,146,008,730
South America	153,496.707	1,089,345,440
North West Europe	134,131.326	971,751,271
Caribbean, Central America	83,217.752	581,257,222
South East Asia	60,240.238	447,369,319
North Asia	16,477.295	122,347,658

Table 1.3: Crude Oil Import by Discharge Zone/Region

	Volume	
Discharge Zone	Volume (KT)	Barrels
North Asia	729,192.123	5,335,390,908
Mediterranean, North Africa	282,786.985	2,068,852,204
North America	244,275.211	1,769,421,378
North West Europe	224,737.565	1,650,202,177
South Asia	212,141.830	1,542,812,132
South East Asia	173,538.415	1,276,491,467
South America	54,207.385	392,472,586
Middle East	52,577.293	384,224,135
Caribbean, Central America	32,475.726	232,420,024
Russia, FSU	23,475.907	168,744,424

1.2.3 Freight Contract Agreement

There are four main types of agreements, also referred to as charter-parties, between shipowners and charterers (Stopford, 2009). They are voyage charter, contract of affreightment, time charter and bareboat charter. The difference is the distribution of costs and responsibility. The voyage charter (VC) contract provides transport for a specified shipload of cargo from the load port to the discharge port. VC contracts are therefore just for a single route specific voyage, and often referred to as a “spot” contract. In this case, the charterer pays the shipowner a pre-agreed freight rate based on a per-tonne basis or a lump sum. Contractual terms include the freight rate, load port, discharge port, cargo type, and quantity, laytime, and demurrage. In terms of trade-related costs, the owner is fully responsible. Among tankers, VC contracts are the most common, especially for VLCC's. Contract of Affreightment (CoA) is a contract similar to a VC, yet somewhat more complex. Where a CoA concerns a series of shiploads, for example, a delivery each month for the next 6 months, as opposed to just one voyage. The transport is fixed at a price per tonne, and again the owner pays all costs. Time Charter (TC) is an agreement where the charterer pays the owner an agreed daily rate over a certain period, and gains operational control over the ship. The owner pays the operational cost (OPEX) while the charterer pays the voyage specific costs (VOYEX). VOYEX includes bunkers, port and canal fees. TC contracts can be further divided into spot and term contracts, where the spot contracts usually have a duration of less than 3 months. Often agreed to in a few days before start of operation, and re-negotiated frequently. Spot TC rates can therefore vary on a day-to-day basis. Lastly, TC contracts are most common in the dry-bulk fleet. Bareboat charter (BBC) is a contract in which the charterer gets full operational control of the ship, and all associated costs, with the exception of capital cost, as this is still the owner's responsibility. These contracts usually have a longer duration, stretching over several years.

Charter Party	**Party paying for**		**Freight rate**
	OPEX	**VOYEX**	
VC & CoA	Shipowner	Shipowner	$ per tonne
TC	Shipowner	Charterer	$ per day
BBC	Charterer	Charterer	$ per day

Table 1.4: Charter Party Details

Year	**Voyage/Spot**	**TC**
2020	1,134	28
2021	1,063	24
2022	1,136	25
2023 (YTD)	918	13

Table 1.5: Fixture Type Overview U/VLCC

1.2.4 Worldscale

Oil tankers are, as mentioned, often fixed on spot VC contracts, where the charterers often are oil and gas companies, oil traders, or oil refineries. VC freight rates are generally settled based on the Worldscale index in the tanker industry. Worldscale is an index based on the freight rate of transporting a tonne of cargo using a standard vessel on a round voyage. This freight rate is on a dollar per tonne basis, and the above description represents the Worldscale 100 (WS 100) commonly referred to as the flat rate. The functioning of the flat rate is that it represents a break-even scenario for a given voyage. As there are numerous different voyages and vessels, there exist a large number of WS flat rates. This gives the participants an easily comparable number across routes. There is also an element of economics of scale to consider. The larger vessels, VLCCs, usually trade at lower WS, as they can profit at lower dollar per tonne. For example, if a route's flat rate is 25$ per tonne, a WS75 amounts to 18.75$ per tonne.

In our case, the dependent variable, the freight rate on the TD3C route, is given in WS units.

2 Literature Review

2.1 Literature review

This chapter will present literature that will serve as a fundamental basis for further analysis. Additionally, it will give insight into the current field of study of the application of AIS-related data in combination with machine learning, as well as the core understanding of freight rate dynamics. To gain the entire perception, the chapter is divided into two parts. First, we introduce the general literature regarding freight rate dynamics. Second, we delve further into the literature concerning AIS data and machine learning methodology.

2.1.1 Freight rate dynamics

2.1.2 Supply & Demand

The foundation for freight rate dynamics can be said to have been laid in 1939 by Koopmans through his analysis of cyclical fluctuations. The findings suggested that the cyclicality is a product of the delay between supply and demand. More specifically, the delay of which the ship capacity meets the demand. His concept has influenced a series of empirical models exploring the relationship between freight rates and the supply of seaborne transportation. Notable studies include Zannetos (1964), Hawdon (1978), Strandenes (1984), Hampton (1990), Beenstock and Vergottis (1989), Beenstock and Vergottis (1993), Engelen et al. (2006) and Randers and Göluke (2007).

Their investigations have revealed that freight rates are intricately tied to several factors, including global economic activity, growth in industrial production, seaborne commodities trade in general, oil prices, availability of ship tonnage, new-builds on orders and deliveries, and scrapping rates. The essence of these findings is that the equilibrium between demand and the size of the active fleet determines the freight rate.

More comprehensively, Beenstock and Vergottis (1993) made an econometric model for the tanker market, using the insights from Beenstock (1989) about the expected second-hand value of ships. In this model, supply was modeled as a function of freight rate, fleet size, OPEX, and the cost of lay-up, while demand was exogenous. In more recent years

Randers and Göluke (2007) have applied macroeconomic variables to model and forecast freight rates, where they also treated demand as exogenous.

In addition, Randers and Göluke (2007) highlighted the flexibility of supply, measured in tonne-miles, which is influenced by the operations of the vessels. In times of excess demand, the owners or operators can enhance profitability by adjusting operational aspects of shipping, such as speeding up their fleet and increasing capacity utilization.

Not to mention, Aadland et al. (2017) observed a positive relationship between capacity utilization and freight rate. They hypothesized that owners compete over cargo sizes lower than their ships' optimal size in poor market conditions with low freight rates. This suggests that changes in freight rates lead to changes in utilization and not necessarily vice versa. However, economic theory suggests that higher utilization leads to lower available supply, which should increase freight rates. Therefore, there is some uncertainty about the load factor's effect on rates.

Moreover, the studies conducted by Beenstock and Vergottis (1993), as mentioned above, resulted in a development towards more advanced time-series models as a consequence of implementations of complex econometrics. This development resulted in the ability to investigate the non-linearity of freight rates. In this context, a significant number of research has been conducted on, for instance, the stationarity of freight rate dynamics. A notable study by Koekebakker et al. (2006) delved into this, with findings that align with maritime economic theory, which states that freight rates are non-linear stationary. The result is helpful since the predictability of freight rates is enhanced since the long-term fluctuation would return and stabilize in a predictable pattern.

Concerning the tanker market, Lyridis et al. (2004) used Artificial Neural Networks to predict VLCC spot freight rates. The findings from the study indicated that there were several valuable variables. Among those deemed most important were the demand for oil transportation, the active fleet, crude oil production and price, TC rates, new build, second-hand and scrap prices, bunker oil prices, and oil stock building. Talley and Alizadeh (2011) also found many of the same variables, although to determine supply. Adding to the aforementioned are tonnage available for trading, scrapping rate, and fleet productivity.

A thesis from NHH by Olsen and da Fonseca (2017) investigating the predictive ability of AIS data for tankers in the AG. They focused on two sets of variables. One set with "publicly available data" and the other with the addition of AIS-derived variables. The first set included average production of crude oil in the Middle East, VLCC fleet development, VLCC tonnage available in the AG, bunkers price in Singapore, refinery margin in Asia, and Brent-Dubai crude oil price spread. These measures were chosen to capture the dynamics of supply and demand, both locally and globally. From the AIS data, they derived the following variables for supply measures: fleet speed, voyage speed, and load factor. On the demand side, they derived the number of VLCCs in the AG, the number of other tankers in the AG, the number of VLCCs elsewhere, and the number of other tankers elsewhere. The demand measures were included to account for the distance effect on demand, which should be possible to measure using AIS data (Aadland et al., 2017).

Moreover, Aamann et al. (2015) investigated the relationship between vessel speed under high freight rates and low bunker prices. The discovery unveiled that speed optimization was more significant on backhaul trips than on laden trips, where the optimization was insignificant This suggest that speed optimization ultimately can affect the supply of ships, and thereby the availability of load factor in specific areas based on the underlying conditions presented.

Lastly, the study by Tham (2008), investigating the leading indicators for Arabian Gulf oil tanker rates, applied Bayesian logistic regression to find predictors for TD3 rates. His resulting predictors were the refining margin in Asia, crude oil production in the Middle East, capacity utilization, and the Brent-Dubai crude oil spread.

2.1.3 Financial and Non-fundamental

Studies by Alizadeh and Talley (2011) and Bakshi et al. (2010) found a connection between real and financial markets by the predictive power of the Baltic Dry Index (BDI). Both studies show that the BDI can serve as an indicator of stock market returns. Additionally, Bakshi et al. (2010) also found that the BDI can predict global economic growth and commodity indices. These findings show the connection between shipping related indicators, the BDI as an example, and broader economic and financial trends. Furthermore, a study by Kavussanos and Visvikis (2006) also highlights the currency risk present in shipping. Owners usually have U.S. dollars income, while payments in local currencies.

2.1.4 AIS & Machine learning

Moreover, this section emphasizes research conducted in the area of AIS-derived data and machine learning methodologies. In the early stages, Li and Parsons (1997) demonstrated that neural networks, compared to Autoregressive Moving Average (ARMA) series models, perform better when predicting freight rates, especially in long-term forecasting horizons.

Equally important, Olsen and da Fonseca (2017) utilized the Vector Autoregression (VAR) model in sequence with AIS-derived data to predict the freight rate in the Arabian Gulf tanker market. The research compared multivariate models against univariate models, unveiling that the multivariate model performed better than the univariate model when predicting the freight rate. However, it was also noted that there was limited evidence supporting the use of AIS data from a forecasting perspective. Despite the result of this research suggesting that AIS-derived data might not enhance the predictable capability, investigation using alternative methodologies might argue the contrary.

Han et al. (2014) employed Support Vector Models (SVM) to predict the dry bulk freight index. They compare the SVM model against other forecasting methods, and the findings proposed that SVM has an improved accuracy when predicting in a short-term horizon. Accordingly, the SVM model would be attractive when investigating whether the AIS data contributes additional value when forecasting freight rates. Moreover, Gao and Lei (2017) used neural networks to predict crude spot prices by developing a new stem learning algorithm enabling continuous data processing for a non-stationary process. The model proposed by the researchers is a univariate neural network that considers past price history. The outcome demonstrated that the model attained the highest accuracy compared to other models, considering both mean squared prediction error and directional accuracy ratio across a range of forecast horizons. As a result, neural networks will be one of the models to consider when predicting freight rates. Equally important, Næss (2018) investigated whether multivariate machine learning and AIS data could add additional information to the short-term prediction of weekly spot prices in the LPG market. The discoveries presented evidence leveraging features of AIS data and additionally gave evidence in favor of utilizing multivariate machine learning models.

2.2 Thesis contribution

When examining the existing literature, it becomes evident that using AIS-derived data in combination with machine learning methods to predict freight rates in the VLCC market is a relatively unexplored avenue. Currently, only a limited proportion have investigated this relationship. With the recent emergence of accessibility of AIS data within the shipping industry, the literature review underscores various methods and applications, revealing the potential for expansion in this field. Machine learning methods, particularly neural networks, have exhibited promising results in forecasting time-series-related data. Consequently, this thesis will combine the different models presented to examine if the AIS-derived data adds additional value when predicting freight rates on the TD3C route.

3 Data

This chapter will start by introducing AIS data, and its applicability. Furthermore, we briefly explain our data gathering and handling, before presenting the freight rate determinants chosen, and their reasoning. Lastly, we present descriptive statistics of the variables as well as the correlation matrices.

3.1 Introduction to AIS Data

Automatic Identification System (AIS) is a coastal tracking system based on short-range Very High Frequency (VHF) maritime band ("NATO Shipping Centre", 2023). It is installed on more or less all merchant vessels today by regulations determined by The International Maritime Organization (IMO, 1974). Facilitating data exchange among vessels, AIS stations, and satellites, the system records both static and dynamic data to enhance safety.

AIS functions as a ship-to-ship and ship-to-shore reporting system, utilizing VHF radio waves transmitted and captured by terrestrial land-based antennas (T-AIS) and Low Earth Orbit Satellites (S-AIS). The introduction of S-AIS expanded the reception of global AIS messages. These messages encompass static information such as MMSI, voyage-related data, for instance, draught, and dynamic data such as speed, and rate of turn(Adland et al., 2017). Contrary to this, the AIS data was collected through land-based receivers (Skauen et al., 2013).

Furthermore, AIS delivers different types of messages containing distinct information. The information sent is the MMSI number, IMO number, name and call sign, length and beam, type of ship, and location of position fixing antenna. Second, the dynamic information depends on the speed and course alteration. The content of these messages is, for instance, the ship's position with accuracy indication, position timestamp in UTC format, and course over ground (COG).

The system identifies the distinct vessels by the Maritime Mobile Service Identity (MMSI) number. The MMSI number is connected to the AIS system onboard the vessel and changes only if there is a new vessel owner. Additionally, there is an identification number, defined as an IMO number, that is mandatory for all ships over 100 gross tonnage, with some exceptions (IMO, 1974). It is considered the best available identification system for vessels because it is a number permanently implemented on the ship's hull and is never reassigned regardless of the vessel's owner, country of registration or name. This number follows the ship until it is scrapped (Vosburgh, 2017).

There are various concerns when it comes to the quality of AIS data, which will be further discussed. Some of the main concerns regarding AIS data are the quality issue of S-AIS data, general deficiency, and human errors. Overall, S-AIS data can be unreliable due to factors such as signal blockage in certain areas. Traffic variations from different time periods can have more exposure which results in a distorted view of traffic density (Smestad, 2015). General deficiencies can be a result of poorly maintained equipment or external factors. One of the external factors can be that in a time span of 24 hours, the High North and South are covered up to 15 times, whilst the central area around the equator is covered significantly to a lesser extent (Smestad, 2015). With new AIS satellites in orbit, the coverage is extended, and the data quality is increased. However, there is a possibility that interference problems occur when the satellite has an extended area to cover. Combined with high traffic and low orbiting rates over the given area, this could lead to significant gaps in the data (Næss, 2018). Moreover, human error could affect the quality of AIS data. The message contains manually reported data, such as draught, destination, ETA, route plan, and navigational status. These data can be exposed to human error, leading to deviations from reality.

3.2 Data Gathering & Handling

3.2.1 AIS Data

The AIS data was provided by Assistant Professor Gabriel Moises Fuentes at the department of Business and Management Science. The data was dated from 2013 to the end of July 2019 with worldwide coverage of tanker vessels. Due to the lack of the same time frame for other variables, the final data set excludes the year 2013. To prepare for the analysis, R scripts were created to load and increase the manageability of the raw data. The raw data contained the variables shown in Table 3.1, for both bulk carriers and tankers worldwide.

The preparation of the data involved matching the AIS data with the IMO number from the VLCC/ULCC fleet registry (Clarksons WFR, 2023) only to keep the AIS data provided for the VLCC fleet. Also added via the fleet registry were the gross tonnage (GT), deadweight tonne (DWT), and the ship's designed maximum draught. The fleet registry applied is the fleet as of November 2023. This entails the probability that some VLCCs have been excluded from the earlier observations, as they might not appear in our retrieved fleet registry. However, our conclusion is that this is negligible for the findings. Table 3.2 shows the size of our data after filtering with the scripts explained above.

Table 3.1: AIS data set

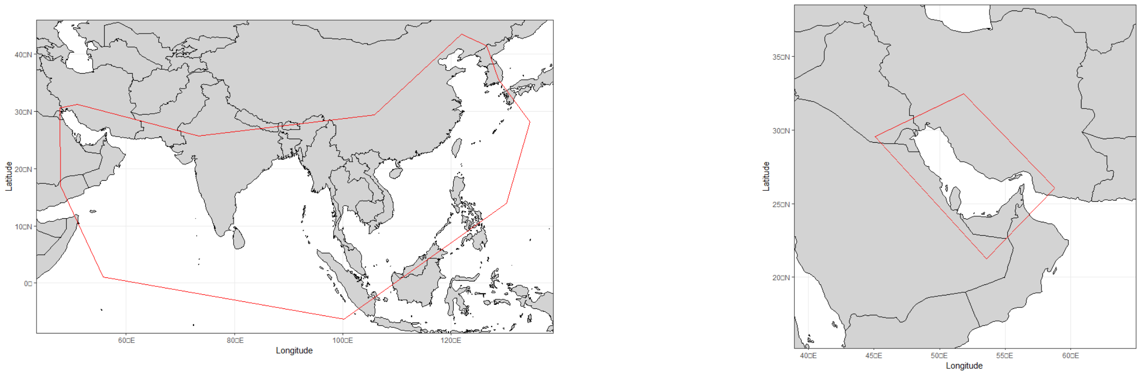
Variable	**Description**
"name"	Latest known AIS name (not historical)
"imo"	IMO Number
"mmsi"	Latest known MMSI
"callsign"	Latest known IMO
"length"	Length in meters as sent via AIS
"width"	Width in meters
"timestamp_position"	Timestamp that belongs to the position
"source_position"	T-AIS or S-AIS
"lon", "lat"	Position
"speed"	Speed in knots
"course"	Course over ground in degrees (0 north, 90 east,...)
"heading"	Compass course, direction of the bow of the ship. 511 if no compass is connected to AIS
"nav_status"	AIS-navigational status
"timestamp_voyage"	Timestamp where the voyage data (eta / destination / draught) was received
"source_voyage"	T-AIS or S-AIS
"draught"	Draught in meters
"destination"	Destination text as sent by the ship
"eta"	ETA as sent by the ship.
"aisshiptype"	Shiptype as described in AIS
"shiptype"	Additional shiptype we know for most vessels with IMO number
"dimA", "dimB", "dimC", "dimD"	Distance from the GPS antenna on board to all 4 directions. Length is DimA + DimB, width is C + D

Table 3.2: Data Size

Year	Observations
2014	7,654,309
2015	13,557,193
2016	15,188,376
2017	13,389,346
2018	13,558,817
2019	7,945,072
Total	71,293,113

In addition to sorting for VLCC, we wanted to create variables based on position. To achieve this, we applied geographical polygons, presented in Figure 3.1. These were used, together with the reported ship's position, to identify those within the selected areas.

Figure 3.1: Polygons



(a) Polygon of "local" area

(b) Polygon for the Arabian Gulf (AG)

3.2.2 Non-AIS Data

The time-series data for the non-AIS variables are retrieved via Refinitiv Eikon (2023) or Clarksons Research's Shipping Intelligence Network (SIN) (2023). The dependent variable, for instance, the freight rate from the Arabian Gulf to China, was gathered via Eikon through the ship-brooking firm Simpson Spence Young (SSY). The handling of this data is explained in section 3.2.3.

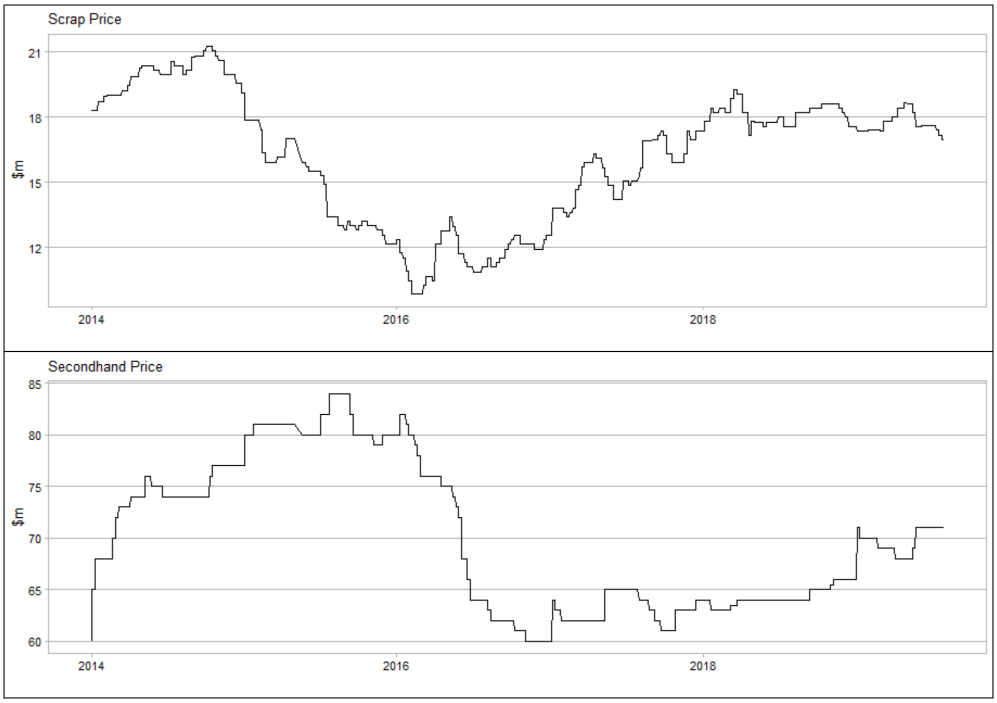
3.2.3 Missing Data & Other Challenges

During the data investigation, we discovered only a few missing values for the destination and estimated time of arrival (ETA) columns. Nevertheless, most of the other information, such as speed and position, was present. We decided to keep the observations with missing data as it is not uncommon for a ship to have a non-defined destination or an ETA.

The manual input of the destination also presented other issues, mainly abbreviations and spelling errors. This led to the need for corrections, which was done with an R script; an example of the type of corrections can be seen in Table 3.3. Further explanation and usage of destination can be found in section 3.3.2.

Concerning the financial variables, a reoccurring problem was missing values concerning weekends and holidays. We decided to continue the last observed observations until a new one occurred. Meaning that an observation n would be repeated daily until a new observation m replaced it. When converting it into daily observations, the same method was applied to weekly data. This is because decision-makers have the same available information as our data after transforming. Another aspect supporting this transformation is the low short-term variation, especially in the scrap- and secondhand prices, as seen in Figure 3.2.

Figure 3.2: Time-series of Scrap & Secondhand Price



3.3 AIS Derived Variables

3.3.1 Fleet Productivity

In Table 2.1, Stopford (2009) identifies fleet productivity as one of supply's most important factors. This productivity largely depends on the speed of the ships and their utilization of cargo capacity (DWT). To capture fleet productivity, we have included variables for speed and load factor (cargo utilization), both for the local and global fleet.

The inclusion of these variables is based on the fact that when there is excess demand, the fleet will be fully employed, and freight rates will be high (Koekebakker et al., 2006). Profitability can be further improved by increasing fleet productivity. Thus suggesting a link between high freight rates and high sailing speed and load factor. On the other hand, theory suggests that increased supply puts downward pressure on freight rates. However, this depends on the current demand. Another factor to consider is the increased cost of operation that occurs as a consequence of increased fleet productivity. Again, this depends on factors such as bunker price and attainable freight rate (demand). As a result of a significant number of factors to consider there is uncertainty regarding the direct effect on the freight rate.

The speed variable was derived from the average of all daily speed observations. The load factor was derived as the average of the current draught divided by the maximum draught for all observations per day. These methods were applied to the whole data set, as well as to the one filtered for a position within our local polygon, to derive the global and local measures.

A possible weakness of the load factor is the manual input of the draught, as it could be wrongly reported. In addition, ships usually ballast when not transporting cargo, which could lead to a falsely high load factor. Regarding the speed variable, it could be affected by ships at port, which ideally should be excluded as it do not reflect an actual decision.

3.3.2 Local Fleet Activity

To account for the local fleet activity and, more specifically, the local supply, we created variables to measure the number of ships in the area. This was done in two ways. The first is the number of ships in the Arabian Gulf (AG), and the second is the number of ships headed for the AG. The reasoning behind them is that this represents the current supply as well as future development. In this context, we expected a negative relationship between the freight rate and the number of ships in and headed to the AG. More ships looking for cargo leads to bidding between the owners, putting downward pressure on freight rates. On the contrary, fewer ships would lead to bidding amongst the charterers.

When deriving the measures, we used the polygon for the AG in order to count the number of ships with positions reported within the area on a daily basis. The variable of ships headed to the AG was found by matching their reported destination with a port in the area. This entailed some corrections as mentioned in section 3.2.3, and an example is provided in Table 3.2. For this variable we counted the number of ships by their expected time of arrival (ETA), sorted on a daily basis.

Possible weaknesses include the number of ships present not accounting for their current status, i.e. whether the ships are fixed or not. The excess supply is better represented by those present without being fixed. Furthermore, there is almost a guarantee that more ships are headed to the AG than those we have reported, as we have relied solely on destination. However, we still believe that the variables provide good information about the current situation.

Table 3.3: Destination Correcting Example

Correct Destination Name	AIS Destination Input
ARABIAN GULF	PER GULF P.GULFP P.GULF FOR ORDER PERSIAN GULF-KHBK!P AR.GULF ORDER PERS.GULF P_GULF PERSIAN GULF FOR ORD PER.GULF A.GULF FOR ORDERS P-GULF FOR ORDERS ARABIAN GULF 4 ORDER A.GULF PER...GULF P GULF P G

3.4 Non AIS Derived Variables

3.4.1 Supply and Demand

When examining supply-driven factors, we identified bunker price, oil price, secondhand ship price, and scrap price as key elements. Beginning with bunker price, it is widely recognized as a primary determinant of a ship's voyage cost (VOYEX). In their pursuit of maximizing profits, ship owners are influenced indirectly by bunker prices, as optimizing fuel consumption is a critical component of profit maximization. When demand surpasses supply, there is an incentive to increase supply by accelerating sailing speed. However, this results in higher fuel consumption. The price of bunkers, therefore, sets a threshold beyond which it is not profitable to boost speed. While the impact of bunker price is significant, its direction on supply is ambiguous. Higher bunker prices could lead to reduced supply, potentially driving rates up.

Moving to oil prices, we note that bunker prices often mirror oil price fluctuations. Poulakidas and Joutz (2009) observed that rising oil prices exert upward pressure on weekly spot rates. Conversely, Stopford (2009) suggests, following classic supply and demand theory, that an increase in oil price may shift the energy mix towards alternatives like coal, thereby reducing tanker demand. This leaves us with uncertain effects on the freight rate from both bunker and oil prices.

When considering secondhand ship prices, it is crucial to note that while the actual trading of ships do not alter fleet size, these prices often rise in high freight rate markets. Owners bid higher prices to expand their fleets and leverage high rates, suggesting a positive relationship between secondhand prices and freight rates.

As for scrap prices, they play a different role. Higher scrap prices usually encourage more ship scrapping, thus reducing supply. However, this decision is influenced by current market conditions, operational profitability, and the state of the secondhand market. Furthermore, scrap prices also depend on the external demand for metal scrap. While a positive relationship between higher scrap prices and reduced supply is expected, it is a complex interplay of various factors, making a strictly positive relationship to freight rate overly simplistic.

3.4.2 Financial

With regard to the financial variables, we incorporated the exchange rate between the U.S. dollar and the Chinese Yuan to address the currency risk highlighted by Kavussanos and Visvikis (2006). This inclusion is pertinent because revenue is often in dollars, while expenses and income are typically in foreign currencies, leading to a currency effect. Specifically, strengthening the dollar against the yuan is expected to result in higher freight rates. However, it is important to note that a stronger dollar might also encourage new-building, thereby increasing supply, although with a delayed effect. As a result, we anticipate a positive relationship in this context. We initially considered including the U.S. dollar to the Saudi Riyal exchange rate but excluded it due to the Riyal's peg to the dollar.

Moreover, we included the Baltic Dirty Index (BDTI). It reflects the price level of all major routes, including TD3C. We expect a positive relationship with the freight rate.

Table 3.4: Variable Descriptions

Variable	Unit	Description
Dependent Variable		
Rate	WS	Spot freight rate on the TD3C route (MEG-CHN).
AIS Derived Variables		
speed_global	knots	Average speed of global VLCC fleet
speed_local	knots	Average speed of local VLCC fleet
lf_global	fraction	Average load factor (dwt utilization) of global VLCC fleet
lf_local	fraction	Average load factor of local VLCC fleet
n_in_ag	Integer	Number of VLCC's at origin port location; the Arabian Gulf (AG)
n_to_ag	Integer	Number of VLCC's headed to AG
Non-AIS Derived Variables		
fx_usd_cnh	USD/CNH	U.S dollar - Chinese Yuan exchange rate
bdti	index	Baltic Exchange Dirty Tanker Index
oil_price	$/bbl	Price of Brent crude oil
bunker_price	$/mt	Bunker oil price in Singapore
sechand_price	$m	Second-hand price of 5-year-old VLCC
scrap_price	$m	Estimated VLCC scrap value

3.5 Descriptive Statistics

In this section, we present descriptive statistics in addition to correlation matrices for our variables. The data is a daily time-series from 2014 to the end of July 2019. The resulting data after processing consists of 2037 observations across 13 variables. In addition, time-series plots of all variables can be found in Appendix A.

Regarding skewness, a positive number indicates a longer right tail, and a concentration of the distribution on the left, and vice versa. While kurtosis is a measure of the presence of outliers, meaning that a higher kurtosis implies a higher probability of of extreme observations occurring. Generally higher absolute values of skewness and kurtosis indicate that the distribution is non-normal. This can be a problem as some models assume that the variables are normally distributed. Skewed distributions could therefore negatively affect performance. Another issue to consider is bias, as the model could give more weight to extreme values, leading to poor generalization.

3.5.1 AIS Derived Variables

For the AIS-derived variables, the relevant Tables are 3.5 and 3.6, descriptive statistics and correlation matrix.

Looking at the speed variables, we observe a positive correlation with the rate. Although we presented some uncertainty regarding the effect of speed, the positive correlation is not surprising. The more surprising aspect is the large difference between them, as the global speed has a considerably higher correlation, while local speed exhibits almost no correlation at all. Nevertheless, the global speed correlation of 0.28 suggests that there is a link to the rate. This is aligned with the theory about fleet productivity, and the measure seems to capture this aspect. With regards to the local speed, we suspected that the low correlation was due to the calculation of the measure being affected by the number of ships in the AG. We expect low sailing speed for ships awaiting cargo, and the large number of ships in the AG will likely affect our local speed measure. This can be observed in Table 3.5, and could eliminate some of its relevance. A solution to this could be to only include vessels that have been fixed (arranged for cargo) in the speed variable while excluding them in the measure of ships in the AG. However, we did not have access to such data.

When assessing the load factor variables, we noted a positive correlation for both the local and global variables. This is in line with our expectations. Again we noted that the local measure is lower in comparison to the global, and we suspected the number of ships in the AG to be the cause. However, the discrepancy is lower, indicating that the load factor is a relevant variable to consider when predicting the rate.

Lastly, we evaluate the number of ships in and headed to the AG. We observe a negative correlation with the rate, which is according to our expectation. However, the correlation is close to negligible for both variables. This is somewhat contradicting with theory, and our expectations, as the variables should provide a good measure of the local supply.

3.5.2 Non AIS Derived Variables

The non-AIS derived variables generally have a greater correlation with the freight rate. This suggests that these variables provide more information for the prediction. However, correlation does not imply causation.

Starting with the exchange rate the variable appears to have been quite stable in the period, and the correlation is positive. The highest correlation is found between the bdti variable and the rate, which is expected as the rate of the TD3C is present in the bdti. With regard to the price variables, we observe a similar correlation for both the oil and bunker variables. We expected these to be similar, and the correlation between them is close to 1. We presented evidence supporting both positive and negative relationships, and for our sample period, we observed a negative relationship. As for the secondhand price, we observe a positive relation, although somewhat small. Scrap price, on the other hand, shows a larger correlation, but a negative one, opposite to our expectation.

Table 3.5: Descriptive Statistics of AIS Variables

	Mean	Std.Dev.	Median	Min	Max	Skew	Kurtosis
speed_global	7.89	0.96	7.78	4.91	10.96	0.57	0.48
speed_local	6.40	1.25	6.13	4.00	11.59	1.82	3.91
lf_global	0.72	0.02	0.71	0.64	0.79	0.06	-0.13
lf_local	0.67	0.02	0.67	0.60	0.74	0.06	-0.09
n_in_ag	84.22	16.42	87.00	12.00	120.00	-1.48	3.12
n_to_ag	15.43	4.06	15.00	2.00	29.00	0.17	0.06

Table 3.6: Correlation Matrix of AIS Variables

	rate	speed_global	speed_local	lf_global	lf_local	n_in_ag	n_to_ag
rate	1.00	0.28	0.04	0.40	0.27	-0.11	-0.09
speed_global	0.28	1.00	0.79	0.56	0.53	-0.34	0.03
speed_local	0.04	0.79	1.00	0.20	0.39	-0.51	0.07
lf_global	0.40	0.56	0.20	1.00	0.78	0.09	0.06
lf_local	0.27	0.53	0.39	0.78	1.00	-0.15	-0.01
n_in_ag	-0.11	-0.34	-0.51	0.09	-0.15	1.00	0.26
n_to_ag	-0.09	0.03	0.07	0.06	-0.01	0.26	1.00

Table 3.7: Descriptive Statistic of Non-AIS Variables

	Mean	Std.Dev.	Median	Min	Max	Skew	Kurtosis
rate	56.10	15.55	53.50	27.50	111.50	0.74	0.03
fx_usd_cnh	6.52	0.28	6.55	6.04	6.98	-0.01	-1.42
bdti	778.11	137.50	763.00	496.00	1344.00	0.99	1.53
oil_price	65.03	20.16	61.32	27.88	115.06	0.94	0.14
bunker_price	372.69	119.73	360.50	147.00	620.00	0.34	-0.63
sechand_price	70.03	7.21	68.00	60.00	84.00	0.36	-1.30
scrap_price	16.19	2.97	17.14	9.85	21.25	-0.39	-0.98

Table 3.8: Correlation Matrix of Non-AIS Variables

	rate	fx_usd_cnh	bdti	oil_price	bunker_price	sechand_price	scrap_price
rate	1.00	0.24	0.75	-0.34	-0.28	0.10	-0.29
fx_usd_cnh	0.24	1.00	0.06	-0.45	-0.28	-0.62	-0.33
bdti	0.75	0.06	1.00	-0.02	0.05	0.10	0.05
oil_price	-0.34	-0.45	-0.02	1.00	0.96	0.00	0.81
bunker_price	-0.28	-0.28	0.05	0.96	1.00	-0.12	0.87
sechand_price	0.10	-0.62	0.10	0.00	-0.12	1.00	-0.02
scrap_price	-0.29	-0.33	0.05	0.81	0.87	-0.02	1.00

4 Methodology

This chapter explores the selected forecasting techniques and machine learning methodologies implemented in the analysis. The machine learning models subjected to evaluation in this thesis entail, for instance, artificial neural networks, selected for their demonstrated effectiveness in complex time-series forecasting. See studies done by (Li and Parsons, 1997), Kulkarni and Haidar (2009), and Gao and Lei (2017) for further reference. Additionally, there are other machine learning models selected. The primary emphasis of this thesis lies in the potential additional insights gained by incorporating machine learning methodologies combined with AIS data in forecasting.

4.1 Machine Learning

Machine learning dates back to the 1940s when McCulloch and Pitts (1943) laid the foundation for what has evolved into today's machine learning and artificial neural networks. Its adoption has increased in recent years as a result of advancements in computer capabilities, enhanced computational power, and the accessibility of vast amounts of data. The principles of machine learning models revolve around the application of model refinement, where it can make predictions and decisions without the requirement for explicit programming. Moreover, machine learning models consist of an application of a function f to an input x, resulting in the corresponding optimal output $\hat{y}$ as expressed by the simple equation.

$$\hat{y} = f(x) \tag{4.1}$$

The output $\hat{y}$ represents the optimal value of the true value of y. To get the optimal representation, it involves instructing on the anticipated output subsequent to a specific input. This could be conveyed as two sets of map features X and Y, where

$$\hat{y} = f(x) \quad y \in Y \quad \forall x \in X \tag{4.2}$$

The model receives feedback on the inequality between its predicted and expected output throughout the training phase. Consequently, the network iteratively adjusts its parameters to minimize its error. This paradigm characterizes supervised learning since both the input and output are equally known (Roth, 2016). On the contrary, unsupervised learning transpires only when the input is apparent, with the objective of uncovering formerly unknown patterns in the data. Given that this thesis centers on estimating the optimal mapping function between input and output variables, it aligns with the principles of supervised learning. In the forthcoming sections, the different methodologies used in the thesis are presented.

4.1.1 Artificial Neural Network

An artificial Neural Network, also known as neural networks (NN), emulates the structure and functions of biological neural networks through mathematical models. The basis of the model is governed by three sets of rules: multiplication, summation, and activation (Suzuki, 2011). Moreover, it is distinguished by its non-parametric, non-linear, and assumption-free nature (Kulkarni and Haidar, 2009). More specifically, it avoids making predetermined assumptions about the problem, making it well-suited for addressing complex issues e.g., time-series problems. The network is composed of layers and nodes, wherein complex computations unfold. Figure 4.1 illustrates the basic architecture of a neural network.

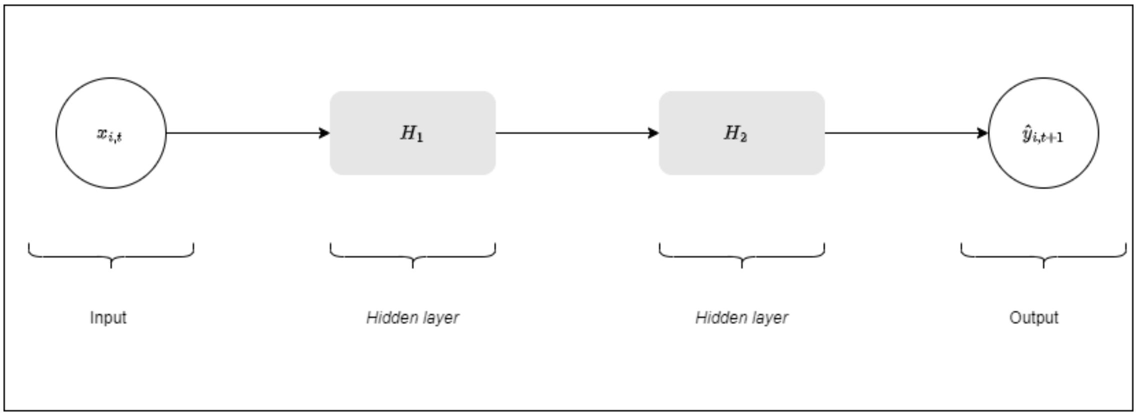


Figure 4.1: Basic Neural Network architecture

The interconnection of individual neurons within a network is referred to as the architecture of the neural network. There are two underlying architectural classes: feedforward architecture and recurrent architecture. In feedforward architecture, information flows unidirectionally from inputs to outputs. Opposed, the recurrent architecture allows the information flow to go the opposite direction, which creates a source to evaluate and train on information that has been through the flow.

The equation displayed underneath represents the mathematical expression.

$$\hat{y}_{i,t+1} = F(X)\left[\sum_{i=1}^{n} w_{i,t}x_{i,t} + b\right] \quad \forall\, n \in \mathbb{N} \tag{4.3}$$

where $\hat{y}_{i,t+1}$ is equal to the output variable in discrete time $(t+1)$, $F(X)$ is the activation function, $x_{i,t}$ is the input variable, $w_{i,t}$ is the weight value, and b is the bias.

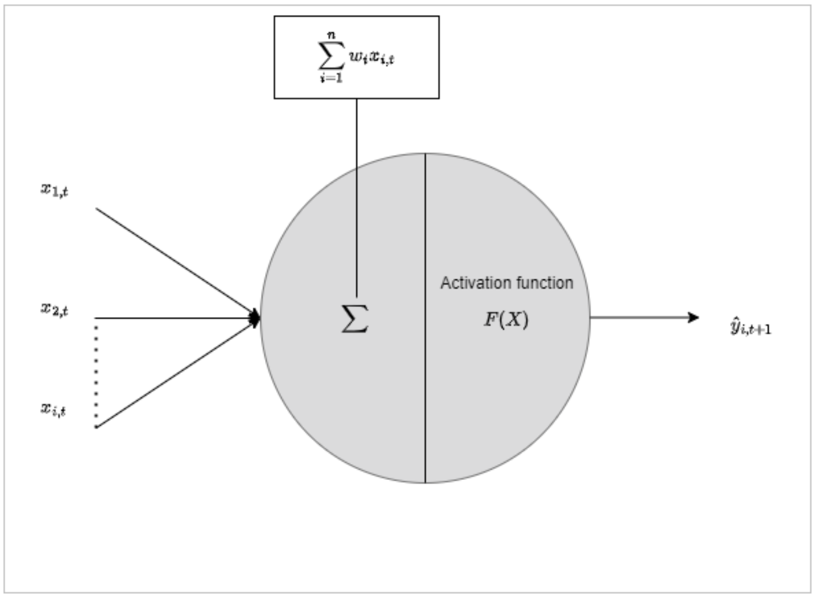


Figure 4.2: Neuron in hidden layer

In Figure 4.2, the layout of the individual neuron is displayed. The neuron combines input $x_1, x_2, ..., x_i$ from i neurons, multiplied with a weight that amplifies or dampens the respective input. Next, it gets through an activation function that determines the properties of the neuron. The activation function is typically selected from a prerequisite set of functions, including the Step, Linear, and Sigmoid functions (Suzuki, 2011). Each depends upon the specific problem which is going to be solved. The Step function produces a binary output based on a threshold and is suitable where clear decisions are needed. Linear function produces an output proportional to the input, making it suitable if a linear relationship exists between the inputs and outputs. Sigmoid function is usually applied when there is non-linearity in the model, and when tasks have complex patterns or relationships that need to be captured.

An overview of a more sophisticated neural network architecture is displayed in Figure 4.3

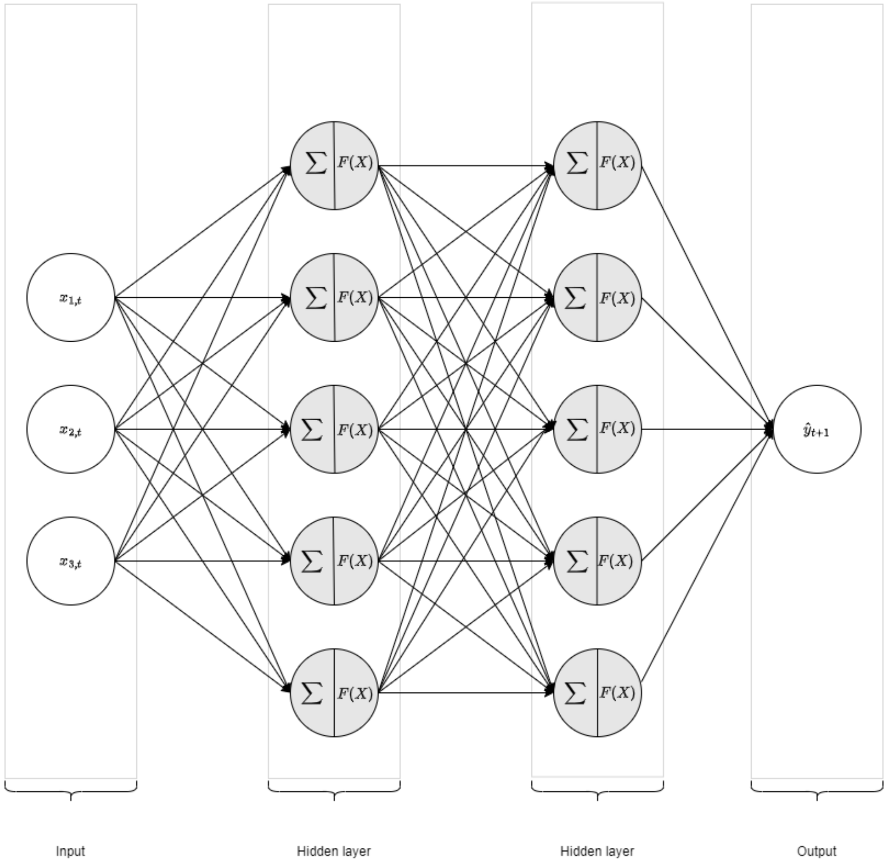


Figure 4.3: Advanced Neural Network architecture

4.1.2 Long Short-Term Memory

Long Short-Term Memory (LSTM) differs from traditional feedforward NN since it can preserve information from previous events. More specifically, it is a sub-model of Recurrent Neural Networks (RNN) which have the ability to utilize previous information from other networks combined with new information to create a new output. Albeit, the original RNN can encounter a problem called the vanishing gradient problem (Yilmaz and Poli, 2022). This originates from the backpropagation of the gradients, where the gradients vanish, which results in low performance of the model. LSTM solves this issue by implementing a four-network layer, also called gates, to store and forget information.

Table 4.1: LSTM unit gates

Number	Gate
1	Forget gate
2	Input gate
3	Update gate
4	Output gate

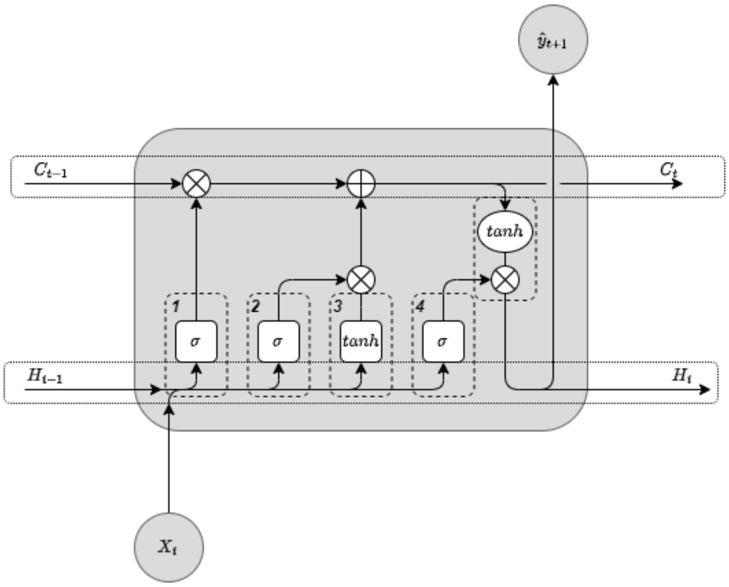


Figure 4.4: LSTM unit

In the forget gate, the unit undergoes a decision on which information to discard and store. The input gate and update gate modify the previous and current information through two different activation functions, *sigmoid* and *tanh*. Lastly, the output gate determines the information that goes through and into a new unit. This ongoing evaluation of information prevents unnecessary information from being stored, eventually degrading the performance of the model.

4.1.3 Linear Regression

Linear regression is, in its simplest form, a linear relationship between the dependent variable y and an independent variable x. When there are two or more independent variables, it is defined as a multiple linear regression model.

The mathematical notation is given by:

$$\hat{y}_t = \sum_{t=1}^{n} \beta_t x_{i,t} + \varepsilon_t \quad \forall i = 1, 2, ..., m \tag{4.4}$$

Here, the dependent variable is represented by $\hat{y}_t$ given by the sum of the independent variables x_i together with its estimated coefficient β_t which impacts whether the dependent variable is going in the positive or negative direction for one unit-change of x. Lastly, we have an error variable ε_t, which is represented by the difference between the observed data and the predicted data. The overall objective is to minimize the error or, more specifically, fit a line that has the shortest distance between the observed and predicted value.

4.1.4 Ridge Regression

Ridge regression is frequently used in multiple regression where there is an instance of multicollinearity. Multicollinearity occurs when a high correlation exists between two or more independent variables. There are several adverse effects of multicollinearity. First, the standard error increases, which results in unreliable parameter estimates. Second, higher sample variability will vary from one sample to another, which causes unreliable cross-validation estimates between predictions. Third, independent variables may be eliminated because it is determined as insignificant (Kidwell and Brown, 1982). Solving this issue involves introducing a bias into the equation and generating new lines that do not align with the original training data. This necessitates balancing bias and variance for improved performance and sustained accuracy in long-term forecasting.

The objective is to minimize the penalty of the sum of squares. The equation, with inspiration from Tibshirani (1996) is equal to:

Minimize:

$$\sum_{i=1}^{n}\left(y_i - \sum_{i=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{4.5}$$

Subject to:

$$c > 0, \sum_{j=1}^{p}\beta_j^2 < c \tag{4.6}$$

The constraint works alongside a penalty λ which penalizes the optimization function if there are instances of large β values. Instead of removing variables or degrading the variables to zero, there is an instance of a penalty for those variables that are significantly far away from zero.

4.1.5 Lasso Regression

Lasso Regression, or Least Absolute Shrinkage and Selection Operator, differs from Ridge regression by which the β is a constraint. It takes the absolute size of the regression parameters and adds it up as a penalty.

If we compare to the Ridge regression model, which has a constraint equal to $\sum_{j=1}^{2} \beta_j^2$, the lasso model will be subject to $\sum_{j=1}^{2} \left|\beta_j^2\right|$.

The equation, with inspiration gathered from Tibshirani (1996) is equal to:

Minimize:

$$\sum_{i=1}^{n} \left(y_i - \sum_{i=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j^2\right| \tag{4.7}$$

Subject to:

$$c > 0, \sum_{j=1}^{p} \left|\beta_j^2\right| < c \tag{4.8}$$

4.1.6 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric regression algorithm where predictions are based on the length of a point distance to a training set, defined as the target variables (Rigopoulos, 2022). Generally, a distance metric is one of the central mechanisms concerning KNN. The most common is the Euclidean distance, which measures the relationship between two data samples. The idea of KNN is to predict an output value of a given input sample, together with a smoothing parameter rooted in the input-output training samples (Kumbure and Luukka, 2022). The main objective is to find the optimal value of the predicted variable $\hat{y}$, which in mathematical denotation is equal to

$$\hat{y} = \frac{1}{k} \sum_{j=1}^{k} y_j \tag{4.9}$$

Before we estimate the arithmetic mean of the output variable y_i from $j = 1, 2, \ldots, k$, we have to find the distance, which makes it possible to calculate the optimal variable $\hat{y}$. This is enabled if the dataset containing training variables T equals $T = (X_i, y_i) \mid i = 1, 2, \ldots, N$, where N is the number of samples. Next, we let $X_i = x_{i,1}, x_{i,2}, \ldots, x_{i,m} \in R^m$ which is a given input sample i from a dimensional feature space m. This allows us to use the Euclidean equation, which calculates the diagonal distance d between two sets of variables x_j and $x_{i,j}$ in space m.

$$d(X, X_i) = \sqrt{\sum_{j=1}^{m} (x_j - x_{i,j})^2} \tag{4.10}$$

$$N_{X,k} = (X_i, y_i) \quad \forall i = \{1, 2, ..., k\} \tag{4.11}$$

Once we obtain k cases that represent the nearest value representation compared to the training set, we want to learn the forecasting function to predict $\hat{y}$

4.1.7 Support Vector Machine

Generally, Support Vector Machines (SVM) are utilized to solve classification and regression problems. When considering time-series problems, SVM takes advantage of the general idea of classification solving. Nevertheless, the difference lies in the data-mapping where the vector is transitioned into a higher-dimensional space, separated by using hyperplanes (Cortes and Vapnik, 1995). In total, there are three hyperplanes. Two are in a parallel format on each side of the third hyperplane, which divides the classes. The main objective is to minimize the total error by identifying the hyperplane that maximizes the margin between the two-dimensional classes. The mathematical denotation is given by:

Minimize:

$$\frac{1}{2}|w|^2 + C\sum_{i=1}^{n} \varepsilon_i \tag{4.12}$$

Subject to:

$$y_i - (w \cdot x_i + b) \leq \varepsilon_i \tag{4.13}$$

$$(w \cdot x_i + b) - y_i \leq \varepsilon_i \tag{4.14}$$

$$\varepsilon_i \geq 0 \tag{4.15}$$

4.2 Data Preprocessing

Generally, many machine learning methods benefit from feature scaling. The reason for this is that the data often contain variables varying in both, magnitude, range and unit. This can lead machine learning models to interpret these variables incorrectly. To negate this problem, we need the variables or features on the same scale. There are a few methods for achieving this, and among the more common ones are standardization and normalization. Standardization centers the data around a mean of 0, with a standard deviation 1. Normalization scales the data to a set range, often between 0 and 1.

4.2.1 Normalization

The choice between these two methods depends on the data distribution. Standardization is used under the assumption of the data being normally distributed, i.e., a Gaussian distribution. Normalization is applied when this assumption fails. In order to test this we used the Shapiro-Wilks test, which simply stated is a test of goodness-of-fit test (King and Eckersley, 2019). Furthermore, it is a hypothesis test, with a null hypothesis that the sample comes from a normal distribution, and the alternative opposite. Dependent on a chosen degree of confidence, in our case a 95% - level of confidence. From Table 4.2 we can observe that the p-value associated with each observation, apart from lf_local, is much smaller than a significance level of 0.05, thus rejecting the null-hypothesis and concluding that the variables are not following a normal distribution.

The distribution of the data is used as the decider upon which scaling method to use. With different scales of the features the normalization is an important step, in order to get statistically sound models that can generalize the research problem. In addition, normalization does not change the shape of the variable distribution (Næss, 2018). We therfore adopted a simplistic approach with regard to comparison of models and applied normalization in all models.

Table 4.2: Shapiro-Wilk Test

Column	Statistic	P-value
lf_local	0.998417	0.049958
speed_global	0.968272	8.73098e-21
speed_local	0.831761	5.70749e-42
n_in_ag	0.893882	2.43192e-35
lf_global	0.995742	1.55926e-05
n_to_ag	0.992649	1.39266e-08
fx_usd_cnh	0.923496	5.40207e-31
bdti	0.942124	1.59582e-27
oil_price	0.900632	1.91848e-34
bunker_price	0.963449	2.78916e-22
sechand_price	0.902382	3.337e-34
scrap_price	0.943938	3.848e-27
rate	0.951356	1.79638e-25

4.3 Evaluation of Models

To evaluate the different models, we split the data into training and test, we utilized a train-test split of 80-20, which means that 80% of the data was used for training the models, while the remaining 20% was used to make predictions. As the goal is to assess the performance of a multitude of machine learning methods, we need metrics to assess their relative performance. For this purpose, we used the following evaluation metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared. While there is no consensus upon which metrics are most appropriate, the mentioned are often used when evaluating prediction models (Zhang & Hu, 1998).

Both the MAE and the RMSE provides a score in the same unit as the target variable, but due to the aforementioned normalization that unit is changed. As a consequence, the interpretability is somewhat poorer. However, the use of these metrics is for comparing the relative performance of the models, and they still provide valuable insights.

4.3.1 Mean Absolute Error

The MEA metric is a prominent metric, one of the reasons being that the error value matches the predicted target units, and the value increases linearly with the increase in error (Schneider and Xhafa, 2022). This means that a large error has the same relative effect as a small error, which sets it apart from the RMSE. The mathematical function for MAE is:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |\text{Actual} - \text{Predicted}| \tag{4.16}$$

4.3.2 Root Mean Square Error

The RMSE measures the root of the mean squared errors from the predicted values, to the actual values. Similar to the MAE, the RMSE also provides a metric of the error expressed in the same unit as the target variable, and a lower score implies better predictive performance. As it involves squaring the error value, it penalizes large errors in comparison to small errors. Its mathematical formula is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\text{Actual}_t - \text{Predicted}_t)^2} \tag{4.17}$$

4.3.3 R-squared

The R-squared or R^2 measure informs about the goodness-of-fit. In terms of regression, the measure is how well the regression line approximates the actual observations. It means that it measures the proportion of the variance in the dependent variable explained by the independent variables. The formula is as follows:

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sum of Squares}} \tag{4.18}$$

Where the numerator is the sum of the squared difference between the actual and predicted values, and the denominator is the sum of the squared difference between the observed values and the mean of the observed values. The R-squared metric, therefore, ranges from 0 to 1. A value of 0 indicates that the model does not explain any variance in the dependent variable, and a value of 1 perfectly explains it.

4.4 Cross-Validation

In order to further evaluate the performance of the models we utilized k-Fold Cross-Validation (CV). CV is a technique used to asses how well a model generalizes to independent data. The k-fold approach involves randomly dividing the set of observations in to k folds. The first fold is subsequently treated as the validation set while the model is fit on the remaining $k - 1$ folds (James et al., 2013). The evaluation metrics is further computed on the held-out fold. This process is repeated k times, and a different fold is used as the held-out, resulting in k estimates of the evaluation metrics. The final metrics are the averages across the folds.

5 Results & Discussion

5.1 Model Results

In the following chapter, the results of the models are presented using the chosen evaluation metrics. When addressing a Time-series split, the data is split by its index, and training is done on the first 80% of observations and testing on the last 20%. When a random split is used, the train-test split picks random observations to train and test. The same split sizes are used in both cases, and a random state was used to ensure that all models are evaluated on the same train and test data. In the following, we provide tables of the evaluation metrics for each model. Plots of the predicted vs. the actual rate can be found in Appendix B.

Table 5.1 shows that all models except KNN have a relatively high R-squared measure, thus explaining the dependent variable's variance. Regarding MAE and RMSE, the goal is a low value, and there is some slight variation between them. Suggesting that the models have similar predictive performance.

Table 5.1: Model's with Time-series Split

Model	MAE	RMSE	R-squared
Neural Network	0.09	0.112	0.718
Linear Regression	0.085	0.105	0.753
Support Vector Machine	0.083	0.105	0.752
K-Nearest Neighbor	0.137	0.167	0.367
Ridge	0.086	0.105	0.75
Lasso	0.08	0.102	0.767
Long Short Term Memory	0.092	0.102	0.724

As mentioned, we wanted to investigate the predictive power of AIS data and check for its effect. We therefore ran the models with the exclusion of the AIS-derived variables. The results are presented in Table 5.2. Compared to the models with all variables included, we observe some minor changes. Among the most improved models are the Neural Network and KNN, SVM on the other hand, performs somewhat worse. The rest of the models only see slight differences in both directions. These findings suggest that the AIS variables have low predictive power and, in some cases, only add noise to the model, affecting them negatively. In addition, Table 5.3 shows the coefficients from the linear regression. A negative coefficient suggests a negative relationship between the variable and the target variable. As the variable increases, the target variable decreases, and vice versa. With the opposite for positive variables. The magnitude of the coefficient represents the influence on the target variable; a higher value has a more significant impact on the target variable. In our case, we can observe smaller magnitudes for the AIS variables, as expected from the minor change in the evaluation metrics. This implies less effect on the predicted freight rate. Also worth noting is the dominant coefficient of the *bdti* variable.

Table 5.2: Excluding AIS-variables, Time-series Split

Model	MAE	RMSE	R-squared
Neural Network	0.079	0.094	0.802
Linear Regression	0.087	0.109	0.732
Support Vector Machine	0.1	0.117	0.693
K-Nearest Neighbor	0.098	0.128	0.631
Ridge	0.088	0.109	0.732
Lasso	0.084	0.105	0.751
Long Short Term Memory	0.086	0.105	0.737

In the literature review we mentioned studies that found predictive power in AIS-variable, which raises the question as to why our findings do not seem to do so. A reason for this could be the deriving of the AIS variables. With regards to the speed factors, we have only accounted for the VLCC fleet, while it has a significant impact on the tanker market, the exclusion of other tankers, mainly Suezmax and Aframax vessels could be questioned. In addition, the measures of local speed are bound to be affected by other trade routes in the area, mainly TD2 from the Arabian Gulf to Singapore and TD3 to Japan. The latter is mostly done on long-term deals and, therefore, likely has a higher incentive for reducing speed and increasing profitability. Adding unrelated noise to our speed measure. With regard to load factor, we encounter the same issues, with the addition of ballasting, likely giving a falsely high estimate. Lastly, the number of VLCC's in and headed to the AG has not accounted for the use of ships as storage. Let alone the number of ships that have been fixed, which could result in numbers that do not reflect the true local supply.

Table 5.3: Coefficients from Linear Regression

Variable	Coefficient
lf_local	-0.0386544
speed_global	0.0731834
speed_local	-0.0683629
n_in_ag	-0.00596538
lf_global	0.151613
n_to_ag	-0.0255841
fx_usd_cnh	0.1849
bdti	0.799895
oil_price	0.121192
bunker_price	-0.174443
sechand_price	0.0971764
scrap_price	-0.0210375

With this in mind we believe more detailed work with regards to AIS data could improve its applicability. However, such analysis is beyond the scope of this thesis.

5.1.1 Cross-Validation Results

In Table 5.4 are the results from the k-fold Cross-Validation (CV) performed. In comparison to a single train-test split, the use of CV gives a better measure of how the model will generalize to unseen data. As the estimates prove, the models retain good predictive power when tested on different samples of the data. Suggesting that they do in fact generalize well to unseen data.

A consequence of using CV is the deviation from the time-series split of the data used previously. While there is consensus about shipping cycles, there is also consensus about each cycle being different. Therefore in order to capture all aspects of a cycle, the data needs to be sufficiently large. With this in mind the argument could be made to use a random split, as this would train the model on all parts of a shipping cycle, while using a time-series split could end up removing parts of a cycle. Although our models seem to perform well with both split types, suggesting that they generalize well to unseen data. However, due to the nature of shipping it is not given that they would perform well on

Table 5.4: Results from Cross Validation

Model	MAE	RMSE	R-squared
Neural Network	0.065	0.085	0.787
Linear Regression	0.078	0.102	0.692
Support Vector Machine	0.064	0.078	0.82
K-Nearest Neighbor	0.054	0.073	0.842
Ridge	0.079	0.102	0.692
Lasso	0.079	0.104	0.684
Long Short Term Memory	0.059	0.078	0.827

future data. An example relatively fresh in mind are the extreme freight rates achieved during the early month of the covid-19 pandemic. Showcasing a situation that likely would be hard for any model to predict.

6 Conclusion

The main objective of this thesis has been to investigate the use of AIS data in predicting a route-specific freight rate. This route was the TD3C, originating in the Arabian Gulf with the destination in China. In order to investigate this, we implemented a series of machine learning methods. Consequently, it was essential to verify the predictive power of the AIS variables in various models due to potentially differing results. To address this, we utilized, as mentioned, diverse machine learning models to assess the effectiveness of AIS variables in the prediction. This approach also considers the variability in model performance and the potential added value.

The sample period used is from 2014 to the end of July 2019. From a shipowner's standpoint, the ability to anticipate the fluctuations in freight rate greatly influences their profitability, highlighting our thesis's relevance. Furthermore, this thesis adds to the existing literature by implementing AIS-derived variables and non-AIS data in a series of machine learning models.

Through our analysis, it was found that AIS variables offer some insight into freight rate movements. However, compared to the other chosen variables, they fall short, and for some models, they reduce the overall predictive power. This could be due to the weaknesses of AIS data and the methods applied when deriving the variables, and a more in-depth analysis of AIS data could be beneficial. Specifically, AIS data is on a per-ship basis, which provides valuable information regarding the individual ship. However, when creating variables for machine learning, aggregating them on a per-day basis is necessary. This could lead to helpful information getting lost in the process. Nonetheless, based on our findings, AIS data provides some valuable insight, and the implementation could be beneficial.

Generally, all models show good predictive power, implying that they explain the variance in the freight rate. This was true when applying a time-series train-test split and k-fold cross-validation, underscoring the relevance of machine learning methods in shipping.

Further research could explore more advanced machine learning methods, for example, by adding less quantifiable variables such as political events (e.g., sanctions and wars) to the modeling. In addition, other routes and shipping segments could be explored. While our thesis found that the AIS variables have negligible additional value, other papers suggest the opposite. Thus, further research and a more thorough analysis of AIS data are proposed. Additionally, incorporating a more extended sample period could enhance the overall ability to capture the cyclicality of shipping. Lastly, the use of other independent variables is also encouraged.

7 Appendix

7.1 Appendix A - Time-series

Figure 8.1 to 8.7 includes time-series plots of all variables. The dependent variable (Freight Rate) is illustrated in Figure 7.7

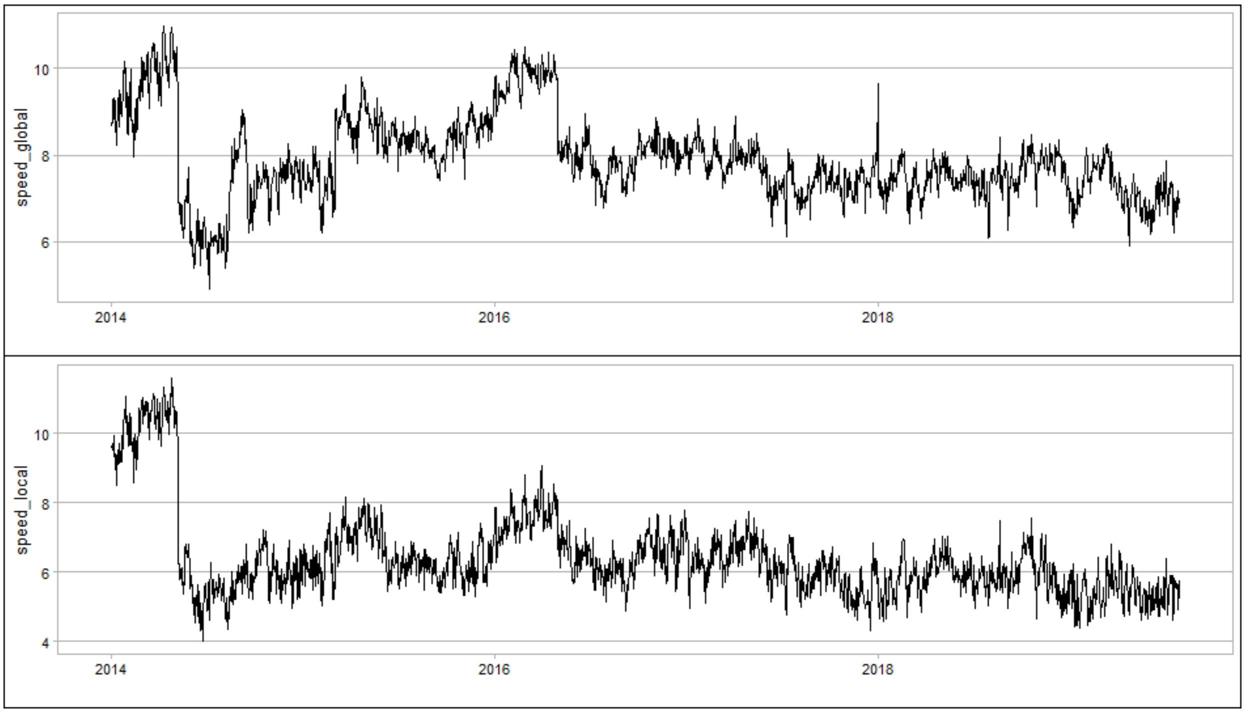


Figure 7.1: Speed variables

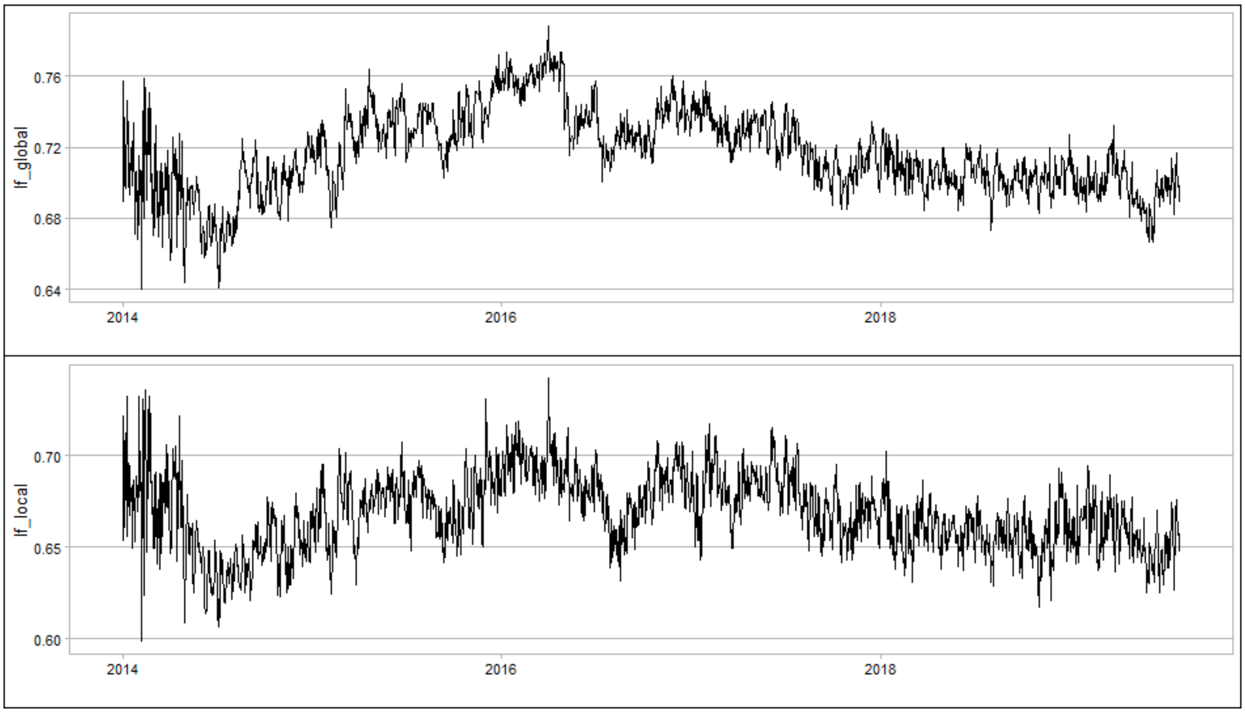


Figure 7.2: Load Factor variables

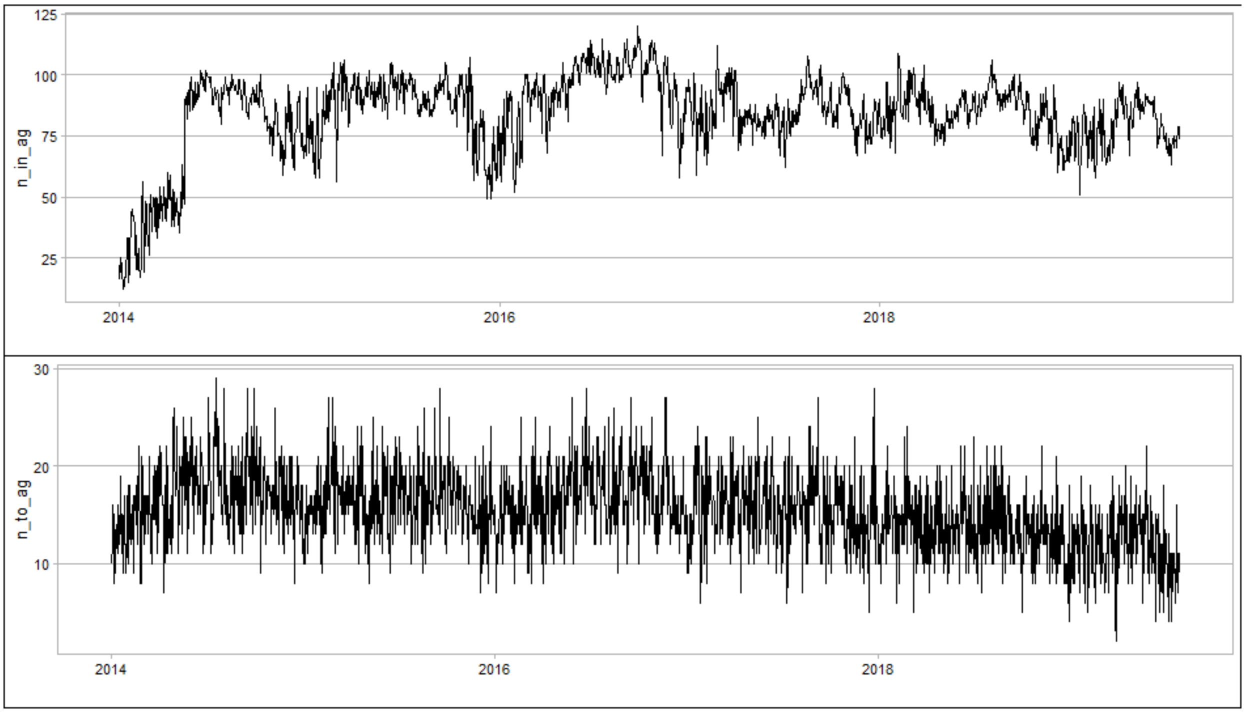


Figure 7.3: Ships in/to AG variables

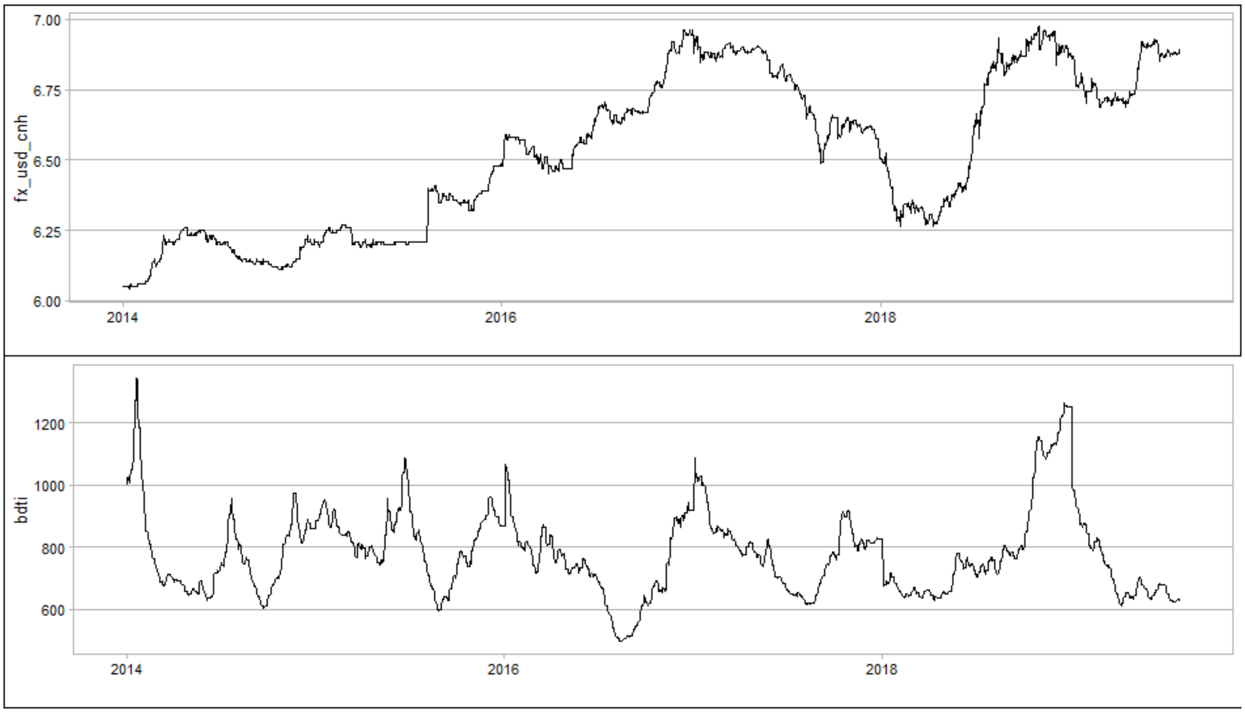


Figure 7.4: Exchange Rate & BDTI

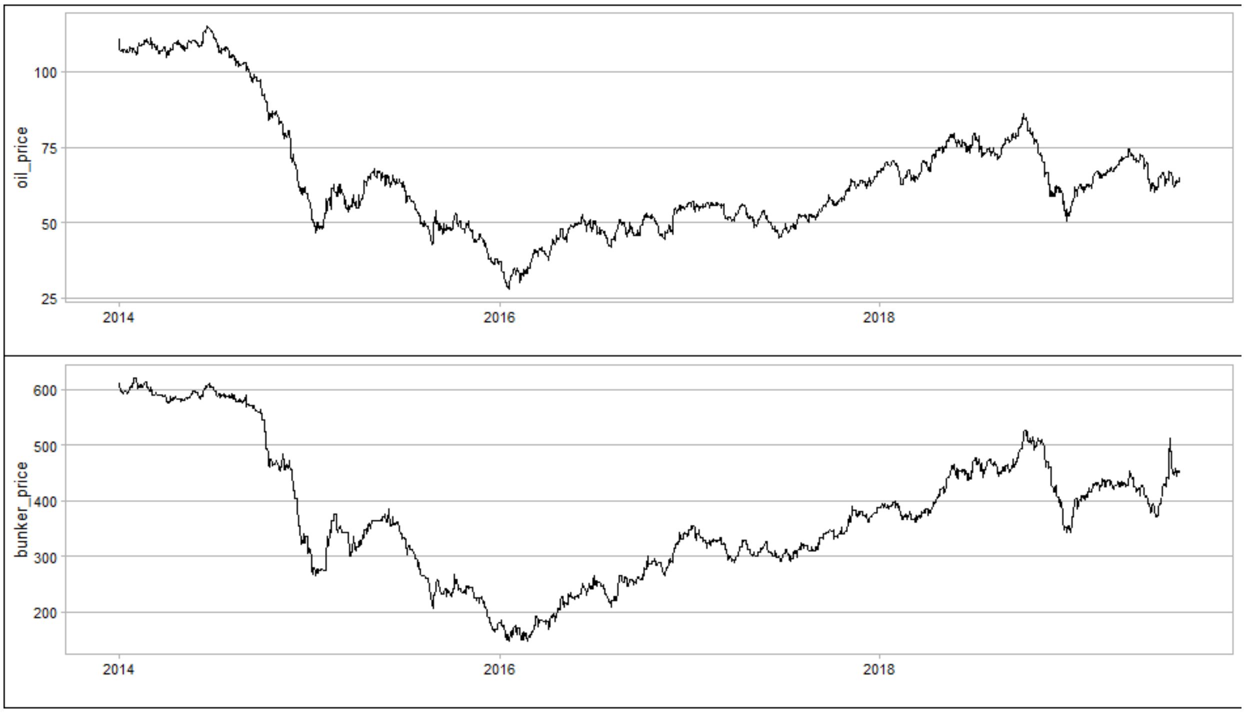


Figure 7.5: Bunkers & Oil

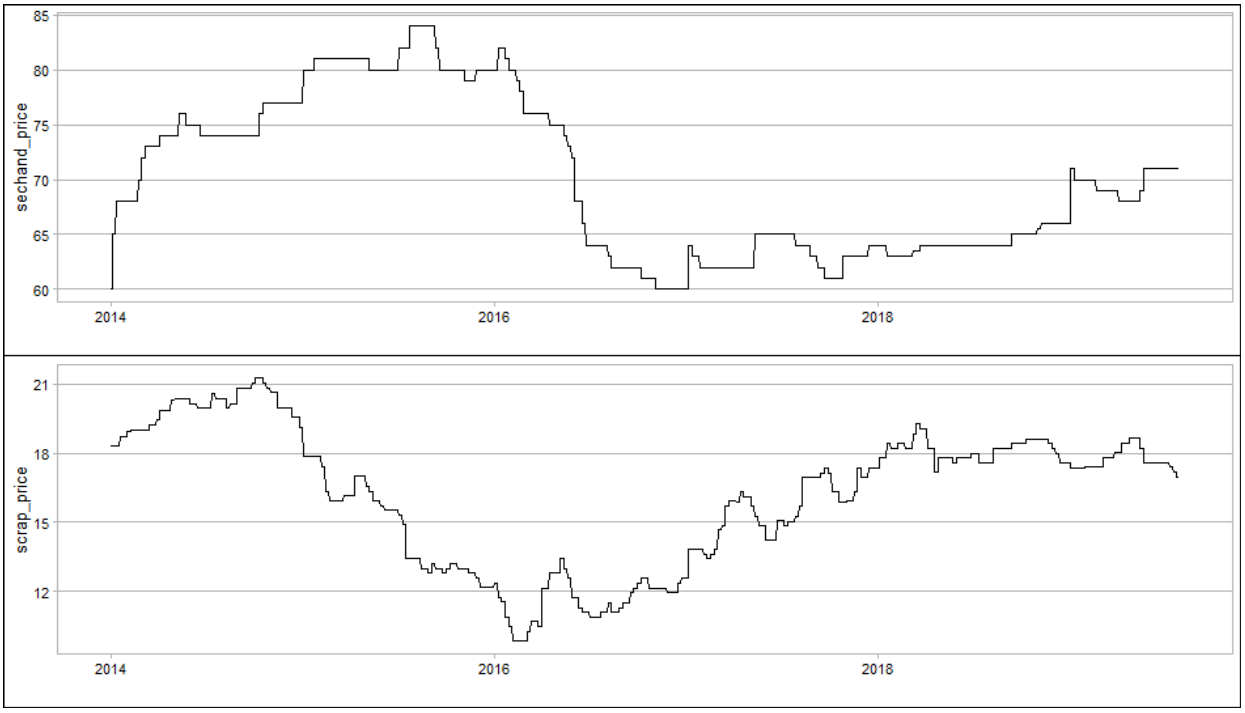


Figure 7.6: Secondand & Scrap

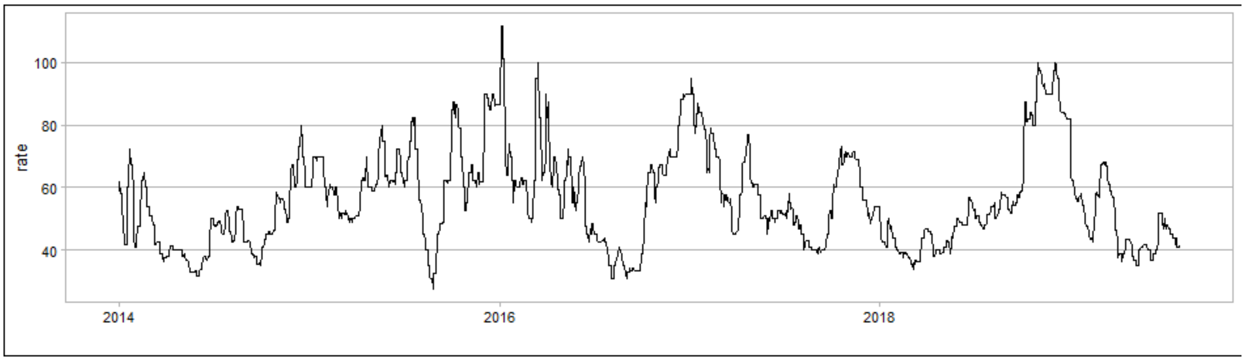


Figure 7.7: Freight Rate

7.2 Appendix B - Predictions vs Actual

Figure 8.8 to 8.14 includes the graphs of the actual values of the rate for the last 0.2 observations of our data with the predictions by the different trained models. These are from the models with all variables included.

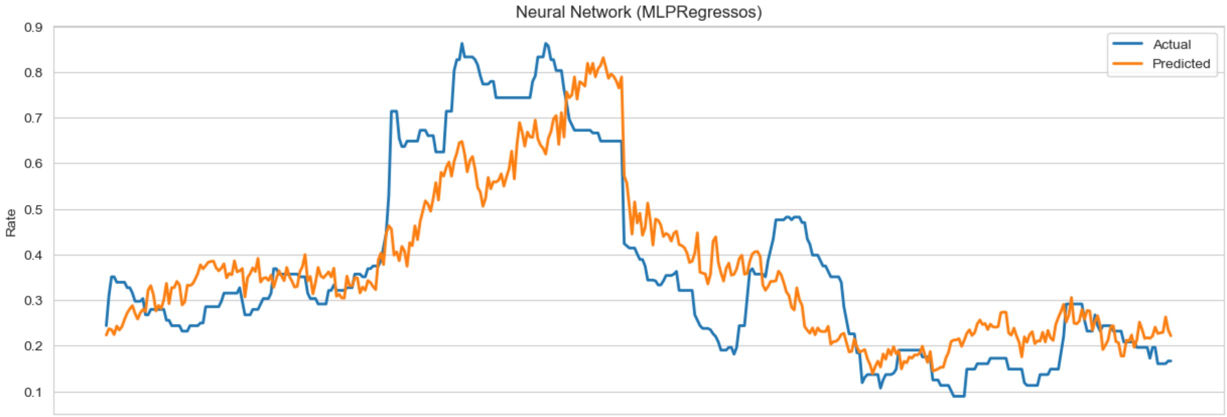


Figure 7.8: Neural Network

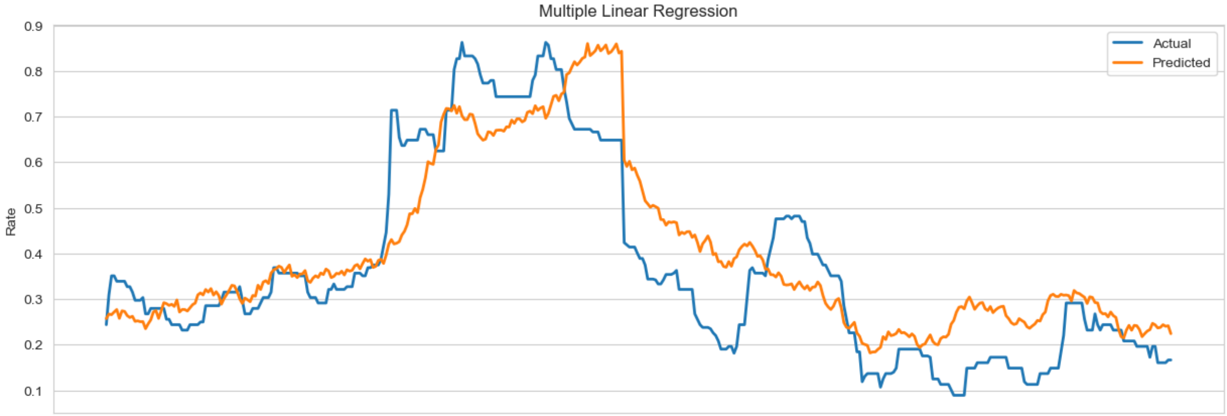


Figure 7.9: Multiple Linear Regression

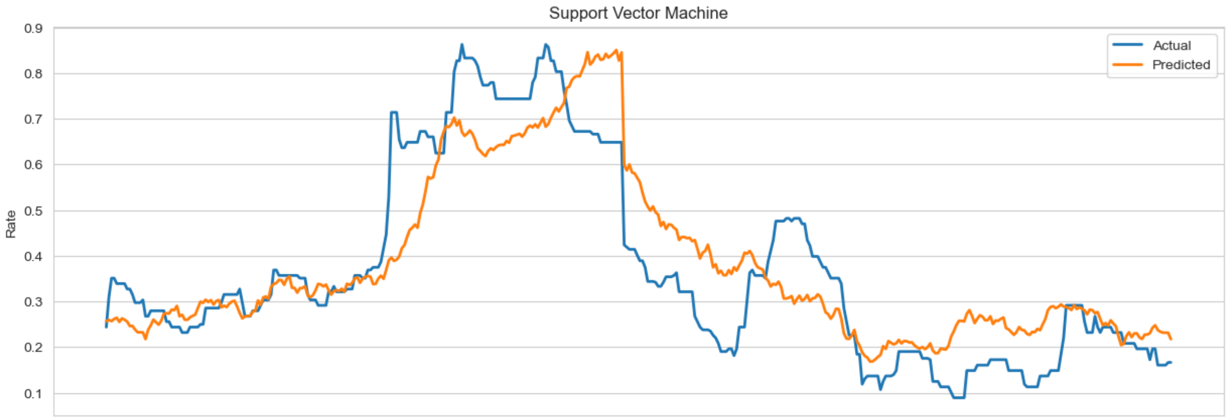


Figure 7.10: Support Vector Machine

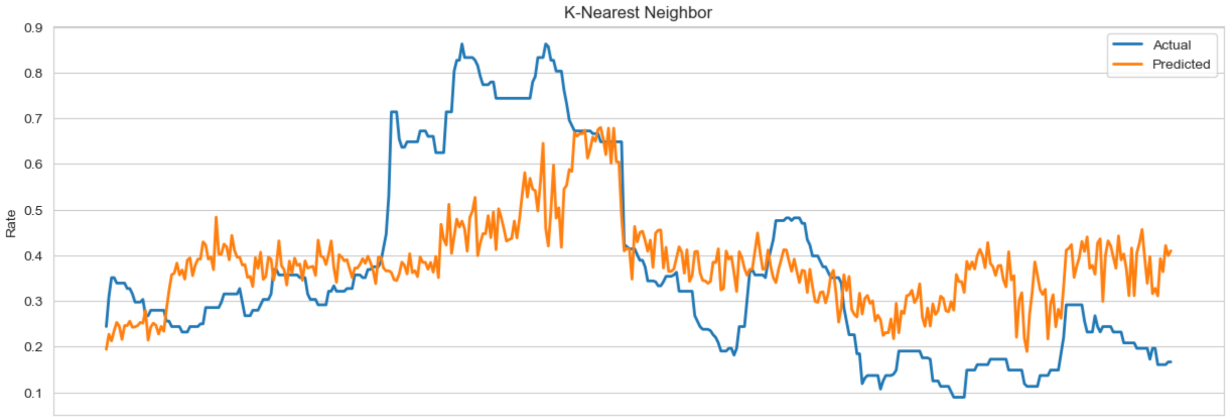


Figure 7.11: K-Nearest Neighbor

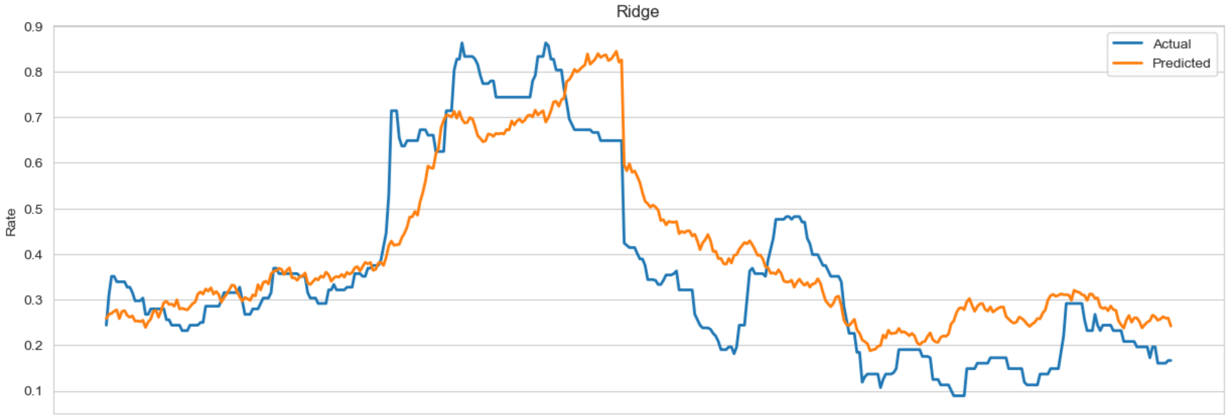


Figure 7.12: Ridge

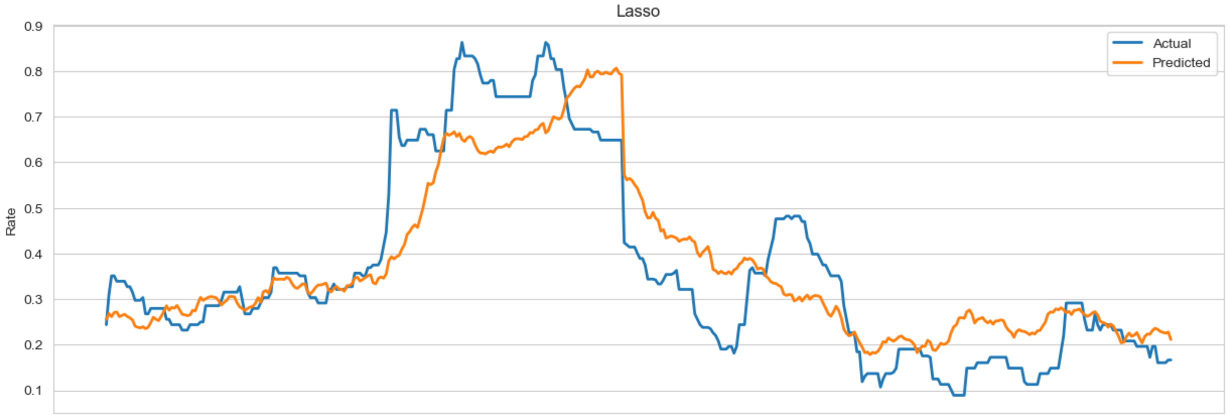


Figure 7.13: Lasso

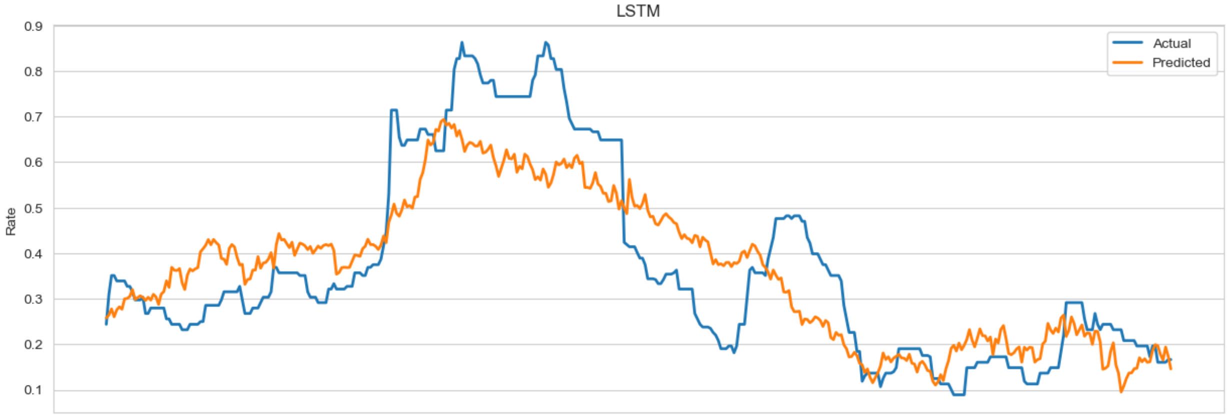


Figure 7.14: Long Short-Term Memory

References

AAmann, L., Andersson, J., & Eskeland, G. (2015). Missing in action? speed optimization and slow steaming in maritime shipping. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2577240

Adland, R., Haiying, J., & Strandenes, S. (2017). Are ais-based trade volume estimates reliable? the case of crude oil exports. *Maritime Policy & Management*, 1–9. https://doi.org/10.1080/03088839.2017.1309470

Alizadeh, A. H., & Talley, W. K. (2011). Vessel and voyage determinants of tanker freight rates and contract times. *Transport Policy*, *18*(5), 665–675.

Bakshi, G., Panayotov, G., & Skoulakis, G. (2010). The baltic dry index as a predictor of global stock returns, commodity returns, and global economic activity. *Commodity Returns, and Global Economic Activity (October 1, 2010).*

Beenstock, A., Michael & Vergottis. (1993). The interdependence between the dry cargo and tanker markets. *Logistics and Transportation Review, 29 (1), 3.*

Beenstock, M., & Vergottis, A. (1989). An econometric model of the world tanker market. *Journal of Transport Economics and Policy*, 263–280.

Clarksons Resarch | Shipping Intelligence Network. (2023).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *20*, 273–297. https://doi.org/https://doi.org/10.1007/BF00994018

Engelen, S., Meersman, H., & Voorde, E. V. D. (2006). Using system dynamics in maritime economics: An endogenous decision model for shipowners in the dry bulk sector. *Maritime Policy & Management*, *33*(2), 141–158.

Gao, S., & Lei, Y. (2017). A new approach for crude oil price prediction based on stream learning. *Geoscience Frontiers*, *8*(1), 183–187. https://doi.org/https://doi.org/10.1016/j.gsf.2016.08.002

Hampton, M. J. (1990). *Long and short shipping cycles: The rhythms and psychology of shipping markets*. Cambridge Academy of Transport.

Han, Q., Yan, B., Ning, G., & Yu, B. (2014). Forecasting dry bulk freight index with improved svm. *Mathematical Problems in Engineering*, *2014*, 1–12. https://doi.org/10.1155/2014/460684

Hawdon, D. (1978). Tanker freight rates in the short and long run. *Applied Economics*, *10*(3), 203–218.

IMO, 1. (1974). International convention for the safety of life at sea (solas), 1974.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kavussanos, M. G., & Visvikis, I. D. (2006). Shipping freight derivatives: A survey of recent evidence. *Maritime Policy & Management*, *33*(3), 233–255.

Kidwell, J. S., & Brown, L. H. (1982). Ridgde regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and Family, Vol. 44, No. 2 (May, 1982), pp. 287-299 (13 pages)*. https://doi.org/https://doi.org/10.2307/351539

King, A. P., & Eckersley, R. (2019). *Statistics for biomedical engineers and scientists: How to visualize and analyze data*. Academic Press.

Koekebakker, S., Adland, R., & Sødal, S. (2006). Are spot freight rates stationary? *Journal of Transport Economics and Policy (JTEP)*, *40*(3), 449–472.

Kulkarni, S., & Haidar, I. (2009). Forecasting model for crude oil price using artificial neural networks and commodity futures prices. *arXiv.org, Quantitative Finance Papers*, *2*.

Kumbure, M. M., & Luukka, P. (2022). A generalized fuzzy k-nearest neighbor regression model based on minkowski distance. *07*, 657–671. https://doi.org/https://doi.org/10.1007/s41066-021-00288-w

Li, J., & Parsons, M. G. (1997). Forecasting tanker freight rate using neural networks. *Maritime Policy & Management*, *24*(1), 9–30. https://doi.org/10.1080/03088839700000053

Lun, Y., Lai, K.-h., & Cheng, T. C. E. (2010). *International trade and shipping*. Springer London. https://doi.org/https://doi.org/10.1007/978-1-84882-997-8_1

Lyridis, D., Zacharioudakis, P., Mitrou, P., & Mylonas, A. (2004). Forecasting tanker market using artificial neural networks. *Maritime Economics & Logistics*, *6*, 93–108.

Næss, P. (2018, June). *Investigation of multivariate freight rate prediction using machine learning and ais data* [Doctoral dissertation].

Nato shipping centre [Accessed: 2023-12-11]. (2023).

Olsen, M., & da Fonseca, T. R. K. (2017, September). Investigating the predictive ability of ais-data: The case of arabian gulf tanker rates. https://openaccess.nhh.no/nhh-xmlui/handle/11250/2454692

Poulakidas, A., & Joutz, F. (2009). Exploring the link between oil prices and tanker rates. *Maritime Policy & Management*, *36*(3), 215–233.

Randers, J., & Göluke, U. (2007). Forecasting turning points in shipping freight rates: Lessons from 30 years of practical effort. *System Dynamics Review: The Journal of the System Dynamics Society*, *23*(2-3), 253–284.

Refinitiv Eikon. (2023). SSYTDS270MEGNBD | Simpson Spence Young Daily Dirty Middle East Gulf to Ningbo Zhoushan 270 World Scale.

Rigopoulos, G. (2022). Univariate Time Series Forecasting Using k-Nearest Neighbors Algorithm: A Case for GDP $_2$022. *International Journal of Scientific Research and Management (IJSRM)*, *10*(09), 3807–3815. https://doi.org/10.18535/ijsrm/v10i9.em01

Schneider, P., & Xhafa, F. (2022). *Anomaly detection and complex event processing over iot data streams: With application to ehealth and patient data monitoring*. Academic Press.

Skauen, A. N., Helleren, Ø., Olsen, O., & Olsen, R. B. (2013). Operator and user perspective of fractionated ais satellite systems. https://api.semanticscholar.org/CorpusID:127414004

Smestad, B. B. (2015, June). A study of satellite ais data and the global ship traffic through the singapore strait. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2350756

Stopford, M. (2009). *Maritime economics* (3rd). Routledge.

Strandenes, S. P. (1984). Price determination in the time charter and second hand markets. *Center for Applied Research, Norwegian School of Economics and Business Administration, Working Paper MU*, *6*, 15.

Suzuki, K. (2011, April). *Artificial neural networks - methodological advances and biomedical applications*.

Tham, E. (2008). Leading indicators for arabian gulf oil tanker rates. *OPEC energy review*, *32*(2), 139–149.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1. (1996), pp.267-288.* https://websites.umich.edu/~jizhu/jizhu/wuke/Tibs-JRSSB96.pdf

Vosburgh, K. (2017). The imo number explained.

Warren S. McCulloch, W. P. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of matematical biophysics*, *5*, 115–133.

Wijnolst, N., & Wergeland, T. (1996). *Shipping*. Delft University Press. https://books.google.no/books?id=TtFaAAAAYAAJ

Yilmaz, A., & Poli, R. (2022). Successfully and efficiently training deep multi-layer perceptrons with logistic activation function simply requires initializing the weights with an appropriate negative mean. *Neural Networks*, *153*, 87–103. https://doi.org/https://doi.org/10.1016/j.neunet.2022.05.030

Zannetos, Z. S. (1964). The theory of oil tankship rates: An economic analysis of tankship operations.