Master Thesis Business Analytics and Management

The effectiveness of combining XBRL reporting data with machine learning approaches for fundamental earnings movement forecasting

Jelto de Jong – 448942

448942jj@student.eur.nl

Coach: Dr. Dion Bongaerts

Co-reader: Yonghan Li

Preface

Acknowledgements

Before the research topic of this thesis is introduced, I would like to express my gratitude towards my thesis coach: dr. Dion Bongaerts. Not only did he provide me with continued assistance throughout the process of writing this thesis and supplied me with helpful input before and during our feedback sessions, but he also sparked my interest for conducting a financial analytics research by teaching the Financial Modelling and Fintech electives. I would furthermore like to thank my co-reader, Yonghan Li, for his input during our early-stage feedback session and dr. Ana Martinovici for her guidance to set-up and structure the GitHub repository linked to this thesis.

Executive Summary

This thesis proposes the combined application of machine-readable XBRL data and easy-to-reproduce machine learning tools, which are both available to retail and smaller institutional investors, to produce fundamental earnings movement predictions. By providing an understanding of how publicly available XBRL data can be obtained and tested, this study first of all shows that the XBRL data provides an accurate source for accounting information relative to traditional data sources. A transparent methodology is then provided that shows how fundamental earnings movements can best be predicted by applying specific machine learning algorithms on the XBRL data. The most suitable algorithms have then been applied on XBRL sample data, where the Random Forests and Gradient Boosting Machines models show to provide good one-year forward earnings movement predictions with an accuracy of 73.8% and 75.5%, respectively. Moreover, the Gradient Boosting Machines model is able to outperform professional analysts' earnings movement predictions when outcomes within the same set of test observations are compared. Besides showing that XBRL data is well suited as input for machine learning models to perform fundamental earnings movements forecasting, both in terms of error-metrics as well as when predictions are compared to professional analysts' forecasts, this study also shows that the performance of the models is consistent across fiscal years and firm sizes. Lastly, no sizable difference in predictive power is observed when the best performing model on the XBRL data, being the Gradient Boosting Machines algorithm, is applied on data obtained from a traditional source of accounting fundamentals.

Throughout the thesis, references are made to the https://github.com/Jelto48/MasterThesis repository on GitHub. The supplementary documents supplied here aim to provide a deeper understanding of the applied approach and contains the R files needed to replicate the results. This has been done to increase both the interpretability as well as the transparency of the research and is in line with the guiding principles of the Open Science Community Rotterdam.

Table of Contents

1. Introduction

With 2021 being said to be a breakthrough year for retail investors, triggered by the boredom of COVID-19 lockdowns, commission-free trading and the power of social media platforms, the need for more awareness regarding the fundamental drivers of security value has become visible (Martin and Wigglesworth, 2021). The recent GameStop share price surge has shown that everyday investors, inspired by online communities, can drive the market value of a loss-making chain to unrealistic heights (Smith and Wigglesworth, 2021) and shows proof of the lack of knowledge these retail investors have regarding the fundamental drivers of firm value (Mackenzie, 2021). Where often a more technical or noise-trading approach is taken by these investors to outperform the market, the performance of these approaches generally falls short for them in providing long-term profits (Barber and Odean, 2013). An alternative for these investors is to base their investment decisions on the fundamental drivers of value, which can be achieved by following professional analysts' recommendations or through self-estimation of these value drivers. For the latter approach, the limited availability of data and tools needed to do this has shown to be a major obstacle (Strampelli, 2018). Although some off-the-shelf methods for fundamental earnings forecasting exist, they mostly rely on complex approaches that are difficult to apply and interpret for the user. This indicates that there is a call for a transparent methodology that shows how fundamental earnings predictions can be made with publicly available data and tools. Moreover, no evidence is available whether such earnings forecasts can compete with professional analyst's predictions, being an alternative approach used by investors for fundamental investment decision making. This thesis therefore proposes the combined application of publicly available XBRL accounting data and easy-to-reproduce machine learning tools to provide earnings movement predictions that are useful for fundamental investment approaches. By testing this approach in terms of predictive power and the ability to outperform professional analysts' forecasts, the uncertainty that smaller investors face when competing with better-informed investors can be addressed.

1.1. Problem Background

In the last decade, developments in the availability of computer-readable accounting information have cleared the way for investors to apply new fundamental earnings forecasting techniques. The mandatory adoption of electronic financial reporting in the XBRL format for listed companies around the world is an example of such a development and has created advantages for investors in terms of information accessibility and readiness to apply machine learning (ML) forecasting approaches (Palas and Baranes, 2019; Wang and Seng, 2014). These advantages have however not yet been fully acknowledged by retail and smaller institutional investors, increasing the information barriers that limit their ability to compete with more sophisticated investors (Guo and Yu, 2020). A key problem for small-scale investors that want to perform fundamental earnings movement predictions themselves with their available data and tools has shown to be the limited availability of step-by-step procedures to do so (Strampelli, 2018). Although some literature shows the basic procedure of applying machine learning algorithms on accounting data for earnings forecasting (e.g., Hunt et al., 2019 and Anand et al., 2019), these studies mostly rely on 'black box' machine learning approaches that are difficult to interpret for the user, leaving smaller investors with questions regarding the actual drivers of earnings movements. This causes uncertainty for the investors related to the ability to compete with other market participants with more technical capacity and resources at their disposal and often results in them withholding in applying fundamental investment techniques (Seth et al., 2020). Furthermore, no evidence is provided in these studies whether the forecasts can compete with professional analysts' predictions, which is commonly used by investors as a substitute for self-assessed fundamental analysis (Guo and Yu, 2020). Since fundamental earnings forecasting can predict future earnings movements based on historical financial data, another hesitation for smaller investors to adopt fundamental approaches relates to the fact that not all company-specific information needed for an earnings prediction can be captured in this financial statement information (Monahan, 2017). Although the importance of non-financial information is slowly being acknowledged, the current accounting taxonomy is still lacking standards to capture all sources of value within a firm (Maama and Mkhinze, 2020).

1.2. Problem Statement and Research Questions

In this thesis, the discussed problems related to fundamental earnings forecasting will be addressed to clear the way for smaller investors to benefit from the easier and less costly access to XBRL accounting information combined with reproducible machine learning algorithms. This can be summarised into the following problem statement that will be fundamental for the thesis:

Retail and smaller institutional investors are often unaware of how the fundamental drivers of firm value can be estimated by using publicly available data and tools, which limits their ability to compete with more sophisticated investors.

The following research questions aim to help in solving this stated problem:

RQ1: *How accurate is XBRL data as a source for fundamental earnings forecasting compared to traditional data sources?*

RQ2: *How can XBRL data best be combined with machine learning approaches for earnings movement forecasting?*

RQ3: *To what extent can these earnings movement forecasts compete with professional analysts' predictions on a one-year forward time horizon?*

RQ4: *For which specific companies is improving the models most needed, and can this be related to the XBRL data-quality issues?*

1.3. Academic Relevance

The first academic contribution this thesis makes relates to the established fundamental earnings forecasting literature. Although this approach of forecasting a fundamental value driver of a firm has been around for decades, new technological developments provide room to further develop this area of finance (Monahan, 2017). In line with this, the thesis develops a methodology to combine publicly available accounting information with machine learning approaches similar to Hunt et al. (2019) and Anand et al. (2019), but by using models that are built on machine-readable XBRL data and that will be tested by comparing the outcomes to professional analysts' benchmarks. This not only adds valuable insights to the literature regarding the ability of easy to reproduce forecasting methods to outperform other forecasts but also assists in overcoming the knowledge gap that smaller investors with limited technical capacity and resources at their disposal face. As stated by Talwar et al. (2021), with the growing share of retail investors being active in capital markets, it will be important to monitor and test the possible approaches behind their investment decisions.

This thesis also adds a fair share to the XBRL literature, especially to the application possibilities of this computer-readable source of accounting information. By showing if this accounting information format can be used as a suitable input for investment decision making, which is one of the key reasons the reporting language has been developed, the importance of adopting this source of information by global companies and regulators will be evaluated. This is important since Guo and Yu (2020) show evidence that based on the SEC's server log, where the XBRL data can be found, small-scale investors do not show to be downloading the data yet. They also highlight the fact that this might be related to a lack of knowledge by these investors to derive information from the raw XBRL data for investment purposes. Besides this, the thesis also aims to cast further light on the importance of the quality of XBRL reports by comparing the data to traditional data sources for fundamental earnings forecasting. Although a wide range of standards has been set in place to ensure the XBRL data quality, there is still room for large improvements in this area which is highlighted by Birt et al. (2017).

From a more data analytical perspective, the comparison and application of different machine learning approaches for earnings forecasting will be of additional value to the established literature. As far as it is known, this will be the first study to compare and synthesise different machine learning approaches on fundamental accounting information, where the perspective of the increasingly dominant small-scale investor is taken in terms of data input and computational resources. Hunt et al. (2019) have shown the success of applying the Logistic regression and Random Forests models to predict earnings movements based on Compustat data, which is often not available to investing individuals. Moreover, there is a significant time interval between the publishing of a financial report by a company and the availability of the included fundamentals within Compustat (Guo and Yu, 2020). Although an out-of-sample prediction accuracy of respectively 62.3% and 76.8% is shown by Hunt et al. (2019) based on 60 predictive variables, they highlight the fact that both the use of more accounting variables as well as the deployment of other non-parametric machine learning models might be beneficial.

This study also allows scholars and academics to understand the trade-offs between different types of machine learning approaches for earnings forecasting purposes. By comparing the earnings movement predictions to professional analysts' consensus forecasts, an approach applied by Babii et al. (2020) to benchmark regularised panel data regressions, insights are added to the literature regarding the relative performance of machine learning models compared to more sophisticated earnings forecasts. Finally, by emphasising the characteristics of the correctly and incorrectly predicted earnings movements in the developed models, this thesis aims to provide a deeper understanding of the potential sources of these forecasting errors. Since Nichols et al. (2017) discovered that forecasting approaches based on accounting information tend to fall short in fully predicting earnings movements for specific types of firms, it will be relevant to investigate if this also holds if the predictions are obtained from machine learning models.

1.4. Managerial Relevance

The managerial relevance of this thesis is manifold. Firstly, this study aims to show managers the benefits of utilising the publicly available and machine-readable XBRL accounting information. It aims to show how well this data can be combined with machine learning approaches to conduct fundamental earnings movement forecasts, an approach that is often overlooked by retail and small-scale institutional investors (Guo and Yu, 2020). Besides this, the provided step-by-step procedure to perform earnings movement predictions with readily available data and tools can be beneficial to adopt in practice, dependent on the outcomes. For the reason that this study also takes the interpretability of the machine learning approaches into account, the problems surrounding 'black box' approaches for smaller investors can also partly be resolved. The comparison of the established models with the forecasts of professional analysts will show how well the provided approach performs relative to a benchmark and can provide managers with far better insights about performance than just the prediction metrics of the model.

Despite the fact that mainly smaller investors are addressed in this thesis, it does not take away the benefits that other practitioners might perceive based on the outcomes. For example, with the rise of low-cost investing platforms that sometimes provide automated advice on portfolio construction (Beioly, 2019), the developed approach might be a valuable source of fundamental earnings forecasting information for such platforms. Furthermore, the results can provide other more advanced investors with an alternative approach to fundamental earnings forecasting, which can potentially result in cost reduction or more accurate predictions.

1.5. Research Approach

The next chapters have been constructed in the following way. In the first part of the literature review, the usage and relative benefits of fundamental analysis and earnings forecasting to derive company value will be discussed. Moreover, the debate in the literature why retail and smaller institutional investors do not tend to incorporate this approach in their investment decisions will be elaborated on, to show what areas need to be improved to overcome the barriers they face. The development and benefits of using the publicly available and machine-readable XBRL information standard for fundamental analysis will then be covered, which provides a bridge to various machine learning algorithms that can be deployed on the XBRL data. Different algorithms that can be used for fundamental earnings forecasting will be covered in the second part of the literature review, combined with the findings in the literature regarding different optimisation techniques such as cross-validation and regularisation. The literature review will end with a synthesis of the discussed machine learning algorithms, that shows the benefits and drawbacks of usage in combination with XBRL data to produce fundamental earnings movement forecasts.

By using the insights obtained from this synthesis, the data section will first of all discuss the development of a web-scraping algorithm that has been used to obtain the XBRL data. After this discussion, the relevant financial statement items can be compared to the known financial values obtained from Compustat, to determine the quality of the scaped data and answer the first research question. Next, the methodology section will discuss the implementation of the most suitable algorithms for fundamental earnings movement forecasting based on the scraped XBRL data and provides a description of how to train, validate and test the models. By doing this, the second research question regarding how XBRL data can best be combined with machine learning approaches will be addressed.

The results and discussion section will be used to answer the third and fourth research questions. First of all, the applied machine learning models will be tested and compared based on different test statistics to be able to show which machine learning approach can best be used for one-year forward earnings movement forecasting**.** The obtained forecasts will then be compared to professional analysts' forecasts to see how well the models perform relative to a benchmark. Furthermore, the outcomes of the third research question will be used to answer for which specific companies the models perform relatively well and poor. The patterns derived from the correctly and incorrectly classified earnings movements within the models and the professional analysts' predictions will be used to show what models perform best in which specific instances. Based on these outcomes, a deeper understanding can be obtained regarding in which instances specific models perform better or can outperform professional analysts' forecasts. Moreover, by comparing the outcomes of the best performing XBRL data machine learning model to the same model with accounting fundamentals from a traditional source, it can be concluded if improved XBRL data quality can increase the predictive performance of the model.

Finally, after discussing the limitations of the study, where additional directions related to this thesis will also be pointed out, the main findings of the thesis will be summarised in the conclusion section.

2. Literature Review

In this chapter, the literature covering various relevant topics for this thesis will be elaborated upon. First, the difference between basing investment decisions on rules-based quantitative analysis or on comprehensive fundamental analysis of the underlying securities to capture the economic reality will be discussed by using the leading paper by Sloan (2019). This will then be linked to literature that covers the importance of earnings forecasting for fundamental analysis and the barriers for retail and smaller institutional investors that this approach is surrounded by. A possible way to overcome these barriers will then be proposed by casting light on the literature covering the publicly available and machine-readable XBRL data, followed by the findings in the literature regarding various machine learning algorithms that could be suitable to apply on this specific data for fundamental earnings forecasting purposes. Furthermore, various methods to improve the machine learning algorithms will be covered based on the findings in the established literature. This results in a synthesis of the discussed machine learning algorithms to determine the most suitable models in terms of predictive power and transparency for the user. The chapter will end with a short elaboration on the expected shortcomings of fundamental earnings forecasting methods that have been found in previous research.

2.1. Quantitative and Fundamental Investment Strategies

In a recent paper by Sloan (2019), the finding that quantitative investment strategies dominate the academic and practical financial worlds is discussed extensively. Quantitative investment strategies are referred to as 'rules-based' strategies that can be used to quickly select securities from an investment universe, based on pre-defined technical rules such as only investing in stocks with a relatively high book-to-market ratio. The benefits of such an approach are that behavioural biases of human investors can be overcome, and the historical performance of rule-based strategies can objectively be verified by using back-tests. Besides being actively deployed by large investment management firms, small-scale investors also increasingly base their investment decisions on these quantitative investment strategies (Walker et al., 2020). A major downside of this strategy has shown to be the fact that once the rules used for the quantitative strategy are known and implemented by a large enough proportion of investors, the ability to provide a consistent superior investment performance ceases to exist. Smaller investors with less information at their disposal are often too late to capture the discovered anomalies and tend to overlook real-world implementation costs such as lending fees and liquidity costs (Sloan, 2019). This causes them to perform poorly compared to more sophisticated institutional investors that utilise their advanced back-testing techniques and market knowledge. Moreover, this investment strategy is based on the strong assumption that capital markets are efficient in the long run and therefore ignores the basic procedure of fundamental analysis, which is a method to evaluate the fair value of a security based on economic, financial and various other quantitative and qualitative factors (Sloan, 2019). Since an important role of capital markets is to efficiently allocate scarce resources, marginal investors need to have accurate estimates of the fair value of the underlying securities. With the increasing dominance of less informed retail investors and noise-traders on capital markets, significant mispricing effects might therefore emerge (Kelley and Tetlock, 2013). Unlike quantitative investment strategies that use technical analysis, fundamental analysis aims to for example forecast the future earnings of a firm to determine the fair value of a security. Skilled investors can earn abnormal profits by estimating which securities show deviations from this fair value and capture this in an investment portfolio. However, in order to benefit from fundamental analysis, broad business skills and knowledge are required to synthesise both qualitative and quantitative information for providing accurate forecasts of future earnings.

2.2. Earnings Forecasting

Earnings forecasting can be defined as the process of analysing historical financial statement and economic data with the purpose to predict the movement of, or the exact, future earnings (Bartram and Grinblatt, 2018) and is central to determining the fair value of companies and the securities they have issued. When dividend policy irrelevance is assumed, indicating that the pay-out of dividend will result in a decline in the stock price by the amount of dividend per share, expected earnings are the fundamental determinant of equity value. Moreover, empirical evidence has shown that accrual-accounting earnings, and not dividends or free cash flows, are what investors forecast when they estimate the fair value of a firm (Monahan, 2017). In line with this, a useful earnings forecasting approach needs to be replicable, objective and should generate low-cost and accurate forecasts for a large sample, which can be validated when benchmarking the outcomes to experts' forecasts.

Since fundamental analysis involves the usage of present and historical financial statements, combined with industry and economy-specific data to determine the true value of securities, the actually observed prices at which a firm's securities trade in the market are only important to validate the assumptions made in the models. When actual prices deviate from this calculated true value, investors can benefit by capturing these perceived differences in a well-diversified investment portfolio. Babii et al. (2020) showed that investors typically use professional analysts' recommendations to obtain these perceived differences to pick stocks or rebalance their investment portfolios. Often, consensus forecasts of earnings are used for this, which combine the expectations of a pool of experts to come to a buy, hold or sell recommendation.

Another and less common approach for smaller investors is to directly look into the issued financial statements by the companies in the investment universe. Ou (1990) has shown evidence for a predictive information link between non-earnings numbers on a company's financial statements and the changes in future earnings. She also showed that a valuation link exists between these predicted earnings movements and the stock returns in the subsequent period. Skogsvik (2008) has chosen an alternative approach that separates winners from losers in the stock market based on accounting-based fundamental signals. Both these studies show that accounting information, and especially earnings movement forecasts based on these values, can be used as valuable input for fundamental analysis by investors. In this way, a hybrid approach is taken that uses earnings movement predictions as the input for a quantitative investment strategy that builds an investment portfolio based on a predicted increase or decrease in earnings of the analysed firms.

2.3. Barriers for Small-scale Investors

As with quantitative investment strategies, smaller and retail investors generally lack the ability to perform profound fundamental analysis given their limited knowledge and resources. It requires expert judgement since the predictions are highly contextual and can not easily be simplified by applying generalising models (Sloan, 2019). This has proven to be a major obstacle for small-scale investors to apply fundamental analysis, that generally do not have the knowledge and time needed to analyse large datasets and observe the economic and financial situation of the companies to determine the fair value (Prokopowicz, 2019). Although low-cost computer programs and recommendations published by brokers, that both provide estimates of fair value, are widely available for retail investors, the usage of these information sources is often neglected by smaller investors due to the lack of validity and transparency regarding how the numbers have been calculated (Groysberg et al., 2008). Smaller investors therefore mainly limit themselves to technical analysis for investment decisions, that does not require specialised economic and financial knowledge and can be backtested to assess the performance (Walker et al., 2020). The risk of overlooking significant mispricing effects if market efficiency is violated therefore emerges, which could have been captured by applying a form of fundamental analysis. Although the expectation that if sufficiently many large investors conduct fundamental analysis, there is no need for small investors to do so too, an immediate and precise prediction of fundamental value drivers can still result in major benefits for this group of investors. However, in order to benefit from fundamental analysis, the information barriers smaller investors face have to be overcome. Since it would be impossible to teach broad business skills and knowledge to all investors, allowing them to be able to synthesise the information needed for fundamental analysis, a more suitable approach is needed. In line with this, Bartram and Grinblatt (2018) have shown that fundamental analysis does not necessarily require forecasts using explicit economic models and parameters, but can be achieved by solely looking at all recently reported accounting information, essentially making it a quantitative approach. The hybrid approach derived from the discussed papers by Ou (1990) and Skogsvik (2008) goes one step further than this and uses earnings movement forecasts purely based on past and current accounting information. For the reason that accounting information is publicly available, this could be a potential solution for smaller investors to be able to implement fundamental analysis in their investment strategy. However, even with the possibility to construct a portfolio based on fundamental earnings movement forecasting, two major obstacles still remain. This first of all relates to the availability of large-scale accounting information that can be used as input for the fundamental analysis. Secondly, smaller investors need the right statistical tools to be able to immediately transform this accounting information into insights regarding the movement of fundamental earnings forecasts. Developments in the field of business analytics may provide an outcome for these barriers, which will be covered in the next sections.

2.4. eXtensible Business Reporting Language (XBRL)

The eXtensible Business Reporting Language (XBRL) is an extensible markup language that is used for digital business reporting. It includes a standard list of tags to describe business and financial information represented in the financial statements or other business reports (FASB, 2018). With the mandatory adoption of XBRL for listed companies in more than 50 countries, financial reporting data can now move digitally, accurately and rapidly between organisations and the users of this data. This means that the users, such as investors, now have access to accounting information in a machine-readable format to conduct large-scale analyses such as earnings movement forecasting for fundamental analysis. Although at first, this was mainly done by using data in human-readable formats, such as annual reports in paper or PDF, this data is now immediately available in a machine-readable format. As mentioned by the SEC (2009) as one of the purposes of XBRL: “in this format, financial statement information could be downloaded directly into spreadsheets, analysed in a variety of ways using commercial off-the-shelf software, and used within investment models in other software formats’’. Besides this, the

XBRL data is standardised and publicly available, showing the potential benefits for retail and smaller institutional investors. Typically, accounting information used for fundamental analysis approaches had to be manually obtained by analysts from human-readable reports or bought from commercial data vendors such as Compustat that are issued with a delay. With the introduction of XBRL, all investors now have access to the same large-scale and machine-readable information for fundamental analysis. However, since it is not yet necessary to audit the XBRL documents, many errors have initially shown to be present in the filings (XBRL U.S., 2015). This is one of the reasons why only a small proportion of analysts and investors showed to use XBRL data for their investment decisions (Harris and Morsfield, 2012). Nevertheless, error rates have decreased vastly over time and indicate the usefulness of this data source for investors with no access to large and often expensive datasets with accounting data (XBRL U.S., 2015). A study by Guo and Yu (2020) showed that based on evidence from the SEC's server log where XBRL data can be found, smaller investors do not show to be downloading XBRL data. A clear reason for this is not provided in the study, but it is argued that it might be due to the lack of knowledge regarding how to transform the hundreds of XBRL tagged financial items into useful insights for investment decisions.

2.5. Machine Learning Algorithms

With the developments in the field of data analytics over the last decades, the investment industry has seen many technological advances. Especially with the revolution in computing power and data-generation, analytical methods have shifted the industry towards algorithmic and high-frequency trading practices, where machine learning approaches are used to obtain predictions of fundamentals, price movements and market conditions (Jansen, 2018). Algorithms can exploit fundamental and market data in a more efficient way, allowing investors to determine the value of securities in an objective and replicable way with minimal resources (Sloan, 2019). Although at first, these techniques were only available to larger institutions that had enough computing power and storage available to deploy the algorithms, nowadays basically all investors with a computer at hand can run advanced machine learning models if the right data is available. Although this should enable retail and smaller institutional investors to compete with more sophisticated investors, evidence suggests that computerised investment decision making is mainly taken advantage of by larger institutional investors (Lo, 2017). Besides the perceived unavailability of accurate input data for the models, which has been addressed in the previous two sections, Jansen (2018) discusses the difficulties in selecting the right machine learning algorithm that cause these investors to withhold from applying these tools for investing purposes.

The next subsections will discuss several machine learning algorithms and their application possibilities to provide fundamental earnings movement predictions based on accounting information. Emphasis will be placed on three dimensions that address the restraints that smaller investors might have to adopt these approaches. First of all, it is important to stress the model family to which a specific algorithm belongs. Since this defines the assumptions regarding the nature of the functional relationship between the input and output variables, this ultimately determines how accurate the results for a specific task can be (Jansen, 2018). The algorithms that will be discussed belong to the linear model and tree-based ensemble model families. While deep learning models such as neural networks have also been proven to provide accurate earnings movement forecasts based on accounting information (Xinyue et al., 2020), these models require many complex calculations to observe why a specific prediction has been made. This so-called ‘black-box-ness’ is less embedded in the linear and ensemble models (Jansen, 2018), and since this showed to be a major barrier for smaller investors to apply machine learning algorithms for investment decision, it is important to consider the interpretability of the applied model. The deep learning models have therefore been left out of scope, and the next sections pay special attention to the interpretability dimension.

The importance of the bias-variance trade-off will also be discussed per model, which shows how the generalisation error can be reduced. Error due to bias emerges if the algorithm attempts to learn the true function but fails to capture the full complexity of this function. This can also be defined as underfitting (Géron, 2019). On the other hand, error due to variance arises if the algorithm is overly complex and overfits the noise embedded in the data used to train the model (Géron, 2019). Since the risk of under- and overfitting differs per model and can be balanced by means of general optimisation techniques such as cross-validation and model tuning or model-specific optimisation techniques such as regularisation, it is important to cover the findings in the literature related to this. Finally, different findings regarding the application of the specific algorithms on accounting data for earnings forecasting purposes will be covered.

2.5.1. Linear Models

The first and most basic machine learning algorithms belong to the family of linear models. Many financial machine learning applications rely on linear predictions due to their robustness when deployed on noisy financial data and their easiness to interpret (Jansen, 2018). Linear models can be applied to both classification problems, which can for example be the directional prediction of future earnings, as well as regression problems, that contain the prediction of the exact amount of future earnings (Jansen, 2018).

The ordinary least squares (OLS) model and generalised linear models (GLM) assume that a linear combination of the input variables result in an output, where various error and distributional assumptions aim to guarantee that the estimates are unbiased and efficient (Géron, 2019). However, in the context of earnings forecasting, the assumptions fundamental to these models are almost always violated (Sloan, 2019). The generalised least squares model (GLS) is an alternative if some of the assumptions are violated but does not take into account that the linear regression model may contain multicollinearity that causes high variance and subsequently, causes the model to be likely to overfit the training data (Jansen, 2018).

The linear models that control for this overfitting are the ridge and lasso regression models. They build on the regularisation technique that adds a penalty term to the error function that creates a size constraint for the coefficients (Géron, 2019). By doing this, the low bias of the least squares estimates is slightly increased by ‘shrinking’ some coefficients. On the other hand, the relatively high variance of the least squares estimates is thereby reduced, increasing the overall prediction accuracy. Especially when forecasting earnings movement based on many financial statement items, the large number of predictors can make the interpretation of why a specific prediction has been made complicated. By applying a shrinkage model, a smaller subset of predictors is used that increases the interpretability and penalises correlated predictors that cause multicollinearity (Jansen, 2018). The main difference between the ridge and lasso regression models relates to how the regularising penalty term is calculated. For the ridge regression model, the sum of squared coefficients is used, whereas for the lasso regression model, this is based on the sum of absolute values of the coefficients. This regularisation parameter can for both models, like all hyperparameters, be tuned by applying cross-validation. This technique of using part of the training data one or several times to validate the model assists in providing an unbiased estimate of the error rate (Géron, 2019).

A small advantage of using the lasso regression model over the ridge regression model is that some coefficients are reduced to zero, allowing the model to be used as a selection method for choosing a subset of features that will be included in the model. In line with this, Babii et al. (2020) have shown that by using this lasso regression on financial statement and economic data for predicting price-earnings ratios, several benchmarks, including analysts’ predictions, can be outperformed. It will therefore be interesting to see how well this model performs when applied on XBRL data with the aim to predict future period earnings movements.

2.5.2. Ensemble Models

Ensemble models are machine learning models that build upon multiple individual models combined and produce an aggregated prediction to lower the error variance. They are among the most successful algorithms to apply on standard numerical data like financial statements, but they also come with several disadvantages that need to be considered (Jansen, 2018). Two models that have been proven particularly suitable for forecasting purposes will be discussed here, being the Random Forest model and Gradient Boosting Machines.

The Random Forests model is an ensemble model that combines many different decision trees that predict the value of a target variable based on learned rules from the training data. An individual decision tree starts at the root for all samples, after which a threshold on one specific feature is used to split the samples into two groups. This process is continued until a final node, where a prediction is made, is reached, creating transparency regarding how different features and their values result in a specific model outcome (Jansen, 2018). A major downside of this model is the tendency to overfit the training data, which increases the prediction error of the model (Géron, 2019). This especially holds when a large number of features are included, such as in earnings forecasting based on financial statement items. Although this overfitting can be reduced by decreasing the complexity of the tree by applying regularisation or tree-pruning techniques, a more powerful approach is to combine multiple trees while randomising the construction of these trees, creating a Random Forest.

This Random Forests model builds on the averaging method that trains several independent base estimators and averages their predictions. This can be achieved by randomly sampling many different training sets from a population with duplicates, also known as ‘bagging’, and decreases the generalisation error of the model by significantly reducing the variance (Jansen, 2018). Moreover, the Random Forests model further decreases the variance by randomly sampling the predictive features used in the model. This randomised reduction of a pre-defined number of features de-correlates the prediction errors of the grown individual trees and can be configured by using cross-validation. A major benefit offered by Random Forests is that this cross-validation is already built-in since part of the observations is not included in a training set due to the bagging. These ‘out-of-bag’ observations can be used to validate the model and provide an unbiased estimate of the generalisation error (Jansen, 2018). Where regular decision trees can be visualised, greatly increasing their interpretability, this is not the case for Random Forests since this ensemble model averages the outcomes of numerous individual decision trees. However, by looking at the overall importance of different features in the bagged models, the model can be made somewhat interpretable (Jansen, 2018). Another disadvantage is the relatively high computational costs for the user, making the predictions slower to generate. By running the model in parallel, the speed of running the Random Forests model can be increased greatly.

Many applications of the Random Forests model for fundamental analysis can be found in the literature. For example, Anand et al. (2019) use Random Forests to generate out-of-sample predictions for the directional change of five profitability measures and show that this machine learning method offers better predictive performance than regression-based methods. Similarly, Hunt et al. (2019) apply Random Forests, among two other machine learning approaches, to predict the sign of one-year ahead earnings changes based on Compustat data from 1976 to 2015, including 60 financial ratios. Their findings show that the Random Forests model can significantly improve the prediction of the direction of earnings changes compared to Logistic regression, with an out-of-sample accuracy of 78.8% relative to 62.3%.

The Gradient Boosting Machines model is another ensemble model that builds upon combining several weak decision tree models. It can be considered as one of the most useful machine learning algorithms due to its ability to train a new tree based on the cumulative errors in the previous decision tree (Jansen, 2018). Whereas in a Random Forests model, the trees are trained independently using different 'bagged' training sets, the Gradient Boosting Machines model relies on sequentially 'boosting' the model using a reweighted version of the data that reflects cumulative learning results (Géron, 2019). The main idea behind the model is to train the base learners to learn the gradient of the loss function, allowing it to reduce the overall training error made by prior learners. The final model then makes a prediction based on the weighted sum of the predictions of the individual trees that have all been trained to minimise the ensemble loss function given prior predictions. In this way, a complex functional relationship can be learned in an incremental way, but the risk of overfitting has to be managed (Jansen, 2018). This can first of all be achieved by tuning the complexity of the embedded decision trees to avoid it from learning highly specific rules. By ensuring a minimum number of samples to split or accept a node, or a minimum improvement in node quality, the risk of the model to overfit can be reduced. Furthermore, by tuning the size of the combined ensemble and by making sure the model training is stopped when the validation error no longer decreases, the generalisation error can be minimised. Lastly, a shrinkage penalty can be included and tuned to scale the contribution of each new ensemble member down by a factor between 0 and 1, which decreases model complexity (Jansen, 2018). Like in the Random Forests model, much of the interpretability found in decision trees gets lost when many of these trees are ensembled within the Gradient Boosting Machines algorithm. However, feature-importance scores can again be obtained from the model that show which model features produce the highest gains in terms of the total reduction in loss, usefulness for model splits and decrease in prediction error (Géron, 2019).

In the literature, the application of the Gradient Boosting Machines algorithm for fundamental forecasting has not yet been covered in detail. Hunt et al. (2019) mention the untabulated application of this algorithm in their research and the found underperformance compared to the extensively covered Random Forests model. A comparison of this algorithm to the other discussed ensemble and linear models, especially when deployed on the publicly available and machine-readable XBRL data to make earnings predictions, is therefore needed to see how well it performs.

2.6. Synthesis

Table 1 provides an overview of the discussed machine learning algorithms and their expected advantages and disadvantages for applying them on XBRL data for earnings movement forecasting purposes.

Although all machine learning algorithms seem to offer a viable tool for making earnings movement predictions based on XBRL data, the Rigid and Lasso linear regression models clearly provide more benefits than the GLM and GLS linear models. When comparing these Rigid and Lasso models, a minor benefit is found in the Lasso regression due to its ability to be used as a feature selection method. Similarly, the Random Forests and the Gradient Boosting Machines models outperform the Regression Trees model. Both these models have unique features for which it will be interesting to see their results when applied on XBRL data to make earnings movement predictions.

Based on a comparison of the models, it has been decided to compare the predictions of the Lasso linear regression model and the Random Forests and Gradient Boosting Machines ensemble methods in the remainder of the thesis. The methodology section will provide a more detailed description of how earnings movement forecasts can be made by using these models.

Table 1
Synthesis of Machine Learning Algorithms

Summary table of the discussed algorithms in section 2.5. that includes the name of the model, to which model family it belongs (linear or tree-based) and a summary of the advantages and disadvantages of applying the model. In the table, the GLM/GLS models have been used as a benchmark for the other linear models. Similarly, the Decision Tree model has been used as a comparison model for the tree-based ensemble models.

Machine Learning Algorithm	Model Family	Advantages	Disadvantages
GLM/GLS Models	Linear models	-Low bias -Easiness to interpret	-High variance if dimensions are not scaled -Too many correlated features cause overfitting
Ridge Regression	Linear models	-Reduced variance and increased interpretability through regularisation	-Slightly increased bias
Lasso Regression	Linear models	-Reduced variance and increased interpretability through regularisation -Can be used as feature selection method	-Slightly increased bias
Decision Trees	Tree-based models	-Low bias -Easiness to interpret through visualisation -Non-parametric model	-High variance, even with tree-pruning techniques
Random Forests	Tree-based ensemble models	-Reduced variance through bagging -Built-in cross-validation	-Partly lost interpretability -High computational costs
Gradient Boosting Machines	Tree-based ensemble models	-Reduced training error through boosting -Iterative learning function	-Risk of overfitting if parameters are tuned incorrectly -Partly lost interpretability

2.7. Shortcomings of Fundamental Analysis for Investors

Although the application of well-founded machine learning algorithms for earnings movement forecasting has proven to provide accurate predictions in the literature, the shortcomings of using these forecasts as an investor to determine an investment strategy need to be considered. Nichols et al. (2017) have shown that accounting fundamentals do not capture all the relevant information needed to determine the fair value of all firms, which relates to the fact that the book values of a firm do not faithfully represent the firm’s net asset values. In line with this, Monahan (2017) showed proof that not all company-specific information that is needed to predict the movement of earnings is captured in historical financial statement data. This mainly relates to the ability of the current accounting taxonomy to capture additional factors that have an impact on value, such as industry or macroeconomic conditions (Maama and Mkhinze, 2020). This causes the approach to predict earnings movements for investment purposes, purely based on accounting information, likely to be flawed, and the resulting predicted movement in earnings are therefore not expected to be fully accurate. Nevertheless, as with almost all investment strategies, the ability to outperform other investors by having an information advantage is crucial. The importance to compare the outcomes of the developed earnings movement forecasting approach to the forecasts made by professional analysts hereby becomes visible, which can show retail and small institutional investors further proof to adopt a more fundamental approach and compete with other and more sophisticated investors that base their investment decisions on these analyst's forecasts.

A specific procedure of using machine learning results has proven to be useful in the context of determining where the fundamental earnings forecasting approach falls short the most. By grouping together the correctly and incorrectly predicted earnings movements, the resulting characteristics of these groups can show which specific companies need special attention when their future earnings are predicted based on the approach. Dagilienė and Nedzinskienė (2018) showed that with the clustering of observations where the wrong outcome was predicted based on accounting fundamentals, a link could be drawn to firms with the tendency to be mispriced due to the possession of items not reported in the financial statements.

3. Data

In this section, the approach to scrape the XBRL data needed for the fundamental earnings movement forecasting will be discussed which is part of the tool that small-scale investors can use for investment decision-making purposes. First of all, the characteristics of the companies used as the sample for this research will be discussed, which can also be classified as the hypothetical investment universe for investors applying this approach. The scraping algorithm will then be explained, showing how the process of obtaining accounting fundamentals for the specified firms can be automated by scraping XBRL data with only a list of company symbols as input. Subsequently, part of the scraped data will be compared to the financial reporting data obtained from Compustat, to be able to assess the quality of the scraped data. This will be used to answer the first research question to provide an indication of how accurate the XBRL data is compared to traditional data sources. Following this, the data cleaning and transformation steps will be explained to make sure the output of the scraping algorithm is usable for predicting the movement of future earnings. A discussion of what data from the I/B/E/S database will be used to compare the predicted earnings movement forecasts to professional analysts' predictions will conclude this section.

3.1. Data Characteristics

The data sample used for this research includes all publicly listed U.S. companies between January 2012 and January 2021, except for financial services and regulated utilities companies. This specific sample has been chosen for several reasons. First of all, financial services and regulated utilities companies have a significantly different financial statement composition than other firms. Investment firms, for example, can report extremely high earnings per share with having an almost empty balance sheet at the end of the year. For regulated utilities, 'allowable revenues' can be set in place by authorities also impacting their reported financial statements. Including these companies is expected to reduce the predictive power of the models that will be applied, and they have therefore been left out of scope. Secondly, since all public companies in the U.S. have to tag quantitative amounts in their annual and quarterly reports in the XBRL format since 2012, only XBRL data from U.S. companies has been obtained. Although many more countries around the world adopted this standard in the years following the adoption by the U.S., the main advantage of focussing on solely U.S. companies is that all the XBRL reports are available at a centralised location, being the SEC's website. For the reason that the aim of this research is to also provide a reproducible tool for smaller investors to conduct fundamental earnings movement forecasting, the perspective of this group of investors is taken and therefore, the only input needed to obtain the data is a list of company tickers the investor is interested in. In this case, this is the complete investment universe of more than 7,000 U.S. firms that showed to be publicly listed between 2012 and 2021, which has been obtained through Compustat.

3.2. Scraping Algorithm

An XBRL data scraping algorithm has been built in R to obtain the relevant and publicly available data needed for this research. The first supplementary coding document in the GitHub repository shows the final scraping algorithm, that is readily available for investors to obtain financial reporting data from listed U.S. firms, and requires only a text file with tickers as input. The algorithm uses these tickers to obtain the XBRL instance document from the U.S. Securities and Exchange Commission (SEC) website. It then identifies the 10K annual report filing in the machine-readable XBRL format and stores the subsequent XML hyperlink. From the approximate 7,000 companies from which filings were requested, close to 5,000 showed to return one or more valid 10K-filing XML links[1]. The algorithm then uses these XML links to download all the tagged information from the annual reports. Within the XML file, all the relevant company-specific information that is captured in the annual report can be found as tagged items, which are both the standardised U.S. GAAP items as defined by the XBRL U.S. committee, as well as non-standard items that the company tagged additionally. Within the algorithm, only the U.S. GAAP items have been obtained that show comparable financial information that can be used for machine learning purposes. From the 38,115 XML filings requested, 37,733 returned successful which resulted in a final dataset with over 5,000 distinct ‘us-gaap’ tagged items.

3.3. XBRL Data Quality

To assess the quality of the scraped data, the 20 tagged items with the least missing values have been compared against the matched values found on Compustat, to give an indication of how accurate the XBRL data is. It is hereby assumed that the Compustat fundamental data provides a fully accurate representation of the annual accounting reality. The data quality assessment requires the manual linking of the XBRL tags with variables from the Compustat U.S. annual fundamentals database, of which a full list with matched tags can be found in Appendix 1. The steps taken in R for the quality assessment have been provided in the second supplementary document in the GitHub repository.

[1] Improvements in the error-prone ‘edgarWebR’ package could have resulted in more successful requests, but since the total number of XML links at this stage is 38,115, this is considered sufficient to continue the scraping.

Table 2 shows the outcome of the data quality assessment, where for each chosen tagged item, the number of matched observations and the correct and incorrect values have been quantified. It can be seen that, on average, 79% of the scraped items are exactly equal to the data obtained from Compustat. The approximate deviation of a single item is close to $22 million dollars, and the RMSPE for the items is 1.49%. Although not generalisable to the complete dataset, the table provides an indication that the scraped fundamentals are of decent quality for usage in predicting earnings movements based on fundamentals. This immediately answers the first research question by showing that compared to traditional and often costly data sources, the publicly available and machine-readable XBRL data shows proof to be reasonably accurate.

Table 2
Summary Table of XBRL Data Quality

The table shows the quality of the 20 most frequently reported financial statement items in the XBRL reports relative to the same values found on Compustat, that are assumed to be fully accurate. The number of matched observations has been provided and the percentage of items equal and unequal to the value obtained from Compustat have been shown. The average deviation and Root Mean Squared Percentage Error (RMSPE) values quantify these differences in dollars and percentages, respectively. The RMSPE is calculated by squaring the errors and calculating the average square root divided by the feature average.

XBRL Tag	Matched Observations	Equal	Unequal	Average Deviation	RMSPE
AccountsPayableCurrent	18,040	74%	26%	$27.78M	1.88%
AccountsReceivableNetCurrent	16,061	73%	27%	$35.66M	1.92%
AccumulatedDepreciation DepletionAndAmortization	23,541	68%	32%	$70.38M	2.46%
AccumulatedOtherComp. IncomeLossNetOfTax	19,864	91%	9%	$1.61M	0.40%
Assets	29,223	81%	19%	$10.42M	1.90%
AssetsCurrent	22,652	83%	17%	$16.47M	1.01%
CashAndCashEquivalentsAt CarryingValue	26,278	79%	21%	$21.39M	1.23%
CommonStockValue	24,533	83%	17%	$7.16M	0.94%
Depreciation	19,601	84%	16%	$13.65M	1.18%
Goodwill	16,985	94%	6%	$1.00M	0.22%
LiabilitiesCurrent	22,420	83%	17%	$1.87M	0.92%
PropertyPlantAndEquipment Gross	19,577	76%	24%	$78.62M	1.27%
StockholdersEquity	27,430	79%	21%	$7.76M	2.39%
NetIncomeLoss	25,968	81%	19%	$0.89M	0.70%
EarningsPerShareBasic	23,838	96%	4%	$13.17	1.27%
OperatingIncomeLoss	26,883	80%	20%	$8.73M	1.33%
Liabilities	27,221	67%	33%	$40.38M	3.14%
OtherAssetsNoncurrent	20,471	77%	23%	$24.33M	2.56%
ShareBasedCompensation	27,516	70%	30%	$18.40M	1.58%
CurrentFederalTaxExpense Benefit	20,579	63%	37%	$4.2M	1.54%
Average:	**23,096**	**79%**	**31%**	**$21.94M**	**1.49%**

Interestingly, it can be observed in Figure 1 that the quality of the scraped XBRL tagged items in terms of RMSE and average deviation reduces over time. This is in line with the findings of the XBRL U.S. (2015) that showed that large improvements in the quality of XBRL tagged statements are observable since the mandatory adoption in 2012.

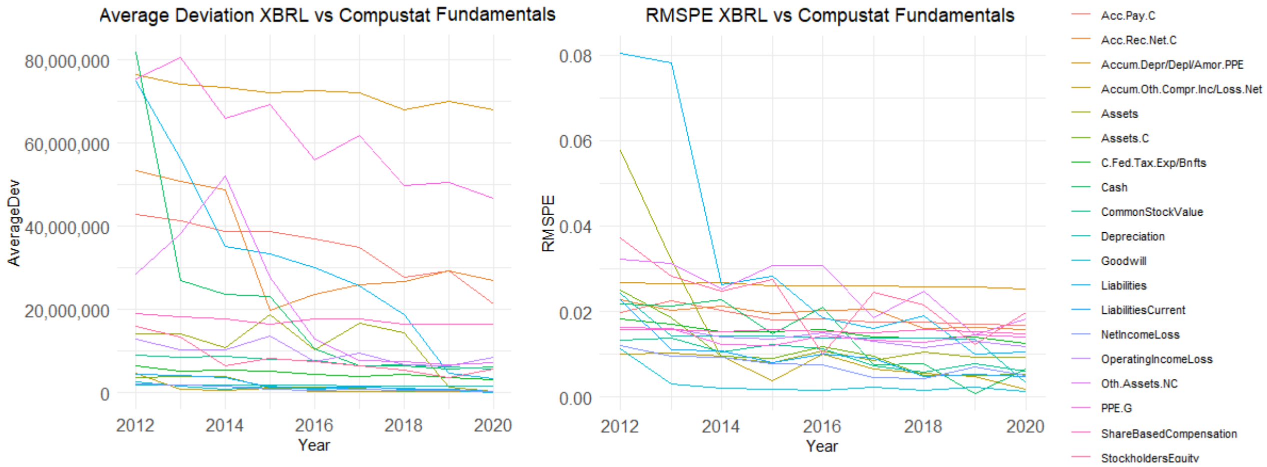


Figure 1: Movement of the RMSPE and Average deviation. The left figure plots the average deviation in dollars over time of the 20 XBRL financial statement items compared to Compustat fundamentals. The right plot shows the Root mean squared percentage error (RMSPE) of the same items. Both these figures have a year range from 2012 to 2020 and include values reported in table 2.

A side note with regards to the data quality assessment is that many of the existing financial statement items obtained from the Compustat database have not been XBRL tagged. This especially holds for ‘long-tail’ reporting items that have a relatively small frequency in the complete dataset. This is shown by the fact that a large increase in the average deviation and RMSPE is observed if all the missing values of the tagged items are replaced by zero before the XBRL data is matched and compared with the Compustat data. When an actual value is then compared to zero instead of a missing value, the error metrics largely increase. Nevertheless, it will still be interesting to see how suitable the scraped fundamentals data is to predict earnings movements, which can be utilised by all types of investors.

3.4. Data Cleaning and Transformation

Now that the quality of the scraped XBRL data has been evaluated, it can be cleaned and transformed to increase the quality and prepare it to be used in the selected machine learning algorithms. Multiple steps have been taken to achieve this, which will be discussed in this section. The taken steps in R have been provided in the third supplementary document.

A crucial first step to be able to utilise the scraped dataset is to significantly reduce the number of variables included. Processing data with too many dimensions demands excessive amounts of computational resources, and since this research aims to show the possibilities of smaller investors to conduct fundamental analysis, that do not tend to have these resources, dimensionality reduction has been applied. Although including all the items and letting the algorithms themselves do the reduction through regularisation is likely to provide better earnings forecasts, it has been decided to only focus on a selection of tagged ‘us-gaap’ items with limited missing values. By removing these columns, close to 2,500 individual ‘us-gaap’ items remain.

After combining the duplicate columns and removing the empty rows, the summary statistics show that the data contains a large number of missing values, which is in line with the expectations since not all companies in the sample report the same items in their financial statement. For example, it is unlikely that a communication service provider would report ‘Unprocessed Materials’ on their balance sheet or that a company purely focussed on the domestic market reports ‘Deferred Foreign Income Tax Expense’ on their income statement. For the reason that the ‘Total Assets’ feature will be used to scale the data, the 2,856 observations that show to have a missing or negative amount of total assets have been removed. Since lagged variables and percentage increases in the items will also be used as predictive variables, it is important that three subsequent years of data are available for a company to be included in the sample. This includes two years of data that make up the predictive features of the model and the year that follows to be able to predict the earnings movement in that year. Observations that are not part of a three or more year cluster have therefore been removed and after this selection, 15,241 observations remain for the data transformation steps. Summary statistics of these observations, including the size, year and sector distributions have been provided in Table 3.

To make sure that the observations are not impacted by the size of the companies, the first data transformation that is made is dividing all included financial statement items by the total assets. Since the dataset has already been filtered to only include observations of which total assets are available, no data points are lost in this transformation. An exception is made here for tagged items that are provided as a percentage, such as ‘tax rate’ or per-share items such as ‘common stock value per share’. Also, the ‘total assets’ variable itself has been log-transformed to ensure a normal distribution. The second transformation to the dataset has been to obtain the one-year lagged variables of all the observations that are usable for the earnings movement predictions.

A final dimension that is added to the data is the percentage change of the financial statement items compared to the previous year. Subsequently, the variables have been winsorised at a 1% level to adjust the data for outliers and all missing observations have been replaced by zero, showing the made assumption that if a financial item is not reported by a company, it is also non-existent. This is required by the machine learning algorithms that will be deployed on the data to avoid having too few data points to train and test the models with.

Table 3
Summary Statistics of the XBRL sample

The table shows the summary statistics of the obtained and cleaned XBRL sample. A distinction has been made between the fiscal years of the observations (Year), the size measured by total assets (Size) and the 2-digit SIC sector classifications (Sector). Per subgroup, the number of observations in the sample data has been included, and the subsequent frequency of this subgroup has also been shown.

	Categories	**Observations**	**Frequency**
Year	2012	1,180	7.74%
	2013	1,661	10.90%
	2014	1,762	11.56%
	2015	1,913	12.55%
	2016	2,108	13.83%
	2017	2,248	14.75%
	2018	2,172	14.25%
	2019	2,197	14.42%
	Total	**15,241**	**100%**
Size *(Total Assets)*	0-10M	2,123	13.94%
	10-50M	1,537	10.08%
	50-200M	1,909	12.53%
	200-500M	1,593	10.45%
	500-1B	1,799	11.80%
	1B-2B	1,789	11.74%
	2B-5B	1,415	9.28%
	5B-10B	2,017	13.23%
	>10B	1,059	6.95%
	Total	**15,241**	**100%**
Sector (*SIC Code)*	Mining	978	6.42%
	Construction	247	1.62%
	Manufacturing	8,326	54.63%
	Transportation, communication and utilities (non-regulated)	810	5.31%
	Wholesale and retail trade	1,468	9.63%
	Services	3,187	20.91%
	Other	225	1.48%
	Total	**15,241**	**100%**

Besides cleaning and transforming the predictive features of the model, the outcome variable, being the one-year ahead earnings per share movement, also needs to be derived. More detailed information on why this specific variable is considered as the outcome variable will be discussed in the next chapter. Since part of the XBRL tagged earnings per share items showed to deviate from the earnings per share variable provided by Compustat, as can be observed in Table 1, it has been decided to use the Compustat ‘actual’ earnings per share as the dependent variable. This one-year forward earnings per share movement variable has been assigned a value of ‘1’ if an increase in the earnings per share is visible and ‘0’ if it remained the same or decreased. The resulting dummy variable will be used to test the predictive power of the built models and will be used to compare the predictions of the XBRL data machine learning models to the predictions of professional analysts.

3.5. Professional Analysts’ Forecasts

In order to answer this third research question regarding how well the earnings movements predictions perform compared to professional analysts’ predictions, the earnings forecasts of these analysts have been obtained through the I/B/E/S database in WRDS. This database has been used to gain access to one-year forward earnings per share predictions by professional analysts and brokers.

A dataset including all the scraped publicly listed companies that have received an earnings per share forecast in the 2012 to 2020 period has been acquired from I/B/E/S, which contains the individual one-year forward predictions by different analysts. A selection of these predictions has been made by looking at a time interval of one day after an annual filing has been published to match the speed at which predictions can be made by using the XBRL and machine learning approach. If multiple analysts provided an estimation in this time interval, the average prediction has been used to compare the quality of the approach applied in this research. Although the aggregates of these predictions are likely to provide an accurate and robust classification of earnings movements, they can also be used as a benchmark to see how well other models perform and where improvements can be made. In total, around 80% of the observations from the final XBRL dataset could be matched to a one-year forward earnings per share prediction by professional analysts.

4. Methodology

Now that the XBRL data has been scraped and the quality of this data has been assessed, the methodology section will be used to show how the XBRL data can be transformed into earnings movement predictions by applying machine learning approaches. This section is structured as follows. First of all, the dependent and independent variables of the models will be elaborated on, after which the procedure to create a train and test set for the raw XBRL data will be covered. The Lasso Regression, Random Forests and Gradient Boosting Machines machine algorithms, that will be used to make the earnings movement predictions, will then be covered. Moreover, the methodology to regularise and tune these models will be discussed to show how the predictions of these models can be optimised. This immediately provides an answer to the second research question by explaining the statistical intuition behind the applied machine learning algorithms. A detailed description of how these models can be operated and tuned in R has been provided in the fourth supplementary coding document on GitHub. Besides showing the methodology that will be used to make earnings movement predictions, that will be compared to that of professional analysts, this section also serves as showing a ‘best practice’ for small-scale investors to use publicly available and machine-readable accounting information for making fundamental earnings forecasting purposes.

4.1. Model preparations

4.1.1. Dependent and Independent Variables

Since earnings can be seen as a fundamental driver of firm value (Monahan, 2017), machine learning algorithms will be applied to predict the movement of these earnings. As discussed, it has been decided to predict the up or down movement of earnings since this binary classification helps to overcome the low out-of-sample performance of the models if actual earnings per share would be predicted. It also allows investors to easily create a portfolio based on predicted earnings movements. In order to compare the predictions of the models to analysts’ forecasts, it is important that not the earnings itself, but the earnings per share (EPS) are predicted. Bradshaw et al. (2018) have shown that these movements are the leading indicator for valuation used by financial analysts and investors. Since XBRL filings have been obtained on a yearly basis, the data will also be used to predict one-year ahead movements.

The independent variables used in the model have all been obtained or derived from the scraped XBRL data. The first category of predictive features includes the tagged XBRL items of the current reporting year, which have all been divided by the reported total assets of that year and have received adjustments to account for missing values and outliers. This also holds for the second category, which contains the lagged variables reported in the annual report of the previous year. Finally, the delta variables are the percentage difference between the current and lagged features, which are expected to of good predictive value within the models.

4.1.2. Train and Test Set Creation

Fundamental in machine learning practices is the usage of part of the available data to train the model and test this model on data it has never encountered before to obtain an unbiased estimate of the generalisation error. Avoiding data leakage of test observations into the training set is key, and it is therefore important that an initial split of the data is made. In this case, 80% of the data is used for training and the remaining 20% for testing, which provides a good balance between model optimisation and generalisability. All the earnings per share movements in the test set have been replaced by the earnings per share obtained from Compustat, to make sure the test outcomes are fully accurate. Furthermore, to avoid this random sampling procedure creating an unbalanced train-test split, where companies with an increase or decrease in earnings per share are overrepresented, stratified sampling has been used. This creates a train and test dataset with a proportion of earnings per share up and down moves equal to that of the full sample distribution. To make sure the test results are comparable to the outcomes of professional analysts' forecasts, the data also considers that all observations in the test set should occur in the obtained I/B/E/S predictions dataset. Finally, no distinct validation set has been chosen upfront for the reason that model tuning will be done slightly differently for the three selected machine learning algorithms.

4.2. Machine learning algorithms

4.2.1. Lasso Regression

The first machine learning algorithm that will be applied is the Lasso regression model. As mentioned in section 2.5.1, this model builds upon the assumption that there is a linear relationship between the (lagged) financial statement items and the future earnings per share of a specific company. Since the aim is to predict the one-year ahead movement of this earnings per share, which can be defined as a classification problem, the logistic regression model that predicts an earnings per share increase or decrease needs to be understood before the 'least absolute shrinkage and selection operator' (lasso) adaptation can be applied.

Logistic regression provides class probabilities, where the output of the model is a value between zero and one. A threshold probability then determines if the output is classified as either class one or class two. More specifically, the model is a linear regression of the log-scaled odds of the conditional probability that the outcome variable belongs to a specific class. This has been summarised into equation 1, where β_1 is a vector containing the predictor coefficients.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \ldots \tag{1}$$

A maximum likelihood estimation is then used to determine which combination of beta coefficients most accurately classifies the observations, which is done by minimising the ‘log likelihood’ as shown in equation 2.

$$-\log\left(\prod_{i:Y_i=+} p(X_i) \prod_{j:Y_j=-} (1 - p(X_j))\right) \tag{2}$$

Since the coefficient estimates are likely to overfit the data if many predictors are added to the model, which is the case when predicting earnings movements with XBRL data, regularisation techniques can be applied to overcome this problem. The Lasso regression provides such a technique by adding a penalty term to the log likelihood function, that is shown in equation 3.

$$-\log\left(\prod_{i:Y_i=+} p(X_i) \prod_{j:Y_j=-} (1 - p(X_j))\right) + \lambda \sum |\beta_1| \ldots \tag{3}$$

In this equation, the λ parameter is used to shrink the coefficients that contribute to the error term the most to zero, which causes the model to be applicable as a feature selection method since the most predictive features remain. This shrinking parameter will be tuned by using 10-fold cross-validation, which is a resampling method that divides the training data into ten different groups and is used to minimise the generalisation error. This will be done by using an error grid containing 100 shrinkage parameters between 0.0001 and 0.1.

4.2.2. Random Forests and Gradient Boosting Machines

The Random Forests and Gradient Boosting Machines models are the second and third machine learning algorithms that will be applied on the scraped XBRL data to make earnings movement predictions. As discussed in section 2.5.2., the Random Forests model builds on the bagging of decision trees, that introduce a random component by growing many deep and independent trees on bootstrapped copies of the training data and aggregates the found predictions. On the other hand, the Gradient Boosting Machines model builds an ensemble of shallow decision trees that sequentially learns and improves. Both these models build on a combination of individual decision tree models, which will be explained next.

Within an individual tree model, a tree split is searched in the selected candidate variables that minimise the Gini impurity, which quantifies the probability of incorrectly classifying an observation if it is classified based on the class distribution in the dataset. The Gini impurity function is provided in equation 4.

$$Gini\ Impurity = \sum_{i=1}^{C} p(i) * (1 - p(i)) \qquad (4)$$

In this equation, C is the number of classes, in the case of this study, being an increase or decrease in earnings per share. $P(i)$ is the probability of randomly picking an outcome of a specific class. When a decision tree is trained, the best split is chosen based on maximising the Gini Gain, which can be calculated by subtracting the weight-adjusted Gini impurities of the tree branches from the original impurity. This is shown in equation 5.

$$Information\ Gain = -\sum_{i}^{C} p(i) * log_2 * p(i) \qquad (5)$$

Since the Random Forests and Gradient Boosting Machines ensemble models build upon multiple individual decision tree models to produce an aggregated prediction, the logic behind tree splits also holds for both these models. However, the models build on different machine learning logic, for which different parameters need to be selected to apply them on XBRL data.

For the Random Forests model specifically, split-variable randomisation is embedded in the model to reduce tree correlations, which is the random limitation of the number of variables available for a tree split. As proposed by Géron (2018), the square root of the number of predictors should be used as the split-variable parameter (mtry), being the default value of this parameter for classification problems. Although using this proposition would result in an mtry of around 60, 10-fold cross-validation has been used to tune this parameter between a range of 50 and 70. Another parameter in the model is the number of trees in the forest (ntree), which needs to be sufficiently large to stabilise the prediction error of the model. A rule of thumb is to use ten times the number of predictors, but since the XBRL data is expected to be noisy, this parameter will also be tuned by using 10-fold cross-validation. The grid that has been used for this is set to contain 1,000, 2,500 and 5,000 trees. Finally, tree complexity can be specified within the model, which in this case is done by setting the tree node size of the model equal to one. According to Jansen (2018), this small number of nodes gives the best results for financial predictions. It does however result in longer model run-time, which can partly be resolved by running many ensemble trees in parallel.

The Gradient Boosting Machines model grows new decision trees by using information from the previously grown tree, which is done in several steps. First of all, a single decision tree is fitted to the data. The next decision tree is then fitted to the residuals of the previous model and is then added to the algorithm. A new decision tree is then fitted to the residuals of this new model, and the process is continued until a stopping parameter in the model says that no further improvements can be made. Equation 6 shows these steps in mathematical form, where $f(x)$ indicates the stagewise additive model and b the individual trees added to this model.

$$f(x) = \sum_{b=1}^{B} f^{b}(x) \tag{6}$$

An important parameter in the Gradient Boosting Machines model is the learning rate that is used to determine the contribution of each tree to the final outcome, or in other words, how quick the algorithm learns. For the application on financial data, a range between 0.001 and 0.3 is recommended by Jansen (2018). A second parameter is the total amount of trees in the model. Since there is no standard value for both these parameters, they will both be tuned by using 10-fold cross-validation. For the learning rate, a tuning grid has been used containing values between 0.001 and 0.3 and for the number of trees, this is a grid of length 10 containing values between 500 and 5,000. Furthermore, the tree depth parameter that controls how deep the individual trees can grow has also been tuned using 10-fold cross-validation by looking for the optimal tree depth in the values 1, 2 and 3. Finally, the number of observations in terminal nodes determines when the algorithm stops learning, for which a value of 1,000 has been selected based on Géron (2018).

The combined discussions in chapter 3 and 4 have shown how XBRL data can best be combined with machine learning approaches and have thereby provided an answer to the second research question, being how XBRL data can best be combined with machine learning approaches for earnings movement forecasting. The next chapter will show if the predictions of these models are of qualitative value, both in terms of standard machine learning test metrics, as well as when the predictions are benchmarked to professional analysts’ forecasts.

5. Results and Discussion

In this section, the prediction results obtained from the machine learning models will be discussed and compared. First, an overview of the optimal tuned parameters for the specific models will be provided, combined with a detailed description of the resulting test error metrics and confusion matrices of the models. Moreover, since regularisation has been applied in all the models, the most important features for the individual models will also be provided. The outcomes of these models will then be compared to the earnings movement predictions of professional analysts, to answer the third research question regarding how well the combination of XBRL data with machine learning approaches performs relative to professional analysts' forecasts. A discussion is then provided regarding for what observations the specific models perform well or poor compared to both each other and the predictions made by professional analysts. This will be done on three dimensions, being fiscal year, size and industry and will show when specific models can best be applied and if potential steps can be taken to increase the performance. Finally, the outcomes of the best performing model based on XBRL data will be compared to the same model using Compustat fundamental data to test if improvements in the XBRL data quality can result in better predictive results.

5.1. Machine Learning Predictions

Table 4 provides an overview of the predictive characteristics of the applied machine learning models, that have all been tested on the same test data set including observations the models have never encountered before. The table includes the chosen and tuned parameter specifications of the individual models, the confusion matrix and chosen machine learning error metrics. The fifth supplementary coding document shows how these values have been obtained.

5.1.1. Lasso Linear Regression Results

Based on the results provided in Table 4, it can first of all be observed that the Lasso linear regression model shows the best results in terms of accuracy with a shrinkage parameter of 0.005. Figure 1 in Appendix 2 shows a visual representation of how this parameter has been tuned. With 1,162 and 813 correctly predicted earnings up and down moves in the test set respectively, the model shows a moderate performance. This can further be quantified by looking at three different test metrics. Firstly, the accuracy shows the percentage of predictions the model classified correctly, which is calculated by dividing the sum of the true positives and the true negatives by the total amount of predictions. For the Lasso regression, the accuracy is 64.8% and shows that the model is able to correctly predict a fair majority of the observations.

Since the no information rate of this test set is 53.1%, indicating that by basing predictions on the most frequently occurring class, an accuracy of 53.1 percent is achieved, the model performs considerably better than this benchmark. The model also shows to produce more false positives than false negatives, showing that the model struggles with correctly classifying earnings decreases. Secondly, the area under curve (AUC) error metric shows the probability that the model will rank a randomly picked earnings up-move higher than a randomly picked earnings down-move. The AUC can visually be represented as the area under the ROC curve that shows the true positive rates against the false positive rates. With an AUC of 64.3%, the model again shows a fair predictive performance. The final error metric that has been calculated is Cohen's Kappa, which shows how well the classifier performs over a classifier that would randomly guess an outcome based on the class distribution. It is calculated by first of all taking the overall accuracy of the model minus the measure of agreement between the predictions and the actual class distribution. This is then divided by one minus this measure of agreement, and a Cohen's Kappa between -100% and 100% is then obtained. In other words, the measure removes the possibility that the classifier and a random guess agree and thus shows the number of correct predictions it makes that can not be explained by randomly guessing. With a Cohen's Kappa of 28.8%, the Lasso model shows a fair prediction of earnings movements.

Table 4
Machine Learning Results Table

The table shows the outcomes when the tuned machine learning models are applied on the test data set, including observations it has never encountered before. In the table, the forecasting method has been specified and the optimal tuning parameters have been provided, where * indicates that the parameter has been tuned by using 10-fold cross-validation. The table also shows the true positive (TP), true negative (TN), false positives(FP) and false negative (FN) prediction outcomes, that have been obtained from the confusion matrix. The accuracy ($= \frac{TP+TN}{\text{Total}}$), area under the ROC curve (AUC) and Cohen's Kappa test error metrics show the performance of the models. The last row of the table shows the aggregate professional analysts' forecasts for the exact same observations in the test set, that have been obtained from the I/B/E/S database.

Forecasting Method	**Parameters**	**TP**	**TN**	**FP**	**FN**	**Total**	**Accuracy**	**AUC**	**Cohen's Kappa**
Lasso Regression	Shrinkage penalty = 0.005*	1,162	813	617	457	3,049	64.8%	64.3%	28.8%
Random Forests	ntree = 5000* mtry = 65* nodes = 1	1,218	1,033	397	401	3,049	73.8%	73.7%	47.5%
Gradient Boosting Machines	Learning rate = 0.03* ntree = 1,500* depth = 2* stop = 1000	1,219	1,084	346	400	3,049	75.5%	75.6%	51.0%
Professional Analysts' Forecasts		1,243	1,024	406	376	3,049	74.4%	74.2%	48.4%

Since the Lasso linear model has a built-in regularisation technique that adjusts non-value-adding predictive features to zero, only a selection of features has been used in the final model. To gain an understanding of which of these features add the most predictive value to the model, variable importance scores have been calculated. The subsequent 20 most important features have been visualised in Figure 2 of Appendix 2 and show that the lagged ‘Income Tax Reconciliation Other Adjustments’ is the most important variable in predicting the one-year forward earnings movement. This is followed by the ‘Unrecognized Tax Benefits Reductions’ and lagged current year ‘Income Tax Reconciliation Other features. Although these features might not be the most fundamental within a company’s financial statements, they do allow the model to provide reasonable earnings movement predictions.

5.1.2. Random Forests Results

Compared to the Lasso linear model, the non-parametric random forests model shows to outperform this model in all test data metrics. First of all, it can be observed that the model does so by using a tuned 5,000 number of grown trees and 65 randomly sampled variables at each split. The visualised tuning outcomes of this parameter can be found in Figure 3 of Appendix 2. Based on the resulting accuracy of the model that shows that 73.8% of the predictions are correct, it can be seen that the model provides far better predictions than the Lasso model in terms of accuracy. Although this is sizably lower than the findings of Hunt et al. (2019), that found an accuracy of 78.8% based on Compustat fundamentals, it should be noted that their sample covered a far longer period from 1976 to 2015, where the impact of years of unexpected economic downturn on the made predictions is likely to be far less severe.

Within the random forests model, an almost equal number of false positives and false negatives show to be present. The AUC of the model of 73.7% indicates that the model provides good earnings movement predictions, which is again a sizable improvement compared to the Lasso regression model. Finally, the Cohen’s Kappa of 47.5% confirms the relatively good performance and shows that when the chance of randomly guessing a movement correctly is adjusted for, the model still provides good predictions. Although making the random forests model interpretable is more difficult than doing so for the Lasso regression, the variable importance scores can still be obtained. This is provided in Figure 4 of Appendix 2 and shows that within the model, the current year earnings per share is by far the most important feature, followed by the ‘Net Income Loss’ and previous year’s ‘Total Assets’. Noticeably, the most important features within the Random Forests model are mostly earnings-related or are items reported on the profit and loss statement.

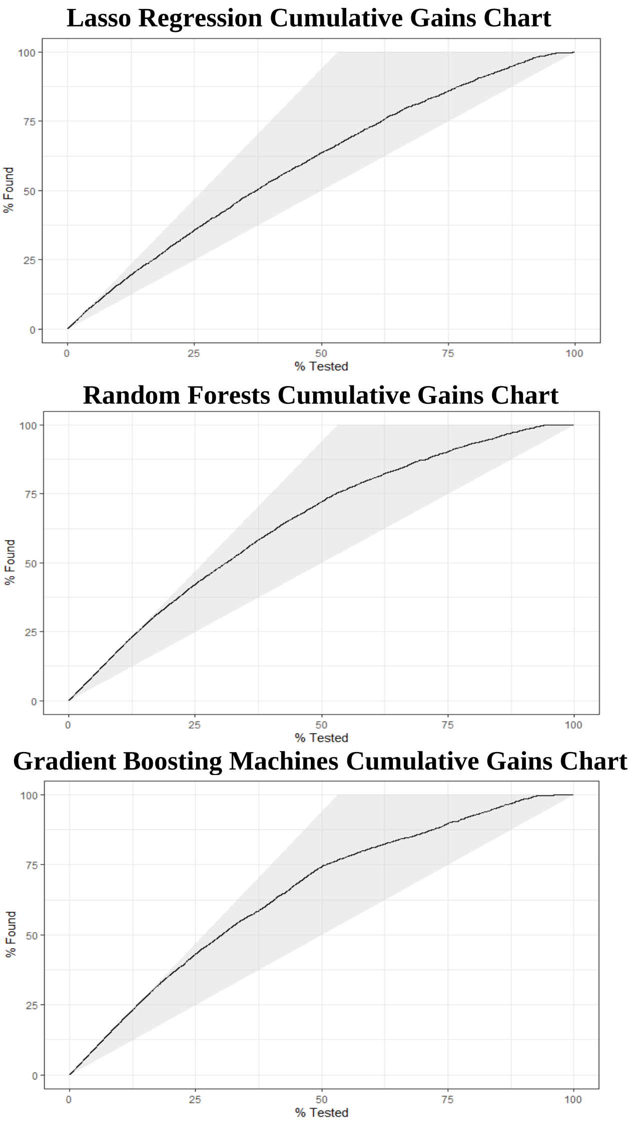
5.1.3. Gradient Boosting Machines Results

The final machine learning model of which the results are presented in Table 4 is the Gradient Boosting Machines model. With a tuned learning rate of 0.03, 1,500 number of trees and a tree depth of 2, this model outperforms both the Random Forests and Lasso Regression models in all error metric dimensions, making this the best machine learning model to apply on XBRL data for one-year ahead earnings movement predictions within this sample. A tuning plot for the number of trees, tree depth and learning rate parameters can be found in Figure 5 of Appendix 2.

In Table 4, a slight imbalance can be observed when looking at the confusion matrix, where relatively more false negatives show to be present than false positives. In terms of accuracy, 75.5 percent of earnings movements in the test set have been correctly predicted, showing a good generalised model performance. The AUC of 75.6% and Cohen's Kappa of 51.0% confirm this finding and provide proof that the model is better suited to apply on the scraped XBRL data than both the Lasso and Random Forests models. Figure 6 in Appendix 2 shows the most important features within the Gradient Boosting Machines algorithm, that have been obtained from the variable importance measurement. It can be observed that within the model, the current year 'Earnings per share' variable is used most often when a tree split needs to be made, making it the most important variable included. The 'Income Tax Expense Benefit' and 'Net Income Loss' variables show to be the second and third most important features respectively. These variables are similar to those found to be most important in the Random Forests model, but overall, the Gradient Boosting Machines produces better predictions based on these features. Moreover, the most predictive features in this model again show to be mostly obtained from the company's profit and loss statement, highlighting the importance of including these variables in the machine learning models.

Overall, it can be observed that all applied machine learning models provide decent earnings movement predictions when applied on XBRL data, where the Random Forests and Gradient Boosting Machines ensemble models perform the best. With approximately 75% prediction accuracy, these models are well suited to predict next year's earnings per share movements. A potential explanation for why the ensemble models outperform the lasso linear model could be due to their non-parametric nature, causing them to not be impacted by the non-linearity embedded in some of the XBRL features. The higher prediction scores for the Gradient Boosting Machines model relative to the Random Forests model show that the 'boosting' procedure of sequentially improving weak base learners outperforms the 'bagging' procedure of averaging base learners trained on bootstrapped samples. The most important features in these two models are comparable and show to mostly be related to either the current year earnings items or items obtained from the profit and loss statement. This also shows the importance that these items are accurately reported in the XBRL filings.

Figure 2 shows the cumulative gains charts of the discussed machine learning models. These charts are used to visualise that given a percentage of tested and ranked predictions, a certain percentage of the total amount of these class predictions is also found. It can for example be observed that within the Lasso regression, the top 25% of ranked earnings up-move predictions uncover approximately 30% of all up-moves. In comparison, for the Gradient Boosting Machines model, this is close to 50%.


Lasso Regression Cumulative Gains Chart

Random Forests Cumulative Gains Chart

Gradient Boosting Machines Cumulative Gains Chart

Figure 3: Cumulative Gains Charts. Visualisation of the percentage of the total amount of found correct predictions given a percentage of ranked test outcomes. This has been shown for the Lasso Regression, Random Forests and Gradient Boosting Machines models. The x-axis shows the percentage of test predictions, while the y-axis shows the percentage of found true observations in these tested predictions. The upper bound of the grey area in the background shows the best possible gains chart. The diagonal lower bound is the baseline curve.

5.2. Professional Analysts’ Forecast Comparison

Now that the applied machine learning models have been discussed and compared to each other, the performance of these models can be benchmarked to actual earnings movements predictions made by professional analysts to answer the third research question of this thesis. For the reason that the observations in the used test set are also present in the I/B/E/S dataset that contains one-year ahead aggregated earnings movement predictions of professional analysts, the performance can directly be compared. The aggregated predictions of professional analysts are provided in the bottom row of table 4 and show to outperform the Lasso and Random Forests machine learning models in terms of all three calculated metrics. However, the accuracy of 74.4% and area under curve of 74.2% are only slightly higher than that obtained from the Random Forests model, showing that applying this machine learning model for earnings per share movement predictions provides comparable outcomes to that of professional analysts in terms of accuracy and AUC.

When the professional analysts’ forecasts are compared to the predictions of the Gradient Boosting Machines algorithm, an outperformance of the machine learning approach is visible in terms of all error metrics. The accuracy and AUC of both models are comparable, showing that they are both able to accurately classify approximately 75% of one-year forward earnings per share movements. Interestingly, the aggregated professional analysts’ forecasts show to predict the most earnings up-moves correctly with 1,243 true positives.

Evidence is hereby provided that machine learning models, and especially the Gradient Boosting Machines variant, are able to compete with the aggregated earnings per share movement predictions of professional analysts. Although the limitations of the applied approach are yet to be discussed, there is reason to assume that smaller investors with minimal resources are able to accurately predict the earnings movements of companies in their investment universe by combining XBRL data with machine learning approaches, and thereby bridge their information gap relative to investors using professional analysts’ forecasts. This can provide advantages for smaller investors when used for investment purposes, when this approach is not picked up by a majority of the professional analysts themselves. This furthermore shows that XBRL data is well suited as input for machine learning models to perform fundamental earnings movements forecasting, highlighting the usefulness of this publicly available and machine-readable source of financial information.

Based on this discussion, an answer to the third research question, being to what extent the applied machine learning models on XBRL data can compete with professional analysts' predictions on a one-year forward time horizon, can be formulated. It has been shown that the earnings movement predictions of the applied ensemble models are well suited to compete with professional analysts' forecasts. With the Gradient Boosting Machines model even slightly outperforming the aggregated analysts' forecasts within the sample, the advantage in predictive power has been quantified. Furthermore, when taking into account that all investors have free access to the publicly available XBRL data that is obtainable the instance a financial report is published, combined with the tools to reproduce the machine learning models, major additional advantages can be captured. This holds in particular for small-scale investors that would otherwise not have access to the required data and resources needed for fundamental earnings forecasting.

5.3. Model Improvements

Since the applied machine learning models have now proven to provide general insights into the one-year forward earnings movements, both in terms of test metrics as well as when compared to professional analysts' forecasts, a more in-depth approach can be taken. Since historical accounting fundamentals do not tend to capture all the relevant information needed to fully predict the movement of earnings, which mostly relates to industry or macroeconomic conditions, it will be interesting to see if differences in predictive power from the obtained earnings forecasts of the models are visible in three different dimensions. This has been achieved by grouping together the observations in the test set based on reporting year, size and industry, and by looking at the predictive accuracy of all four forecasting approaches within these clusters. In Table 5, the results of this analysis have been provided.

It can first of all be observed that for the fiscal years included in the test set, no major deviations can be found in terms of accuracy for the forecasting approaches. The Lasso linear regression model shows a relatively good performance in 2019, for which the opposite is true for the Gradient Boosting Machines and professional analysts' forecasts. This could be related to macroeconomic conditions in this year, since the observations of the 2019 fiscal year have been used to predict 2020 earnings per share, which are likely to be affected by the COVID-19 pandemic. Since the Lasso linear model shows to perform well in this year, it could be an indication that this model is more robust in times of economic downturn. However, with only 420 test observations in this year, it is difficult to draw conclusions based on this observation.

Table 5
Clustering Results Table

The table shows the predictive outcomes of the machine learning models on the test data when the 3,049 predictions are clustered by Year, Size (based on total assets) and Sector (based on 2-digit SIC code). The table shows the amount of test data observations and the subsequent accuracy of the Lasso regression (Lasso), Random Forests (RF), Gradient Boosting Machines (GMB) models and the professional analysts' predictions (Analyst FC).

			Predictive Accuracy:			
	Categories	**Observations**	**Lasso**	**RF**	**GBM**	**Analyst FC**
Year	2012	269	69.5%	75.8%	78.4%	74.0%
	2013	301	60.1%	72.8%	72.8%	78.1%
	2014	366	61.2%	73.2%	74.9%	77.3%
	2015	380	62.4%	71.8%	75.5%	74.7%
	2016	411	65.5%	72.7%	77.9%	71.5%
	2017	456	67.1%	75.7%	77.9%	77.2%
	2018	430	62.1%	74.9%	77.2%	73.3%
	2019	420	70.0%	73.6%	70.5%	69.0%
	Total	**3,049**	**64.8%**	**73.8%**	**75.5%**	**74.4%**
Size *(Total Assets)*	0-10M	228	72.0%	80.1%	81.8%	76.3%
	10-50M	382	66.2%	77.2%	77.7%	74.6%
	50-200M	407	73.5%	82.6%	80.3%	76.2%
	200-500M	362	60.5%	66.9%	69.6%	71.8%
	500-1B	275	59.6%	68.4%	74.5%	65.1%
	1B-2B	333	55.9%	66.7%	69.1%	64.9%
	2B-5B	339	62.8%	68.4%	69.6%	67.0%
	5B-10B	416	57.4%	74.5%	78.2%	78.5%
	>10B	306	68.0%	75.5%	77.1%	77.0%
	Total	**3,049**	**64.8%**	**73.8%**	**75.5%**	**74.4%**
Sector (*SIC Code)*	Mining	185	66.5%	80.5%	76.8%	69.7%
	Construction	43	39.5%	60.5%	62.8%	55.8%
	Manufacturing	1,675	65.3%	78.4%	79.0%	71.8%
	Transportation, communication and utilities (non-regulated)	177	63.3%	66.1%	71.8%	81.9%
	Wholesale and retail trade	286	58.4%	65.4%	66.4%	79.4%
	Services	646	67.2%	66.9%	72.1%	77.9%
	Other	37	81.1%	66.7%	66.7%	85.2%
	Total	**3,049**	**64.8%**	**73.8%**	**75.5%**	**74.4%**

It can also be observed that the Random Forests and Gradient Boosting Machines models show a similar performance for all the year subgroups, where in all fiscal years besides 2019, the Gradient Boosting Machines model shows a better accuracy. This confirms that the underlying assumptions of these ensemble models are similar and that consistent predictions can be made in all included years. The professional analysts' predictions show to outperform the machine learning models for earnings predictions made for the 2013 and 2014 observations, but in all other years, there is at least one machine learning model that outperforms these aggregated forecasts. A possible improvement for the models based on the observed fiscal year clusters could be to include macro-economic trends in the model. However, due to the inability to predict sudden economic downturns such as the COVID-19 pandemic, no major improvements in this regard are expected to be possible. As shown by the large drop in accuracy for professional analysts' one-year ahead earnings per share forecasts in 2019, even those with the expectancy of having the right tools available to predict economic trends were affected.

When looking at the firm size subgroups, larger accuracy deviations are visible compared to the fiscal year subgroups. For example, the Lasso linear model provides forecasts with an accuracy of only 55.9% for firms with total assets between 1 and 2 billion, and an accuracy of 57.4% for firms with total assets between 5 and 10 billion. Above-average predictions for this linear model can be found for the largest size cluster, where 68% of the earnings movements are correctly predicted. Again, the Random Forests model and Gradient Boosting Machines models show similar performance within the size clusters, where the Gradient Boosting Machines model only proves to be of better use for companies with a total asset value between 500 million and 1 billion. All machine learning models show a relatively poor prediction performance for companies in the three clusters with total assets between 1 and 10 billion. However, since the analysts' forecasts also underperform for these firms, no major model improvements are expected to be possible.

Finally, the sector clustering shows a major class imbalance in the test set, with more than half of the observations being manufacturing companies. This is in line with the data characteristics as provided in Table 3, but when comparing the results between sectors, this should be taken into account. Based on the results, earnings movements in the mining sector show to be best predicted by the Random Forests model, with an accuracy of 80.5%. Within this sector, professional analysts' predictions show to be of low quality, only providing slightly better predictions than obtained from the Lasso linear model. Earnings movements of manufacturing companies show to be best explained by the Random Forests and Gradient Boosting Machines models. Remarkably, professional analysts provide by far the best predictions for the transportation, communication and non-regulated utilities companies, as well as for wholesale and retail trade and services companies. This outperformance relative to all the machine learning models shows that for these specific companies, the machine learning models need to be improved in order to compete with professional analysts' forecast. It could for example be the case that for the companies within these sectors, the obtained financial reporting items are simply not suitable to predict the earnings movements of these companies. This could be because their financial reports deviate significantly from companies in other sectors or because not all value delivering assets or activities are correctly captured in the XBRL tagged financial statements. If this is the case, then a link with the quality of the XBRL reports can be drawn.

In line with this, it will be interesting to see how well the best performing machine learning model on XBRL data, being the Gradient Boosting Machines model, performs when it is applied on the assumed to be fully accurate Compustat data. The sixth supplementary coding document shows the steps taken to perform both the data cleaning and model building steps fo this approach, where the input data is the file used to compare the XBRL data quality to the Compustat fundamentals. Within this dataset, over 400 accounting fundamentals are present on which the exact equal cleaning and transformation steps have been applied as for the XBRL data. This includes only looking at data points that are observed for three consecutive years, dividing all non-ratio items by the total assets, and calculating the lagged and percentage changes of all the variables. The Gradient Boosting Machines model is then configured by tuning the parameters on a training set, after which test set predictions have been made. With an accuracy of 75.1%, the model shows to slightly underperform compared to the Gradient Boosting Machines model applied on XBRL data. It should however be noted that not the exact observations have been used to train and test the models. Nevertheless, the similar predictive power in terms of accuracy shows that no major benefits can be expected if the XBRL data quality is improved even further, but additional analysis is required to validate this claim. Figure 4 shows the cumulative gains chart of the Gradient Boosting Machines model applied on Compustat fundamentals data.

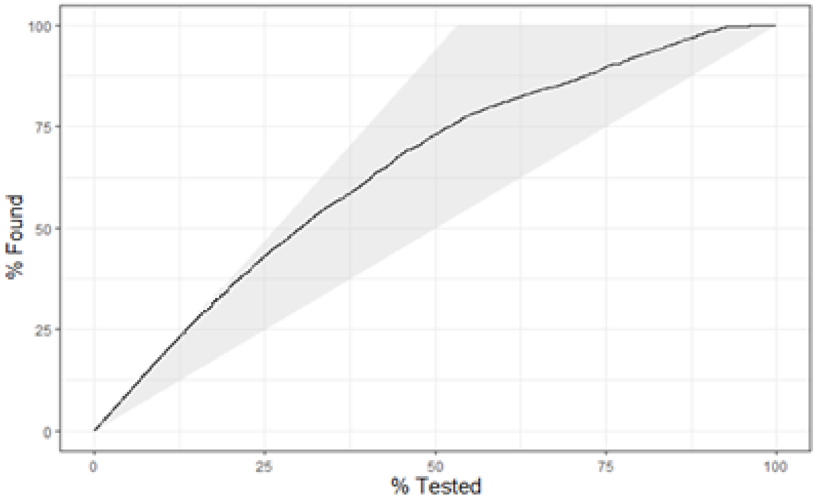


Figure 4: Compustat Data Cumulative Gains Chart using GBM. Visualisation of the percentage of the total amount of found correct predictions given a percentage of ranked test outcomes. This graph shows the outcomes from the Gradient Boosting Machines model applied on Compustat data. The x-axis shows the percentage of test predictions, while the y-axis shows the percentage of found true observations in these tested predictions. The upper bound of the grey area in the background shows the best possible gains chart. The diagonal lower bound is the baseline curve.

In sum, the clustering of the test data by fiscal year, size and sector, and the subsequent calculation of the predictive accuracy per subgroup has shown that different forecasting methods show to work best for specific instances. The found deviations are not sizable and for most observed subgroups, the machine learning models still provide similar predictive accuracy as that of professional analysts. In terms of XBRL data quality improvements, the comparison of the XBRL data based Gradient Boosting Machines model to the same model applied on similar Compustat data showed no sizable differences. Although this indicates that the benefits of using XBRL data in terms of timely availability at minimal costs do not come at the expense of losing predictive power, it also shows that improving the XBRL data quality is not the main priority for improving the earnings movement predictions.

The fourth and final research question, regarding for which specific companies model improvement is most needed and if this can be related to the XBRL data-quality issues, is therefore answered by first of all showing that based on the clustering of the observations in the test set, no major underperforming subgroups could be discovered. Moreover, no evidence is found that improvements in the quality of the XBRL data can result in increased predictive performance. The main drawbacks of predicting earnings movements by applying machine learning models on XBRL data are therefore expected to be due to the inability of financial reporting items to capture macro-economic trends or other non-financial information, which will further be elaborated upon in the next chapter.

6. Limitations and Future Research

Despite the fact that applying machine learning models on XBRL data has shown to provide good out-of-sample predictions for one-year forward earnings movements, which is validated by the finding that the Gradient Boosting Machines model is able to outperform professional analysts' forecasts, several limitations regarding the taken approach and used dataset need to be discussed. Besides this, recommendations for future research will also immediately be provided in this section.

First of all, since the proposed methodology to combine publicly available and machine-readable XBRL data has been developed from the perspective of a small-scale investor, several limitations related to this taken approach need to be pointed out. One of the goals of this thesis is to provide a detailed approach on how to obtain large scale and publicly available accounting information that is otherwise not available to all investors. Even though the XBRL initiative addressed this need and has proven to provide a relatively accurate source of accounting fundamentals, the scraping algorithm built is not able to obtain tagged XBRL reports from all requested companies. This is mainly due to the many bugs in the official SEC Filings Access API, causing the request for many XBRL tagged statements to be error-prone. It is therefore still required to have the necessary coding skills to handle these issues, causing the replicability of the approach to be called into question. Furthermore, a major obstacle for smaller investors showed to be the understanding of what analytical tools can be used to predict future earnings to perform fundamental analysis. While this research aimed to provide an understanding of how the applied machine learning models work by taking the interpretability of the algorithms into account in the model selection phase, the 'black-box-ness' of these models could not fully be unveiled. Besides showing what features are most important for a model, additional steps are needed to make the machine learning outcomes fully interpretable. Future research that uses a similar approach can apply the 'LIME' technique as proposed by Ribeiro et al. (2016). This explanation technique can be used to explain a made prediction of any machine learning model in an interpretable way by looking at the local behaviour of a model.

Another limitation related to the usefulness of the provided approach for investors is that the outcomes only provide a one-year forward earnings per share movement prediction. Although these predictions are commonly used by investors to construct a portfolio, a predicted up-move of earnings is by no means a guarantee that the share price of a firm will do the same. This immediately shows a limitation of applying fundamental analysis since even if earnings movements can fully be predicted, all that matters towards an investor is if this movement is also reflected in a return on investment percentage. For this reason, future research can use the applied earnings movement predictions based on XBRL data to construct an investment portfolio aimed at capturing the expected up- and down-moves of earnings per share. If this fundamental investment approach provides a consistent profit, the usefulness of combining XBRL data with machine learning approaches can be quantified into an investment return.

Besides these practical issues, the limitations related to the applied approach to compare the outcomes of the machine learning models to professional analysts' forecasts also need to be addressed. Firstly, the used sample within this research consists of only U.S. publicly listed firms, due to the availability of the XBRL filings in a centralised place. This causes the made claims to only hold for U.S. companies and limits the external validity of the research. Since the XBRL standard has been adopted by a majority of countries across the world, the potential investment universe for investors applying this approach consists of many more companies than just this U.S. sample. Furthermore, due to the fact that the XBRL standard has been adopted from 2012 onward, the sample data can not be used to confirm good predictive performance over a large period of time. Although the outcomes based on the used sample look promising, future research in times when more annual XBRL reports are available is needed that can test the longer-term performance. This limitation also shows the call for a worldwide database where tagged reporting data can be obtained, which eases the ability to obtain fundamental accounting information that can be used by academics or investors to predict future earnings.

In terms of improvements for the applied machine learning models, the first important limitation of this study that needs to be addressed is the full utilisation of the XBRL data. Whereas in the current approach, only US GAAP tagged items have been stored, the XBRL filings provide numerous amounts of so-called footnote disclosures that are often not tagged in a standardised way by all companies. If these items are also somehow included in the machine learning models, possible improvements can be achieved. Moreover, besides the currently used financial statement items from the current and previous year, combined with variables providing the percentage change between these two years, additional predictive features can be included in the model. For example, different ratios can be calculated and included to increase the predictive performance of the models. Finally, the established research only takes into account the linear and tree-based ensemble models. With many more machine learning possibilities available such as time-series models and neural networks, better performing models on XBRL data can be discovered. Future research can build on the methodology provided in this thesis and take into account all the discussed limitations related to the feature and model selection.

7. Conclusion

For the reason that retail and small institutional investors are becoming increasingly dominant on capital markets but often withhold from applying fundamental investing approaches, this thesis has shown the application possibilities of combining publicly available and machine-readable XBRL data with machine learning approaches to predict fundamental earnings movements. The provided approach has first of all shown how the XBRL data can be scraped and transformed into a useful dataset. Moreover, by comparing part of the scraped XBRL data to traditional sources for fundamental earnings forecasting obtained from Compustat, this research has shown that, on average, 79% of the scraped items are fully accurate and that the average deviation in terms of RMSPE is 1.49%. When this RMSPE for the items is calculated distinctively over the years 2012 to 2020, a steady decrease in this error rate can be observed, showing the improvements in XBRL data quality overtime. Based on these observations, the usage of XBRL data as a source for fundamental earnings forecasting has proven to be useful, especially for investors with no access to large and often expensive datasets with accounting data.

This thesis has also shown that several machine learning approaches can be applied on accounting fundamentals. By providing a synthesis of different machine learning algorithms that can be applied on the XBRL data, insights have been provided into the advantages and disadvantages of different machine learning models. Three models have then been selected to apply on an obtained sample of XBRL data, being the Lasso linear model and the Random Forests and Gradient Boosting Machines ensemble models. When the earnings per share movements predictions of these models are compared based on different machine learning error-metrics, it is observed that the Gradient Boosting Machines ensemble model performs the best, closely followed by the Random Forests model. With an approximate 75% prediction accuracy, these models are well suited to predict the one-year forward earnings per share movements based on XBRL data. The most important features in these ensemble models showed to be either a current or lagged earnings ratio or a feature obtained from the income statement of the firm. When the outcomes of these models are compared to the earnings movement predictions made by professional analysts, the Gradient Boosting Machines model shows to outperform these analyst forecasts. Although the differences are small, evidence is provided that machine learning models are able to compete with the earnings movement predictions of skilled analysts. This also shows that XBRL data is well suited as input for machine learning models to perform fundamental earnings movements forecasting, which highlights the usefulness of this publicly available source of accounting information.

Based on the clustering of the test data by fiscal year, size and sector, it has been observed that the machine learning predictions show to be reliable in different years and for different firm sizes. However, earnings movement predictions for transportation, communication and non-regulated utilities companies by the ensemble machine learning models show to be outperformed by professional analysts' predictions, providing evidence that the obtained XBRL reporting items for these companies are not in all cases suitable to predict the earnings movements. When the predictive power of the Gradient Boosting Machines algorithm applied on XBRL data is compared to the same algorithm on Compustat fundamental data, no sizable differences in accuracy could be observed. This shows that the benefits of using XBRL data, being timely available at minimal costs, are not penalised by decreased predictive power. It also shows that improvements in the quality of the XBRL reports are not necessarily needed when the data is applied for earnings movement predictions. Moreover, if the XBRL U.S. committee truly wants to allow accounting data to be 'analysed in a variety of ways using commercial off-the-shelf software, and used within investment models in other software formats' (XBRL U.S., 2015), this thesis shows that improvements in the accessibility of the data are more important than improving the quality of the reporting data that is currently aimed for by the committee.

Finally, although applying fundamental earnings forecasting techniques on historical financial information showed to provide accurate earnings movements forecasts, there is still room for large improvements. Since no major underperforming subgroups could be identified in the sample data of this thesis, the main area where improvements can be made is pointed out to be the current accounting taxonomy, which is still lacking standards to capture all sources of value creation within a firm. It is however expected that even if major advances are made in the ability to use data that captures all sources of value, which can then also be made machine-readable by improvements in the area of data analytics, there will always be a need for human forecasting insights.

8. References

Anand, V., Brunner, R., Ikegwu, K. and Sougiannis, T. (2019). Predicting profitability using machine learning. Working Paper. University of Illinois at Urbana-Champaign.

Arnold, V., Bedard, J. C., Phillips, J. R., & Sutton, S. G. (2012). The impact of tagging qualitative financial information on investor decision making: implications for xbrl. *International Journal of Accounting Information Systems, 13*(1), 2–20.

Babii, A., Ball, R. T., Ghysels, E., & Striaukas, J. (2020). Machine learning panel data regressions with an application to nowcasting price earnings ratios. *Ssrn Electronic Journal*.

Barber, B. M., & Odean, T. (2013). The behavior of individual investors. *Handbook of the Economics of Finance, 2*(Pb), 1533–1570.

Bartram Söhnke M, & Grinblatt, M. (2018). Agnostic fundamental analysis works. *Journal of Financial Economics, 128*(1), 125–125.

Beioley, K. (2019). Free trading apps - investment freedom or false economy? Retrieved on 23-03-2021 from: https://www-ft-com.eur.idm.oclc.org/content/8be69c5e-5f6b-11e9-b285-3acd5d43599e

Birt, J. L., Muthusamy, K., & Bir, P. (2017). Xbrl and the qualitative characteristics of useful financial information. *Accounting Research Journal, 30*(01), 107–126. https://doi.org/10.1108/ARJ-11-2014-0105

Bradshaw, M., Christensen, T., Gee, K. and Whipple, B. (2018). Analysts' GAAP earnings forecasts and their implications for accounting research. *Journal of Accounting and Economics 66*, 46-66.

Dagilienė, L., & Nedzinskienė, R. (2018). An institutional theory perspective on non-financial reporting. *Journal of Financial Reporting and Accounting, 16*(4), 490–521.

FASB (2018). SEC reporting taxonomy technical guide. Retrieved on 05-04-2021 from: https://www.fasb.org/cs/ContentServer?d=Touch&c=Document_C&pagename=FASB%2FDocument_C%2FDocumentPage&cid=1176169716122.

Géron, A. (2019). Hands-on machine learning with scikit-learn, keras, and tensorflow : concepts, tools, and techniques to build intelligent systems (Second). O'Reilly Media

Groysberg, B., Healy, P., & Chapman, C. (2008). Buy-side vs. sell-side analysts' earnings forecasts. *Financial Analysts Journal, 64*(4), 25–39.

Guo, K. H., & Yu, X. (2020). Retail investors use xbrl structured data? evidence from the sec's server log. *Journal of Behavioral Finance,* (2020).

Huang S, Liu S. (2019). Machine learning on stock price movement forecast: The sample of the Taiwan stock exchange. *International Journal of Economics and Financial Issues, 9*(2), 189-201.

Hunt, J., Myers, J., & Myers, L. (2019). Improving earnings predictions with machine learning. Working Paper. Mississippi State University.

Jansen, S. (2018). Hands-on machine learning for algorithmic trading: design and implement investment strategies based on smart algorithms that learn from data using python. Packt Publishing.

Jaiyeoba, H. B., Abdullah, M. A., & Ibrahim, K. (2019). Institutional investors vs retail investors. *International Journal of Bank Marketing, 38*(3), 671–691. https://doi.org/10.1108/IJBM-07-2019-0242

Kelley, E., & Tetlock, P. C. (2013). How wise are crowds? insights from retail orders and stock returns. *The Journal of Finance, 68*(3), 1229–1265. https://doi.org/10.1111/jofi.12028

Lo, D. (2017). On the limit order behaviour of retail and non-retail investors. *Pacific-Basin Finance Journal, 44*, 1–12.

Maama, H. & Mkhize, M. (2020). Integration of non-financial information into corporate reporting: a theoretical perspective. *Academy of Accounting and Financial Studies Journal, 24*(2), 1-15.

Mackenzie, M. (2021). Beware the madness of markets. Retrieved on 22-03-2021 from: https://www-ft-com.eur.idm.oclc.org/content/cdf7b2a1-1a45-4fad-afa5-667a5d0e38be

Martin, L., Wigglesworth, R. (2021). Rise of the retail army: the amateur traders transforming markets. Retrieved on 30-3-2021 from: https://www.ft.com/content/7a91e3ea-b9ec-4611-9a03-a8dd3b8bddb5

Monahan, S. J. (2017). Financial statement analysis and earnings forecasting. *Foundations and Trends in Accounting, 12*(2)*,* 1–115.

Nichols, D. C., Wahlen, J. M., & Wieland, M. M. (2017). Pricing and mispricing of accounting fundamentals in the time-series and in the cross section. *Contemporary Accounting Research, 34(*3), 1378–1417.

Ou, J. A. (1990). The information content of nonearnings accounting numbers as earnings predictors. *Journal of Accounting Research, 28*(1), 144–144.

Palas, R. & Baranes, A. (2019). Making investment decisions using xbrl filing data. *Accounting Research Journal, 32*(4), 587–609.

Prokopowicz, D. (2019). How can the level and significance of the emotions of investors operating on the stock exchange market be measured? Retrieved on 04-04-2021 from: https://www.researchgate.net/post/How-can-the-level-and-significance-of-the-emotions-of-investors-operating-on-the-stock-exchange-market-be-measured

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should I trust you?" Explaining the predictions of any classifier. ACM SIGKDD international conference on knowledge discovery and data mining. 1135-1144.

SEC (2009). Interactive data to improve financial reporting. Final rule. Retrieved on 05-04-2021 from: https://www.sec.gov/rules/final/2009/33-9002.pdf

Seth, H., Talwar, S., Bhatia, A., Saxena, A., & Dhir, A. (2020). Consumer resistance and inertia of retail investors: development of the resistance adoption inertia continuance (raic) framework. *Journal of Retailing and Consumer Services, 55.*

Skogsvik, S. (2008). Financial statement information, the prediction of book return on owners' equity and market efficiency: the Swedish case. *Journal of Business Finance & Accounting, 35(*7/8), 795–795.

Sloan, R. G. (2019). Fundamental analysis redux. *The Accounting Review, 94*(2), 363–363. https://doi.org/10.2308/accr-10652

Smith, I. & Wigglesworth, R. (2021). GameStop's wild ride: how Reddit traders sparked a 'short squeeze'. Retrieved on 04-04-2020 from: https://www-ft-com.eur.idm.oclc.org/content/47e3eaad-e087-4250-97fd-e428bac4b5e9

Strampelli, G. (2018). The EU issuers' accounting disclosure regime and investors' information needs: the essential role of narrative reporting. *European Business Organization Law Review, 19*(3), 541–579

Talwar, M., Talwar, S., Kaur, P., Tripathy, N., & Dhir, A. (2021). Has financial attitude impacted the trading activity of retail investors during the covid-19 pandemic*? Journal of Retailing and Consumer Services, 58*.

Walker, T., Gramlich, E., Bitar, M., & Fardnia, P. (2020). *Ecological, societal, and technological risks and the financial sector.* Palgrave Macmillan.

Wang, T. & Seng, J. (2014). Mandatory adoption of XBRL and foreign institutional investors' holdings: evidence from China. *Journal of Information Systems, 28*(2).

XBRL U.S. (2015). Avoiding common errors in XBRL creation. Retrieved on 05-04-2021 from: https://xbrl.us/wp-content/uploads/2015/12/2015AvoidingCommonErrors.pdf

Xinyue, C., Zhaoyu, X., & Yue, Z. (2021). Using machine learning to forecast future earnings. *Atlantic Economic Journal, 48(4),* 543–545.

Appendix 1. XBRL and Compustat tag links

Table A
XBRL and Compustat Matching Table

The table shows the matches of XBRL tagged items and the equal items found in the Compustat Fundamentals database. Both the items have been provided with a short description of how the variable can be defined.

XBRL	Compustat	Description
us.gaap_AccumulatedOtherComprehensiveIncomeLossNetOfTax	acominc	Accumulated Other Comprehensive Income (Loss)
us.gaap_AssetsCurrent	act	Current Assets - Total
us.gaap_AccountsPayableCurrent	ap	Accounts Payable - Trade
us.gaap_Assets	at	Assets - Total
us.gaap_CashAndCashEquivalentsAtCarryingValue	ch	Cash
us.gaap_CommonStockValue	cstkcv	Common Stock-Carrying Value
us.gaap_AccumulatedDepreciationDepletionAndAmortization	dp	Depreciation and Amortization
us.gaap_Depreciation	dpc	Depreciation and Amortization
us.gaap_ OtherAssetsNoncurrent	oanct	Other Assets Non-current
us.gaap_EarningsPerShareBasic	epspi	Earnings Per Share (Basic)/Including Extraordinary Items
us.gaap_Goodwill	gdwl	Goodwill
us.gaap_LiabilitiesCurrent	lct	Current Liabilities - Total
us.gaap_Liabilities	lt	Liabilities - Total
us.gaap_NetIncomeLoss	ni	Net Income (Loss)
us.gaap_OperatingIncomeLoss	oiadp	Operating Income After Depreciation
us.gaap_PropertyPlantAndEquipmentGross	ppegt	"Property, Plant and Equipment - Total (Gross)"
us.gaap_AccountsReceivableNetCurrent	rect	Receivables/Total
us.gaap_StockholdersEquity	seq	Stockholders' Equity - Total
us.gaap_ShareBasedCompensation	stkco	Stock Compensation Expense
us.gaap_CurrentFederalTaxExpenseBenefit	txfed	Income Taxes/Federal

Appendix 2. Model Tuning and Variable Importance

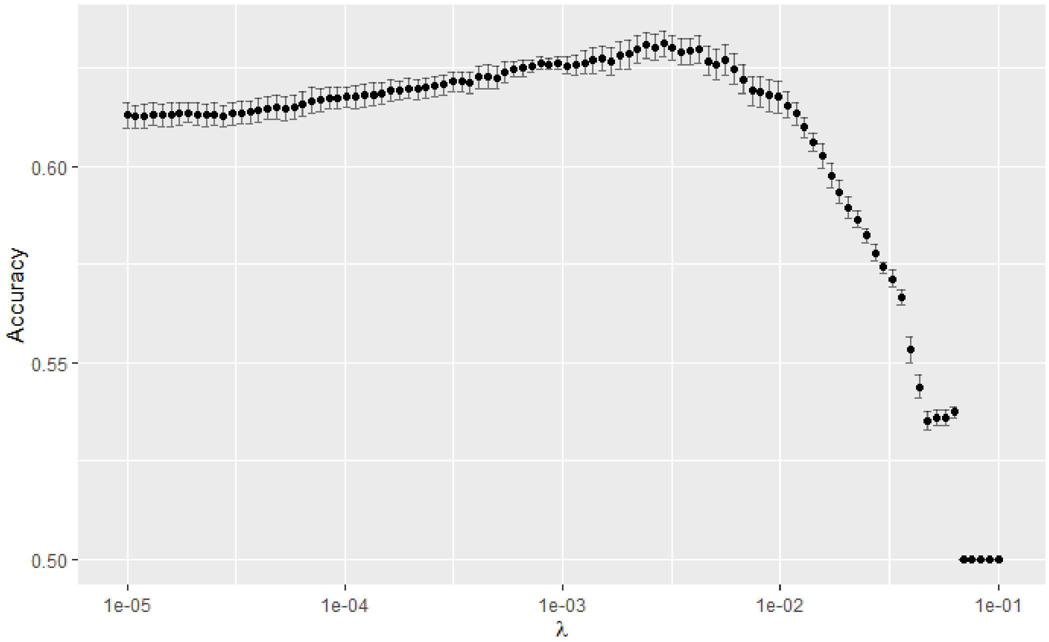


Figure 1: Tuned shrinkage parameter for Lasso Regression model

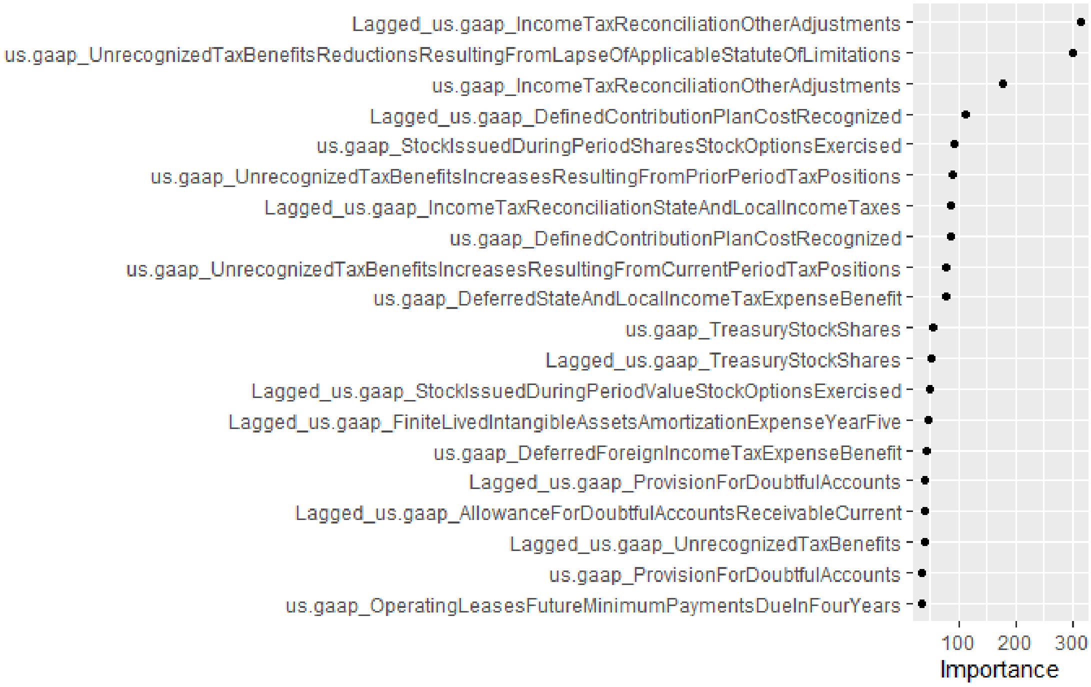


Figure 2: Lasso Regression variable importance metrics

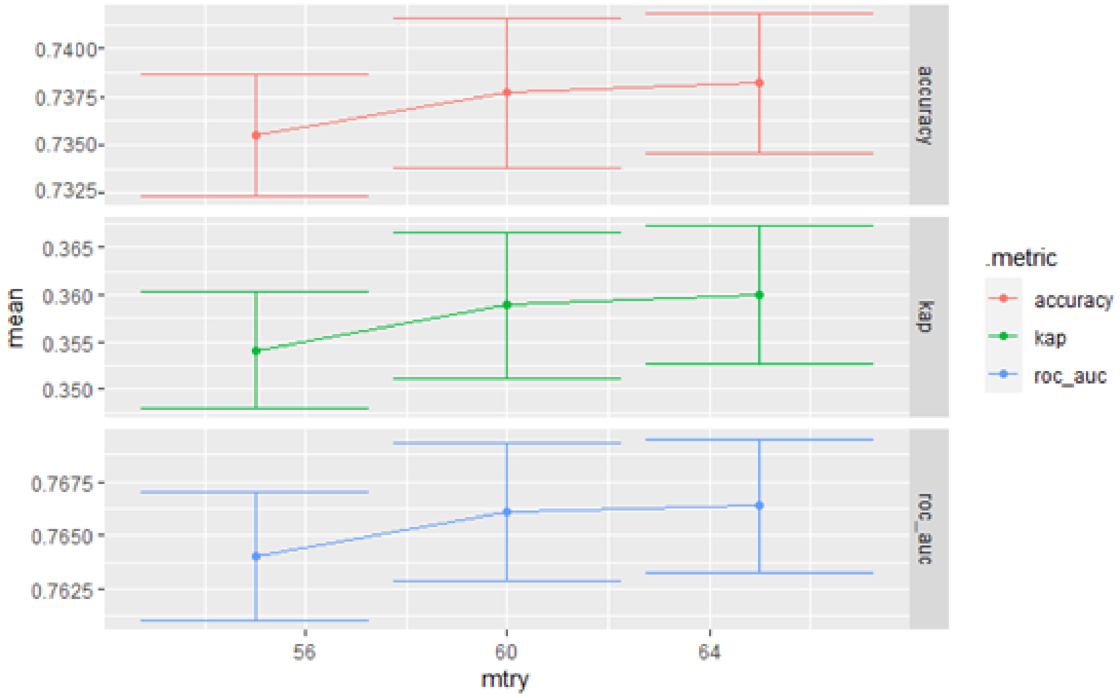


Figure 3: Tuned mtry parameter for Random Forests

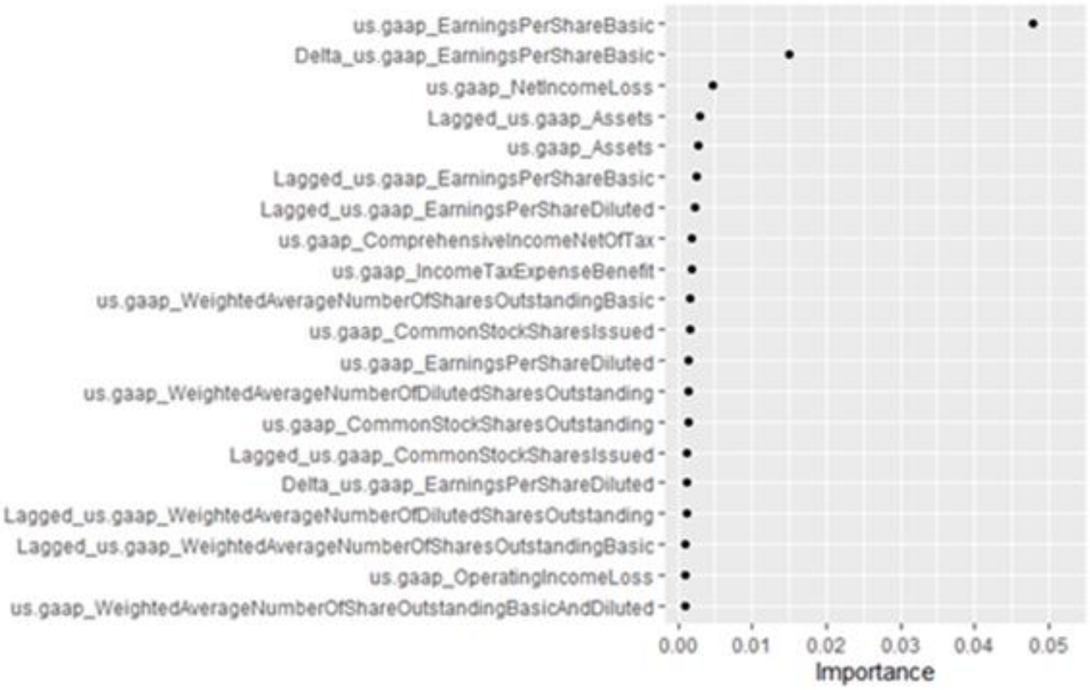


Figure 4: Random Forests variable importance metrics

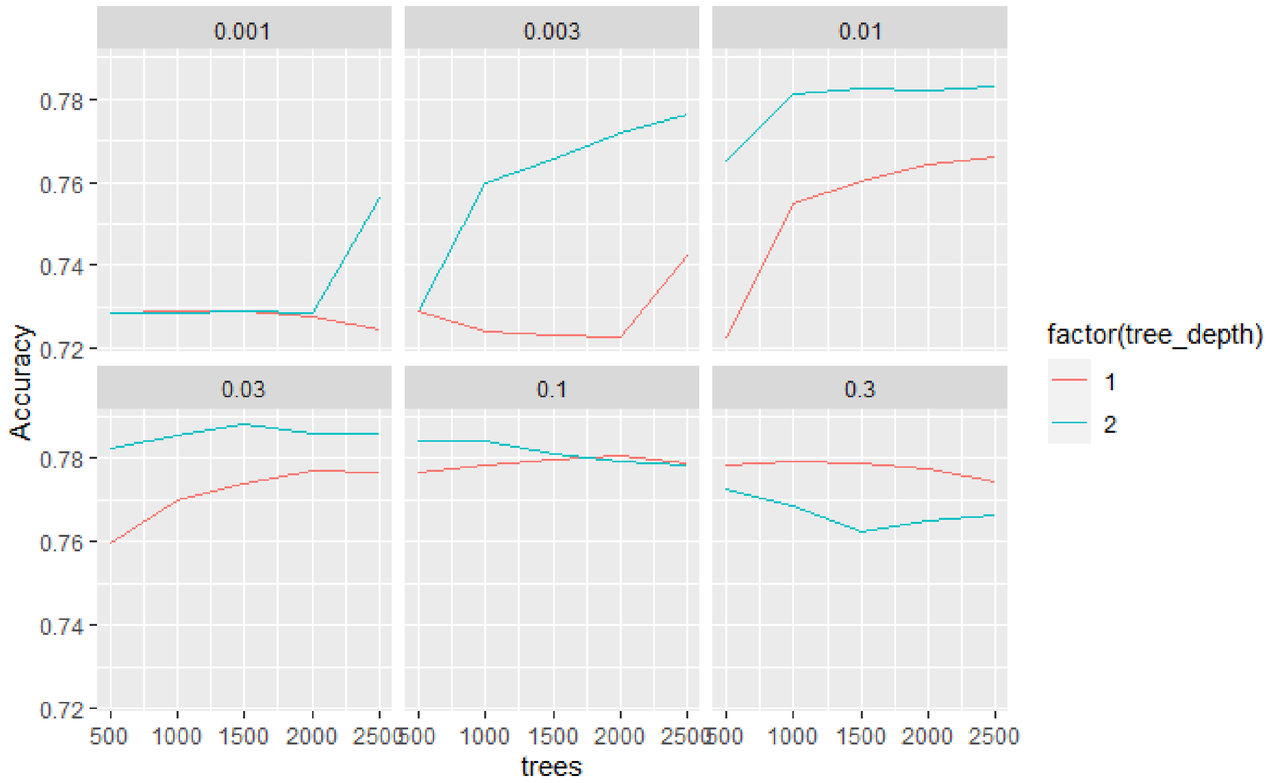


Figure 5: Tuned tree depth, number of trees and learning rate parameters for Gradient Boosting Machines model

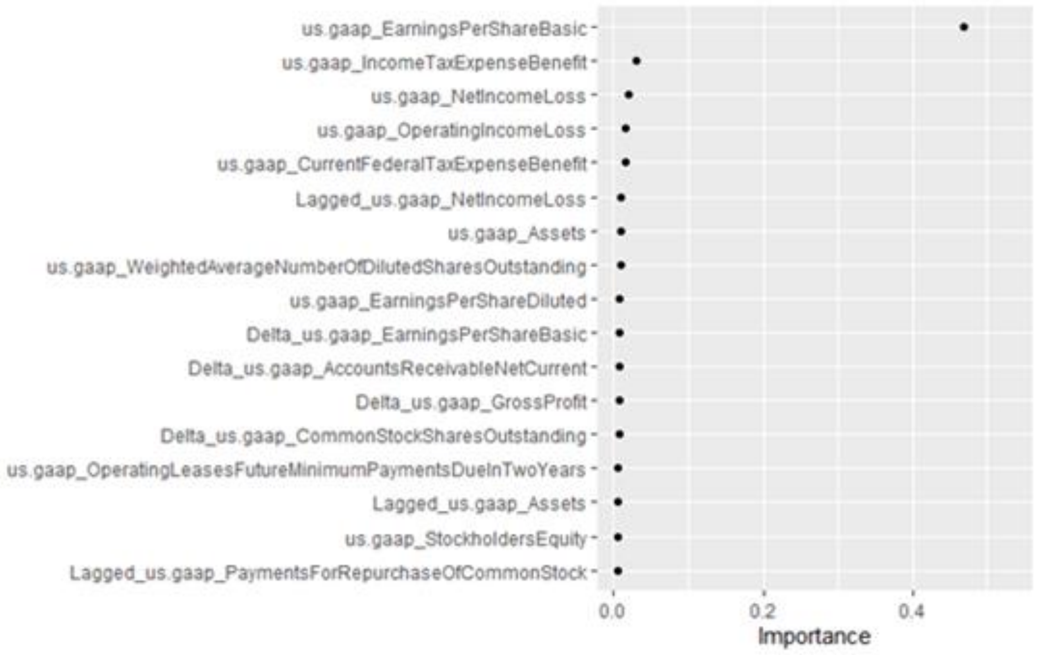


Figure 6: Gradient Boosting Machines variable importance metrics