UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Master Thesis

Prediction of Stock Returns Using Accounting Data with a Machine Learning Approach

GM1460 Research in Accounting and Financial Management

Authors: Ludvig Ekmark, Tobias Frisell

Supervisor: Jan Marton

2022-06-07

Abstract

The relationship between accounting data and stock price prediction has been a hot topic for over half a century. Researchers have been trying to identify the relationship and investigate how it may be useful when trying to improve prediction accuracy. The non-linear relationship and unpredictable stock market environment translate to a complex forecast and prediction procedure. However, recent developments in statistics and machine learning allows for earlier technical limitations to be solved. It has been argued that machine learning models can assist in identifying and translating patterns that previously were not comprehensible. This study tests this statement by utilizing the traditional logistic regression along with a newly introduced machine learning library called CatBoost, based on the gradient boosting decision tree algorithm. This study provides evidence of the usefulness of the two models and how they improve the prediction accuracy of directional stock price movements. In addition, the relevance of using accounting data for prediction purposes is supported by the results of the study. Further, the predictive capability of individual performance measures is presented where risk and growth proxies together with profitability proxies are identified as the most important and influential predictor variables.

Keywords: Stock price prediction; Accounting data; Machine learning; Gradient boosting decision trees; CatBoost classifier; Logistic regression; Feature importance

---

We would like to express our thanks to Jan Marton for providing excellent supervision and support in all aspects of the writing process. We are also grateful to Catalin Starica who has provided valuable feedback when setting up our machine learning models. Lastly, we would like to thank our fellow students at the School of Business, Economics and Law at the University of Gothenburg, for their input and helpful comments throughout this term.

Table of contents

1. Introduction

1.1 Background

Investors, traders, shareholders as well as researchers have been trying to predict stock price movements for decades with the incentive to make economic gains and improve investment decisions. Despite researchers' efforts and all the research made on the subject, the practical tools and methods developed to predict stock price movements are limited and there is no one-fit model that will always be suitable (Na & Kim, 2021). A prediction model or trading strategy that seems to work well is usually limited to a shorter time frame as the market is constantly changing as presented by Holthausen and Larcker (1992). This constant change or development is further noticeable when investigating the relationship between the global equity markets and the accounting data of performance measures (e.g., the corporation's cash flow, capital value, and profits), which currently sits at a value substantially higher than the historical average value (Prazak & Stavarek, 2018). The price-to-equity (P/E) ratio for the total S&P 500 in 2021 had a current ratio of 35.96 which can be compared to the historical value that has been ranging between 12 to 15 (Multpl, n.d.). This indicates that the total equity performance (i.e., share price) will further be driven by two other factor variables; expected continued growth and dividends. However, history has shown the importance of financial performance measures and their relation to share price development. Financial ratios from the accounting data and the financial statements are still the basic foundation on which investment decisions are based. The financial performance measures are necessary for investors to analyze corporations and understand future patterns in total equity performance (Prazak & Stavarek, 2018).

The stock market return prediction is an attempt to confirm future values of security, stock, or other financial tools traded in the global stock exchanges. According to Zheng and He (2021) the predictions are of interest for the shareholders, investors, and traders due to the fact that a more accurate prediction of a share price movement may produce higher profits for the investors, both long term and short term (i.e., stock trading). There is always uncertainty and complexity within the stock market and the prediction of stock market returns is intensely challenging with its dynamic, volatile, and nonlinear characteristics (Vijh, Chandola, Tikkiwal & Kumar, 2020; Zheng & He, 2021). Stock prices are affected by internal as well as external factors, for instance, global economic conditions, politics, or unexpected events, which makes the prediction of stock prices highly complicated with multilayered independent

factors. That being said, if it is possible to find even the slightest improvement and accuracy in the prediction of share prices, it could bring a substantial effect on the returns of investments (Zheng & He, 2021).

Technological limitations had previously been identified by Holthausen and Watts (2001), who argued that valuation models and statistical tools needed further development to contribute to the research area. 20 years later, Starica and Marton (2021) provide evidence supporting the argument that recent advancements in statistical methods may allow for new findings within the field of stock price prediction to be made. This study will put this to the test by utilizing newly developed machine learning mechanisms and test whether the new computational power can help us do what humans alone have failed to do. That is, with increased accuracy being able to predict stock price movements based on publicly available accounting data. This will be done by using a machine learning model based on the gradient boosting decision tree algorithm and a logistic regression algorithm.

1.2 Purpose and Research Questions

This research aims to analyze the fundamental part of the market analysis spectrum and study the predictive power of the financial performance measures when it comes to predicting stock market returns. It is a field that is heavily researched with several different angles and approaches. However, Holthausen and Watts (2001) argue that the majority of previous researchers in accounting theory are using methods, especially within the field of linking stock movements to accounting data, that are not fully developed to reflect the knowledge in descriptive accounting theory. They also state that a new focus area is to develop methods and advance the models within the research area and the linkage between accounting data and stock price movements (Holthausen & Watts, 2001). As of today, advancements in models and methods within the field are still highly topical as Bertomeu (2020) states that machine learning has become increasingly more important in accounting research as it is being acknowledged as a powerful and useful tool. Our report will be a part of this new focus area as we try to link accounting data and stock price movements by a newly introduced machine learning tool, thus contributing to the advancements of models and methodologies used in the research area of stock price prediction and how it may be linked to accounting data. However, our study does not only aim to investigate stock price prediction and provide evidence supporting or disproving the concept, but a motivating factor is also to examine the

importance of accounting data. For this purpose, the machine learning approach chosen will be particularly useful since the importance of individual predictor variables will be rated by the model.

In the last decades, technology has become more advanced. The development of data technology and more specifically; artificial intelligence and machine learning models have given the stock market stakeholders an unprecedented computational power to detect previously hidden patterns in the financial data and the opportunity to create a predictive model that can enhance the decision-making process within investment decisions (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018; Vijh et al., 2020; Kotsiantis, Zaharakis & Pintelas, 2007). Accordingly, the purpose of this master thesis project is to exploit this computational power to examine the possibility to predict whether a stock price will increase or decrease based on the quarterly reported accounting data. This would provide convincing evidence of the usefulness of accounting data and also motivate a machine learning approach as viable for prediction purposes in a complex stock market environment. Thus, the authors of this study shall attempt to answer the following research questions:

R1: Can a machine learning classification model be utilized to predict whether the stock price will go up or down based on accounting data features?

R2: Can a supervised machine learning classification model based on a newly introduced gradient boosting on decision trees algorithm be more accurate in its predictions than the traditional logistic regression model?

R3: What are the most important fundamental data and performance measure features of the predictive models?

Although for a different research issue, the motivation behind our paper is supported by arguments made by Starica and Marton (2021), who express that the recent developments in statistics and machine learning can solve the shortcomings in the methodology identified by Holthausen and Watts (2001). Based on newly introduced methods in statistical analytics and machine learning, together with the availability of large finance data, it is possible to construct this research with a new and interesting approach. Since the relationship between

stock returns and financial performance measures is known to be non-linear, a machine learning method can provide us with a less biased result (Starica & Marton, 2021).

This research aims to bring knowledge to both investors and managers on the global stock exchange market and provide decision-makers with useful information regarding firms' stock prices and their expected movements. A regular estimation procedure made by the management is exposed to the risk of being driven by personal judgments and incentives, however, the machine learning algorithm makes its prediction solely based on known information. The information and also the estimation made by the machine learning approach is therefore considered unbiased, as it is purely based on accounting data and financial performance measures (Bertomeu, 2020). Consequently, after adopting a machine learning approach the importance of specific performance measures will also be revealed based on their predictive capabilities. This is particularly useful for investors and traders. If you are able to use a machine learning algorithm and get better predictive power in your investment management practices, it can bring a substantial effect on the investors' returns. This may also prove valuable for decision-makers of future business activities as well as managerial and economic strategies. Evidence supporting the usefulness of accounting data may also motivate increased accounting quality as firms have more incentives to present high-quality accounting information if a direct relation to the company's stock price is recognizable.

1.3 Machine Learning & Gradient Boosting Decision Trees

1.3.1 Introduction to Machine Learning

The concept of machine learning is usually context-specific without a definite and unified definition of the term and what it signifies (Gu, Kelly & Xiu, 2020). To grasp the general concept of machine learning, we refer to the following definition, expressed by IBM (International Business Machines Corporation).

"Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy." (IBM Cloud Education, 2020, 15th July)

In recent years, machine learning has become increasingly more important in empirical accounting research (Bertomeu, 2020). Machine learning allows for enhanced flexibility

relative to traditional economical prediction techniques. Although this additional flexibility also adds risks of overfitting the data, the potential for better approximations of the unknown and complex data generated processes is expected to outweigh these risks. Simply put, the problem of overfitting arises when a model or algorithm has learned and fits the test data too well, which increases the room for error once it is applied outside the test data. Usually, this problem is managed by adding a penalty related to increased complexity, thus motivating simpler solutions ahead of complex ones (Open Data Science, 2018). The risk of overfitting the data can be reduced by making refinements in the implementation of machine learning algorithms and the technique that is applied (Gu et al., 2020).

There are multiple machine learning methods available. Gu et al., (2020) examine the most commonly used ones and concludes that Boosted Regression Trees, Random Forests, and Neural Networks appear to be the most suitable machine learning techniques for the prediction of stock prices, comfortably outperforming the market and other techniques including linear regression, generalized linear models with penalization, dimension reduction through principal component regression (PCR) and partial least squares (PLS). Machine learning methods and especially nonlinear methods can be used to identify conditional expectations as expressed by Starica and Marton (2021). Thus, by developing this further we believe it might be possible for the model to also make powerful predictions based on these expectations which can be particularly exploited by an active investor.

1.3.2 Why Apply a Machine Learning Method?

Starica and Marton (2021) investigate the relation between stock prices and accounting earnings. They argue that machine learning and recent developments in statistics allow for earlier limitations and issues identified within the field to be solved. These issues were previously highlighted by Holthausen and Watts (2001), who argued that valuation models used in accounting research at the time were not sufficiently elaborated. While machine learning and algorithmic modeling are relatively commonly used in finance, accounting research has yet to fully embrace these models and the opportunities they may bring. Using machine learning for stock market prediction is not only less tedious and time-consuming than the manual procedure it has been historically, but the prediction is also based purely on numbers and data, therefore ignoring potential emotional bias (Shen & Shafiq, 2020). Since earlier machine learning methods were introduced, training costs and time required for

learning have also reduced as both computer power and machine learning programs have developed further and become more efficient (Shen & Shafiq, 2020).

1.3.3 Limitations to Machine Learning

Although machine learning has the potential to improve risk premium measurement, which is an essential part and a difficulty associated with prediction, these improvements are only measurements and do not contain information about actual economic mechanisms or equilibriums. Machine learning may be useful to understand economic mechanisms, however, these tools do not on their own identify these associations and relationships between conditional variables and asset prices (Gu et al., 2020). In short, machine learning is a useful tool, although it still requires someone to design, maneuver, structure, and control the process and apply the best-suited algorithms for specific uses.

One of the major limitations of machine learning models is their "black-box" nature. You provide inputs to the algorithm and get outputs in return. The machine learning algorithm itself is technically complicated and details about the process are usually left out with a limited explanation. For instance, it may exclude information about why the algorithm would choose one scenario over another or how it came up with certain decisions (Chan, Reddy, Myers, Thibodeaux, Brownstone & Liao, 2020)[1]. Related to the black-box nature of machine learning, there is also the risk of overreliance on the machine learning model as well as the risk of misinterpretation. The user still has to carefully examine the algorithms in case of small errors that may affect the result or cause the algorithm to fail. Without the necessary skills to interpret the result and enough knowledge to revise and edit the model inputs, the machine learning model will not be particularly useful to the user (Armstrong, 2015; Malik, 2020).

Machine learning is limited to the information that is provided. Thus, there might be a lack of data or a lack of "good" data. As stated by Chan et al. (2020), machine learning is prone to the maxim "garbage in, garbage out". Accordingly, it is up to the user to decide which variables to include and what data to use. Some of the most important variables could be left out since their importance is unexpected to the user. In that sense it is also not always

[1] This article is not a financial article, however it provides a good explanation of the machine learning model.

completely unbiased as the user may include variables based on expectations and personal judgments (Chan et al., 2020; Malik, 2020).

1.3.4 Gradient Boosting Tree Model

Neural networks are commonly regarded as the most powerful and suitable machine learning model, including when it comes to predicting stock prices (Gu et al., 2020). However, the authors also highlight the random forest method and boosted regression trees as effective machine learning methods when predicting stock returns. The focus of this paper will be on the latest mentioned, the boosted regression trees or more specifically; the gradient boosting decision tree model (GBDT), a classification algorithm. The fact that the GBDT is simpler than neural networks to construct (Gu et al., 2020), makes GBDT an attractive machine learning method, especially for the average investor. Compared to the random forest method, decision trees also require less time to learn (Pradeepkumar & Ravi, 2017). GBDT is a commonly used model in research outside the accounting and finance fields (Daoud, 2019). However, previous research within the accounting and finance field is lacking when it comes to utilizing gradient boosting on decision trees, which makes the angle of approach in this study different from existing literature.

The motivation for conducting this research is to establish if it is possible to predict whether a stock price will go up or down based on accounting data with a machine learning classifier, more specifically, if a gradient boosting decision tree algorithm model can be utilized to predict stock price movements, using accounting data as the predictor variables. If so, the GBDT can be an appropriate choice to implement in an investment decision-making process. The technical parts of the classification model will be discussed in the method section of the paper.

1.4 Thesis Structure

The structure of the thesis consists of several chapters. Starting with a theoretical section consisting of a literature review where previous literature on stock prediction and its relationship with accounting data is presented. To conclude this chapter, the main hypothesis of the report is presented and motivated. In Chapter 3 the data is presented including the selected sample as well as the predictor and response variables used in our models. In addition, the methodology for this study is explained and the machine learning tool and

approach used are further described. In Chapter 4 the results of the study are presented. This result is later discussed in Chapter 6 where the results of the models are compared against each other and how it relates to previous research. Lastly, the final section includes a conclusion of the report and our findings. Further, limitations within the study are discussed and guidance for future research is expressed.

2. Theoretical section

2.1 Literature Review

In the following section, relevant literature and their respective findings will be presented. More specifically, research papers that have attempted to determine the relationship between financial ratios and performance measures, and the stock price movements. For starters, it is important to address the Efficient Market Hypothesis (EMH). A theoretical approach that is relatively basic, but frequently expressed. The EMH states that all information instantly is reflected in stock prices. This suggests that stocks always trade at a fair value and that finding undervalued stocks is impossible. Therefore, the technical and fundamental analysis would be pointless. Ultimately, EMH suggests that one cannot consistently outperform the market (Fama, 1965a; 1965b; Samuelsson, 1965; Schmidt, 2011; Delcey, 2019). This makes EMH highly controversial with a substantial amount of academics and researchers both supporting and criticizing the theory and its fundamentals.

Fama (1970) identifies three different subcategories of efficient markets. Namely, weak, semi-strong, and strong form. The forms differ in terms of information available and how strong or extreme of a market efficiency that is considered. The weak form suggests that today's stock prices reflect all data of past prices, the semi-strong form suggests that all publicly available information is reflected in current stock prices and the strong form acknowledges all information, whether its public or private, to be reflected in the current stock price. With performance measures being publicly available through income statements, numerous studies examine whether this information of performance indicators can be used to predict stock returns to outperform the market. However, with mixed results, as later presented by Basu (1983), Schrimpf (2010), Dimitropoulos and Asteriou (2009), Lewellen (2004), Dzikevičius and Šaranda (2011), and more. As mentioned, most of the financial ratios and performance measures that are interesting to include are usually presented in the

corporation's financial statements. If they are not reported directly, the numbers can at least be manually computed from the reported raw numbers in the financial statements.

In short, the theory of the efficient market hypothesis argues that security prices and share prices will immediately decrease or increase in response to the knowledge gained from a number of different variables. These variables include, but are not limited to; net assets, cash flow, return on capital employment, D/E-ratio (debt-to-equity), dividends, and earnings (Fama, 1970). Although EMH has been heavily researched and widely covered, the number of supporters, as well as critics, makes this theoretical section constantly relevant for new studies to be designed and performed. Technical developments in statistical tools further enable discoveries and new evidence to be found, which will contribute to existing knowledge within the theory of efficient markets.

Several different research papers have made an attempt to predict stock market performance based on three different types of predictor variables. In the earlier stages, research papers within the field focused on predictor variables related to the past prices (e.g., moving average) and multiple researchers could identify significant patterns and relations between the predictors related to past prices and stock prices movements (Fama, 1965a; 1965b; Jegadeesh & Titman, 1993). The second type of predictor variable that is broadly utilized within the field of predicting stock prices is the accounting data of performance measures and other economic fundamental variables. In particular, Jegadeesh and Livnat (2006) analyzed revenue surprises and stock returns. They confirmed that analysts are slow to incorporate information about revenue and earnings surprises into their forecasts. This allows for profitable trading strategies to be exploited, especially in the case of more extreme surprises. Other previous studies that also used fundamental analysis and accounting data in predicting stock performance are Basu (1983), Rosenberg, Reid, and Lanstein (1985), and Cooper, Gulen, and Schill (2008). These articles raise evidence towards market inefficiency and argue that fundamental data can be useful to predict stock return, thereby increasing the likelihood of outperforming the market. The third type of predictor variable that has been utilized within the field is non-financial information, more specifically this could include text and social media sentiments (Chen, De, Hu, and Hwang , 2014; Kim & Kim, 2014; Renault, 2017). This study will focus on the second type of predictor variables, which includes accounting data of performance measures and economic fundamental variables.

Historically, accounting numbers were for a long time generally regarded as somewhat meaningless. However, when Ball and Brown (1968) provided evidence supporting the fact that accounting earnings were in fact useful, they set the foundation for future capital markets research in accounting by inspiring researchers to investigate the phenomenon further. The research used databases with accounting data and monthly stock prices which at the time recently became possible to collect in machine-readable form (Kothari & Wasley, 2019). Thus, the founding motivation and methodology behind the following presented literature, as well as this study in its whole, can in many regards be traced back to the study by Ball and Brown (1968). Just like the EMH, these authors and their conclusion that motivates the usefulness of accounting earnings were faced with numerous supporters as well as critics (Kothari & Wasley, 2019).

Stock price prediction includes numerous unpredictable components, both internally and externally. The components affecting the stock price may also change over time and each firm's stock price is not only affected by different factors, but they also react differently to these factors. In addition to these complications, there is also intense competition from other traders. This means that if successful forecasting techniques are discovered, they would be copied and widely used by others. Thus, eliminating the ability of the forecasting model (Rapach & Zhou, 2013). Although forecasting models will most likely never be able to explain more than a very limited part of the stock returns, forecasting methods do have evidence supporting the fact that they improve forecasts in a way that will result in economic gains that outperform an investor that does not rely on these methods (Rapach & Zhou, 2013).

In 1983, Basu (1983) concluded that earlier research had suggested that firms with higher earnings to price ratio most times had higher stock returns than firms with lower earnings to price ratio (E/P). However, during his study, he identified a clear difference between larger and smaller firms, where the larger than average firms had an insignificant or only marginally significant E/P effect, which he argued was primarily suppressed indirectly by firm size. He emphasizes that earnings, firm size, and expected earnings have a complicated relationship. Takeaways from Basu (1983) suggest that firm size should be considered and that the result of our study might differ depending on the size of investigated firms.

Schrimpf (2010) examined the validity of corporations' financial ratios in their predictive abilities toward the profitability of corporations in the stock exchange market. The author could conclude that the predictability of profitability with the use of financial ratios has a high correlation with a stock price analysis. These results indicate that financial ratios, most specifically profitability ratios, do have a close relationship with the share price movements. These findings are akin to a study by Dimitropoulos and Asteriou (2009) where the article's findings confirmed that profitability ratios are the most important determinants of predicting stock returns. The findings of this particular study suggest that leverage is not a key variable for predicting stock returns, instead the authors conclude that the most productive firms are rewarded with higher stock returns, and accordingly, the productivity variables were the key variables for prediction.

Lewellen (2004) made a study with regressions and results using different variables compared to the previously mentioned studies. The results of this study focused on the long horizon and the researcher found significant results in the dividend yield, P/E-ratio, and book-to-market ratio as predictor variables of stock returns. In connection with the previously mentioned articles, Dzikevičius and Šaranda (2011) studied 20 different financial performance ratios as predictors of stock returns and resulted in other predictor variables that better can explain the stock returns. More specifically, variables including return on assets (ROA), return on equity (ROE), and total-liabilities-to-equity.

Based on mentioned studies, numerous different performance measures have been presented and all have different predictor variables as the most significant variables. Although, it seems that the performance ratios measuring profitability have been the most significant predictors of stock return movements. This gives valuable insights into which variables to test. Based on this part of the literature review, D/E, leverage, ROE, and ROA are the most intriguing financial ratios that should be included in the research. The leverage ratio is interesting to include as a predictor variable because even though the study by Dimitropoulos and Asteriou (2009) concluded that it is not a key variable for predicting stock returns on Greek stocks, Dzikevičius and Šaranda (2011) did in fact identify a relationship between leverage and stock price movements in their study. This suggests that leverage might still be a key variable in our study and should not be excluded without testing.

In a more contemporary article, Ball and Nikolaev (2022) study earnings and cash flows as predictors of future cash flows. The authors conclude that earnings will actually provide better information about future operating cash flows, compared to using historical operating cash flow as a predictor. Consistent with Dechow and Dichev (2002) and Dechow (1994), the authors argue that operating cash flow is a noisy measure of performance because of timing and matching problems. This is especially apparent in firms in which cash flows are not in a steady continuous flow. Firms that are in a growing phase or have irregular income are especially noisy and cash flow is even less useful as a predictor (Dechow, 1994). Nonetheless, even for stable firms, Ball and Nikolaev (2022) and Dechow (1994) suggest that operating earnings will give a more accurate prediction of future cash flow. Further, they conclude that accrual-based measures can effectively reduce noise. Our study will relate to these studies in the sense that accrual-based performance measures, for instance, gross profit, ROE, ROA, and operating income, will be included as predictors for stock return.

Accounting literature agrees on the fact that price and earnings have a nonlinear relationship (Starica & Marton, 2021; Vijh et al., 2020). This was hinted at already 50 years ago by Ball and Brown (1968). This suggests that there is no simple straight-line correlation between the dependent and independent variables. Instead, the relationship can take a variety of differently curved forms, making it less straightforward to identify and interpret. Using a linear regression model to interpret a nonlinear relationship would lead to a misspecification of the functional form, thus a nonlinear model is required to effectively capture said relationship (Brooks, 2014). The nonlinear regression model used is further described in the methodology chapter of this paper (Chapter 3.4.5). Because of recent advancements in technology, Starica and Marton (2021) propose a modern machine learning method for nonlinear modeling. For their study, the authors use a nonlinear regression Random Forest and ultimately present price to book ratio, size, and total liabilities to the market value of equity as the most important and influential variables in the relation between price and earnings.

Advanced intellectual techniques are proven to be useful in finding hidden patterns in large data sets with complex relations (Vijh et al., 2020). Machine learning methods are more efficient than past methods for this task and there are researchers and investors effectively using these tools for forecasting purposes. Multiple different machine learning methods exist where some may perform better than others. The complexity of each model also differs. New

machine learning packages and techniques are being developed with the ambition of being more user-friendly or accurate than previous techniques (Vijh et al., 2020; Gu et al., 2020).

2.2 Hypothesis

This report aims to investigate whether performance measures from accounting data can be used to predict stock prices by using machine learning and statistical tools. This study will test the phenomenon that is stock price prediction. It will test whether the prediction is possible and if accounting data is useful for this purpose. Any evidence of this would be remarkable as it would suggest that it is possible to exploit this information and outperform the market.

We motivate a null hypothesis that is in line with the efficient market hypothesis which suggests that accounting data cannot be used to predict stock price movements.

H^0: Performance measures and accounting data cannot be used to predict stock price movements.

If the null hypothesis is rejected, results must show significant evidence of the predictive capability of accounting data, which further can be used to predict stock price movements and outperform the market. We reject the hypothesis if the two models can predict stock price movements with an accuracy over 50%. As we base our study on publicly available information, rejecting this hypothesis would also suggest that the semi-strong form of EMH does not hold.

3. Data and Methodology

3.1 Data

The data sample used in this thesis is the US market, more specifically, large companies within the S&P 500. The choice of this sample is based on the foundation that we wanted companies that were relatively large and financially stable with many years of historical financial data. Further, investment companies have been excluded as their accounting data substantially differs from regular companies due to ownership in other firms and different financial reporting. Consequently, the research is based on 418 firms from the S&P 500. The availability of data is crucial, and since US-listed firms are required to publish this

information, collection of data was no issue. Databases are being used for this task and the firms within the selected sample are large. If our sample would have also included smaller firms and newly founded firms, this could have been more of an issue. The industry distribution of the sample is presented in Table 1.

Industry	Number of firms
Consumer Discretionary	83
Information Technology	69
Industrials	61
Health Care	52
Energy	43
Consumer Staples	40
Utilities	33
Materials	29
Telecommunications Services	8

Table 1. Industry Distribution

Further, the data collection period ranges from 2007 to the end of 2021. It could have been of interest to exclude 2020 and 2021 due to the complicated situation with the pandemic that had a substantial effect on the stock market. However, market conditions may not always be perfect, and excluding these years may result in a machine learning model that solely works in less turbulent market conditions. Thus, there would be an increased risk of ending up with a beautified model that does not work in practice. It is also interesting to see how well the models are performing in an uncertain time on the macro front. The chosen classification model is efficient and does not require much data for learning. Therefore, the selected time period did not need to expand further.

The fundamental data collected is based on the literature review. In addition to the share price and earnings, it includes risk and growth proxies. Namely, size, price-to-book, book-growth, revenue growth, and earnings growth. The predictor variables further include profitability proxies, financing proxies, and investment proxies. All predictor variables are listed and described in Chapter 3.2. Sample data for the study was collected from the WRDS database. More specifically, the accounting information was gathered from Compustat and the share prices were obtained through CRSP. An overview of the sample data and the descriptive statistics can be found in the Appendix of this paper.

3.2 Predictor Variables

3.2.1 Risk and Growth Proxies

Existing literature suggests that risk, growth, economic rent, and accounting conservatism affect the relation between prices and earnings (Holthausen & Watts, 2001; Kothari & Shanken, 2003; Liu & Thomas, 2000; Starica & Marton, 2021; Biddle, Chen & Zhang, 2001). Price-to-book (P/B) is one of the identified proxies for risk, economic rent, growth opportunities, and unconditional conservatism (Fama & French, 1992; Roychowdhury & Watts, 2007). Apart from share price and earnings, Starica and Marton (2021) classify this proxy as the most important explanatory variable in shaping the price-earnings relation, although when using a different machine learning approach. Another well-known risk and growth factor is size. The importance of including a size proxy was particularly motivated by Basu (1983). Inspired by Starica and Marton (2021), two different estimations of size will be used and tested as explanatory variables. These are based on total assets as well as previous-year market size.

Cash volatility could have been included as a proxy for risk and growth but was ultimately excluded. This decision was based partly on Starica and Marton (2021) who identified cash volatility as less relevant in explaining the price-earnings association than previously mentioned proxies of risk. In addition, numerous studies have motivated cash flow as a noisy measurement with timing and matching problems that are especially common in firms with volatile and irregular cash flow (Ball & Nikolaev, 2022; Dechow, 1994; Dechow & Dichev, 2002). Therefore, its predictive ability of share price might be limited.

Multiple direct measures of growth were used. Namely, revenue growth, growth in earnings, change in total assets, and change in equity. All proxies for risk and growth, as well as their respective calculations, are specified in Table 2. Using lagged values in the denominator reduces heteroscedasticity in terms of unwanted biases and unwanted correlation effects (Algharaballi & Albuloushi, 2018).

Variable Name	Definition
P/B	Previous year price to book ratio (mkvaltq/ceqq)
Size 1 - TA	Firm's ventile of total assets in the cross-section (atq)
Size 2 - MV	Firm's ventile of previous years market value in the cross-section (mkvaltq)
Revenue growth	Change in sales (saleq) / lagged sales (saleq)
Earnings growth	Change in earnings per share (epsxq) / lagged EPS (epsxq)
Book growth	Change in common equity (ceqq) / lagged equity (ceqq)

Table 2. Risk and Growth Proxies

Risk and growth also include more indirect proxies, where investment and financing are proxies of risk, while profitability and payout policy are proxies of growth (Starica & Marton, 2021). These proxies are large enough to be identified as their own category which includes numerous proxies and variables. Three of these categories are included in this study as pay-out proxies have been excluded. Even though payout proxies have been identified as significant predictors of stock returns (Lewellen, 2004), this predictive capability is primarily long-term (Beaver and Ryan, 2005). This report focuses on short-term movements and is based on quarterly data which makes these proxies less useful for our purpose.

3.2.2 Profitability Proxies

Dimitropoulos and Asteriou (2009) acknowledge profitability ratios as the single-handedly most important determinants of predicting stock returns. The importance of profitability ratios is further supported by Dechow, (1994), Ball and Nikolaev (2022), Starica and Marton (2021), and Schrimpf (2010), who all identify a close relationship with stock price movements. Therefore, multiple profitability ratios were included.

Starica and Marton (2021) further explore profitability ratios inspired by previous studies, particularly the studies made by Fama and French, (2015), Novy-Marx (2013), and Ball, Gerakos, Linnainmaa, and Nikolaev (2016). Because of the relative importance of these ratios compared to other explanatory variables investigated, multiple profitability ratios were included in this study. These ratios include gross profit-to-book value, gross profit-to-assets, operating income-to-assets, and R&D-adjusted OI-to-assets.

Besides mentioned proxies, return on assets (ROA) and return on equity (ROE) were also included as their respective explanatory abilities have been highlighted as significant (Dzikevičius & Šaranda, 2011). All profitability proxies and their respective decomposition is presented in Table 3.

Variable Name	Definition
Operating income-to-Assets	OI before depreciation - taxes and interest (oibdpq) / lagged assets (atq)
Gross profit-to-Book value	Revenue-Cost of goods sold (saleq) / lagged book equity (ceqq)
Gross profit-to-Assets	Revenue-Cost of goods sold (saleq-cogsq) / lagged assets (atq)
R&D adjusted Operating income-to-Assets	OI (oibdpq) / Assets (atq) +R&D (xrdq) / lagged assets (atq)
Return on Assets	Earnings (niq) / lagged assets (atq)
Return on Equity	Earnings (niq) / lagged book equity (ceqq)

Table 3. Profitability Proxies

3.2.3 Financing Proxies

Debt-to-equity (D/E) is a commonly used ratio to evaluate a company's financial leverage. It reflects the degree to which a firm's operations are financed through debt. Thus, higher leverage indicates a higher risk to the shareholders. D/E is calculated by using total liabilities as the numerator and shareholder's equity as the denominator. The inclusion of this variable is motivated by Dzikevičius and Šaranda (2011). Additional financing proxies used as explanatory variables in this study include interest-to-equity and debt-to-asset.

Variable Name	Definition
Debt-to-Equity	Total debt (dlttq) / total equity (teqq)
Interest-to-Equity	Interest expense (xintq) / equity (teqq)
Debt-to-Asset	(Long term debt (dlttq) + debt in current liabilities (dlcq)) / total assets (atq)

Table 4. Financing Proxies

3.2.4 Investment Proxies

Fama and French (2015) define an investment factor as the change in total assets, however, this variable is already included as an explanatory variable, being a direct measure of growth. Following Starica and Marton (2021), other investment variables that are included in our data are the investment-to-asset (I/A), cash-to-assets, tangibility, R&D intensity, and inventory-to-assets. Chen, Novy-Marx & Zhang (2011) define investment-to-asset as the annual change in property, plant, and equipment plus inventory change, divided by lagged total assets.

Variable Name	Definition
I/A	Investment-to-assets, (PPE (ppegtq) + inventory change (invtq) / lagged total assets (atq)
Cash-to-Assets	Cash (cheq) / total assets (atq)
Tangibility	Fixed assets (ppentq) / total assets (atq)
R&D instensity	R&D spending (xrdq) / lagged assets (atq)
Inventory-to-Assets	Inventory (invtq) / total assets (atq)

Table 5. Investment Proxies

3.3 Response Variable

As the chosen model in this study is a classification model, it is specifically designed for a binary response. Therefore, we created binary variables. The two binary variables will be defined as 1 and 0. The response variable (or dependent/target variable) is the stock return and if the returns are greater than zero, the binary variable will be 1. If the returns are smaller or equal to zero, the binary variable will be 0. Return is based on whether the stock price has increased or decreased, which also can be phrased as a profit or loss on an investment.

The two binary variables:

Total stock return > 0 = 1

Total stock return ≤ 0 = 0

When estimating stock returns, logarithmic returns are used. In contrast to regular (or simple) returns, logarithmic returns are continuously compounded. This makes it easier to compare returns between different assets and companies. Logarithmic returns are commonly used when discussing and estimating returns over time as the compounding effect is effectively captured. In practice, a positive and negative return of the same magnitude will cancel each other out when using logarithmic returns. It is assumed that stock prices follow lognormal distribution, therefore logarithmic returns are suitable. This further ensures that stock prices can not become negative (Brooks, 2014).

The logarithmic return is estimated as follows:

$$ln(\frac{New\ Closing\ Price}{Previous\ Closing\ Price})$$

3.4 Methodology

3.4.1 Supervised Machine Learning

To understand the predictive methodology that was used in this thesis, it is necessary to introduce machine learning and more specifically, supervised machine learning. To begin with, supervised learning is the most utilized form of machine learning. It basically involves having input variables (X) and an output variable (Y), and then utilizing an algorithm to train the mapping function on the output variable with the data from the input variable. The main goal is to approximate the function so well that when you collect new input data, you can

predict the output variables for that data. As the algorithm is created on historical data, we do have the correct answers, then the algorithm makes predictions repeatedly on the training data and then tests the model on the output data to see how well the model has learned the patterns. Examples of supervised learning models include regression models and classification models (Kotsiantis et al., 2007).

3.4.2 Classification

If the task is to classify data points into a specific number of classes (Total stock return > 0 = 1, Total stock return ≤ 0 = 0), a classification model is the most appropriate method to utilize. An example of a classification model is given below where the purpose is to separate data pointers whether they lie below or above the separator curve with the function 1/X (Javapoint, n.d.).

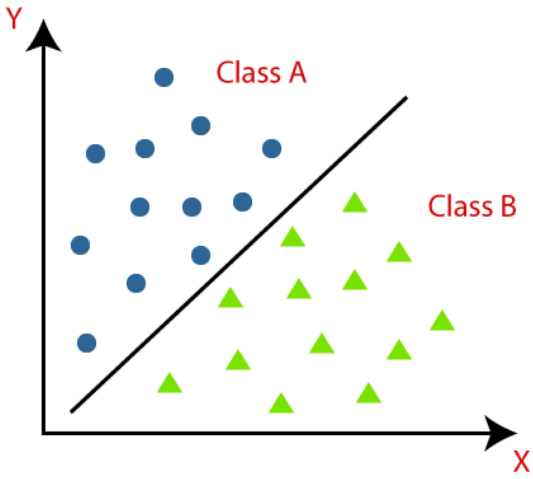

Figure 1: Example of a Classification Chart (Javapoint, n.d.).

Visualized in Figure 1, two different classes are represented in blue and green depending on their location relative to the curve. We call these the blue and green classes. The data points below the separator belong in the green class and the data points above belong in the blue class. The main task with this type of classification algorithm is to process a chosen amount of data observations and make an attempt to simulate the behavior in the general case and thereby approximate the unknown separator function (Javapoint, n.d.).

As the purpose of this study is to predict if a stock return will increase or decrease based on the accounting data, a model with a binary classification model is preferred. As mentioned, a classification model builds on a number of classes, and this type of model is fitting as the

main method used in this paper. The choice of utilizing classification models is motivated by the purpose of this paper which is to predict if a stock price increases or decreases at a given time when the quarterly reports are released. This makes the classification models most suitable for our research questions. A regression model would have been relevant if the aim was to predict continuous values, such as share prices and market trends. However, as we only want to predict if the stock price will go up or down at a given time, a classification model where the model predicts (i.e., classify) the discrete values of total stock return > 0 and total stock return ≤ 0 is well fitted.

Furthermore, within the classification models, there are several different machine learning classification algorithms. For instance, Logistic regression, Random Forest Classification, Kernel SVM, Support Vector Machine, and Decision Tree Classification (Javapoint, n.d.). In this paper, the chosen classification models are a Logistic regression model and a decision tree classification (gradient boosting on decision trees). The basis of this model selection will be argued in the sections below.

3.4.3 Logistic Regression Model

As known from previous research, stock price prediction is a regression analysis problem if you want to predict stock prices or market trends. However, if you want to predict movements of the stock increasing or decreasing based on different factors, it is more of a classification analysis problem. Therefore, the research design for this thesis was a predictive research design, which is empirical research with models that aims to forecast future events (e.g., stock returns). The most commonly used models in this type of research design are OLS-regressions, logistic regressions, decision trees, and neural networks (Mitchell, n.d.).

Starica and Marton (2021) are pointing out that the relation between stock return and accounting performance measures is known to be nonlinear, and as the aim is to predict the discrete values of total stock return > 0 and total stock return ≤ 0, only the classification models are suitable for this task. Using a linear probability model would be problematic since the relationship between our binary outcome and continuous independent variables are nonlinear, leading to misspecification of the functional form. This may generate biased estimates. In practice, this model would allow for probabilities greater than one or even negative probabilities, which should not be possible (Brooks, 2014). This makes a logistic

regression model more suitable, which is the first model that was applied in this research to examine the different performance measures' impact on stock returns.

Different from the linear model, logistic regression uses a cumulative distribution function where probabilities of the coefficients take a value between 0 and 1. This makes the model S-shaped instead of linear. When interpreting a nonlinear model, the magnitude of the coefficient estimates cannot be translated directly to marginal effects. Instead, observation is made on whether the relation between the dependent and independent variables is positive or negative. In addition, the coefficients can be compared to each other on an ordinal scale based on their influence on the dependent variable (Brooks, 2014). This suggests that the degree of difference between coefficients cannot be specified, but it is observable whether one coefficient is more important and influential than another.

Within machine learning, the logistic regression model is a classification model that predicts the probability of the dependent variable (i.e., target variable) and the dependent variable will only have two possible classes. For this thesis, stock return is the dependent variable where total stock return > 0 is one class and total stock return ≤ 0 is the second class. The performance measures were the logistic regression's suitable independent variables. The logistic regression describes the relationship between the independent variables (predictor variables) and the dependent variable (stock returns). A logistic regression model is specifically designed for a binary response (Brooks, 2014). As previously expressed, we created a binary variable. If the returns are greater than zero, the binary variable will be 1. If the returns are smaller or equal to zero, the binary variable will be 0.

The classification model described is solely predictions made within the data that we have collected. However, as the main aim of this study is to predict future share price movements, we need to use a model that can predict data outside the collected data. Using a program like Python for this task, it is possible to create a model with machine learning for the predictive analysis. This basically means that we trained a model with a percentage of the data we have collected to see the accuracy of the predicting model. Then we tested the model with the untrained data. The results of the logistic regression model were then compared to the result of the gradient boosting algorithm to see which model yields the best results. When training the model in Python and then testing it, the model is providing us with an accuracy score. The score from this model is compared later in this study with the purpose to investigate whether

the machine learning algorithm could predict the stock returns better and outperform the traditional logistic regression model as well as an uninformed investor.

3.4.4 CatBoost: A Machine Learning Library

When creating the other machine learning classifier model that builds on decision trees, the most important code package utilized in this study was the CatBoost package. CatBoost was developed by Yandex and is an open-source algorithm that enables machine learning and predictions. The machine learning package works well with different types of data. For instance, text, audio, images and historical data. The latter (historical data) is the most significant data source in the case of this study (Ray, 2017). CatBoost is based on gradient boosting algorithms which is a type of machine learning algorithm that is used in several different business challenges, most notably in forecasting and prediction tasks. The algorithm can also provide the user with good results with a relatively small amount of data, which is the main difference between this machine learning algorithm and other Deep Learning models that need a massive amount of data to train on (Prokhorenkova et al., 2018).

There exist several different types of algorithm libraries within gradient boosting, for instance, LightGBM, H2O, and XGboost. However, the selection of CatBoost is based on the same advantages of the algorithm library. Firstly, the most important advantage is the performance. The CatBoost library has a very good performance and provides ultra-modern results. Although the library is easy to use, it still performs on the same level as any other machine learning algorithm. Secondly, as already mentioned, CatBoost is simple to use and the library is doing a lot of the work in pre-processing data by itself. The algorithm is handling the categorical feature automatically which means that instead of converting categorical data into numbers, which you have to do in other machine learning models, the algorithm uses various statistics on combinations of numerical and categorical features. This makes the CatBoost model very user-friendly (Ray, 2017).

Furthermore, the algorithm reduces the chance of overfitting and decreases the need for parameter tunings which makes the model more robust than others. Overfitting is often mentioned as the main disadvantage in a gradient boosting model compared to Random Forest. However, the CatBoost algorithm is adequately solving this dilemma of overfitting by using oblivious decision trees as predictor bases (Prokhorenkova et al., 2018). The term oblivious suggests that each tree level uses the same splitting criterion, making them more

balanced than a standard decision tree and therefore less prone to overfitting. An additional benefit is that this also allows for the machine learning testing time to be significantly faster (Prokhorenkova et al., 2018). Lastly, the choice of CatBoost was also based on a comparison between CatBoost and other boosting algorithms libraries, showing that CatBoost is providing results with the lowest log-loss values. In addition, when using CatBoost there is no requirement of conversion data sets to any specific format, which is the case in both LightGBM and XGBoost (Ray, 2017). This is supported by Prokhorenkova et al. (2018) who identified that the CatBoost gradient boosting algorithm, which builds on ordered boosting with ordered target statistics, solves some of the prominent problems within predictions and outperforms previous leading gradient boosting packages.

3.4.4.1 Gradient Boosting on Decision Trees

The CatBoost model, also known as the CatBoost classifier, is based on a gradient boosting tree algorithm (GBDT). The GBDT is the main method that was utilized to test the prediction capability of accounting data. It is a good fit machine learning model when the data is heterogeneous, which is the case in this research. There are other machine learning tools, primarily neural networks or random forests. However, they can be too complex and not that user-friendly which makes the choice of selecting CatBoost easier as the CatBoost model and the GBDT is rather simple to use and is still a powerful model for predictive analysis (Ray, 2017). Friedman (2001) was the first paper to introduce the concept of GBDT. The author proposed a new type of classification method that would combine multiple weak classifiers (i.e., weak trees) and then create a robust classifier. Daoud (2019) also identifies that a weak decision tree is a classifier that cannot do any accurate predictions. However, the weak classifier is still better than a purely random prediction. That said, it is when combining weak decision trees into an ensemble, that the model can make meaningful predictions with accurate results (Daoud, 2019).

The model is constructed on the same binary variables as the logistic regression model. What the CatBoost model does is that it creates a gradient boosting algorithm that combines several weak learning systems and creates a strong learning model that actually can see patterns in the data (Prokhorenkova et al., 2018). Below is visualized how a gradient boosting tree works.

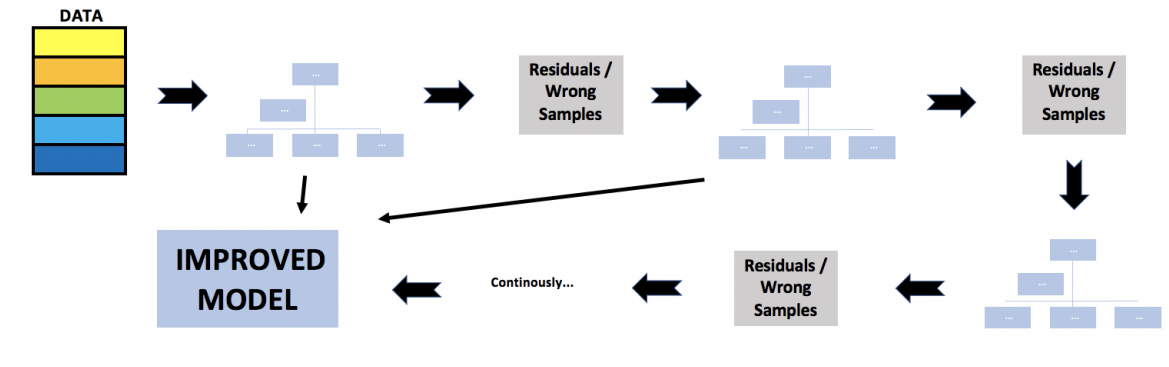

Figure 2. Gradient Boosting Tree Model

The model tests different configurations to test the data and delivers the best configurations for the model to us by itself (Prokhorenkova et al., 2018). The model then gives an accuracy of how well the performance measures and the model manages to predict stock price movements. Another feature that can be used with this gradient boosting algorithm from the CatBoost library is that one can get the importance of the variables in the model (Ray, 2017). Meaning that it is identifiable which variables from the accounting data that have the most influence on our prediction results. This could further guide and pinpoint future research within the field. As aforementioned, if we can see just a small improvement in the prediction from the machine learning tool, it can make a great impact on the equity market and how investors operate.

3.4.5 Feature Importance

According to Billiau (2021) and Chan et al., (2020), machine learning models are commonly referred to as "black-box" models. As the nickname suggests, these models are very complex. It is a challenge to understand and interpret what is going on inside these models. The challenge of interpretation exists in deep learning models like neural networks as well as in ensemble models like gradient boosting trees. The models can yield really good prediction results; however, the drawback is the difficulty to understand how the inputs (predictor variables) are combined and how they influence the predictions (Billiau, 2021).

In traditional statistical models like linear regression, it is possible to make precise statements based on statistical inferences. The statistical inferences can tell us how the inputs relate to the output. In the case of these black-box models, it is almost impossible to make these types of statements due to the complexity of trainable weights that link the input variables with the

output variable. Feature importance techniques are relatively new within machine learning that solves this interpretability problem (Billiau, 2021). Feature importance techniques assign each predictor variable a score based on its ability to improve predictions. This makes it possible to rank the predictor variables in our machine learning model based on the predictive power of each variable (Billiau, 2021; Bertomeu, 2020).

The feature importance method used in this study is a technique provided by the CatBoost library. The method builds on three different formulas: loss function change, prediction value change, and internal feature importance (CatBoost, n.d.).

It is important to point out that feature importance is not a statistical inference and the method can only tell us how important a feature is for the prediction. The results from the feature importance analysis do not say anything about the relationship if it is linear, non-linear, quadratic, or the magnitude of the predictor effect (Billiau, 2021; Bertomeu, 2020). However, it is still a good alternative solution when utilizing these black-box models where it is nearly impossible to perform a traditional statistical inference (Billiau, 2021).

3.4.6 Python and Feature Selection

3.4.6.1 Data Manipulation and Creating the Data Frame

The programming language used throughout this master thesis is Python. Python is a popular programming language and is considered one of the most user-friendly programming languages. Although it is relatively simple to use and learn, it is still a programming language that can be powerful in tasks of machine learning and data analytics. The main advantage of Python is that it allows code packages that facilitate the utilization of the language and encourage code reuse and program modularity (Python, n.d.). The data that was extracted from Compustat and CRSP could easily be imported into Python with CSV files. Later, the Pandas library was utilized to process the imported data. Panda library is a software library constructed for the Python programming language and can provide codes for data manipulation and analysis. The codes in the Pandas library were utilized to load the CSV-file of the quarterly fundamental data from all firms in the S&P 500. Following this, manipulation of the data was necessary to get the data structured in a way that could be used for the prediction. We started by setting the date as the index, before merging the fundamental data file with the stock data file. Lastly, predictor variables were calculated and added to the fundamentals data file. Calculations of all predictor variables are detailed in Chapter 3.2.

To get the stock data, the Pandas library was used once again to read all the daily prices from the companies from the CSV-file retrieved from CRSP. To get this data file in the same structure as the fundamentals data file, we set the date as the index. These two files were then merged to get a data frame with all the quarterly fundamentals and the quarterly stock prices in the same data frame. Lastly, the log returns were calculated and added to the data frame as this was our response variable for the prediction models. We created a variable for the logarithmic returns on a quarterly basis as well as dropped the 'NAN' values (i.e., the missing values, otherwise the models would not work). We chose to include the logarithmic returns since that is what typically is being used in this type of analysis (thereby excluding the raw returns and returns percentage). The decision of using logarithmic returns rather than simple returns is further described in Chapter 3.3. As the purpose was to predict future stock prices, we used the pandas' shift function to get the next quarterly log return on the row of the previous quarter of fundamental data.

In order to get the complete data frame, the pandas.concat() function in Python was being utilized. Following, a data frame with all companies and data stacked below each other was created. As the aim was to train the prediction models with a specific number of years to predict the next period, a modification of the data frame was necessary to have the data frame sorted by date instead of by companies, which was being made within the concatenated data frame.

Lastly, as the chosen models in this study were two classification models, we created a binary variable. The binary variable will be defined as 1 or 0. The response variable is the stock return (logarithmic return) and if the returns are greater than zero, the binary variable will be 1. If the returns are smaller or equal to zero, the binary variable will be 0. The variable is based on whether the stock price has increased or decreased. The data frame was now finished and configured to make it possible to put the data frame into the machine learning models.

3.4.6.2 Training and Testing the Machine Learning Models

The next step was to set up the training data set and testing data set. As we were running the machine learning model for ten years. The training and testing data is moving for each year. Starting with predicting 2012, the training data was set to the five previous years, 2007 to the

end of 2011. Then the model tested its predictability in 2012. The same procedure was repeated for the following years up to 2021. Then we ran the CatBoost model and the logit regression for all these periods. As the models are running several iterations to train themselves and achieve the best possible prediction result, we set a code to get the best iteration possible. This means that the model was running several iterations and then provided us with the best configuration. The two models then provided us with prediction scores for each test period. Each test period and prediction is kept separate from each other. For instance, the prediction of 2012 is not related to the prediction of future years, which means that there is no risk of potential leakage of data. Thus, we run the prediction models several times, one for each quarter.

3.4.6.3 Feature Importance in Python

Lastly, we utilized the CatBoost library to get the feature importance of the CatBoost classification model. The feature importance is based on the training data and the scores given from the code are basically how important each feature is for the predictions and to what extent they influence the response variable, that is whether the stock price is predicted to move up or down.

4. Experimental Results

4.1 Results of the CatBoost Classifier

Results of the CatBoost Classifier are presented in Table 6. The presented values represent the prediction accuracy achieved by the model for each year. The accuracy score for each year consists of four predictions made (one in each quarter) for all the stocks included in the sample that were active during the period. In total, this means that each accuracy score is based on approximately 1700 standalone predictions of directional stock price movements. The highest accuracy of correct predictions was achieved in 2015, with a score of 60.46%. This prediction was based on what the CatBoost Model had learned during the most recent five years prior to the prediction. The lowest score was in 2016 at 52.32%. Predictions made during this year and these quarters were based on data from 2011 until 2016. The results suggest that in all ten years, the CatBoost model performed better than what would be the expected result by an uninformed investor who is blindly guessing (50%). The average accuracy score over the entire period was 56.71%.

Prediction Accuracy	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Avg	SD
Catboost Classifier	54,65%	57,47%	58,14%	60,46%	52,32%	53,48%	58,13%	57,47%	58,13%	56,88%	56,71%	2,48%
Training Data	2007-2011	2008-2012	2009-2013	2010-2014	2011-2015	2012-2016	2013-2017	2014-2018	2015-2019	2016-2020		

Table 6. CatBoost Classifier Prediction Accuracy

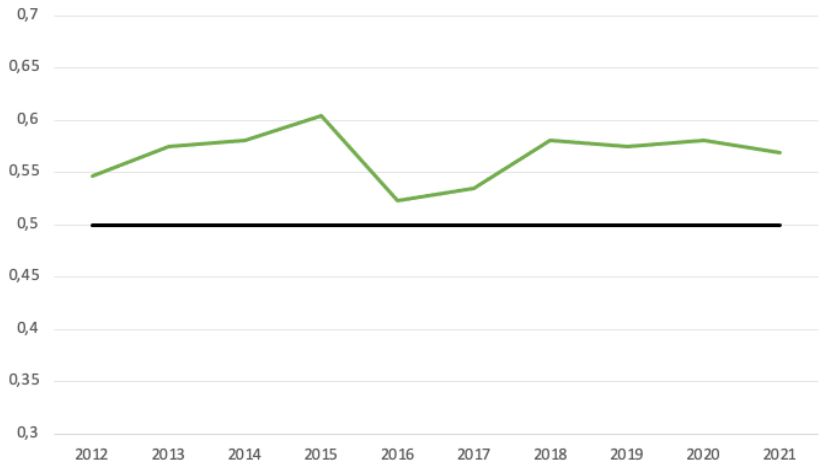

Figure 3. CatBoost Classifier Prediction Accuracy

As visualized in Figure 3, neither a positive nor negative trend in the accuracy scores is observable. The change in accuracy seems to be random between each year. However, the results are relatively stable with a low spread of values. The standard deviation of the accuracy scores is 2.48%. The figure further shows that the model consistently achieves accuracy scores above 50%, which in the diagram is represented by a constant dark line.

4.2 Results of the Logit Regression

The accuracy scores from the logit regression range between 46.51% and 64.36% during the ten-year period. The worst performing year was 2017 while the best performing year was 2019. There were two years where accuracy scores were above 60%, but also two years where accuracy was below 50%. As presented in Table 7 and further displayed in Figure 4, the accuracy scores were drastically different between the years with a standard deviation of 5.73%. The ten years, or 40 quarters, of predictions and accuracy scores, resulted in an average accuracy of 54.57%. This means that the model successfully achieved an accuracy score above 50% looking at the time period as a whole.

Prediction Accuracy	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Avg	SD
Logit Regression	52,32%	55,17%	59,30%	48,83%	51,16%	46,51%	61,63%	64,36%	51,16%	55,23%	54,57%	5,73%
Training Data	2007-2011	2008-2012	2009-2013	2010-2014	2011-2015	2012-2016	2013-2017	2014-2018	2015-2019	2016-2020		

Table 7. Logistic Regression Prediction Accuracy

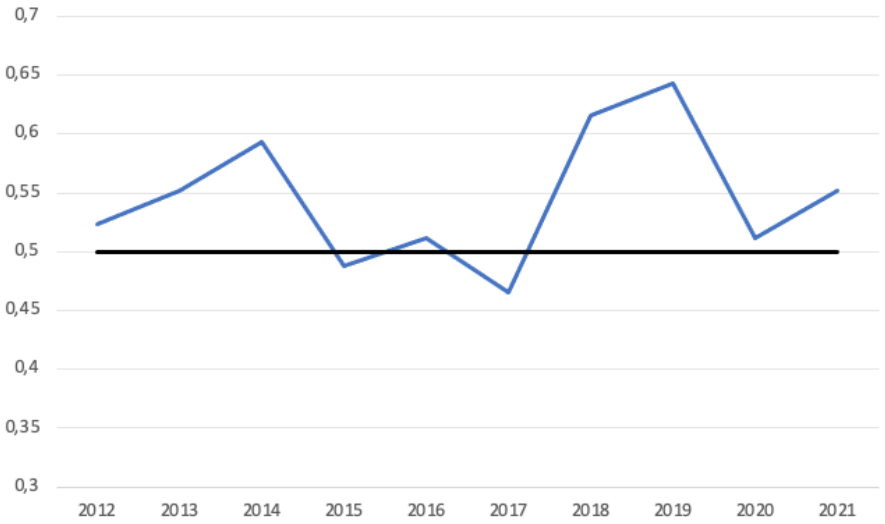

Figure 4. Logit Regression Prediction Accuracy

4.3 Results of Feature Importance

The average feature importance in the CatBoost classification for each of the 20 different predictor variables that have been considered is visualized in Figure 5. The bars display the most important features for the prediction from the left side and the least important features to the right side of the bar chart. The results do not show a clear ordering between the different fundamental categories. However, it shows five risk and growth proxies among the ten most important predictor variables. Book growth is by far the most important predictor variable and has a score of 22 of 100. The profitability proxies are the second most important fundamental category with four proxies among the ten most important predictor variables. Among the profitability proxies, ROA is the most important followed by ROE. Investment proxies and financing proxies are approximately equally important and share the third place in the importance of predictability. However, cash-to-assets is by far the most important of these two categories and places itself as the second most important feature for the prediction. The other proxies within these two categories all fall within the ten least important features for the prediction.

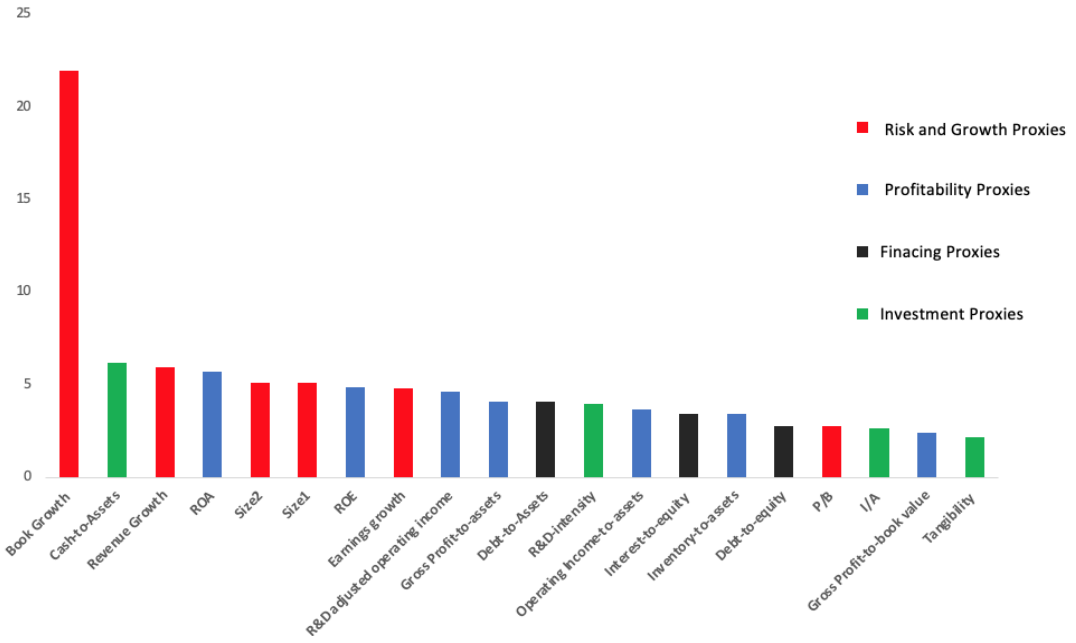

Figure 5. Average Feature Importance of the Predictor Variables

In the following four diagrams (Figure 6a-d.), it is visualized how each feature's importance has increased or decreased in importance over the ten years of prediction. As shown by the first diagram (Figure 6a.), book growth has always been the most important feature during these ten years. In the profitability proxies (6b), ROA has been the most important feature during the period except for 2017 when ROE and R&D adjusted operating income was more important. Among the investment proxies (6c), cash-to-assets has been the most important proxy during the period. However, the inventory-to-assets proxy has moved from a score of two up to a score of eight which is a noteworthy increase in the feature importance. Among the financing proxies (6d), debt-to-assets has been the most important proxy for the majority of the years.

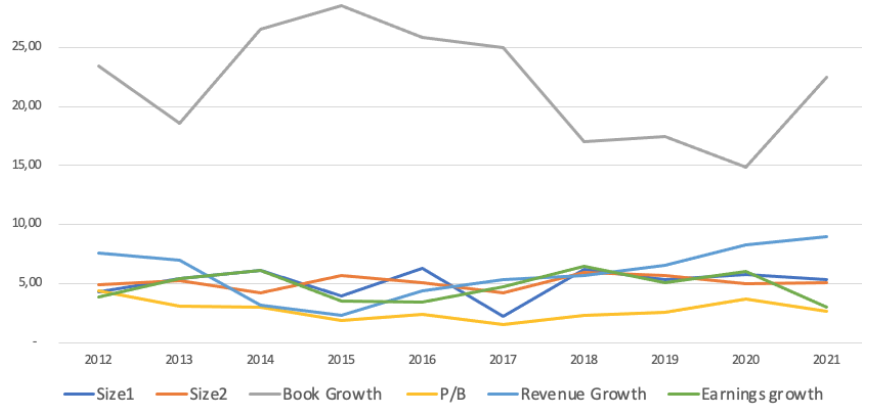

Figure 6a. Predictor Variables Importance Over 10 Years of Prediction: Risk and Growth Proxies

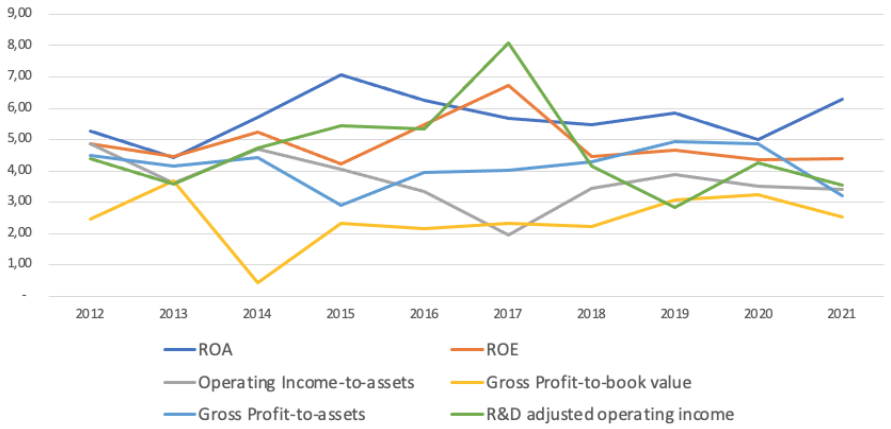

Figure 6b. Predictor Variables Importance Over 10 Years of Prediction: Profitability Proxies

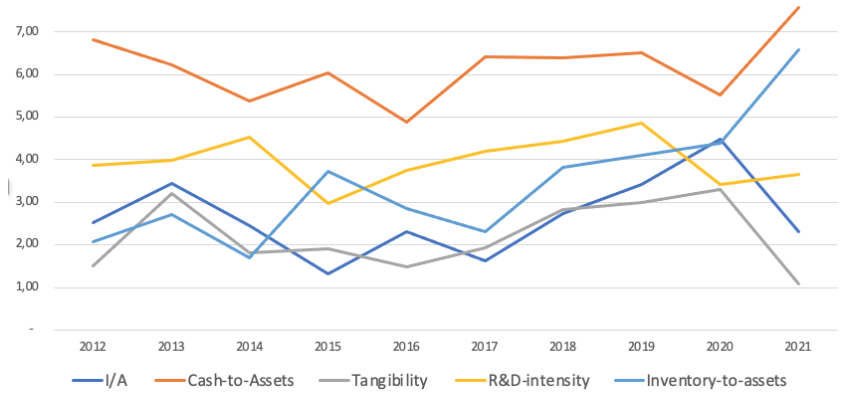

Figure 6c. Predictor Variables Importance Over 10 Years of Prediction: Investment Proxies

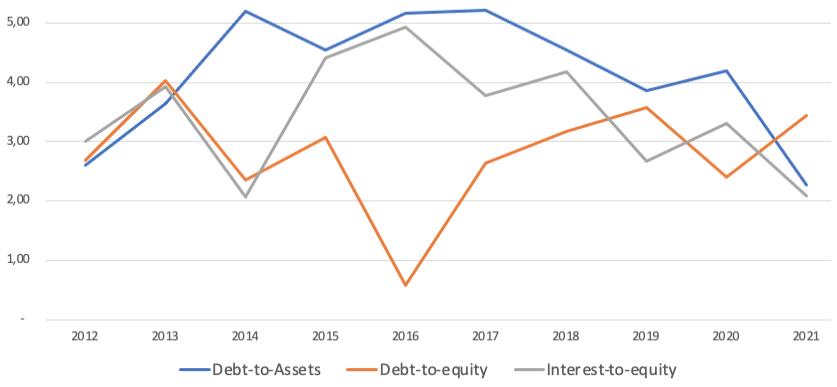

Figure 6d. Predictor Variables Importance Over 10 Years of Prediction: Financing Proxies

5. Discussion and Analysis of the Results

5.1 Can a machine learning classification model be utilized to predict whether the stock price will go up or down based on accounting data features?

The results of the predictions made by both machine learning models scored an overall accuracy of close to 57% and 55% respectively, between the years 2012 and 2021. Thus, both models achieve a total accuracy above 50%, which an uninformed investor theoretically would have achieved in the long run. As previously stated, our method ensures that each prediction is kept separate from each other and that no data is overlapping from one prediction to another. The results in 2020 and 2021 were particularly impressive since we anticipated stock price movements during these years to be heavily influenced by external factors. Therefore, the model might have been trained with accounting data consisting of numbers and measures that were largely uncorrelated with the current stock price movement, causing inaccurate predictions. For instance, stock prices could suddenly drop considerably when the price would have gone up or not react at all during more “normal” circumstances. Continuing on the concept of abnormal market conditions, the machine learning models did also achieve accuracy scores above 50% in 2012 and 2013, in which the training data partially consisted of data from the financial crisis in 2007-2008. The presence of these unnatural market conditions further triggers a discussion about whether there really exists something that can be labeled as a "normal” market condition. As motivated by Holthausen and Larcker (1992), the market is constantly evolving and static models will not work for a longer period. A successful model or trading strategy, therefore, needs to be adaptive to the changes in the market or it will be outdated in short order.

A complete error-free prediction should not be expected and stock price movements are still largely due to external factors (Rapach & Zhou, 2013; Vijh et al., 2020). Nonetheless, the results of the study are in line with previous research suggesting that one can use machine learning methods to increase the likelihood of accurately predicting stock price movements (Daoud, 2019; Prokhorenkova et al., 2018; Gu et al., 2020; Vijh et al., 2020; Kotsiantis, et al., 2007). Besides motivating the usefulness of machine learning and particularly the gradient boosting tree model and CatBoost library, the results also provide evidence supporting research arguing for the usefulness of accounting data, as previously introduced by Ball and

Brown (1968), but also supported by Basu (1983), Rosenberg et al. (1985) Cooper et al. (2008), Dimitropoulos and Asteriou (2009), and Schrimpf (2010).

For one of the machine learning models, the logistic regression model, the accuracy scores in two years were lower than 50%. This questions the reliability of this particular model. However, it is not too surprising because of the unpredictable nature of the stock market (Fama, 1970; Rapach & Zhou, 2013). Although the scores in these years were disappointing for the logistic regression model, the accuracy scores achieved in all other years for both models motivate the usefulness and opportunities that machine learning methods may bring. Our results for both models are therefore in line with Rapach and Zhou (2013) and Gu et al. (2020) who argue that users of forecasting models will outperform investors who do not rely on these methods.

If accounting data can be used to outperform the market, the semi-strong form of the efficient market hypothesis does not hold. That is because publicly available information is used, which according to the theory should be instantly reflected in the stock prices. This would remove any potential advantage an informed investor would have over an uninformed investor (Fama, 1970). Since our result from both different machine learning models suggests that the likelihood of making correct decisions do improve by on average almost 7% and 5% respectively, this goes against the semi-strong form of EMH. Completely bust or disregarding this well established hypothesis based on these results would be senseless, but we do however believe our evidence to be sufficient to at least question this very strict argument of public information not being advantageous to use to outperform the market and investor competitors. Thus, we argue that a machine learning model certainly can be utilized to predict stock market movements by using accounting data as predictor variables.

5.2 Can a supervised machine learning classification model based on a newly introduced gradient boosting on decision trees algorithm be more accurate in its predictions than the traditional logistic regression model?

There are several things to note when comparing the results from the different approaches. For starters, the average accuracy scores for both models are above 50%, which suggests an improved accuracy over blindly guessing. However, the average score generated by the machine learning model based on a gradient boosting decision tree algorithm is higher than

the one for the logistic regression. The difference is approximately two percent, which in the environment of stock price prediction can be considered substantial where even a tiny percentage of better accuracy may result in large economic gains (Zheng & He, 2021). The biggest difference was in 2015, where the CatBoost model did its best performing year of above 60% accuracy while the logistic regression model even failed to reach 50%. This year the CatBoost model outperformed the logistic regression model by almost 12%. Also noteworthy, in CatBoost's worst performing year (2016) it still outperformed the logistic regression by one percent. Thus, this year in particular seems to have been difficult to predict.

Interestingly there was a noticeable difference in consistency between the two models. While the logistic regression peaked with the highest accuracy scores over the ten years, this model also had the lowest scores of the two. The range and also the standard deviation were higher for the logistic regression model. The standard deviation of accuracy scores generated by the logistic model was almost twice as high as the CatBoost model. In 2021 as well as 2020, the pandemic situation included unforeseen external factors severely affecting the stock market. As a result, heavily influential information was outside the training data which the models based their predictions on. Surprisingly, there was no noticeable difference in accuracy scores for these years. Nonetheless, it can be noted that the machine learning model performed better than the logistic regression during both these years, perhaps due to learning and re-calibrating the model and the importance of each separate predictor variable, while the logistic regression did not do it as successfully.

Looking at the results, there are arguments to be made for both models. If one prefers a riskier model to follow with higher highs, then the logistic model may be preferred. However, the risk-averse investor should prefer the CatBoost classifier over the logistic regression model, where the user gets more consistent accuracy scores that are above 50%. The CatBoost model built on a gradient boosting decision tree algorithm achieved higher accuracy than the logistic regression model in 7 out of 10 years. In addition, the CatBoost model also had a higher score over the entire period. The continuous learning further allows for a more adaptive model, which is better suited for the volatile and ever-changing stock market environment (Holthausen & Larcker, 1992). For this, the CatBoost model demonstrated better capabilities for adaptation compared to the logistic regression model. Ultimately, in line with Gu et al. (2020), we can confirm that a machine learning model based on gradient boosting

decision trees definitely can be more accurate in its predictions compared to the logistic regression model.

5.3 What are the most important fundamental data and performance measure features of the predictive model?

There are several interesting outcomes from the average feature importance. The risk and growth proxies were the most important predictor variables in the predictions followed by the profitability proxies. This can be compared to the results by Starica and Marton (2021) and their average feature importance. The results from their report are pointing out risk and growth proxies as the most important variables, followed by financing and profitability variables. The authors do have a different research approach and are utilizing another machine learning model (random forest). However, the price-earning association approach is still within the topic which makes the results comparable to some extent. The results do have some similarities. The risk and growth variables are the most important in our study which is in line with Startica and Marton (2021), followed by the profitability proxies. Further, our results are in line with Schrimpf (2010) who concluded that profitability ratios have a close relationship with the stock price movements. Our results are also partly supported by Dimitropoulos and Asteriou (2009) where the article findings argue that profitability ratios are the most important determinants of predicting stock returns. Our results suggest that risk and growth ratios are the most important indicators, however, profitability ratios were identified as the second most important proxies for the prediction of stock price movement.

The financing proxies in our case show low importance, which contradicts the results by Starica and Marton (2021). However, our result is supported by Dimitropoulos and Asteriou (2009) who have a research approach more akin to ours. The findings of this study suggest that leverage is not a key variable for predicting stock returns, instead the authors conclude that the most productive firms are rewarded with higher stock returns. Accordingly, the productivity variables and profitability variables were the key variables for prediction in our study. The leverage variables included in our model were within the ten least important variables for predicting stock price movements.

Three out of four variables among the investment proxies were among the ten least important variables for predicting stock price movements. This is in line with the result by Starica and

Marton (2021). However, cash-to-assets was in fact the second most important predictor variable during these ten years of predictions. This outcome is surprising as none of the previously mentioned authors have pointed out cash-to-assets as a variable with that magnitude of importance in predicting stock movements.

6. Conclusions, Limitations, and Future Research

6.1 Conclusions

Judging by the results of our study, a machine learning method is a useful method to predict stock price movements by using accounting data. Therefore, we reject our null hypothesis stating that performance measures and accounting data cannot be used to predict stock price movements. The machine learning method by a gradient boosting regression tree model did also outperform the traditional logistic regression model. The difference between the models may not seem large at first glance, but in this environment, every sole percentage of improved accuracy is highly significant for the investor. Our result does not discourage the traditional logistic regression model as this model also provides an accuracy that is above 50%, which would be obtainable by blindly guessing or flipping a coin. However, if one wishes to make the most accurate prediction, our result suggests that the machine learning method through the CatBoost Classifier is the more appropriate method of the two to use.

Regarding the third research question, the feature importance is a useful tool in order to understand which predictor variable contributes the most to the prediction. Judging by the results of the feature importance, risk and growth proxies were the most prominent predictors followed by the profitability proxies. Looking at the specific variables, book growth was the most important feature followed by cash-to-assets (2), revenue growth (3), return on assets (4), and size (5,6). Feature importance is not a statistical inference and the method can only tell us how important a feature is for the prediction. The results from the feature importance analysis do not say anything about the relationship itself, whether it is linear, non-linear, quadratic, or the magnitude of the predictor effect. However, the information gathered from the feature importance is still useful since it provides details about which variables the model is primarily basing its prediction on. In addition, it provides guidance towards which variables that could be further investigated if one wants to study the relationship closer.

6.2 Limitations

There are a couple of limitations related to this study. Firstly, there are a few limitations within the data sample. The result is heavily dependent on the data chosen, for instance, which market and time frame that is investigated. Besides the time frame chosen for data collection and machine learning, the result may also differ depending on the forecasted horizon. This study focuses on quarterly data and tries to predict a positive or negative return in the following quarter. However, some investors may prefer to have a longer forecasting horizon, for instance, half a year, a year, or multiple years. For these purposes, it could also be interesting to include other variables that historically have proven useful for forecasting purposes, particularly dividends and other pay-out proxies, which we did not include in our data sample consisting of quarterly data. In this case, the importance and predictive capacity of individual variables may change. Limitations within the data also include the predictor variables. Even though we base our variables on previous research, there might still be important variables that are left out. There could be important variables that have not been investigated properly simply because their importance and relevance are largely unexpected.

Secondly, the ever-changing nature of the market. A model that functions well today may be irrelevant already after a few years. We tried to manage this issue by using a model that is not completely static but is based on the most recent years to determine the variables that are important. This way the model should in theory not become outdated, however drastic and unforeseen changes in the market may still occur and reduce the accuracy of predictions made. In addition, a proven and functioning model or trading strategy that outperforms the market is sure to be mimicked by competitors, which will eliminate potential trading advantages it might bring.

Thirdly, the technical aspect of our study. It should be noted that the authors of this paper are not programmers. Our knowledge and education are related to finance and accounting. Thus, our expertise in the technical parts of machine learning is limited. However, we believe our knowledge to be adequate to perform this study as we did not develop the code from scratch, but rather applied an existing machine learning library. Our study is also limited to only two machine learning models, other alternative models may yield different results.

Fourthly, the simplistic output of a binary variable. This limits the usefulness of the model for an investor. Simply knowing that a stock price movement is predicted to either be positive or negative might not be enough for the investor to outperform the market. The model does not indicate how much of an increase or decrease in stock price that is expected. Thus, even if the majority of the predictions are correct, the result for the investor might still be negative if the largest stock price movements happen at the incorrect investment actions. To make the model more useful, some indication of how big of a movement that is predicted would be of great assistance for the user. In addition, one could improve the model further and make it more tailored to the investor by adding a confidence level. Then the investor could choose to act and follow the model only if a certain confidence level is reached, which may result in better trading results.

6.3 Future Research

This study has raised evidence towards the possibility of predicting stock price movements based on accounting data and performance measures by utilizing machine learning models built on the gradient boosting decision tree algorithm and the logistic regression model. Future research should aim at developing this knowledge and machine learning methods further to make it more practically usable for the investor. This could be done by solving mentioned limitations of this study by adding a confidence level for the predicted output together with an estimation of how big of a movement is expected. With these additions to the model, it would be interesting to see the results an investor would be able to achieve and compare it to the market index or a simple buy-and-hold strategy of selected stocks. One could also construct a similar study to ours but outside the US market. Thereby providing further evidence in the area of machine learning and stock price prediction by using accounting data.

As this study solely includes two machine learning approaches in the logistic regression and the CatBoost Classifier based on the gradient boosting decision tree algorithm, future studies could compare results generated by different learning models. Research made on stock price prediction by using the CatBoost machine learning library, in particular, is extremely limited and it would be intriguing to test how well it would perform compared to other machine learning algorithms. Both against other machine learning methods in neural networks and random forests, but also compared to other algorithms or libraries based on the gradient

boosting tree. Besides investigating why the accuracy scores might differ, future research could also focus on the predictor variables. For instance, how their ranked importance differs across models and why that is the case.

References

Algharaballi, E., and Albuloushi, S. (2008) Evaluating the specification and power of discretionary accruals models in Kuwait. *Journal of Derivatives and Hedge Funds* 14, 251–264. https://doi.org/10.1057/jdhf.2008.23

Armstrong, H., (2015). Machines that learn in the wild: Machine learning capabilities, limitations and implications. Nesta. Retrieved April 6, from https://media.nesta.org.uk/documents/machines_that_learn_in_the_wild.pdf

Ball, R., and Brown, P. (1968). An Empirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research*. 6, 159–78.

Ball, R., and Nikolaev, V. V. (2022) On Earnings and Cash Flows as Predictors of Future Cash Flows. *Journal of Accounting & Economics* (JAE). Elsevier, vol. 73(1).

Ball, R., Gerakos, J., Linnainmaa, J. T., and Nikolaev, V. (2016). Accruals, cash flows, and operating profitability in the cross section of stock returns. *Journal of Financial Economics*, 121(1):28 – 45.

Basu, S. (1983). The relationship between earnings yield, market value and return for NYSE common stocks: Further evidence. *Journal of Financial Economics*. 12(1), 129-156

Beaver, W. H. and Ryan, S. G. (2005). Conditional and unconditional conservatism: Concepts and modeling. *Review of Accounting Studies*, 10(2):269–309.

Bertomeu, J. (2020). Machine Learning Improves Accounting: Discussion, Implementation and Research Opportunities. *Review of Accounting Studies* 25 (3):1135-1155.

Biddle, G., Chen, P., and Zhang, G. (2001). When capital follows profitability: Non-linear residual income dynamics. *Review of Accounting Studies*, 6(2):229–265.

Billiau, S. (2021). Importance for ML Interpretability. Retrieved: April 24, 2022, from: https://towardsdatascience.com/from-scratch-permutation-feature-importance-for-ml-interpretability-b60f7d5d1fe9

Brooks, C. (2014). Introductory Econometrics for Finance. Third Edition. *Cambridge University Press*, UK.

Chan, S., Reddy, V., Myers, B., Thibodeaux, Q., Brownstone, N., and Liao, W. (2020). Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. *Dermatology and Therapy* 10, 365–386.

CatBoost (n.d.) Feature Importance. Retrieved April 26, 2022, from: https://catboost.ai/en/docs/concepts/fstr#prediction-diff

Chen, H., De, P., Hu, Y. J., and Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies.* 27(5), 1367-1403.

Chen, L., Novy-Marx, R., and Zhang, L. (2011). An alternative three-factor model. Technical report, Working Paper (Washington University in St. Louis).

Cooper, M. J., Gulen, H., and Schill, M. J. (2008). Asset growth and the cross‐section of stock returns. *Journal of Finance*. 63(4), 1609-1651.

Daoud, E.A. (2019). Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset.

Dechow, P. M. (1994). Accounting earnings and cash flows as measures of firm performance: The role of accounting accruals. *Journal of Accounting and Economics*. 18(1), 3–42.

Dechow, P. M., and Dichev, I. D. (2002). The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*. 77(s-1), 35–59.

Delcey, T. (2019). Samuelson vs Fama on the Efficient Market Hypothesis: The Point of View of Expertise. *Œconomia - History/Methodology/Philosophy, NecPlus/Association) Economia*, 2019, Varia, 9 (1), pp.37-58. ffhal-01618347v2f

Dimitropoulos, P., and Asteriou, D. (2009). The value relevance of financial statements and their impact on stock prices. *Managerial Auditing Journal*. 24(3), 248-265.

Dzikevičius, A., and Šaranda, S. (2011). Can financial ratios help to forecast stock prices?. *Journal of security and sustainability issues*. 1, 147-157.

Fama, E. F. (1965a). The Behavior of Stock-Market Prices. *The Journal of Business* 38 (1): 34‑105.

Fama, E. F. (1965b). Random Walks in Stock Market Prices. *Selected Papers of the Graduate School of Business*, University of Chicago, reprinted in the *Financial Analysts Journal* (September - October 1965), *The Analysts Journal*, London (1966), *The Institutional Investor* , 1968., October, 55‑59.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance.* 25 (2): 383‑417.

Fama, E., and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*.

Fama, E. F., and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1 – 22.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies*, Volume 33, Issue 5, May 2020, Pages 2223–2273, https://doi.org/10.1093/rfs/hhaa009

Holthausen, R. W., and Larcker. D. F. (1992). The prediction of stock returns using financial statement information. *Journal of Accounting and Economics*. Volume 15, Issues 2–3, June–September 1992, Pages 373-411

Holthausen, R. W., and Watts, R. L. (2001). The relevance of the value-relevance literature for financial accounting standard setting. *Journal of Accounting and Economics*. 31: 3–75.

IBM Cloud Education. (2020). Machine learning: What is machine learning?
Retrieved: Februari 2, 2022, from https://www.ibm.com/cloud/learn/machine-learning

Javapoint. (n.d.). Classification algorithm in machine learning. Retrieved: Mars, 5, 2022, from: https://www.javatpoint.com/classification-algorithm-in-machine-learning

Jegadeesh, N., and Livnat, J. (2006). Revenue surprises and stock returns. *Journal of Accounting & Economics*. 41(1), 147-171.

Jegadeesh, N., and Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance* (New York). 48(1), 65-91.

Kim, S. H., and Kim, D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior & Organization*. 107, 708-729.

Kothari, S. P., and Shanken, J. (2003). Time-series coefficient variation in value-relevance regressions: A discussion of Core, Guay, and Van Buskirk and new evidence. *Journal of Accounting and Economics*, 34:69–87.

Kothari, S. P., and Wasley, C. (2019). Commemorating the Fifty-Year Anniversary of Ball and Brown (1968): The Evolution of Capital Market Research over the Past Fifty Years. (July 5, 2019). Available at SSRN: https://ssrn.com/abstract=3417149

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics*. 74(2), 209-235.

Liu, J., and Thomas, J. (2000). Stock returns and accounting earnings. *Journal of Accounting Research*, 38(1):71–101.

Malik. M. (2020). A Hierarchy of Limitations in Machine Learning. Retrieved: April 10, 2022, from https://arxiv.org/pdf/2002.05193.pdf

Mitchell, M. (n.d.). Selecting the correct predictive modeling technique. Retrieved: January 9, 2022, from https://towardsdatascience.com/selecting-the-correct-predictive-modeling-technique-ba459c370d59

Multpl. (n.d.). S&P 500 PE Ratio by Year. Multpl. Retrieved: January 9, 2022, from https://www.multpl.com/s-p-500-pe-ratio/table/by-year

Na, H., and Kim, S. (2021). Predicting stock prices based on informed traders' activities using deep neural networks. *Economics Letters*. 204, 109917.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, 108(1):1 – 28.

Open Data Science. (2018). What Overfitting is and How to Fix it. *Published in Predict*. Retrieved: March 7, 2022 from https://medium.com/predict/what-overfitting-is-and-how-to-fix-it-887da4bf2cba

Prazak, T., and Stavarek, D. (2018). Importance of financial ratios for predicting stock price trends: Evidence from the Visegrad Group. *International Journal of Trade and Global Markets*. 11(4), 293-305.

Pradeepkumar, D. and Ravi, V. (2017). Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network. *Applied Soft Computing*, 58:35–52

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Python. (n.d.). What is python?. Retrieved: February 14, 2022, from https://www.python.org/doc/essays/blurb/

Rapach, D., and Zhou, G. (2013). Chapter 6 - Forecasting Stock Returns. *Handbook of Economic Forecasting.* Vol 2, Part A, 2013, Pages 328-383

Ray, S. (2017). CatBoost: A machine learning library to handle categorical (CAT) data automatically. Retrieved: February 14, from https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/#h2_5

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. *Journal of Banking & Financ*e. 84, 25-40.

Rosenberg, B., Reid, K., and Lanstein, R. (1985). Persuasive evidence of market inefficiency. *The Journal of Portfolio Management*. 11(3), 9-16.

Roychowdhury, S. and Watts, R. L. (2007). Asymmetric timeliness of earnings, market-to-book and conservatism in financial reporting. *Journal of Accounting and Economics*, 44(1):2 – 31.

Samuelson, P. A. (1965). Rational Theory of Warrant Pricing. *Industrial Management Review.* 6 (2): 13-39.

Schmidt, A. B. (2011) Financial Markets and Trading: An Introduction to Market Microstructure and Trading Strategies. *John Wiley & Sons*. Ch.7. p. 65-74.

Schrimpf, A. (2010). International stock return predictability under model uncertainty. *Journal of International Money and Finance*. 29(7), 1256-1282.

Shen, J., and Shafiq, M.O. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of Big Data* 7, 66 (2020). https://doi.org/10.1186/s40537-020-00333-6

Starica, C. and Marton, J. P. (2021). A rigorous research design for the assessment of the price-earnings relation.

Vijh, M., Chandola, D., Tikkiwal, V. A., and Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia Computer Science*. 167, pp. 599-606

Zheng, L., and He, H. (2021). Share price prediction of aerospace relevant companies with recurrent neural networks based on PCA. *Expert Systems with Applications*. 183, 115384.

Appendix

	mean	std	min	25%	50%	75%	max
ROA	0,0189	0,0242	-0,2778	0,0097	0,0186	0,0292	0,1987
ROE	0,0618	0,6792	-10,5882	0,0229	0,0452	0,0763	14,5897
Operating income-to-assets	0,0398	0,0216	-0,0693	0,0272	0,0366	0,0491	0,1733
Gross profit-to-book value	0,3594	1,9839	-31,2585	0,1450	0,2163	0,3541	45,9895
Gross profit-to-assets	0,0956	0,0506	-0,0599	0,0606	0,0849	0,1178	0,4129
R&D adjusted Operating income-to-assets	0,0509	0,0256	-0,0681	0,0339	0,0479	0,0641	0,1804
DtA	0,2540	0,1487	0,0000	0,1471	0,2335	0,3353	1,0151
DtE	1,1263	8,5106	-142,4773	0,2600	0,5728	1,1163	193,7047
ItE	0,0142	0,1120	-0,8357	0,0026	0,0056	0,0126	2,5906
Size1	37728,3483	59540,6540	982,0670	6832,3595	16260,0000	39484,0000	551669,0000
Size2	10,1418	1,3733	6,0513	9,1648	9,9748	11,0954	14,6590
Revenue Growth	0,0125	0,1864	-0,6093	-0,0616	-0,0102	0,0496	2,0556
Earnings Growth	0,2509	8,7448	-64,0000	-0,3572	-0,0659	0,2247	282,0000
Book Growth	-0,0696	2,4097	-91,5625	-0,0436	-0,0134	0,0262	10,0000
P/B	6,7115	43,0078	-690,0175	2,4396	3,8255	6,0677	1127,7934
I/A	0,5782	0,3780	0,0270	0,2858	0,4628	0,8008	2,4177
Cash-to-assets	0,1502	0,1333	0,0024	0,0557	0,1129	0,2005	0,7338
Tangibility	0,2139	0,1671	0,0116	0,0866	0,1511	0,2877	0,7784
R&D-intensity	0,0111	0,0133	0,0000	0,0000	0,0085	0,0155	0,1282
Inventory-to-assets	0,1301	0,1192	0,0000	0,0429	0,0896	0,1891	0,5734
median shareprice	103,5404	115,8833	1,8400	42,3200	69,1700	123,2850	1736,5150
log returns	0,0195	0,1507	-1,7106	-0,0371	0,0276	0,0947	0,8191
positive_return	0,4997	0,5001	0,0000	0,0000	0,0000	1,0000	1,0000

Appendix. Descriptive Statistics