Norwegian School of Economics
Bergen, Autumn 2020

Transforming data into profit

Building a transformer neural network to predict Golden Ocean stock price based on Forward Freight Agreements

Lars Mølmann and Ulrik Aasen

Supervisor: Roar Os Ådland

MSc in Business Administration, Business Analytics and Finance

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

We would first and foremost like to thank our supervisor Roar Os Ådland for providing us with guidance during the entire process. From selection and forming of the thesis during the spring and summer of 2020 until the very end, Ådland has given us valuable input and guidance with the thesis. The availability, quick replies, and expertise from Ådland during the entire process has been critical for us to be able to deliver this thesis.

Abstract

The purpose of this study is to investigate the predictive relationship between Forward Freight Agreements (FFA) and the Golden Ocean Group stock price, using a Transformer Neural Network. Under the assumptions that the market for FFA-rates is efficient and an unbiased predictor of future spot rates, it should also provide reliable information about the future earnings of shipping companies. This relationship between the FFAs and the stock market can possibly be taken advantage of by applying the right trading models. This paper contributes to the literature by investigating the relationship between movements in the FFAs and the stock market, as well as the Efficient Market Hypothesis (EMH) through the notion that excess profit should not be possible to acquire in an efficient market. The results suggest that the transformer neural network has some predictive power on Golden Ocean Group stock price with the use of our selected FFA-rates and other non FFA-features. Our transformer model generated a profit of 58.44% from December 2019 till October 2020, and has an annualized Sharpe ratio of 1.12, thoroughly beating the benchmark models.

Contents

1 Introduction

The global shipping market has faced massive challenges in the aftermath of the financial crisis of 2008 and the oil price crisis in 2015. Freight rates have plunged, contributing to the demise of several substantial shipping companies. Large market fluctuations and pressing conditions are not new phenomena to most shipowners, who have become accustomed to weathering the storm. But after the storm comes the calm, and dry bulk shipping continues to be the most important form of seaborne trade in terms of volume of cargoes traded. By now, the characteristics of shipping markets have been thoroughly researched, and it has become evident that the different shipping markets inherit some characteristics which so far have been troubling to accurately account for through economic theory. Investors are not bound by the same obligations as shipowners and the pressing conditions affecting shipping companies can provide opportunities rather than strains.

In our thesis, we aim to display an ability to produce excess profit in the stock market by utilizing data on trading activity in the market for FFAs and the application of technical analysis and a machine learning model in the shipping securities market, in this case, Golden Ocean Group (GOGL). We will explore the compared accuracy of a Transformer Neural Network model, hereby mentioned as the transformer model, by using other models with diverse complexity as benchmarks. Our motivation surrounding the topic of our thesis is two-folded. Firstly, we believe there are underlying characteristics that tie together the movements of the FFA market and the market for shipping securities. An informational spillover will allow us to draw some knowledge of the future market direction of shipping stocks and prove there is excess profit to be made from this information. There is a large body of research arguing for the predictability of the spot rate, even though there are uncertainties regarding the efficiency of the spot market. Contrary to the spot rate, FFAs are tradable and should therefore not inherit the characteristics of the spot rate. Although, it is seemingly possible to effectively trade using information from the FFA market. Kavussanos, Visvikis, and Menachof (2004) investigate the unbiasedness hypothesis in the FFA market and find FFA prices to be unbiassed predictors of future spot freight rates. This notion provides a rationale where the FFAs act as a proxy for future earnings of shipping companies, because of close ties between the spot rate and earnings. Because the FFA market has to be efficient and an unbiased predictor to act as a valid proxy for the earnings of GOGL, we will provide a circumstantial discussion of the EMH concerning the degree of efficiency in the freight markets, and the expectation hypothesis because it provides understanding regarding the FFA market as a predictor for future spot rates. To draw a greater understanding of the unclear relationship between the FFA market and shipping stocks, we will investigate whether there is excess profit to be made from trading the GOGL stock based on information from the FFA market.

Secondly, the motivation to apply machine learning stems from a belief that the application of machine learning models in the financial markets has not been fully explored academically, although there are real-life examples of these models consistently outperforming the market. The implication of more widespread use is the high degree of advanced knowledge over multiple fields needed to construct and efficiently take advantage of such models. Exploiting techniques involving machine learning in the physical markets require advanced knowledge of mathematics, and coding, while also being able to comprehend the real-life applicability and the psychology of the market. Investors who have mastered these fields have been able to yield unmatched returns, e.g the Medallion fund of Renaissance Technologies (Dewey and Moallemi, 2019). Although the public has limited insight into the techniques used in the Medallion fund, by looking at data from 1988 to 2010 an exploitation of statistical arbitrage is suggested, using algorithms recognizing advanced patterns in the market. Although we by no means are of a belief that our model touches upon the intricacy of the models used by the Medallion fund and others, the pure existence of such anomalies sparked an interest into the reach of more traditional machine learning models. As Professor Bradford Cornell of UCLA firmly states in a paper reflecting on the performance of the Medallion fund:

"The performance of Renaissance Technologies' Medallion fund provides the ultimate counterexample to the hypothesis of market efficiency." (Cornell, 2019)

Inefficient markets constitute possibilities, possibilities which we intend to exploit. By applying a machine learning model and feeding it training data of a composed data set we hope the model will be able to take advantage of patterns unrecognizable to the human eye, maximizing risk-adjusted return by acting independently on buy and sell signals. Furthermore, the use of information from the market for FFAs is especially beneficial because we believe there exist underlying connections between this market and GOGL for us to take advantage of. The forward freight market is currently the only existing forward market in the shipping sector. Since the spot and FFA prices are inherently correlated by the designed mechanisms, their fluctuations are inevitably subject to the same comparative power of market demand and supply. We intend to evaluate and compare the profitability and risk-return performance of our transformer model against several benchmark models including a Moving Average, VAR, ARIMA, and Random Walk. In applying our constructed trading models using both machine learning and technical trading strategies we hope either will be successful in creating excess profit in a market which in recent years have left many investors in dismay due to high volatility and an unforgiving cyclicality.

The data used for the analysis covers the period 2012-2020. We used daily FFA data to explore any potential informational advantages that might be present in the market for FFAs. An FFA is a financial instrument used to hedge against fluctuations in the market for freight rates. In the past, the market was driven by mainly shipowners who used the FFA to secure their business against changes in freight, but in recent times more speculative investors have entered the market. Because an FFA is a derivative of an underlying asset it can be acquired as a purely speculative position as well as a hedge, which has the potential to largely increase the liquidity of the FFA-market (Alexandridis et al, 2017). Additionally, the increased variance in spot rates led to higher volatility as a result of allowing a larger number of investors to easily enter or exit the market, although this effect was mild (Pelagidis and Panagiotopoulos, 2019). Previously FFAs were traded strictly in the physical market between shipowners and charterers, but the market barriers and transaction costs were close to eradicated when the derivatives market became available.

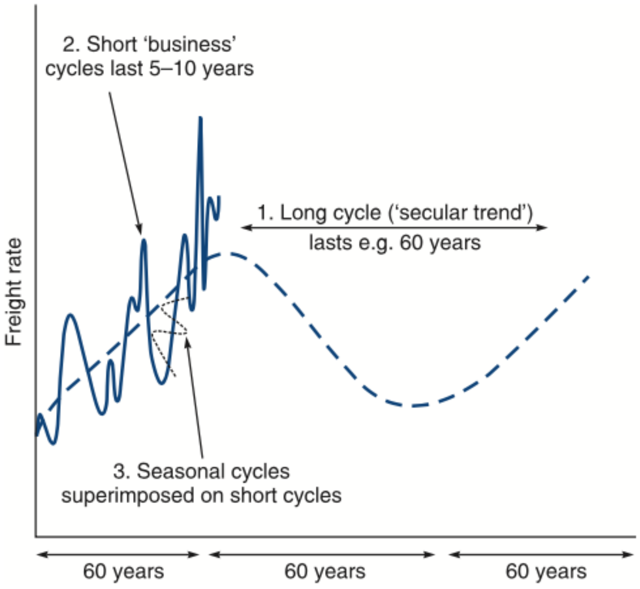
2. Short 'business' cycles last 5–10 years
1. Long cycle ('secular trend') lasts e.g. 60 years
Freight rate
3. Seasonal cycles superimposed on short cycles
60 years
60 years
60 years

Figure 1 - Seasonal, short and long cyclical components

The FFA market is constructed through the future levels of freight rates. Which would mean that FFA rates can be applied as an indicator for future spot freight levels and the market direction. Because of the characteristics of the FFA market, especially the ability to trade on forward freight derivatives, the characteristics inherited by the FFAs should not be consistent with those of the spot market. Although there is evidence of the reality being inconsistent with this theoretical rationale, where FFAs inherit some of the cyclicality and predictability of the spot market. Figure 1 illustrates the cyclicality of the freight rates market, the short-term cycles as well as the long-term and seasonal. There is research that suggests that FFAs generally are good indicators of future market direction (Kasimati and Veranos, 2017), which seemingly provides an advantage on decisions surrounding market entry and exit, beneficial for accumulating excess profits.

The findings in this paper provide clarity surrounding the effectiveness of trading in the stock market based on information from the FFA market, with our results substantiating the notion of the FFA market as a proxy for the future earnings of shipping companies. Additionally, it provides further proof of the effectiveness of the predictive power of the FFA market. Finally, evidence of producing excess profit will contribute to the existing literature on the EMH in the stock market.

The structure of this thesis is as follows: The next section discusses existing research and literature on several relevant economic factors of both the shipping market and our methodical approach. Due to the extensive body of literature on the economic theories which will be discussed, the different theoretical discussions will comprise a revision of the freight markets, both spot, and FFA, as well as the stock market, in that order. The section on literature is followed by a short overview of GOGL. Followed by an overview of feature extraction. Then we provide an elaboration on the methodology we used to conduct our research. Before finally providing and discussing the results of our analysis, including the limitations of our study and the conclusion we derive from our research.

2 Literature Review

The market efficiency is highly relevant because the FFA market must be efficient to be an unbiased predictor of future spot rates. (Kavussanos et al, 2004). Through the research of Fama (1970) and Samuelson (1965), three different categories of Efficient Markets were established (Jensen, 1978):

(1) The weak form of the Efficient Market Hypothesis, where asset prices contain only the information of past price history in the market as of that point in time.

(2) The semi-strong form of the Efficient Market Hypothesis, where asset prices contain all publicly available information at that point in time, although this form has been criticized for its undefined boundaries.

(3) The strong form of the Efficient Market Hypothesis, which states that asset prices contain all available information at that point in time.

An implication regarding the semi-strong form is to what extent information can be regarded as public, and the inability to firmly assert these boundaries may limit the validity of the research. The solidity of the EMH has been at the center of much research and is heavily disputed. There is a lack of convincing research to indicate a detachment from the EMH in the modern financial markets, considering the number of investors who successfully have accumulated excess profits continuously over longer periods.

"*Although there is evidence inconsistent with the Strong Form of the Efficient Market Hypothesis, if anything is surprising about it, it is the fact that such inconsistent evidence is so scarce." (Jensen, 1978)*

Research by Adland and Strandenes (2006) suggest the freight market is characterized by a semi-strong form, according to the literature by Fama (1970) and Samuelson (1965) on the three forms of market efficiency. The research continues by commenting on the terms of the EMH, and how some of these do not apply to the market for freight rates, especially the spot market, but the example is also relevant for the short-term forwards market. The example consists of how the spot price for freight rates would not be fully reactionary to the Organisation of Petroleum Exporting Countries, OPEC, signaling a reduction of output of oil in three months. The rationale here is that since the information will not have direct market implications in three months, the spot rate and prices of short term contracts, below 3 months, will not be dependent on these market conditions. The notion that the freight rate market is semi-strong should therefore not be correct since this publicly available information will not be priced into the market, but Adland and Strandenes (2006) state that these inconsistencies have to be overlooked because of the nature of the shipping market. The spot market can be predicted without the market being efficient, contrary to the FFA market. There is a lack of research into the efficiency of the FFA market, which Kavussanos et al, (2004) credits to the identification of risk-less arbitrage opportunities in the FFA market, which makes the research into the EMH very challenging.

If stocks always trade at fair value, trading strategies involving market timing and momentum trading should not outperform the market consistently. There has been a lot of research into the viability of the efficient market hypothesis in the shipping sector, mostly relating to vessel prices, with the results being two-sided. Although there is evidence that the EMH holds, it is generally for shorter periods and not across all vessel types. Research by Adland (2000) and Adland and Koekebakker (2004) argues that technical trading strategies should not produce excess profit using generic buy and hold strategies as a benchmark. Although Alizadeh and Nomikos (2007) argue that firstly, the results are dependent on the included variables and the set of rules used to construct the technical trading strategies, meaning the outcome may contain some form of bias. Secondly, not taking into account the underlying principles of economic theory, the results from this research are hard to redeem sufficient in being able to extrapolate information regarding the future behavior of market prices. The historical information contained is not satisfactory because of the common consensus that vessel prices follow random walk processes. Regarding vessel prices, although this market is not the main focus of our thesis, the prices of ships are closely linked to the intrinsic value of GOGL and its stock price, and are therefore highly relevant to the discussion of trading strategies. Because of the relationship between vessel prices and company value, our model may struggle to recognize changes in stock price caused by changes in vessel prices from scrapping or new shipbuilding. Due to the characteristics of the vessel market we might experience shifts in

GOGL`s stock price which seem inexplicably detached from the information integrated into our models.

Lo and McKinlay (1999) find that the existence of too many successive moves of stock prices in the same direction enable them to reject the hypothesis that these behave as true random walks. Additionally, they find that serial correlation in the short-term is unequal to zero, which provides evidence that there is some momentum effect in the short run. Lo, Mamaysky and Wang (2000) use sophisticated nonparametric statistical techniques to recognize patterns in stock price, arguing that signals used in technical analysis such as “double bottom” and “head and shoulders” may have some predictive power. Finally, economists such as Shiller (2000) study the field of behavioral finance and uncover short-term momentum to be consistent with psychological feedback mechanisms. As well as a tendency for market participants to underreact to new information, leading to a positive serial correlation following the reveal of impactful news, as the full effect of this news is not covered until later. An isolated review of these inefficiencies may not provide an economically reasonable point, and it could be argued that there is an important difference in statistical significance and economical significance. With the introduction of transaction cost and cost-of-carry, although the cost-of-carry is close to non-existent in the shipping- and stock market (Kavussanos and Visvikis, 2004), the inefficiencies that are uncovered in this research is not of a stature that would allow investors to design trading strategies to exploit them for excess profit.

An assessment that needs to be made is how the expectation hypothesis is used to characterize the freight markets. There are two different versions of this hypothesis, the expectation hypothesis, and the pure expectation hypothesis; the latter being a stronger proposition. The pure expectation hypothesis asserts that the forward rates exclusively represent the expected future rates (Nasdaq, 2020). Applying this to shipping would mean that forward freight rates represent the expected future spot freight rates. In theory, this would effectively equalize the expected earnings across all charter durations, creating zero risk premium. The regular expectation hypothesis is a slightly weaker proposition, only suggesting that the difference in expected earnings is constant, but not necessarily zero. An alternative hypothesis is that expectations of freight rates are adaptive or extrapolative, meaning they are backward- rather than forward-looking and depend on past values of these variables (Beenstock and Vergottis, 1989). Earlier research on the expectation hypothesis of the term structure recognized a relationship between short- and long-term contracts in the shipping market. Glen *et al* (1981) as well as Strandenes (1984) found evidence of an existing relationship but did not test for validity, meaning we cannot say for certain whether it is significant. According to Batchelor *et al* (2007), efficient non-storable commodity markets dominated by speculative investors are often characterized by two things; the forward prices are unbiased forecasts of future spot prices, and the price changes of fixed future dates are random, which is a reflection of the reaction to news. The absence of arbitrage between the FFA- and spot market means that if spot rates converge with the forward rate, it is because the forward rate embodies expectations of the future level of the spot rate (Batchelor *et al*, 2007).

Cullinane (1992) was one of the first papers researching the predictability of spot freight rates, using univariate ARIMA models to forecast the Baltic Freight Index, BFI, through data from the now outdated BIFFEX. The research concludes that the optimal forecasting horizon is very short-term which discourages the use of BFI forecasting to predict directional fluctuations in forward freight rates. Instead, Cullinane (1992) implies that his model could be used as a basis for investments in short-term BIFFEX contracts on short-term contracts in the spot market, seeing as these BIFFEX contracts provided insight into the future movements of the spot contracts. Kavussanos and Nomikos (2001) examined the causal relationship between the BFI and BIFFEX by trying to forecast the performance of the BIFFEX using daily data from the BFI. By applying a test for causality as well as using a generalized impulse response analysis, which describes the reaction of a variable to a standard deviation shock to the residual of another variable, they found that future prices are more responsive to market changes than the spot rates. Consequently, this research has led to a belief that the derivatives market contains useful information in successfully uncovering a price discovery function. Meaning the FFAs should contain information that could be used to trade in the spot market, however the inability to trade spot prices directly forces us to evaluate the effectiveness of the price discovery in the market for shipping stocks. Additionally, Kasimati and Veranos (2017) also find that FFAs generally are good indicators of future market direction but fail to accurately predict the turning points of market cycles.

There has been much research devoted to the lead-lag relationship of the future and spot market. The lead-lag effect in the shipping sector displays how quickly markets adapt to new information, seeing as both the spot and forward markets are influenced by similar factors. This rationale would imply that in a perfectly efficient market, both the spot and forward market would react simultaneously to events and shocks. Should one of these markets react before the other, due to lower transaction costs or other market microstructure effects, we could also experience spill-over effects in addition to lead-lag effects. Kavussanos and Visvikis (2004) study the lead-lag relationship in returns and volatilities between the FFA- and spot market, arguing that the FFA market contributes to discovery of new information in the spot market, as well as uncovering a bi-directional lead-lag relationship between FFAs and spot prices. Research provided by Zhang, Zeng, and Zhao (2014) used a hybrid forecasting method to assess a lead-lag relationship between freight rates in the spot and forward market. The empirical results from the study indicated an existing cointegration between spot and FFA rates, which means FFA rates are helpful in forecasting spot freight rates. Although we do not deal with forecasting the spot or FFA market, the notion that the FFA market provides valuable information in predicting future market behavior is fundamentally positive for our research. The lead-lag effect between the FFA and spot market would indicate that we could buy or sell the GOGL stock based on information from FFAs and capture the lagging effect of the spot market which would be closely linked to the behavior of the GOGL stock.

Investors in shipping securities have always had to deal with high volatility and complex surroundings. The exploitation of the cyclicality of the market is a factor that is often coherent with success but figuring out where the tops and bottoms of the cycle are something which few have consistently succeeded with. Quantitative market timing strategies have traditionally been tested on liquid commodities and financial futures, often with mixed results concerning their performance. Michail and Melas (2019) use co-integration techniques to establish a relationship between returns on shipping stock and the Baltic tanker Index, to test a Moving Average trading strategy. Their results indicate that the simple MA strategy outperforms a standard buy-and-hold by a substantial margin.

3 Golden Ocean Group Limited

The fleet size and structure of GOGL are depicted in figure 2 together with a diagram of main commodity exposure. (Golden Ocean, 2020). The fleet composition is of interest because it helps provide an understanding of the risk related to the company. As Adland, Ameln, and Børnes (2020) argue, asset diversification is the primary tool to hedge investments and control for balance sheet fluctuations due to the volatility of ship prices. GOGL has diversified its fleet by focusing on Capesize and Panamax, although this strategy will only be effective if the prices of these ship categories are not positively correlated.

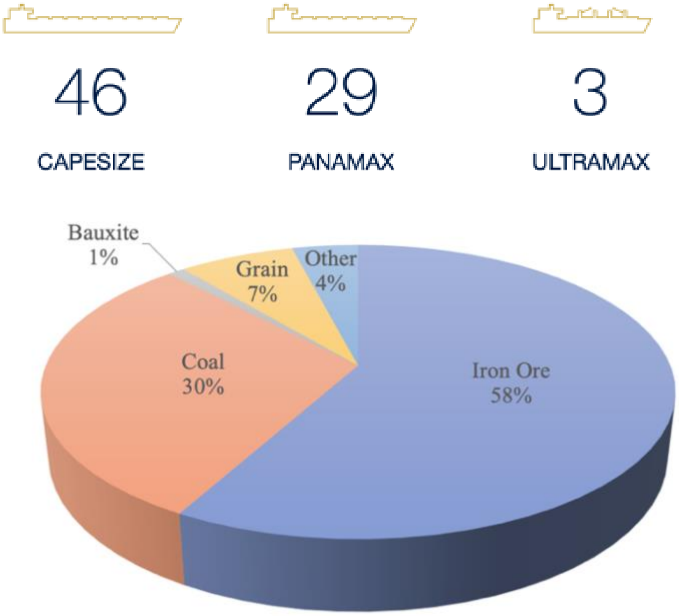

Figure 2- Structure of GOGL fleet and main commodities

Observing the fleet structure of GOGL we notice that they are mainly focused on the larger vessel sizes, Capesize being the second largest vessel size in deadweight tonnage and Panamax being a medium-sized vessel. Kavussanos (1996) argues that smaller vessels have a more flexible trading pattern and are less volatile, meaning these vessels are less exposed to ship price fluctuations than larger vessel sizes. Alizadeh and Nomikos (2009) investigate the correlation of vessel returns between Capesize, Panamax, and Handysize. The highest correlation was revealed to be between Capesize and Panamax, with a correlation coefficient of 0.510. While the lowest correlation was found between Capesize and Handysize, 0.370, which makes sense given these are the furthest from each other in deadweight tonnage. Given these positively correlated vessel price relationships, the balance sheet will be exposed to fluctuations in ship prices. The issue of balancing risk may lead to participation in the financial markets through charter coverage in the physical and financial markets. This means the fluctuations in FFA rates could have a direct impact on the intrinsic value of GOGL.

Looking further into the composition of the fleet we can derive some of the focal points of GOGL`s business operations. Ultramax vessels, which GOGL only has three of, are mainly engaged in the transportation of grain commodities from North and South America, in addition to Australia. They mainly deliver to Europe and Asia but are also involved in the transport of more minor commodities such as bauxite, aluminum, and sugar around the world. These commodities do not account for much of the volume in the dry-bulk industry and the Ultramax vessels account for a small percentage of the GOGL fleet. In the final model, none of the indices linked to the most common commodities contributed significantly to our model, seeing as price changes would most likely have an insignificant effect on the GOGL stock.

4 Feature Extraction

To improve the accuracy of our model we included several features that would increase the predictability of the movements of the GOGL stock. First, we must address the most relevant FFA contracts which were included in our model. The 4TC_C contract, which is an average of the most important capsize routes, had the highest relevance. Unsurprisingly, the contracts for Capesize vessels proved more effective than the contracts for Panamax, 4TC_P, seeing GOGL's fleet consists of more Capesize vessels than Panamax. We have also included the Baltic dry index (BDRY) because it is a good representation of the levels of the dry bulk shipping industry. We have touched upon the freight rates' close relation to the macroeconomic situation and seeing as the level of the S&P 500 is a good reflection of the situation of the Western economy we have included it in our model. As Klovland (2002) states in his research, there is a close timing relationship between the upper turning points of business cycles, commodity prices, and freight rates. Although this research was conducted on a period predating the second world war the basic principles of the shipping industry have not changed much, hence we include all relevant major size commodities such as SGX Iron Ore 62% Futures, Generic 1st XW Futures (Coal), and Generic 1st No. 2 Wheat Future. Tsioumas, Vangelis, and Papadimitriou (2016) investigate the lead-lag relationship between the most major commodities and the relevant vessel sizes, with their results implying a bidirectional relationship in the cases of iron ore and coal, while the wheat price only had a unidirectional relationship with the Baltic Panamax index. To further implement the balance between markets that affect the shipping economy we included the USD/Yuan relationship. Finally, crude oil prices are also included through the brent futures because it is an important cost factor for shipping companies. We will further elaborate on which features were included in the final model in the Feature Selection paragraph 5.3.

5 Methodology

5.1 Data preparation

The transition from raw data files to clean, interpretable data applicable for use in a machine learning model is extensive. The process consists of understanding the data we extract, evaluating the relevant parts, transforming the data from raw data into clean data frames, checking for stationarity, normalizing it – and finally splitting the data into training, validation, and test data. Most of these processes can be statistically and computationally calculated, but the combinations of them are found through trial and error.

The first part of the process is finding out which parts of the data we have gathered that are relevant, both regarding features to use in our model, but also in a time aspect. On one side, it is desirable to have enough data so our model is trained on at least an entire shipping cycle of approximately seven years, but on the other side, will more recent data observations perhaps be more relevant than observations from 20 years ago. A general rule of thumb is to keep most of the data, as the transformer model we are building can assign more relevance to newer data points than older ones, but we will through trial and error investigate how performance shifts with different collections of data. As for the features, we will perform several statistical tests to find out which FFA contract types and other non-FFA-features are most influential regarding the stock price of GOGL.

After filtering out the most relevant data, we must format the data in a stationary form. The transformer model assumes and requires that the time series is stationary. In a stationary time series, statistical measurements like mean and variance are constant over time, meaning that trends and seasonality are absent. To check if our data set fulfills the requirement of stationarity, we perform an Augmented Dickey-Fuller test. This is a statistical test called a unit root test. It uses an autoregressive model and optimizes a criterion across several lag values. When performing the ADF-test, we found out that only 2 of the 50 features were stationary and the other 48 were therefore not optimal for use in the transformer model. To cope with this problem, we performed a differencing to the closing price of GOGL and the features to produce a stationary data set. This differencing increases the stationarity of our dataset, contributing to information extracted from the dataset by our transformer model having a higher validity for future predictions. See appendix

A.2 for a thorough review of the ADF-test and the differencing method, together with the results of the tests.

To further prepare the data set for the transformer model, we had to normalize it through a min/max normalization. This is a widely used data preparation technique that assigns a 0 to the minimum value and a 1 to the highest value. Every other value then gets transformed to a decimal between 0-1. In a data set like ours where the scales of our features differ by tens of thousands, and the range of values within each feature is large, features with larger scales will have a greater impact on the output predicted by our model. (Angelov and Gu, 2019). Normalization of data also has the benefit of increasing the computational speed when training the model, which is of great significance without access to external servers and several million data points to create a neural network from. A drawback of using the min/max method occurs if we have outliers present in the dataset. When treating outliers, the model will weight these heavily, which could be damaging for the prediction. However, these outliers can symbolize dramatic events that can have a great effect on the stock price. Removing these could harm the model's ability to react quickly to news. We performed a z-test to detect any outliers which we define by using the number of standard deviations each value is away from the mean value and put the threshold at 3 (Sharma, 2018). No outliers were detected, and we continue by computing the normalization using equation 1.1.

$$\frac{Value-MIN}{MAX-MIN} \qquad (1.1)$$

5.2 Splitting the data set

The final task to make the data ready for use in our model was splitting it into a training-, validation- and test data. Our goal when splitting the data set is to get the highest possible test accuracy. This is obtained when we have just enough data in our training data set, without overfitting the model. This is known as the bias-variance trade-off and is a decision surrounding flexibility in the model. The variance is measured by how much the predicted stock price changes when predicted on another data set. Bias is measured by the mean squared error. In general, the more training data we have, the more variance the prediction gets. This leads to overfitting of the model because it creates a model that fits perfectly to the training data, but cannot adapt to other scenarios, hence making it underperform on test data. Models with insufficient training data get oversimplified and lead to a biased model which results in a higher error on test data. This leads to underfitting the model, making it unable to see patterns between variables.

We seek a balanced split of data that leads to a low bias and variance. To achieve this, we start by making our training set 80% of the data. The validation set is 10% of the data, which we use to provide an unbiased evaluation of the models fit on the training data set. We use this data to adjust the hyperparameters, so we get the best possible fit for our test data set (Shah, 2017). The test data is the last 10% and is the data we use to test the prediction accuracy of our transformer model up against the other prediction models and methods. The size of the different data sets depends largely on the number of hyperparameters. As the validation set is used to adjust these, the more hyperparameters we have, the larger validation set we need (James et al, 2017). Because the model does not have many hyperparameters we decided to start with this 80/10/10 split of the data, and after some trial and error with lower percentage training data, the original 80/10/10 split proved most effective.

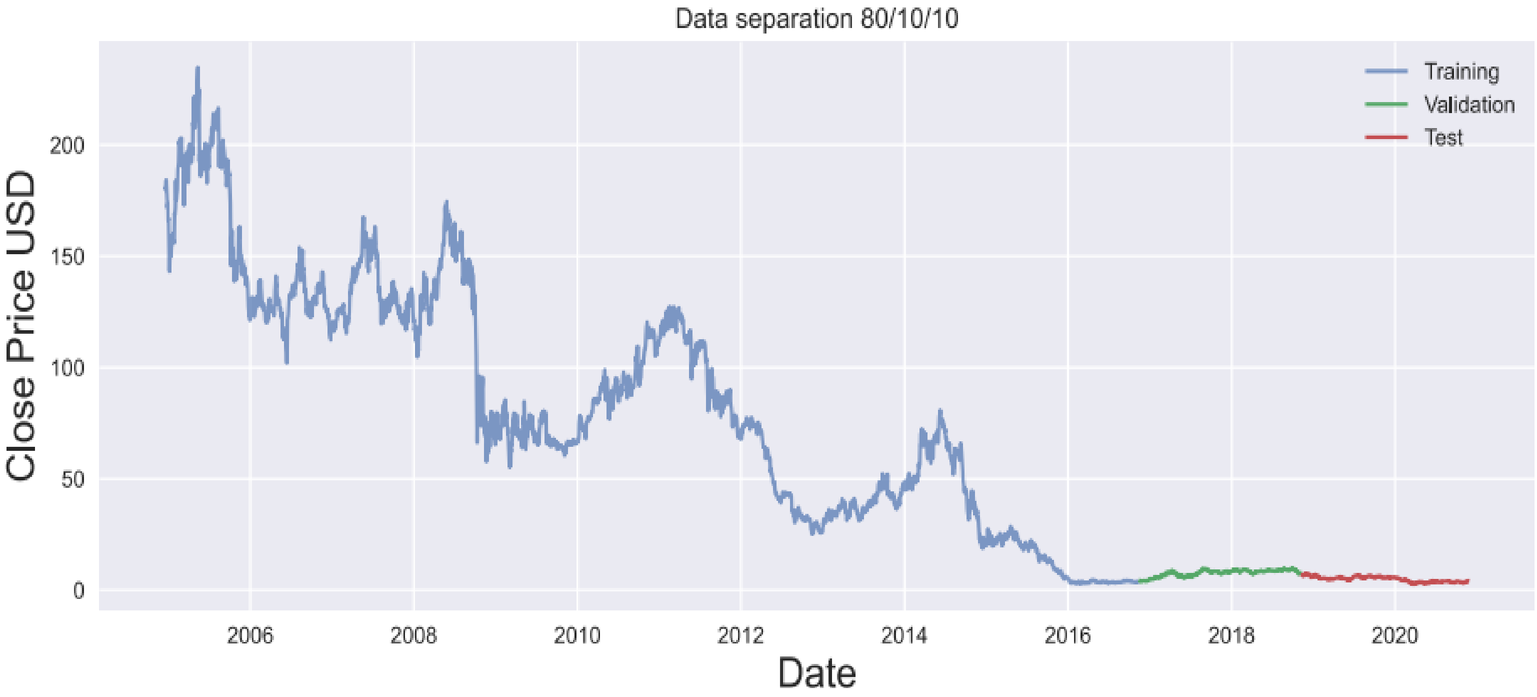

Figure 3 - Splitting of data-set

5.3 Feature selection

As mentioned earlier, we extracted several different features to be used in the transformer model. As we commenced with the process of constructing an optimal data set, we adjusted our feature selection to optimize the prediction performance of our model. This process involved an exclusion of the features which were deemed irrelevant through testing, and in the end only including the most relevant features for our model.

When choosing features it is important to remove redundant data for two reasons. One, it reduces overfitting and two it makes for less noise in the model. This will lead to higher accuracy and a reduction in training time. To select our features, we performed a chi-squared test, a ridge regression, and a recursive feature elimination test. We normalized these test scores and made a combined average score to see which features had the greatest impact on the GOGL stock price. We then selected the top 30 performing features and used them in the model. From the feature selection, it shows that the 4TC_C FFA-rates are what affects the GOGL stock price the most, hence being the most important features in our model. Furthermore, we see that some of the non-FFA features performed well enough to be considered an influential feature, but none of them were high on the ranking. S&P500, Brent futures, and the Baltic dry index are the top non-FFA features, and the only three with high enough statistical influence on the GOGL stock price to be included in our model. The complete list and results from the feature selection are explained in appendix A.3.

Chi-squared test

The chi-square statistic is used for testing relationships between two variables. With the formula:

$$x_c^2 = \sum \frac{(o_i - e_i)}{e_i} \qquad (1.2)$$

It tests how the expected count e deviates from the observed count o. This helps us identify which features are influential on the response, and thereby important for the prediction model (Gajawada, 2019).

Ridge regression

The ridge regression is quite similar to the least-squares method. The coefficients are estimated by minimizing a different quantity. The ridge regression coefficient estimates are the values that minimize the loss function where λ ≥ 0 is a tuning parameter. The second term of the formula is a shrinkage penalty, with the effect of shrinking the estimates of beta towards zero. This penalty is equivalent to the square of the magnitude of the coefficient. (James et al. 2017).

$$\sum_{t=1}^{n}(y_i - B_0 - \sum_j B_j x_{t,j})^2 + \delta \sum_{j=1}^{p} B_j^2 \qquad (1.3)$$

Recursive feature elimination (RFE)

RFE is a wrapper-type feature selection algorithm. This algorithm is used in the core of the model to rank features by importance and re-fits the model. It is a backward selection of the predictors. It begins by building the model on the entire list of features that we have available. Then it computes the importance of each feature and removes the least important feature. This process continues until we have reached our desired number of features. (Johnson, 2019).

5.4 Machine learning methodology

Machine learning is the use of statistical algorithms to find patterns in large amounts of data, without being directly programmed. Neural networks go under the category of deep learning, which is a subgroup of machine learning that uses techniques that gives machines an enhanced ability to figure out even the smallest patterns. Deep learning has many layers of computational nodes working together to find these patterns and create predictions. Here is a severely simplified illustrated version of a neural network with three input layers, one output layer, and four hidden layers.

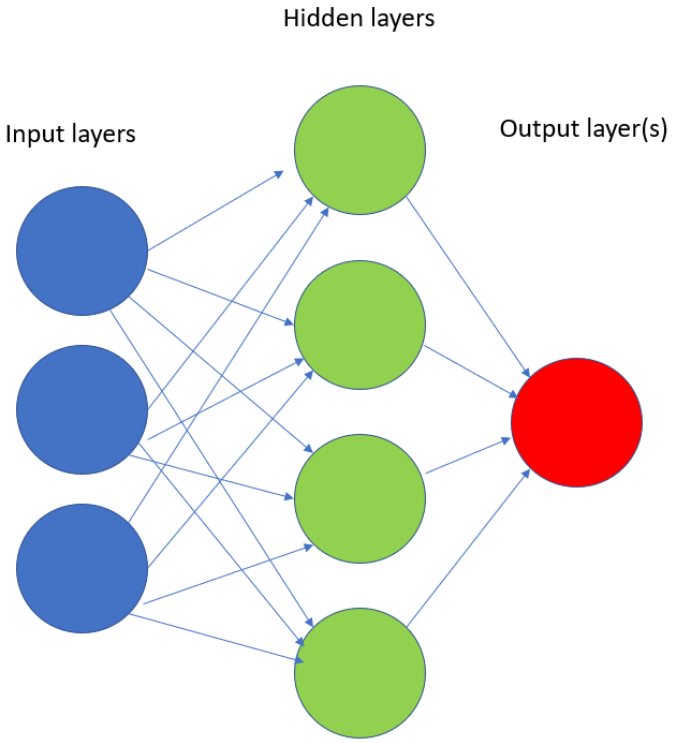

Figure 4 - Basic neural network

These neural networks were inspired by the human brain, where the nodes in the model represent the neurons in our brain, signaling between them to recognize patterns. Information in neural networks goes from the input layers, through several hidden layers which determines the weighting of the input. When the output is calculated, the model compares the estimation with the actual observed value and gets feedback from a loss function, for example, mean squared error. The model`s learning process is based on minimizing this loss function by adjusting weights and parameters, while also adjusting for bias (Haykin, 2009). To exemplify for our case, the input layers are the different features we use. These get sent into the transformer model where the hidden layers calculate the weights for each feature and then predicts the GOGL stock price which is the output layer. Equation 1.4 shows how a basic neural network calculates the value of an output node. We got the neuron value x, weights w, and bias b.

$$x_{l,f} = \sum_{f=1}\left(w_{l-1,f} * x_{l-1,f}\right) + b_{l,f} \qquad (1.4)$$

The parameters, which determine the weighting of the input features and biases, are decided in the model. The hyperparameters are decided by the developer and are individual for different models and use cases. A way to optimize the hyperparameters, often used in machine learning, is a randomized grid search cross-validation (Benner, 2020). This process consists of us creating a list of hyperparameter values we want to explore and the cross-validation choosing random hyperparameters from our list and evaluating the model`s performance with different combinations.

The latest addition to deep learning is reinforced learning. A reinforcement algorithm learns from its previous errors to achieve better predictions. The method we use, a transformer model, is in the category of recurrent neural networks. This is a feed-forward neural network rolled out over time, making it able to deal with sequenced data (Hao, 2020). Recurrent networks are looped, enabling them to store information from previous calculations. It can be thought of as multiple copies of the same network, delivering information to the next layer (Olah, 2015). The transformer model was first introduced in 2017 and is a new method of computing the hidden layers in the model, with a specific focus on attention layers to determine which features are the most important. One of the main challenges with neural networks is the computational power needed to perform the complex training process. An advantage with the transformer model is that it allows more parallelization than regular recurrent neural networks, thereby reducing training time. (Vaswani et.al, 2017).

Our transformer model and hyperparameter tuning are thoroughly explained in appendix A.5.

5.5 Benchmark models

To better assess the performance of our model, we have benchmarked it against several different trading strategies and statistical models to see if it is the best performing model for our case. We have included an autoregressive integrated moving average (ARIMA) model, a vector autoregressive (VAR) model, and a random walk model for our trading models. In addition to this, we have also implemented two different trading strategies, a long/short strategy with a moving average and a simple buy and hold. First, we compared our model to a 50 and 200-day moving average, buying when the 50-day average crosses the 200 from below and selling when it crosses from above. To illustrate the buy and sell signals, we plotted it below, with green arrows ∧ marking a buy signal and red arrows down ∨ signaling a short, as illustrated in figure 5.

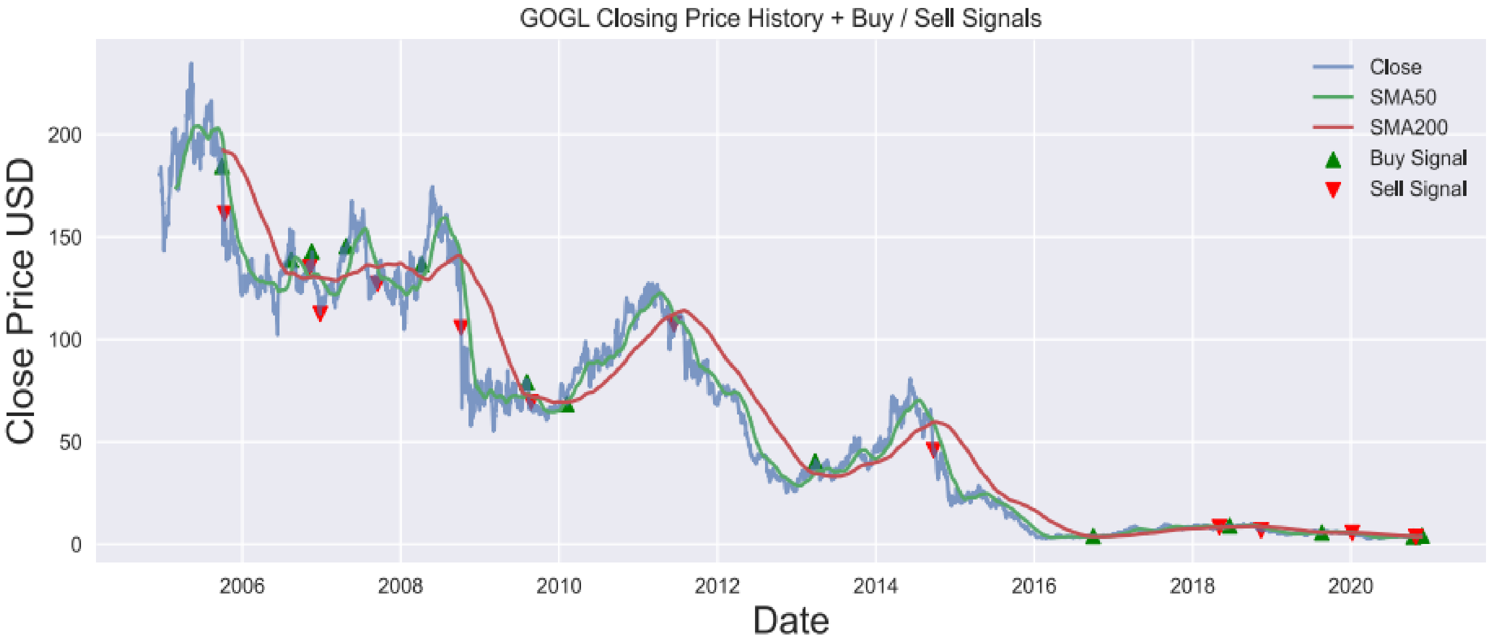

Figure 5 - Moving average strategy trading signals

ARIMA

The autoregressive integrated moving average (ARIMA) model is a class of models that explains a time series based on its own earlier values. It is a linear regression model that estimates the lags of the time series and uses these as predictors. The model is composed of the autoregressive part (**AR**IMA) where Y is a function of its lags. Where Y(t-1) is the lag 1 of the series, beta is the coefficient of lag 1 and alpha is the intercept term, which is also estimated by the model. The number of lags included is determined by the parameter p and the integrated term is defined by a parameter d, which determines the order of differencing.

The next part of the model is the moving average part (ARI**MA),** where Y depends only on the lagged forecast errors**.** The error terms are the errors of the autoregressive models of the respective lags. The parameter q defines the number of error terms to include in the equation.

$$Y_t = a + B_1Y_{t-1} + B_2Y_{t-2} + \cdots + B_pY_{t-p}e_t + \emptyset_1 e_{t-1} + \emptyset_2 e_{t-2} + \cdots + \emptyset_q e_{t-q} \qquad (1.5)$$

The complete ARIMA model is simply the autoregressive and moving average equations put together, where we have the predicted Yt which is a product of a constant + the linear combination lags of Y(up to p lags) + the linear combination of lagged forecast errors (up to q lags). In our model, we have used 2-day lag as we can see from figure 6, GOGL is autocorrelated with 2 days lag and does not need any higher order of differencing (Hyndman and Athanasopoulos, 2013). This can also be seen by performing an AIC-test. In the process of both ARIMA and VAR modeling, the AIC test is a criterion selection to select the number of lags (p) in the models. The AIC penalizes models with too high complexity, even though more complex models may perform slightly better on other criteria. We expect to see an inflection point, meaning that the AIC score should get lower, before turning after a certain point. We perform a grid search to find that the optimal number of lags corresponds with what we saw when plotting autocorrelation.

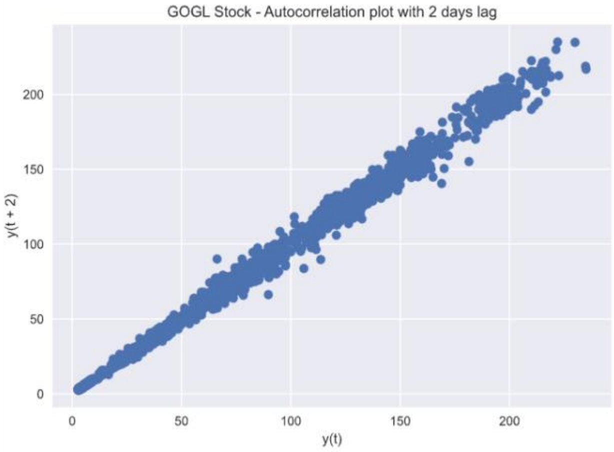

Figure 6 - Autocorrelation of GOGL vs features

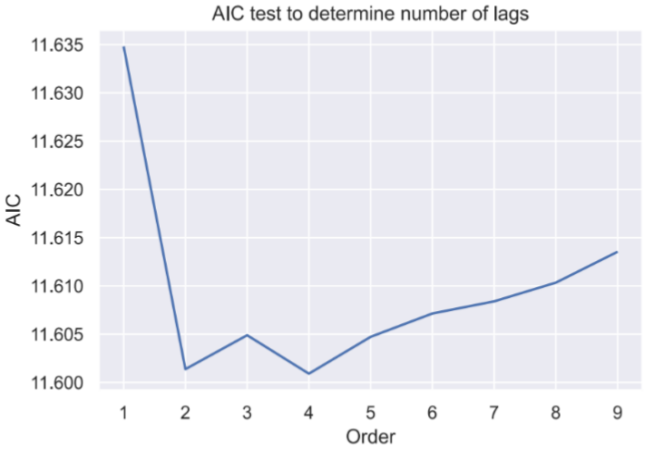

Figure 7- AIC test to determine lag

VAR

The vector autoregression (VAR) model is a multivariate prediction model that facilitates the inclusion of previous values of features and has proven to be a powerful forecasting tool for financial time series (Zivot and Wang, 2006). The difference between linear regression and vector autoregression is that in the VAR model, the variables influence each other. A VAR model is a generalization of the univariate autoregressive model for forecasting a vector of time series. It is a system of equations where every variable is calculated as linear combinations of past values. It compromises one equation per variable in the system. The right-hand side of each equation includes a constant and lags all of the variables in the system. (Hyndman and Athanasopoulos, 2018). The lags are set to 2 days, as determined by the tests explained above for the ARIMA model.

$$y_{1,t} = c_1 + \emptyset_{11,1} y_{1,t-1} + \emptyset_{12,1} y_{2,t-1} + e_{1,t} \qquad (1.6)$$

$$y_{2,t} = c_2 + \emptyset_{21,1} y_{1,t-1} + \emptyset_{22,1} y_{2,t-1} + e_{2,t} \qquad (1.7)$$

Where $e_{1,t}$ and $e_{2,t}$ are white noise processes that may be contemporaneously correlated. The coefficients $\emptyset_{ii,l}$ captures the influence of the lth lag of variable y on itself, while the coefficient $\emptyset_{ij,l}$ captures the influence of the lth lag of variable y_j on y_i.

Random walk

The random walk is a much-used benchmark for prediction models. It simply uses the last actual value and uses it to forecast. The model is denoted with formula 1.8, where H is the forecasting horizon. This model did generate better results than the buy and hold, but it performed a lot worse than any of the other models and strategies. (Hyndman and Athanasopoulos, 2013).

$$\hat{y}_t + H = y_t \qquad (1.8)$$

5.6 Measuring the quality of fit

One of the key aims of prediction modeling is to measure the quality of fit in our model. To evaluate the performance of the model on a data set, we need measures for how well its predictions match the observed data, and we have selected these four metrics to determine predictive accuracy.

MSE

The most used method is the mean squared error (MSE) given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad (1.9)$$

Where $\hat{f}(x_i)$ is $\hat{f}$`s prediction for the *i*th observation. The MSE in the equation is computed using the training data used in our model (FFA-rates + other features) and is referred to as the training MSE. However, the most important aspect is not how well the model works on training data, but how well it works on test data. When predicting stock prices, the ability to predict stock prices from last week is not relevant, but rather to forecast tomorrow`s development. The trade-off between minimizing the test MSE and training the model sufficiently is a fundamental part of statistical learning and is the same bias-variance trade-off we have when deciding on the size of the data set. As the flexibility of the model increases, the training MSE will decrease, but the test MSE might not. When we obtain a small training MSE but a large test MSE we are overfitting the data, which was exactly what happened when we first ran our model. After finetuning the hyperparameters and reducing the flexibility of our model, we ended up with a better fitted model with a lower MSE on the validation and test set.

MAE

The second measure of model fitting are mean absolute error and mean average percentage error. MAE is a measure of errors between the estimated and the observed value(Willmott and Matsuura, 2005).

$$MAE = \frac{\sum_{i=1}^{n}|e_i|}{n} \qquad (1.10)$$

MAPE

MAPE is a statistical measure of the accuracy of a prediction. It measures this accuracy by calculating the average percentage error for each time period, in our case each day, minus actual observed values and dividing this by the actual values. Here, A is the actual value and F is the prediction value. (Myttenaere et.al 2016)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \qquad (1.11)$$

Diebold-Mariano test

The Diebold-Mariano (DM) is a test used to measure the directional accuracy of predictions. The test compares the different models against a random walk, which indicates if the model is better performing than a random walk with 50% theoretical directional accuracy. Because our trading strategy only takes the direction of predictions, and not how much the price changes into account, it is a valuable measure to evaluate performance. The DM test with results is explained in appendix A.4

6 Results

Table 1 and 2 show the overview of the performance and trading results of the transformer model together with the different benchmarks. When talking about performance, we refer to the different evaluation metrics of MSE, MAPE and MAE. The trading results are the profit gained on the GOGL stock based on trading signals from the different models. When testing our trading results, we use the predictions from December 2019 until late October 2020. We use a long/short strategy for the VAR-, ARIMA-, and transformer model, trading on signals from the models on a day-to-day basis. When a model predicts a downfall in stock price by more than 1%, it short-sells, and when it predicts a rise of more than 1%, it buys. Meaning if our model predicts a price change in the interval [-1%, 1%], it holds the position as it is. With the moving average, it buys and short sells when the 50- and 200-day averages cross. When the period is done, in October 2020, we liquidate the positions. We assume a standard brokers fee of 1% and calculate the Sharpe ratio using the risk-free rate of a high-interest savings account in DNB called “Superspar” with an interest of .35%.

The random walk and moving average strategies were thought of as the lower tier benchmarks for performance, being relatively simple in their ways. Regarding the VAR and ARIMA models there was evidence from another master thesis (Farbrot and Kalvik, 2019), that especially the VAR-model could generate a decent profit from FFA-rates. We would consider our transformer model to be successful if it could outperform both the VAR- and ARIMA models.

	Mean average percentage error	Mean squared error	Mean absolute error
Transformer model	50.95%	**0.22**	**0.051**
Random walk	96.08%	29.32	0.74
VAR	57.86%	1.87	0.16
ARIMA	**49.65%**	0.39	0.12

Table 1 - Evaluation metrics trading models

Overall, the models are not performing particularly well regarding the magnitude of price shifts, with high values throughout the three evaluation metrics in table 1. However, the most important concept in this trading strategy is not the prediction of price shift magnitudes, but the directional accuracy of our predictions. Meaning, if the GOGL stock price increases by 10% and the model indicated a 2% price increase the day prior, we are still buying and making a profit. On the DM evaluation metric, the transformer model displays better directional accuracy with a 99% confidence interval than the random walk, the VAR with 95%, and ARIMA only with a 90% confidence. This difference seems to be mirrored in the trading results.

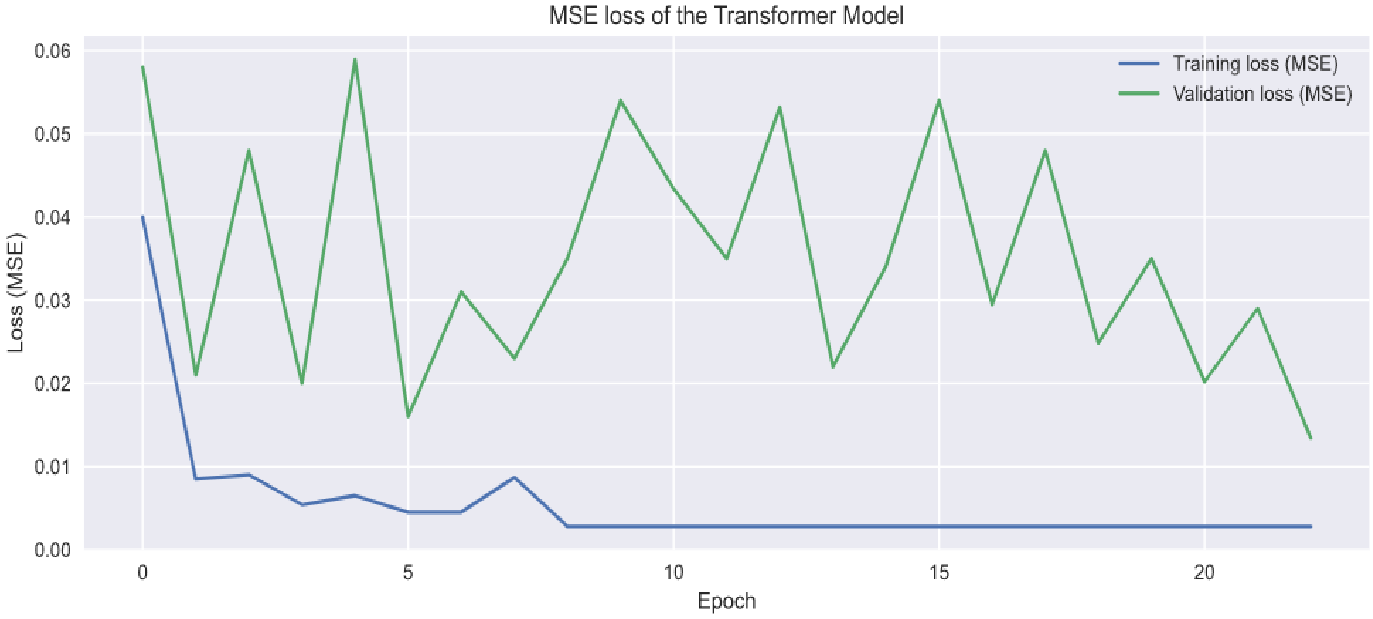

Figure 8 - Loss function for transformer model

Above is the evolution of the MSE in the validation set. This is as mentioned used to give an evaluation of how well the training data fits while tuning our hyperparameters. We observe that the validation set achieves a low MSE which indicates that the model is unbiased, the hyperparameters are tuned, and the model can be implemented on test data. The transformer model has the best score on mean squared error and mean absolute error, even though these metrics rise slightly from validation to test data. On the mean average percentage error, it is marginally beaten by the ARIMA model. This is evidence that the transformer model outperforms the benchmarks on most metrics. However, seeing as the directional accuracy outweigh MSE, MAE, and MAPE as the most important evaluation metric, it is more interesting to see how the different models perform regarding trading results in our test period.

	Profit	Annualized Sharpe ratio
Buy and hold	-18.40%	0
Moving average (50 vs.200)	27%	0.51
Transformer model	**58.44%**	**1.12**
Random walk	-9.47%	0
VAR	23.67%	0.45
ARIMA	16.89%	0.32

Table 2 – trading results

The moving average strategy gained quite a large profit over the entire period, especially in the period after the financial crises, and is here plotted against a buy and hold strategy. In the test period, it gained a 27% profit. This is second best of all our models and supports Micheal and Melas (2019) research on the performance of the moving average strategy on shipping securities.

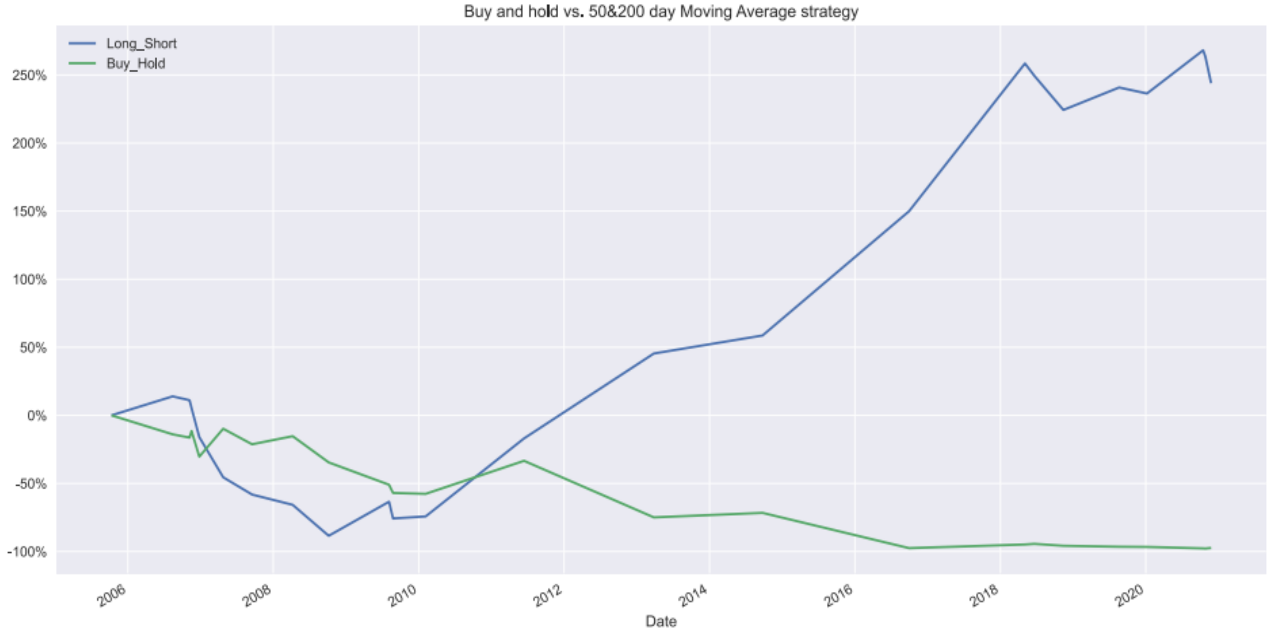

Figure 9 - Buy and hold vs. moving average strategy

It must be said that the buy and hold is a questionable comparison regarding performance, due to the long negative trend in GOGL's stock price. The moving average does however provide a more valid benchmark for the transformer model, displaying how effective a simple trading strategy can be.

The ARIMA model generated a 16.89% profit over the testing period, which is in the lower range of the models, and the lowest of the three algorithmic ones. The evaluation metrics are also sub-par, especially the directional accuracy swerving around the 50% mark, similar in theoretical accuracy to a coin toss. Tweaking of lags and features does make the ARIMA somewhat better, but it is overall not a satisfactory prediction model in this case.

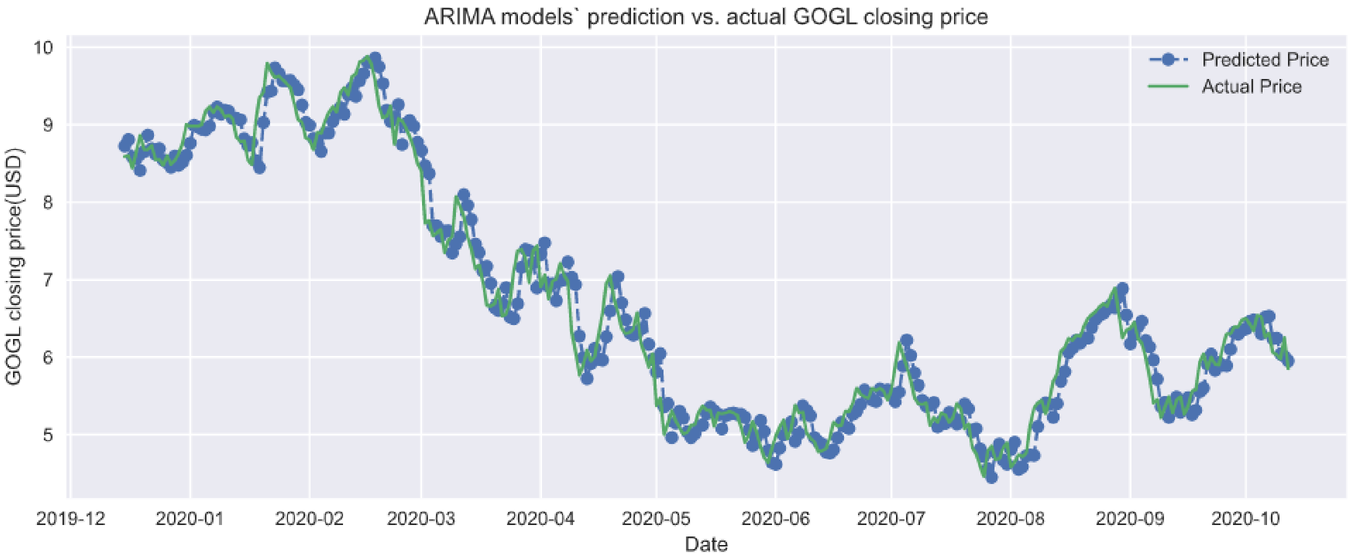

Figure 10 – ARIMA model predictions

The VAR model, as with the ARIMA, did in hindsight not perform as well as expected. Both being outperformed by the moving average strategy. The trading results from the VAR and ARIMA models are not important, but the lack of high performing benchmarks to the transformer model makes these results disappointing.

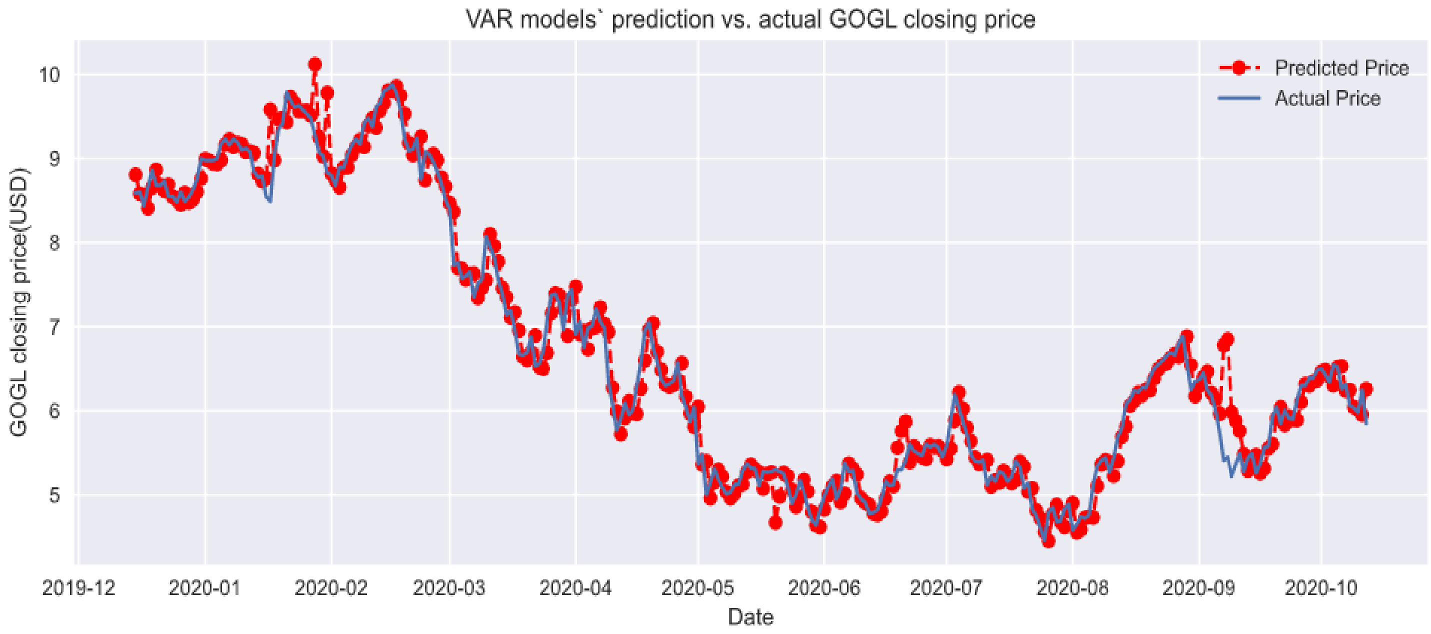

Figure 11 – VAR model predictions

The transformer model with selected features is the highest performing model, even though, as we can see from table 1, the price prediction itself is limited in accuracy. However, due to the high score in directional accuracy, it generates a 58% profit over the test period, which is the highest of all our models. It also has the best overall score on annualized Sharpe ratio.

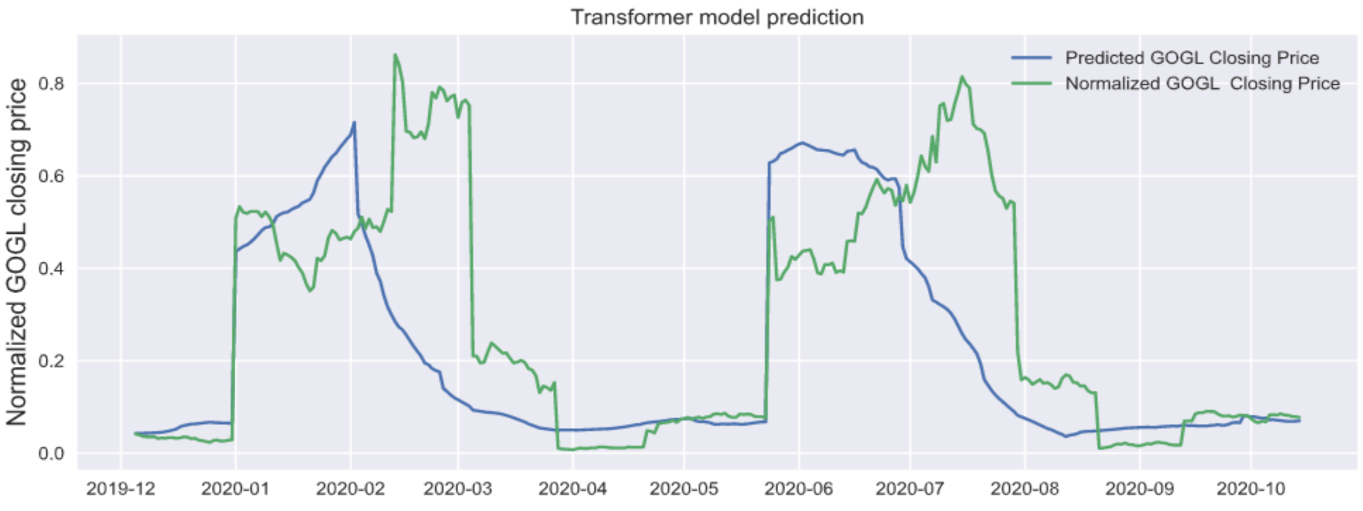

Figure 12 - Transformer model prediction

Originally, we only used FFA-rates in our prediction models, but after including other relevant non-FFA features like S&P500 and Brent futures the performance of the models increased, both in terms of accuracy and trading results. We also experimented with using every feature, but this increased the computational challenge and time consumption without increasing the performance of the models, and we therefore decided to stick with the 30 best features from the feature selection test. We did not train our model on any fewer than 30 features.

The general trend of the GOGL stock is since the mid-2000s a long negative one. Since peaking at 225 dollars in 2005 the stock price has been going down, and today`s level of 4.16 dollars is severely lower than what it historically has been traded on. Also, we must address that there in 2016 was a 1 to 5 reverse split of the GOGL stock, most likely a consequence of the negative trend of GOGL`s stock price. Stock splits are widely used as a way of guiding the stock price to a level which the company board finds satisfactory, often following a large growth or decline. Because of the already large fluctuations of the GOGL stock price in our data set we did not believe this split would have further implications for our model’s accuracy. It would be interesting to investigate the model’s performance in a perhaps slightly less volatile scenario, to see how well it would perform.

7 Concluding remarks

7.1 Limitations

We have raised concern about the fall of GOGL stock price in recent years, regarding the relevance of the entirety of our data set. The downfall in GOGL stock price does not seem to be mirrored in the FFA-rates, and it could therefore be difficult to predict today`s price based on the FFA-rates. This could indeed prove our data set to be less relevant, and harder to make accurate predictions on than with a stock that has had a steadier development. To adjust for this, we originally opted to utilize intraday data for the FFA-rates, but unfortunately were unable to get a hold of data of a sufficient quality.

One of the most effective ways of optimizing a model is through trial and error with different combinations of the factors included in the model. There is a vast amount of different data splits, features, and hyperparameters which can be tested together for increased performance. Due to the almost endless number of combinations, we worked through the different factors one step at a time, by first experimenting with data splits, then features, and then with hyperparameters through cross-validation. Hereby determining one step before experimenting with the next. This leaves out several combinations which argues for possible undiscovered potential within the model.

It would also be interesting to see how intraday data would perform in the model. Being more compact in time and hence providing more data from a more relevant time period, it would probably increase the model's performance. On the other hand, shipping is cyclical, and a typical cycle lasts for 7 years. This implies that a longer training period would make the model more prepared for cyclical changes in the future. Our study covers over 8 years which is more than the usual short-term shipping cycle and should be enough time for our model to pick up underlying trends and tendencies in the market.

We also believe that the amount of benchmark models could be higher. To test how well our transformer model performed, we should perhaps have tested it up against other machine learning models or neural network types, to see how models with the same level of complexity perform. Due to the time-demanding task of learning and programming a neural network and other models, time was not sufficient to create other machine learning benchmark models.

7.2 Conclusion

The results from this study indicate that the transformer model has some predictive power on the GOGL stock price's directional movements, using FFA-rates and other relevant non-FFA features. The transformer model most effectively exploited the informational relationship between FFAs and GOGL, outperforming every other model during the testing period. The usage of feature selection to select the most relevant and influential features improved the evaluation metrics, directional accuracy and trading profit compared to a model which use every single FFA-rate and other non-FFA feature. These results indicate that it to a certain degree is possible to predict the directional movements of the GOGL stock based on FFA data, arguing for a proxy relationship between the FFA-market and GOGL performance.

The directional predictability is in line with our theoretical review which argues for FFAs holding a predictive power subject to market direction. The source of the effectiveness of the transformer model is unsure and could have grounds in several of the economic principles we have discussed. The true advantage of applying such a model is its effectiveness in recognizing patterns and predicting the market's reaction to certain types of information, whether these are lead-lag effects between spot and FFA market or a predictive power inherent in the FFAs which can be applied to shipping securities. The success of the moving average model can also hold some information about the discussed short-term momentum effect which seemingly is present in both the freight- and stock market. These strategies become advantageous when there exist trends within the market, which may be the reason the MA-strategy was relatively effective.

Although the thesis was not a direct test of the EMH in neither the stock- or FFA market, we were successful in generating excess profits compared to our benchmark models, having eliminated existing biases within the data set. This could provide some insight into the EMH in a stock market perspective since that is where we physically acquired profits. Although, keeping in mind the size and fragmentation of the stock market it would be unreasonable to state our thesis as proof of an inefficient stock market. Regarding the EMH in the FFA-market, this market must be efficient if it is to provide valid information on future earnings of GOGL, and even though our models accuracy in directional predictions of GOGL cannot be directly linked to an earnings increase, it certainly provides some evidence of FFA-markets inheriting information on short-term stock performance, suggesting an efficient FFA-market.

The results from our feature extraction show that it is the 4TC_C contracts that perform best out of all the selected relevant features. Out of the non-FFA features, the S&P500, Brent Future, and Baltic Dry Index made the cut. Somewhat surprisingly for us, none of the common commodities like coal, grain and iron ore made the cut of influential features. Feature selection proved to be a pivotal process during our work, as it was one of the efforts that brought down prediction errors the most.

Finally, even though our transformer model displays excellent directional predictive power, the testing period is somewhat short, and not robust evidence of the transformer model outperforming the benchmark models. With this in mind, we recommend further research on the relationship, and predictive power on dry bulk shipping stocks using FFA-rates and other relevant features.

8 Bibliography

1. Adland, R.O., Strandenes, S.P., (May, 2006), *Market efficiency in the bulk freight market revisited,* Retrieved from MARIT. POL. MGMT, VOL. 33, NO. 2, 107–117
2. Adland, R.O., Ameln, H,. Børnes, E, A,. (2020), *Hedging ship price risk using freight derivatives in the drybulk market*, Retrieved from Journal of Shipping and Trade, 2020.
3. Adland, R.O., (2000), *Technical trading rule performance in the second-hand asset markets in bulk shipping*. Retrieved from http://hdl.handle.net/11250/166566
4. Adland, R.O., Koekebakker, S., (2004), *Market efficiency in the second-hand market for bulk ships*. Retrieved from Maritime Economics and Logistics 6 (1), 1–15.
5. Alexandridis, G., Sahoo, S., Visvikis, I., (2017), *Economic information transmissions and liquidity between shipping markets: New evidence from freight derivatives.* Retrieved from Logistics and Transportation Review, 98, 82–104.
6. Alizadeh, A., Nomikos, N., K., (2009), *Shipping Derivatives and Risk Management*, Faculty of Finance, Cass Business School, City University, London
7. Alizadeh, A., Nomikos, N., (2007), *Investment timing and trading strategies in the sale and purchase market for ships*. Retrieved from Transportation Research Part B: Methodological. 41. 126-143.
8. Andersson, J., (2020), Lecture slides, exercise sets.
9. Batchelor, R., Alizadeh, A., Visvikis, I., (2007), *Forecasting spot and forward prices in the international freight market*. Retrieved from International Journal of Forecasting 23 (2007) 101–114
10. Beenstock, M., Vergottis, A., (1989), *An econometric model of the world market for dry cargo freight and shipping*. Retrieved from Applied Economics 21, 339–359.
11. Brock, W., Lakonishok, J., LeBaron, B., (1992), *Simple technical trading rules and the stochastic properties of stock returns*, Retrieved from Journal of Finance 47 (5), 1731–1764.
12. Brownlee, J., (2016), *Deep learning with Python: Develop deep learning models on Theano and TensorFlow using Keras*. Place of publication not identified: Machine Learning Mastery.

13. Brownlee, J., (2016), *Deep learning with Python: Develop deep learning models on Theano and TensorFlow using Keras*. Place of publication not identified: Machine Learning Mastery.

14. Brownlee, J., (2020, August 14), *How to Check if Time Series Data is Stationary with Python*. Retrieved from https://machinelearningmastery.com/time-series-data-stationary-python/

15. Brownlee, J., (2020, August 14), *How to Check if Time Series Data is Stationary with Python*. Retrieved from https://machinelearningmastery.com/time-series-data-stationary-python/

16. Brownlee, J., (2020, June 22), *How to Remove Trends and Seasonality with a Difference Transform in Python.* Retrieved from https://machinelearningmastery.com/remove-trends-seasonality-difference-transform-python
17. Company Introduction Golden Ocean, (2020), Retrieved from https://www.goldenocean.bm/company- introduction/
18. Cornell, B., (October 2019), *Medallion Fund: The Ultimate Counterexample?*, Retrieved from:https://www.cornell-capital.com/blog/2020/02/medallion-fund-the-ultimate-counterexample.html
19. Cullinane, K., (1992), *A short-term adaptive forecasting model for BIFFEX speculation: a Box—Jenkins approach.* Retrieved from Maritime Policy & Management, 19(2), 91–114.
20. Devlin, J., Chang, M., Lee, K., Toutanova, K., (2019, May 24), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from https://arxiv.org/abs/1810.04805
21. Devlin, J., Chang, M., Lee, K., Toutanova, K., (2019, May 24), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from https://arxiv.org/abs/1810.04805
22. Devlin, J., Chang, M., Lee, K., Toutanova, K., (2019, May 24), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from https://arxiv.org/abs/1810.04805
23. Dewey, R., Moallemi, C.,(2019), *The Unsolved Mystery of the Medallion Fund's Success*, Retrieved from: https://www.bloomberg.com/news/articles/2019-11-12/the-unsolved-mystery-of-the-medallion-fund-s-success
24. Diebold, F.X., Mariano, R.,S., (1995), *Comparing Predictive Accuracy*. Retrieved from Journal of Business and Economic Statistics, 13: 253-63.
25. Fama, E., (1970), *Efficient Capital Markets: A Review of Theory and Empirical Work*. Retrieved from Journal of Finance. **25** (2): 383–417.
26. Fleet composition GOGL, (2020), Retrieved from https://www.goldenocean.bm/fleet/
27. Fuller, R.J., Kling, J.L., (1990), *Is the stock market predictable?* Retrieved from The Journal of Portfolio Management 15 (4), 28–36.
28. Fuller, R.J., Kling, J.L., (1994), *Can regression-based models predict stock and bond returns*. Retrieved from The Journal of Portfolio Management 19 (3), 56–63.
29. Gajawada, S. K., (2019, October 20), *Chi-Square Test for Feature Selection in Machine learning*. Retrieved from https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223
30. Gajawada, S. K., (2019, October 20), *Chi-Square Test for Feature Selection in Machine learning*. Retrieved from https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223
31. Glen, D. R., Owen, M., van der Meer, R., (1981*), Spot and time charter rates for tankers 1970-1977,* Journal of Transport Economics and Policy, 15 (1) (1981), pp. 45-58

32. Glen, S. (2020, September 05), *Mean absolute percentage error (MAPE*). Retrieved from https://www.statisticshowto.com/mean-absolute-percentage-error-mape/
33. Hao, K., (2020, April 02), *What is machine learning?* Retrieved from: https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/
34. Hao, K., (2020, April 02*), What is machine learning?* Retrieved from: https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/
35. Haykin, S., (2009), Neural Networks and Learning Machines Third Edition. New Jersey: Pearson Education.
36. Hyndman, R. J., Athanasopoulos, G., (2018), *Forecasting principles and practice*. Melbourne, Australia: OTexts.
37. James, G., Witten, D., Hastie, T., Tibshirani, R., (2017), *An introduction to statistical learning with applications* in R. New York: Springer.
38. Jensen, M. C., (1978), *Some Anomalous Evidence Regarding Market Efficiency*, Retrieved from Journal of Financial Economics, Vol. 6, Nos. 2/3 (1978) 95-101.
39. Johnson, M. K., (2019, June 21), *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Retrieved from https://bookdown.org/max/FES/recursive-feature-elimination.html
40. Kasimati, E., Veraros, N., (2017), *Accuracy of forward freight agreements in forecasting future freight rates*. Retrieved from Applied Economics 50(7), Pages 743-756
41. Kavussanos, M. G., Visvikis, I., (2004), *Over-the-Counter Forward Contracts and Spot Price Volatility in Shipping*. Retrieved from Transportation Reasearch Part E Logistics and Transportation Review
42. Kavussanos, M. G., (1996), *Comparisons of volatility in the dry-cargo ship sector: spot versus time charters, and smaller versus larger vessels*. Retrieved from J Transp Econ Policy 30(1):67–82
43. Kavussanos, M. G., Visvikis, I., Menachof, D., (January, 2004), *The Unbiasedness Hypothesis in the Freight Forward Market: Evidence from Cointegration Tests*, Retrieved from Review of Derivatives Research
44. Kavussanos, M. G., Visvikis, I. D., (August, 2004), *Market interactions in returns and volatilities between spot and forward shipping freight markets*, Retrieved from Journal of Banking & Finance,
45. Kavussanos, M. G., Nomikos, N. K., (2001), *Price Discovery, Causality and Forecasting in the Freight Futures Market*, SSRN Electronic Journal
46. Klovland, J. T., (2002), *Business cycles, commodity prices and shipping freight rates: Some evidence from the pre-WWI period.* Retrieved from http://hdl.handle.net/11250/165223
47. Lo, A. W., MacKinley, A. C., (1999), *A Non-Random Walk Down Wall Street,* Princeton University Press

48. Lo, A. W., Mamaysky, H., Wang, J., (2000), *Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation,* NBER Working Paper No. w7613
49. Michail, N. A., Melas, K. D, (2019), *A Cointegrating stock trading strategy: application to listed shipping companies*, Journal of Shipping and Trade 2019 4:9
50. Myttenaere, A. D., Golden, B., Grand, B. L., Rossi, F., (2016, March 10), *Mean Absolute Percentage Error for regression models*. Retrieved from https://www.sciencedirect.com/science/article/pii/S092523121600332
51. Normalization. (n.d.), Retrieved September 25, 2020, from https://www.codecademy.com/articles/normalization
52. Olah, C., (2015, August 27), *Understanding LSTM Networks*. Retrieved from Colah's blog: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
53. Pelagidis, T., Panagiotopoulos, G., (2019), *Forward Freight Agreements and Market Transparency in the Capesize Sector*. Retrieved from The Asian Journal of Shipping and Logistics, 35(3), 154–162.
54. Prabhakaran, S., (2020, September 17), *ARIMA Model - Complete Guide to Time Series Forecasting in Python*. Retrieved from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
55. Pure expectations theory, Retrieved on October 5th 2020 from https://www.nasdaq.com/glossary/p/pure-expectations-theory
56. Samuleson, P. A., (1965), *Proof that properly anticipated prices fluctuate randomly*. Retrieved from Industrial Management Review, 6, 41–49.
57. Shah, T., (2020, July 10), *About Train, Validation and Test Sets in Machine Learning*. Retrieved from: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7
58. Sharma, N., (2018, May 23), *Ways to Detect and Remove the Outliers*. Retrieved from https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
59. Shiller, R. J., (2000), *Irrational Exuberance,* Princeton University Press
60. Stopford, M., (2009), Maritime Economics. 3rd ed. London: Routledge.
61. 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
62. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I., (2017, December 06), *Attention Is All You Need*. Retrieved from https://arxiv.org/abs/1706.03762
63. Willmott, C. J., Matsuura, K., (2005, December 19), *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance*. Retrieved from http://www.int-res.com/abstracts/cr/v30/n1/p79-82/
64. Zhang, J., Zeng, Q., Zhao, X., (Oct 2014), *Forecasting spot freight rates based on forward freight agreement and time charter contract,* Retrieved from Applied Economics 46(29)

A Appendix

A.1 Descriptive statistics

Table A.1 – Non-FFA features

Feature	Mean	Std.	Min	Max
Golden Ocean	20.22	19.07	2.70	80.80
Coal	79.74	17.99	48.80	119.90
Grain	5.35	1.13	3.61	9.03
Baltic dry index	1056.08	412.45	290.00	2518.00
Brent futures	72.60	24.66	27.88	118.90
CNY/USD	0.15	0.01	0.14	0.17
S&P500	2204.38	429.03	1353.33	3025.86
EUR/USD	1.19	0.10	1.04	1.39
Iron ore	91.99	25.09	47.28	161.62

Table A.2 -FFA-features

Feature	Mean	Std.	Min	Max
4TC_C+1CAL_x	13219.96	4176.53	6155.00	25970.00
4TC_C+1Q_x	12563.17	5904.61	2467.00	33087.00
4TC_C+2CAL_x	13693.66	3330.29	7795.00	22490.00
4TC_C+2Q_x	12335.34	5445.11	3790.00	32300.00
4TC_CCURQ_x	12256.54	6647.67	676.00	34200.00
C4+1CAL_x	8.11	2.13	4.11	11.31
C4+1MON_x	7.34	2.27	2.47	13.59
C4+2CAL_x	8.35	1.94	5.01	11.04
C4+2MON_x	7.43	2.26	2.74	13.50
C4CURMON_x	7.21	2.39	2.40	14.71
C7+1CAL_x	9.33	2.39	4.84	13.97
C7+1MON_x	8.67	2.45	3.47	16.54
C7+2CAL_x	9.56	2.23	5.39	13.11
C7+2MON_x	8.74	2.44	3.76	15.73
C7CURMON_x	8.59	2.59	3.28	17.00
4TC_C+1CAL_y	13219.96	4176.53	6155.00	25970.00
4TC_C+1Q_y	12563.17	5904.61	2467.00	33087.00
4TC_C+2CAL_y	13693.66	3330.29	7795.00	22490.00
4TC_C+2Q_y	12335.34	5445.11	3790.00	32300.00
4TC_CCURQ_y	12256.54	6647.67	676.00	34200.00
4TC_P+1CAL_x	8957.54	2263.25	5060.00	14850.00
4TC_P+1Q_x	9014.28	2746.26	4133.00	15917.00
4TC_P+2CAL_x	9202.38	1741.82	5890.00	13504.00
4TC_P+2Q_x	8904.53	2566.67	4495.00	14600.00
4TC_PCURQ_x	8667.13	3023.98	2783.00	17463.00
C4+1CAL_y	8.11	2.13	4.11	11.31

C4+1MON_y	7.34	2.27	2.47	13.59
C4+2CAL_y	8.35	1.94	5.01	11.04
C4+2MON_y	7.43	2.26	2.74	13.50
C4CURMON_y	7.21	2.39	2.40	14.71
C7+1CAL_y	9.33	2.39	4.84	13.97
C7+1MON_y	8.67	2.45	3.47	16.54
C7+2CAL_y	9.56	2.23	5.39	13.11
C7+2MON_y	8.74	2.44	3.76	15.73
C7CURMON_y	8.59	2.59	3.28	17.00
P2A+1MON_x	15655.56	4255.68	6088.00	27550.00
P2A+2MON_x	15734.59	4147.01	7113.00	25638.00
P2A+3MON_x	15716.97	4048.41	7625.00	25833.00
P2ACURMON_x	15457.55	4631.66	5463.00	29115.00
P3A+1MON_x	8390.76	2716.40	2738.00	16038.00
P3A+2MON_x	8482.52	2629.57	3413.00	15188.00
P3A+3MON_x	8453.96	2534.35	3963.00	14767.00
P3ACURMON_x	8120.26	2915.94	2281.00	16583.00
4TC_C+1CAL	13219.96	4176.53	6155.00	25970.00
4TC_C+1Q	12563.17	5904.61	2467.00	33087.00
4TC_C+2CAL	13693.66	3330.29	7795.00	22490.00
4TC_C+2Q	12335.34	5445.11	3790.00	32300.00
4TC_CCURQ	12256.54	6647.67	676.00	34200.00
4TC_P+1CAL_y	8957.54	2263.25	5060.00	14850.00
4TC_P+1Q_y	9014.28	2746.26	4133.00	15917.00
4TC_P+2CAL_y	9202.38	1741.82	5890.00	13504.00
4TC_P+2Q_y	8904.53	2566.67	4495.00	14600.00
4TC_PCURQ_y	8667.13	3023.98	2783.00	17463.00
C4+1CAL	8.11	2.13	4.11	11.31
C4+1MON	7.34	2.27	2.47	13.59
C4+2CAL	8.35	1.94	5.01	11.04
C4+2MON	7.43	2.26	2.74	13.50
C4CURMON	7.21	2.39	2.40	14.71
C7+1CAL	9.33	2.39	4.84	13.97
C7+1MON	8.67	2.45	3.47	16.54
C7+2CAL	9.56	2.23	5.39	13.11
C7+2MON	8.74	2.44	3.76	15.73
C7CURMON	8.59	2.59	3.28	17.00
P2A+1MON_y	15655.56	4255.68	6088.00	27550.00
P2A+2MON_y	15734.59	4147.01	7113.00	25638.00
P2A+3MON_y	15716.97	4048.41	7625.00	25833.00
P2ACURMON_y	15457.55	4631.66	5463.00	29115.00
P3A+1MON_y	8390.76	2716.40	2738.00	16038.00
P3A+2MON_y	8482.52	2629.57	3413.00	15188.00
P3A+3MON_y	8453.96	2534.35	3963.00	14767.00
P3ACURMON_y	8120.26	2915.94	2281.00	16583.00

A.2 ADF-test and differencing

An augmented Dickey-Fuller test is a unit root test that uses an autoregressive model and optimizes a criterion across several different lag values. The null hypothesis of the ADF test is that the time series can be represented by a unit root, in other words, that it is not stationary. This means that our data has some sort of time-dependent structure. What rejects the null hypothesis is that the time series indeed is stationary:

H_0: If not rejected, the time series has a unit root and is non-stationary. It has a time-dependent structure. P-value >0.01 will fail to reject H_0, the data is not stationary.

H_1: The null hypothesis is rejected; the time series does not have a time-dependent structure and is therefore not stationary. P-value <=0.01, H_0 is rejected, the data has a unit root and is stationary.

We use the Akauke Information Criterion (AIC) to determine the lag of the time series. The adfuller function returns a tuple of statistics including the p-value, number of lags used, number of observations, and a dictionary of critical values. When first running the ADF-test, we found that only 2 of 50 features were stationary. Therefore, to make our data stationary, we performed a differencing. A differencing computes the differences between consecutive observations and subtracts the previous observation from the current. This helps to stabilize the mean of the time series by removing level changes, thereby reducing trend and seasonality. (Brownlee, 2020)

Table A.3 -ADF test and results after differencing on features

Feature	P-Value before differencing	P-Value after differencing	ADF-score	Accept H0 at 1%
4TC_C+1CAL_x	0.45	0.00	-11.83	FALSE
4TC_C+1Q_x	0.05	0.00	-38.35	FALSE
4TC_C+2CAL_x	0.39	0.00	-11.47	FALSE
4TC_C+2Q_x	0.15	0.00	-30.42	FALSE
4TC_CCURQ_x	0.01	0.00	-13.46	FALSE
C4+1CAL_x	0.74	0.00	-25.44	FALSE
C4+1MON_x	0.08	0.00	-8.11	FALSE
C4+2CAL_x	0.73	0.00	-11.38	FALSE
C4+2MON_x	0.32	0.00	-25.32	FALSE
C4CURMON_x	0.06	0.00	-9.20	FALSE
C7+1CAL_x	0.74	0.00	-36.32	FALSE

C7+1MON_x	0.16	0.00	-16.15	FALSE
C7+2CAL_x	0.77	0.00	-17.81	FALSE
C7+2MON_x	0.16	0.00	-11.06	FALSE
C7CURMON_x	0.14	0.00	-9.62	FALSE
4TC_C+1CAL_y	0.45	0.00	-11.83	FALSE
4TC_C+1Q_y	0.05	0.00	-38.35	FALSE
4TC_C+2CAL_y	0.39	0.00	-11.47	FALSE
4TC_C+2Q_y	0.15	0.00	-30.42	FALSE
4TC_CCURQ_y	0.01	0.00	-13.46	FALSE
4TC_P+1CAL	0.58	0.00	-24.93	FALSE
4TC_P+1Q	0.19	0.00	-12.87	FALSE
4TC_P+2CAL	0.43	0.00	-23.67	FALSE
4TC_P+2Q	0.37	0.00	-12.36	FALSE
4TC_PCURQ	0.31	0.00	-36.08	FALSE
C4+1CAL_y	0.74	0.00	-25.44	FALSE
C4+1MON_y	0.08	0.00	-8.11	FALSE
C4+2CAL_y	0.73	0.00	-11.38	FALSE
C4+2MON_y	0.32	0.00	-25.32	FALSE
C4CURMON_y	0.06	0.00	-9.20	FALSE
C7+1CAL_y	0.74	0.00	-36.32	FALSE
C7+1MON_y	0.16	0.00	-16.15	FALSE
C7+2CAL_y	0.77	0.00	-17.81	FALSE
C7+2MON_y	0.16	0.00	-11.06	FALSE
C7CURMON_y	0.14	0.00	-9.62	FALSE
P2A+1MON	0.33	0.00	-34.04	FALSE
P2A+2MON	0.29	0.00	-25.05	FALSE
P2A+3MON	0.49	0.00	-12.01	FALSE
P2ACURMON	0.14	0.00	-20.76	FALSE
P3A+1MON	0.05	0.00	-20.73	FALSE
P3A+2MON	0.03	0.00	-7.07	FALSE
P3A+3MON	0.12	0.00	-7.28	FALSE
P3ACURMON	0.02	0.00	-16.75	FALSE
COAL	0.48	0.00	-8.70	FALSE
IRON ORE	0.50	0.00	-7.46	FALSE
GRAIN	0.42	0.00	-30.91	FALSE
BDRY	0.18	0.00	-10.00	FALSE
BRENT_FUTURES	0.70	0.00	-45.93	FALSE
CNY_USD	0.97	0.00	-19.21	FALSE
S&P500	0.85	0.00	-10.77	FALSE
EUR_USD	0.74	0.00	-43.92	FALSE

A.3 Diebold-Mariano test for directional accuracy

We performed a Diebold-Mariano test for directional accuracy, comparing the transformer-, VAR- and ARIMA model against the random walk with the following hypotheses.

H_0: $E(dt\ )$ = 0 ∀t (Same accuracy for the two forecasts)

P-value > 0.01 will fail to reject H_0 and the prediction accuracy of the transformer model is not significantly better than a random walk.

H_1: $E(dt\ )$ ≠ 0 (different level of accuracy for the two forecasts) (Diebold & Mariano, 1995)

P-value <0.01 will reject H_0, and it will be true that the transformer model has significantly better prediction accuracy than a random walk.

Table A.4 – Diebold-Mariano score

Model	P-value	10% interval	5% interval	1% interval
Transformer	0.00435	True	True	True
VAR	0.0428	True	False	False
ARIMA	0.0869	False	False	False

The DM-test is two tailed, meaning that the significance level must be split in the upper and lower tail. As we can see from table A.4 the transformer model shows a better directional accuracy than the random walk at a 1% confidence level. This indicates that the directional predictions for the GOGL price is consistent and trustworthy, better than the random walk. As for the VAR model a 10% confidence level will perhaps not be enough to use these models for prediction. The ARIMA has no statistical confidence in arguing its directional accuracy is better than the random walk with a 10% confidence interval and is hence rendered useless for this predictive cause.

A.4 Feature selection results

This is the ranked list of features based on the influence on GOGL stock price. We normalized chi-squared-, ridge regression-, and RFE tests and ranked them from most to least influential. We chose to use the top 30 features shown in table A.5, where the last one to be included is C4+1CAL FFA. In table A.6 we see the next 20 features that were in contention to be used but were discarded due to higher prediction accuracy with only 30 features.

Table A.5 – Used features

Features	Chi squared	Ridge regression	RFE	Mean score
4TC_C+1Q_y	1.00	0.51	0.96	0.82
4TC_C+2Q_y	0.88	0.51	0.89	0.76
4TC_C+1Q_x	1.00	0.51	0.74	0.75
4TC_C+2Q_x	0.88	0.51	0.63	0.67
4TC_CCURQ_y	0.92	0.51	0.57	0.66
4TC_CCURQ_x	0.92	0.51	0.54	0.66
P2ACURMON	0.43	0.51	0.98	0.64
4TC_C+1CAL_y	0.60	0.51	0.80	0.64
4TC_C+2CAL_y	0.38	0.51	1.00	0.63
4TC_P+2Q	0.29	0.51	1.00	0.60
4TC_C+1CAL_x	0.60	0.51	0.67	0.60
P3ACURMON	0.34	0.51	0.91	0.59
P2A+1MON	0.42	0.51	0.83	0.59
4TC_C+2CAL_x	0.38	0.51	0.87	0.59
P2A+3MON	0.41	0.51	0.72	0.54
P3A+2MON	0.34	0.51	0.78	0.54
4TC_PCURQ	0.34	0.51	0.76	0.54
4TC_P+1CAL	0.24	0.51	0.85	0.53
4TC_P+2CAL	0.14	0.51	0.93	0.53
S&P500	0.04	0.50	1.00	0.52
4TC_P+1Q	0.33	0.51	0.70	0.51
P2A+2MON	0.43	0.51	0.59	0.51
P3A+1MON	0.33	0.51	0.65	0.50
C7+1CAL_x	0.00	1.00	0.46	0.49
C7+1CAL_y	0.00	1.00	0.43	0.48
P3A+3MON	0.31	0.51	0.61	0.48
BRENT_FUTURES	0.00	0.60	0.50	0.37
C4+1CAL_y	0.00	0.66	0.41	0.36
BDRY	0.04	0.51	0.52	0.36
C4+1CAL_x	0.00	0.66	0.39	0.35

Table A.6 – Discarded features

C7CURMON_y	0.00	0.63	0.28	0.30
C7CURMON_x	0.00	0.63	0.26	0.30
COAL	0.00	0.41	0.48	0.30
C4CURMON_x	0.00	0.66	0.13	0.26
C4+2MON_x	0.00	0.70	0.09	0.26
C4CURMON_y	0.00	0.66	0.11	0.26
C4+2MON_y	0.00	0.70	0.07	0.25
C7+1MON_x	0.00	0.51	0.17	0.23
C4+1MON_x	0.00	0.63	0.04	0.22
C7+2CAL_y	0.00	0.34	0.33	0.22

C7+1MON_y	0.00	0.51	0.15	0.22
C4+1MON_y	0.00	0.63	0.02	0.22
C7+2CAL_x	0.00	0.34	0.30	0.22
C7+2MON_y	0.00	0.38	0.24	0.21
C7+2MON_x	0.00	0.38	0.22	0.20
CNY_USD	0	0.51	0	0.17
GRAIN	0	0.19	0.20	0.13
C4+2CAL_y	0.00	0.00	0.37	0.12
C4+2CAL_x	0.00	0	0.35	0.12

A.5 Transformer neural network

In the first step of our transformer implementation, we must consider how to deal with the notion of time in the model. This being a time series, time is an important feature. When processing sequential data with a transformer model, the sequences are all being run through the model at once, making extraction of sequential dependencies challenging. The transformer model requires a notion of time when processing the GOGL stock price. To make the model able to understand the time aspect, we use time embeddings, making it so that a stock price from 2020 is more relevant to the model than one from 2012.

To overcome the time sequence issue and make the model able to understand how to weight observations based on when it happened, we will implement the Time to Vector method, based on the paper *BERT: Pre-training of Deep Bidirectional Transformers for Language*, on how to learn a vector the representation of time. (Devlin, J., et.al, 2019).

$$t2v(\pi)[i] = w_i\pi + \vartheta_{i,}, if\ i = 0 \qquad \text{(A.1)}$$

$$t2v(\pi)[i] = F(w_i\pi + \vartheta_{i,}), if\ 1 \leq i \leq k \qquad \text{(A.2)}$$

The time vector *t2v* is based on two components, where $w_i\pi + \vartheta_{i,}$is the linear and $F(w_i\pi + \vartheta_{i,})$, is the periodic feature of the time vector. The linear function can be written much easier, as $y = a_i x + b_i$ which is a standard linear function, where a defines the slope and b is the constant which defines where our time-series intercepts the y-axis.

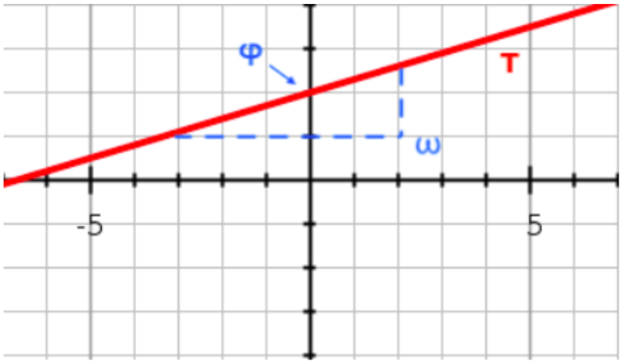

Figure A.1 – 2D representation of non-periodic time feature

The second component represents the periodic feature of the time vector and has the same linear term, but is here wrapped in the $F(w_i\pi + \vartheta_{i,})$, function. This can be visualized like this:

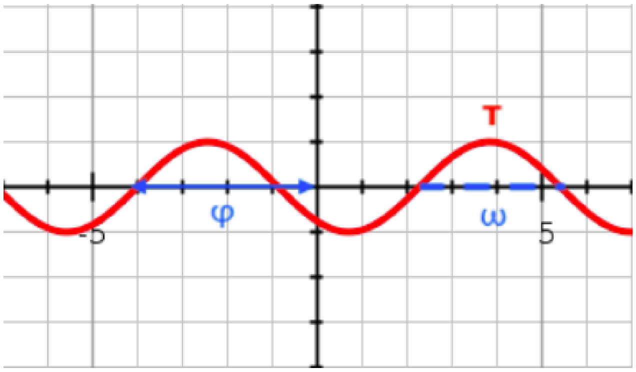

Figure A.2 – 2D representation of periodic time feature

To implement the time 2 vector in our code, we use the Keras framework and define the time vector as a Keras layer. We initiate 4 matrixes, 2 for w_i and 2 for $\vartheta_{i,}$ since we need matrixes both for the linear and periodical features.

Transformer

The next step in our model will be the main part, the transformer. The transformer model architecture is based on a self-attention algorithm. This helps the model focus on the most relevant parts of the time series to improve the accuracy of its predictions. This self-attention mechanism consists of both a single- and multi-head attention. These self-attention mechanisms are connecting the time series sequences and thereby leading to the creation of long-term dependency understanding. This is what elevates the transformer model from models like the LSTM.

The model uses the time vectors in combination with the Golden Ocean stock price and every FFA-feature as input for the transformer. The Time2Vector layer calculates the periodic and non-periodic time features. In the visualization below, we can see how the time features are concatenated with the FFA-rates and GOGL price data, forming a matrix with the shape (16, 64, 30) which is the batch size (16), the sequence length (64) and the, now increased by two, number of features (30).

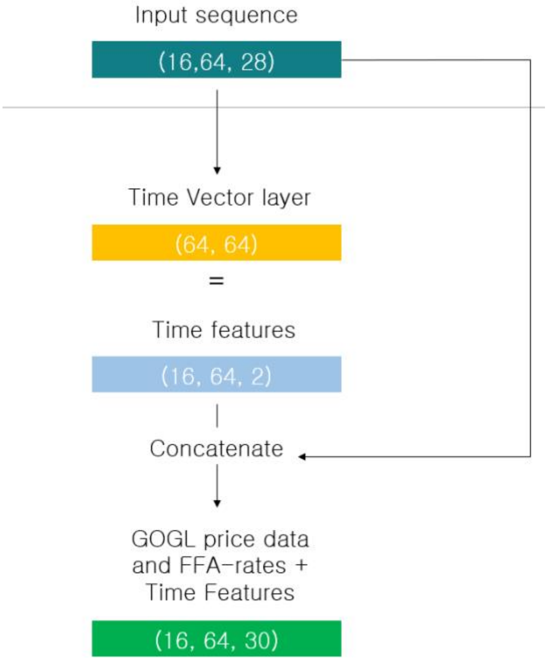

Figure A.3 – Time to Vector and feature concatenation

This combined time, FFA-rates + GOGL price data featured is the initial input in the single-head attention layer. This attention layer takes a query, a key, and a value as inputs. Each query, key, and value represent the FFA-rates, GOGL price data, and the time features. These go through a separate linear transformation through three separate dense layers. Deciding to have 72 layers put here was mostly a trial-and-error process, and something we will touch on more deeply in the hyperparameters section. This linear transformation through these dense layers is what ultimately decides the attention weights. This means how much the model weights the change in values in different features when predicting GOGL stock price. First, we calculate the scalar product of the query and the key inputs. The scalar product for these vectors is the product of the magnitude (the length of the vector) which is multiplied by the cosine of the angle between the vectors. Afterward, the scalar product is divided by 72, which is the number of dense layers, to avoid exploding gradients. These attention weights are then calculated by using the softmax function to get weights that sum up to 1 in total. Finally, the softmax matrix is multiplied with the transformed value matrix.

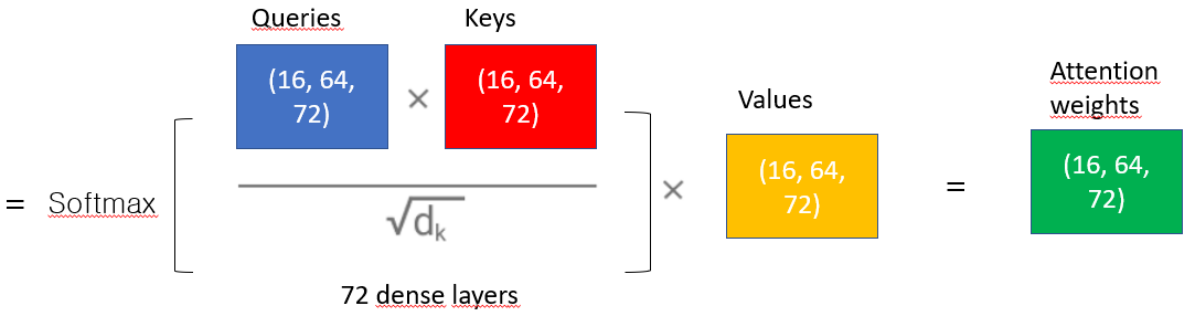

Figure A.4 – The single head attention layers calculation of the attention weights

To further improve upon the accuracy of these attention weights, we added multi-head attention, based on the paper *Attention is all you need* where the authors propose implementation of this. The functionality is to merge the attention weights of several single head attention layers and perform a non-linear transformation with a dense layer. We illustrate the process here:

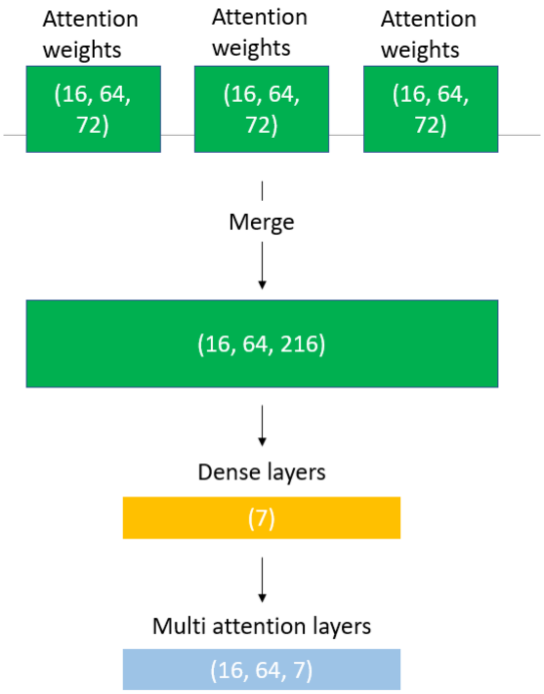

Figure A.5 – Merge of attention weights

We have now created both a single- and multi-head attention mechanism and aggregate this into a layer in the transformer model. Each layer incorporates a self-attention sublayer and a feed-forward sublayer. The self-attention mechanism can connect all steps of the time-series at once, creating a long-term dependency understanding. The self-attention mechanism is then aggregated into a transformer encoder layer. Each layer has a self-attention- and a feedforward sublayer. Each layer is followed by a dropout layer, and after the dropout, we have a residual connection which is formed by adding the initial query input to both sublayer outputs.

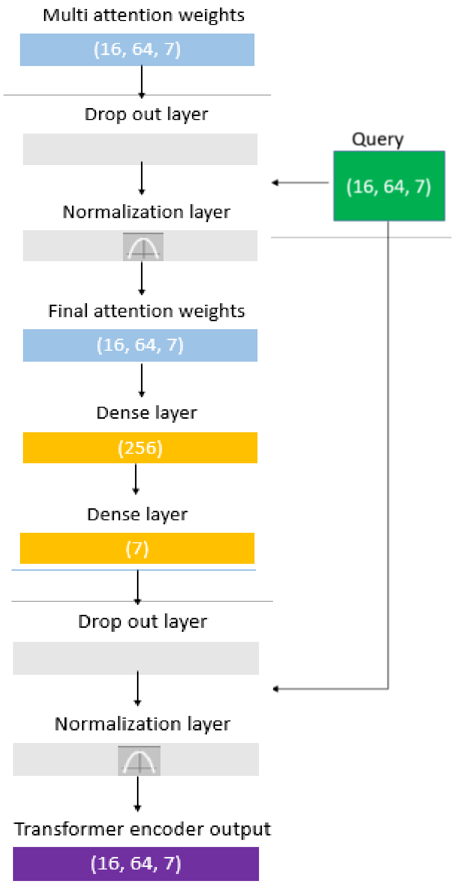

Figure A.6 – Self-attention mechanism and path to transformer output

To summarize the walkthrough of the transformer model, we first initialized the time embedding layers by vectorizing the time series using the time to vector model. We then created the three transformer encoded layers with time features, time vector, and the input sequence to form a concatenated layer with GOGL price data, time, and FFA-features. After this, we calculate the attention weights with single- and multi-head attention layers. The model is then ready to begin the training process.

Hyperparameters

While the internal parameters like attention weights are created by the neural network itself, we have six different hyperparameters in our model which we control and adjust to best fit the model. As mentioned under chapter 3.5 have we used cross-validation to tune our hyperparameters. This method is a part of the sklearn framework in python, and takes as input a list of the range of which we would like to explore the different hyperparameters, and calculates the best combination by training the model on different combinations. The six different hyperparameters in our transformer model is:

- Batch size 16
- Sequence length 64
- Dense layers key 152
- Dense layers value 152
- Number of single-head attention layers 12
- Dense layers 72
- Epochs 50

Batch size is the number of sequences we feed into the model simultaneously, which we after some trial and error chose to test in the interval 8-128 by doubling the number each step. Sequence length is the number of days in each batch size, we used the same range here as for the batch size. We then use cross-validation and found that smaller samples of batch size and sequence lengths were more beneficial for model performance. The number of epochs refers to the number of times we train the model to adjust weights. 50 epochs were not necessary, since the MSE did not improve notably after a couple of runs, but we decided to keep it at a high number as the model will save the run with the best evaluation metrics. Dense layers are the layers of neurons in the model, change in these values did not have a significant impact on model accuracy, and were only tested with a range of three values.