Machine Learning based Financial Statement Analysis predicting the Market Reaction to Earnings Announcements

Daniel Kaiser

Mansfield College

University of Oxford

A thesis submitted for the degree of

Master of Science (by Research)

Trinity 2019

Abstract

This thesis considers the question of whether machine learning models can utilise the data contained in past financial statements to predict the market reaction to a future earnings announcement. The theoretical motivation for this hypothesis is drawn from a literature review encompassing the research domains of financial statement analysis, and earnings forecasting.

For the empiric evaluation of the hypothesis, a range of machine learning models and traditional linear models are trained to predict the abnormal return of a stock following its earnings call based on 121 variables from the balance sheet, income-, and cashflow-statement. These quarterly financial statement variables for the entire North American stock market are taken from the Compustat FUNDQ data set between the years 1991 and 2017, and the critical problem of missing values in the data is addressed with a comprehensive pre-processing strategy. The abnormal stock return serving as the dependent variable is computed as the Buy-And-Hold Abnormal Return (BHAR) of a particular stock until 30 days after the earnings announcement.

The design of the experiments involves model setups formulated as regressions and classifications, which are evaluated using an out-of-time and out-of-sample test sample. For it, the models demonstrate an ability to predict the correct sign of the abnormal market reaction in the majority of the cases using a newly introduced metric called PC. Additionally, the model predictions are demonstrated to translate into positive quarterly returns when employed in a simulated trading strategy over the study period. These task-specific evaluation metrics confirm a superior performance of the random forest and neural network over traditional linear models like the Lasso regression.

Machine Learning based Financial Statement Analysis predicting the Market Reaction to Earnings Announcements

Daniel Kaiser

Mansfield College

University of Oxford

A thesis submitted for the degree of

Master of Science (by Research)

Trinity 2019

Acknowledgements

Thank you.

I would like to thank my thesis advisors Jan-Peter Calliess, Amir Amel-Zadeh, and Steve Roberts, for the countless ways in which they made this thesis possible. Their advice, knowledge, encouragement, and funding, are the factors that enabled this work. Particularly in the beginning of my studies when I faced an episode of health problems, their understanding and support were pivotal for the continuation and success of my studies.

Further, I would like to thank my colleagues at the Oxford-Man Institute, like Daniel Poh, Bryan Lim, Xingyue Pu, Yin-Cong Zhi and Utkarsh Sharma, for their feedback, discussions, friendship, and company that they provided during the weekends and nights we spent working and drinking coffee together at the institute.

Additionally I'm grateful for the constructive feedback made by the assessors of my academic transfer of status, Xiaowen Dong and Nir Vulkan, as well as the suggestions and questions raised Stefan Zohren and many others during the brown bag seminar where I was able to present these results.

It was a great honour to have learned and benefited from such a diverse community of excellent minds, inspiring scholars, exceptional intellects, and genuine friends.

Abstract

This thesis considers the question of whether machine learning models can utilise the data contained in past financial statements to predict the market reaction to a future earnings announcement. The theoretical motivation for this hypothesis is drawn from a literature review encompassing the research domains of financial statement analysis, and earnings forecasting.

For the empiric evaluation of the hypothesis, a range of machine learning models and traditional linear models are trained to predict the abnormal return of a stock following its earnings call based on 121 variables from the balance sheet, income-, and cashflow-statement. These quarterly financial statement variables for the entire North American stock market are taken from the Compustat FUNDQ data set between the years 1991 and 2017, and the critical problem of missing values in the data is addressed with a comprehensive pre-processing strategy. The abnormal stock return serving as the dependent variable is computed as the Buy-And-Hold Abnormal Return (BHAR) of a particular stock until 30 days after the earnings announcement.

The design of the experiments involves model setups formulated as regressions and classifications, which are evaluated using an out-of-time and out-of-sample test sample. For it, the models demonstrate an ability to predict the correct sign of the abnormal market reaction in the majority of the cases using a newly introduced metric called PC. Additionally, the model predictions are demonstrated to translate into positive quarterly returns when employed in a simulated trading strategy over the study period. These task-specific evaluation metrics confirm a superior performance of the random forest and neural network over traditional linear models like the Lasso regression.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivation

Financial statements provide a quantitative perspective on a firms' business operation. Comprising of a balance sheet, an income-, and a cashflow-statement, they constitute a comprehensive record of past business operations for management and investors. Corporations traded on stock exchanges in the United States are legally required to disclose financial statements to the public in quarterly periodicity by filing reports to the *Securities and Exchange Commission (SEC)*. Critical components of financial statements, such as earnings, total sales, and other information that analysts are interested in, are initially published on the day of the earnings announcement, while the complete financial statements are usually disclosed a few days after this date.

Investors anticipate the information content of these publications with forecasts,

and their reaction to the published actual numbers gives rise to the documented phenomenon of the *earnings surprise*, referring to the degree to which the stock price of a firm experiences abnormal returns based on whether pre-announcement expectations are met, exceeded, or missed. While this initial reaction of the stock price is consistent with the efficient market hypothesis that suggests a market movement due to the incremental information gain, financial research found that usually, the stock price continues to drift in the direction of the reaction for up to a month after the announcement. This anomaly is referred to as the *Post Earnings Announcement Drift (PEAD)* and was first noted in the study by Ball and P. Brown (1968). Since then, it has become one of the most widely researched stock market anomalies in financial literature.

The significant abnormal returns that result from PEAD suggest the viability of trading strategies that take long or short positions in firms before an earnings announcement. Since these announcements are scheduled events, traders that expect a particular direction of the stock price reaction could enter positions in liquid equities on the day before the announcement. If they expect a negative price reaction over the period where PEAD is observed, they could enter a short position, and for an expected positive stock price reaction to the event, they hold a long position. The prior knowledge upon which traders could base their expectation includes public and private information about the firm. While the latter in many cases constitutes illegal insider trading, the former is contradicted by the semi-strong-form efficient market hypothesis which suggests that stock prices fully reflect public information.

Recently, accounting researchers adopted machine learning methods for problems relating to financial statements like fraud detection (Perols, 2011), firm bankruptcy prediction (Barboza et al., 2017), and textual sentiment analysis (Hajek and Henriques, 2017). These applications show promise in demonstrating how traditional problems related to accounting can be addressed by computational forecasting methods. The information contained in financial statements is of high dimensionality, has intertemporal relations to past and future financial statements, and can contain patters across industries. These features of the data might present a challenge

to human analysts that aim to derive insights from it to trade over earnings announcement events. Therefore the task lends itself to test the viability of a machine learning-based approach to generate signals based on which positions are taken. The existence of these signals could defy the semi-strong-form efficient market hypothesis as computational methods might be superior to the average trader at synthesising the data available in financial statements. The purpose of this thesis is to evaluate this hypothesis.

A separate but related challenge that has been considered in the accounting literature is the task of automated earnings forecasting. In their seminal paper, Ou and Penman (1989) were among the first to investigate whether past financial statements contain information that is useful for that task. Compiling 28 variables from financial statements, they successfully trained a logit model forecasting the one-year earnings increase (decrease) of firms. Their study received much attention as they were able to derive a profitable investment strategy based on their findings by holding (shorting) stocks over 12 and 24 months as guided by the model.

Earnings prediction is related to predicting PEAD, in that the market reaction (i.e. the share price response) to earnings correlates with the direction, and magnitude of the unexpected earnings (decrease) increase (Foster et al., 1984; Bartov et al., 2002). As the studies following Ou and Penman (1989) suggest that earnings can be predicted based on financial statements, it appears plausible that with sufficient reliability, financial statements might also facilitate the prediction of just the market reaction. Particularly as this market reaction is also sensitive to other published accounting figures, e.g. revenues (Ertimur et al., 2003; Jegadeesh and Livnat, 2006), financial statement based predictions could be better suited for the prediction of the market reaction than to merely forecast future earnings.

Of particular interest for this thesis is the question of whether besides the direction, also the magnitude and severity of the future market reaction can be predicted. Skinner and Sloan (2002) and Kinney et al. (2002), notice that already minor misses of expectations can yield significant negative stock reactions, and Kasznik and McNichols (2002) show that minor positive surprises lead to

disproportionately positive price reactions. This suggests the existence of significant non-linearities. Therefore, one would expect non-linear (machine learning) methods to outperform traditional linear models in this domain.

1.2 Contribution

This thesis demonstrates that machine learning methods can be employed to generate models that are capable of forecasting the sign and magnitude of the abnormal stock market abnormal + PEAD, as measured by Buy-And-Hold Abnormal Returns (BHARs), on the basis of financial statement data alone. Training a range of models on a time series of financial statements to forecast significant directional PEADs reactions, it compares their relative out-of-sample prediction performances in an expanding window setup for the quarters between 1990 and 2018. To illustrate the ability of the forecasts to yield abnormal return, multiple trading strategies are backtested that use generated signals to make adaptive decisions on long-short-neutral positions for a comprehensive portfolio of publically traded stocks in North America.

The experiments are constructed so that the quantitative part of the four last published financial statements of a particular firm is used to predict the 30 day market reaction (i.e. BHAR) to the publication of the next upcoming financial statement. This prediction is done without knowing the actual figures contained in the upcoming financial statement publication so that this study doesn't require the hypothetical investor to trade in the brief window between the publication and when the stock market opens again but so that it is possible to decide on positions on quarter ahead. The small time window between the release and market open is marked by large big-ask spreads, out of market trading, and volatility.

While the literature supports the notion that there is a direct relationship between the just released financial statements and the stock market returns following it (see Earnings Response Coefficient), evidence that earlier financial statements could predict the future stock market reaction contradicts the efficient market

hypothesis which assumes that all public information ought to be priced into the stock. Sophisticated analysts however might explicitly model the to be released financial results and thereby see if the stock will be undervalued or overvalued after the release of the information. The machine learning conducted in this thesis follows this intuition but leaves it to the internals of the computational model to implicitly these future results in order to predict the market reaction.

A thorough review of the relevant literature suggests that this thesis is the first study that attempts to forecast stock market reactions to earnings announcements directly by using a financial statement analysis approach. Further to the best of the authors knowledge, there have not been works investigating the viability of modern machine learning methods for financial statement analysis beyond the studies using logistic regression models following the original study by Ou and Penman (1989). While the trading strategy is a simple simulation, the thesis also includes a comprehensive discussion of how signals can be derived from the predictions of models in both a regression and classification setup.

While the trivial simulated trading strategy yielded abnormal returns, these returns are not statistically significant to disprove the efficient market hypothesis. The performance of a more sophisticated strategy that encompasses technical signals, risk weighting, and other sources of data, could possibly do that to a significant degree.

The thesis is structured as follows. After this introduction chapter, the literature review in chapter 2 discusses the existing financial literature in which this thesis is embedded. It includes the finance and accounting theory that surrounds earnings announcements and their stock market reaction and concludes by discussing the findings of the most closely related studies.

Chapter 3 introduces the data set used in this study and how it is pre-processed to be used for the experiments. Most importantly, it addresses the problem of missing values in financial statements and explains how the problem was adressed. Chapter 4 explains in detail the choice of machine learning methods used in the study. It

outlines the hyper-parameter combination of the employed models and serves as an introduction to the model types that one would find in a technical literature review.

The *Experiments* chapter 5 contains all the empirical results of the study. The first two sections discuss the raw results of the regression type model setup in contrast to models specified as a classification. Following these discussions of the results concerning the main hypothesis of the thesis, the subsequent parts of the chapter focus on resulting questions that are of interest to financial research. These sections focus on an investigation of the role particular segments of market capitalisations across firms play for the returns and the questions which variables have been most important for the model predictions. The final chapter 6 summarises the findings in this thesis, discusses its implications for theory, and points to potential directions for future research.

2

Literature Review

This review serves as an introduction to the background and existing literature or financial statement analysis, earnings forecasting, and statistical modelling by previous work in that context. These fields form the corpus of finance and accounting research that this work ties in and contributes. The machine learning models used in the thesis are introduced in chapter 4 together with their specific hyper-parameter configurations, and discussions of how the models relate to the problem setting of the study.

2.1 Financial Statement Analysis

The term financial statement analysis refers to the process of analysing a company's financial statements to conduct economic decisions. These usually concern investment decisions by investors comparing firms to allocate capital, and operational decisions by the management of a firm to improve profitability. As such, the release

of financial statements represents a scheduled important incremental information gain for capital markets.

2.2 Efficient Market Hypothesis

Developed by Fama, Fisher, et al. (1969) and Malkiel and Fama (1970), the efficient market hypothesis provides a framework in which to interpret how stock prices respond to information. Defining an efficient market as a state where "security prices fully reflect all available information" (Fama, 1991), the efficient market hypothesis notes that multiple forms of efficiency exist:

1. A weak-form that implies that markets are efficient enough that future prices are independent of past prices and thereby follow a random walk (Bachelier, 1900). In this case, a fundamental analysis could achieve abnormal returns while technical analysis that just interprets a time series of stock prices ought to fail.

2. A semi-strong-form, where neither fundamental analysis nor technical analysis could achieve abnormal returns as all public information is reflected in the stock price instantaneously after it is published.

3. A strong-form, which is similar to the semi-strong-form, but assumes that private insider information leaks and is also reflected in the stock price instantaneously.

Models like the Gordon Growth Model (Gordon, 1962) derive the value of a firm in an efficient market by discounting risk-adjusted future earnings and cash flows. The original "Statement of Financial Accounting Concepts No. 1" by the Financial Accounting Standards Board (1978) explicitly states this focus of investors on future cash flows in it's practical guidance to accountants. Paragraph 37 about the objective of financial reporting reads:

> Financial reporting should provide information to help present and potential investors ... in assessing the amounts, timing, and uncertainty of prospective cash receipts from dividends .. of securities. The prospects for those cash receipts are affected by an enterprise's ability to generate enough cash .. and may also be affected by perceptions of investors and creditors generally about that ability, which affects market prices of the enterprise's securities. Thus, financial reporting should provide information to help investors, creditors, and others assess the amounts, timing, and uncertainty of prospective net cash inflows to the related enterprise.

Subsequently, we should be able to assume that financial statements have the imperative to include sufficient and reliable information for investors to determine how the cash flows of a firm might change and persist over the future.

Earnings and cash flows differ by accrual adjustments which are included in earnings. This suggests cash flows to be more predictable than earnings as they are not estimated and can only be manipulated by accountants via real incurred costs to the firm. The study by Sloan (1996) highlighted this by finding a differential in the persistence of earnings through cash from operations (more persistent) and accruals (less persistent). Although the paper demonstrated that the cash components in earnings are more predictable than the transitory accrual components, it also shows that earnings alone are a better predictor of cash flows than cash flows themselves are. Dechow and Dichev (2002) provide further evidence for this by demonstrating that 'accrual quality' (influenced by firm characteristics) is positively correlated to earnings persistence. Seeing how accruals as the more volatile component of earnings can be related to firm characteristics, and given that earnings are predictable based on past accruals and cashflows, supports the hypothesis that financial statement analysis ought to further improve the predictability of earnings.

2.3 Earnings Announcements

As expectations about future earnings and their actual realised values can differ, the moment of when realised earnings and financial statements are published has been the focus of many works in accounting research. Among the first studies

looking at these earnings announcements were Ball and P. Brown (1968) and W. H. Beaver (1968) conducting pivotal work into a form of research now referred to as 'event studies'.

Ball and P. Brown (1968) found a significant positive correlation between the sign of the abnormal returns in the month when a firm conducts an earnings announcement with the sign change of earnings in relation to the previous year. As such, they find evidence that revised expectation of better earnings represents good news to investors yielding positive abnormal returns, while a negative surprise of earnings leads to negative abnormal returns.

W. H. Beaver (1968) looks at the association between the price volatility and trading volume of stocks around earnings announcements. Finding a correlation between the two, he argues that earnings announcement have information content based on the observation that compared to weeks in which no reporting happens, weeks in which earnings are announced have abnormally high trading volume and price volatility.

The study by Ball and P. Brown (1968) was also the first study to document the phenomenon of the post-earnings announcement drift (PEAD). Called the "best-documented and most resilient capital markets anomaly" (Livnat and Mendenhall, 2006) for its resilience against the efficient market hypothesis , it refers to a drift of the stock price in the direction of the initial reaction to unexpected earnings for many weeks after the announcement. It contradicts the efficient market hypothesis as it would suggest a single immediate price reaction (as in a stock price jump) to the published information. Because of this phenomenon, this thesis extends the trading window for up to 30 days after the announcement to include the majority of the PEAD drift. As most of the stock price reaction should happen around the release of the information, these 30 days should be sufficient.

Many studies about the PEAD have attempted to explain it by:

- behavioural considerations of investors (e.g. PEAD is the result of a delayed price response (Bernard and Thomas, 1989));

- information uncertainty (e.g. PEAD is a function of how much media and analyst coverage a stock receives and how high the composition of institutional investors is (Z. Yan and Y. Zhao, 2011; L. D. Brown and Han, 2000));

- limits of the profitability of exploiting the anomaly because of transaction costs (e.g. PEAD exists predominantly among smaller (Bhushan, 1994; Foster et al., 1984) and illiquid (Chordia et al., 2009) stocks.

Mendenhall (2004) corroborates all three hypotheses and points out that PEAD varies strongly with the real world arbitrage risk of investors that takes a position in the stock and hedges the position via market indexes. These studies together imply that there are cross-sectional firm characteristics (usually expressed in financial statements) and market conditions that are correlated with PEAD. Besides the original north american studies, multiple works have found evidence of PEAD in equities in international equity markets like New Zealand (Truong, 2010), Poland (Forbes and Giannopoulos, 2015), or China (Truong, 2011).

2.4 Earnings Response Coefficient

Following the findings of Ball and P. Brown, 1968, there has been a range of works investigating the relationship between earnings and share prices under the name of "Earnings Response Coefficient" (ERC).

The pioneering paper on this line of research has been by Kormendi and Lipe (1987), where they focus on the magnitude of the relation between earnings and stock returns and examine how this magnitude relates to the time-series properties of the firm's earnings. They build on the work by M. H. Miller and Rock (1985) that found that the magnitude of the market reaction to an earnings innovation (i.e. the novel information in earnings) should be a function of the past earnings persistence. Concretely, the study tests the hypothesis whether the magnitude of the relation between stock returns and earnings (i.e. the earnings response coefficient) depends on the earnings persistence. The results of their cross-sectional study support this

relation across firms, while also finding that stocks are not abnormally sensitive to earnings innovations, exhibiting no "excess volatility" of stock prices relative to what classical valuation theory would predict.

Consistent with the findings of Kormendi and Lipe (1987), Easton and Zmijewski (1989) find that the more persistent earnings are (i.e. the fewer earnings innovation the time series exhibits), the bigger the earnings response coefficient is. They investigate this through the earnings revision coefficient, which relates current earnings to future earnings.

In their empirical study of cross-sectional determinants of the ERC, Collins and S. Kothari, 1989 find that the earnings response coefficient varies negatively with the risk-free interest rate and systematic risk (also noted by Easton and Zmijewski (1989)) while varying positively with a firm's growth prospect and earnings persistence. Their intuition behind the risk-free interest rate stems from the logic that if this rate rises then the discounted present value of expected future earnings innovations falls, (all things equal) leading to a negative association between the rate levels and the ERC. They note that ERC is positively related to earnings persistence and economic growth opportunities. In their study Collins and S. Kothari (1989) note that:

1. The ERC is positively related to earnings persistence and economic growth opportunities.

2. The ERC is negatively related to the securities' future expected discount rates.

3. The discount rate is made up of (i) the risk-free interest rate R, and the market risk premium, and the firms' CAPM beta risk.

4. Because R, and the market risk premium are the same for all firms, they obviously are not a source of cross-sectional variation in ERCs.

5. The ERCs are negatively related to the interest rate levels through time and the CAPM beta risk in the cross-section.

Based on these initial findings on the determination ERC, a range of other works also found interesting factors that play a role. These factors are important for our study, as the magnitude of our independent variable is closely related to the magnitude of the ERC. Ideally, we have financial statement variables, that act as proxies for some of these ERC determining factors.

Notably, Ahmed (1994) looks into variables related to financial statements as he investigates the hypothesis by Biddle and Seow (1991) of whether the ERC is related to cost structure (i.e. the ratio of fixed costs to total costs) of firms. Controlling for risk measures, they find that cost structure has a positive effect on ERCs and additional results demonstrating that growth opportunities have a negative effect on ERCs. The finding of growth opportunities is contrasted in the study by Anthony and Ramesh (1992) looking to explain ERC based on the firm life cycle and strategy. Proxied by sales growth and capital investment, and using cross-sectional data for various life cycle groups, they find evidence of a positive correlation between the ERC and young firms. Established "old" (low sales growth) firms have relatively lower ERCs.

Kinney et al. (2002) conduct a statistical analysis on the relation between the abnormal stock returns surrounding the announcement of annual earnings and the magnitude of the earnings surprise. They define the earnings surprise as the difference between the realised earnings and the analyst consensus forecast from a dataset by First Call Corporation. In line with Skinner and Sloan (2002) and Kasznik and McNichols (2002) they find small negative earnings surprises that resulted in disproportionally large negative returns, and small positive surprises yielding large positive returns. However, they find that the sign of surprises and returns only have the same direction in 55% to 57% of cases. The lack of a clear relationship between earnings surprise and the significant market reaction suggests that there could be other factors influencing the market reaction. Ideally, the models employed in this thesis are able to identify these factors.

The study by Freeman and Tse (1992) finds the marginal stock price response to unexpected earnings to decline as the absolute magnitude of unexpected earnings

increases. This evidence stands contrary to the assumption that the ERC is linear instead, suggesting that studies employing linear regression models on the ERC to underestimate the returns-earnings relation. The found non-linearity supports our hypothesis that machine learning methods are better suited for our model than traditional statistical methods.

Looking at *contrarian* share price reactions, Johnson and R. Zhao (2012) investigate cases where earnings surprises and market reactions have opposite directions. They find that such contrarian stocks exhibit little post-earnings announcement drift and that they are usually related to noise in the measured earnings surprise and the market response. The factors they identify as driving noise in the earnings surprise are pre-announcement stock returns and the presence of GAAP exclusions, while the noise in the share price response is primarily driven by surprises in revenue changes, analyst earnings forecast revisions, and discordant prior quarter earnings surprises. While their study does not look at the phenomenon in the framework of earnings response coefficients, their findings highlight interesting cases of when the direction of earnings surprises and market reactions diverge.

2.5 Modelling Earnings

Due to the widespread focus of investors and the public on earnings, a large corpus of literature has emerged over the last decades looking into their time-series properties and potential forecasting methods. L. D. Brown (1993) provides a comprehensive review of these.

2.5.1 Random Walk (with drift)

A large corpus of early studies tried to explaining earnings using a stochastic random walk (with drift). While the efficient market hypothesis and economic theory support the assumption for stock prices to follow a random walk, there is however, no theoretical basis for why earnings should follow a random walk.

Ball and R. Watts, 1972 were among the first to find evidence suggesting that accounting income represents a submartingale or some very similar process. Multiple subsequent studies (Albrecht et al., 1977; R. L. Watts and Leftwich, 1977) confirm these findings by comparing an autoregressive moving average (ARMA) Box-Jenkins model (Box et al., 1970) against a random walk model, finding the random walk model to perform better. The hypothesis of this thesis stands in contrast to the random walk model however, as it assumes a level of predictability of future earnings.

The big problem with time series models applied in traditional earnings studies is that they usually require long study periods of firms as they concern yearly earnings realisations. As Fama and French (2000) note, this creates survivorship bias because it excludes firms that go bankrupt before they reach that period. This makes model estimates just applicable to a subset of firms and on average leads to large errors for all firms, providing evidence for the discussed random walk properties.

2.5.2 Mean Reversion

Some research questioned the random walk properties by arguing that economic theory suggests that profits and its components should be mean-reverting within and across industries. In a competitive free-market economy, the argument stands that abnormal profitability should not be sustainable, due to firms entering the market. Abnormal losses should, however, also not be sustainable as it leads firms to the prospect of bankruptcy and liquidation, enticing management to direct efforts to return back to mean profitability.

The early study by Brooks and Buckmaster (1976) provided the first evidence for this noticing that time series of income indeed have a tendency for income to revert back to previous levels in the period subsequent to one where it has a substantial deviation. They, however, do not determine whether this can be attributed to the firm management smoothing the income numbers, or whether it was emerging due to the inherent income determination process. Fama and French (2000) pick up on these results and conduct a cross-section study finding that the rate of mean reversion is about 38% per year. Yet they find it to be highly

non-linear by depending on whether profitability is below the mean, leading to a generally faster mean reversion than if it's above, and when it's extremely far from the mean in either direction. Their results concerning profitability also extend to earnings which they find to reverse even faster.

2.5.3 Price based predictions

As market prices of securities react to news and macroeconomic conditions during the year between the announcements of earnings, some works have looked into whether prices can be used to predict earnings. From an investor's perspective, prices have an imperative to approximate the estimated value of a firm and expected earnings as accurately as possible ex-ante, since any difference between the ex-ante price and assumed value ex-post the release of the earnings, would open profit opportunities. On the contrary, however, from a fundamental valuation perspective, prices are understood to incorporate all future earnings with their assumed time-series dynamics. Therefore they might not be particularly indicative about the upcoming earnings in relation to last earnings.

To test this hypothesis of using prices to predict earnings W. Beaver et al. (1980) invert the price-to-earnings ratio simply using the price as a predictor for earnings. Doing that it was the first study that used exogenous factors to explain the time-series properties of earnings instead of using merely previous earnings realisations. They find that security prices behave as if earnings do not follow a simple random walk, and present preliminary evidence that, according to their nomenclature, 'raises the possibility' that price-based prediction models are more accurate than plain random walk (+drift) models. Collins, Stephen P Kothari, et al. (1987) and Freeman (1987) look at this predictive ability more accurately in the context of firm size in relation to the predictive information content in security prices and find that price based forecasts work better for large firms than for smaller firms.

2.5.4 Financial Statement Analysis

In their foundation paper Ou and Penman (1989) (O&P) laid the grounds for financial statement based predictions of earnings. They were able to show that accounting variables selected based on univariate significance can predict whether future earnings increase or decrease through a Logit model. Estimating this probability (of an earnings change) through a summary measure called Pr, they constructed portfolios and which realised positive returns over a subsequent 12 and 24 month period between the years of 1973 and 1983. Their study was pivotal as it demonstrated that the information in financial statements can be used to conduct earnings forecasts. Due to the studies popularity, multiple further works re-examined and replicated their findings up until today.

Questioning the O&P results Greig (1992) found that once Pr has been controlled for firm size and CAPM (Capital Asset Pricing Model) risk, no significant incremental explanatory power for abnormal returns can be attributed to it. The study notes that all returns found by O&P are a manifestation of a 'size effect' and not evidence for a market inefficiency based on poor earnings forecasts.

Holthausen and Larcker (1992) follow a similar model as O&P, but instead of predicting unexpected earnings, they directly model excess returns similarly to our study. Their approach differs from the one in this thesis, as they use data from the end of the fiscal year to predict excess returns for 12 months following the release of the financial statement data in April. This thesis, in comparison, restricts itself to predicting the 30 days returns following the future quarterly earnings announcement for which the information content is unknown. Thereby it predicts a stock market reaction to a quarterly event, Holthausen and Larcker (1992) concern themselves with predicting what the 1 year stock return will be from the publication of a financial statement. Taking a subset of the variables proposed by O&P, their strategy achieves annual excess returns of 4.3% to 9.5% for the period of 1978 to 1988. Replicating the original O&P study for this period, they found no significant excess returns.

This thesis decides to take a 30 day limit for the measuring of PEAD as it seems appropriate for quarterly results. While studies find PEAD to persist for up to 90 days, these were usually yearly earnings publications with slower reacting stock markets multiple decades ago. Most of the PEAD movement these days happens shortly after the release of the information. Since everything that comes after the (abnormal) market reaction is just noise that distorts the returns, it makes sense to rather limit the PEAD window to around 30 days. No experiments have been conducted for any other choice of the PEAD window.

Lev and Thiagarajan (1993) also examine the relations between financial statement variables (i.e. fundamentals) and excess returns. They find that their selection of fundamentals adds about 70% to the explanatory power a mere earnings-based prediction model. Compared to O&P, they find their set of value relevant fundamentals not through a statistical search but guided by analysts' descriptions, publications, professional commentaries on corporate financial reporting, and newsletters of securities firms. They note the set of 12 fundamentals they select to be value relevant between the period of 1974 to 1988, demonstrating (consistent with Holthausen and Larcker (1992)), that fundamentals are predictive of future abnormal returns.

Abarbanell and Bushee (1998) also demonstrate that the information contained in financial statements can provide strong enough signals to yield abnormal returns of 13.2% over 12 months. Relevant for our study, they find that abnormal returns over their one-year forecasts are concentrated around a three-day window around subsequent quarterly earnings announcements. While this small window size captures the last two days of the market movement anticipating the earnings announcements, the initial earnings reaction, and the first day of post-earnings announcement drift, this evidence suggests the fundamental based model could have predicted these reactions alone as well.

Seng and Hancock (2012) corroborate the study by Abarbanell and Bushee (1998) and extends the methodology to global data from 33 countries. They find that a model with fundamental signals from financial statements and the

current change in earnings is superior in predicting future earnings changes to a model merely using earnings changes as an input. Methodologically they conduct the study using regressions and evaluate results via the R^2 score. They also found evidence that incorporating variables about the macroeconomic context like inflation and GDP improves the models.

Bird et al. (2001) replicate the OP model for a more recent period (1983-1997) and extended the analysis to the UK and Australian market. While they found a similar predictive performance as the O&P study, their 12 significant fundamental variables (compared to the original 18 by O&P) didn't yield a profitable investment strategy.

Stober (1992) investigates the difference between the predictions about one-year earnings changes of the O&P model and analysts forecasts (in the I/B/E/S dataset). They find that in general analysts' forecasts are marginally superior to the O&P Pr measure, correctly predicting 54% of the cases compared to about 46% correct predictions via Pr. Subsequently, they looked at the cases distinctly where both predictions differ. In the samples where they disagree, despite a superior prediction performance of analysts' forecasts, the Pr strategy generates significant abnormal returns. The paper explains this by suggesting that Pr might be a proxy for cross-sectional differences in expected returns, capturing information that the market is slow to recognise.

Some methodological studies have looked into whether neural network-based predictions of earnings can improve upon the results that traditional Logit models achieve. Zhang et al. (2004) find more accurate neural net-based forecasts for EPS via fundamentals than with traditional linear models. Cao and Parry (2009) improved upon the neural network approach with a genetic algorithm that was significantly more accurate.

Modelling earnings realisations alone, similar to the traditional studies testing the time series properties of earnings to confirm (reject) the random walk hypothesis, Callen et al. (1996) compare neural networks to linear time series models and find an inferior performance. Similar to this study, Shen (2012) uses the components

of earnings (i.e. accruals and cash flows) of the last five quarters as an input in a neural network to predict the earnings of a future quarter.

Falas et al. (1994) compare artificial neural networks (ANN) to the Logit model of the O&P study using just nine accounting variables. They train their models on a period from 1982 to 1986 and test it on the years 1987 to 1991. On average, the Logit model predicts the correct direction of the future earnings change in 62% of cases, while the ANN predicts it correctly in 62,9% with just one hidden layer of 3 neurons.

While the early studies looked at market movements related to predicted year-over-year earnings changes in a long term of many months, some recent works started to look at the earnings surprise and the immediate earnings reaction to quarterly earnings.

The study by Zhang et al. (2004) looks into whether the returns of a "good (bad) news" portfolio (with actual EPS higher (lower) than estimated ones) correspond with cumulative abnormal returns over a (t-32,t+1) anticipation window, and (t-32, t+32) anticipation & reaction window relative to the announcement day. Unfortunately, they do not investigate the reaction window (t-1,t+30) alone (i.e. the period our study investigates). They find the multivariate neural net model to outperform all of the seven other models they specify in terms of all the error measures.

Very closely related to this thesis is the study by Dhar and Chou (2001) in which, they compare four non-linear models (i.e. artificial neural nets, a genetic algorithm, CART, Näive Bayes) on their ability to predict whether firms deliver an earnings surprise. Similarly to this thesis, they then take positions based on predictions, that they liquidate 40 days after the announcement. Their study differs from a financial statement analysis strategy, like our work, as they merely include two synthetic fundamental variables (i.e. cash flow return on investment, discount rate differential). The rest of their 18 variables are earnings expectation based variables derived from consensus analyst estimates, and technical stock price variables derived from industry trends. Consistent with our findings, their study finds a relationship between earnings forecast errors and company size showing

that larger companies with more analysts covering them tend to have smaller absolute earnings surprises than small firms. In terms of the models, they find the genetic algorithm to deliver the best results, followed by Näive Bayes, CART, Neural Nets, and a linear model. Unfortunately, their study does not investigate whether their best performing model (i.e. genetic algorithm) provides sufficiently strong results for a profitable investment strategy if the results vary within the time period, and how the test sample was created.

The study by X. Yan and Zheng (2017) investigates the general ability of fundamentals from financial statements to predict stock returns. They create over 18,000 fundamental signals from financial statements with a data mining approach that creates various ratios and combinations of about 220 original variables. Predicting cross-sectional stock returns, they find evidence that a large number of the generated fundamental signals are significant predictors. While their study doesn't concern the abnormal returns around the earnings announcements as our work does, it suggests that financial statements contain information that is indicative of future stock returns. The new variables the study find to be most indicative are changes in interest expense, levels of short-term debt, tax loss carryforward, and selling, general, and administrative expense. In the study, the authors provide plausible explanations of why these factors are economically meaningful to predict returns.

2.5.5 Natural Language Processing

Besides the noted studies looking into applying machine learning methods to the quantitative parts of financial statements, we also found works using NLP (Natural Language Processing) methods to model the textual content of quarterly filings. We include some of these works in our literature review to assess the potential usefulness of features extracted from the text. It could be the case that the textual parts of financial statements contain better and more comprehensive information than what is expressed in accounting variables. Studies like Bryan (1997), G. S. Miller and Piotroski (2000), Callahan and R. Smith (2004), and Davis et al. (2006), suggest that future earnings can be related to the tone (i.e. "What managers

mean") of management disclosures , while studies like Li (2008) and Mayew and Venkatachalam (2012) relate them to transparency and truthfulness (i.e. "How managers say it") of statements.

The study by Li (2010) analyses the information content of forward-looking statements (FLS) in the management discussion analysis section of quarterly (i.e. 10-Q) and yearly (i.e. 10-K) Securities and Exchange Commission (SEC) filings. Using a Naïve Bayesian machine learning algorithm, the study finds indeed evidence that the average tone of FLS is positively associated with future earnings. Further, they note that also the 'constitution' of a firm is reflected in the tone of FLS. 'Good' firms with better current performance, lower accruals, smaller size, lower market-to-book ratio, less return volatility, and longer history tend to have more positive FLS. These results suggest a potential extension to our work by including features extracted from FLS in quarterly reports.

2.6 Discussion

Since the 1970s, an extensive corpus of Accounting and Finance literature has formed around earnings forecasts, financial statements analysis, and the market reaction to earnings announcements. This extensive literature review suggests that the research direction of this thesis is novel as no study has looked into applying machine learning-based financial statement analysis to predict the market reaction to earnings announcements directly.

The most closely related work to this thesis is the study by Dhar and Chou (2001) in which an array of inputs that include a few fundamental signals are used in conjunction with a set of non-linear machine learning models in a classifier to predict whether a firm delivers an earnings surprise. Since they predict the direction of the earnings surprise, in particular, their performance at this task will be compared to the approach of this thesis in this setting.

The works discussed in this literature review support the study design of this thesis since

- financial statement information has been shown to help the predictability of earnings,
- earnings surprises are correlated with the market reaction, and factors impacting the earnings response coefficient include items expressed in financial statements,
- machine learning methods can outperform traditional linear and logistic regression models in earnings-related prediction tasks, and,
- the focus on quarterly financial statements provides a large sample size to train viable machine learning models.

3

Data

This chapter discusses the data sources for the experiments of this thesis. To support reproducibility, the most important part of the subsequent sections is a detailed description of how the data is prepared and preprocessed. These operations are necessary as the data in its original form can not be used directly for the following experiments. The concluding section outlines the expanding window construction of the training- and testing-set used for the validation of the results.

3.1 Financial Statement Data - FUNDQ

The data source for the financial statement related data is the Compustat FUNDQ file provided for research purposes by the Wharton Research Data Services (WRDS)[1]. This data represents the independent variables in the model setup, which are taken as predictors for the abnormal returns following the earnings announcement.

[1] https://wrds-web.wharton.upenn.edu/wrds/index.cfm

The FUNDQ file covers the entire North American stock market from 1987 to 2018. For this thesis all variables from the balance sheet, income statement, and cash flow statement are considered in quarterly periodicity[2]. In the data set one row corresponds to a company quarter for 27,410 companies resulting in a total of 1,567,486 rows.

A crucial issue related to the Compustat data set is the problem of missing values. Due to changes in accounting standards over the long time span of the data set, and different accounting requirements among industries, about 63% of values in the original data set are missing. Figure 3.1 represents an attempt to visualise the temporal structure of these missing values, while figure 3.2 is a figure following a similar format on the role the industry of a company plays.

Very significant in figure 3.1 is the year 2001 that is approximately in the horizontal centre of the figure. In it, a large set of novel financial statement items have been introduced and were mandated to be reported. Most of these new financial statement items (e.g. XOPTQP (Implied Option Expense), RDIPAQ (In Process RD Expense After-tax), ..) have however been faded out again in subsequent years.

Figure 3.2 reveals through red and green horizontal lines that there are sets of variables that are just reported in particular industries. For instance, about a third of all variables are just used for companies that are identified as financial institutions (see the top right corner), and that there are sets of variables are usually reported, but some industries are exempt (see the repeating horizontal red lines in the first quarter of mostly green cells).

[2]A comprehensive set of variable descriptions for all variables can be found at `https://intranet.johnson.cornell.edu/Portals/36/Departments/Research%20Computing/Variable%20Descriptions/Compustat%20Quarterly%20Variables.pdf`

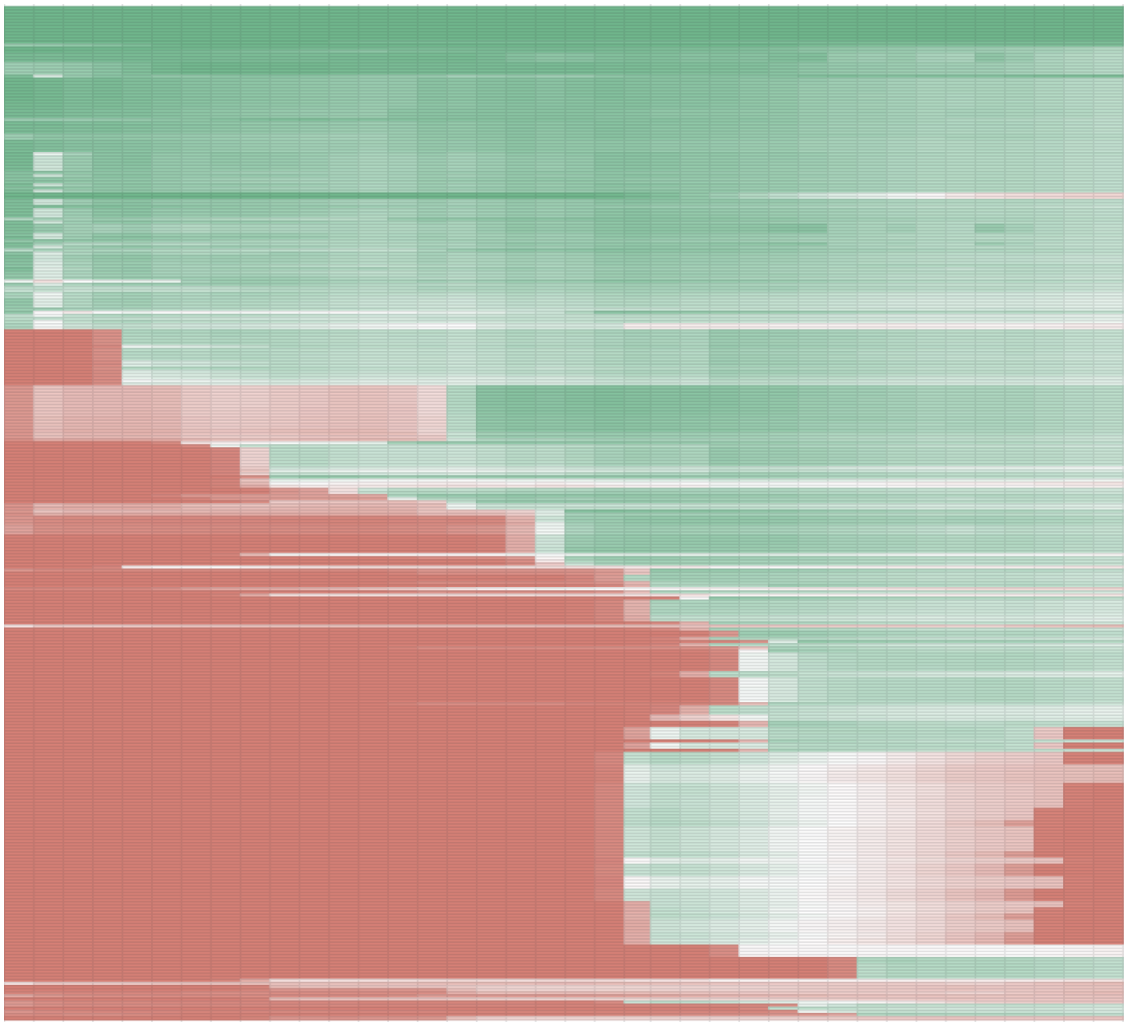


Figure 3.1: This figure attempts to visualise the structure of missing values in the Compustat FUNDQ data. The rows correspond to a variable (e.g. total assets), and the columns divide years from 1983 to 2018. The colour of the cells indicates the relative frequency of missing values of a particular variable in a year. A pure red cell represents a rate of 100% missing, a green cell a rate of 0% missing, and a white cell 50% missing. The rows are sorted based on their overall rate of missing values. Red clusters across variables indicate that the particular financial statement items have been introduced together after a particular year.

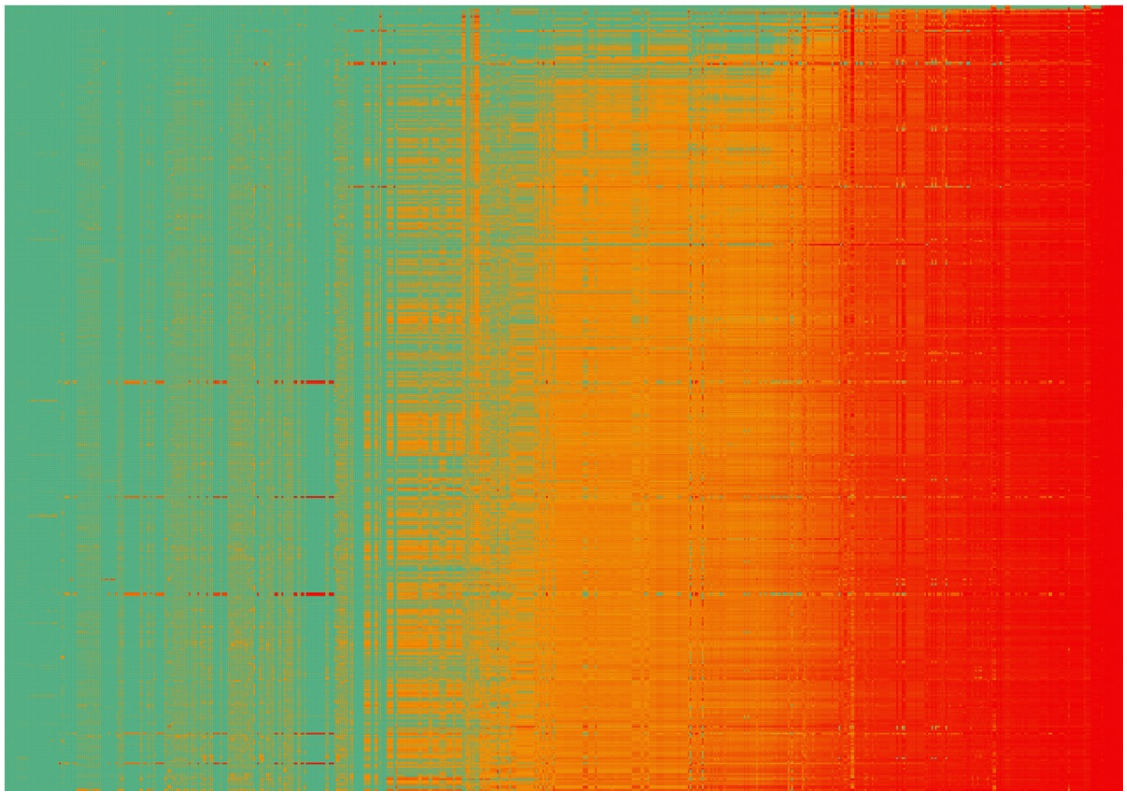


Figure 3.2: This figure follows the format of figure 3.1 to visualise the relationship of missing values in the Compustat FUNDQ data according to the industry of companies. The columns correspond to a particular variable (e.g. total assets), and the rows indicate the sectoral classification according to the Standard Industrial Classification (SIC) code. The colour of the cells indicates the rate of missing values of a particular variable in a year. A pure red cell represents a rate of 100% missing, a green cell a rate of 0% missing, and a yellow cell 50% missing. The rows are sorted based on their overall rate of missing values. Coloured horizontal lines indicate that a particular set of financial statement items is used (green) or irrelevant (red) to a particular industry.

Based on the sectoral and temporal analysis of missing values the total set of available financial statement variables is manually limited to a set of 41 balance sheet variables, 29 cash flow statement variables, and 51 income statement variables as listed in table 3.1. Among the remaining company quarters, all rows where more than 50% of values are missing are also excluded.

In contrast to these raw variables the models could also be trained on ratios that have been identified by research to be predictive. This could include traditional accounting ratios such as in ... or even ratios with temporal macroeconomic variables like in Gu et al. (2018). Ideally the machine learning models would learn to synthesise these concepts or latent economic variables on their own, but providing them explicitly might improve the models anyway. Future studies might explicitly model accounting ratios identified by Ou and Penman (1989), Holthausen and Larcker (1992), and Abarbanell and Bushee (1998) together with results from some mode modern data mining approaches such as by X. Yan and Zheng (2017).

During the period of financial statement reporting that this thesis investigates the north american accounting standard GAAP has been in place. In Europe, on the contrary, the International Accounting Standards (IAS) were replaced by the International Financial Reporting Standards (IFRS) in 2001. Currently the European Securities Market Authority (ESMA) is also implementing a mandate for public corporations in the EU to report their yearly financial statements in the standardised XBRL format also used by the SEC. This change will make studies like this one replicable for the first time using public IFRS accounting data.

Table 3.1: Selected Compustat FUNDQ Variables

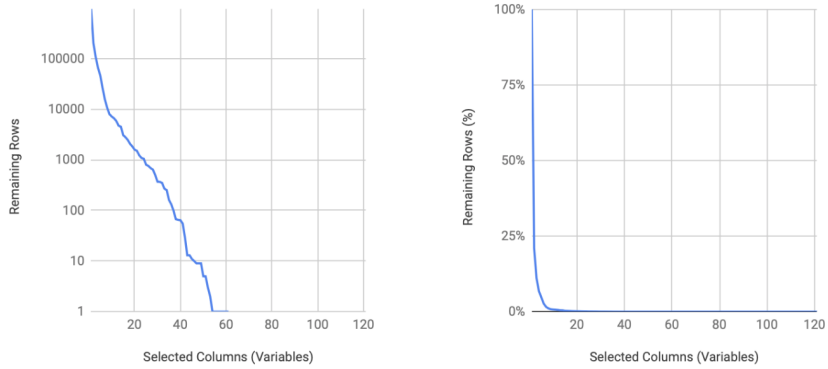
Category	Compustat Variables [3]
Balance Sheet	acoq, actq, ancq, aoq, apq, atq, capsq, ceqq, cheq, csh12q, cshoq, cshprq, cshpry, cstkq, dlcq, dlttq, dpactq, epsx12, icaptq, invtq, lcoq, lctq, lltq, loq, lseq, ltmibq, ltq, mibq, mibtq, ppegtq, ppentq, pstknq, pstkq, pstkrq, rectq, req, seqq, tstkq, txditcq, txpq, wcapq
Cashflow Statement	aolochy, apalchy, aqcy, capxy, chechy, dltisy, dltry, dpcy, dvy, esubcy, exrey, fiaoy, fincfy, fopoy, ibcy, intpny, invchy, ivacoy, ivchy, ivncfy, ivstchy, oancfy, prstkcy, recchy, sivy, sppivy, sstky, txdcy, xidocy
Income Statement	acchgq, cogsq, cogsy, cstkeq, doq, doy, dpq, dpy, dvpq, dvpy, epsfiq, epsfiy, epsfxq, epsfxy, epspiq, epspiy, epspxq, epspxy, ibadjq, ibadjy, ibcomq, ibq, iby, miiq, miiy, niq, niy, nopiq, nopiy, oiadpq, oiadpy, oibdpq, opepsq, piq, piy, revtq, revty, saleq, saley, spiq, spiy, txtq, txty, xidoq, xidoy, xintq, xiq, xiy, xoprq, xopry, xsgaq

[3] A comprehensive set of variable descriptions for all the listed variables can be found online in the documentation at `https://intranet.johnson.cornell.edu/Portals/36/Departments/Research%20Computing/Variable%20Descriptions/Compustat%20Quarterly%20Variables.pdf`

Something that has been criticised by Greig (1992) in the original Ou and Penman (1989) study is an over-fitting of the model on companies of a particular size. To prevent this and to normalise the data, all the independent variables (i,e, financial statement variables) are normalised before being used in the models. This happens on a company quarter basis based on total sales (i.e. variable SALEQ) and total assets (i.e. variable ATQ). All values in the balance sheet are divided by total assets and all values in the income and cash flow statement by total sales. About 2,000 quarterly financial statements where either of these variables are missing or zero are excluded from the sample that constitutes the training and test data. The ranking or grouping of variables has not been conducted but might be potential way to rank missing values in a category.

This initial selection leaves the data set at 868,125 records (55% of the original length) and 121 of originally 720 financial variables where still about 23% of values are missing. For the use of the dataset in the machine learning models, these remaining missing values need to be addressed.

A common strategy in finance and accounting studies has been to limit the number of samples further so that only samples remain where all input values exist. This approach is futile in the setting of this thesis for the reasons that machine learning methods derive their exceptional predictive ability by being able to learn and synthesise insights from large amounts of data, and since such a reduction of samples of would be disproportionally severe with respect to the remaining rows relative to the low amount of selected features. Figure 3.3 illustrates this reduction of training samples with respect to the number of selected columns where values must not be missing.



(a) Absolute number of the remaining rows on a logarithmic scale.

(b) Relative number of remaining rows in percentage of the original amount.

Figure 3.3: These figures chart the reduction of available rows in relation to the number of selected columns where no missing values are allowed to occur. They demonstrate that the approach of reducing the size of the data-set to have a smaller set of training samples with no missing values is futile as already selecting only the twenty most populated columns would result in only 1000 training samples (i.e. 0.2%) of the original data set. For the 107 quarters, this would result in an average of just nine samples per quarter that satisfy the requirement of having no missing values among the selected twenty columns.

The chart in figure 3.3 is created using figures generated by algorithm 1 listed below in pseudo-code. The code demonstrates how the most populated columns are selected iteratively, and the amount of remaining rows (where all values in the selected columns have to exist) is printed.

Algorithm 1 Compute the amount of remaining rows per number of selected columns

```
1: procedure VarRed
2:     remaining_columns ← columns of dataset
3:     selected_columns ← empty array
4:     while elements in remaining_columns do
5:         column ← select the column from remaining_columns
                        where rate of missing values is lowest
6:         add column to selected_columns
7:         remove column from remaining_columns
8:         print amount of selected_columns
9:         print amount of rows with no missing values in selected_columns
```

An alternative strategy to sample reduction employed in the literature is the imputation of missing values. In traditional linear regression settings often this imputation is supplemented with a boolean one-hot encoding of whether a particular value is missing or not (effectively doubling the dimensionality of the input space). Experiments using the reduced FUNDQ dataset and the subsequently employed machine learning models, however, demonstrated a decreased performance when the boolean encodings were added. The diminished performance suggests that negative effects of doubling the dimensionality of the input vector from 121 to 242 variables might outweigh the incremental information about the of missing (i.e. imputed) values.

There is a large body of literature discussing missing value imputation and works like (Garcıa-Laencina et al., 2010; Pigott, 2001) suggest model-based imputation

to be the most promising approach. While statistical imputation imputes values like the mean or median, model-based methods usually conduct a dimensionality reduction to find redundant information in variables that can be exploited.

The tested imputation methods for this thesis include a neural net based auto-encoder, MICE, wkNN, and ultimately the matrix factorisation method Soft-Impute.

The tested auto-encoder option consisted of a denoising autoencoder architecture with a loss function that was adapted to be invariant to missing values (Beaulieu-Jones and Moore, 2017). While this approach appeared promising, it fell short as the neural network based method needed its inputs to be normalised to zero mean and unit variance. Once this operation was undone for the model prediction to arrive at the reconstructed values, the errors induced by the method were exacerbated so that not even the existing values were reconstructed to a promising degree from the lower dimensional encoding.

The MICE (Azur et al., 2011) approach attempts to iteratively impute every variable through other variables filling in the mean as an initial starting value. This approach was unsuccessful as often values were missing and existing together (i.e. in the case of a particular industry where a set of financial statement items weren't necessary to report). Therefore the approach modelled one of these variables directly through the other so that in the case of a missing, the mean of the other variable was simply imputed.

The wkNN nearest neighbour imputation (Hechenbichler and Schliep, 2004; Beretta and Santaniello, 2016) uses values from similar samples to impute missing data. It combines these values for similar samples using a weighing calculated on the mean squared error between the existing features. The approach appeared promising as it is widely used in many studies. However, the large amount of missing values and the size of the FUNDQ data made the computation unfeasible in finite time on the available computational infrastructure.

The eventually employed method was a form of matrix factorisation called "Soft-Impute". It uses nuclear norm regularisation for matrix completion by iteratively replacing missing values by the figures computed from a soft-thresholded (Donoho,

1995) singular value decomposition. The algorithm was introduced by Mazumder et al. (2010) and the implementation from the fancyimpute package [4] for the Python programming language was used. Originally developed for the problem of recommender systems where unknown ratings are similar to missing values, the approach was devised for the Netflix dataset where the inputs have a dimensionality of 10^6 rows with 10^6 columns where only 0.001% of values are known. It is, therefore, a performant algorithm devised for a harder problem then the FUNDQ dataset. The missing value estimates generated by the approach had a similar mean and variance as the existing values indicating a meaningful imputation for the subsequent experiments.

The quality of the imputation was validated by comparing the mean and standard deviation of the imputed values in a particular financial statement item to the mean and standard deviation to the existing values. Since these metrics were of similar realistic magnitude, the predictions are taken to be valid substitutes for the remaining 23% of missing values.

To prevent look-ahead bias in the imputation, the imputation was conducted for every calendar quarter separately by only taking into account data samples that were published before the imputation period. For example, to impute Q1 2001 the only other values used for the imputation were rows in the FUNDQ dataset belonging to earlier quarters. After each of these quarterly imputation steps, the imputed rows belonging to the quarter are saved and finally combined to form the total dataset of 545.387 rows spanning 107 quarters between Q1 1991 and Q4 2017.

The imputation also benefited from the large quantity of data samples in the original data set that do not have a market reaction date. These samples correspond to financial statements where the announcement date is not set in the FUNDQ dataset and which are therefore not of use for the subsequent experiments. These rows, however, aid the imputation as it increases the size of the available training data based on which the matrix factorisation into a low-rank representation is conducted.

[4] https://github.com/iskandr/fancyimpute

Table 3.2: Missing value related data set size reduction

Reduction Step	Rows (= samples)	Columns (= variables)	Missing Values
1. Original data set	1,567,486	720	64%
2. Selection of the most populated columns	1,567,486	121	46%
3. Dropping all rows where more than 50% of values are missing	870,492	121	23%
4. Dropping all rows where SALEQ or ATQ is missing	868,125	121	23%
5. Imputation via SoftImpute	868,125	121	0%
6. Dropping all rows where SALEQ or ATQ is 0	810,407	121	0%
7. Excluding quarters where Y_{Q-0} has no announcement date and limiting the data to events between 1991 and 2017	545.387	121	0%

On step 6 of the table: These firms exist probably due to data entry errors and most likely not traded. The reduction is necessary as the financial statement variables are being normalised by SALEQ and ATQ.

The announcement date used for the subsequent computation of the abnormal market reaction is contained in the `rdq` variable of the FUNDQ file. The COMPUSTAT definition notes that it "represents the date in which quarterly earnings and earnings per share figures are first publicly reported in the various news media". It

is, therefore, an appropriate date based on which the market reaction is computed.

3.2 Abnormal Returns - CRSP

In the field of event studies, which concerns investigations of the impact particular events have on stock prices, a commonly used measure to quantify a stock market reaction to an event are Buy- and Hold Returns (BHARs). These returns represent the movement of a stock relative to the movement of the entire market in a particular time period. It is, therefore, a realistic measure of the abnormal return an investor would earn over the market by going long or short on a particular position.

Usually, the BHAR is calculated over a period of days. Since the financial literature suggests abnormal returns to persist to up to 60 days following the earnings announcement (Ball and P. Brown, 1968), this thesis investigates a period of 30 days following the announcement where it assumes the majority of the abnormal returns to occur. Since the `rdq` column in the FUNDQ data doesn't indicate whether the earnings were announced before market open or after market close, the period is also conservatively assumed to start one day before the announcement. This assumption, therefore, includes in the worst case also the last day of the anticipation of the event, at the benefit of certainly including the initial reaction in all cases.

Mathematically the BHAR represents the product of daily returns $R_{i,t}$ of the stock i on day t computed from daily prices P_t from the beginning ($\tau_1 = -1$) up until the end ($\tau_2 = 30$) of the event period, minus the return of a market portfolio $R_{m,t}$. It can be described as

$$R_{i,t} = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{3.1}$$

$$\text{BHAR}_{i,\tau_1,\tau_2} = \prod_{t=\tau_1}^{\tau_2}(1 + R_{i,t}) - \prod_{t=\tau_1}^{\tau_2}(1 + R_{m,t}) \tag{3.2}$$

The data to compute the BHARs were taken from the CRSP file provided by the Center for Research in Security Prices [5] which contains daily stock returns for

[5] `http://www.crsp.com`

listed US corporations stored in Open/High/Low/Volume format. When a stock is delisted due to liquidation, CRSP denotes the daily return as -1, which results in some calculated BHARs of less than -1. The market portfolio is computed as the mean return of all companies included in the CRSP file on a particular day.

The BHARs are not beta adjusted in the sense that they are calculated relative to market returns. That means that the alpha (i.e. the abnormal return) of the stock is not calculated relative to the beta adjusted return of the stock, but relative to the market movement.

The FUNDQ file and CRSP file are linked using the publicly available *CRSP-COMPUSTAT_link* file via the unique company identifiers `permno` and `gvkey`. The CRSP file and the link file were also obtained was obtained from the Wharton Research Data Services.

3.3 Data Window Construction

To allow the models to learn the past company-specific dynamics, multiple financial statements of past quarters are concatenated to form the input vector of independent variables. Including more quarters comes at the benefit of potentially learning more intricate past growth dynamics at the trade-off of inducing additional noise and a linearly growing curse of dimensionality. Conducting empiric tests suggests a period of 4 quarters, effectively including the last year of business activity, to be a reasonable compromise.

The data samples, therefore, are constructed using a sliding window of five quarters (i.e. four quarters of inputs and one quarter for the dependent variable). This window traverses the history of records of all companies in our data set. One sample consists of four successive quarterly reports denoted as, e.g. X_{Q-4} to X_{Q-1}, from which the published financial statements are used as independent variables, and the subsequent quarter (Y_{Q-0}) from which the market reaction is taken as the dependent variable.

Let X_{Q-n} denote a report of company f that is filed in calendar quarter $Q-n$ where $Q \in \mathbb{N}$ represents an index that pertains to a unique year and calendar quarter combination since year 0 AD in the Gregorian calendar starting at 1, and n denotes the offset of how many quarters before Q the report was filed. The set of reports spanning 4 quarters before Q (i.e. X_{Q-4}, X_{Q-3}, X_{Q-2}, X_{Q-1}) act as the independent variables to construct one sample for our models while Y_{Q-0} (or Y_Q) denotes the market reaction in quarter Q that is the dependent variable. These windows are company-specific and are constructed for all companies during the time of their operations. The collection of all company-samples for a particular quarter Q constitutes the training and test set of quarter Q so that for every quarter there exists such a test and training set.

The mapping the models therefore conduct takes the form of

$$f : (X_{Q-4}, X_{Q-3}, X_{Q-2}, X_{Q-1}) \mapsto Y_Q \tag{3.3}$$

This setup makes sure that the financial statements published in Q are unknown to the model, constraining the setup to base the prediction of the market reaction on the past financial statements.

Since fiscal years and the period of fiscal quarters vary across companies, the fiscal quarters are remapped to calendar quarters by associating a fiscal quarter with the calendar quarter where at least two months of the operating overlap. If for instance the fiscal quarter Q3 2009 of a particular company ranges from November 1st 2008 to January 31st 2009, it would remap the fiscal quarter to be a calendar quarter Q4 2008 filing, as two of the operating months (i.e. November and December) in this fiscal quarter fall into this calendar quarter which extends from October 1st 2009 to December 31st 2009. These calendar quarters represent the quarters Q mentioned in the previous paragraph along which the financial statements of different companies are grouped.

3.4 Training and Test Set

To prevent look-ahead bias, it is paramount to use only past data to train models, and unknown future data to evaluate them. The test set is, therefore constructed to be out-of-sample and out-of-time. This is achieved by applying an expanding window approach where, by traversing the history of quarters from past to present, upcoming quarters are initially used to evaluate the past model, before including them in the training data for models evaluated on future quarters.

Concerning calendar quarter Q4 2008, for instance, only data up to and including Q3 2008 is used to train the models, whereas the data pertaining to Q4 2008 is taken to evaluate the performance of this model. For the subsequent quarter Q1 2009, the data of quarter Q4 2008 is then included in the training data together with the data from earlier quarters.

This approach implies that the models are tested in realistic conditions and so that the performance can possibly improve over time as more training data becomes available. As the period of the study spans 107 quarters, it is the case that the models trained for Q4 2017 have substantially more training data than a model trained for Q1 1991.

While in general, it could be assumed that more training data would lead to better performance, the recency of the training data might also play a role. The further in the future the evaluated quarters are, the higher the rate of less recent training data is. Two experiments were performed to evaluate this effect with respect to the employed models that are introduced and discussed in the subsequent chapter.

One test considered limiting the training data to include only the previous 4, 12, 20, or 40 quarters for the random forest-based models. In comparison to the complete training set, imposing no limitation on the selection of the past quarters, these models did not perform favourably with regards to the performance metrics (e.g. the total compounded profitability) discussed in chapter 5.1.4. The second-best performing simulation, including 40 quarters, performed about 10% worse in terms

of the total compounded profitability than the complete training set, including all past quarters in the training data.

The second test attempted a similar limitation for the neural networks by reusing the trained weights of a past quarter for the subsequent quarter and biasing the weights towards the more recent quarter. This was achieved by using the weights of the past quarter as a starting point and training on the most recent quarterly data for five epochs. Ideally, the neural net would keep robust weights from the past and adapt some weights for the recent data of the last quarter. This test also failed to yield improved performance. The neural networks that used the randomly shuffled entire history past quarter samples as training data had superior performance to the neural networks where the data of the last quarter was used to induce a bias towards more recent dynamics. A transfer learning approach using frozen layers (Yosinski et al., 2014) was not attempted but appears promising for further studies.

4

Machine Learning Models

The machine learning methods employed in this thesis represent a selection of models that have gained increasing popularity in applied research settings over the last years. They include a Deep Neural Network (DNN), a Recurrent Neural Network (RNN), and a random regression forest (RF). The hypothesis stands that these models are able to take advantage of the non-linear relationships that this thesis assumes to exist. To test this assumption and to, therefore, justify the use of machine learning in this thesis, the dataset is also modelled using traditional linear models. The type of models used in this comparison are an ordinary least squares (OLS) regression and a least and a least absolute shrinkage and selection operator (Lasso). The following sections explain the specifics of these models and the particular configurations used in the experiments of this thesis.

4.1 Deep Neural Network

A Deep Neural Network (DNN) is a particular form of a feed-forward artificial neural network (ANN) that is different by having more than one hidden layer. It may be considered a 'deep' form of the standard multilayer perceptron (MLP) and represents a supervised form of machine learning.

Similar to other types models, its purpose is to approximate a function f^* that maps inputs x to a prediction $\hat{y}$ in the form of $\hat{y} = f^*(x_i)$. The functional specification is usually extended by a set of values known as the hyperparameters θ of a network $f^*(x;\theta)$ that determine how the f^* conducts this mapping. Subsequently, the learning process also known as *training* of a DNN consists of finding a set of optimal values for θ so that $f^*(x;\theta)$ results in the best function approximation (Goodfellow et al., 2016). A good function approximation usually means that the function output $\hat{y}$ is close to what is considered the true value y if it is observed.

In the context of neural networks, θ usually refers to the set of weights w_i and biases b of a model so that the elementary component of a neural net, known as a perceptron (Rosenblatt, 1958), can be explained as:

$$\hat{y} = f(b + \sum_{i=1}^{n} x_i w_i) \tag{4.1}$$

where $\hat{y}$ is the prediction/output, and f is a non-linear activation function (e.g. Sigmoid, tanh, ELU (see equation 4.2, ..), x_i are the n inputs of the perceptron, w_i are the weights by which the inputs are transformed, and b is the bias of a unit. The intuition for the perceptron has loosely been inspired by biological insight of how synapses fire in the brain. The measure of how well the approximation succeeds is referred to as the loss function of a network. This loss function guides the learning process to find the optimal set of hyperparameters through the backpropagation algorithm.

The hyperparameters of the DNN used in this thesis are found during the training process employing the Adam optimiser (Kingma and Ba, 2014). Also known as Adaptive Moment Estimation, Adam is an extension of RMSProp (Tieleman and

G. Hinton, 2012) that uses running averages of the first and second moments of the gradient to adapt the learning rate for each weight of the neural network. The learning rate is a hyperparameter of the training process that controls how much the weights are updated with respect to the loss gradient. A learning rate value of 0.00005 in combination with the Adam optimiser is used for training the DNN's in this thesis. The learning rate was found empirically by tweaking the learning rate in orders of magnitude much like described by L. N. Smith (2018).

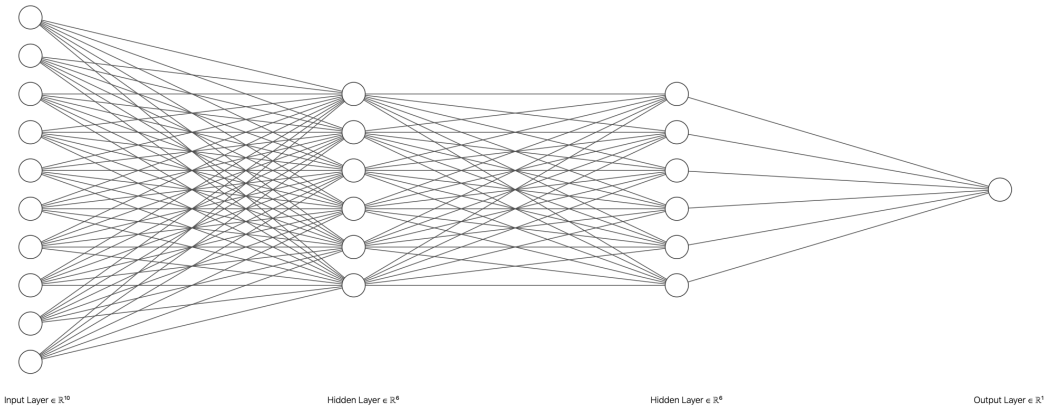


Figure 4.1: Visualisation of a neural network with two hidden layers. The circles denote a neuron where inputs are combined and run through a non-linear activation function with a neuron-specific bias. The lines represent the weighted connections between nodes that turn the outputs of one layer to the inputs of a subsequent layer.

As the loss function that is optimised is usually a non-convex function of the DNN hyperparameters θ, it is important to note that it can have many local minima. Therefore, θ, as found through the training process of the model, are not necessarily the best parameters. Choromanska et al. (2015) address this issue and assert that this is no major problem as the found local minima are usually of high quality and finding the real global minima of training data would be over-fitting.

A layer of a DNN consists of multiple neurons (see eq. 4.1) with non-linear activation functions that implement a function mapping of the outputs of the previous layer to the inputs of the subsequent layer. Generally, layers take the outputs of the previous layers as inputs while providing their own output for

a subsequent layer, as demonstrated in figure 4.1. A technique that is applied between the layers of the DNN used in this thesis is batch-normalisation (Ioffe and Szegedy, 2015). It addresses the problem of internal covariance shift in which inputs of hidden layers follow different distributions. This is of concern due to the heterogeneity of firms and samples that compile the training sample. For every batch during training, batch-normalisation behind the scene standardises the inputs to zero mean and unit variance.

Trying various numbers of layers and sizes of layers empirically, the deep neural net used in this thesis consists of three hidden layers that follow the input layer of dimension 484 (i.e. 4 concatenated quarters of financial statement variables) with respectively 100, 50, and 33 units, using an Exponential Linear Unit (ELU) (Clevert et al., 2015) as the activation function per layer. Deeper neural nets with up to 15 layers have been tried and more units per layer have been tried but didn't perform favourably. The configuration of the actual architecture was found empirically via grid searching iteratively over the hyper-parameter space.

The ELU denoted as the activation function f in the previous equations, has the mathematical properties of:

$$\mathrm{ELU}(x) = \max(0, x) + \min(0, \alpha * (\exp(x) - 1)) \tag{4.2}$$

where $\alpha = 1$, and x is the input.

In contrast to the popular rectified linear unit (ReLU) (Nair and G. E. Hinton, 2010) the ELU function can have negative values and therefore is able to learn on examples for which the activation is zero. Due to this problem of the normal ReLU function, multiple techniques alternative like the ELU have been proposed such as the leaky ReLU (Maas et al., 2013) or the PReLU (He et al., 2015). While conducting grid search on the learning rate and network parameters, the activation function was a parameter and the ELU non linearity was found to perform best among the activation functions supported in the PyTorch (Paszke et al., 2017) library version 1.3.1 . The data for the DNN is passed with a batch size of 256, and a total training period of 10 epochs per training set.

4.2 Recurrent Neural Network - Gated Recurrent Unit

The distinct characteristic of a Recurrent Neural Networks (RNNs) (Rumelhart et al., 1988) compared to a traditional feed-forward DNN is that it is designed to process sequential data $x_1, x_2, ..., x_t$ by sharing a tensor called hidden state between all sequence steps x_t. Such a standard RNN model can be described as:

$$h_t = \tanh(W_{hh} \cdot h_{t-1} + W_{xh} \cdot X_t + b_h) \tag{4.3}$$

$$y_t = W_{hy} \cdot h_t + b_y \tag{4.4}$$

where t denotes the sequence step, X_t the input at step t , h_t denotes the hidden state at step t, tanh the tanh non-linearity [1], and W denotes the weight matrices that are randomly initialised. Concretely the weight matrix W_{hh} is used to transform the past hidden state h_{t-1} to h_t, W_{xh} is used when transforming the input X_t at step t to h_t, and W_{hy} is used when transforming the computed hidden state h_t to the output y_t. b represents randomly initialised column matrices added as biases to the calculation of h_t (b_h) and y_t (b_y).

These type of neural network models have become state of the art in a range of natural language processing (NLP) tasks like speech recognition (Graves et al., 2013), language modelling (Mikolov, 2012), and machine translation (Kalchbrenner and Blunsom, 2013). In these applications, embeddings of words that constitute a sentence are passed as sequential arguments while the hidden state models relationships among them.

The decision to try an RNN architecture was made because the financial statement data used as inputs in the models follow a temporal sequence. With other machine learning models, the variables of all inputs are concatenated to form a vector of 484 elements (i.e. 121 variables * 4 sequence steps). In the case of

[1] $\tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

the RNN, this concatenation is not necessary as the model is designed to take a sequence of four reports (with respectively 121 variables) as an input so that one report represents to one sequence step. In comparison to many published applications of RNNs where the type of neural network is applied to settings with inputs of a relatively long sequence length, the input sequence in this application is just four steps (i.e. four quarters) long.

Multiple types of RNNs exist in the literature, and the particular one chosen for this thesis is a Gated Recurrent Unit (GRU) (Cho et al., 2014) which is functionally very similar to the popular Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). These architectures were developed to deal with the exploding and vanishing gradient problem in traditional RNNs by introducing memory cells and forget gates (Gers et al., 1999). The benefit of the GRU over the LSTM is that it has a smaller number of parameters and gates to be learned by combining the forgetting gate and the decision to update the unit state into a single update unit. Works such as (Chung et al., 2014) suggest that GRUs are on par with LSTMs in applied settings.

The mathematical properties of a GRU are best described as the following set of equations:

$$h_t = (1 - z_t)n_t + z_t h_{(t-1)} \tag{4.5}$$

$$r_t = \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \tag{4.6}$$

$$z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \tag{4.7}$$

$$n_t = \tanh(W_{in}x_t + b_{in} + r_t(W_{hn}h_{(t-1)} + b_{hn})) \tag{4.8}$$

with h_t being the hidden state at time t, x_t the input at time t, $h_{(t-1)}$ the hidden state of the layer at time $t-1$ or the initial hidden state at time 0, and z_t, r_t, n_t being the update, reset, and new gates. σ represents the sigmoid function $\sigma(x) = \frac{e^x}{e^x+1}$, W and b are learned weight matrices and biases belonging to various gates as denoted in the second letter of the subscript. The weights and biases where the first letter

of the subscript is h (e.g. W_{hr}, b_{hr}) transform the hidden state h in a gate, whereas a subscript starting with the letter i (e.g. W_{ir}, b_{ir}) transform the input x.

The GRU initially computes the hidden state h_t according to the current input vector x_t, and then uses this information to compute the update gate z_t and reset gate r_t. Then, it uses current reset gate r_t, input x_t and previous hidden state $h_{(t-1)}$ to calculate new memory content n_t. The previous hidden state $h_{(t-1)}$ and new memory content n_t are then combined to form the final hidden state h_t. At the first iteration h and n are initialised as zeros.

The employed GRU architecture follows the custom of stacking multiple GRU cells (i.e. 10 GRUs in this setting) on top of each other. Stacking means that one GRU takes the output of a GRU below as input until the GRU on top of the stack computes the final output. Figure 4.2 demonstrates how this stacking computes the hidden state and output. The blue cells in the figure represent w LSTM cells that are stacked for n sequence steps. It demonstrates how the input sequence x and the hidden state h is transformed through the network to form the final hidden state $h_n^{(w)}$ and cell state $c_n^{(w)}$ (the hidden state of the bottom GRU would be $h_n^{(0)}$).

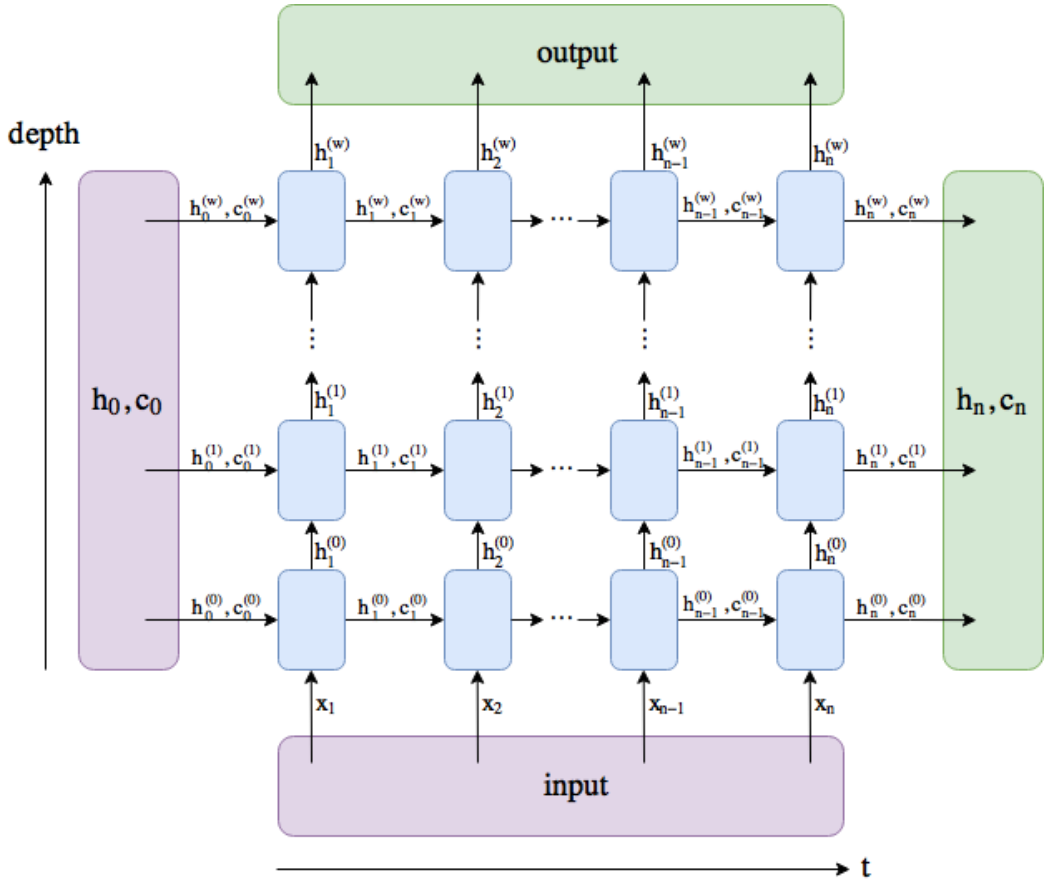


Figure 4.2: This figure has been adopted from StackOverflow User "nnnmmm" (2018). See section 4.2 for an explanation.

Each GRU in the setup has a hidden state dimension of 20 units, and the hidden state of the topmost GRU $h_n^{(w)}$ is linked to a fully connected linear unit for prediction. The RMSProp optimiser is used to train the model with a learning rate of 0.001 training for five epochs with a batch size of 128 elements.

4.3 Random Regression Forest

A random forest is a supervised ensemble learning approach that combines multiple classification and regression trees (CART) for a non-linear prediction. It has been introduced by Breiman (2001) and the thesis uses the scikit-learn (Pedregosa et al., 2011) implementation that is based on the paper. It was chosen in favour of the

popular XGBoost algorithm (Chen and Guestrin, 2016) so that the implementation of all models is done in sci-kit learn. Many practitioners argue that extreme gradient boosting trees and random forests have similar performance characteristics and there is no clear bias towards using either.

A CART tree is a hierarchical structure with every "node" representing a binary split of the data space into pieces based on the value of a feature. During the construction of a particular tree, all of its splits are made based on a random selection of features. Starting from the root node, the feature is taken as a criterion for the split which yields the "best" segmentation of the target variable. "Best" in this context is often referred to as solving an optimisation problem that maximises the homogeneity of the target variable in the segments resulting from the binary split.

The original paper by Breiman et al. (1984) or the recent paper by Loh (2011) explain the entire intricacies of the CART algorithm. In particular, they go into detail how the tree is constructed by discussing the type of optimisation problem that is solved for each split in case of a regression tree and classification tree.

The output for a random forest depends on the type of its dependent variable. For a classification tree, each tree makes a vote for the most likely class for the input vector, while the votes of all trees are combined to form a majority vote on the class that is predicted by the forest. For the regression tree, every tree predicts the mean of the terminal leaf that the input is classified as, whereas the whole forest combines the predictions of all trees as a simple mean.

The two hyperparameters for the random forest are the number of regression trees that it consists of and their maximum depth. The amount of regression trees makes the forest more robust as the dimensionality of the inputs increases as every regression tree gets assigned a random set of features. This method is referred to as feature bagging. A higher amount of trees, therefore, means that variables get reused often in permutations with other variables. The computational complexity of training the model scales linearly with the number of trees it consists of and was decided to be 200 trees per forest. While the computational complexity of the random forest training increases with the number of trees per forest, the benefit of more trees

is that it allows features to be picked multiple times during bagging and growing the tree. The original paper by Breiman (2001) explicitly notes that over-fitting should not be a problem. The number of trees was also optimal for the number of available training CPUs for the experiments as it scaled linearly with CPU threads.

The maximum depth rather induces a tradeoff between overfitting and modelling capacity instead of more trees. A deeper tree can model a more complex and intricate combination of inputs with hypothetically unlimited depth of a tree perfectly learning the entire training data to the point where one leaf just contains one sample. Since this over-fitting might be problematic for the prediction of unseen samples, and as the memory requirements of the trees grow as well, in this thesis, the depth was limited to ten splits.

One benefit of using a random forest is that it provides a potentially interesting variable importance measure. This variable importance measure can help to identify which factors contain the signal based on which the predictions are made. Neural net-based machine learning models lack such a measure making the output less interpretable. The mathematical construction of this variable importance measure is also explained in the original paper by Breiman et al. (1984).

4.4 Linear Models

Linear regressions represent a classic example of the standard forecasting methodologies in Econometrics and Finance. They are benchmarked as a reference point in this thesis to determine the benefits of utilising sophisticated non-linear machine learning algorithms.

4.4.1 Ordinary Least Squares (OLS)

The most trivial linear regression model is estimated via ordinary least squares (OLS). It assumes that the target variable y can be approximated through a linear combination of the independent variables x by a set of coefficients β.

For n observations $y_i, x_{i i=1}^n$ of the scalar independent variable y_i and the column vector x_i of d dependent variables $x_i j$ for $j = 1, .., d$, the OLS regression models y_i as:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{4.9}$$

where β is a fitted parameter (i.e. coefficient) that is 'learned' and ε_i is a gaussian noise with zero mean.

It can be written in matrix notation as the operation:

$$y = X\beta + \varepsilon \tag{4.10}$$

where y and ε are vectors of length n consisting of the various dependent variables and errors, and X is matrix of shape $n \times d$ that contains the d explanatory variables for n samples.

The best selection of estimates $\hat{\beta}$ for β is then found by solving the minimisation problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| y - X\beta \right\|_2^2 \tag{4.11}$$

4.4.2 Least absolute shrinkage and selection operator (Lasso)

The least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) is a linear regression and was introduced to improve the model fitting by selecting only a subset of coefficients (instead of all of them) in the final prediction model.

For n observations $y_i, x_{i i=1}^n$ of the scalar independent variable y_i and the column vector x_i of d dependent variables $x_i j$ for $j = 1, .., d$ lasso aims to solve the quadratic minimisation problem

$$\min_{\beta_0, \beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t \tag{4.12}$$

with the tuning parameter $t \geq 0$.

Rewriting this optimisation using the matrix notation introduced in the previous chapter in the Lagrangian form gives

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \alpha \|\beta\|_1 \right\} \tag{4.13}$$

where a relationship existing between the respective tuning parameters t and α. This form of optimisiation is also referred to as basis pursuit in the mathematical literature.

In a simple interpretation, the parameter α and t control the number of selected variables in the model. If $\alpha = 0$ the Lasso regression finds the same coefficients as an OLS regression, and as α becomes larger, fewer independent variables are selected as more coefficients become zero.

5

Experiments

This chapter discusses the experiments and their results that are conducted in the context of this thesis. It is structured in two major blocks and a supplementary experiment.

The first section concerns experiments where the market reaction to earnings announcements is modelled in a regression setup so that the models predict a scalar, while the second section conducts experiments of a classification setting where the dependent variable is discrete. This distinction in model specification is made as both types of model output lend themselves to distinct experiments to generate signals based on financial statement information as discussed in the upcoming sections.

Following the main sections of this chapter the third section takes the results of the best performing models and conducts experiments on the role of the market capitalisation in these results. This analysis investigates whether the discovered abnormal returns can solely be attributed to firms of a very low market capitalisation

(i.e. penny stocks or micro caps) or whether even firms of larger market capitalisation contribute to the overall abnormal profits.

The concrete metrics investigated are mentioned in the sections based on whether they belong to the regression or classification. Since the returns are of a quarterly nature and not simulated on a daily basis, the Sharpe ratio (Sharpe, 1994) of the portfolio has not been calculated as a performance metric. For this reason also no regression on the Fama-French factors has been conducted. It is for future studies to simulate different portfolio strategies which would allow such a comparison.

5.1 Regression

The first type of model specification is of a regression setup that uses the continuous scalar value of the BHARs as the dependent variable. In the context of earnings forecasting, this would be equivalent to directly forecasting the earnings surprise as the difference between earnings expectation and realised earnings. The benefit of this type of setup is that the models are able to learn from the nuances in the market reaction by being trained to distinguish the whole spectrum of major and minor reactions in both directions.

Contrary to classification models, there is usually no certainty provided with the prediction as the output of the model is not a probability of a stock movement. This lack of confidence limits the ways the results can be utilised and motivates the introduction of ε thresholds that are introduced later in the chapter. While intuition suggests that predictions of a large magnitude should be more certain than predictions of small magnitudes, dedicated tests aim to substantiate this hypothesis.

5.1.1 Loss function

The models are trained for the regression setup using the mean squared error (MSE) as the loss function. For the prediction prediction $\hat{y}_i$ and ground truth label y_i of sample index i in the set of training indices $i \in T$ with $|T| = n$ the MSE is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.1}$$

5.1.2 Goodness Of Fit

The predictive ability of the models is analysed by looking the errors e that are calculated for all predictions $\hat{y}_i$ and ground truth labels y_i of sample i in the test set $i \in I$ with $|I|$ = the number of samples, as

$$e_i = y_i - \hat{y}_i \tag{5.2}$$

The set I includes all the indices of test samples from all quarters between the second quarter of 1991 and the fourth quarter of 2017. The construction of the test indices has been introduced in section 3.4.

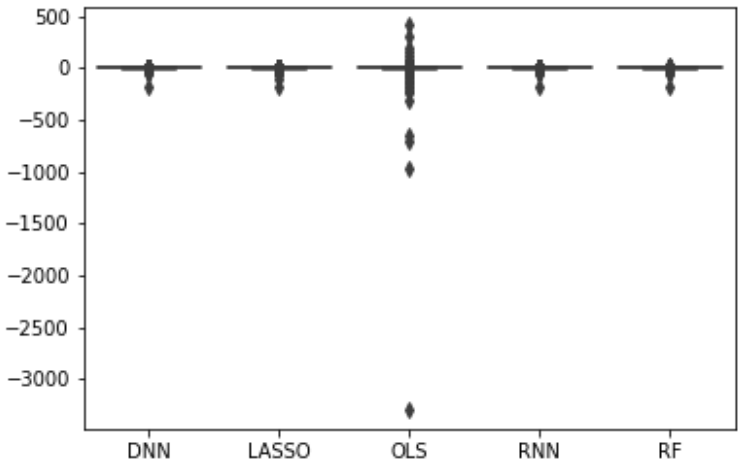


Figure 5.1: Boxplot of the errors (Y-Axis) across model types (X-Axis).

The boxplots of the errors e across the model types are depicted in figure 5.1. They indicate some extreme values and outliers among the OLS model that skew the figure and might potentially make performance statistics biased. Therefore the subsequent results in this section are computed based on the 99.99^{th} percentile of the absolute value of errors per model type. For a quantitative description

of the distribution the errors take across model types (without the outliers in figure 5.1) please consult table 5.1.

The percentile is constructed using the ordered set E that is composed of the absolute values of the errors $E = \{|e_0|, .., |e_n|\}$ of the n predictions contained in the complete test set I per model type. The set E is sorted (from smallest to largest) so that for every monotonically rising integer index j of the elements $k \in E$ it is true that $k_j \leq k_{j+1}$. Then the ordinal rank r for percentile $p \in [0, 100]$ is computed as:

$$r = \left\lceil \frac{p}{100} \times |E| \right\rceil \tag{5.3}$$

So that taking the value k_r from E yields the number that is used to limit the selection of predictions and ground truth labels used in the computation of the subsequent statistics by redefining the set of test sample indices I of every model to samples where

$$|y_i - \hat{y}_i| \leq k_r \tag{5.4}$$

This operation excludes 0.0001% of samples at the choice of $p = 99.99$ to exclude a few most extreme outlier errors which are not representative of the general predictive ability of the models. The figure k_r is selected based on p, and I is limited per model.

The first way to compare the predictive ability of the models is to compare the value of the loss function and similar metrics on the test set predictions. For this purpose the raw value of the mean squared error (MSE) (see equation 5.1) , the root mean squared error (RMSE), the mean absolute error (MAE), the median absolute error (MedAE), and the R^2 metric is computed.

Using the notation introduced earlier in this chapter, and $\bar{y}$ representing the mean of the dependent variable, the metrics are defined as:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{|I|} \sum_{i \in I}^{|} I|(y_i - \hat{y}_i)^2} \tag{5.5}$$

$$\text{MAE} = \frac{1}{|I|} \sum_{i \in I}^{n} |y_i - \hat{y}_i| \tag{5.6}$$

$$\text{MedAE} = \text{median}(|y_i - \hat{y}_i|_{i \in I}) \tag{5.7}$$

$$\text{R}^2 = 1 - \frac{\sum_{i \in I} (y_i - \hat{y}_i)^2}{\sum_{i \in I} (y_i - \bar{y})^2} \tag{5.8}$$

These metrics differ slightly in their interpretation. The MSE is generally used during the training of a model because the squaring operation disproportionally penalises large errors in comparison to small ones. This way, the models are tuned to be more sensitive to learn from large errors than from minor errors.

However, because of its squaring operation, the MSE is not clearly interpretable with regards to the original unit of measure. Therefore the RMSE is also computed, which undoes the squaring operation by taking the square root.

The MAE, on the other hand, directly computes the mean absolute error of the predictions. This makes the MAE the simplest metric to interpret as it denotes how much the predictions deviate from the ground truth on average. As the mean could be skewed by a few extremely bad predictions, the MedAE is also included in the array of metrics as it that represents the median of the mean absolute errors. It is more robust and should be interpreted in conjunction with the MAE.

From the perspective of a practitioner, the MSE, MAE, MedAE, and RMSE, are *better* if the error is smaller and *worse* if larger. Perfect predictions achieve an error of zero.

The R^2 (Wright, 1921), also known as the *coefficient of determination*, denotes the proportion of the variance in the predicted variable that can be explained through the model. It has been introduced in the context of ordinary least squares regression to measure and compare the goodness of fit of a model. A value of one suggests the prediction perfectly fits the data while the interpretation of the values smaller than one depends on the specific application. The formula also allows for a negative R^2 which is interpreted to be worse as it gets more negative. It indicates

that the mean of the dependent value provides a better fit than the predictions. The formula used is provided by scikit-learn[1] and allows for negative values of the R^2.

Model	MSE	RMSE	R^2	MAE	MedAE
RF	0.055	0.234	0.007	0.153	0.099
DNN	0.058	0.241	-0.051	0.16	0.106
RNN	0.07	0.265	-0.273	0.174	0.111
Lasso	0.055	0.235	-0.002	0.154	0.100
OLS	0.062	0.250	-0.063	0.158	0.101

Table 5.1: Mean squared error, root mean squared error, R^2, mean absolute error, median absolute error of the employed model types. The metrics have been computed for the collection of the quarterly test sets over the entire study period between 1991 Q2 and 2017 Q4.

The results of the metrics as calculated over the predictions of the entire test samples across all periods are presented in table 5.1. They indicate the Random Forest to be the best performing model across all metrics, closely followed by the Lasso regression and the deep neural net. The ordinary least squares regression and the recurrent neural net are the poorest performing model. Based on these results, the hypothesis that machine learning models outperform traditional linear models cannot be confirmed.

5.1.3 PC Metric

While the mean-squared error is a meaningful metric to compare the predictive performance across models, it lacks in economic interpretability. In this context,

[1] `https://github.com/scikit-learn/scikit-learn/blob/7e85a6d1f/sklearn/metrics/_regression.py#L488`

the question of whether models predict the correct direction of the market reaction is more useful in the context of the research question. For this purpose, this section introduces a metric called PC which quantifies the rate of correct sign predictions. The models are still trained using the mean squared error as the loss function, but the discussed performance evaluation metric is different.

Epsilon Thresholds

For weak BHARs close to zero, the meaningfulness of the sign is ambiguous. A reaction of -1% might be driven by a very similar information content as a reaction of $+1\%$ so that the sign might be the result of noise in the price movement. Therefore a set of epsilon (ε) thresholds are created which are used to rectify the value of the prediction and ground truth to zero. Through this operation, which is detailed in the subsequent paragraphs, the continuous predictions and ground truth values can be separated into significant positive, non-indicative neutral, and significant negative ones, based on which the performance of the models is evaluated.

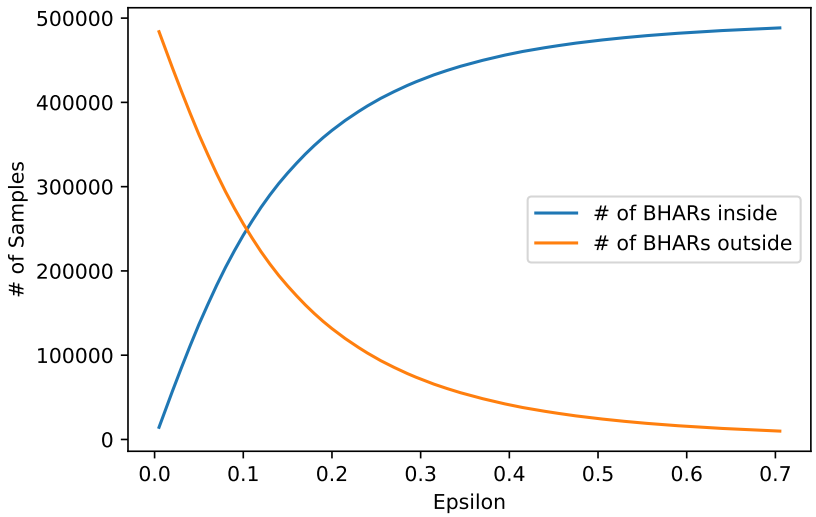


Figure 5.2: Ratio of ground truth values of BHAR inside and outside particular ε thresholds

If the absolute value of the reaction r is smaller than a given ε so that $|r| < \varepsilon$ then, in the context of evaluating the prediction it assigned the value zero $r = 0$.

$$r = \begin{cases} 0 & \text{if } |r| < \varepsilon \\ r & \text{otherwise} \end{cases} \tag{5.9}$$

This adjustment separates the BHARs into significant negative $r < -\varepsilon$, ambiguous $r > -\varepsilon \wedge r < \varepsilon$, and significant positive reactions $r > \varepsilon$.

Seven epsilon thresholds are created $\varepsilon \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and applied to both the prediction of models $\hat{y}$ and the actual ground truth y. This is done after the prediction so that if $\hat{y}$ and/or the ground truth y, falls into the respective epsilon threshold (e.g. $|\hat{y}| < \varepsilon$, $|y| < \varepsilon$) they are set it to 0. This way the choice of ε, does not impact training as the epsilon threshold check occurs after $\hat{y}$ is predicted.

Rectifying the ground-truth value to zero based on ε before the training of the models as a preprocessing step might be preferable as models would only be explicitly exposed to significant reactions (outside of ε) and non-reactions (inside of ε) as training inputs. Models would not learn to distinguish ambiguous minor *non*-reactions as they are all set to zero. This would however drastically increase the amount of necessary computation seven-fold as instead of merely fitting one model (with the non-epsilon adjusted values), it would be necessary to train one model per level of ε. For lack of computational resources, this pursuit was not undertaken. Table 5.2 and figure 5.2 illustrate how many BHARs fall inside and outside the particular ε thresholds.

ε	# outside ε	# inside ε	% inside	% outside
0	545,387	0	0.00	1.00
0.05	391,567	153,820	0.28	0.72
0.1	272,912	272,475	0.50	0.50
0.2	137,070	408,317	0.75	0.25
0.3	73,771	471,616	0.86	0.14
0.4	41,904	503,483	0.92	0.08
0.5	25,270	520,117	0.95	0.05

Table 5.2: Number and proportion of ground truth BHAR values inside and outside of the selected epsilon thresholds

PC Metric

With regards to the intended use of the PC metric to interpret the rate of correct prediction of the sign of the market reaction, the introduction of ε thresholds complicates matters. As table 5.2 demonstrates, a growing ε leads to an increasing amount of samples to fall into ε threshold. This would skew a metric that measures the correctly predicted BHARs greater than ε, less than $-\varepsilon$, or within the *insignificance range* $(-\varepsilon, \varepsilon)$, not comparable across the levels of ε.

Therefore, in a more detailed consideration, the *PC* metric is constructed for two purposes:

- It quantifies the rate of *correct* (i.e. having the right sign) predictions outside the ε threshold (e.g. predicting a BHAR of 20% at a $\varepsilon = 0.1$ where the true BHAR has a positive sign would be a *correct* prediction) , and

- the metric is penalised for the cases where a true BHAR is outside of the ε threshold and predicted to be within.

With $i \in I_t$ being the index of the ground truth value y_i and predicted value $\hat{y}_i$ in the set of test sample indices I of a quarter t, $\varepsilon \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$, $|I|$ denoting the cardinality of the set, the helper functions $sign(x)$, $g(x, \varepsilon)$, $h(x, z, \varepsilon)$ are defined as:

$$sign(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{5.10}$$

$$g(x, \varepsilon) = \begin{cases} 1 & \text{if } x < \varepsilon \wedge x > -\varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

$$h(x, z, \varepsilon) = g(x, \varepsilon) * g(z, \varepsilon) \tag{5.12}$$

so that the PC metric can be computed as

$$PC = \frac{\sum\limits_{i \in I_t} (|h(y_i, \hat{y}_i, \varepsilon) - 1| \, \frac{|sign(y_i) + sign(\hat{y}_i)|}{2})}{|I_t| - \sum\limits_{i \in I_t} h(y_i, \hat{y}_i, \varepsilon)} \tag{5.13}$$

Based on this definition, a perfect PC metric of 1 indicates a case where the models predicted significant (i.e. outside of ε) BHARs have the correct sign, and there are no true BHARs that were predicted to be non-significant (i.e. inside of ε). As already stated, it is important to note that the PC metric is merely used as an evaluation metric for the test loss, and not in any way during the training.

Results

Table 5.3 lists the mean and standard deviation of the PC measures computed in the experiments. The random forest has the highest percentage of correct predictions averaging 55% of the cases correctly with an epsilon $\varepsilon = 0$, an average of 59% correctly predicted directions at $\varepsilon = 0.3$, and an average of 59% correctly

predicted directions at $\varepsilon = 0.5$. These results run contrary to the hypothesis that predictions with higher magnitudes are more correct on average than predictions of smaller magnitudes as a linear increase of the PC metric cannot be observed with growing ε. For this hypothesis, the mean PC would have to increase up until the $\varepsilon = 0.5$ case monotonically. However, finding the higher mean PC value at $\varepsilon \in \{0.2, 0.3\}$ than at $\varepsilon = \{0, 0.5\}$ suggests that reactions of medium severity are the ones where the direction can be most confidently predicted. The results of other models also follow a similar pattern.

Epsilon	**0**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
RF	0.55	0.56	0.57	0.59	0.59	0.58	0.56
	(0.06)	(0.05)	(0.06)	(0.09)	(0.11)	(0.13)	(0.14)
	1.0	*0.74*	*0.5*	*0.25*	*0.13*	*0.08*	*0.05*
DNN	0.53	0.54	0.55	0.56	0.56	0.56	0.54
	(0.03)	(0.04)	(0.05)	(0.07)	(0.09)	(0.11)	(0.12)
	1.0	*0.77*	*0.52*	*0.26*	*0.14*	*0.08*	*0.05*
RNN	0.54	0.55	0.56	0.57	0.57	0.56	0.55
	(0.05)	(0.06)	(0.07)	(0.09)	(0.11)	(0.12)	(0.13)
	1.0	*0.75*	*0.51*	*0.25*	*0.13*	*0.08*	*0.05*
Lasso	0.55	0.56	0.57	0.57	0.56	0.54	0.51
	(0.07)	(0.08)	(0.1)	(0.14)	(0.16)	(0.17)	(0.19)
	1.0	*0.72*	*0.49*	*0.25*	*0.13*	*0.07*	*0.04*
OLS	0.54	0.55	0.56	0.56	0.55	0.54	0.52
	(0.05)	(0.06)	(0.08)	(0.11)	(0.13)	(0.15)	(0.16)
	1.0	*0.75*	*0.51*	*0.25*	*0.14*	*0.08*	*0.05*

Table 5.3: Mean, (standard deviation) of the PC measure across the quarters of the study period per model and epsilon. The third row in *cursive* represents the mean *proportion* of all prediction outside the given ε threshold of the entire test sample.

Another interesting finding is how the variance of the quarterly PC measure monotonically increases with ε, from a standard deviation (std) of 0.06 at $\varepsilon = 0$ up to a std of 0.14 at $\varepsilon = 0.5$. This evidence supports the hypothesis that predictions at larger ε levels are harder to predict. The mean proportion of samples based

on which the PC metric has been computed might be another strong contributing factor for this phenomenon. Quarters with just a few predictions outside of a high ε threshold would inadvertently yield a higher std than lower levels of ε where the majority of the samples are used to compute a relatively *stable* level of PC. For most models, only about 5% of predictions fall into the $\varepsilon = 0.5$ case.

Apart from the best performing RF model, the comparison between the neural net based models (i.e. DNN, RNN) and the linear models (i.e. Lasso, OLS) exhibits interesting dynamics. For $\varepsilon \in \{0, 0.05, 0.1\}$ the performance of the linear models is on par and in the case of Lasso even superior to any neural network. While the RNN is about 1% point better at the mean PC than the DNN, Lasso is about an additional 1% better than the RNN. At higher levels of $\varepsilon \in \{0.2, 0.3, 0.4, 0.5\}$ the neural net based models outperform the linear models clearly. At $\varepsilon = 0.5$ for instance, Lasso and OLS are barely more accurate than a random guess with a mean PC of 51% and 52%. At these levels, the DNN and RNN are performing substantially better with a mean PC of 54 and 55%. When trained for the MSE, and assessed by PC, these metrics suggest that the neural net based models are better at predicting larger reactions than linear models.

The results in table 5.3 concerning the standard deviation of the quarterly PC metric also demonstrate a significant difference among linear models and neural nets. Across all thresholds of ε, Lasso and OLS have a higher std than neural net based models. The difference is most significant at higher values of ε but is also exhibited at the lower thresholds where the linear models have a better mean PC.

Discussion

These initial results confirm the main hypothesis of the thesis as in general $PC > 50\%$ among the tested machine learning models. This metric indicates that for over 50% of the stock market reactions outside a particular ε threshold, the direction was correctly predicted. Assuming that this direction of the market reaction is driven primarily by the information content of the newly released financial statement, this prediction based on just the past financial statement data is

suggestive that the models are able to model the unreleased future results to make this decision implicitly.

The evidence presented in table 5.3 also provides support for the hypothesis that machine learning (ML) models perform better than plain linear models at the task of predicting the direction of the BHAR. The difference in the standard deviation of quarterly PC indicates that ML models are more consistent right and robust in their prediction performance, while the better PC metric across higher thresholds of ε suggests that ML models are also relatively better at identifying strong market reactions and their direction.

5.1.4 Compounded Profitability

This section evaluates whether the results exhibited in the PC metric translate into trading signals that yield profits if traded on. To this end, a simple trading strategy is devised that takes long and short positions over the event horizon of the earnings announcement. This performance metric also takes into account the seven ε thresholds introduced in the last section. The interpretation of ε, in this case, is that it limits the trading to only *significant* reactions so that no trading position is taken made for *ambiguous* reactions that fall into the threshold.

Trading Strategy

The trading strategy assumes that for every quarter Q, across the test indices I_Q there exists a subset of trading indices $T_{Q,\varepsilon} \subseteq I_Q$ which contain the indices outside the ε threshold that is being used for the trading strategy. The set is constructed based on the value of the prediction $\hat{y}$ as:

$$T_{Q,\varepsilon} = \{i \in I_Q | \hat{y}_i > \varepsilon \vee \hat{y}_i < -\varepsilon\} \tag{5.14}$$

Every position based on index t of $t \in T_{Q,\varepsilon}$ is a position that is being traded on and is weighed equally using a nominal quarterly portfolio size s.

$$w_t = \frac{s}{|T_{Q,\varepsilon}|} \tag{5.15}$$

with w_t denoting the weight of the position taken for t.

Across these positions based on the sign of $\hat{y}_t$ either a long position (if $\hat{y}_t > 0$) or a short position (if $\hat{y}_t < 0$) is taken. Depending on the sign of y_t these positions yield a positive or negative return of p_t depending on whether the correct sign was predicted.

$$p_t = \begin{cases} 1 + y_t & \text{if } \hat{y_h} > 0 \\ 1 - y_t & \text{if } \hat{y_h} < 0 \end{cases} \tag{5.16}$$

This p_t metric is then multiplied by the weight w_t and summed up to determine the overall quarterly return $P_{Q,\varepsilon}$

$$P_{Q,\varepsilon} = \sum_{t \in T_{Q,\varepsilon}} w_t p_t \tag{5.17}$$

Because of the use of BHARs as the dependent variable y, the quarterly profit $P_{Q,\varepsilon}$ can be directly interpreted as the abnormal profit of a given quarter and ε threshold. It has to be interpreted in relation to the initial nominal portfolio of size s to determine whether the quarter was profitable or not.

For the analysis of the compounded profitability in a later section, the profit is compounded by starting the analysis at the nominal portfolio size of 1 in Q1 1991 and reinvesting the grown (or declined) portfolio in subsequent quarters. Therefore in the case of compounded profit:

$$s_{Q,\varepsilon} = P_{Q-1,\varepsilon} \tag{5.18}$$

If for example in Q1 1991 the nominal portfolio of size 1 grows by 10% to 1.1, then the nominal portfolio for Q2 1991 is assumed to be of size 1.1 so that the profit after the quarter can be interpreted in relation to the initial portfolio.

If the number of quarterly trading positions $|T_{Q,\varepsilon}|$ is smaller than two, then the trading strategy does not trade in that quarter so that $P_{Q,\varepsilon} = 1$. The purpose

of this restriction is to avoid the high volatility that a small diversification on two positions could produce.

A further assumption is that the extreme end of BHAR reactions that are greater than +100% or smaller than −100% are rectified to is rectified to these limits so that:

$$y_t = \begin{cases} 1 & \text{if } y_t > 1 \\ -1 & \text{if } y_t < -1 \\ y_t & \text{otherwise} \end{cases} \tag{5.19}$$

These extreme BHAR values arise due to the market movement over the long holding period of 30 days. The rectification can be justified as the positions could be closed immediately after the initial market reaction before the BHAR reaches such large magnitudes. The purpose of this limitation is not to let profits/losses of a quarter be biased by these rare outlier cases.

Results

Quarterly Profitability

The results presented in table 5.4 are analogous to interpret as table 5.3 of the previous section. They outline the mean, standard deviation of the quarterly profitability per model and ε and the *total number of trades* that comprise the strategy. The values that comprise the statistic are calculated by subtracting 1 from the quarterly profit of $P_{Q,\varepsilon}$ of the nominal portfolio. Thereby they represent the % profit (loss) of a quarter.

Epsilon	**0**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
RF	2.1%	6.7%	10.2%	10.4%	11.3%	11.9%	13.2%
	(3%)	(8%)	(9%)	(12%)	(14%)	(18%)	(22%)
	540,995	*90,447*	*21,308*	*3,345*	*1,371*	*766*	*515*
DNN	1.6%	3.0%	4.8%	7.0%	9.9%	8.9%	6.5%
	(2%)	(3%)	(5%)	(9%)	(12%)	(16%)	(17%)
	540,995	*211,504*	*79,233*	*14,981*	*3,853*	*1,307*	*584*
RNN	1.2%	2.5%	3.4%	4.2%	4.9%	5.2%	5.7%
	(2%)	(3%)	(5%)	(6%)	(6%)	(6%)	(7%)
	540,995	*210,689*	*119,791*	*54,055*	*27,148*	*14,291*	*7,733*
Lasso	1.7%	7.1%	9.6%	9.4%	7.6%	5.7%	2.8%
	(4%)	(9%)	(11%)	(13%)	(17%)	(17%)	(18%)
	540,995	*41,334*	*7,900*	*1,685*	*761*	*425*	*277*
OLS	1.4%	5.2%	7.3%	7.7%	6.5%	5.3%	4.3%
	(3%)	(6%)	(8%)	(8%)	(9%)	(10%)	(11%)
	540,995	*87,536*	*24,877*	*7,167*	*3,837*	*2,529*	*1,894*

Table 5.4: Mean and (std) of the quarterly profitability, and *total number of trades* per model and epsilon.

Consistent with results of the PC metric, the best performing model overall is the random forest with the highest mean and median profitability across the various levels of ε. In the case of $\varepsilon = 0$, equivalent to a model where every position is traded, the mean/median quarterly profitability is 2.1%. The highest mean of 13,2% profitability is found at $\varepsilon = 0.5$. The difference in the quarterly profitability across the ε levels suggests that the model is able to distinguish market reactions of higher magnitude from smaller market reactions.

The second place in the ranking is shared by the DNN and Lasso model, which achieve a significantly higher mean profit than the RNN and OLS. The Lasso model performs favourably at lower thresholds of $\varepsilon \in \{0, 0.05, 0.1, 0.2\}$ while the DNN outperforms at higher thresholds. This difference suggests that the DNN is better at predicting profitable large reactions than the linear Lasso model. This dynamic between neural networks and linear methods can also be observed in the comparison of the RNN and the OLS model, at lower ε thresholds, the OLS model wins, while at higher thresholds the RNN wins. While the RNN, in general, has a lower quarterly mean return, the high number of trades it enters yields a low standard deviation in returns and a linearly with ε increasing quarterly profit.

The poor performance of the linear models at the higher ε thresholds is consistent with the poor performance in the PC metric and confirms the hypothesis that linear models are inferior at predicting market reactions of a large magnitude. The low amount of predictions (at higher profitability) in comparison to the neural net models at $\varepsilon \in \{0.05, 0.1, 0.2\}$ indicate that the linear models are better at distinguishing the non-reaction cases close to zero.

Compounded profitability

The next analysis concerns how quarterly profits translate into compounded profits over the study period. Table 5.5 presents the final portfolio value $P_{Q,\varepsilon}$ of the last period of the study (i.e. Q4 2017) if a nominal portfolio of size 1 was invested at the beginning of the study period (i.e. Q1 1991). Table 5.6 translates this total return into the compound annual growth rate (CAGR) of the portfolio as

$$\text{CAGR} = \text{FV}^{\frac{1}{26.75}} - 1 \tag{5.20}$$

with FV denoting the final compounded return of the nominal portfolio, and 26.75 being the total number 107 of quarters translated to years.

Epsilon	**0**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
RF	8.77	797.2	23,396	22,203	40,782	42,083	71,527
DNN	5.34	23.11	127.47	957.16	11,747	3087	239.84
RNN	3.53	13.15	32.79	69.21	136.09	189.22	287.32
Lasso	5.62	994.86	10,617	6,784	403.05	86.88	2.4
OLS	4.31	184.09	1,380	2,156	587.69	155.89	48.55

Table 5.5: Final compounded value on a nominal portfolio (of size 1) from 1991 and 2017 per epsilon and model

In this analysis, the choice of ε demonstrates its benefits in the difference it makes for the total portfolio value. The random forest generates the highest abnormal return of $71,527$ in the case of $\varepsilon = 0.5$. The vanilla DNN again outperforms the RNN among all ε thresholds except for the highest $\varepsilon = 0.5$ case. The Lasso model again performs stronger than the plain OLS model with the highest profit of all models in the $\varepsilon = 0.05$ scenario.

The decline in total profit of the linear models at higher ε-thresholds is consistent with the reported results in table 5.4 suggesting a poorer accuracy for the reactions of predicted higher magnitude. The OLS and Lasso model yields the worst result in comparison to the other models as well. It is particularly interesting how the high standard deviation of the Lasso model at $\varepsilon = 0.3$, leads to a smaller total profit than the OLS model, which has a lower mean quarterly profitability. While all machine learning models manage to increase their total profitability from the 0.2

to the 0.3 ε threshold at least twofold, the profitability of both traditional linear models in comparison sinks to less than half of it.

Epsilon	**0**	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
RF	8%	28%	46%	44%	49%	49%	52%
DNN	6%	12%	19%	29%	41%	35%	23%
RNN	5%	10%	13%	16%	19%	21%	24%
Lasso	7%	29%	40%	38%	25%	17%	3%
OLS	6%	21%	31%	33%	27%	20%	15%

Table 5.6: Compounded annual growth rate between 1991 and 2017 per epsilon and model.

These results also undermine the findings of the model performance statistics presented in table 5.2. While these findings suggested no major differences in the predictive ability between the RF and Lasso, and equivalent performance of OLS to the neural nets, the results in this section contrast these findings. This divergence of theoretic model performance metrics (e.g. MSE, R^2, ..) and directly applied results of the compounded profitability indicates that it makes an important difference whether machine learning methods are evaluated based on the task at hand, or in general metrics that have been used traditionally.

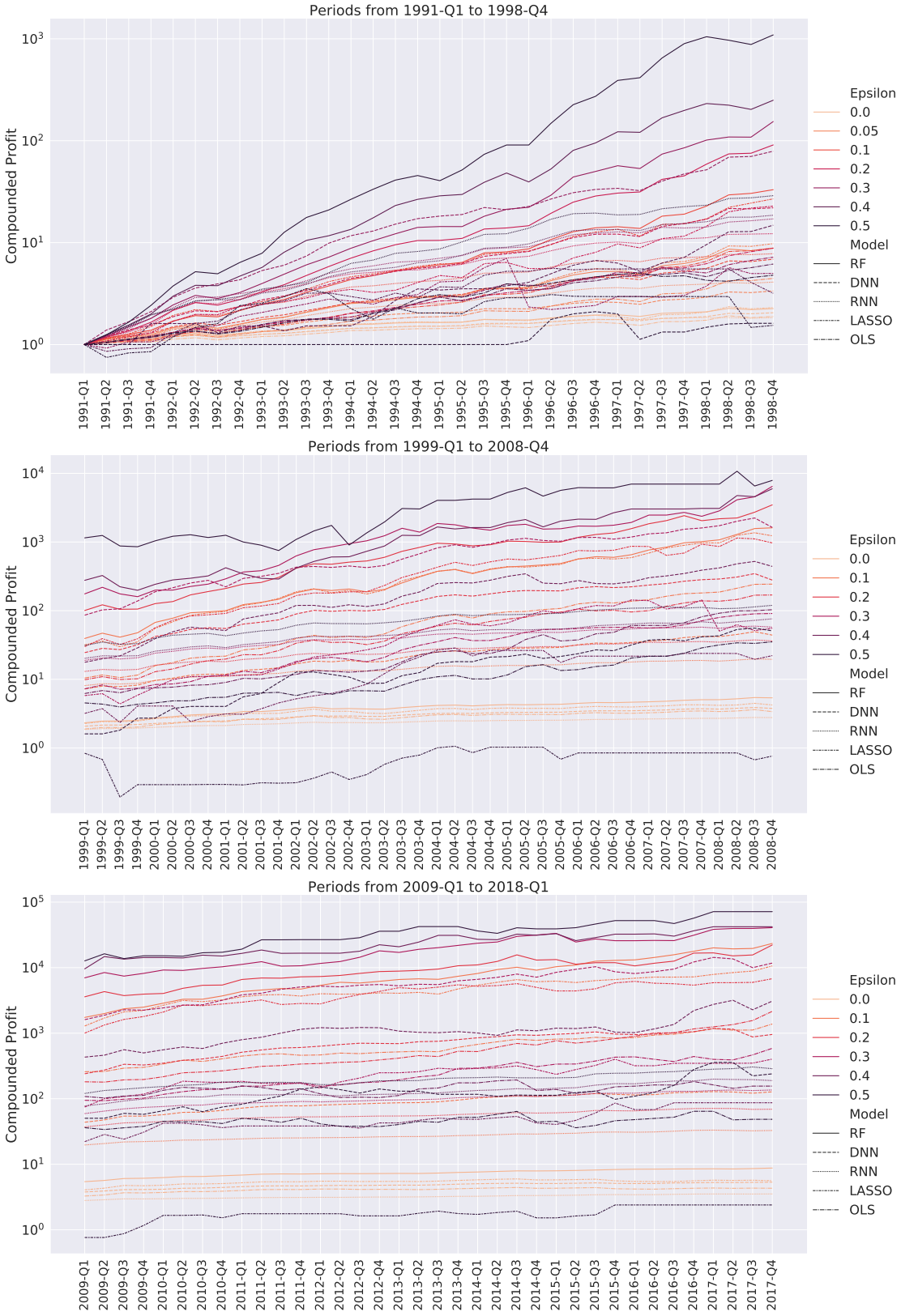


Figure 5.3: Compounded quarterly returns from 1991 to 2017 separated in thee charts of periods. Please note the different scale of the logarithmic Y-Axis among the charts.

Because of the time-series nature of these compounded profits, it is also of particular interest to investigate these results in a historical perspective. Therefore figure 5.3 demonstrates how the profits grow (decline) from quarter to quarter over the study period. Since the growth (loss) of the portfolio value in the quarters is in relative percentages and since potential losses reduce the portfolio infinitesimally, a logarithmic scale is chosen for these charts.

As expected, the choice of ε appears to control the volatility of the profit movements in both directions. The RF distinguished itself as the best performing model from the initial quarters up to the end of the study period. Both the $\varepsilon \in \{0.3, 0.4, 0.5\}$ scenario take the lead early on, while even the $\varepsilon \in \{0.1, 0.2\}$ case outperforms all other models in the years after 2009. This ranking of profitability according to the value of ε is not shared among other models. For instance, the $\varepsilon = 0.5$ scenario of the DNN doesn't trade (i.e. make a prediction outside the threshold) for four years until 1995. The Lasso model and the DNN represent two very strong models that compete for the second place. While there are only a few periods where Lasso outperforms the total profitability of the DNN, it trails closely.

With a large margin, the OLS and RNN model deliver the poorest performance. The $\varepsilon = 0.5$ scenario of the OLS also delivers the poorest performance overall. It suffers major losses in 1999 and struggles to make sufficient profits to recover. In general, models perform poorly in particular quarters like Q3 1999 or Q2 2009. This phenomenon could be attributed to macroeconomic anomalies in these periods like the global financial crisis or the dot-com crash. The dynamics of the RNN follow the low standard deviated as reported in table 5.4, it has barely any major swings but slight and consistent gains.

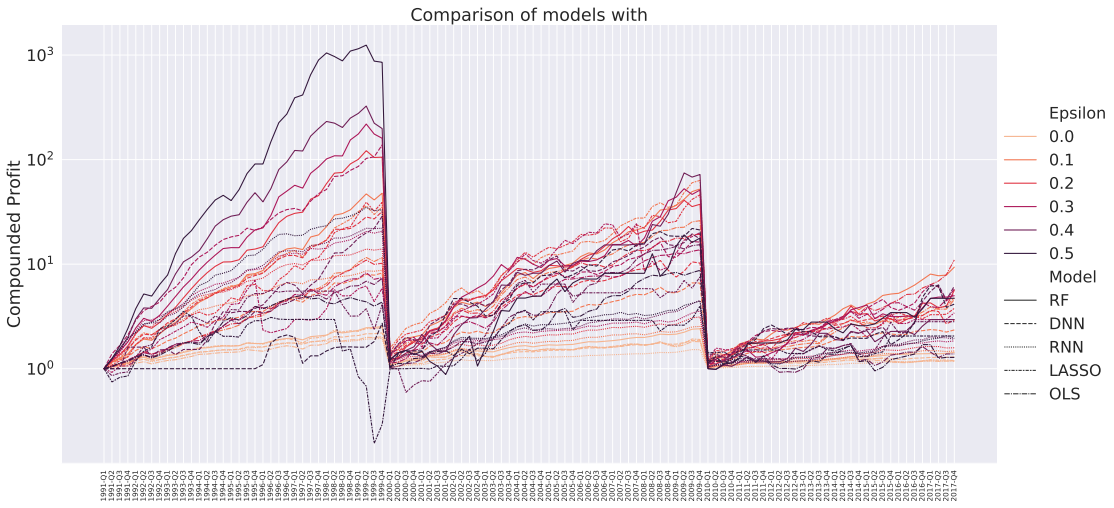


Figure 5.4: Compounded profits that are reset to 1 every 10 years using a logarithmic y-axis.

Since figure 5.3 foreshadows the growth rate of compounded profits to sink by the later quarters, figure 5.4 conducts the same profitability simulation but resets the compounded profits to 1 every ten years. While this visualisation suffers from a high information density, it is preferable to the previous figures that are split into three graphs as it allows to compare and visualise the maximal compounded profit and its growth on the same y-axis.

The figure also suggests that the reason for the high total profits of the RF $\varepsilon = 0.5$ model can mostly be attributed to the first decade from 1991 to 2000. In this period it, together with the other ε thresholds, has an exceptionally high accumulated return of up to 10^3. In the subsequent period, it only achieves an incremental compounded return of $10^{1.5}$ with the RF $\varepsilon = 0.4$ model outperforming it at a performance of close to 10^2. In the last, albeit slightly shorter period from 2010 to 2017, it then reaches a return of only $10^{0.8}$.

This sinking profitability, also evident across other models, is unfortunate as it foreshadows the trading strategy to be less profitable in eventual future quarters. However, the phenomenon is consistent with the efficient market hypothesis, which suggests that the abnormal profit should disappear as information about firms is better disseminated, and investor sophistication increases.

The results of the $\varepsilon = 0$ case demonstrated in figure 5.5 represent the low risk, low return, and high diversification scenario. It is a good example of the general ability of the models to generate signals that achieve steady quarterly returns.

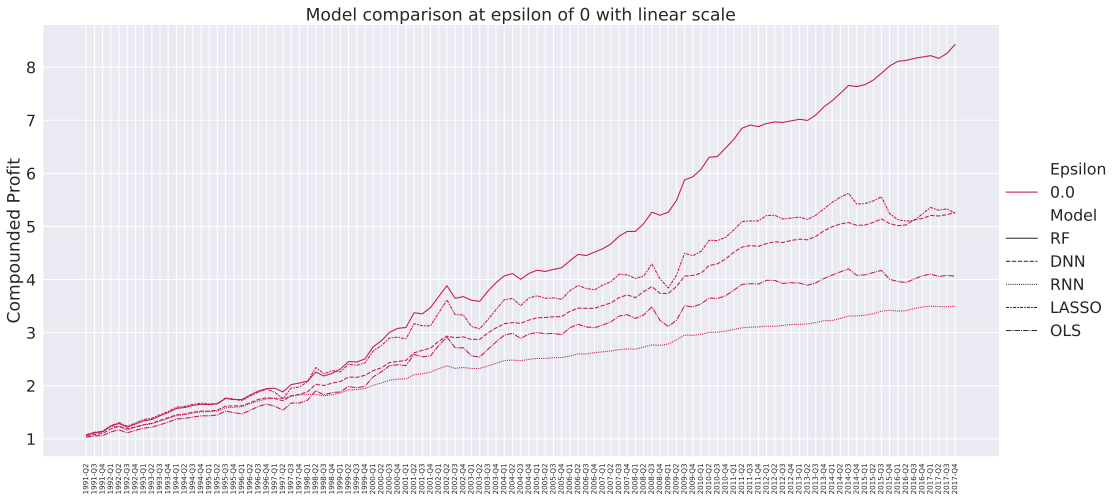


Figure 5.5: Compounded profits at $\varepsilon = 0$ on a linear y-scale.

Even with the low volatility in returns, the figures indicate the machine learning models to be less sensitive to economic anomalies than linear models. During the global financial crisis from Q3 2008 to Q1 2009, both the RNN and RF are relatively stable, while OLS and Lasso suffer significant losses. Interestingly, during the dot-com crash from Q3 2002 to Q1 2003, the RF, and DNN, are effected whereas the RNN Lasso and remain steady.

This ability of the RNN to have steady (albeit lower) returns in this setup is consistent with the relatively lower mean standard deviation of quarterly returns reported in table 5.4. It might be due to the explicit memory cells the GRU architecture allows and warrants further investigation into why the model was better able to anticipate and navigate these exceptional crisis periods.

5.2 Classification

Contrary to the regression setup, the prediction problem can also be constructed as a classification setup. Traditionally neural networks perform favourably in such

problem setups with many popular applications in areas like image recognition following this format.

The original Ou and Penman (1989) study also is constructed as a (binary) classification setting. The independent variable of this study represents the probability of an earnings increase for one year ahead, which the researchers also found to be predictive of positive stock returns for 12 and 24 months. The study in this thesis, however, is different in that it directly predicts the returns (not the yearly earnings change) and that the window over which the returns are predicted concerns just 30 days following a quarterly announcement. Because of these factors, a replication of the original study is not appropriate.

The most closely related study to this work by Dhar and Chou (2001) also categorises the dependent variable, which represents the future 20-day risk-adjusted return from 1 day after the announcement until 20 days after the announcement. They pick three bins based on whether Z score was less than -0.5 (negative earnings surprise), between -0.5 and 0.5 (no surprise), or greater than 0.5 (positive earnings surprise). Distribution of samples in this bin was 1:6:1 with 1,624 in the negative surprise, 8,994 samples in the no surprise, and 1,546 records in the positive surprise. The authors don't control the model and training data for this class imbalance. Their results will be used as a benchmark for the performance of the classification models in this thesis.

5.2.1 Loss function

For the classification setting the models will be trained using the entropy for the random forest and cross-entropy for the neural networks. These functions are similar, and their slight differences are explained in this section.

Introduced by Shannon (1948) in the context of information theory as the Shannon entropy, it refers to the *average rate of information produced from a stochastic process*. This information can be expressed mathematically as

$$I(E) = -\log_2[Pr(E)] = -\log_2(P) \tag{5.21}$$

with $Pr(E)$ representing the probability of event E, and $-\log_2(P)$ being a an equivalent short form for it. A very likely event with a probability of 99% happening $Pr(E) = 0.99$ would thereby have a low information content of $-\log_2(0.99) = 0.0144$ bits while a rare event of $Pr(E) = 0.01$ would be more informative with $-\log_2(0.01) = 6.64$ bits.

Using the formula for the expected value

$$E[X] = \sum_{i=0}^{n} p_i x_i \tag{5.22}$$

where x_i represents a finite outcome that occurs with the probability p_i allows to extend the information content formula to define the expected value of information so that

$$H(X) = E[I(X)] = E[-\log(P(X))] = -\sum_{i=0}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i) \tag{5.23}$$

for n classes of labels where $\mathrm{P}(x_i)$ are the ratio of elements of the respective label (representing the probability). A high level of $H(X)$ is considered a low level of purity (i.e. a high entropy).

The random forest (and consequently also the classification trees it consists of) uses this metric to chose the optimal next split to get maximally pure leaves.

The loss function of the neural net-based models is related to the idea of entropy and is referred to as the cross-entropy. Whereas the entropy in itself just concerns a single probability distribution as a parameter, the cross-entropy is defined as a distance measure for two discrete probability distributions p and q:

$$H(p, q) = -\sum_{i=0}^{n} \mathrm{p}(x_i) \log \mathrm{q}(x_i) \tag{5.24}$$

In the context of training models, it is used to find a *true* probability distribution p using a model that produces an approximated distribution q. The cross-entropy thereby becomes the difference measure that is to be minimised through the training process, much like the mean squared error in the regression setting.

Technically the distance measure in the cross-entropy is adopted from the Kullback-Leibler divergence (D_{KL}) which is also referred to as the relative entropy

of p with respect to q. It has the useful property that it is non-negative and that it is 0 if p and q are the same distribution. Therefore from an applied perspective, D_{KL} is low if p and q are close, and high if they are not. It is mathematically defined as

$$D_{\mathrm{KL}}(p \parallel q) = -\sum_{i=0}^{n} p(x_i) \log\left(\frac{q(x_i)}{p(x_i)}\right) \tag{5.25}$$

The equation for D_{KL} contains the log likelihood ratio LLR that expresses for the sample x_i of $i \in \{0, 1.., n\}$ how likely it is from distribution p rather than from q.

$$LR = \sum_{i=0}^{n} \log\left(\frac{q(x_i)}{p(x_i)}\right) \tag{5.26}$$

Adding the term $p(x_i)$ in the D_{KL} thereby makes the measure sort of an expected value (see equation 5.22) of the LLR.

Before being simplified algebraically to the form noted in equation 5.24, the cross-entropy consists of the entropy of p and the KL divergence $D_{KL}(p \parallel q)$ taking the form of

$$H(p, q) = H(p) + D_{KL}(p \parallel q) \tag{5.27}$$

The term $H(p)$ as the entropy of the true probability distribution p is fixed during training while only D_{KL} changes during training so that the optimal cross-entropy can't reach zero but only the entropy of p.

In an applied context the discrete category of the training label is encoded as a one-hot encoded binary variable. The class labels of a ternary classifictation c_0, c_1, c_2 would take the forms $c_0 = [0, 0, 1], c_1 = [0, 1, 0], c_2 = [1, 0, 0]$.

5.2.2 Model adaptions

Reformulating the problem as a classification requires minimal adaptions on the machine learning models. While the core dynamics and architectures of the models used will remain similar, some details need to be changed.

For the neural networks, the number of output nodes will be increased from one to the number of class labels. These new output nodes will be passed through a

softmax layer that normalises an n-dimensional input vector to an n-dimensional output vector in which the sum of the elements equals one.

Analytically the softmax function for element x_i in the n-dimensional output vector is defined as:

$$\text{Softmax}(x)_i = \frac{e^{x_i}}{\sum_{j=0}^{n} e^{x_j}} \tag{5.28}$$

Due to the one-hot encoded class labels, $\text{Softmax}(x)_i$ can be interpreted as the probability that the inputs are of the class i.

The linear models (i.e. OLS and Lasso) are replaced with a logistic regression that is also referred to as a Logit model. Due to the shared linear nature of the β-coefficients it is the appropriate replacement in the subsequent benchmarks. The development Logit is discussed by Cramer (2002) and it has been adopted in a wide range of disciplines to model a binary outcome probability. Theil (1969) created a multinomial extension of the original binary model, and this type of setup is used for the ternary classification setup in this thesis.

The regression tries to model a probability of the target variable y being one as the linear function of its predictors x_i and coefficients β so that

$$\Pr(y = 1 \mid x) = [1 + e^{-x'\beta}]^{-1} \tag{5.29}$$

Due to this specification, the predicted probability is within $[0, 1]$ and the learning process consists of finding the optimal set of values for β. The logistic regression implementation of scikit-learn (Pedregosa et al., 2011) is used for the experiments.

5.2.3 Binary Classification

The binary classification categorises the continuous dependent variable into two discrete classes depending on whether it is positive or negative. A positive

market reaction will be interpreted to be class 1, and a negative market reaction will be class 0.

Due to this transformation of the BHARs into classes, the models lose information that the regression models are exposed to. For example, both an extremely positive reaction of +50% is encoded for the model as in the same way as a minuscule positive reaction of +0.5%. While this restriction means that the models don't have information about the magnitude, they could learn factors that distinguish positive and negative reactions more accurately.

Receiver Operator Characteristic

The results of the binary classification are compared using the Receiver Operator Characteristic (ROC). This metric plots the true positive rate (TPR) on the y-Axis and false positive rate (FPR) on the x-Axis at various thresholds based on which the probability of a positive classification is decided. Both metrics are defined as following:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.30}$$

with TP representing the number of true positives (i.e. classification as class 1 with the ground truth also being class 1), and FN representing the false negatives (i.e. classification as class 0 with the ground truth being class 1) that is also known as a type II error.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5.31}$$

with FP representing the number of false positives (i.e. classification as class 1 with the ground truth being class 0) that is also known as a type I error, and TN representing the true negatives (i.e. classification as class 0 with the ground truth also being class 0).

The curves for the respective models on the ROC are compared to a random guess at various thresholds which represents the diagonal in the plot. Integrating the area under the respective curves (i.e. the AUC of a particular model) in the

ROC figure gives a number that summarises the performance of the model. A perfect prediction model would have an AUC of 1, while a random prediction model would have an AUC of 50%.

Figure 5.6 plots the ROC statistics of the various model types. It demonstrates that the random forest (RF) is the best performing model across all discrimination thresholds having an overall AUC of 55%. The second-best model is the deep neural net (DNN) with an AUC of 54%. The last places are the recurrent neural network (RNN) and the logistic regression (Logit) with an AUC of 52%. The RNN outperforms the Logit model for low discrimination thresholds, while the Logit model performs better at higher discrimination thresholds.

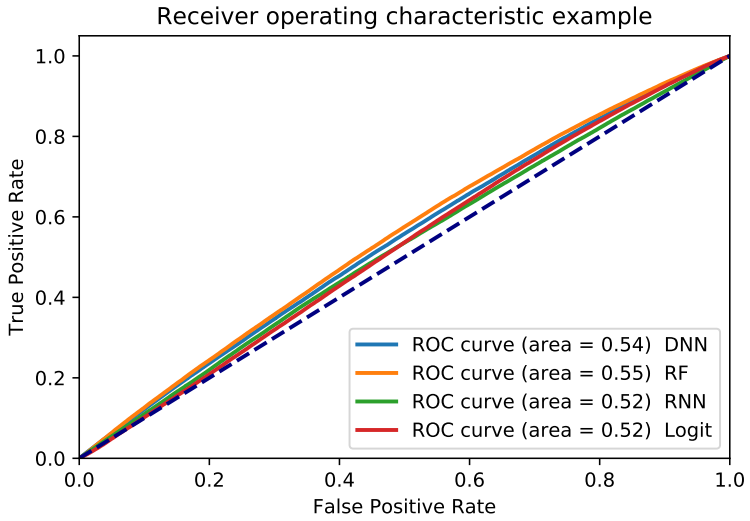


Figure 5.6: Receiver Operator Characteristic (ROC) curves of the tested model types

Using the method introduced by Hanley and McNeil (1983) it is possible to verify if the AUC values are significantly different. Using the H0 that *AUC of model y is not significantly better than AUC of model x* yields the p-values noted in table 5.7. They confirm the visual analysis of the ROC curves by demonstrating that the RF and DNN are better than other models while the RNN and Logit perform similarly poor. The necessary correlation coefficient r for the Hanley and McNeil (1983) method is calculated based on the Kendall rank correlation coefficient (Kendall, 1938).

y \ x	DNN	RF	RNN	Logit
DNN	0.5	0.99	***0.4e-29	*0.03
RF	***4.7e-11	0.5	***8.6e-11	**0.001
RNN	1	1	0.5	0.57
Logit	0.96	0.99	0.425	0.5

* p < 0.05, ** p < 0.01, *** p < 0.001

Table 5.7: p-values for the null hypothesis: AUC of model y is not significantly better than AUC of model x

Profitability

To translate the predictions of the binary classifier into an eventual profitability, this thesis again assumes a naive trading strategy similar to the one proposed for the regression models in section 5.1.4.

For every quarter Q there exists a set of test-set indices I_Q which contains sample indices $i \in I_Q$ for which there exists a ground truth label $y_i \in \mathbb{R}$ and a model-based prediction $\{\hat{y}_i \in \mathbb{R} : 0 \leq x \leq 1\}$ of the probability of a positive BHAR reaction.

These indices i comprise the total number of $|I_Q|$ positions that are taken in a quarter and that are being held over the BHAR horizon. The trading strategy weighs the nominal portfolio s over these positions so that the weight of one position w can be computed as

$$w = \frac{s}{|I_Q|} \tag{5.32}$$

According to whether this predicted probability is larger or equal 50%, either a long position ($\hat{y}_i > 0.5$), or a short position ($\hat{y}_i \leq 0.5$) is entered over the event horizon. The relative profit p_i of a position is therefore computed as:

$$p_i = \begin{cases} 1 + y_i & \text{if } \hat{y}_i >= 0.5 \\ 1 - y_i & \text{if } \hat{y}_i < 0.5 \end{cases} \tag{5.33}$$

The total profit of a quarter P_Q relative to the nominal portfolio s at the beginning of the quarter is then computed as the sum of all weighted positions:

$$P_Q = \sum_{i \in I_Q} w_i p_i \tag{5.34}$$

In comparison to the profitability of the regression models, the strategy trades on the entire set of test indices of a particular quarter due to the absence of ε thresholds. Therefore the results are most comparable to the $\varepsilon = 0$ threshold of the regression models. Compounded profits are calculated based on P_Q in line with the former definition in equation 5.18.

Table 5.8 compares the results of the compounded profit of a nominal portfolio of 1 over the whole study period. The first column contains the already discussed results of the regression models at $\varepsilon = 0$, and the second column contains the total compounded profit of the binary classification, as discussed in this section.

	Regression ($\varepsilon = 0$)		Binary Classification	
	CP	CAGR	CP	CAGR
RF	8.77	8.45%	8.05	8.11%
DNN	5.34	6.47%	7.24	7.68%
RNN	3.53	4.82%	4.93	6.15%
Lasso / Logit	5.62	6.66%	5.18	6.34%

CP = Compounded Profit, CAGR = Compound annual growth rate (see equation 5.20)

Table 5.8: Total compounded abnormal return and annualised compound growth rate between Q2 1991 and Q4 2017 per model type.

These results indicate that the reformulation of the problem as a binary classification yields improvements for the neural nets. The DNN increased its compounded profit from 535% to 724% of the original nominal portfolio of 100% resulting in an improvement of the compound annual growth rate of +1.19%. Also, the RNN improved its performance by a similar amount. The linear models and the RF worsened in their predictive ability. This evidence confirms the hypothesis that neural net models perform relatively better if predicting in a classification setting rather than in a regression setup.

5.2.4 Ternary Classification

Similar to the binary classification, the ternary classification turns the continuous dependent variable into three discrete classes. This approach is similar to the regression model with introduced ε in that based on the model output, it is possible to distinguish positive, neutral, and negative reactions.

For classification models of this kind, it is important to provide an even distribution of class labels so that none of the classes is over-represented in the training- and test-sets. Traditional approaches in machine learning address this concern by oversampling (undersampling) under- (over-) represented classes so that the data is synthetically augmented to be equally distributed.

In the problem setting of this thesis however, it is possible to augment the boundary of the classes so that they achieve an even distribution of samples. This is done by selecting the boundaries of the no-reaction class based on the $\frac{1}{3}$ and $\frac{2}{3}$ quantile. If the boundary is set according to these quantiles of the BHAR reaction, it is certain that an equal amount of samples fall are represented in all three categories.

The quantiles are constructed dynamically every quarter based on the distribution of the training data at the respective point in time. An empiric investigation demonstrated that in most quarters these boundaries resorts to the lower boundary of -0.08 for distinguishing negative from non-reactions, and an upper boundary of $+0.04$ to distinguish non-reactions from positive reactions. Subsequently in the training and test-samples where the ground truth $y < -0.08\%$ the class label 0

(negative reaction) is applied, for samples where $+0.04 > y > -0.08\%$ the class label 1 (neutral reaction) is applied, and for samples where $y > +0.04$ the class label 2 (positive reaction) is assigned.

This assignment of classes is different to the ε thresholds introduced in the regression setup in distinct ways: These classification labels are applied *before* training, they are are not symmetric around zero, and the boundaries are computed dynamically per quarter. Because of this distinction, the evaluated profitability simulation just represents one scenario which however follows the same intuition of the trading strategy introduced in section 5.1.4.

Confusion Matrix / ROC

Training and evaluating the models as discussed yields the normalised confusion matrices depicted in figure 5.7. The rows represent the ground truth class, and the columns denote the prediction. C0, C1, and C2 are the abbreviated class labels.

	C0	C1	C2
C0	**0.47**	0.24	0.29
C1	0.27	**0.43**	0.3
C2	0.36	0.31	**0.33**

(a) Confusion Matrix RNN

	C0	C1	C2
C0	**0.56**	0.24	0.20
C1	0.28	**0.52**	0.2
C2	0.41	0.34	**0.25**

(b) Confusion Matrix RF

	C0	C1	C2
C0	**0.51**	0.21	0.28
C1	0.33	**0.38**	0.29
C2	0.42	0.27	**0.31**

(c) Confusion Matrix DNN

	C0	C1	C2
C0	**0.49**	0.22	0.30
C1	0.26	**0.38**	0.36
C2	0.39	0.28	**0.34**

(d) Confusion Matrix Logit

Figure 5.7: The respective confusion matrices of the ternary classification are depicted here. The columns of the tables represent the predicted labels, while the rows represent the true label. The rows are normalised not to state the absolute number of samples, but the rate of samples.

The confusion matrices indicate that generally, the models are very capable of predicting negative and neutral reactions (i.e. C0, C1) but struggle with the positive reactions (i.e. C2). The RF performs best among the models, and the RNN and Logit model share the last place.

Comparison to Dhar and Chou (2001)

To compare the findings to the results of Dhar and Chou (2001), their tables have to be translated to a normalised confusion matrix. Additionally, the models in their study classify the ground truth into only positive and negative reactions, while

they differentiate the ground truth into three classes. Therefore the 'predicted' non-reaction labels have to be calculated from the ground truth number of samples in respective categories and the total number of predictions for C0 and C2. Table 5.8 presents the original and missing figures using their best performing model - a genetic algorithm (GA).

	C0	**C1**	**C2**	**Total**
C0	592	*862*	170	1,624
C1	718	*7,569*	707	8,994
C2	313	*807*	426	1,546
Total	1,623	*9,238*	1,303	12,164

(a) Original Confusion Matrix - Table IX GA - Dhar and Chou (2001)

Figure 5.8: The original confusion matrix from the Dhar and Chou (2001) paper. The *cursive* figures were not directly reported in the authors paper and therefore have been calculated indirectly based on the totals and other class samples. The matrix also has been transposed to follow the format of this thesis whereby the column represents predictions and rows the ground truth cases.

In their study Dhar and Chou (2001) identify the middle class (i.e. non-reaction) by using Z-Scores. These Z-Scores might be biased by a few extreme values (e.g. arising with liquidation events in the BHAR time horizon) which is why this thesis considers quantiles for the class boundaries instead. Choosing the Z-Scores -0.5 and $+0.5$ their data has a distribution of 13%/74%/12% as opposed to the 33%/33%/33% split in this thesis. The authors note in their study, "We focused only on bin 1 and bin 3 forecasts since they represent earnings surprises.". This incomplete analysis of the prediction performance, excluding the true positive and false positive

rate in the non-reaction bin, could mask the potential negative effects of the class imbalance in their data. Table 5.8 infers these not explicitly stated figures through the total number of samples and number of ground truth labels.

Comparing the best performing model, a genetic algorithm (GA), with the random forest model of this thesis yields the results presented in figure 5.9 in the format of a normalised confusion matrix. It indicates a substantially stronger performance by the method in this thesis in comparison the best results of the paper. This differential across the classes manifested in the bold figures might be caused by the unequal distribution of class labels in the experimental setup of their paper. The GA predominantly predicts non-reactions across all three ground truth labels; however, the true positive rate is higher than the false positive rate of a reaction of the opposite direction.

	C0	**C1**	**C2**
C0	**0.47**	0.24	0.29
C1	0.27	**0.43**	0.3
C2	0.36	0.31	**0.33**

(a) Confusion Matrix (RF) - Thesis

	C0	**C1**	**C2**
C0	**0.36**	0.53	0.10
C1	0.08	**0.84**	0.08
C2	0.20	0.52	**0.28**

(b) Confusion Matrix (GA) - Dhar and Chou (2001)

Figure 5.9: A comparison of the confusion matrices of the best performing models in this thesis and the work by Dhar and Chou (2001). The RF model refers to the Random Forest of this thesis while GA refers to a Genetic Algorithm. The columns of the tables represent the predicted labels while the rows represent the true label. The rows are normalised to sum up to 1.

Profitability

The benefit of applying the ternary classification to a trading setup lies in its ability to predict the non-reaction class in order not to enter a position. While in the regression setup, this selection is made ex-post based on the value of the prediction, the ternary classification is trained with this option not to enter a short/long position. Adopting the dynamic class boundaries for an even class separation from the previous section, another trading strategy is simulated similar to the one introduced in section 5.1.4. The difference lies in the selection of the trading indices $t \in T_Q$, and the decision of whether a long or short position is taken as the variable p_t.

Instead of equation 5.14 the trading indices are selected based on the predicted probability for the negative reaction class 0 ($\hat{y}_i^0$), non-reaction class 1 ($\hat{y}_i^1$), and positive reaction class 2 ($\hat{y}_i^2$)

$$T_Q = \{i \in I_Q | \hat{y}_i^1 \neq \max(\{\hat{y}_i^0, \hat{y}_i^1, \hat{y}_i^2\})\} \tag{5.35}$$

so that only indices of i are contained in T_Q where the highest predicted probability is either a positive or negative reaction.

The decision of whether a long or short position is entered is also different from equation 5.19 in that it is now based on whether the probability of a positive reaction is higher or lower than the probability of a negative reaction so that:

$$p_t = \begin{cases} 1 + y_t & \text{if } \hat{y}_t^2 > \hat{y}_t^0 \\ 1 - y_t & \text{otherwise} \end{cases} \tag{5.36}$$

Except for these adjustments, the trading strategy and the compounding of profits is simulated equally for the ternary classification.

Table 5.9 presents the results of the results of the ternary classification in comparison to the binary classification and the regression of the $\varepsilon = 0$ scenario. The figures illustrate that the RF benefits most from the ability to distinguish a non-reaction class achieving a compound annual growth rate of 11.53%. Interestingly the DNN has a poorer performance ranking even behind the Logit model. The

RNN, on the contrary, improves significantly in its predictive ability over to the binary classification specification.

Model	Regression ($\varepsilon = 0$)		Binary		Ternary	
	CP	AR	CP	AR	CP	AR
RF	8.77	8.45%	8.05	8.11%	18.62	11.54%
DNN	5.34	6.47%	7.24	7.68%	5.62	6.66%
RNN	3.53	4.82%	4.93	6.15%	7.39	7.7%
Lasso / Logit	5.62	6.66%	5.18	6.34%	5.75	6.76%

CP = Compounded Profit, AR = CAGR = Compound annual growth rate (see equation 5.20)

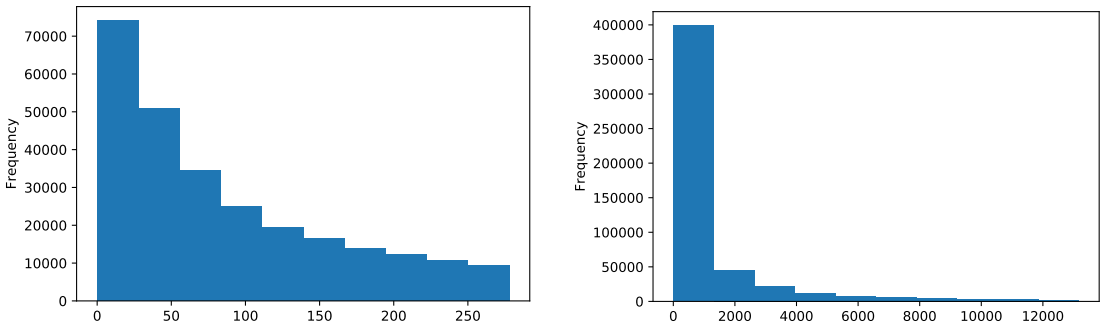
Table 5.9: Total compounded abnormal return and annualised compound growth rate between Q2 1991 and Q4 2017 per model type.

The profitability results of the ternary classification suggest that for most models this setup is preferable over a simple binary classification setup. It allows the models to predict samples as *non-reactions* to improve the return of the portfolios. This evidence is conclusive with the ε thresholds of the regression models and suggests that machine learning based weighing of portfolio positions could improve the returns further.

5.3 Market Capitalisation

Because of the high returns exhibited in the trading strategies, the upcoming chapter investigates what role the market capitalisation of firms plays that are being traded. For this purpose, analyses follow that address the question for which sizes of firms the models make the best predictions, and whether a set of firms can be identified that primarily drives the profits.

The market capitalisation of a firm is calculated by multiplying the share price with the total amount of shares outstanding. Both these values are contained in the Compustat FUNDQ file under the variable names `prccq` and `cshoq`. The distribution of the computed market capitalisation is visualised in figure 5.10.



(a) Histogram up to the 50% quantile.

(b) Histogram of up to the 99% quantile.

Figure 5.10: Histograms of the market capitalisation with a different upper quantile cutoff with respectively 10 buckets. The X-Axis depicting the market capitalisation is in thousand USD units.

For about 1.3% of the remaining firm-quarters, either the shares outstanding or the share price is no set so that the market capitalisation is missing. These data samples are excluded in the subsequent analysis unless stated otherwise.

In finance, jargon firms are often binned into groups along their market capitalisation. These bins are known as nano caps, micro caps, medium caps, and large caps. The limits for the market capitalisation, determining whether a firm is classified as being in one group or another, have changed over time as the mean market capitalisation of stocks has changed. Figure 5.11 demonstrates this based on the mean market capitalisation in of the firms in the data set of this thesis. It rose from $1,000,000$ USD in the second quarter of 1991, up to $8,500,000$ USD in the fourth quarter of 2017.

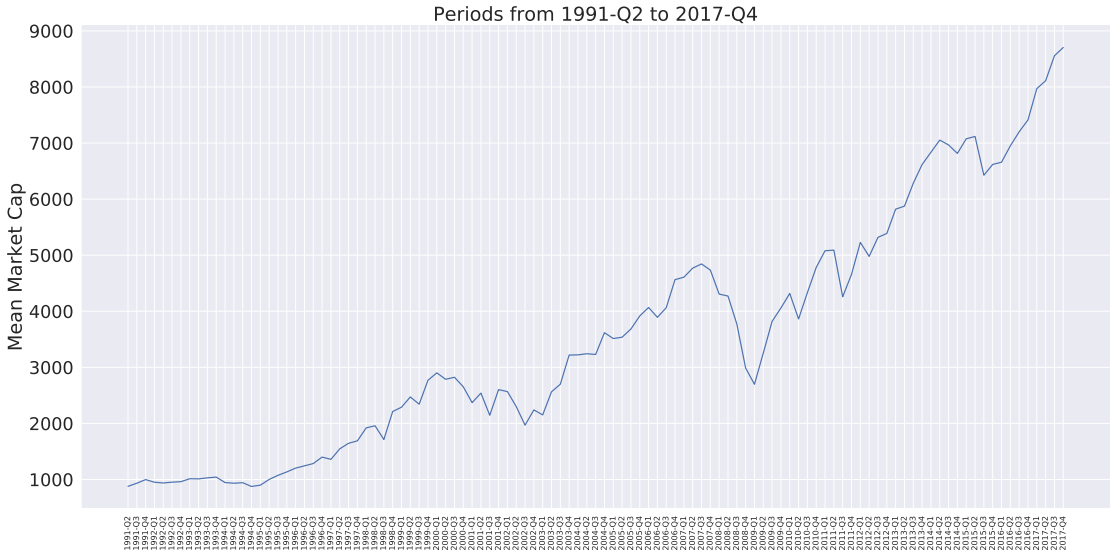


Figure 5.11: Mean market capitalisation from 1991-Q2 to 2017-Q4.

Due to this long period of three decades that the data set spans, the subsequently selected borders for the market capitalisation bins are chosen as a compromise. Table 5.10 depicts the selected limits for the traditional bins according to financial jargon, and table 5.11 contains limits for more granular bins that are also used in the subsequent analysis. The NaN Cap refers to firms where the market capitalisation could not be calculated because of at least one missing value.

Bin Name	Lower bound	Upper bound	Samples
NaN Cap			1.3%
Nano Cap		10	4.7%
Micro Cap	10	100	27.8%
Mid Cap	100	1,000	37.4%
Large Cap	1,000		28.8%

Table 5.10: Traditional market capitalisation bins in million USD

Bin Name	Lower bound	Upper bound	Samples
NaN Cap			1.3%
5m Cap		5	1.7%
10m Cap	5	10	3%
50m Cap	10	50	16.9%
100m Cap	50	100	10.9%
500m Cap	100	500	26.5%
1b Cap	500	1,000	10.9%
5b Cap	1,000	5,000	18.3%
+5b Cap	5,000		10.5%

Table 5.11: Granular market capitalisation bins in million USD

5.3.1 BHAR and the Market Cap

The first analysis concerns the question of whether there exists a relationship between the Buy-And-Hold Abnormal Returns (BHARs) and the market capitalisation. For this purpose three OLS regressions are constructed following the specifications:

$$\text{abs(BHAR)} = \beta_0 + \beta_1 \text{ SP} + e \tag{5.37}$$

$$\text{abs(BHAR)} = \beta_0 + \beta_1 \text{ SP} + \beta_2 \text{ SO} + e \tag{5.38}$$

$$\text{abs(BHAR)} = \beta_0 + \beta_1 \text{ SP} + \beta_2 \text{ SO} + \beta_3 \text{ MC} + e \tag{5.39}$$

with SP referring to the stock price, SO to shares outstanding, and MC to the market capitalisation. The error term e is normally distributed random variable with zero mean $E(e) = 0$. The coefficients for shares outstanding and market capitalisation refer to numbers that are denoted in thousands.

The results of the specified regressions are presented in table 5.12. While the R^2 measure is low, which is common in these setups, all coefficients are significant. In particular, they demonstrate a negative relationship between the market capitalisation (including its components explicitly) and the absolute value of the BHAR.

Table 5.12: OLS regression on the absolute value of BHARs in three versions

Variable	**Model 1**	**Model 2**	**Model 3**
Stock price	-2.665e-06***	-2.65e-06***	-1.218e-06*
	(6.9e-07)	(6.89e-07)	(6.93e-07)
Shares outstanding		-2.863e-05***	-6.925e-06***
		(1.23e-06)	(1.71e-06)
Market cap			-8.26e-07***
			(4.52e-08)
Intercept (β_0)	0.1652***	0.1685***	0.1690***
	(0.001)	(0.001)	(0.001)
N	498,184	498,184	498,184
R^2	0.000	0.001	0.002

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

These findings can be interpreted so that on average and ceteris paribus, firms with a larger market capitalisation have a smaller BHAR, firms with a higher amount of shares outstanding have a smaller BHAR, and firms with a smaller stock price have a larger BHAR. This evidence supports the hypothesis that the market reactions of larger magnitudes, primarily yielding the high profits in the trading strategy, might be the be concentrated among smaller firms and penny stocks.

5.3.2 Abnormal Profits of Bins

The ensuing investigation considers the question of what role the various bins of the market capitalisation play for the abnormal returns of the previous sections. The best performing model (i.e. the random forest in the regression setup) is picked for this analysis at various ε thresholds.

Initially, the experiment was conducted by forming dedicated portfolios based on all firms of a quarter in a particular bin. Unfortunately, this approach yielded unsatisfactory results as there was a large number of calendar quarters without or with only very few tradeable events (see equation 5.14). These quarters made the results not representative as they either led to no trading positions, or to high volatility for lack of diversification as the entire portfolio was just invested in one or two events.

A better and more robust way to measure the phenomenon is to construct portfolios that include all firms and differ from each other by leaving particular market cap bins out. Through this 'leave one out' approach the profit of the *standard portfolio*, which includes all firms, can be compared to the portfolios with omitted bins. The relative difference in the overall profit between these portfolios indicates the contribution that the bin made in the standard trading strategy.

If the profit of the *standard portfolio* is lower than the one where the bin is excluded the bin made a negative contribution, whereas if the profit of the *standard portfolio* is higher the bin made a positive contribution. As just one bin is omitted, this strategy doesn't suffer from the portfolio strategy problems induced by a low amount of events.

Let the standard portfolio denote a set SP_ε of all trading signals generated by the random forest regression model at various ε thresholds. These signals can be grouped along the market capitalisation of the firms they belong to. These bins, as defined in tables 5.11 and 5.10, have a lower bound (lb) and upper bound (ub) that allow to differentiate market cap specific portfolios

$$MP_{b,\varepsilon} = \{x \in SP_{\varepsilon} | lb \leq MC(x) \leq ub\} \tag{5.40}$$

per bin b, where the function $MC(x)$ retrieves the market capitalisation of the firm that the signal belongs to. The *leave-one-out* portfolio for bin b is then constructed as:

$$LP_{b,\varepsilon} = \{SP_{\varepsilon} \cap {MP_{b,\varepsilon}}^{c}\} \tag{5.41}$$

For each of these portfolios, the total compounded profit is calculated using the previously defined trading strategy in section 5.1.4. These profits are then normalised by subtracting and dividing by the total profit of the standard portfolio SP_{ε}. The result of this operation is presented in table 5.13.

	ε threshold						
Bin (b)	0	0.05	0.1	0.2	0.3	0.4	0.5
NaN cap	0.0	0.0	0.0	-0.06	0.1	0.58	0.14
nano cap	-0.12	-0.02	-0.19	-0.51	-0.81	-0.98	-0.97
micro cap	-0.21	0.49	0.11	-0.39	-0.49	-0.7	-0.77
mid cap	0.06	-0.2	0.04	-0.14	-0.01	-0.32	0.62
large cap	0.39	-0.04	0.01	0.11	0.32	-0.4	-0.57
1m cap	-0.0	-0.01	0.0	-0.1	-0.1	-0.1	-0.11
5m cap	-0.05	-0.02	-0.05	-0.24	-0.17	-0.68	-0.48
10m cap	-0.08	-0.01	-0.12	-0.17	-0.63	-0.81	-0.6
50m cap	-0.17	0.37	0.15	-0.56	-0.58	-0.36	-0.34
100m cap	-0.02	-0.03	-0.07	0.72	0.1	0.02	-0.52
500m cap	0.0	-0.17	-0.05	0.02	0.3	-0.41	0.59
1b cap	0.04	-0.03	0.09	0.05	0.0	0.08	0.37
+1b cap	0.39	-0.05	0.01	0.11	0.32	-0.4	-0.57

Table 5.13: Total compounded profit of the $LP_{b,\varepsilon}$ portfolios relative to the total compounded profit of the SP_{ε} portfolio. A positive number indicates that $LP_{b,\varepsilon}$ performs x% better than SP_{ε}, while a negative number indicates a poorer performance if the market capitalisation bin b is excluded.

The results indicate that for the high ε strategies, which have achieved the largest abnormal profit, firms of a smaller market capitalisation are very important. The total profit of the $\varepsilon = 0.5$ scenario would be 97% smaller if nano caps were excluded and 77% smaller if micro caps were excluded. On the contrary, the profits would be 62% larger if mid-caps were excluded, indicating that they diminish the

currently observed profits. However, the large caps contribute positively as the total profits would be 57% smaller if they were excluded.

The non-discriminating $\varepsilon = 0$ portfolio mirrors the observations made for the $\varepsilon = 0.5$ case for the smaller market caps; however, the large caps turn their sign indicating a negative contribution. This turn of sign suggests that through the ε threshold, the regression model can pick out the large-cap reactions that are more likely to be profitable.

This analysis confirms the hypothesis that firms of smaller market capitalisation play an important role in the abnormal profits achieved by the trading strategy. These results of the previous section, showing a negative relationship between the market cap and the magnitude of the BHAR, corroborate these results as correctly predicted higher BHARs, lead to higher profits during trading. A phenomenon deserving further investigation would be why mid-cap firms have a negative contribution to the results while large caps are important.

5.4 Variable Importance

Random regression forests allow computing a variable importance metric that rank the importance of particular variables for the model. The basic idea of this metric is that it can be calculated for any feature by accumulating the improvement in the splitting criterion metric (i.e. the sum of squares) that it yields.

Computed for every tree that constitutes the random forest, the variable importance for the whole forest is seen as the mean of these metrics in the same way as the prediction of the forest is the mean of the tree predictions. Being normalised, to sum up to one over all features that constitute the input vector, the metric helps to evaluate how much a particular feature improved the modelling ability in comparison to the others.

The most interesting variable importance in the context of this thesis concerns the random regression forest as the most profitable model. For it, the variable importance measure could be computed for every distinct quarter. To answer the

question of the generally most important variables, it makes sense to aggregate the variable importance measures over the entire study period.

For every modelled quarter, there are four quarters of financial statement variables that serve as inputs. Therefore the variable importance metric can be analysed along three lines:

1. The importance of particular financial statement variables in a particular quarter ahead of the announcement

2. The importance of the total financial statement of a particular quarter ahead of the announcement

3. The importance of financial statement variables across all quarters before the announcement

The subsequent sections are structured along these lines.

5.4.1 Particular variables in particular quarters

For this analysis, the variable importance measures which are normalised per quarter to sum up to 1, are added up per variable over all quarters of the study period. A higher number, therefore, indicates a higher importance overall.

Table 5.16 outlines the results and presents the top 10 variables of the 484 variables that constitute the input vector for the models. Besides the variable name in COMPUSTAT the table also includes the quarter from which the variable is important, the relative importance measure, and a brief description[2].

The analysis indicates the variables RECCHY (Change in Accounts Receivable), CHEQ (Cash and Short-Term Investments) and EPSX12 (Earnings per share Excluding Extraordinary Items) to be the important variables. Interestingly the quarters in which these variables are important are one quarter before the

[2]This description has been adopted from `http://www.crsp.com/products/documentation/quarterly-data-industrial`

announcement (i.e. the last quarter) and four quarters before the announcement (i.e. the upcoming quarter one year ago).

Variable	Quarter	Importance	Description
RECCHY	-1	62%	Accounts Receivable - Δ
EPSX12	-4	59%	EPS Excluding Extraordinary Items
CHEQ	-1	57%	Cash and Short-Term Investments
CHEQ	-4	56%	Cash and Short-Term Investments
EPSX12	-1	56%	EPS Excluding Extraordinary Items
RECCHY	-4	55%	Accounts Receivable - Δ
INVCHY	-1	54%	Inventory - Δ
APALCHY	-1	54%	Accounts Payable and Accr. Liab. - Δ
RECCHY	-3	54%	Accounts Receivable - Δ
INVCHY	-4	53%	Inventory - Δ
RECCHY	-2	53%	Accounts Receivable - Δ

Table 5.14: Top 10 important variables for the random regression forest over all periods. The column **Quarter** denotes how many quarters before the announcement quarter the financial statement variable is reported. The symbol Δ in the variable name indicates that the variable is a *change variable* that denotes the change in the account.

5.4.2 Particular quarters

As the top 10 of most important variables indicates a concentration of the most important variables in the immediate quarter before the announcement (i.e. the -1 quarter) and the quarter one year before the announcement (i.e. the -4 quarter), this analysis looks at whether this pattern repeats across all variables. Therefore this experiment sums up the variable importance measures along the relative lag

which the quarters they belong to have. Since the input vector is comprised of the last four quarters before the announcement, the past quarters go from -1 to -4.

The metric displayed in table 5.15 is computed along the just discussed motive by summing the variable importances per relative lag quarter over all quarters covered in the study period. It can therefore not be directly interpreted, but the difference between the figures indicates how much more important a particular quarter is than another.

Consistent with the previous section, quarter -1 and quarter -4 are the relatively most important ones. This is consistent with intuition about earnings reactions as analysts usually compare results quarter-over-quarter (i.e. with the -1 quarter) and year-over-year (i.e. with the -4 quarter).

Quarter	**Importance**
-1	29.5
-2	25.3
-3	25.4
-4	27.9

Table 5.15: This table outlines the relative importance of the inputs in a particular quarter before the announcement. The column **Quarter** denotes how many quarters before the announcement quarter the input quarter is reported.

5.4.3 Particular variables

The difference in this analysis is that it ignores the quarter lag and computes the sum of variable importances per financial statement variable. The methodology thereby follows the previous section but differs in that the sum is made per financial statement item, and not per relative lag quarter.

Variable	Importance	Description
RECCHY	223.0%	Accounts Receivable - Change
EPSX12	201.0%	EPS Excluding Extraordinary Items
CHEQ	203.0%	Cash and Short-Term Investments
INVCHY	208.0%	Inventory - Change
APALCHY	206.0%	Accounts Payable and Accrued Liabilities - Change
CHECHY	194.0%	Cash and Cash Equivalents - Change
AOLOCHY	178.0%	Assets and Liabilities Other - Change
CAPXY	174.0%	Capital Expenditures
WCAPQ	165.0%	Working Capital Quarterly
DVY	121.0%	Cash Dividends

Table 5.16: Top 10 important variables for the random regression forest over all periods. This table computes the sum of their importance regardless of the relative quarterly lag to the announcement.

5.4.4 Discussion

This section yielded interesting insights into the source of the signal based on which the abnormal returns of the random regression forest were achieved. All the identified most important variables are meaningful from the perspective of an accountant. These results suggest that the models did not overfit a particular edge case with obscure selected financial statement items, but rather that it is variables that stand in a direct relationship to earnings and firm operations.

The relative importance of the last quarter and the quarter from a year ago is intuitive as expectations by investors are anchored on both last years results (potentially due to seasonal nature of business), and last quarters results (as they are the most recent comprehensive indication about the health of the firm). In

contrast to multiple other works in financial statement analysis, the importance of past quarters for this type of model suggests that the approach of modelling the status of a firm at a particular point in time as the sequence of past financial statements, is a more promising strategy than merely assuming the most recent financial statement to be comprehensive enough.

While the selection of Earning Per Share (EPS), as the second most important variable, is intuitive as they are the primary metric on which the stock market reaction depends, their selection by the model indicates some sort of predictable time-series properties of them. In the context of the discussed literature in chapter 2, this evidence is coherent with traditional studies investigating earnings as predictors alone, and the works related to financial statement analysis that were successful without considering them.

6

Conclusion and Future work

The thesis began with a thorough review of the existing financial and accounting literature to motivate and demonstrate the theoretical viability of the research hypothesis. Multiple pieces of seminal accounting research demonstrated that components of financial statements are predictive of future earnings. This strategy for making earnings forecasts is known under the terminology of the *financial statement analysis* and has been popularised following the paper by Ou and Penman, 1989. The intuition for this type of model is evident as it uses accounting data from the balance sheet, income statement, and cash flow statement, to forecast another figure (i.e. earnings) that is calculated based on the accounting data of a future period.

Another corpus of research then started investigating the relationship between announced earnings and their corresponding capital market reaction using the terminology of the *earnings response coefficient*. This was motivated by the study of Ball and P. Brown (1968), which demonstrated that the sign change

of earnings is highly correlated to the direction of stock price reaction following their announcement. Later works then also found evidence that markets in these cases do not only react to the *earnings surprise*, but that newly released accounting figures like revenues or accruals also play an essential role.

Perceiving the stock market reaction from this perspective, therefore, begs the questions whether it could be predicted based on past financial statement data just like earnings. While the reviewed literature points towards the viability of the approach, the most potent theoretical opposition to this hypothesis is the efficient market hypothesis that in all its forms proclaims that past public data can not predict future the movement of future prices. This discord of theory around the phenomenon, the lack of research and empirical results along this line of reasoning was the motivation behind this thesis.

Following the introduction of the finance and accounting theory, the subsequent chapters concerned a series of empirical experiments used to evaluate the research hypothesis.

Chapter 3 discussed the dataset used and the way it was prepared for the models. First, the COMPUSTAT FUNDQ dataset was explained, which served as the source of the independent variables (i.e. the financial statement data) for the experiments. The biggest challenge in this context was the problem of missing values in the financial statement data. Multiple approaches have been evaluated to address this problem which arose due to the long time-span of the dataset and different reporting requirements for firms across industries.

The chosen strategy to deal with the missing values involved the classical approach taken in empirical finance and accounting research of reducing the data set by poorly populated columns, and a machine learning based imputation to interpolate missing values from existing ones in the remaining rows and columns. The imputation was supported by a large number of samples in the dataset which did not have an announcement date and would, therefore, have been of no use for the empirical experiments. A number of established imputation methods mostly devised for the field of computational biology, where missing values is a classical

problem, were implemented and tested to no avail. The eventually selected and successful method (i.e. *SoftImpute*) was developed in the context of recommender systems where the task of interpolating unknown ratings is functionally very similar to the inference of missing values.

Following the introduction of the financial statement data, the dependent variable in the model setup, which represent the market reaction, was decided to be the Buy-And-Hold Abnormal Return (BHAR). Since it is computed by subtracting the market return from the real return a stock has over a period, it can be interpreted as the abnormal return the stock delivered over this period. The measure was computed for all the announcement dates of the quarterly reports using the CRSP daily returns file.

To ensure the accuracy and real-world viability, the final part of the data chapter introduced the sliding window construction for the data samples and the construction of the train and test set. A concatenation the last four quarters (i.e. the last year of quarterly reports) preceding an earnings announcement, was empirically found to have the optimal trade-off between the dimensionality of the inputs and predictive ability. The validation following an expanding window that traverses from past to present ensured that the empirical experiments were evaluated out-of-time and out-of-sample for the entire study period from 1990 to 2017.

The methods chapter introduced the machine learning models that were evaluated in this thesis. Additionally, the chapter explained the choice of their specific hyperparameter settings to support full reproducability of the results. The selection of models included a deep neural network, a recurrent neural network, and a random regression forest. To justify the use of machine learning models for the experiments of the thesis, an OLS-, Logit-, and a Lasso regression were also introduced as prominent examples of traditional linear models that the machine learning methods were benchmarked against.

The experiments chapter of the thesis was dedicated to a series of empiric experiments to answer the main hypothesis of the research question. The first two sections of it represented two ways to set up the models for this purpose.

Initially, the models were trained and evaluated as a regression which stands in contrast to training as a classification. Rather than predicting probabilities of a positive or negative stock market reaction or certainty bounds, a scalar is predicted of which its sign indicates the direction, and the magnitude the severity of the abnormal stock market reaction.

First, the predictive ability of the models in a regression setup were evaluated based on traditional goodness of fit metrics like the MSE (Mean Squared Error), RMSE(Root Mean Squared Error), MAE (Mean Absolute Error), MedAE (Median Absolute Error), and the R^2. While these metrics were not directly indicative for the research question, the ranking of the best performing models suggested the best performing models to be the random forest and the lasso regression.

In order to evaluate the relationship between the magnitude of the predictions, and the direction of the true market reaction, the following section introduced the concept of ε-thresholds that distinguished significant positive, neutral, and significant negative reactions. To quantify the models' ability to predict the correct direction of the market reaction given a particular ε threshold, and to not miss predicting true significant reactions, a metric named PC was constructed and introduced. The results of these experiments corroborated the research hypothesis as they presented evidence that the correct direction can be predicted consistently in the majority of cases across the study period.

The third experiment concerning the regression type setup translated the abstract PC metric into concrete abnormal returns that could be achieved if the predictions were traded on. For this purpose, a simple trading strategy was simulated that constructed a portfolio of short and long positions on a quarterly basis. The abnormal returns differed across the models and ε thresholds with the best model, the random forest, achieving a mean quarterly abnormal return of 13%. In a further step, these quarterly returns were also compounded from the beginning of the study period in 1991 up to 2017 to yield a compounded annual growth rate (CAGR) of the simulated portfolio. Again, the random forest performed most favourably achieving returning an annual growth rate of 52%. Besides confirming the main hypothesis of

the research question, this analysis also corroborated the superior predictive ability of machine learning methods. Particular among the significant reactions given the higher ε thresholds, the performance of linear models faltered severely.

The classification based model setup was motivated by a potentially improved performance via an explicit binary (i.e. positive and negative reaction) and ternary (i.e. positive, neutral, negative reaction) classification of the market reaction. The analysis of the binary classification was conducted using a ROC analysis which yielded an AUC of over 0.5 for all models. Investigating whether the AUCs of the model types are significantly different, the random forest was found to be significantly better than the deep neural net ($p \leq 0.001$), recurrent neural net ($p \leq 0.001$), and the logistic regression ($p \leq 0.001$). In addition, the deep neural net in this analysis was found to have a significantly better AUC than the Logit model ($p \leq 0.05$).

Comparing the return of the binary classification to the return of a regression with $\varepsilon = 0$, the performance of the deep neural net increased considerably (from a CAGR of 6.47% to 7.67%) while the logistic lasso regression, as well as the random forest declined in their performance. The neural networks, therefore, disproportionally benefitted from the reformulation of the problem as a classification.

The ternary classification mimiced the ε thresholds as it give the models a way to identify non-significant reactions for which no position was taken during the trading simulation. Since this strategy was also followed by the design of the related study by Dhar and Chou (2001), it allowed for a comparative benchmark that highlighted the superiority of approach followed in this thesis. In terms of abnormal returns, the ternary classification outperformed the binary classification for all models except the deep neural net. Particularly the ternary classification random forest improved substantially from a CAGR of 8.11% to 11.5%.

Generally, the findings following the experiments conducted in the thesis confirmed the research hypothesis. The achieved abnormal returns are a promising signal that equity markets were imperfect in utilising the information of past financial statements when it comes to anticipating the upcoming publication. Results like the compounded return plotted in figure 5.4 indicated however that over the

years this ability has diminished as the potential abnormal returns in the decade from 1990 to 2000 were about two magnitudes higher than the ones in the years following 2010. This evidence is suggestive of increasing investor sophistication in the utilisation of financial statement data.

Since the financial statement data was normalised by total assets and total sales during its pre-processing, the models were not able to distinguish firms based on their market capitalisation. However, the question remained whether it is large or small corporations where this stock market anomaly was exploited for abnormal returns. For this end, experiments have been conducted to see what contribution particular segments of firms (grouped by their market capitalisation) made for the achieved abnormal returns. Simulating the trading strategy just on these distinct sets of firms was problematic due to the low amount of quarterly samples that would have been traded on. Therefore a *leave-out* approach was undertaken that measured how much lower (or higher) the compounded return of the complete portfolio is if a particular market capitalisation segment is not traded on.

For the model with the highest return, the random regression forest at $\varepsilon = 0.5$, the largest contribution to the overall returns were by firms with a market capitalisation of [1]: 5 to 10 million USD (-60%), over 1 billion USD (-57%), and 50 to 100 million USD (-52%). Two segments (100 to 500 million USD at +50%, and 500 million to 1 billion USD at +37%) made a negative contribution to the returns of the complete portfolio.

While the analysis showed that companies of smaller market capitalisation are important for the returns, the importance of large caps in this scenario indicated that the abnormal returns are not exclusively limited to illiquid small caps and penny stocks.

To understand the source of the signal derived from the financial statements, a dedicated section analysed the variable importance metric that can be computed for the random regression forest. The findings confirmed that it is not an exotic

[1]The number here denotes how much lower (higher) the total compounded return would have been if these firms were not traded on

combination of financial statement variables that were predictive, but rather a plausible selection of items that directly relate to the business activity such as the change in accounts receivable, cash- and short-term investments, (past) earnings per share, inventory change, or accounts payable. Additionally, the variables from the most recent quarter before the announcement, and the quarter one year before the announcement were deemed to be most important.

6.1 Future Research

This work has laid the foundations for multiple avenues of future work. One way to extend it is to investigate the abnormal returns in the context of the stock returns attributed to the risk factors identified by Fama and French (1993). Since the stock price anomaly around earnings contributes significantly to the yearly returns and volatility that a stock has, a particularly interesting avenue to extend this work is to translate the abnormal returns into daily raw returns that can be regressed.

Alternatively, it could also be of interest to evaluate how sensitive the computed abnormal returns are to transaction costs that are of particular concern to penny stocks. Studies such as Bhushan (1994) suggest promising ways to analyse transaction costs. Such an analysis might also provide evidence for investors about whether this anomaly is significant enough to be traded on in the real world. In an earnings momentum based strategy, Chordia et al. (2009) attribute 70 to 100 percent of the potential profit to transaction costs.

Overall the result that returns can be directly predicted for the earnings reaction using financial statement data suggests that further factors, like momentum, macroeconomic variables, or the textual content of financial statements, could improve the performance further. In an ideal world, identifying these factors and pricing them into the announcement ex-ante by opening positions, could decrease the volatility of the stock following the earnings call. While this study included just the financial statement items as contained in the Compustat FUNDQ dataset, an explicit

data mining approach similar to the cross-sectional study by X. Yan and Zheng (2017) might also produce synthetic variables that improve the results of this study design.

While in this thesis, the experiments and portfolios were conducted on a quarterly basis, an online machine learning approach might be promising as well. It is plausible that recently published financial statements of a different firm in the same sector (e.g. the Q1 2018 announcements by Microsoft) might be indicative about the results a similar firm might publish (e.g. Apple releasing it's financial statements for Q1 2018 one week later). Along this line, an alternative approach might also include modelling the growth and business dynamics of sectors on a monthly basis using the financial statements.

Another potential route to improve upon the thesis would be to employ transfer learning methods in the training of the quarterly models. A more sophisticated machine learning method thereby might explicitly learn to model both the time persistent and time-variant dynamics that influence the stock market reaction. Besides these firm-specific time dynamics that could be *transfer learned*, the method also lends itself to the various sectors and industries in which the firms operate.

Whereas this thesis was made possible by the early introduction of the standardised XBRL reporting standard in the United States, the upcoming introduction of this standard in the European Union opens the opportunity to replicate the study. If potential abnormal returns of this replication follow the pattern exhibited in the United States over the three decades where the utilisation of this standardised information format increased, then unique profit opportunities might arise in the initial years following this introduction.

References

Abarbanell, Jeffery S and Brian J Bushee (1998). "Abnormal returns to a fundamental analysis strategy". In: *Accounting Review*, pp. 19–45.

Ahmed, Anwer S (1994). "Accounting earnings and future economic rents: An empirical analysis". In: *Journal of Accounting and Economics* 17.3, pp. 377–400.

Albrecht, W Steve, Larry L Lookabill, and James C McKeown (1977). "The time-series properties of annual earnings". In: *Journal of Accounting Research*, pp. 226–244.

Anthony, Joseph H and Krishnamoorthy Ramesh (1992). "Association between accounting performance measures and stock prices: A test of the life cycle hypothesis". In: *Journal of Accounting and economics* 15.2-3, pp. 203–227.

Azur, Melissa J et al. (2011). "Multiple imputation by chained equations: what is it and how does it work?" In: *International journal of methods in psychiatric research* 20.1, pp. 40–49.

Bachelier, Louis (1900). "Théorie de la spéculation". In: *Annales scientifiques de l'École normale supérieure*. Vol. 17, pp. 21–86.

Ball, Ray and Philip Brown (1968). "An empirical evaluation of accounting income numbers". In: *Journal of accounting research*, pp. 159–178.

Ball, Ray and Ross Watts (1972). "Some time series properties of accounting income". In: *The Journal of Finance* 27.3, pp. 663–681.

Barboza, Flavio, Herbert Kimura, and Edward Altman (2017). "Machine learning models and bankruptcy prediction". In: *Expert Systems with Applications* 83, pp. 405–417.

Bartov, Eli, Dan Givoly, and Carla Hayn (2002). "The rewards to meeting or beating earnings expectations". In: *Journal of accounting and economics* 33.2, pp. 173–204.

Beaulieu-Jones, Brett K and Jason H Moore (2017). "Missing data imputation in the electronic health record using deeply learned autoencoders". In: *Pacific Symposium on Biocomputing 2017*. World Scientific, pp. 207–218.

Beaver, William H (1968). "The information content of annual earnings announcements". In: *Journal of accounting research*, pp. 67–92.

Beaver, William, Richard Lambert, and Dale Morse (1980). "The information content of security prices". In: *journal of Accounting and Economics* 2.1, pp. 3–28.

Beretta, Lorenzo and Alessandro Santaniello (2016). "Nearest neighbor imputation algorithms: a critical evaluation". In: *BMC medical informatics and decision making* 16.3, p. 74.

Bernard, Victor L and Jacob K Thomas (1989). "Post-earnings-announcement drift: delayed price response or risk premium?" In: *Journal of Accounting research* 27, pp. 1–36.

Bhushan, Ravi (1994). "An informational efficiency perspective on the post-earnings announcement drift". In: *Journal of Accounting and Economics* 18.1, pp. 45–65.

Biddle, Gary C and Gim S Seow (1991). "The estimation and determinants of associations between returns and earnings: Evidence from cross-industry comparisons". In: *Journal of Accounting, Auditing & Finance* 6.2, pp. 183–232.

Bird, Ron, Richard Gerlach, and Anthony D Hall (2001). "The prediction of earnings movements using accounting data: an update and extension of Ou and Penman". In: *Journal of Asset Management* 2.2, pp. 180–195.

Box, GEORGE EP, Gwilym M Jenkins, and G Reinsel (1970). "Time series analysis: forecasting and control Holden-day San Francisco". In: *BoxTime Series Analysis: Forecasting and Control Holden Day1970*.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Breiman, Leo et al. (1984). "Classification and regression trees. Wadsworth Int". In: *Group* 37.15, pp. 237–251.

Brooks, LeRoy D and Dale A Buckmaster (1976). "Further evidence of the time series properties of accounting income". In: *The Journal of Finance* 31.5, pp. 1359–1373.

Brown, Lawrence D (1993). "Earnings forecasting research: its implications for capital markets research". In: *International journal of forecasting* 9.3, pp. 295–320.

Brown, Lawrence D and Jerry CY Han (2000). "Do stock prices fully reflect the implications of current earnings for future earnings for AR1 firms?" In: *Journal of Accounting Research*, pp. 149–164.

Bryan, Stephen H (1997). "Incremental information content of required disclosures contained in management discussion and analysis". In: *Accounting Review*, pp. 285–301.

Callahan, Carolyn M and Rod Smith (2004). "Firm performance and management's discussion and analysis disclosures: An industry approach". In: *SSRN*. Available via https://dx.doi.org/10.2139/ssrn.588062.

Callen, Jeffrey L et al. (1996). "Neural network forecasting of quarterly accounting earnings". In: *International Journal of Forecasting* 12.4, pp. 475–482.

Cao, Qing and Mark E Parry (2009). "Neural network earnings per share forecasting models: A comparison of backward propagation and the genetic algorithm". In: *Decision Support Systems* 47.1, pp. 32–41.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp. 785–794.

Cho, Kyunghyun et al. (2014). "On the properties of neural machine translation: Encoder-decoder approaches". In: *arXiv preprint arXiv:1409.1259*.

Chordia, Tarun et al. (2009). "Liquidity and the post-earnings-announcement drift". In: *Financial Analysts Journal* 65.4, pp. 18–32.

Choromanska, Anna et al. (2015). "The loss surfaces of multilayer networks". In: *Artificial Intelligence and Statistics*, pp. 192–204.

Chung, Junyoung et al. (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555*.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2015). "Fast and accurate deep network learning by exponential linear units (elus)". In: *arXiv preprint arXiv:1511.07289*.

Collins, Daniel W and SP Kothari (1989). "An analysis of intertemporal and cross-sectional determinants of earnings response coefficients". In: *Journal of accounting and economics* 11.2-3, pp. 143–181.

Collins, Daniel W, Stephen P Kothari, and Judy Dawson Rayburn (1987). "Firm size and the information content of prices with respect to earnings". In: *Journal of Accounting and Economics* 9.2, pp. 111–138.

Cramer, Jan Salomon (2002). "The origins of logistic regression". In:

Davis, Angela Kay, Jeremy Max Piger, Lisa Marie Sedor, et al. (2006). *Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases*. Tech. rep. Federal Reserve Bank of St. Louis St. Louis, MO.

Dechow, Patricia M and Ilia D Dichev (2002). "The quality of accruals and earnings: The role of accrual estimation errors". In: *The accounting review* 77.s-1, pp. 35–59.

Dhar, Vasant and Dashin Chou (2001). "A comparison of nonlinear methods for predicting earnings surprises and returns". In: *IEEE Transactions on Neural networks* 12.4, pp. 907–921.

Donoho, David L (1995). "De-noising by soft-thresholding". In: *IEEE transactions on information theory* 41.3, pp. 613–627.

Easton, Peter D and Mark E Zmijewski (1989). "Cross-sectional variation in the stock market response to accounting earnings announcements". In: *Journal of Accounting and economics* 11.2-3, pp. 117–141.

Ertimur, Yonca, Joshua Livnat, and Minna Martikainen (2003). "Differential market reactions to revenue and expense surprises". In: *Review of Accounting Studies* 8.2-3, pp. 185–211.

Falas, Tasos, Andreas Charitou, and Chris Charalambous (1994). "The application of artificial neural networks in the prediction of earnings". In: *Proceedings of 1994 IEEE*

International Conference on Neural Networks (ICNN'94). Vol. 6. IEEE, pp. 3629–3633.

Fama, Eugene F (1991). "Efficient capital markets: II". In: *The journal of finance* 46.5, pp. 1575–1617.

Fama, Eugene F, Lawrence Fisher, et al. (1969). "The adjustment of stock prices to new information". In: *International economic review* 10.1, pp. 1–21.

Fama, Eugene F and Kenneth R French (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of financial economics* 33.1, pp. 3–56.

— (2000). "Forecasting profitability and earnings". In: *The Journal of Business* 73.2, pp. 161–175.

Financial Accounting Standards Board (1978). "Objectives of Financial Reporting by Business Enterprises". In: *Statement of Financial Accounting Concepts No. 1* FASB Statement of Concepts.

Forbes, William and George Giannopoulos (2015). "Post-earnings announcement drift in Greece". In: *Review of Pacific Basin Financial Markets and Policies* 18.03, p. 1550019.

Foster, George, Chris Olsen, and Terry Shevlin (1984). "Earnings releases, anomalies, and the behavior of security returns". In: *Accounting Review*, pp. 574–603.

Freeman, Robert N (1987). "The association between accounting earnings and security returns for large and small firms". In: *Journal of Accounting and Economics* 9.2, pp. 195–228.

Freeman, Robert N and Senyo Y Tse (1992). "A nonlinear model of security price responses to unexpected earnings". In: *Journal of Accounting research* 30.2, pp. 185–209.

Garcıa-Laencina, Pedro J, José-Luis Sancho-Gómez, and Anıbal R Figueiras-Vidal (2010). "Pattern classification with missing data: a review". In: *Neural Computing and Applications* 19.2, pp. 263–282.

Gers, F. A., J. Schmidhuber, and F. Cummins (1999). "Learning to forget: continual prediction with LSTM". In: *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*. Vol. 2, 850–855 vol.2.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Gordon, Myron J (1962). "The savings investment and valuation of a corporation". In: *The Review of Economics and Statistics*, pp. 37–51.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). "Speech recognition with deep recurrent neural networks". In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 6645–6649.

Greig, Anthony C (1992). "Fundamental analysis and subsequent stock returns". In: *Journal of Accounting and Economics* 15.2-3, pp. 413–442.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2018). *Empirical asset pricing via machine learning*. Tech. rep. National Bureau of Economic Research.

Hajek, Petr and Roberto Henriques (2017). "Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods". In: *Knowledge-Based Systems* 128, pp. 139–152.

Hanley, James A and Barbara J McNeil (1983). "A method of comparing the areas under receiver operating characteristic curves derived from the same cases." In: *Radiology* 148.3, pp. 839–843.

He, Kaiming et al. (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

Hechenbichler, Klaus and Klaus Schliep (2004). "Weighted k-nearest-neighbor techniques and ordinal classification". In:

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Holthausen, Robert W and David F Larcker (1992). "The prediction of stock returns using financial statement information". In: *Journal of accounting and economics* 15.2-3, pp. 373–411.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167*.

Jegadeesh, Narasimhan and Joshua Livnat (2006). "Post-earnings-announcement drift: The role of revenue surprises". In: *Financial Analysts Journal*, pp. 22–34.

Johnson, W Bruce and Rong Zhao (2012). "Contrarian share price reactions to earnings surprises". In: *Journal of Accounting, Auditing & Finance* 27.2, pp. 236–266.

Kalchbrenner, Nal and Phil Blunsom (2013). "Recurrent continuous translation models". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709.

Kasznik, Ron and Maureen F McNichols (2002). "Does meeting earnings expectations matter? Evidence from analyst forecast revisions and share prices". In: *Journal of Accounting research* 40.3, pp. 727–759.

Kendall, Maurice G (1938). "A new measure of rank correlation". In: *Biometrika* 30.1/2, pp. 81–93.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kinney, William, David Burgstahler, and Roger Martin (2002). "Earnings surprise "materiality" as measured by stock returns". In: *Journal of Accounting Research* 40.5, pp. 1297–1329.

Kormendi, Roger and Robert Lipe (1987). "Earnings innovations, earnings persistence, and stock returns". In: *Journal of business*, pp. 323–345.

Lev, Baruch and S Ramu Thiagarajan (1993). "Fundamental information analysis". In: *Journal of Accounting research* 31.2, pp. 190–215.

Li, Feng (2008). "Annual report readability, current earnings, and earnings persistence". In: *Journal of Accounting and economics* 45.2-3, pp. 221–247.

— (2010). "The information content of forward-looking statements in corporate filings—A naıve Bayesian machine learning approach". In: *Journal of Accounting Research* 48.5, pp. 1049–1102.

Livnat, Joshua and Richard R Mendenhall (2006). "Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts". In: *Journal of accounting research* 44.1, pp. 177–205.

Loh, Wei-Yin (2011). "Classification and regression trees". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1, pp. 14–23.

Maas, Andrew L, Awni Y Hannun, and Andrew Y Ng (2013). "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30. 1, p. 3.

Malkiel, Burton G and Eugene F Fama (1970). "Efficient capital markets: A review of theory and empirical work". In: *The journal of Finance* 25.2, pp. 383–417.

Mayew, William J and Mohan Venkatachalam (2012). "The power of voice: Managerial affective states and future firm performance". In: *The Journal of Finance* 67.1, pp. 1–43.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani (2010). "Spectral regularization algorithms for learning large incomplete matrices". In: *Journal of machine learning research* 11.Aug, pp. 2287–2322.

Mendenhall, Richard R (2004). "Arbitrage risk and post-earnings-announcement drift". In: *The Journal of Business* 77.4, pp. 875–894.

Mikolov, Tomáš (2012). "Statistical language models based on neural networks". In: *Presentation at Google, Mountain View, 2nd April* 80.

Miller, Gregory S and Joseph D Piotroski (2000). "Forward-looking earnings statements: Determinants and market response". In: *Available at SSRN 238593*.

Miller, Merton H and Kevin Rock (1985). "Dividend policy under asymmetric information". In: *The Journal of finance* 40.4, pp. 1031–1051.

Nair, Vinod and Geoffrey E Hinton (2010). "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.

Ou, Jane A and Stephen H Penman (1989). "Financial statement analysis and the prediction of stock returns". In: *Journal of accounting and economics* 11.4, pp. 295–329.

Paszke, Adam et al. (2017). "Automatic differentiation in PyTorch". In: *NIPS-W*.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Perols, Johan (2011). "Financial statement fraud detection: An analysis of statistical and machine learning algorithms". In: *Auditing: A Journal of Practice & Theory* 30.2, pp. 19–50.

Pigott, Therese D (2001). "A review of methods for missing data". In: *Educational research and evaluation* 7.4, pp. 353–383.

Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6, p. 386.

Rumelhart, David E, Geoffrey E Hinton, Ronald J Williams, et al. (1988). "Learning representations by back-propagating errors". In: *Cognitive modeling* 5.3, p. 1.

Seng, Dyna and Jason R Hancock (2012). "Fundamental analysis and the prediction of earnings". In: *International Journal of Business and Management* 7.3, p. 32.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *Bell system technical journal* 27.3, pp. 379–423.

Sharpe, William F (1994). "The sharpe ratio". In: *Journal of portfolio management* 21.1, pp. 49–58.

Shen, Kao Yi (2012). "The modeling of earnings prediction by time-delay neural network". In: *Advanced Materials Research*. Vol. 433. Trans Tech Publ, pp. 907–911.

Skinner, Douglas J and Richard G Sloan (2002). "Earnings surprises, growth expectations, and stock returns or don't let an earnings torpedo sink your portfolio". In: *Review of accounting studies* 7.2-3, pp. 289–312.

Sloan, Richard G (1996). "Do stock prices fully reflect information in accruals and cash flows about future earnings?" In: *Accounting review*, pp. 289–315.

Smith, Leslie N (2018). "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay". In: *arXiv preprint arXiv:1803.09820*.

StackOverflow User "nnnmmm" (2018). *StackOverflow - Comment 4830588 to question 48302810*. URL: https://stackoverflow.com/questions/48302810/whats-the-difference-between-hidden-and-output-in-pytorch-lstm/48305882#4830588 (visited on 04/01/2019).

Stober, Thomas L (1992). "Summary financial statement measures and analysts' forecasts of earnings". In: *Journal of Accounting and Economics* 15.2-3, pp. 347–372.

Theil, Henri (1969). "A multinomial extension of the linear logit model". In: *International Economic Review* 10.3, pp. 251–259.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tieleman, Tijmen and Geoffrey Hinton (2012). "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2, pp. 26–31.

Truong, Cameron (2010). "Post earnings announcement drift and the roles of drift-enhanced factors in New Zealand". In: *Pacific-Basin Finance Journal* 18.2, pp. 139–157.

— (2011). "Post-earnings announcement abnormal return in the Chinese equity market". In: *Journal of International Financial Markets, Institutions and Money* 21.5, pp. 637–661.

Watts, Ross L and Richard W Leftwich (1977). "The time series of annual accounting earnings". In: *Journal of Accounting Research*, pp. 253–271.

Wright, Sewall (1921). "Correlation and causation". In: *Journal of agricultural research* 20.7, pp. 557–585.

Yan, Xuemin and Lingling Zheng (2017). "Fundamental analysis and the cross-section of stock returns: A data-mining approach". In: *The Review of Financial Studies* 30.4, pp. 1382–1423.

Yan, Zhipeng and Yan Zhao (2011). "When Two Anomalies Meet: The Post–Earnings Announcement Drift and the Value–Glamour Anomaly". In: *Financial Analysts Journal* 67.6, pp. 46–60.

Yosinski, Jason et al. (2014). "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems*, pp. 3320–3328.

Zhang, Wei, Qing Cao, and Marc J Schniederjans (2004). "Neural network earnings per share forecasting models: a comparative analysis of alternative methods". In: *Decision Sciences* 35.2, pp. 205–237.