AALBORG UNIVERSITET

Exploring the Option-to-Stock volume Ratio and possible Financial Gain

Masters in finance
10th Semester
Aalborg University Business School

Author: Kristian Kajhøj Nielsen - 20163491
Supervisor: Frederik Steen Lundtofte

Number of Characters: 78160
Number of Pages: 52

Table of contents

Abstract

The aim of the thesis is to understand the option-to-stock volume ratio, and how it affects future prices, while simultaneously using this knowledge to create a trading strategy for financial gain. The study uses 5-year period of weekly data for 12 individual stocks.
The findings of the thesis, while none being statistically significant, suggest that there could be a negative correlation between return and O/S ratio. Secondly, the results from the trading strategy suggest that there's a 50-percentage chance of the strategy having higher returns. however, there is no correlation between the results and the corresponding significance of the O/S variable. The final verdict is that the results lack significance, which highlights the importance for further research and improvements...

Keywords: Option-to-Stock volume ratio, Stock returns, Financial gain, Predictive power, XGboosting.

Introduction

The financial markets have undergone significant transformations in recent years, propelled by various technological advancements, regulatory changes, and a general shift in investors' behavior. Furthermore, the recent Covid-19 pandemic created a major earthquake in the financial market, paving the way for an influx of new investors, hoping to make a quick gain on the volatile situation.
The changes have required a deeper understanding of market indicators and variables, to make informed investment decisions. The introduction of options trading has further added a layer of complexity, where investors have the option, not obligation, to buy the underlying security within a pre-specified timeframe.
In this context, this master thesis investigates the relationship between the option-to-stock volume ratio and stock returns. The aim is to explore the potential predictive power of this ratio, by creating a forecasting study and employing a buy/sell trading strategy. Furthermore, the thesis seeks to evaluate whether the information can be effectively used to generate financial gains.

The motivation behind this research lies in the present complexity of the financial markets and the constant search for innovative approaches to optimize investment strategies. While traditional indicators such as price/earnings ratios, price/book ratios, etc. Having been extensively studied, the option-to-stock volume ratio remains relatively unexplored in the field of predicting future stock returns.
Understanding the implications of the option-to-stock volume ratio on stock performance can provide valuable insights for market participants, including individual investors, traders, and portfolio managers, whom all seek to maximize their Sharpe ratio. By identifying patterns and potential cause-effect relationships between the ratio and the underlying stock returns, we gain a deeper understanding of the dynamic at play and potentially uncover an untapped opportunity to create financial gains.

Moreover, the findings of this research may have broader implications for risk managers seeking financial stability, due to the increasing options trading activity. As options trading continues to gain prominence among investors, combined with the interconnectedness of the markets, comprehending the option-to-stock volume ratio might aid in a better assessment of potential risks associated with derivative instruments and develop better and more robust risk management frameworks.

With this in mind, the research questions addressed in this thesis are as follows.

Research questions

1. *How does the option-to-stock volume ratio affect stock returns?*
2. *To what extent can this information be utilized to make informed investment decisions that could potentially yield financial gains?*

The aim of the project is two-fold. First, to fully comprehend the option-to-stock volume ratio, and to understand how it affects future price changes. This is done by reviewing the written literature on the variable and comparing the different studies, their approach, and their results. Secondly, to create a trading strategy based on the O/S ratio and Fama & French 3 factors, to compare the return for the trading strategy, with the return for each stock. This is done by running a machine learning algorithm, namely XGboost, which performs predictions based on the data, followed by incorporating a trading strategy, that buys the stock on positive predictions and sells on negative predictions.
By investigating these research questions, this thesis aims to contribute to the existing body of knowledge in the field of finance, particularly in the domain of options trading and its impact on stock market dynamics.

Project design

To address the research questions, the thesis is comprised of sections, each with its purpose. A brief introduction of each section is provided, at the start.

- Section 2 – Literature review
- Section 3 – Generalized data process
- Section 4 – Methodology
- Section 5 – Theory and Application
- Section 6 – Analysis & Results
- Section 7 – Discussion & Limitations
- Section 8 – Conclusion

Section 2 - Literature review

Informed traders may be inclined to opt for options trading rather than trading the underlying stock, due to the leverage mechanisms of the options market. Options trading provides the investor with tools to leverage their investment, while simultaneously maintaining risk levels that satiate the investors' risk appetite. Studies by (M. Cao & Wei, 2010) examine the information asymmetry in the options market and conclude that the options market generally has a more significant asymmetry than the stock market and this drives investors to options. Information asymmetry, meaning that traders in the options market generally possess favorable information, when initiating the trade of options. The notion that informed traders engage in the options market is supported by evidence that shows an increase in both puts and calls before positive and negative news about the future. In a paper by (C. Cao et al., 2005), an increase in call options volume is noticed when a firm takeover is about to take place in the nearest future. The opposite is examined in(Hao et al., 2013) , where the total trading volume of put options saw an increase the day before the expected negative earnings announcements.
In an article by (Pan & Poteshman, 2006), the informational content of options volume and its effect on future stock prices is examined. They found strong evidence that informed trading takes place in the options market, which is consistent with prior research. Additionally, they found that the stock market takes several weeks to fully incorporate the information stemming from the options market. This is not because of market inefficiency, but because of a disconnect between the stock market and the options market. This disconnect is due to nonpublic information, held by the informed traders in the options market., however, the information only holds firm-specific stocks, as a comparison with the overall market, did not produce significant results. Furthermore, they found that the options market is uniquely suited for creating volatile trades, due to the nature of the leveraging effect. Thus, creating lucrative opportunities, for informed traders.
If informed investors are drawn to the options market, then perhaps the options market can, with some significance, predict the future return of stock prices. The prediction would stem from the increase in volume across puts and calls for the particular stock.

With the knowledge that informed trading taking place in the options market, and the general understanding that options trading is equipped to with some certainty predict future stock prices, this paper examines the literature on different option metrics, used by practitioners and theorists.

The literature on options metrics for trading and predicting future returns can be boiled down to the two most widely used C/P & O/S.
C/P is the volume of call options, divided by the volume of put options.
O/S is the total volume of options calls and puts, divided by the volume of the particular stock.
A few studies have looked at the C/P ratio and found contradicting results but also similarities as concluded by (Houlihan & Creamer, 2019) in which the ratio’s primary role is to act as investor sentiment, showcasing which direction informed traders believe the stock is moving. Therefore, it is used as a bet against or for future price increases or decreases. It may be a ratio used simply for the act of hedging one's portfolio, against dramatic price decreases, as

this is a common strategy amongst options traders in general. However, this last notion is simply speculation by the author and has no grounds in the written literature.
The focus of the project is based on the O/S ratio, which has substantially well-written literature, that generally speaks for a significant metric, regarding information, one of these studies is done by (Roll et al., 2010) . The study, which they view as the first of its kind, helps shed some light on what drives volume in the options market, and the correlation between O/S and the underlying stock price changes. They show that the option-driving forces are significantly related to the following: size, trading costs, implied volatility, option leverage(delta), and institutional holdings. They also found that O/S is less related to analysts' rating & their forecast dispersion. The last part is interesting because it means that they differ from financial analysts, meaning missing information is hidden in the O/S ratio, which is not captured by the analyst/market. Additionally, the study finds evidence of O/S increasing sharply, days before the earnings announcement, and also a linkage that suggests O/S affects prices, thus reinforcing the study done by(C. Cao et al., 2005)
In a 2012 study,(Johnson & So, 2012) examined the information content in stock and option volume. They found that option to stock volume (O/S) ratio is a negative cross-sectional signal of private information. Stocks in the lowest decile of their research portfolio, outperformed the highest decile of the O/S ratio by 0,34% on a weekly factor-adjusted basis. They explain this as being the result of how informed traders navigate between trading in equity and options markets, depending on the short-sale costs. That is the cost associated with short selling makes informed traders more likely to use options for bad signals than for good ones, and as a result, a high O/S ratio is associated with negative private information and a low O/S ratio is the adverse effect. They also found evidence that the information in these trades is skewed in such faction, that traders focus their attention on options more frequently, when the information they have is negative. Furthermore, they find evidence that the O/S ratio is better and clearer in it signaling power and informational content, compared to using the call-to-put or put-to-call ratio.
(Kim et al., 2017) study the relationship between investor sentiment and the O/S relation and find that the ability of the O/S metric to calculate future returns becomes greater and weaker based on investor sentiment, which is due to higher short sale constraints and irrational demand. Kim et al observed, using a four-factor model, that their O/S metric represented a 0,332% alpha per week, but would drop to 0,155% during periods of low investor sentiment. They also tested a long/short investment strategy based on O/S and found that the strategy works for a maximum of 3-week periods, at which the significance of the alphas diminishes.
A study done by (Blau et al., 2014) compared the two metrics and found interesting implications for both. Firstly, they found that the C/P ratio can predict significant negative returns around daily and weekly levels but does not apply to future returns when looking at monthly intervals. Secondly, they found that the O/S ratio significantly predicts negative returns for all intervals, daily, weekly and monthly. And lastly, they found and concluded that the C/P ratio is the best for daily data, whereas O/S is better at predicting negative returns when looking at both weekly and monthly data. This distinction between the frequency of the data used is an important note and is a contributing reasoning that will be discussed later on in section 3. The most recent study done on the O/S ratio, is by (Woo & Kim, 2021) on the Korean stock exchange. By analyzing 36 stocks and their corresponding options, using data points from 2014 to 2021, they found that both the C/P and O/S ratio shows statistically significant predictor for future returns. Additionally, when both ratios are included in their multiple regression analysis, the C/P ratio becomes insignificant, while the O/S ratio remains

consistent, solidifying the findings from the previous comparative study by (Johnson & So, 2012), where predictability was more robust for O/S than that of C/P.

Given the existing literature on the topic of the O/S ratio as a metric for predicting future returns, the project will be based on the hypothesis that a trading strategy based on the O/S, can significantly predict future stock prices, thus rewarding the investor with a substantial abnormal return. However, when analyzing, based on several years of data, one must first take into consideration the abnormal market conditions that has been and is currently changing the financial market. In the following sections, a brief overview of the market situation is explained, and why this is an important topic to cover, to avoid estimation errors.

Before Covid-19, the financial market was generally in an upwards trend, with steady volatility, however, this all changed when the pandemic broke out. In a matter of 1 month, the S&P index plummeted more than 30%, while investors were trying to cut their losses and transition to the growing Bond market. The high volatility of the stock market had spillover effects on the options, as volatility in options trading essentially means more opportunities. The increase in volatility and the influx of new uninformed traders/speculators/gamblers, resulting from millions of people being in lockdown, with cash and checks on hand, resulted in a lot of volume for the options market. (CNBC, 2020) reported that the surge of new money to the options market was primarily new traders, wanting to gamble their way to quick abnormal returns, with minimal investment. This is an important aspect to take into consideration, when the focus is on creating a trading strategy, using 5 years' worth of data, as some of the data will be before, during, and after this period in time, where a substantial amount of the options, is the direct result of uninformed trading, thus increasing the possibility of faulty readings leading to an ineffective trading strategy. To accommodate for this, one could split the data into three different segments, to ensure that the regression provides similar results.

To fact-check the notion that the options market has seen a steeper growth than the general stock market, I've chosen to illustrate this, by looking at a market-wide ETF SPY, which is an ETF that tracks the S&P500 index and is generally considered one of the biggest options plays, by volume. The reason for this choice is that SPY is a market-wide exposure play that lets investors bet on the entire market, with a single option. Secondly, it's one of, if not the largest ETF by options volume, thereby making it easy for investors to open and close positions in real time. Looking at Appendix B, we see the Adjusted close, stock volume, options volume, and calculated O/S ratio, starting in January 2019 and ending in March 2023, the same period used in the data set. It’s clear that the average level of the O/S ratio is substantially higher than that of the individual stock, the average being 0,5-0,6 confirming the enormous amount of volume and also the trending effect of this ETF. From February 2020 to March 2020, we see a 40pts drop in share price, and a low of 0,289 O/S, however short-lived, the share price has been steadily rising since and the average O/S ratio in 2023 is now around 1.0. This increase is not due to the share volume being lowered, as this stays rather consistent, but the options volume has been increasing and is not showing signs of slowing down, meaning that there’s a possibility of an O/S ratio above 1.
This is not necessarily a problem, but if past literature holds, then an increasing O/S ratio is met with negative future price changes, as we see in the study done by (Johnson & So, 2012). This means that investors, generally have a negative outlook for the SPY and thus, a fair statement would be that traders in the SPY ETF, have a general negative outlook on the market, representing the increasing O/S ratio, as a measure of private negative information.

This is only a simple comparison, to get a general picture of how the market situation has developed during the last couple of years, so a more in-depth view could potentially be used if the goal of the thesis was to analyze the financial landscape. However, this is not the purpose of this paper, because a trading strategy does not pick sides, its purpose is to capitalize on information, to generate positive future returns.

With an understanding of how the variable performs, and the implications of the current market situation, the next section of this paper will be categorized as a Generalized Data process (GDP), in which I showcase the reasoning behind the data used, and how it's obtained.

Section 3 - Generalized Data Proces

3.1 Data selection - Stocks & options

The data used in the thesis is historical data from 12 individual stocks. The data is in a weekly format, with a starting point of April 2018 to April 2023. The weekly adjusted close price & weekly volume is collected from Yahoo Finance, under historical data, with the preset described. Secondly, the weekly option volume data is collected from CBOE, under historical data, with the same presets.
The stocks chosen are the following:

Name of the company	Ticker
Apple	AAPL
Amazon	AMZN
Alphabet	GOOG
Johnson & Johnson	JNJ
J.P. Morgan	JPM
Mastercard	MA
Meta Platforms, formerly Facebook	META
Microsoft	MSFT
Nvidia	NVDA
Procter & Gamble	PG
Tesla	TSLA
Visa	V

Table 1 - Table showing chosen stocks and their tickers.

The data selection process was a simple selection based on the size of the companies, their origin/ trading country, and options trading enabled. The 12 companies are the twelve biggest companies, trading in the USA, with options trading enabled. The selection of companies, for the sake of this thesis, however, remains the same, concerning how the framework is being created.

3.2 Data selection - Period & Frequency

As described earlier in section 2, the idea behind choosing the period from 2018-2023, is that the purpose of the thesis is to create a framework process that estimates and forecasts based on historical data. The forecast is based on the estimation window, so the choice behind the data period is an important factor to include when analyzing the results. The author of the thesis wanted to create a forecast based on the current market situation, thus enabling, both the all-time high periods, pre covid-19, and the lows after, to ensure that the estimation is fed the right amount of information, to forecast the existing trend in the market.
The data uses weekly observations, as opposed to monthly or daily, which is commonly used for most time series analyses. The use of weekly data is purely based on the extensive literature view explained in section 2, where weekly data was shown to have a better correlation with future returns than both daily and monthly. Additionally, the use of weekly data is beneficial when working with options. This is due to the nature of how options expiration works. Options expire each Friday of every week, which means that the weekly data captures the overall volume, this is beneficial because options are generally more sought after, when expiration is long, meaning the volume of options is not constant over the week.

3.4 Data selection – Fama & French

The data used for the Fama & French 3 factor are downloaded from Kenneth R French's library (Kenneth French, n.d.) where weekly data is selected. Then the same period as the stock returns, are used to specify the 3 factors. One thing to notice is that there is a slight mismatch in the data dates, where the 3 factors weekly date is slightly off. However, in this thesis, the dates are set to that of the stock returns, and the 3 factors are simply added to the date. This shouldn't cause problems, but it's worth mentioning, as this might lead to faulty readings from the data, because of the inherent mismatch between the actual dates and dates used.

3.5 Data Processing & Preparation

After downloading the raw data for each stock, the data is then processed and prepared so that it can be used. Below is a step-by-step process of how this is done, in the thesis:

1. Adjusted close and stock volume is selected, as the only variables needed from Yahoo Finance.
2. Simple return is calculated with the following equation: (ValueToday-Valueyesterday)/Valueyesterday
3. Options volume is added and the O/S is calculated with the following equation: (OptionVolume/StockVolume)

4. Fama&French 3 factors' are added to the file.
5. Repeat this process for every single stock.

An example of this is shown in Appendix A.
This concludes the GDP section. The next section will cover the methodology used in the thesis, this includes the scientific approach applied, the overall design/framework of the data analysis, and possible limitations and improvements.

Section 4 - Methodology

4.1 The scientific approach

In this thesis, the scientific approach chosen is the critical realism paradigm. The idea behind this is that the absolute truth is unreachable, but will always remain the goal of the approach, to get as close to the truth as possible. The general idea is that the world is split into two sections, the transitive and intransitive dimensions. The critical realism approach is to focus on the ontological part of the world, i.e., the intransitive elements such as the stock market. By gathering as much information as possible about the stock market, the goal is to better understand the causal mechanism that drives the underlying phenom. The knowledge that is created based on this, helps to explain and understand the absolute truth, without fully reaching it (Buch-Hansen & Nielsen, 2012)
The transitive dimension in critical realism is a conglomerate of socially constructed knowledge based on intransitive elements. This knowledge is neither definitive nor perfect, in regard to fully capturing the absolute truth. Once the ontological parts have been defined, epistemology must be covered. Epistemology is the method used to create and produce knowledge, i.e., which models, theories, and concepts are being used in the thesis, to create knowledge concerning the research question. This is an important aspect, as the methods used will determine how the generated knowledge is perceived. The method used in this thesis is described in detail in the following sections.

4.2 Deductive approach

To generate knowledge based on data and deductive approach has been used. The deductive approach is chosen due to the nature of the research question, in which the author seeks to understand and gain better knowledge of the O/S variable and its ability to predict future returns. The reason behind the deductive approach arose from the literature review, which in general postulates that there is a significant relationship between the O/S variable and future return, thus making this approach beneficial, in regard to testing this postulate. The idea is to test the postulate, that is the correlation between the O/S variable and future price changes, and add to it, the trading strategy to see if there is indeed a correlation that can be exploited for investors.
When working with a deductive approach, however, the limitations are the accuracy and validity of the data being used. In this case, using historical data is considered valid, but the

time frame, frequency, and biased selection of stocks will inevitably be the shortcomings of the results generated. (Buch-Hansen & Nielsen 2012)

4.3 Validity & Reliability

To ensure the validity of the project, the literature used in the literature review is extracted from peer-review sources, and credible authors with more than a few written articles in their resumes. The access to these sources has been through the Aalborg University Primo, which is a search engine database, that grants access to peer-reviewed academic articles, written by the most credible sources. Access to the articles is provided by Aalborg University, and therefore reliability in the sense of using the same information in the numerous articles, is a potential problem for future research.
However, it should be noted that despite their credibility, the author of this thesis does not have access to the data used in the articles, nor does he know how the data has been cleaned and processed.
The raw data in the project, as described in section 3, is derived from credible sources, this ensures both the validity of the numbers being used and also the reliability of the thesis. Any person with an internet connection can obtain the data used in the project, and the GDP section of the thesis describes in detail how the cleaning and processing of the data is done.
The last point of validity is to ensure a connection between what is being researched and the actual results. To ensure this connection, the author strives to connect every section, so that a general idea is maintained throughout the thesis. Additionally, a discussion is created to ensure the connections and any possible loose ends.

Section 5 - Theory & Application

5.1 Stationarity

Before any testing and analysis can be done, the data must be checked for stationarity. This means that the time series data must have a constant mean, constant variance, and autocovariance that is not dependent on time. This is essential if any analysis is to be done to the data, as a non-stationary time series is less than ideal to be used as a forecasting material. To test for this simple code in R, we run the ADF test. See Appendix C for the R script with the ADF testing included. Where the null hypothesis is nonstationary, meaning a P-value below 0,05 results in our data being stationary. Why is this important?
Stationarity is important because it essentially means that the data that's being used constantly changes over time, meaning there's no trend or cycle. If the data is non-stationary, then an estimation and forecasting study would give conflicting results because the estimation might have data that is currently undergoing one type of trend and the other part of the data has a different trend. This is a problem because it makes it harder to conclude anything with statistical significance. If data were to be non-stationary, then usually it means that the amount of data is not great enough, to eliminate non-stationary, thus requiring more data. This was a problem in the early stages of this thesis, as the original idea was to use one year of data with a frequency of every 2 weeks, giving a total of 24 observations. This proved to be a challenge, as most of the variables and returns came back non-stationary. There are certain ways around this, but the easiest fix is to simply include more data points.

5.2 Multiple linear regression

The author has chosen to incorporate the three-factor model, by Eugene Fama and Kenneth French, forwarded known as the FF3 model. The FF3 model is built upon the traditional Capital asset pricing model, with two additional factors. We know from written literature, that the FF3 can capture and explain much of the return for a given stock, however, some uncertainty remains, and the project aims to hopefully capture a fraction of the remains, thus explaining future returns. The thesis focuses on the O/S variable, however, If an analysis was purely based on the O/S variable, then there's a possibility that the regression would not be able to capture enough of the uncertainty regarding the future return, resulting in a poor estimation & forecasting later on in the study. Thus, the inclusion of the FF3 factors will inevitably create better groundwork for the forecasting study (Eugene Fama & Kenneth French, 1992).

The model used in the thesis is illustrated below.

$$Ri - Rf = \alpha i + \beta i(Rm - Rf) + SMB + HML + O/S + \varepsilon i$$

Equation 1

Where Ri is the excess return of the individual stock, minus the risk-free rate Rf. Alpha is the intercept, which represents the stock's expected excess return if others are equal to zero. Rm is the excess return of the market minus the risk-free rate Rf. SMB is the Small minus big factor, and HML is the High minus low factor. O/S is the Option to stock volume variable, that we are testing for, and the error term is the idiosyncratic risk, which is not explained by other factors.
In R we compute this regression by using the lm script. The output of this regression indicates how the different variables affect the return of the stock, with the associated significance level. It's important to mention that the author acknowledges that the significance levels recorded in the project exceed the industry standard of 0,05. However, the purpose of the project is to create the framework and discuss the results, so this problem will be discussed in section 7.

The Regression analysis of each stock forms the basis for understanding the relationship between the O/S variable and the return of the stock. The next step in data analysis is the estimation and forecasting study. This process can be done with various techniques, some simpler, and some more advanced. The author of this project has chosen to work with a machine learning algorithm, known as XGboost, and this algorithm will be introduced in the next section.

5.3 Machine learning & XGboost

In this project, the author has chosen to work with XGboost, originally created by Tiangi Chen in 2014 (Chen & Guestrin, 2016) as it's one of the best machine learning algorithms to use, when trying to estimate and forecast regression or when trying to run classification on a large data sample. It's also considered very easy to use and with a wide array of learning possibilities. The thesis is considered a practical framework, so only a brief introduction to the realm of machine learning and how XGboost works will be introduced.

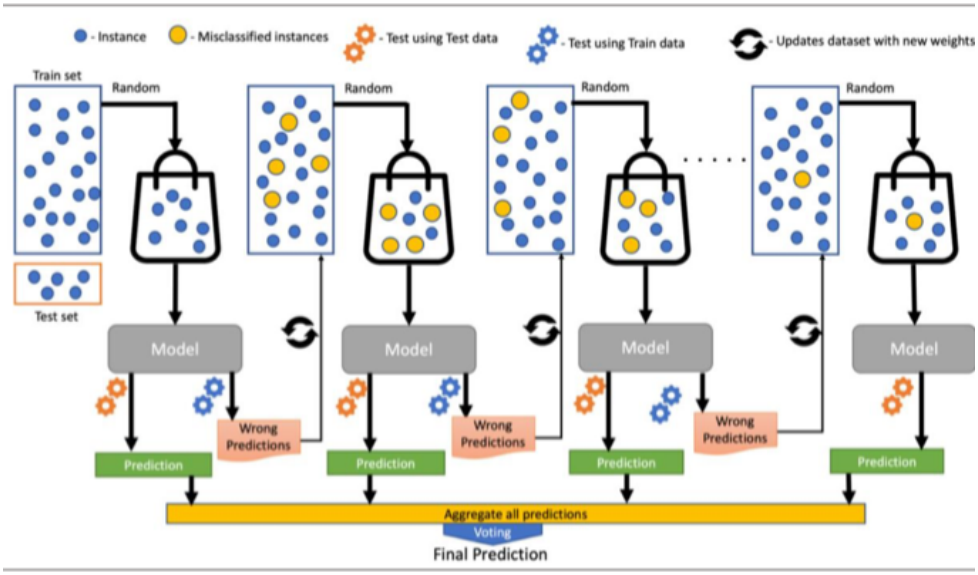
To understand this algorithm, we first have to understand what boosting in data science is. Boosting generally means an increase in performance, in this case, an increase in the performance of our forecasting model. In machine learning, the boosting terminology is described as a "sequential ensemble learning technique" that converts so-called weak learners into strong learners, to increase the accuracy of the model. The term, weak and strong learners generally mean individual models that are either bad or good at guessing. In the case of a forecasting study, a weak learner would be a model with only slightly better accuracy than just guessing, whereas strong learners are models that to a certain degree can predict with good accuracy. The boosting algorithm creates new weak learners that work in a sequential setting, where each new model is based on the last model, thus increasing their predictions to better the overall model. Boosting does not change the previous models/predictors, it only corrects the next model by learning from the mistakes. The general formula for the simple boosting algorithm is shown below.

$$F_i(x) = F_{i-1}(x) + f_i(x)$$

Equation 2

Where Fi(x) is the current model, Fi-1 is the previous model and fi is the weak model.
This computation is then done multiple times, to increase the accuracy.

Since boosting is a greedy form of machine learning, it will overfit and overcomplicate the predictions, thus increasing the errors, and generating bad results. Therefore, it's recommended to set certain boundaries, so-called depth level, which is done to set a limit on the maximum number of iterations. The depth level in the thesis has been tested on different levels, and the consensus is that a depth level between 50/80 iterations produces the best result, which is determined by the lowest Mean-Squared-Error (MSE). After 50 iterations, the boosting algorithm overfits the model increasing MSE. However, in the project, the author has chosen to work with a depth level of 200 for all of the stock, this is done to simplify the process and to make sure that the predictions are fully captured. The continuous model development is called a regression tree, it's illustrated as a tree-like structure, with the roots at the top, and moving downwards, where each leaf/branch is the next iteration. The regression tree has a dependent variable, in this case, return, and several independent variables, here O/S, SMB, HML, and Mkt-rf. The goal is to maximize the dependent variable with the measurement of MSE. The method used in the XGboost is called gradient boosting, which simply means it adjusts the weights of weak predictors in a gradient, a direction in the loss function. Below is a simplified version of how the regression tree is illustrated and how it works.



(Malik et al., 2020)

Image 1 - XGboosting algorithm steps

We see that the data is divided into a test set and a train set, then a random amount of the train set is transferred into the algorithm, the good predictions are stored, and the wrong predictions are cycled through the model again. All this combined is what the XGboost algorithm is capable of doing, which makes it a very efficient tool in analysis.
XGboost machine learning algorithm is much more complex than what's shown above, but the general practicality in how it works and operates is accounted for. The next section focuses on the process of using XGboost to forecast the collected data.

5.4 XGboost practical approach

The practical framework for the analysis consists of splitting and training the data, creating the model, creating signals, predicting returns, back testing with actual data, and plotting the importance matrix and the comparison. This approach is copied for every single stock, to decrease the complexity of the calculations and to simplify the models.
First, the data is split into training and testing, the standard split is 80/20, and this is also the split used in this thesis. The reason behind splitting this particular split is for the algorithm to have 80% of the data to train on, to ensure the most accurate forecast. The parameters for the XGboost model are then set to the following settings:

```
# Train XGBoost model on training data
dtrain <- xgb.DMatrix(as.matrix(train_data[,x_vars]), label = train_data$Returns)
dtest <- xgb.DMatrix(as.matrix(test_data[,x_vars]), label = test_data$Returns)
params <- list(booster = "gbtree", objective = "reg:squarederror", max_depth = 1, eta = 0.1, nthread = 3)
xgb_model <- xgb.train(params, dtrain, nrounds = 200, watchlist = list(train=dtrain, test=dtest), print_every_n = 1)
```

Image 2 - Screenshot of XGboosting settings in R

The objective of the model is to generate models based on the root mean squared error (RMSE), and this is done by creating 200 iterations of the same model, with increasingly lowered RMSE. The depth is set to 1, and multiple levels have been tested, but the consensus

is that a depth above 1, complicates the model and generates worse results, making it harder to compare and back test with the real returns. Eta and the number of threads has not been altered, as these are standard settings for the model when running the XGboost in R.
One important note to consider is that the depth level of 200 is not the perfect level for any of the models but is simply a parameter that the author of the thesis chose, due to some of the models requiring higher levels, to fully minimize RMSE.
In section 6, an overview of each of the RMSE values will be described, additionally, a discussion on this parameter selection will be introduced later in the project, under possible improvements.
After training the data, the predicted stock returns and trading signals are created based on the trained model and are then back tested to the actual values of the stock return.
This is done through both a summary of the model, where min, max, and mean values are compared, and also a plot is created to illustrate the differences.
After model creation is complete, the importance matrix of the factors is plotted. This is done because the XGboost model calculates returns based on all four variables, but the project aims to create a trading strategy, that relies on the O/S variables, thus the importance of this specific variable is ranked above the others, and the data should indicate this.
The full R script is available in Appendix C and will not be rigorously reviewed in the thesis.

The next section will cover the results of the XGboost forecast, the results are split into 2 sections. First, a section where all 12 stocks will be covered, and a walk-through of the results and explanations based on the individual stock will be discussed. Secondly, a comparative overview of all the different outputs and possible outliers.

Section 6 - Analysis & Results

Apple

The Augmented dickey fuller test for Apple confirmed the stationarity of each of the four variables, with only the O/S variable being slightly above 0.05. Therefore, a second test was done on the entire dataset, rather than the training data, to confirm that this value was due to the amount of data, which came back under 0.05, confirming stationarity.

The summary of the regression done on Apple shows that the O/S facto has a negative correlation with the returns of the stock, indicating that perhaps the trading strategy would include some shorting of the stock. The significance is above 0.05. However, this is a common theme in the project, so significance levels for the rest of the regression analysis, will not be commented on, and only serves as a topic of discussion later in the project.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.186318 -0.021514  0.000744  0.020158  0.132703

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01828    0.00784   2.331   0.0209 *
O_S         -0.07660    0.06155  -1.244   0.2150
Mkt_RF       0.01724    0.10966   0.157   0.8752
SMB          0.13638    0.20010   0.682   0.4964
HML          0.04127    0.11171   0.369   0.7122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04016 on 169 degrees of freedom
Multiple R-squared:  0.01271,   Adjusted R-squared:  -0.01065
F-statistic: 0.5441 on 4 and 169 DF,  p-value: 0.7036
```

Image 3 - Screenshot of regression analysis in R

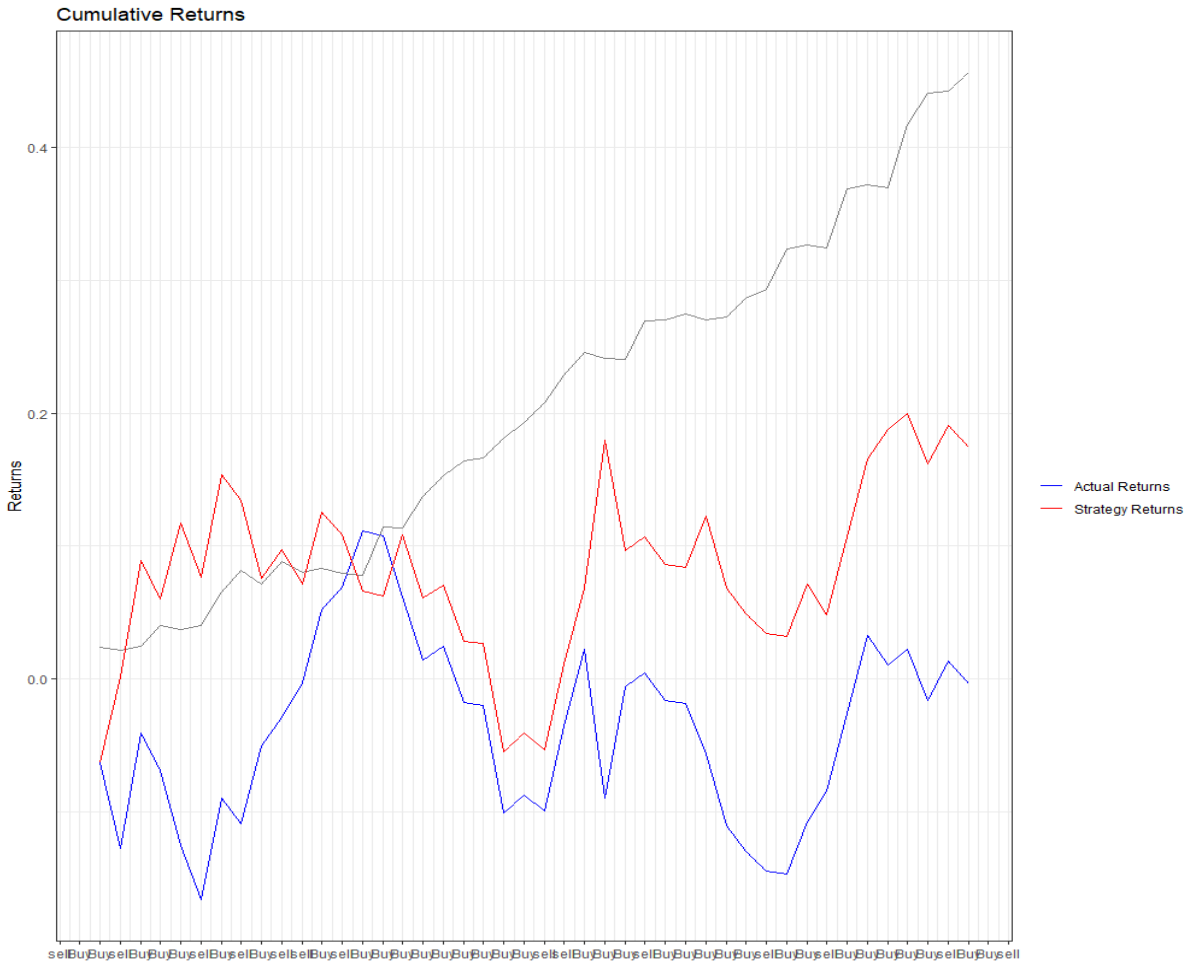
The RMSE values from iterations 1-200, can be found by running the R script, in Appendix C. This indicates that after iteration number 81 of the model, the RMSE values no longer decrease and increase, meaning that the algorithm is overfitting the model. The value used for the prediction is 0.0487204, and the best value is 0.047629, which is only a slight increase in errors.

The return table for Apple is shown below.

Apple	Minimum	Mean	Maximum
Actual Return	-16,6%	-3,7%	11,1%
Predicted return	2,1%	20,4%	45,6%
Strategy Return	-6,3%	8,1%	19,9%

Table 2 - Returns for Apple

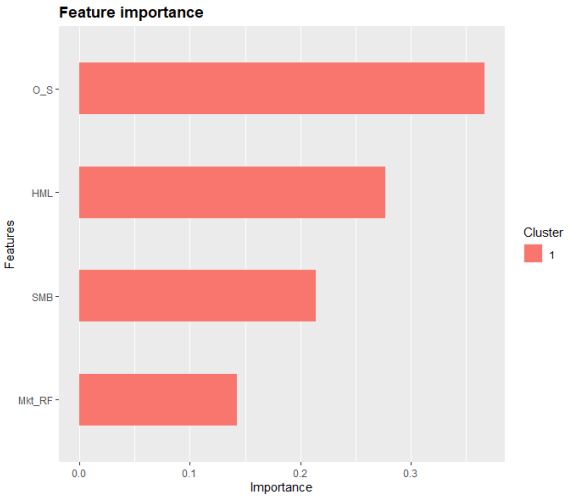
The actual mean return for Apple in the testing period is negative 3,7% with a min/max of -16,6% & 11,1%, this is compared to the cumulative returns for the strategy with a mean of positive 8,1% and min/max of -6,3 & 19,9%. The comparison between these shows that the values based on the model prediction give an overall higher return profile, by incorporating the 12 shorting signals in the trading strategy. It should be noted that the predicted returns by the model alone, without incorporating the trading signals, give a prediction higher than the actual and the strategy returns, which may be an indication that the strategy is missing something. Below are the plot lines for the three different returns.



Plot 1 - Apples return plot

Looking at the plot, we see that the trading strategy follows the actual returns, indicating a decent correlation of the predicted returns when combined with the trading strategy. However, when looking at the predicted returns alone, it's clear that the model generates an overall bullish sentiment on the stock, representing a positive outlook only.

The level of importance for all four variables is illustrated below. We see that the O/S variable is the variable with the most importance, for the XGboost model to estimate and forecast its values. This was also the most significant variable in the estimation summary, so there might be a linkage between the two.



Plot 2 - Importance matrix for Apple

Amazon

The adf test for AMZN did not confirm the stationarity properties of all the variables. The three FF variables were stationary, but the O/S variables proved to have p values of 0,37 for the train data set and 0,23 for the entire dataset. It seems as though more data is required for this stock, to meet the stationarity criterium.

The regression summary gives the following output.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.139218 -0.023025 -0.000565  0.024884  0.152965

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01462    0.01013   1.443    0.151
O_S         -0.22438    0.18155  -1.236    0.218
Mkt_RF       0.08881    0.11109   0.800    0.425
SMB          0.13067    0.20215   0.646    0.519
HML         -0.14130    0.11286  -1.252    0.212

Residual standard error: 0.04058 on 169 degrees of freedom
Multiple R-squared:  0.02352,   Adjusted R-squared:  0.000408
F-statistic: 1.018 on 4 and 169 DF,  p-value: 0.3998
```

Image 4 - Lm regression AMZN

We get the same negative correlation value of the O/S variables, as seen in the example with AAPL. This notion that a decrease in the O/S ratio generates positive future returns, confirms the written literature on the O/S variable.

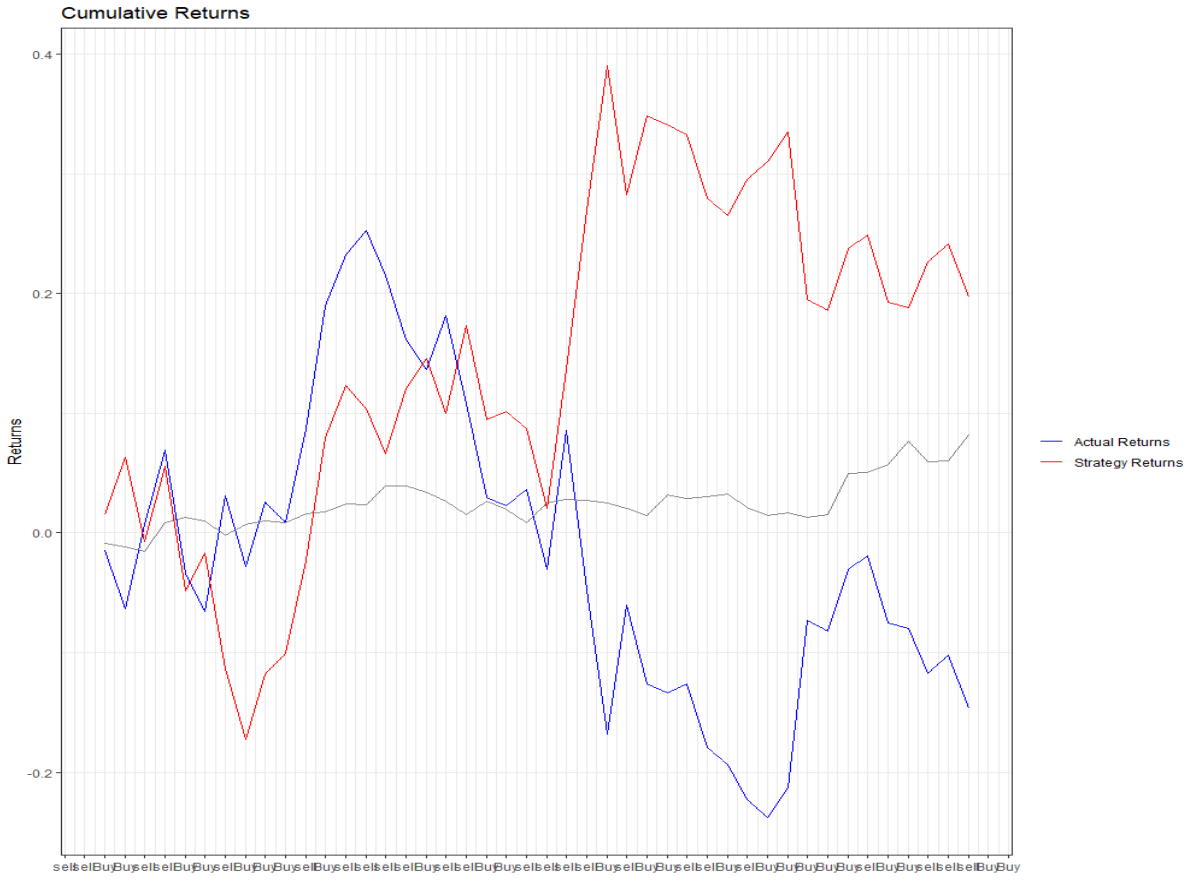
The RMSE values for AMZN peak at the 76 iterations, resulting in 0.064026, but only with a slight increase in value at the chosen iteration 200, of 0.064341, indicating slight overfitting, but nothing concerning regarding the accuracy of the model.

The return table for Amazon is listed below.

Amazon	Minimum	Mean	Maximum
Actual Return	-23,7%	-1,7%	25,2%
Predicted return	-1,5%	2,4%	8,1%
Strategy Return	-17,2%	14,2%	39%

Table 3 - Returns for Amazon

Looking at the returns for Amazon, we have a cumulative strategy mean return of 14,2% and a min/max return of negative 17,2% to 39% this is compared to the actual mean return of -1,7% and min/max of -23,7 to 25,2%. When the returns are plotted, we get a more interesting picture.



Plot 3 - Plotted returns for Amazon

The plot shows that the return for the first half of the dataset follows the actual returns relatively closely, however when entering the second half of the data range, we see that the strategy return becomes significantly higher than the actual return, but with the same amount of volatility in the spikes. The interesting part of the graph is the predicted returns, where a steady

line between the strategy returns and the actual return, shows a much steadier return scheme for Amazon stock, representing an overall neutral to slightly bullish sentiment.

The importance matrix for Amazon is illustrated below.

Plot 4 - Importance matrix for Amazon

Looking at the importance matrix plot for Amazon, we see that the O/S variable is the lowest influencer of the four features, whereas the stock has a higher relationship with the movement of the overall market, captured by the market risk-free rate. In this case, amazon cannot be said to be heavily influenced by the O/S variable, when looking at the specific data range.

Google

The adf testing for GOOG proved to be a problem, regardless of trying to incorporate the full dataset, the stationarity of the O/S variable could not be achieved. However, the results from the boosting model did not show signs of this being a particular problem.
The regression summary for Google as shown below, behave in much the same way as the other stocks, where O/S has a negative correlation with future returns, though a lack of significance is still present.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13074 -0.02241 -0.00149  0.02221  0.12920

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.009176   0.008244   1.113    0.267
O_S         -0.315894   0.537157  -0.588    0.557
Mkt_RF       0.134103   0.100586   1.333    0.184
SMB          0.127987   0.182475   0.701    0.484
HML         -0.086986   0.101969  -0.853    0.395

Residual standard error: 0.03663 on 169 degrees of freedom
Multiple R-squared:  0.02006,   Adjusted R-squared:  -0.003135
F-statistic: 0.8648 on 4 and 169 DF,  p-value: 0.4864
```

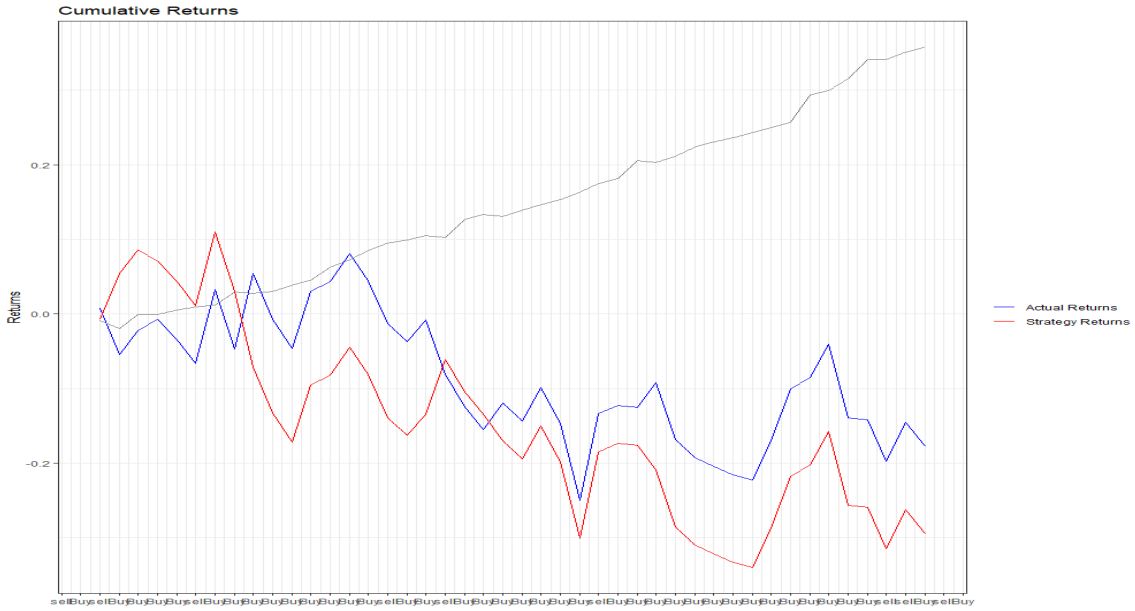
Image 5 - Regression coefficients for Google

The error term for Google was hovering around 0.054-0.055, which again is a testament to the fact that the consensus on the error terms in this project. Their overall scores are pretty low and thus could be an indication of a satisfied model, in regards to fulfilling its objective of minimizing this loss function.
The return table for Google is listed below.

	Minimum	Mean	Maximum
Actual Return	-24,9%	-9,5%	8,1%
Predicted return	-1,9%	14,7%	35,7%
Strategy Return	-33,9%	-15%	11%

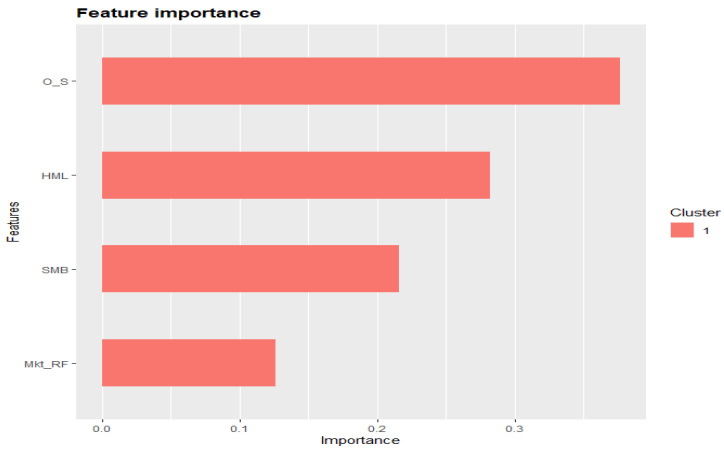
Table 4 - Google return table

For Google, we see that the trading strategy created does a decent job of mimicking the actual returns, and the strategy generates a slightly worse mean return, with both lower minimums and higher maximums. The predicted values show the same pattern, as observed from Apple stock, where a pure bullish trend is visible, indicating a very low visible correlation with the actual returns, thereby questioning the forecasting study done.



Plot 5 - Plotted returns for Google

Lastly, the importance of the factors correlates very well with the regression summary, where the O/S variable has the biggest impact on the predictive model.



Plot 6 – Importance matrix for Google

Johnson & Johnson

For Johnson & Johnson, the adf test showed every variable to have stationary properties, with a very low p-value below 0.01.
The regression summary output, as illustrated below, shows that the O/S variable has the least amount of influence, compared to the other three, additionally, it's a negative relation, which confirms the literature.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.108091 -0.015223  0.002403  0.015389  0.074476

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.005256   0.005861   0.897    0.371
O_S         -0.037917   0.096857  -0.391    0.696
Mkt_RF      -0.099036   0.075014  -1.320    0.189
SMB          0.076286   0.134857   0.566    0.572
HML          0.051134   0.074736   0.684    0.495

Residual standard error: 0.02688 on 169 degrees of freedom
Multiple R-squared:  0.01496,   Adjusted R-squared:  -0.008352
F-statistic: 0.6417 on 4 and 169 DF,  p-value: 0.6334
```

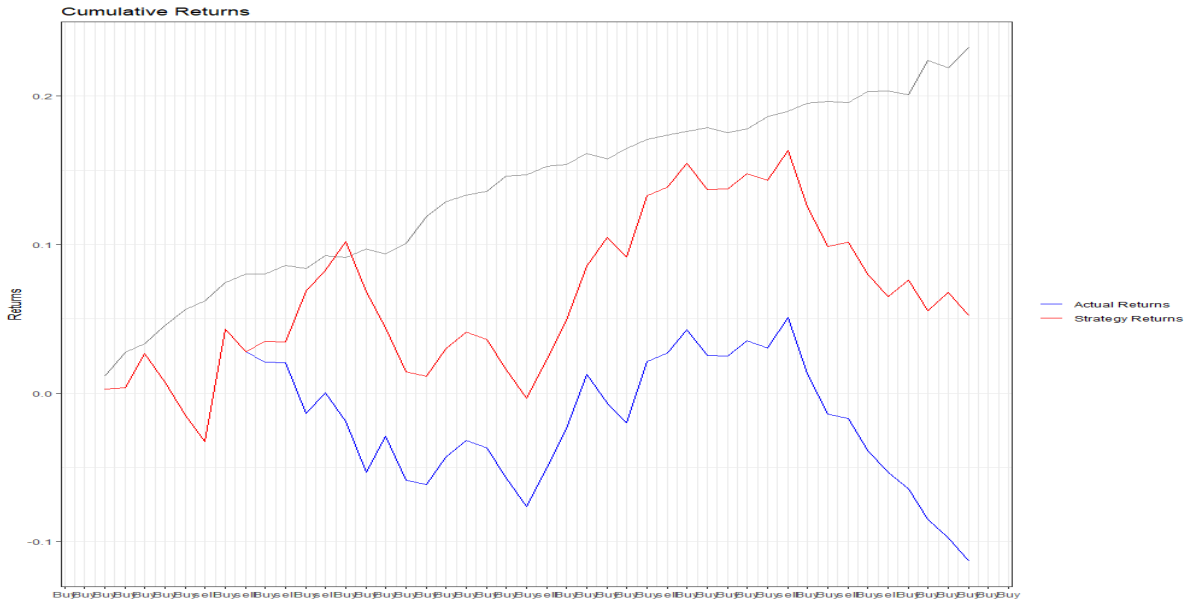
Image 6 - Regression coefficients for Johnson & Johnson

The RMSE of JNJ was measured at 0.024463, which generally speaking indicates a low number of errors for the model.
The return table for Johnson & Johnson is listed below.

Johnson & Johnson	Minimum	Mean	Maximum
Actual Return	-11,3%	-1,5%	5%
Predicted return	1,1%	13,6%	23,2%
Strategy Return	-3,2%	6,5%	16,3%

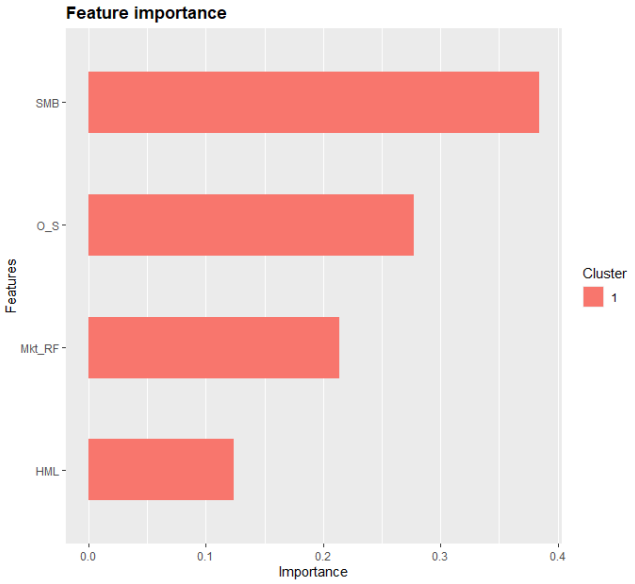
Table 5 - Return table for Johnson & Johnson

The strategy return delivers overall higher returns than the actual return for the stock. Looking at the plot, we see that the strategy return generally follows the same trends as the actual return, with a slightly higher average return. The predicted values follow the same trend as the majority of the others, with a very high return, compared to actual values.



Plot 7 – Plotted returns for Johnson & Johnson

The feature importance matrix for JNJ shows that SMB is the most important feature of the four, with O/S being second.



Plot 8 - Importance matrix for Johnson & Johnson

J.P. Morgan

The adf testing for J.P. Morgan resulted in all four variables being stationary.
The regression table summary showed the following correlations:
Negative O/S variable correlation with future returns, as expected. The Mkt-rf showed the highest correlation and the lowest p-value, which is expected for a Bank.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.209053 -0.021275  0.002254  0.023247  0.202510

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01770    0.01181   1.499   0.1357
O_S         -0.19239    0.13929  -1.381   0.1690
Mkt_RF       0.22307    0.12721   1.754   0.0813 .
SMB         -0.17076    0.22771  -0.750   0.4543
HML         -0.03808    0.12715  -0.300   0.7649
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0457 on 169 degrees of freedom
Multiple R-squared:  0.02466,   Adjusted R-squared:  0.001574
F-statistic: 1.068 on 4 and 169 DF,  p-value: 0.3739
```

Image 7 - Regression coefficients for J.P.Morgan

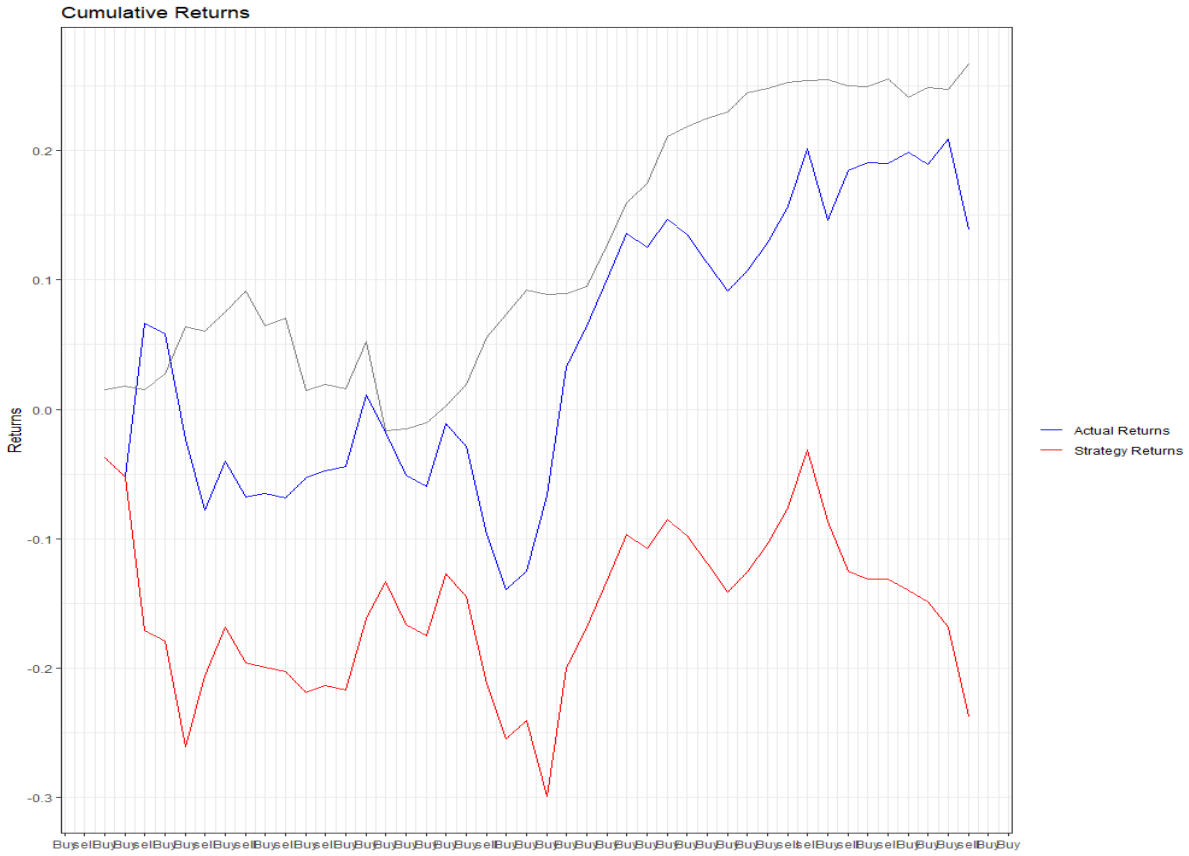
The errors of the prediction model scored 0.047150 with a gap to the best score of 0.041857 at iteration 49. This score is still very low, so the difference does not interfere with the models' predictability.

The return table for J.P. Morgan is listed below.

J.P.Morgan	Minimum	Mean	Maximum
Actual Return	-13,9%	4,4%	20,8%
Predicted return	-1,6%	12,3%	26,6%
Strategy Return	-29,9%	-15,6%	3,1%

Table 6 – Return table for J.P.Morgan

The return for the trading strategy at JPM was lower than the actual return of JPM. The mean return for the trading strategy was negative 15,6%, and min/max values of negative - 29,9% to 3,1% whereas the actual return during the period had better returns on all three points. This was the first stock where the forecasted stock returns based on the trading signals, gave a lower result compared to the actual data.
The plot for JPM shows the predicted returns, are more closely related to the actual returns, than the strategy.



Plot 9 - Plotted returns for J.P.Morgan

The differences in the factors also drastically increased with JPM. In the plot below, we see that the clusters changed colors, meaning there's a big difference between orange and blue. In this case, it can be observed that the HML factor and the mkt-rf factor had a much larger impact on the prediction, than the other two factors. This might explain the negative return for the trading strategy, or it could be the fact that J.P. Morgan's stock is a bank, in which returns might be different from most of the other stocks, which are generally located in technology and IT.

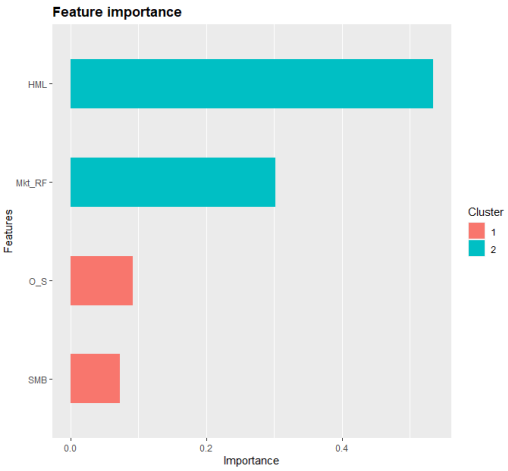


Table 10 - Importance matrix for J.P.Morgan

Mastercard

In the case of Mastercard, the adf testing done resulted in stationarity of the three variables, with close to the significance level of the O/S variable at 0.05281.
The regression summary shows an expected negative correlation with the O/S variable, with a relatively high significance score, compared to the rest of the variables.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.227162 -0.020617  0.000286  0.022388  0.166003

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.014720   0.008522   1.727    0.086 .
O_S         -0.113453   0.084710  -1.339    0.182
Mkt_RF       0.102766   0.120500   0.853    0.395
SMB         -0.077581   0.219887  -0.353    0.725
HML          0.022267   0.123809   0.180    0.857
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0441 on 169 degrees of freedom
Multiple R-squared:  0.01492,   Adjusted R-squared:  -0.008391
F-statistic: 0.6401 on 4 and 169 DF,  p-value: 0.6346
```

Image 8 - Regression coefficients for Mastercard

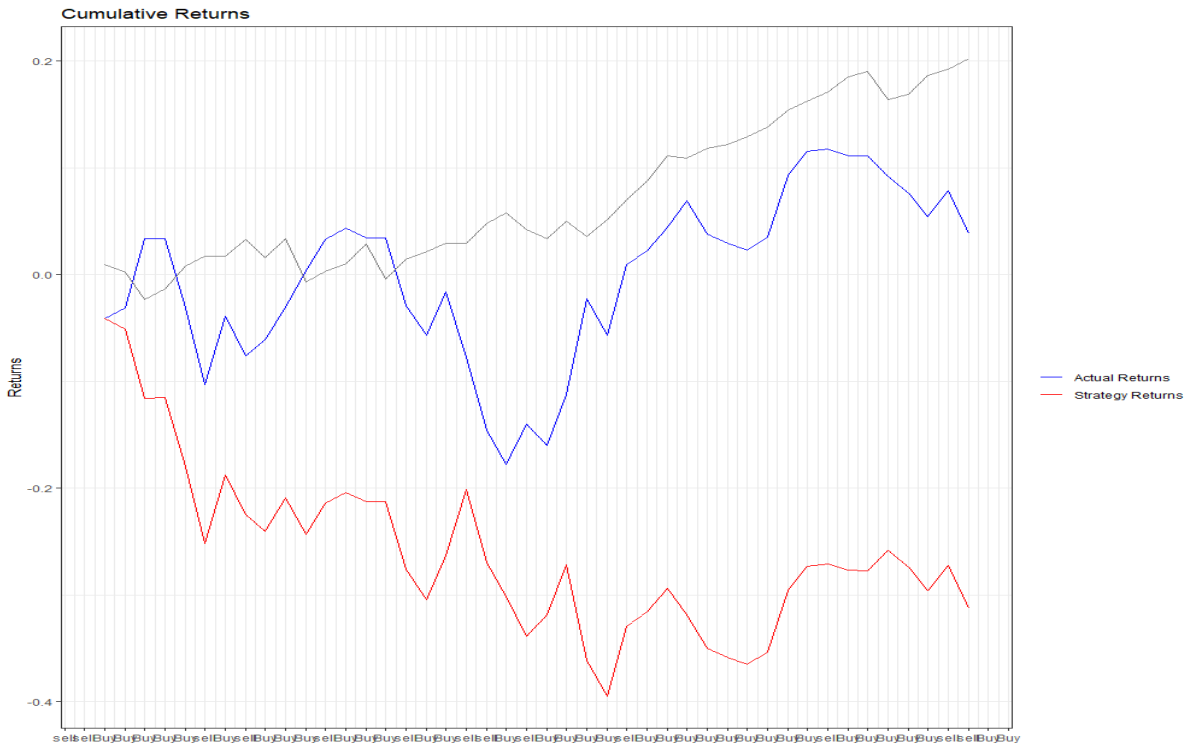
The squared error terms generated the same low score as the other stocks, at around 0.045.

The return table for Mastercard is shown below.

Mastercard	Minimum	Mean	Maximum
Actual Return	-17,7%	-0,09%	4,3%
Predicted return	-2,3%	7,2%	20,2%
Strategy Return	-39,5%	-26,1%	-4,1%

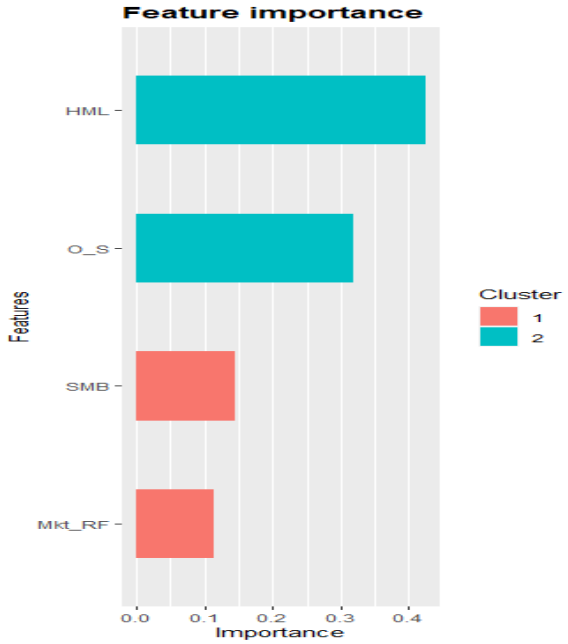
Table 7 - Return table for Mastercard

The trading strategy for Mastercard proved to be a worse strategy than the actual values, with a difference in mean value of -26%. Much like the previous JPM stock, we see that the predicted returns are much closer to the actual values than the strategy, indicating a strategy decision, that does not increase the return.



Plot 11 - Plotted returns for Mastercard

The feature importance of the variables for Mastercard grants high values to both HML and O/S. The same type of pattern is seen in J.P. Morgan, although the variables differ rants high values to both HML and O/S. The same type of pattern is seen in J.P. Morgan, although the variables differ rants high values to both HML and O/S. The same type of pattern is seen in J.P. Morgan, although the variables differ rants high values to both HML and O/S. The same type of pattern is seen in J.P. Morgan, although the variables differ r on importance.



Plot 12 - Importance matrix for Mastercard

Meta platforms

All four variables passed the stationarity testing for the META stock. with only the O/S showing signs of some degree of non-stationary trends, when only using the 4-year training period, rather than the 5-year data period.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.195006 -0.026639 -0.003491  0.026347  0.152251

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.008964   0.014622   0.613   0.5407
O_S         -0.030818   0.065114  -0.473   0.6366
Mkt_RF       0.337941   0.135312   2.497   0.0135 *
SMB          0.252868   0.243093   1.040   0.2997
HML          0.130008   0.135841   0.957   0.3399
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04877 on 169 degrees of freedom
Multiple R-squared:  0.06228,   Adjusted R-squared:  0.04009
F-statistic: 2.806 on 4 and 169 DF,  p-value: 0.02734
```

Image 9 - Regression coefficients for Meta

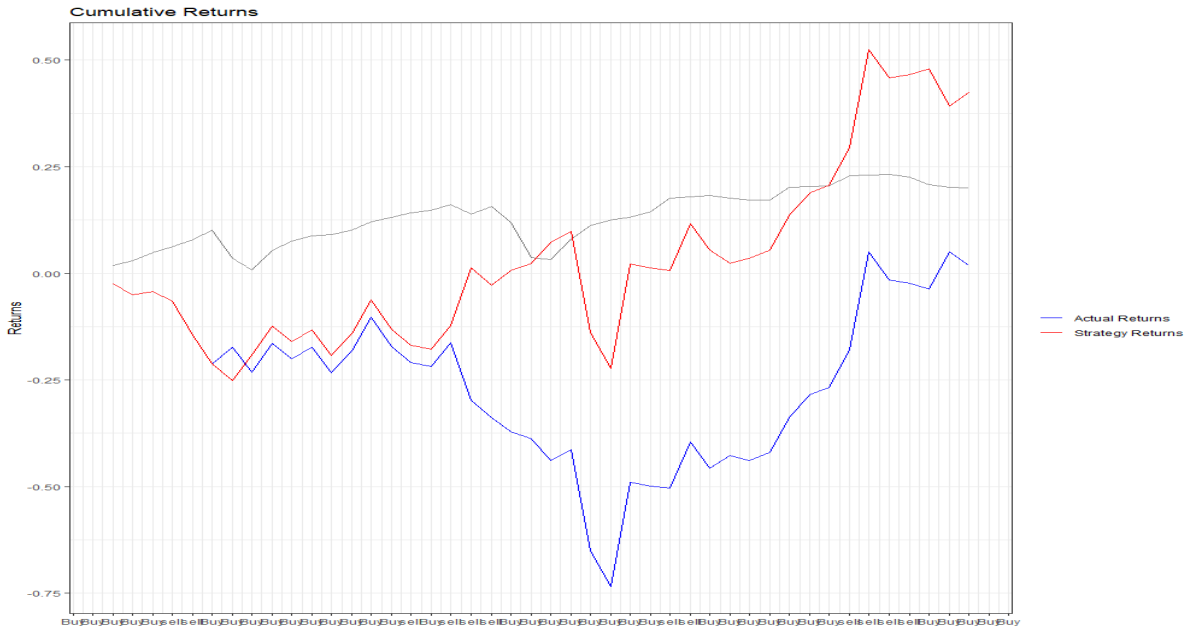
Above is the summary of the regression table, showing the expected negative correlation of the O/S variable, and also the first significant factor mkt-rf. This is an interesting observation to be noted.
The RMSE of Facebook resulted in a slightly higher error term compared to the other stocks of around 0.088, but still, this value is very low, so not a cause for concern.

The return table for Meta is listed below.

Meta	Minimum	Mean	Maximum
Actual Return	-73,4%	-25,51%	5%
Predicted return	0,8%	13,1%	23,1%
Strategy Return	-25,1%	3%	52,4%

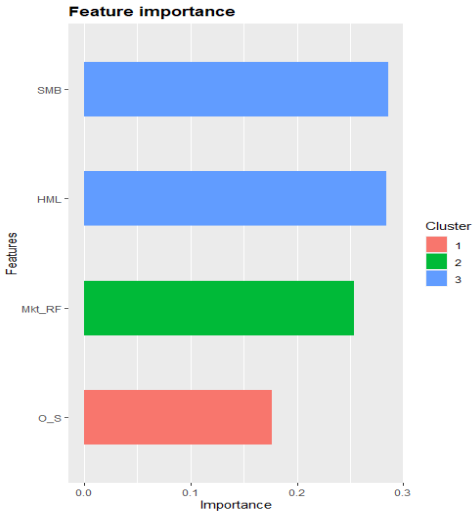
Table 8 - Return table for Meta

The return of both the strategy and the actual return of Facebook showed high volatility in the min/max values ranging from -73% to positive 52,4%. The trading strategy managed to generate a positive mean return of 3% compared to the actual mean of negative 25,51%, which is a big difference. This is not unusual as the period of the data correctly captures a volatile time in Facebooks history, where Facebook became meta, and lots of public backlash, due to negative data sharing leaks. Looking at the predicted returns, we see that it does not capture the negatively skewed volatility of Meta, compared to both the strategy and actual returns. It predicts a steady return for the testing period, which could be seen as conservative, compared to other predictions, but when adjusting for the high lows, it seems extreme.



Plot 13 - Plotted returns for Meta

The importance matrix shows that the O/S variable is the least important factor and both the SMB and HML prove to be very important in predicting future returns for Facebook. Compared with the regression coefficients, the O/S variables were the least impacting and least significant, however, this is only an observation.



Plot 14 - Importance matrix for Meta

Microsoft

The adf testing for Microsoft showed all variables to be stationary, with no deviations regarding testing the O/S variable for the train data and the full data.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.137712 -0.020598  0.000832  0.018733  0.095205

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006046   0.009150   0.661    0.510
O_S          0.006568   0.065755   0.100    0.921
Mkt_RF      -0.116059   0.093282  -1.244    0.215
SMB          0.167008   0.168984   0.988    0.324
HML          0.028605   0.094248   0.304    0.762

Residual standard error: 0.03391 on 169 degrees of freedom
Multiple R-squared:  0.01218,   Adjusted R-squared:  -0.0112
F-statistic: 0.5211 on 4 and 169 DF,  p-value: 0.7203
```

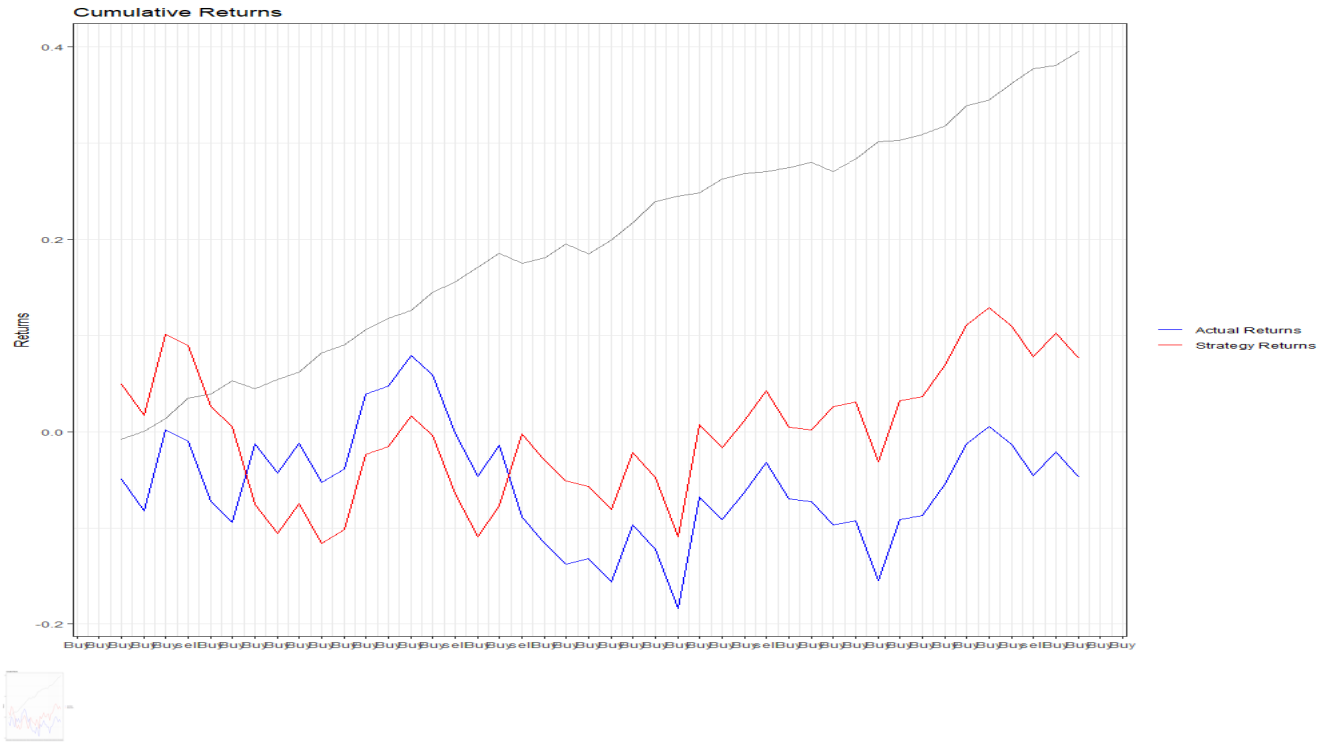
Image 10 - Regression coefficients for Microsoft

The regression analysis on MSFT showed a positive relationship with the O/S variable and future price changes, this is the first time this occurrence has happened; however, it must be noted that the relationship, has a high p-value of 0.9 and the relationship is very minuscule compared to the other stocks.
The RMSE showed the expected results, of a low score of 0,046.
The return table for Microsoft is listed below.

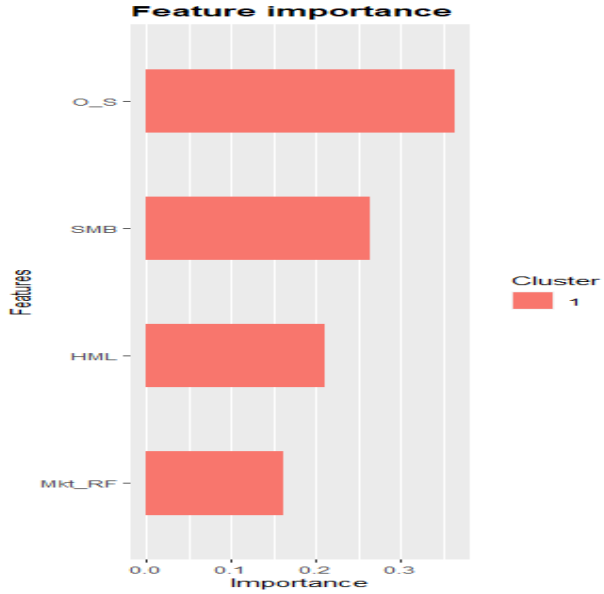
Microsoft	Minimum	Mean	Maximum
Actual Return	-18,3%	-5,5%	7,9%
Predicted return	-0,7%	19,7%	39,5%
Strategy Return	-11,5%	0,09%	12,9%

Table 9 - Return table for Microsoft

Looking at the returns, we see the expected pattern, where the trading strategy returns slightly above the actual returns, with an overall higher mean. The predictive returns trade way higher than both, with the same overshooting bullish signals, that's recurring in many of the testing samples.



The factor importance matrix shows a high dependency on the O/S variable, which is odd if one were to compare it to the regression coefficients and the significance level of the O/S variable.



Plot 16 - Importance matrix for Microsoft

Nvidia

Both the testing data and complete data resulted in nonstationary properties of the O/S variable, for the NVDA stock. The regression on Nvidia shows a negative correlation on all but the SMB variable. but with no significant variables in play.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.162611 -0.035136 -0.003477  0.039180  0.206808

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.019130   0.008982   2.130   0.0346 *
O_S         -0.081422   0.088530  -0.920   0.3590
Mkt_RF      -0.105257   0.167547  -0.628   0.5307
SMB          0.121652   0.307251   0.396   0.6927
HML         -0.038128   0.171952  -0.222   0.8248
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06158 on 169 degrees of freedom
Multiple R-squared:  0.008649,  Adjusted R-squared:  -0.01481
F-statistic: 0.3686 on 4 and 169 DF,  p-value: 0.8308
```

Image 11 - Regression coefficients for Nvidia

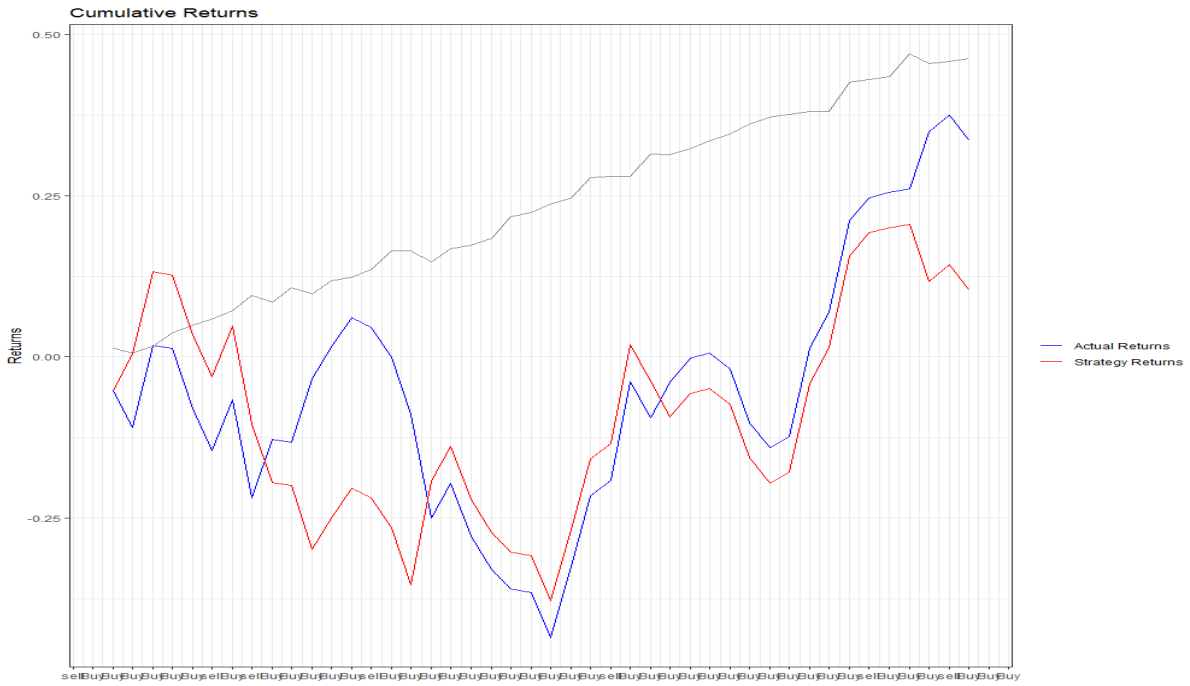
The RMSE of Nvidia was captured at 0,077.

The return table for Nvidia is listed below.

Nvidea	Minimum	Mean	Maximum
Actual Return	-43,5%	-5,1%	37,1%
Predicted return	0,6%	23,6%	46,9%
Strategy Return	-37,8%	-8,9%	20,5%

Table 10 - Return table for Nvidea

For Nvidia, we see the same pattern as observed for J.P. Morgan, where the actual return shows a higher return than the strategy. However, the predicted return Is not similar, instead, we observe the expected overshooting of the model, with very high returns, not able to catch much downside movement.



Plot 17 - Plotted returns for Nvidia

The feature importance of the variables indicates that the SMB is the highest importance factor, with O/S being a close second. Comparing this to the J.P.Morgan stock, there's no similarity between the negative output and the factor importance

Plot 18 - Importance matrix for Nvidia

Procter And Gamble

The Adf testing on PG stock shows no indication of nonstationary properties, regarding either the training set and the complete dataset.
The regression analysis below proves that the O/S relationship is negatively correlated with future price changes, which reiterates the consensus in the written literature. It also indicates that the risk premium captured by the mkt-rf variable is close to being a statistically significant factor.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.108319 -0.011648  0.000906  0.013222  0.079576

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.004788   0.004177   1.146    0.253
O_S         -0.013215   0.108435  -0.122    0.903
Mkt_RF      -0.126177   0.071871  -1.756    0.081 .
SMB         -0.047591   0.130523  -0.365    0.716
HML          0.015651   0.072823   0.215    0.830
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0262 on 169 degrees of freedom
Multiple R-squared:  0.02278,   Adjusted R-squared:  -0.0003485
F-statistic: 0.9849 on 4 and 169 DF,  p-value: 0.4173
```

Image 12 - Regression coefficients for Procter & Gamble

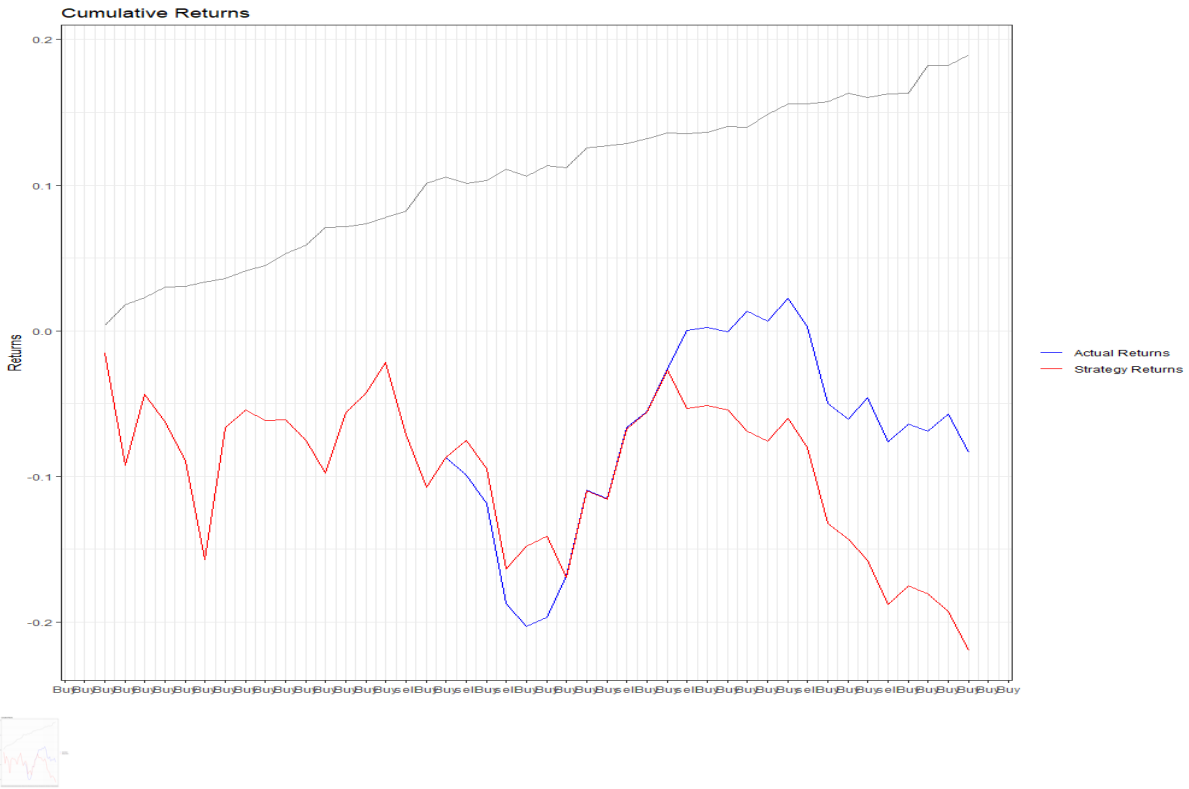
The RMSE for Procter & Gamble showed a low RMSE value of 0,035.

The return table for Procter & Gamble is listed below.

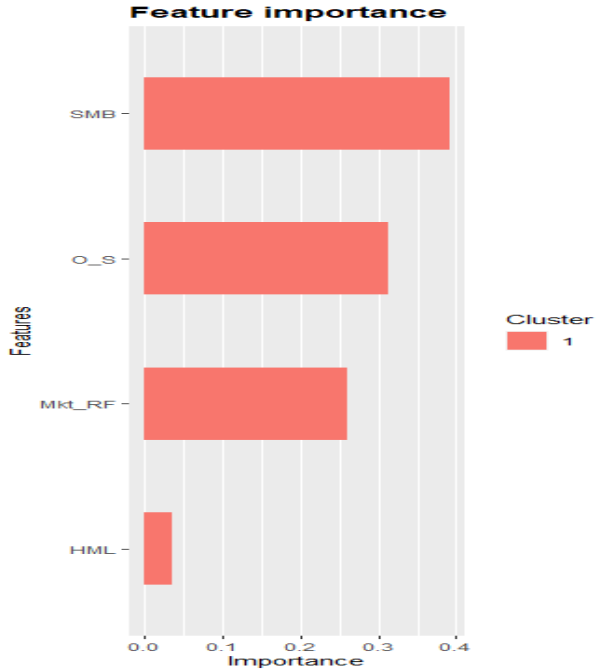
Procter & Gamble	Minimum	Mean	Maximum
Actual Return	-20,3%	-5,6%	2,2%
Predicted return	0,3%	10,5%	18,9%
Strategy Return	-21,9%	-9,6%	-1,5%

Table 11 - Return table for Procter & Gamble

The return for P&G can be compared to the latest stock Nvidea, where strategy return is lower than actual return, and predictive returns are overshooting by a lot, and not capturing significant downside movement. However, an important observation can be seen in the plotted returns for P&G, where the strategy return and actual return are equal to each other for the first half of the testing data, this is not observed on any of the other stocks.



The feature importance of the variables concludes that SMB shows the highest importance, followed by O/S



Plot 20 - Importance matrix for Procter & Gamble

Tesla

The analysis for Tesla shows that the O/S variable does not contain stationary properties, despite testing for the whole dataset. The four-factor regression results are shown below, '

where the O/S contributing factor is negatively correlated with future price changes, which is the main theme of this thesis. None of the variables are statistically significant.

```
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train
_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.289171 -0.062200 -0.004617  0.046903  0.290174

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02695    0.01164   2.315   0.0218 *
O_S         -0.09433    0.09638  -0.979   0.3291
Mkt_RF       0.36364    0.25818   1.408   0.1608
SMB         -0.06345    0.47409  -0.134   0.8937
HML         -0.17786    0.26569  -0.669   0.5041
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09489 on 169 degrees of freedom
Multiple R-squared:  0.02031,   Adjusted R-squared:  -0.00287
3
F-statistic: 0.8761 on 4 and 169 DF,  p-value: 0.4796
```

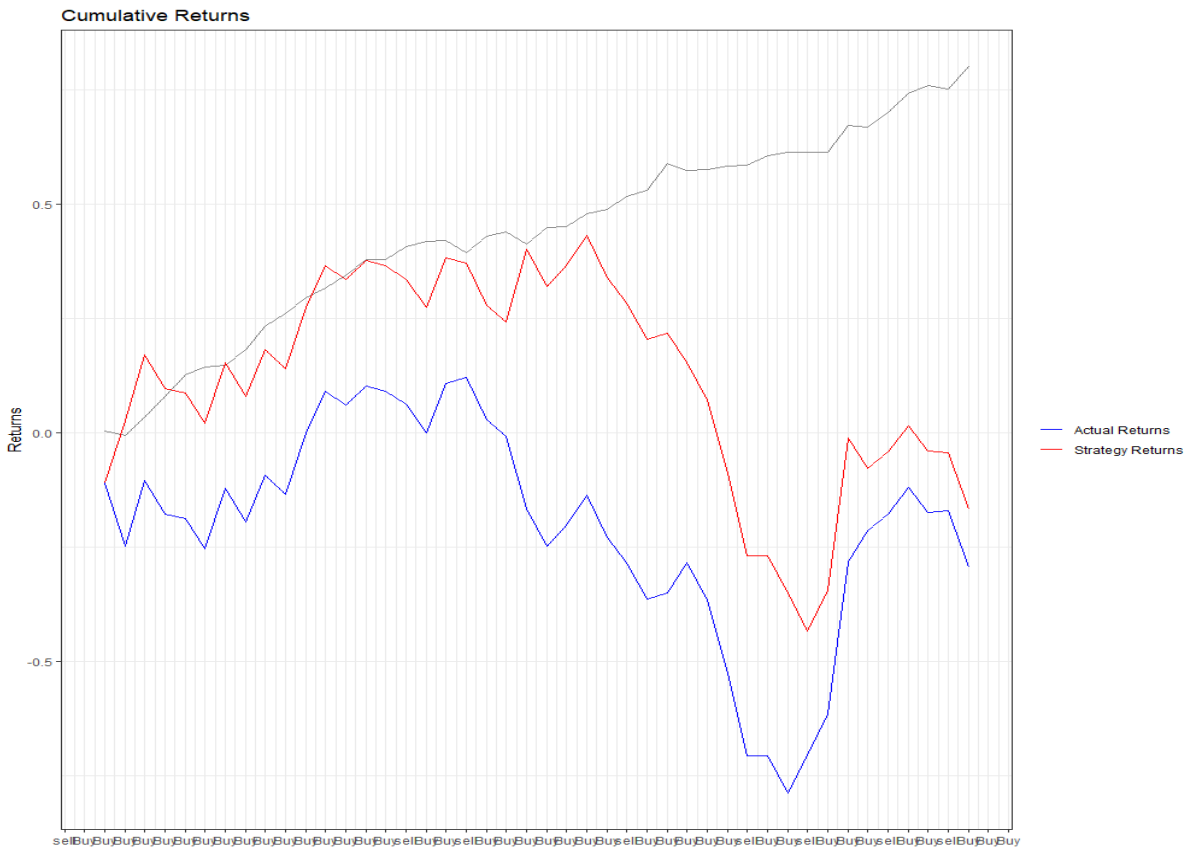
Image 13 – Regression coefficients for Tesla

The RMSE of Tesla is the highest of all the stocks, with a score of 0,099.
The return table for Tesla is listed below.

Tesla	Minimum	Mean	Maximum
Actual Return	-78,8%	-20,6%	12%
Predicted return	-0,5%	43,6%	80,1%
Strategy Return	-43,4%	11,5%	43,1%

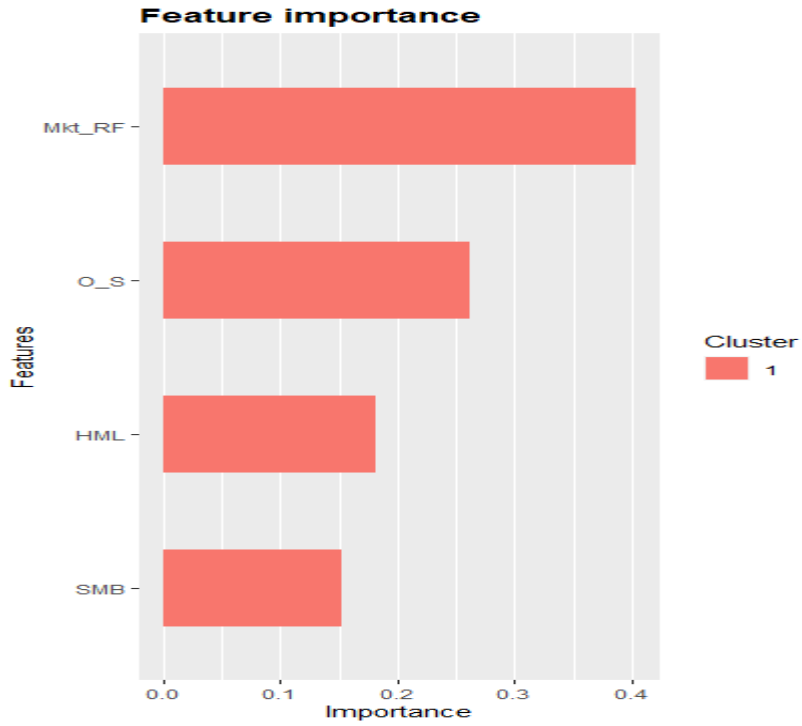
Table 12 - Return table for Tesla

The general return for Tesla over the period is very volatile, thus the trading strategy is expected to generate similar volatile results, and this holds, with a mean return 32% higher than the actual mean. Looking at the plotted returns, we see that for the first half of the testing set, predicted returns and strategy returns travel along the same path, but in the second half, the prediction soars even further while the trading strategy diminishes.



Plot 21 - Plotted returns for Tesla

The feature importance plot shows that the market beta is of the highest importance for generating the predicted returns.



Plot 22 - Importance matrix for Tesla

Visa

The four variables for Visa were all stationary, with statistical significance.
The regression results as listed below, once again highlight the negative correlation between option to stock volume ratio and future price changes, as observed by the majority of the stocks.

```
Call:
lm(formula = Returns ~ O_S + Mkt_RF + SMB + HML, data = train
_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.173470 -0.017436 -0.000198  0.018445  0.121640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.012112   0.008027   1.509    0.133
O_S         -0.126590   0.101697  -1.245    0.215
Mkt_RF       0.124391   0.104120   1.195    0.234
SMB         -0.105754   0.190337  -0.556    0.579
HML          0.012235   0.106172   0.115    0.908

Residual standard error: 0.03819 on 169 degrees of freedom
Multiple R-squared:  0.01706,   Adjusted R-squared:  -0.00620
3
F-statistic: 0.7334 on 4 and 169 DF,  p-value: 0.5704
```

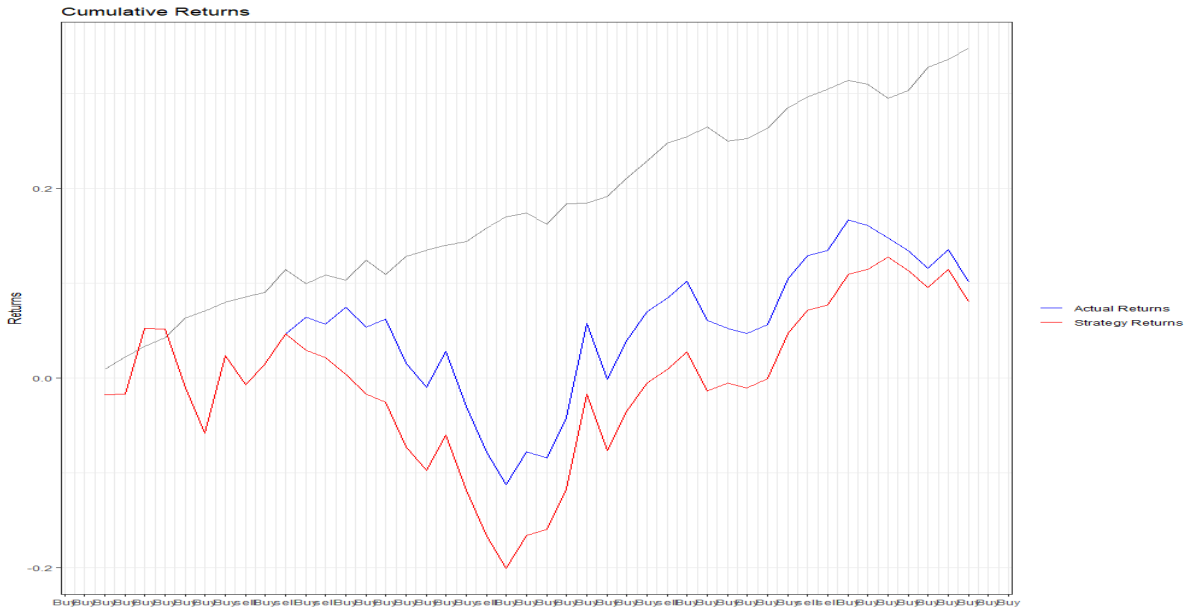
Image 14 - Regression Coefficients for Visa

The RMSE value for Visa's iteration is 0,038.
The return table for Visa is listed below

Visa	Minimum	Mean	Maximum
Actual Return	-11,2%	0,43%	16,6%
Predicted return	0,9%	18,2%	34,7%
Strategy Return	-20%	-0,5%	12,7%

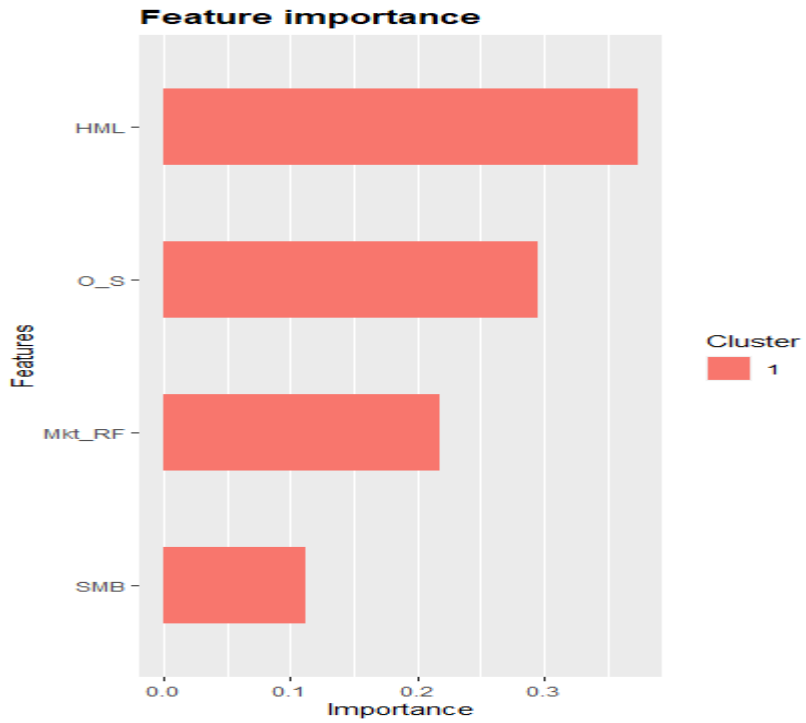
Table 13 - Return table for Visa

Visa stands out, because of the fact the strategy return is lower than the actual return of the stock, whereas the predicted values are better than both the strategy and the actual returns, showcasing some inference. Looking at the plot, we see that the trading strategy follows the actual returns, for the first couple of months, but then diminishes, and trades lower than the actual return for the remainder of the period.



Plot 23 - Plotted returns for Visa

The factor importance shows that the HML factor and the O/S factor are at the top of the features.



Plot 24 - Importance Matrix for Visa

6.2 Findings Summary

Stationarity testing

The summary for adf testing done on all 12 stocks indicates that the O/S variable has some trouble with being a stationary variable when testing on the four and five-year period, with 50 percent of the O/S variables being non-stationary. Therefore, it is either a problem of the quantity of data or simply the variable itself being problematic. The summary of the testing can be seen below in Table 14, where X indicates the failure to pass the augmented dickey fuller test.

AAPL	AMZN	GOOG	JNJ	JPM	MA	META	MSFT	NVDA	PG	TSLA	V
	x	x			x	x		x		x	

Table 14 - Summary of augmented Dickey-fuller test, to test for stationarity.

Regression Correlations

The summary of the correlation between the O/S variable and future price changes is illustrated in Table 15 below. The correlation summary clearly shows that the correlation between the option-to-stock-volume variable is negatively correlated with future price changes, with only one of the 12 stocks indicating otherwise. These results reiterate and confirm the written literature examined in the review, and thus for the data used in this project, it's concluded that O/S negatively impacts the future prices of the 12 stocks. However, for a conclusion to be accepted in an academic paper, the significance of these correlations must be at a certain level, which is not achieved in the project, therefore any real remarks about the O/S variable cannot be stated, based on this analysis.

AAPL	AMZN	GOOG	JNJ	JPM	MA	META	MSFT	NVDA	PG	TSLA	V
-0,076	-0,22	-031	-0,037	-0,19	-0,11	-0,03	0.006	-0,08	-0,013	-0,09	-0,12

Table 15 -based regression correlations with the O/S variable.

Returns & Errors

Below is a table showing the actual mean return of the stock (AR), the mean strategy return (SR), and the root mean squared error term (E). Green color indicating a trading strategy that outperforms actual returns.

	AAPL	AMZN	GOOG	JNJ	JPM	MA	META	MSFT	NVDA	PG	TSLA	V
AR	-3,7	-1,7	-9,5	-1,5	4,4	-0,09	-25	-5,5	-5,1	-5,6	-20,6	0,43
SR	8,1	14,2	-15	6,5	-15,6	-26,1	3	0,09	-8,9	-9,6	11,5	-0,5

E	0,048	0,064	0,055	0,024	0,047	0,045	0,088	0,046	0,077	0,035	0,099	0,038

Table 16 - Summary of returns and errors

Comparing the results of the returns, we see that for half of the stocks, the trading provides a better mean return than the actual return, and for the other half, the opposite is observed. Looking further into the spread of positive and negative strategy returns, based on their representative stocks, we don't see a clear indication of why certain stocks might favor this trading strategy.

Factor importance

Below is the summary of the factor important for the O/S variable. The numbers indicate in what position of importance the variable was placed, where 1 is most important and 4 is least important. We see that in general, the O/S variable placed roughly around the top 2, indicating that it is indeed a factor with some degree of information, this is compared to the three well-known factors, widely used in academic papers. However, while this might be an indication, the results are not significant.

AAPL	AMZN	GOOG	JNJ	JPM	MA	META	MSFT	NVDA	PG	TSLA	V
1	4	1	2	3	2	4	1	2	2	2	2

Table 17 - Summary of important factors, ranking O/S.

Section 7 – Discussion & Limitations

The discussion of the thesis consists of two sections. First, we discuss the results generated and how they may be interpreted, secondly, we explore the limitations of the project and ideas for future research.

7.1 Discussion

Looking at the results covered in section 6.2, some questions arise regarding the validity of the data generated and the data used. This starts with the ADF test, where 6 out of 12 of the stocks' O/S variable does not qualify as having stationary properties, which translates to a higher chance of resulting in faulty readings. Attempts to increase the percentage of stocks with stationary variables, by testing for the entire dataset, resulted in better readings, however not satisfactory readings. This could indicate that perhaps a 10-year period of data would eliminate the non-stationarity of the O/S variable. The reason this is not incorporated into the thesis is due to the authors' goal of trying to create a short-term trading strategy, which means a shorter time frame to capture the market trend, is essential for getting the right results.
The regression coefficients for the stocks were incorporated in the project, to shed some light on how much each variable affected the price and in which direction. If one were to extract the correlations in a vacuum, without the added significance levels, then it becomes clear that on a general level, the O/S variable negatively impacts future price changes. The only stock to generate positive price influence is Microsoft, however only with a very low coefficient. This

negative price influence is also explored and tested in the literature, proving that high options volume compared to stock volume, is indicative of negative future price changes.
The summarized return table 16, records a 50% chance for the trading strategy to generate better returns, compared to the actual returns, based on the XGboosting predictions. Additionally, there's no real indicative evidence of why certain stocks perform better or worse with the incorporated trading strategy, as the linkage between performance, coefficients, stationarity, and importance matrix, does not show any sign of connection. The Importance matrix shows that the O/S variable is on average the top two of the four factors, this would indicate that the information in the variable is important for the creation of the prediction model. However, there's no significant correlation between the high importance of the variable and a good predictive model. Additionally, there's no connection between the O/S variable in the importance matrix, and the regression coefficients, which does raise some concerns regarding the validity of the XGboosting model, used in this project.
Regarding the significance levels of not only the O/S variable, but all the variables, an acceptable level of significance was hard to achieve throughout the project. This means that the data used did not satisfy the criterium for academic testing to generate substantial proof of evidence for the thesis.

7.2 Limitations & Future Research

Due to the limited time, the thesis does not contain a comparison of different trading strategy approaches, however, this is something that could be implemented for further research.
An idea of this would be to incorporate different trading strategies, that trade based on the value of the O/S itself.
An interesting addition could be to test the postulate, from the literature, that the weekly data has the most predictive power, compared to daily and monthly, and see if this idea holds in today's market situation.
A limitation of the project concerns the amount of data collected, initially, the idea was to test more than one variable, however much time was spent and wasted trying to manipulate the lack of data into a model, which ultimately led to the decision of going with only the O/S variable. Ideally, 40-50 stocks could be chosen, each varying in size and industry, to create more robust testing of the trading strategy. Moreover, future research should incorporate a bigger timeframe, thus potentially eliminating the possible problems arising from non-stationary variables.
Additionally, a comparison between the O/S variable and the C/P or P/C ratio, as described in Section 2, might result in some interesting findings.

Lastly, it is important to note that the project required a more advanced approach to programming than the author anticipated, and thus the lack of experience in this field might be reflected in the results.

Section 8 - Conclusion

The conclusion seeks to answer the research question:

1. *How does the option-to-stock volume ratio affect stock returns?*
2. *To what extent can this information be utilized to make informed investment decisions that could potentially yield financial gains?*

The literature on the option-to-volume ratio (O/S), generally concludes that the ratio contains information that is negatively correlated with future price changes. This is showcased by low O/S ratios resulting in higher returns and vice versa. The analysis shows signs of this negative correlation, with 11/12 stocks indicating negative coefficients, however, none of the readings were significant enough to meet the criterium, and thus a supportive conclusion cannot be drawn, based on this dataset.
The trading strategy developed by the XGboosting algorithm created improved mean return for 50% of the individual stocks and inferior results for the other half. This 50/50 split makes it hard to draw any real conclusion, as this result could easily be interpreted as random. Combining the returns with the importance matrix, the author fails to observe any substantial connection between positive returns, their respective correlations, and the importance of the variables. The final verdict is that the O/S variable seems to contain negative information, however not statistically significant information. The thesis does not support the applicability of the trading strategy, and thus the results should be treated purely speculative.

References

Blau, B. M., Nguyen, N., & Whitby, R. J. (2014). The information content of option ratios [Article]. *Journal of Banking & Finance*, *43*(1), 179–187. https://doi.org/10.1016/j.jbankfin.2014.03.023

Hubert Buch-Hansen & Peter Nielsen (2012). Kritisk Realisme.

Cao, C., Chen, Z., & Griffin, J. M. (2005). Informational Content of Option Volume Prior to Takeovers [Article]. *The Journal of Business (Chicago, Ill.)*, *78*(3), 1073–1109. https://doi.org/10.1086/429654

Cao, M., & Wei, J. (2010). Option market liquidity: Commonality and other characteristics [Article]. *Journal of Financial Markets (Amsterdam, Netherlands)*, *13*(1), 20–48. https://doi.org/10.1016/j.finmar.2009.09.004

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

CNBC. (2020). *Pandemic-induced options trading craze shows no signs of slowing down*. https://www.cnbc.com/2020/12/04/pandemic-induced-options-trading-craze-shows-no-signs-of-slowing-down.html

Eugene Fama & Kenneth French. (1992). *the_cross-section_of_expected_stock_returns*.

Hao, X., Lee, E., & Piqueira, N. (2013). Short sales and put options: Where is the bad news first traded? [Article]. *Journal of Financial Markets (Amsterdam, Netherlands)*, *16*(2), 308–330. https://doi.org/10.1016/j.finmar.2012.09.005

Houlihan, P., & Creamer, G. G. (2019). Leveraging a call-put ratio as a trading signal. *Quantitative Finance*, *19*(5), 763–777. https://doi.org/10.1080/14697688.2018.1538563

Johnson, T. L., & So, E. C. (2012). The option to stock volume ratio and future returns [Article]. *Journal of Financial Economics*, *106*(2), 262–286. https://doi.org/10.1016/j.jfineco.2012.05.008

Kenneth French. (n.d.). *Kenneth French*. Retrieved May 29, 2023, from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Kim, J. S., Kim, D.-H., & Seo, S. W. (2017). Investor Sentiment and Return Predictability of the Option to Stock Volume Ratio [Article]. *Financial Management*, *46*(3), 767–796. https://doi.org/10.1111/fima.12155

Malik, S., Harode, R., & Singh, A. (2020). *XGBoost: A Deep Dive into Boosting (Introduction Documentation) Phishing Detection in E-mails using Machine Learning View project Automatic Railway Barrier System, Railway Tracking and Collision Avoidance using IOT View project*. https://doi.org/10.13140/RG.2.2.15243.64803

Pan, J., & Poteshman, A. M. (2006). The Information in Option Volume for Future Stock Prices [Article]. *The Review of Financial Studies*, *19*(3), 871–908. https://doi.org/10.1093/rfs/hhj024

Roll, R., Schwartz, E., & Subrahmanyam, A. (2010). O/S: The relative trading activity in options and stock [Article]. *Journal of Financial Economics*, *96*(1), 1–17. https://doi.org/10.1016/j.jfineco.2009.11.004

Woo, M., & Kim, M. A. (2021). *Option volume and stock returns: evidence from single stock options on the Korea Exchange*. https://doi.org/10.1108/JDQS-06-2021-0012

Finance.Yahoo.com
https://finance.yahoo.com/

CBOE.com
https://www.cboe.com/us/options/market_statistics/historical_data/

Appendix

Appendix A

Example of setup, using Apple.

Date	Returns	O_S	Mkt_RF	SMB	HML
31-12-2018	0,02718186	0,059117982	0,0314	0,0112	-0,0138
07-01-2019	0,029745904	0,062221504	0,0214	0,0128	0,0125
14-01-2019	0,005994146	0,059489542	0,0295	0,0221	-0,0157
21-01-2019	0,05552737	0,054411448	0,028	-0,0084	0,0089
28-01-2019	0,023360471	0,072732984	-0,0028	0,001	-0,0016
04-02-2019	0,00434782	0,085030786	0,0157	-0,0036	-0,0076
11-02-2019	0,014962926	0,055362845	0,0009	0,003	-0,0126
18-02-2019	0,011562945	0,073942371	0,0273	0,0158	0
25-02-2019	-0,011773438	0,051861713	0,0064	0,0099	-0,0084
04-03-2019	0,076398086	0,059605129	0,005	-0,0041	-0,0129
11-03-2019	0,026488114	0,085750331	-0,0255	-0,0196	-0,0026
18-03-2019	-0,005757584	0,095306765	0,0283	-0,009	-0,0047
25-03-2019	0,037115127	0,073457625	-0,0111	-0,0145	-0,0273
01-04-2019	0,009492259	0,076386448	0,0127	0,0106	-0,0021
08-04-2019	0,025091702	0,082064016	0,0214	0,0054	0,01
15-04-2019	0,002158544	0,066162717	0,0055	-0,0071	0,0091
22-04-2019	0,03646583	0,077568788	-0,0018	-0,0105	0,0048
29-04-2019	-0,068807501	0,086896961	0,0124	0,002	-0,0125
06-05-2019	-0,037793692	0,071295309	0,0027	0,0065	0,0036
13-05-2019	-0,053068711	0,072873903	-0,0222	-0,0013	-0,0003
20-05-2019	-0,021791394	0,069036874	-0,0102	-0,0149	-0,0068
27-05-2019	0,086136813	0,066342126	-0,013	-0,0053	-0,0018
03-06-2019	0,013620933	0,079783891	-0,0271	-0,0049	-0,0105
10-06-2019	0,031337583	0,079705185	0,0427	-0,0125	-0,0054

Appendix B
Excel data file containing the return for SPY Etf and the corresponding O/S Ratio

Date	Adj Close	Volume Stock	Volume Option	O/S ratio
01-01-2019	251.578.995	2048691700	10622844	0,519
01-02-2019	259.734.131	1371716300	8507901	0,620
01-03-2019	263.275.848	1678081300	9715789	0,579
01-04-2019	275.238.373	1209204700	7423648	0,614
01-05-2019	257.686.035	1845593200	11753722	0,637
01-06-2019	274.283.478	1340435600	8098914	0,604
01-07-2019	279.784.698	1110102300	7618987	0,686
01-08-2019	275.100.159	2034004800	11071409	0,544
01-09-2019	279.163.818	1303830000	7697524	0,590
01-10-2019	286.652.313	1386748300	9005790	0,649
01-11-2019	297.028.656	1037123500	7447951	0,718
01-12-2019	304.163.483	1285175800	8143043	0,634
01-01-2020	305.535.553	1392003800	8757432	0,629
01-02-2020	281.347.565	2110214900	10353341	0,491
01-03-2020	244.775.986	5926017600	17155251	0,289
01-04-2020	277.480.652	2819312300	14345620	0,509
01-05-2020	290.701.324	1910460500	12017539	0,629
01-06-2020	294.560.455	2358674500	13427026	0,569
01-07-2020	313.280.334	1505145300	8296742	0,551
01-08-2020	335.146.271	1045563300	6357755	0,608
01-09-2020	321.311.035	1814712700	10042525	0,553
01-10-2020	314.553.680	1629016100	9446631	0,580
01-11-2020	348.769.806	1535244300	11850669	0,772
01-12-2020	360.155.975	1344541500	11083691	0,824
01-01-2021	358.005.371	1402265400	12950446	0,924
01-02-2021	367.959.930	1307806200	11881637	0,909
01-03-2021	383.409.302	2401715800	13916806	0,579
01-04-2021	405.017.822	1462106600	10805648	0,739
01-05-2021	407.677.216	1547235900	10873604	0,703
01-06-2021	415.461.151	1282152400	9503417	0,741
01-07-2021	426.996.033	1422104700	12241980	0,861
01-08-2021	439.703.339	1254001400	11486475	0,916
01-09-2021	417.872.040	1745559600	14887676	0,853
01-10-2021	448.624.054	1508665200	15107540	1,001
01-11-2021	445.019.440	1335351500	13547372	1,015
01-12-2021	463.970.581	1927433900	16172930	0,839
01-01-2022	441.044.281	2485167800	16871215	0,679
01-02-2022	428.025.909	2297975100	16415086	0,714
01-03-2022	442.740.173	2380929500	16705565	0,702
01-04-2022	405.135.986	1856757400	15293043	0,824

Date	Adj Close	Volume Stock	Volume Option	O/S ratio
01-05-2022	406.050.476	2418478100	17605198	0,728
01-06-2022	370.964.935	1958611900	16439911	0,839
01-07-2022	406.876.160	1437748400	15299959	1,064
01-08-2022	390.274.841	1443394400	17961655	1,244
01-09-2022	352.746.490	1998908600	22109387	1,106
01-10-2022	382.982.941	2024732000	21561926	1,065
01-11-2022	404.273.560	1745985300	19727630	1,130
01-12-2022	379.234.528	1735973600	21774285	1,254
01-01-2023	404.934.570	1575450100	22328834	1,417
01-02-2023	394.753.418	1603094700	20928915	1,306
01-03-2023	407.833.527	2515907800	24585590	0,977

A. Appendix C

The R script framework, that is used for all 12 stocks

```
# Loading standard libraries, for the use of the the R script.

library(xgboost)
library(readxl)
library(tidyr)
library(dplyr)
library(readr)
library(purrr)
library(forcats)
library(stringr)
library(lmtest)
library(urca)
library(xts)
library(ggplot2)
library(tseries)
library(caret)
library(Ckmeans.1d.dp)
library(writexl)

# Reading data for stock
data <- read_excel("C:/Users/kristian/Desktop/Datafile.xlsx",
                   sheet = "AAPL")

# Define variables for XGBoost model
x_vars <- c("Mkt_RF", "SMB", "HML", "O_S")
y_var <- "Returns"


# Splitting the data into training and testing sets
train_size <- floor(0.8 * nrow(data))
train_data <- data[1:train_size,]
test_data <- data[(train_size+1):nrow(data),]

# Perform a regression analysis using all variables as predictors to test the significance
regression_model <- lm(Returns ~ O_S + Mkt_RF + SMB + HML, data = train_data)
summary(regression_model)

# Run ADF test on variables
adf.test(train_data$Returns)
adf.test(train_data$O_S)
adf.test(data$O_S)
adf.test(train_data$Mkt_RF)
adf.test(train_data$SMB)
adf.test(train_data$HML)
```

```
# Train XGBoost model on training data
dtrain <- xgb.DMatrix(as.matrix(train_data[,x_vars]), label = train_data$Returns)
dtest <- xgb.DMatrix(as.matrix(test_data[,x_vars]), label = test_data$Returns)
params <- list(booster = "gbtree", objective = "reg:squarederror", max_depth = 1, eta = 0.1,
nthread = 3)
xgb_model <- xgb.train(params, dtrain, nrounds = 200, watchlist = list(train=dtrain, test=dtest),
print_every_n = 1)

print(xgb_model$evaluation_log)
importance_matrix <- xgb.importance(x_vars, model = xgb_model)
xgb.ggplot.importance(importance_matrix)


#  Use model to predict stock returns for test data
dtest <- xgb.DMatrix(as.matrix(test_data[,x_vars]))
test_data$Predicted_Returns <- predict(xgb_model, dtest)
print(test_data$Predicted_Returns)

# Calculate trading signal based on predicted returns
test_data$Trading_Signal <- if_else(test_data$Predicted_Returns > 0, "Buy", "sell")
print(test_data$Trading_Signal)

# Calculate strategy returns based on trading signal and actual returns
test_data$Strategy_Returns <- if_else(test_data$Trading_Signal == "Buy",
(test_data[,y_var]), -(test_data[,y_var]))
test_data$Cumulative_Strategy_Returns <- cumsum(test_data$Strategy_Returns)


Actual_returns <- test_data$Returns
cumulative_actual_returns <- cumsum(Actual_returns)
summary(cumulative_actual_returns)
print(test_data$Cumulative_Strategy_Returns)
summary(test_data$Cumulative_Strategy_Returns)
summary(cumulative_actual_returns)
cumulative_predicted_returns <- cumsum(test_data$Predicted_Returns)
summary(cumulative_predicted_returns)
test_data$cumulative_actual_returns <- cumulative_actual_returns


# Model validation using RMSE
y_test <- test_data[,y_var]
y_pred <- test_data$Predicted_Returns
mse <- mean((y_test$Returns - y_pred)^2)
rmse <- sqrt(mse)
cat("RMSE:", rmse)
```

```
test_data$Cumulative_predicted_returns <- cumsum(test_data$Predicted_Returns)

View(test_data)

# Compare trading strategy to actual returns
Df <- data.frame(Date = test_data$Date, cumulative_actual_returns, Strategy_returns = test_data$Cumulative_Strategy_Returns, Predicted_returns = test_data$Cumulative_predicted_returns)
View(Df)

ggplot(Df, aes(x = Date)) +
  geom_line(aes(y = cumulative_actual_returns, col = "Actual Returns")) +
  geom_line(aes(y = Predicted_returns, col = "Predicted Returns")) +
  geom_line(aes(y = Returns, col = "Strategy Returns")) +
  scale_x_datetime(date_labels = test_data$Trading_Signal, date_breaks = "1 week") +
  scale_color_manual(name = "", values = c("Actual Returns" = "blue", "Strategy Returns" = "red")) +
  labs(title = "Cumulative Returns", y = "Returns", x = "") +
  theme_bw()
```