ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

The Robo-Investors from Graham-Doddsville

Applying Machine Learning to the Investment Choices of Warren Buffett

Bachelor Thesis

Ernst Florian Schweizer-Gamborino

December, 2020

Advisors: Prof. Didier Sornette, Dongshuai Zhao, CFA

Department of Computational Sciences, ETH Zürich

Abstract

Recent advances in machine learning techniques have raised the hopes that the investment decisions of so-called super-investors could be replicated by complex algorithms. This thesis aims to determine how much of this hope is well-founded. In particular, it investigates the performance of several well-established machine learning methods trained on the investment decisions of Warren Buffett as reported by Berkshire Hathaway via their 13-F filings from 1997 to 2019. The data underlying this binary classification task was collected on all companies publicly traded in the US during the pertinent time span, taking solely their fundamentals into account. The results show a definitive advantage of machine learning methods over using linear models, possibly pushing this approach into the realm of feasible and profitable investment strategies.

Acknowledgements

To professor Didier Sornette, who introduced me to the art of rationally looking at markets, and to my resourceful and patient advisor Dongshuai Zhao, who handled the more technical parts of said introduction, I extend my deepest gratitude. A special mention is deserved by Dr. Diana Gamborino – my wife – who worked wonders on my motivation. Additionally I feel most thankful for the time and attention of Dr. Sarah Vermij and Dr. Ruya Yuksel, whose proof-reading and critique was vital to presenting a thesis that has a running chance of being understood by any reader.

Contents

Chapter 1

Introduction

> If a business is worth a dollar and I can buy it for 40 cents, something good may happen to me.
>
> *Warren Buffett*

For anyone working in a contemporary financial institution or an investment firm, they will most probably work with an abundance of data - data which needs collecting, processing, analysing, and most importantly, interpreting. All these stages need the attentive care of specialists, and every little step towards furthering the productivity and accuracy of this process will contribute to the potential profitability of future investments. When it comes to the analysis and interpretation of financial data, the conventional weapon of choice has been linear regression, or, if one has acquired unusually many data-points on the usual amount of samples, its cousin sparsified linear regression. This is an efficient first approach that has yielded many important results that are easily interpretable, but often fails with the complex structure and the noisiness of financial data. Due to the exponential growth of computing power - modern-day laptops surpass supercomputers that have been in existence for a mere 20 years - we have access to machine learning methods that were unimaginable only decades ago. These tools can deal with massive amounts of data and can detect subtle, complex patterns that sometimes escape the notice of the human eye, a feat that may revolutionise financial statement interpretation.

Machine learning algorithms have found widespread use in many fields including medical diagnosis, computer vision, and natural language understanding. In finance and accounting their use has already been proven noteworthy in the areas of fraud detection, lending evaluation, risk assessment, and other fields. Their advantages compared to more conventional statistical models are two-fold: Machine learning algorithms can potentially uncover nonlinear relationships and subtle patterns in the data, and easily deal with cases where two or more variables are highly collinear. A significant drawback can be the risk of over-fitting, which means that due to the high dimensionality of its parameters, a machine learning algorithm can perfectly fit the training data yet predict nothing useful, an outcome that is exacerbated if the data exhibits high levels of noise. Despite the significant progress in the application of machine learning algorithms that study a plethora of financial factors influencing the expected returns (Wang and Luo, 2012; Zimmerman, 2016) [38, 27], this thesis will solely focus on the fundamentals of companies and entirely disregard financial time series data, which exhibits a very low signal-to-noise-ratio. Another reason why we don't directly use stock performance data as a learning target is that it might be heavily influenced by decisions that have little to do with the intrinsic value of a company.

Historically we can trace back the beginnings of automated financial statement report classification to discriminate analysis, (Altman, 1968) [3] which used several ratios to predict company bankruptcy. Other simple models were introduced later to detect financial fraud and earnings manipulation, combining several ratios into a predictive model (Beneish, 1997) [5]. Shortly afterward, machine learning algorithms were proven to be more powerful and useful when it comes to classification (Feroz et al., 2000; Lin et al., 2003; Kotsiantis et al., 2006; Perols, 2011) [14, 22, 21, 32], making the decision process more objective and exploiting the existing data more efficiently. Compared to conventional statistical models they don't require time-consuming processes of proposing relations and then proving or disproving them.

An effective method of improving the signal-to-noise ratio in order to reduce the risk of over-fitting is using domain expertise to enhance the financial statement data. Therefore, the treatment of our data integrates previous insights from accounting and finance research. Since Frazzini et al. (2016) [16] assessed that stock picking is indeed the central skill that drove the performance of Berkshire Hathaways' portfolio - and not solely Warren Buffett's prowess leading private firms - we use various indicators that have been proven to be effective measures of a company's success, returns, and risks. Firms with a higher credit risk tend to under-perform (Altman, 1968; Ohlson, 1980; Campbell, Hilscher, and Szilagyi, 2008) [3, 28, 37]. Ou and Oenman (1989) [30] show that certain ratios can indicate expected earnings, while Sloan (1996) [35] posits that firms with low accrual over-perform and firms with high accrual conversely under-perform. Piotroski (2000, 2012) [33, 34] proposes the F-score, combining nine binary financial ratios, which correlates with the expected returns, and Mohanram (2005) [25] propounds the G-Score, using eight accounting ratio signals, which predict high future stock performance. Montier (2008) [26] reports that firms that manipulated their statements have lower future earnings. Asness et al. (2019) [4] introduce a quality-minus-junk (QMJ) indicator, composed of 3 sub-indicators which are in turn composed of 21 sub-factors, discerning low performing companies from extraordinarily performing companies. We included all these indicators in our data set.

The central question of this thesis is whether it is possible to efficiently predict whether companies will be picked by Warren Buffett by solely using their financial statements. In order to do so, we collect and process the financial statement information of all companies that were traded on public exchanges in the US, and train several different machine learning algorithms on the outcome of whether they were selected by Warren Buffett or not. As we maintained earlier, the fundamental information alone might not be enough, thus we augment the data with the scores and ratios mentioned previously, giving an opportunity to check if traditionally efficient and predictive ratios from asset pricing and accounting literature (Altman, 1968; Kotsiantis et al., 2006) [3, 21] can enhance the performance of our models.

While the list of stocks held by Warren Buffett, as well as the annual financial reports of the respective companies are publicly available, making this data set a prime target for supervised learning, we have to add that Warren Buffett not only possesses the fairly rare skill set of a super-investor - that he humbly credits (1984) [6] to his education - allowing him to single out pertinent information that is off the books, buried in the annual report footnotes, yet deducible by a wit sharp enough. These are factors that our models do not take into account (yet), so a one-to-one correspondence between our predictions and the decisions of Warren Buffett would be highly surprising, if not deeply uncanny.

Chapter 2

Data Collection and Processing

2.1 Data Collection

2.1.1 Warren Buffet Picks

We collected the yearly reports of Berkshire Hathaway, which are required to list all investments according to the reporting requirements according to section 13(f) of the securities exchange act of 1934. [2] Relying on this data is generally enough to mimic the behaviour of investors, especially when their focus is not centered on short trades. [9] We did this via the Thomson Reuters Eikon database [1], where data was available starting from 1997. Requests for data prior to this date directed to the SEC were met with long pauses interrupted by awkward periods of silence. We then transformed each holding into a tuple of the company name and the year the stock was first held, henceforth named ″pick″ and ″pick year″ Some companies were duplicates since they issued different kinds of stock, in which case we dropped all entries but the first. Other holdings were ETF's, which are simply not conducive to the kind of analysis we intend to do. After this, we were left with 187 unique company - year pairs.

2.1.2 Company Screening

For each year, starting from 1997, we then collected all companies publicly traded at US exchanges and were reasonably financially active five years prior to each respective year. As a surrogate for 'reasonable financial activity' we took the fact whether a company has specified earnings before interest (EBIT) at the start and the end of this five-year period.

2.1.3 Company Year Assignment

Since we had many duplicate mentions of companies over the years, sample contamination was going to be a problem: When the five year window of financial information is just shifted by one year, one can easily figure out the picks by checking which entries have no overlaps at all with other entries. To avoid this, if a company was assigned to be considered at a certain year, it can't be assigned to any of the five following years yielding intervals containing unique financial information. The assignment to a year was done randomly.

2.1.4 Company Data Collection

After assigning each financially active company to five-year intervals, we collected their fundamentals as reported at the end of each financial year via the Thomson Reuters Eikon database. [1](c.f. appendix for a list of the fundamentals collected) Additionally, we collected information on the industries the company was active in, and embedded it in a one-hot encoding, which means that if the company i is active in industry j, $X_{i,j}$ would be 1.0, and 0.0 else.

2.2 Data Transformation

We scaled the whole data set to have 0 mean and a standard deviation of 1.0 relative to each variable, making each variable somewhat comparable to the other. Generally, we used pandas [24] and NumPy [29] to process our data, while we made use of matplotlib [20] and seaborn [39] to visualise it.

2.2.1 Company Data Cleaning

Since some companies had a lot of missing entries, we deleted those that missed more than 65% of all entries. The 65% threshold was determined in order to preserve as many positive samples as possible. Furthermore, we dropped columns that contained duplicate information but were less complete (E.g. ″Revenue″ vs. ″Total Revenue″) After this, we are left with 140 positive samples and 6166 negative samples, or a class imbalance of 2.14%.

2.2.2 Company Data Feature Engineering

Many machine learning algorithms cannot easily learn certain mathematical functions, such as ratios, the difference between ratios, or counts of elements above a certain threshold. [19] Therefore, we enhanced the fundamental data with scores and ratios well-known and broadly used in financial analysis: The *f-score*, which attempts to gauge the value of a company based on profitability, leverage, liquidity, and source of funds, as well as on operating efficiency (Piotroski 2000) [33] and the *g-score*, extending the idea behind the f-score, detecting growth signals, stability signals, and accounting conservatism (Mohanram, 2005) [25]. The *z-score* represents the probability that a company will go bankrupt within two years (Altman, 1968) [3], while the *o-score* is a further refined indicator of bankruptcy (Ohlson 1980) [28]. The *m-score* detects whether a company has manipulated its statements (Beneish, 1999) [5]. Additionally, we included ratios that were not directly incorporated by the scores mentioned above, such as the *current ratio*, measuring a company's ability to pay off its short-term obligation, *acidity* which similarly measures how well cash, marketable securities, and accounts receivable can cover current liabilities, *times interest earned* (TIE), relating EBIT to interest expenses, *inventory turnover* representing the number of times a company's inventory is sold per year, *fixed asset turnover* accounting for how well a company's assets are used to generate income, and *gross margin* being the percentage of gross profits in relation to net sales. Moreover, we implemented the company quality measures as indicated by 'Quality minus junk' (Asness et al. 2019) [4] which ranks companies according to three sub-indicators corresponding to *growth*, *safety*, and *profitability*.

2.3 Splitting

2.3.1 Train - Test Split

If we were to naively take the entirety of our data and directly fitted our models on it, we'd have a problem: How would we be able to tell whether our models will be useful for predicting data in the future or whether they just make a very convoluted copy of the data set — useless for anything else than representing the data we already have? In order to assess the predictive power of our models, we split our data into a training and a test set, such that 80% of it was used for training and the other 20% was set aside for testing purposes. Of course we had to make sure that both sets contained picks and non-picks in the same ratio. On top of that we also made sure that each 5-year interval roughly adhered to the same ratio of picks and non-picks, to avoid biasing the models (and their evaluation) toward recent or earlier picks.

2.3.2 Data Layout so far

Since we have fixed intervals, our data can be easily represented by the Matrix X_{train} respectively X_{test}, with the pick results as the vector y_{train} respectively y_{test}: For sample $_i$, the j-th entry is comprised by x_{ij} while the pick decision is represented by y_i, $y_i = 1.0$ when sample $_i$ is a pick, and $y_i = 0.0$ when sample $_i$ is a non-pick.

2.4 Patterns in the Data

Before we dive into the structure of our models, we want to explore how our data set looks like. As a quick way to gain insight into patterns in our data, we took the 5-year sample averages of all variables.

2.4.1 Principal Component Analysis

If we want to have a simple visualisation of how the data looks like, we have a problem: Our data lives in more than 100 dimensions, which is at least 97 too many to make sense to the human eye - if only there was a way to bring most of our data in a form that makes sense to our perception! It turns out that there is such a way, and it's called 'principal component analysis' (PCA). PCA projects our data onto a new set of mutually orthogonal coordinates that are ordered by how much they explain the variance within the data. If we then look at the data as it is projected onto the top n transformed coordinates, these representations account for most of the variation within the data set. [17] Projecting the whole data set onto its first three principal components as depicted in figures 2.1, 2.2, 2.3, and 2.4, shows that while the picks tend to be outliers, they are not part of a distinct cluster, highlighting the need for classifiers that pick up on nonlinear relations.

2.4.2 Class-Conditioned Variable Correlation

If we want to gain a further insight into how our data is structured, we can ask how the change in one variable is related to the change in another variable. This is called correlation and can be easily calculated for each variable pair k and l:

$$r_{l,k} = \frac{\sum_{i=1}^{n}(X_{i,l} - \bar{X}_l)(X_{i,k} - \bar{X}_k)}{\sqrt{\sum_{i=1}^{n}(X_{i,l} - \bar{X}_l)^2 \sum_{i=1}^{n}(X_{i,k} - \bar{X}_k)^2}} \tag{2.1}$$

with

$$\bar{X}_l = \frac{1}{n}\sum_{i=1}^{n} X_{i,l} \tag{2.2}$$

If we did this for the whole data set, without discerning between picks and non-picks, we would not get much additional insight. But if we take the picks and the non-picks and calculate the correlations for each possible variable pair separately, we can discover how the correlations between the variables differ across these two classes. To that end, we define a correlation matrix R_{picked} respectively $R_{not\ \ picked}$, where $r_{l,k,picked}$ and $r_{l,k,not\ \ picked}$ is the correlation between variable l and k if the sample was picked respectively not picked. We plotted the difference between these two correlation matrices, $R_{picked} - R_{not\ \ picked}$, in Figure 2.5.

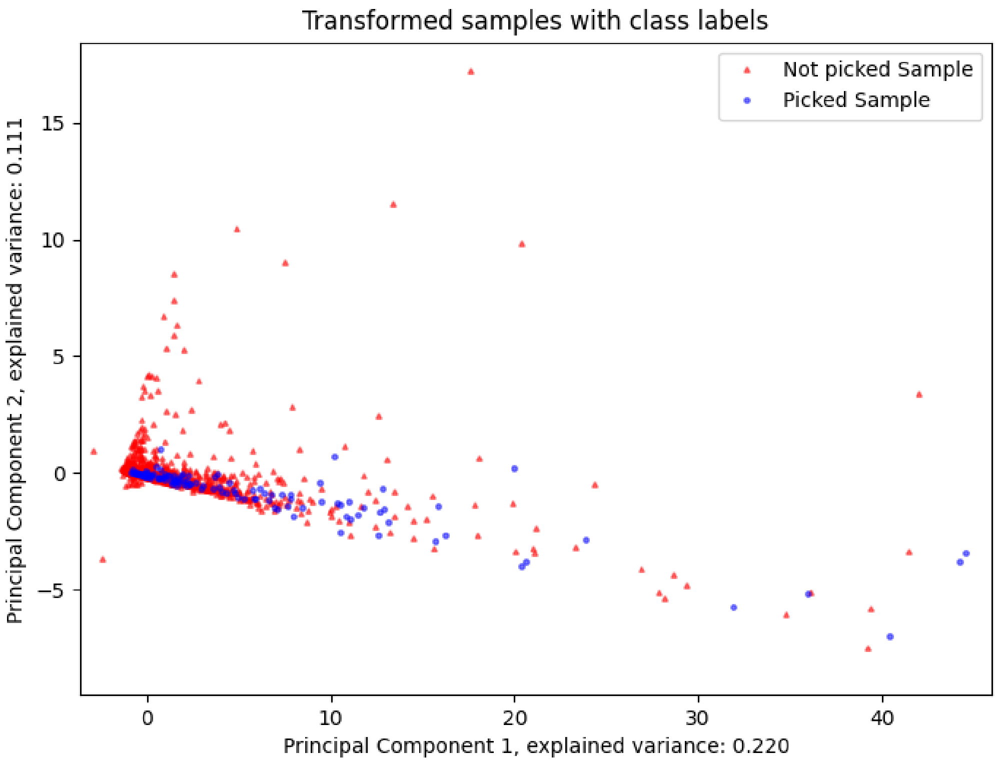

Figure 2.1: Samples projected onto principal components one and two

Surprisingly, many top-level engineered features such as the F-score and the G-score exhibited little to no differences in correlation, while some of their underlying engineered features yielded significant differences.

2.5 Advanced Statistical Properties

2.5.1 Averages

To see how the picks would differentiate themselves from the non-picks, we calculated the class-wise mean of each variable and calculated the difference between these means. The variables with the most relative difference were *Revenue* being income generated from 'normal' business operations, *Gross profit*, which is the revenue minus the costs associated with generating said revenue, *Selling/General/Administrative Expenses*, comprising all expenses except sales expenses necessary to generate revenue, and *Total Assets*, connoting the total of all economic resources belonging to the company. The variables with the least relative differences were *Principal Payments from Securi-*

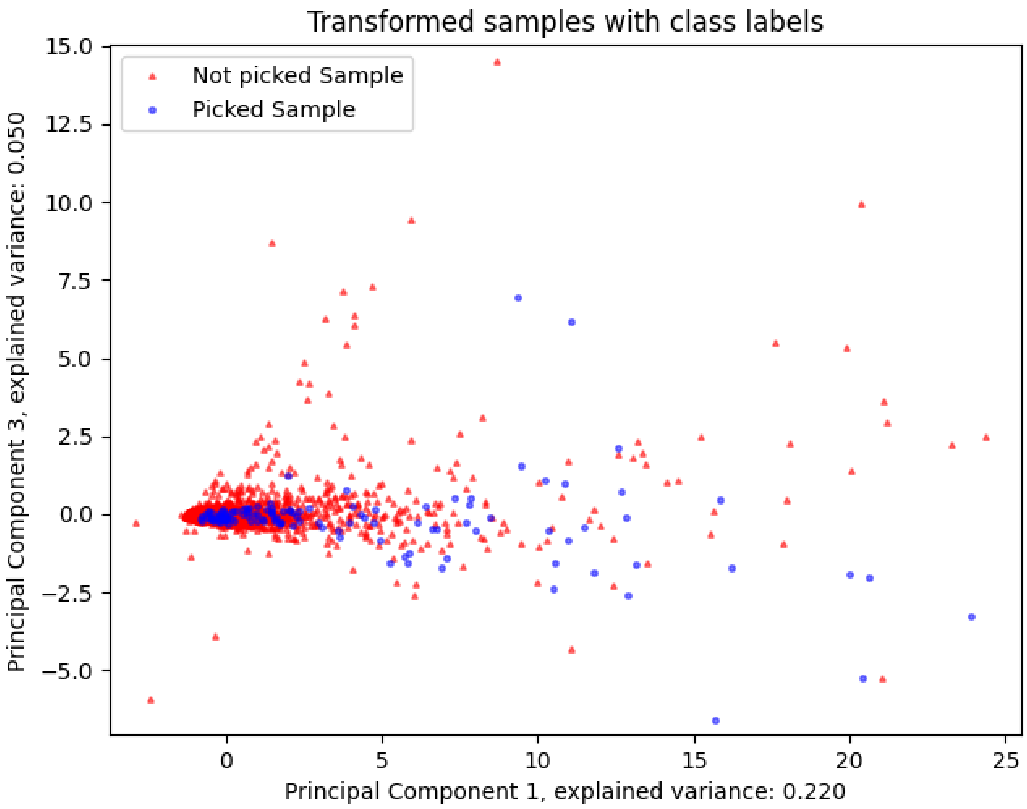

Figure 2.2: Samples projected onto principal components one and three

ties, *Total Interest Expenses*, *Return on Equity* — a ratio that specifies how much net income per shareholder's equity a company generates — and $\Delta_{cashflow}$, which measures the yearly differences in cash-flow. On Table 2.1 we list the top 15 variables sorted by descending absolute difference, the complete list can be found in the appendix, which we provide for each variable comparison.

2.5.2 Feature Importance

To further investigate the importance of single features, we made use of the XGBoost [7] library, which provides feature importance estimates by checking how often and how centrally a variable was used to split the boosted decision trees. (Hastie, Tibsharani, Friedman, 2017) [18] Again, we list the 16 most important variables in table 2.2 by descending order of importance.

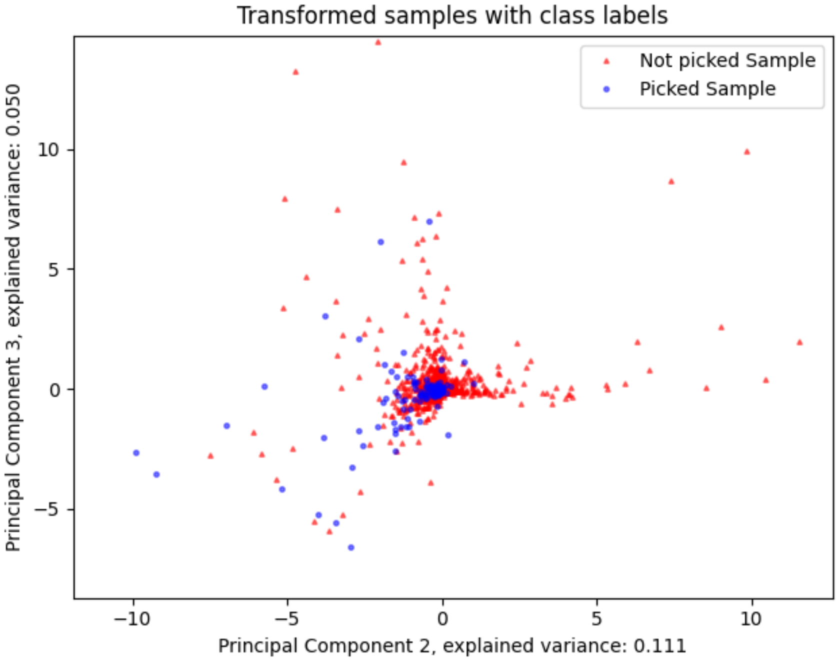

Figure 2.3: Samples projected onto principal components two and three

2.5.3 Feature Importance by Obscuration

To conduct another basic test to indirectly check which variables (or group of variables) influenced the predictive power of our models the most, we cross-validated (refer section 3.2) the omission of a single variable - or variable group, as in the case of the scores and the industries - on 10 folds with the performance of boosted linear learners. We additionally included the possibility that scaling all fundamentals by Total Assets would increase the predictive performance (which it did not). Among all variables, the information on the industry field was performing the best, since its removal lead to a drop in performance of nearly 10%, leaving all other indicators behind. Table 2.3 shows the 15 variables whose absence impacted the cross-validated accuracy the most.

2.5.4 Synopsis

Comparing the different rankings of the variables, we can point out several variables that are crucial for discerning picks from non-picks: we can see that revenue is much higher on average among the picks, meaning that Warren Buffett doesn't bother with businesses that generate little income. In a similar vein, gross profit, earnings before interest, taxes, depreciation, and

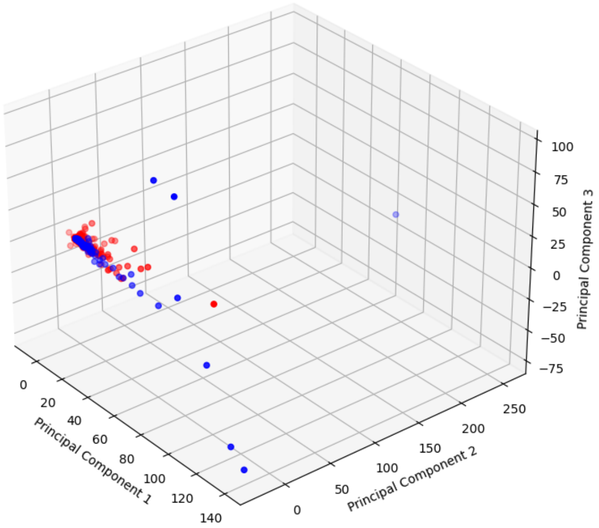

Figure 2.4: Samples projected onto the first three principal components

amortisation (EBITDA), total assets, as well as Selling, general and administrative expenses are on average much higher among firms that were picked, indicating that the size of a company really matters. While research and development is an important discerning factor, difference between picks and non-picks is about one standard deviation, much less than the two standard deviations of the top ten features. This means that the companies picked spend less than the average company on R&D when compared to their size. Surprisingly, the return on assets (ROA), which is the net income divided by total assets, of picked companies is lower than the ROA of not picked companies. This discrepancy, together with the near-zero class-wise difference in ROE, could either hint at picked companies utilising their capacity to shoulder debt more efficiently, or mean that it is more difficult for big companies to consequently attain a high ROA — though this conjecture warrants further investigation. However, when we take a look at the correlation differences (2.5) we can see that ROA accounts for some of the highest differences between within-class correlations. This shows that Warren Buffett mediates common investor knowledge by more subtle differentiations.

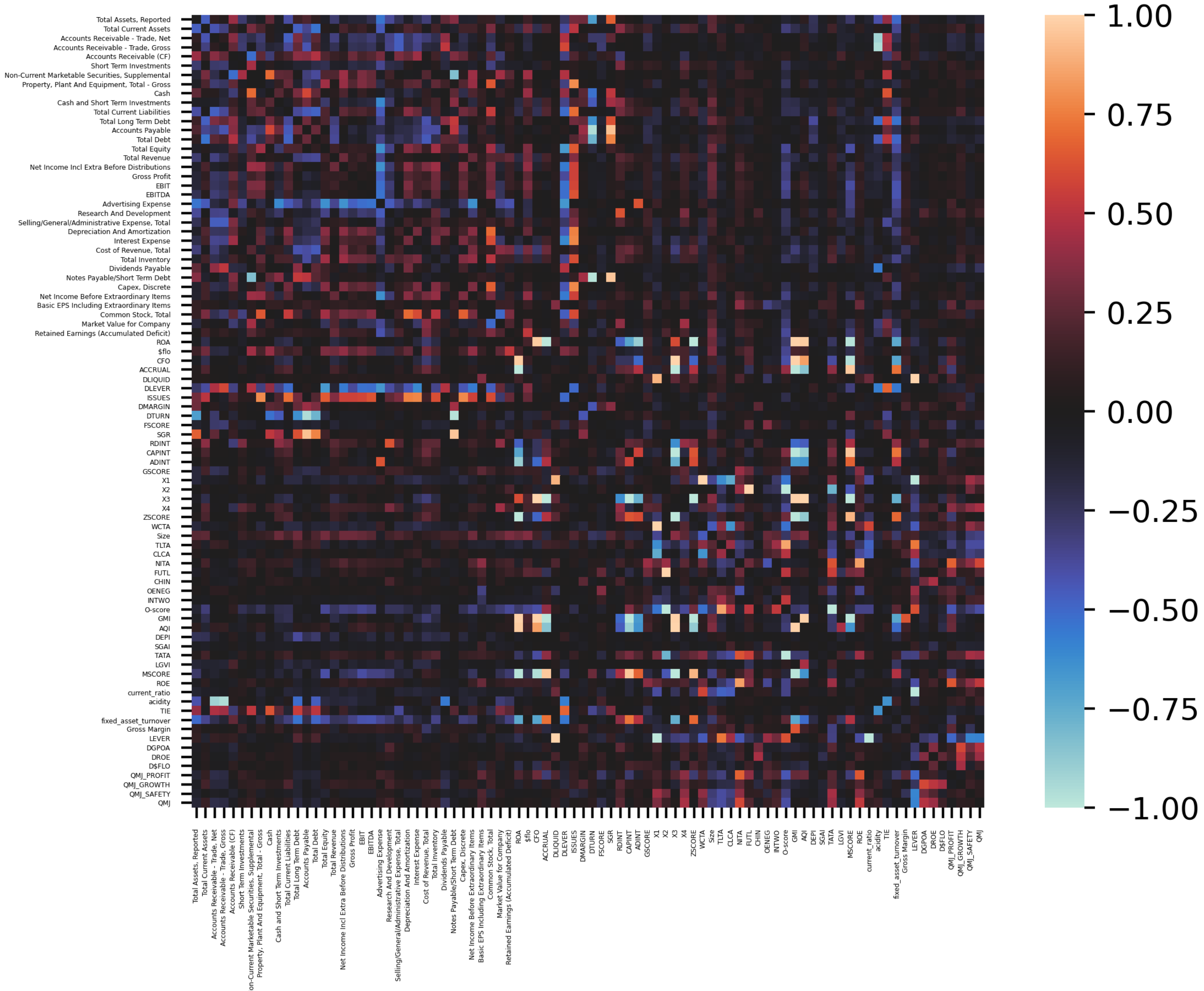

Figure 2.5: The between-class difference of within-class correlations. The closer the value is to 1.0, the more correlated are the variables in the picked samples when compared to the non-picks. The closer the value is to -1.0, the variables in the samples not picked are relatively more correlated when compared to the picked samples.

Variable	*D*ifference
Revenue	2.12
Total Revenue	2.12
Gross Profit	2.09
Selling/General/Administrative Expense, Total	2.02
Total Assets, Reported	2.00
EBITDA	1.92
Operating Expenses	1.91
EBIT	1.85
Property, Plant And Equipment, Total - Gross	1.84
Property/Plant/Equipment, Total - Net	1.82
Total Current Liabilities	1.80
Total Equity	1.79
Interest Expense	1.73
Cost of Revenue, Total	1.67
Total Liabilities	1.66

Table 2.1: Differences between class averages

Variable	Gain
Gross Profit	20.21
Net Income Incl Extra Before Distributions	14.04
EBITDA	7.46
Accounts Receivable - Trade, Net	7.39
Total Current Liabilities	6.48
Property/Plant/Equipment, Total - Net	6.28
Selling/General/Administrative Expense, Total	6.24
Total Inventory	6.07
Accounts Receivable (CF)	6.02
Long Term Debt	3.70
Property, Plant And Equipment, Total - Gross	3.56
Total Assets, Reported	3.49
Interest Expense	3.41
Depreciation And Amortization	3.40
Research And Development	3.39
Advertising Expense	3.39

Table 2.2: Variable Importance as inferred by XGBoost

Variable	Accuracy
Industries	0.811 8
Scaled	0.843 2
Revenue	0.903 3
Operating Expenses	0.907 7
Gross Profit	0.907 7
Net Income Incl Extra Before Distributions	0.907 8
Retained Earnings (Accumulated Deficit)	0.907 9
Operating Income	0.907 9
EBIT	0.907 9
Selling/General/Administrative Expense, Total	0.907 9
Net Income Before Extraordinary Items	0.908 0
Research And Development	0.908 0
TIE	0.908 1
O-score	0.908 1
EBITDA	0.908 1
⋮	⋮
Baseline	0.908 4

Table 2.3: Impact on cross-validation accuracy by variable omission

Chapter 3

Methodology

In this chapter we will introduce the general framework of our approach as well as the tools and models used to analyse our dataset.

3.1 Evaluation Metrics

Defining the evaluation metrics is vital to understanding the performance of our models.

3.1.1 Accuracy, Recall, and Precision

Since we have a binary prediction target, we have four possible outcomes for each prediction: In the case that we are looking at a company that was not picked by Warren Buffett, our models can either (correctly) predict that it will not be picked, yielding a true negative - TN - or (falsely) predict that it will be picked, yielding a false positive, FP. Conversely, if we are looking at a company that was picked by Warren Buffett, the prediction can either be a true positive, TP, or a false negative, FN. From these outcomes, we can construct our basic evaluation metrics. Since our data set is severely imbalanced, *accuracy* as defined by $\frac{TN+TP}{TN+FP+FN+TP}$ will deliver little insight, as an algorithm that consistently predicts negatives will yield an accuracy of 97.78%. An ameliorated version of this metric is *balanced accuracy*, where we weigh the predictions by the number of negative respective positive samples: $0.5 * (\frac{TP}{TP+_FN} + \frac{TN}{TN+FP})$ Another pair of metrics we are interested in is recall vs. precision: *Recall* - often also called sensitivity - measures how well the model identifies positive samples, and is defined by $\frac{TP}{TP+FN}$. *Precision* measures how many of the samples labeled as positive are true positives, and is defined as $\frac{TP}{TP+FP}$ These two metrics specifically interest us because the first tells us how well picks are identified, while the latter tells us how much Warren Buffett we should expect to be contained in a sample that is labeled 'Warren Buffett'.

3.1.2 Receiver Operator Characteristics and Area Under the Curve

The former metrics depend on binary and discrete classifications. What if we want to evaluate our models on a more gradual scale, say probabilistic or ranked predictions? Following metrics do help us with these issues: [13] The Receiver Operating Characteristics (ROC) graph helps in determining the trade-off between recall and loss of specifity. While discrete classifiers can be situated on a single point on the ROC graph, its true evaluative power comes from plotting the recall vs. the specifity of a classifier while varying the decision threshold. As Fawcett (2006) [13] notes, the curves generated in this way have the property that they do not change when the class distribution of the prediction targets changes, making the ROC graph a suitable and desirable way to evaluate the performance of our models. Another interesting property is the Area Under the Curve (AUC) directly calculated from the ROC graph, which is at 0.5 for a completely random classifier, translates into the probability that a randomly chosen positive sample will be ranked higher than a randomly chosen negative sample. This makes it an exceptionally solid metric when determining the overall predictiveness of a model [13]

3.2 The Importance of Cross-Validation

While training appropriate models, most of them will necessitate tuning parameters to account for the differences in underlying reality. If we were to directly tune these parameters according to the performance on the test data set, our models will most likely over-fit due to the richness encoded by the parameter space. Hence we we split our training data set into n folds - such that each fold contains the same proportion of positive to negative samples - and train our models, with varying parameters, n times on $n-1$ folds while holding back the nth fold for validation. Averaging the scores across the n folds will give us a robust estimate of the relatively best parameter set, while greatly reducing the risk of over-fitting and of selection bias. (More on cross-validation in Stone (1974) [36]) In the case of our simpler models, we chose $n = 10$, whilst we chose $n = 5$ for training our neural networks, in order to reduce computation time.

3.3 Models Used

In the following subsections we will detail the models we used to analyse our data. Wherever our models were hampered by the imbalance of our data set described in 2.1, we solved this problem by weighing the samples inversely to their prevalence. This means that the weights of picks vs. non-picks should maintain the ratio of $\frac{6166}{140}$, which equals to saying that a positive sample warrants 44 times more attention than a negative one.

3.3.1 Baseline

In order to determine a baseline for our models, we determine how well our data is linearly separable. This is done via simple linear regression, where we minimise the distance between the product of the sample data and their coefficients and the respective results:

$$\arg\min_{\beta} \sum_{i} (x_{i,train} * \beta - y_{i,train})^2. \tag{3.1}$$

The raw output, $X_{test} * \hat{\beta}$, will yield a ranking descending in Warren-Buffett-ness, while introducing a threshold will yield a discrete binary prediction.

3.3.2 Support Vector Machine

Our first candidate model is a support vector machine (SVM). A SVM works similarly to a linear model, as it introduces a linear boundary in the data, which might have been transformed prior to fitting the SVM. However, unlike the linear model, the SVM does not construct the boundary in perpendicular to the overall variance-minimising direction, but finds an optimal hyper-plane separating positive from negative samples. It does so by minimising the functional

$$\frac{1}{2} * w^{\top} * w + C * \sum_{i=1}^{n} \xi_i \tag{3.2}$$

wrt. constraints

$$y_i(w * x_i + b) \geq 1 - \xi_i \quad \forall i \in 1 \dots n \tag{3.3}$$

$$\xi_i \geq 0 \quad \forall i \in 1 \dots n \tag{3.4}$$

with

$$w_j = \sum_{i=1}^{n} \alpha_i^j * y_i * x_i \tag{3.5}$$

and C a freely choosable penalty term for the slack variables ξ_i.

For the case that we transform our features by an N- dimensional vector function $\varphi : \varphi : \mathbb{R}^n \to \mathbb{R}^N$ we replace all instances of x_i with $\varphi(x_i)$ This (feature-transformed) minimisation problem has the dual quadratic optimisation problem

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} * (\alpha^{\top} * D * \alpha + \alpha_{max}/C) \tag{3.6}$$

wrt. constraints

$$\alpha^\top * y = 0 \tag{3.7}$$

where

$$D_{ij} = y_i * y_j * K(x_i, x_j) \tag{3.8}$$

and K positive semi-definite. (Cortes, Vapnik, 1995) [10]

While the SVM does not directly give a probability estimate, a good surrogate is the distance to the decision boundary, which we can obtain analogously to the linear model. In our case, we evaluated the performance of three different kernels, the radial basis function (RBF), $K(x_i, x_j) = exp(-\frac{|x_i - x_j|^2}{\gamma^2})$, a sigmoid, $K(x_i, x_j) = tanh(\gamma * x_i^\top * x_j + c)$, and a polynomial, $K(x_i, x_j) = (\gamma * x_i^\top * x_j + c)^d$. We also explored how the choice of class weights would influence the performance, and did a cross-validated, iteratively refined grid search on the penalty weight C as well as γ. Concerning the implementation of this algorithm, we used the one offered by Scikit-learn [31].

3.3.3 Boosted Ensemble

Boosting works by iteratively adding a weak learner to the model which aims to correct the error of the ensemble of previous weak learners. More formally, let m be the m-th stage of an ensemble learning process and $f_m(x)$ a weak learner, e.g. $\beta_m^\top * X_{train}$ for a linear model. Now instead of fitting directly to y_{test}, we fit $f_m(x)$ to $y_{m-1} - f_{m-1}(x)$ yielding y_m. (Hastie, 2017) [18] We are making use of the implementation offered by XGBoost (Chen, Guestrin, 2016) [7], where we investigate the performance of regularised regression trees vs. linear models and varying the number of weak models, the learning rate - adjusting the speed of the gradient descent - as well as the strength of regularisation.

3.3.4 Neural Network

As a penultimate candidate model, we tested the performance of several similar neural networks. A neural network is made up from several layers, which themselves are made up of so-called neurons, a layout which is inspired by early research into how our brain might work. [23] These neurons are (quasi-)differentiable functions which map from an m-dimensional space to a scalar, whereas m is the dimension of the prior layer. The input of each neuron is (generally) comprised by the weighted sum of the output of the previous layer plus a bias term. The first layer of neurons is taking its inputs from the data, while the last layer outputs the corresponding decision.

Training is done by iteratively propagating the input through each successive layer (feed-forward), and then comparing the output of the final layer and the pick decision via a loss function giving an error. In a second step, this error then is differentiated with regard to the weights and biases of the previous layer, which indicates the general direction of updating. These differentiations can then be done layer-wise going from the last to the first layer. (Cross, Harrison, Kennedy, 1995) [11] Different methods to implement or approximate this gradient descent exist, and as our goal was to efficiently use existing implementation of well-known algorithms, we decided to use the very efficient NAdam optimiser (Dozat, 2015) [12] as implemented in keras. [8] Since our prediction target is binary, our final layer should produce a single real scalar ranged from 0 to 1, corresponding to the probability of the sample being picked. We constructed a simple sequential neural network with n layers and a layer size of m that decreases by $\frac{2}{3}$ by each subsequent layer. For the number of layers we investigated a range between $n \in 3, \ldots, 17$ and for the number of neurons on the first layer a range of $m \in 100, \ldots, 7000$, whereas the upper limit was given by machine memory restrictions. As networks perform better under some circumstances when the output of the first or second layer is fourier transformed, allowing for efficient convolution, we also investigated the performance of this design feature. As candidate activation functions we investigated Rectified Linear Unit (ReLU), tanh, and sigmoid. While the network size is an implicit source of regularisation by constraining the set of possible parameters, one of the other ways to avoid over-fitting is dropout regularisation. This means that during training for a given level of $p \in [0, 1)$, any neuron is active with a probability of $1 - p$ or does not activate with probability p. After shortly venturing into the territory of $p \in [0.7, 0.99]$ we quickly found out that such networks are barely trainable under the constraints we were given and searched for appropriate $p \in [0, 0.6]$.

3.3.5 Hybrid Model

Our final candidate was a very simple hybrid model which combined the predictions of all the previously described models. This combination was done via a component-wise multiplication of the individual predictions. In the case of the SVM, we transformed the output with the logistic function in order to make its predictions compatible with the other predictions, which were all probabilistic.

Chapter 4

Results and Discussion

> Life is but a walking shadow, a poor player
> That struts and frets his hour upon the stage
> And then is heard no more.
>
> *W. Shakespeare*

4.1 Performance and Comparison

4.1.1 Baseline

The linear model exceeded our expectations with a balanced accuracy of 83%, a recall of 80% but delivered a lacklustre precision of 12%. (If one were to imagine that beer was ordered but kombucha served instead, the amount of disappointment would have been accurately measured.) The predictions had an area under the curve of 0.79, which indicates a predictive performance somewhat better than chance. Taking a look at the predictions of the linear model (to be found in table A.4 in the appendix), we can discern a cluster of false positive outliers leading the predictions, followed by a cluster of true positives. Additionally, in terms of portfolio performance, it leads to returns that up to 2015 greatly exceed the returns of a simple $\frac{1}{n}$ portfolio (see figure 4.2).

	False	True
Not predicted	1074	162
Predicted	10	20

Table 4.1: Confusion matrix of the linear model predictions

4.1.2 Support Vector Machine

The svm performed as well as could be expected from an algorithm that was long considered state-of-the-art before the advent of deep learning. Its balanced accuracy was 89%, its recall 87%, and the precision still somewhat lacking with 21%. The area under the curve was 0.91.

	False	True
Not predicted	1138	98
Predicted	4	26

Table 4.2: Confusion matrix of the support vector machine predictions

4.1.3 Boosted Ensemble

The boosted model performed the best overall, which is congruent with its widespread use in data science tournaments as well as many industries. It yielded a balanced accuracy of 91%, a recall of 87%, and a passable precision of 33%. The area under the curve was 0.96, which is the highest among all models.

	False	True
Not predicted	1182	54
Predicted	4	26

Table 4.3: Confusion matrix of the boosted ensemble predictions

4.1.4 Neural Network

While our results were constrained by the limited time and computing power available, we were able to confirm that given the simple architecture, the network became barely trainable when it reached a depth of 17 layers. Given the limited amount of data, it also tended to over-fit quickly. Its balanced accuracy was 87%, its recall 77% and its precision 38%. The area under the curve amounted to 0.92.

	False	True
Not predicted	1197	39
Predicted	7	23

Table 4.4: Confusion matrix of the neural network predictions

4.1.5 Hybrid Model

The hybrid model had a balanced accuracy of 92%, a recall of 87% and a precision of 42%, which is the highest among all models. It had an area under the curve of 0.94, which is the second highest among all models.

	False	True
Not predicted	1200	36
Predicted	3	27

Table 4.5: Confusion matrix of the hybrid model predictions

4.1.6 Comparison

Comparing all ROC plots (see figure 4.1), we can see that the boosted ensemble dominates all other models, making it a solid model. Yet when it comes to the top accuracy and precision score, (See table 4.6) our hybrid model proved to be superior - coinciding with the small part of its ROC curve that is above the boosted ensemble model ROC curve. Furthermore, our models exhibit surprisingly high metrics across the board, suggesting that Warren Buffett chooses his investments very systematically, and that there is only a very small margin of achievable performance gain left that bridges the gap between our models and a truly comprehensive synthetic investor.

	Balanced Accuracy	Recall	Precision	AUC
Random Predictor	50%	50%	2%	0.50
Linear Model	83%	80%	12%	0.79
SVM	89%	87%	21%	0.91
Boosted Ensemble	91%	87%	33%	**0.96**
Neural Network	87%	77%	38%	0.92
Hybrid Model	**92%**	**87%**	**42%**	0.94

Table 4.6: Comparison of the model performance

4.2 Discussion

We addressed the question whether it is possible to model the investment behaviour of super-investors like Warren Buffett by using machine learning algorithms and have shown that it is well possible to do so, achieving high levels of accuracy and recall. We discovered that the best practice is to combine output from multiple models, and saw that there is still room to improve these hybrid models. We will now discuss the shortcomings of and potentials uncovered by this thesis. While the all metrics of the discussed models are well above the thresholds that make a model competitive, we

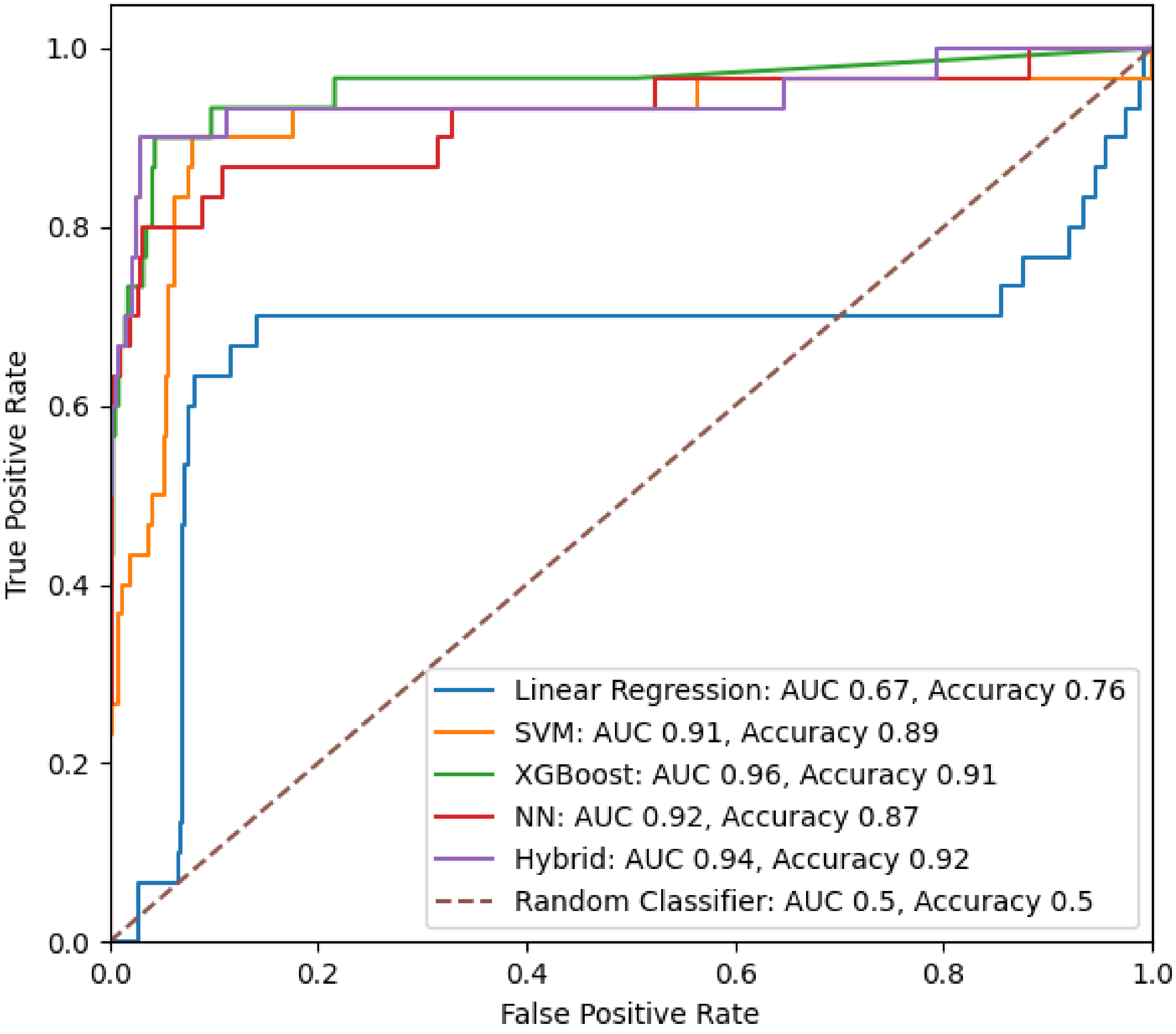

Figure 4.1: The receiver operator characteristics (ROC) of all models combined. The curve is obtained by varying the decision threshold in small enough steps and plotting the respective location of the true positive rate, $\frac{true\ \ positives}{true\ \ positives+false\ \ negatives}$, versus the false positive rate, $\frac{false\ \ positives}{false\ \ positives+true\ \ negatives}$.

still have to be wary of the quality of the predictions, and not directly implement these algorithms without testing their performance as part of a carefully formulated investing strategy. A look at the portfolio performance (4.2) confirms this warning, as the stocks that were both contained in the test set and picked by Warren Buffett showed sub-optimal returns when compared to a simple $\frac{1}{n}$ portfolio, decisively impacting the performance of most of our predictions. Furthermore, since an investor might not face the same restraints and boundary conditions as Warren Buffett or Berkshire Hathaway, there is a high probability we miss many opportunities due to e.g. a company occupying a very small niche where it could generate above-average profits on a limited amount of investment. As a direct counter-example to this cautionary warning we can mention the performance of our linear model, which decidedly out-performed the $\frac{1}{n}$ portfolio. These portfolios are simulated such that each year i we convert all stocks bought in year $i-1$ into the numéraire X_i at their current median price and buy all n_i acceptable stocks by spending $\frac{X_i}{n_i}$ on them, again at their current median price. A stock is acceptable if its value is above zero, and if it has not been held for more than three years. Additionally, in the case of our $\frac{1}{n}$ portfolio, it is only acceptable if it is contained in the test data set, while for the other portfolios, a stock is only acceptable if it was predicted to be picked by Warren Buffet that year by the respective predictor. Since we intended to deliver a rough comparison to the market portfolio, and not absolute performance numbers, we disregarded any inclusion of transaction costs, and in the same vein did not model price effects arising from position building and reduction. As demonstrated in our variable importance section, a drawback of our current models is that they're very sensitive to industrial sector information. This means that predictions about these companies should be taken with a pinch of salt since these models are only guaranteed to hold over the last twenty years. That does neither account for possible innovations that make whole sectors obsolete nor for structural changes in the market. Many possible ways exist to optimise the calibration of our models: The first approach aims to heighten the precision of timing decisions by including the year of divestment in the data set, and by possibly weighing the positive samples by the holding time and amount. Additionally, we could include the years before the investment decision as a positive sample, opening up the possibility of deciding faster than Warren Buffett himself. The second approach would be to take the future relative performance of the company into account, either by trying to directly predict future performance, or by weighing the samples according to their performance. These efforts more concerned with timing would also greatly benefit by integrating signals by thoroughly developed models that concern themselves with questions of bubble and crash formation, such as the log-periodic power law model. [15] What also needs to be investigated is whether more general industry profiles impact the predictive performance greatly, and how they relate to the problem of over-fitting.

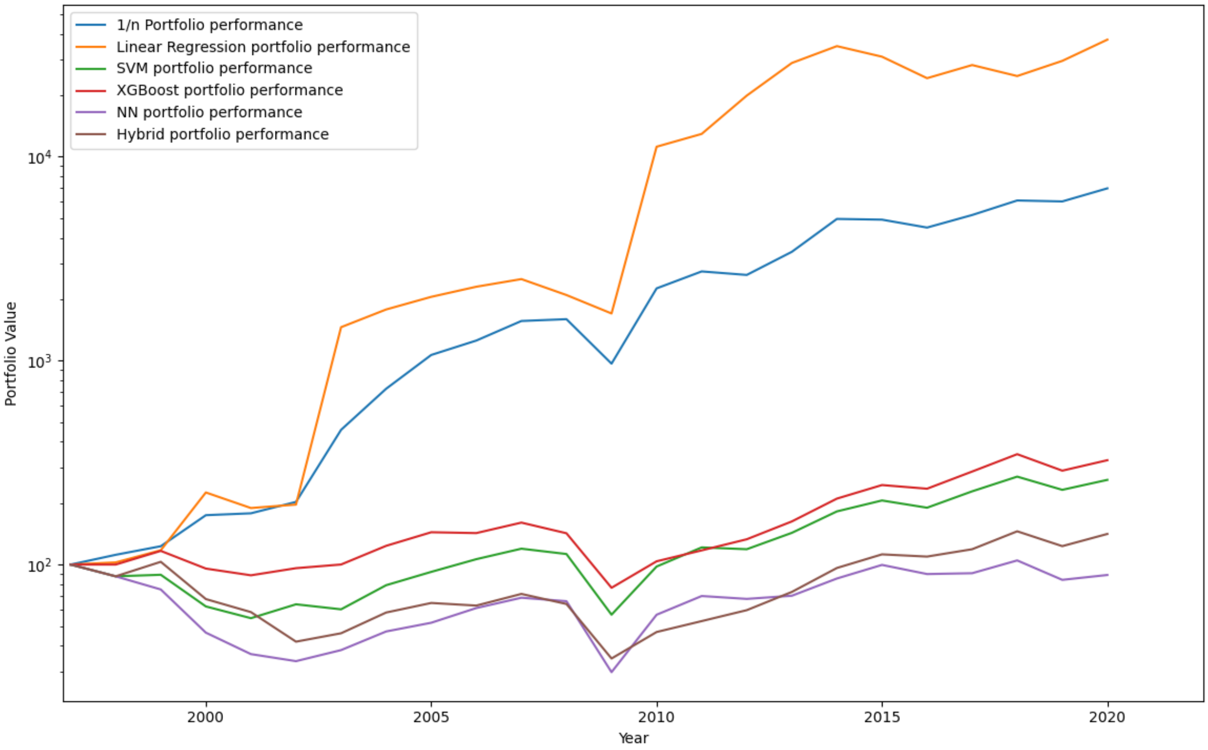

Figure 4.2: The simulation of simple portfolios, where each year i the account is re-balanced to contain all n_i acceptable stocks, held for 3 years. In the case of the $\frac{1}{n}$ portfolio, a stock is acceptable when it is contained in the test data set. For the other portfolios, a stock is acceptable when it is predicted to be picked by Warren Buffett that year.

Lastly, we have left out a whole swathe of deep learning techniques that might prove to perform even better and can handle time series of variable length. This is not such a relevant point since we discovered that our models perform exceedingly well in terms of predictive power. In terms of evaluation, we only examined the performance on the binary classification task, without further asking if these classifications can be weighed in terms of long term profitability, and whether the quality of the predictions could be improved according to these terms. Ultimately, the scope of our research can be extended to all known super-investors. Doing so would make the stock selection criteria even less noisy and less dependent on the unique informational environment and the investment constraints of a single super-investor. Setting all these concerns aside for a moment, we can also think about formulating the problem in wholly different terms: If we assume that we're only really certain about the companies that Warren Buffett did pick, throw in additional examples of very bad investment ideas, and, in a first step, set the decision about all other companies as ’undecided’, we could come up with a semi-supervised learner that gradually discovers what makes great companies valuable investment targets. To conclude, our research shows that using machine learning algorithms on the investment decisions of super-investors is an exciting venue for future research.

Appendix A

Appendix

A.1 Variable Statistics

A.1.1 Average Differences

Variable	*D*ifference
Revenue	2.123
Total Revenue	2.119
Gross Profit	2.088
Selling/General/Administrative Expense, Total	2.021
Total Assets, Reported	1.999
EBITDA	1.921
Operating Expenses	1.910
EBIT	1.847
Property, Plant And Equipment, Total - Gross	1.843
Property/Plant/Equipment, Total - Net	1.818
Total Current Liabilities	1.799
Total Equity	1.789
Interest Expense	1.731
Cost of Revenue, Total	1.670
Total Liabilities	1.657
Operating Income	1.656
Net Income Incl Extra Before Distributions	1.637
Market Value for Company	1.579
Capex, Discrete	1.571
Total Inventory	1.554
Total Long Term Debt	1.540
Long Term Debt	1.529
Net Income Before Extraordinary Items	1.431
Depreciation And Amortization	1.426
Retained Earnings (Accumulated Deficit)	1.418

Variable	*D*ifference
$flo	1.383
Total Current Assets	1.373
Total Debt	1.362
Accounts Receivable - Trade, Net	1.309
Accounts Payable	1.051
Accounts Receivable - Trade, Gross	1.023
Cash and Short Term Investments	0.958
Research And Development	0.920
Common Stock, Total	0.865
Dividends Payable	0.857
Accounts Receivable (CF)	−0.735
ISSUES	0.683
Short Term Investments	0.556
Notes Payable/Short Term Debt	0.533
Cash	0.504
Non-Current Marketable Securities, Supplemental	0.448
ROA	−0.432
CFO	−0.405
DROA	0.398
VARROA	0.371
CAPINT	0.368
ACCRUAL	0.367
X3	−0.337
X5	0.336
ZSCORE	0.329
DLEVER	0.323
TIE	0.317
OENEG	−0.294
Size	0.247
FSCORE	0.244
AQI	−0.235
invtry turnover	0.229
acidity	−0.165
fixed_asset_turnover	0.164
SGAI	0.158
CHIN	0.148
D_ROA	0.147
QMJ_PROFIT	0.130
DSRI	0.121
GSCORE	0.117
MSCORE	0.092
Advertising Expense	0.091
INTWO	−0.073

Variable	*D*ifference
QMJ_GROWTH	0.041
RDINT	−0.037
QMJ	0.035
ADINT	−0.029
QMJ_SAFETY	−0.026
DEPI	0.022
TLTA	−0.020
WCTA	0.020
X4	−0.016
DTURN	0.016
SGR	−0.016
FUTL	−0.014
O-score	0.012
Basic EPS Including Extraordinary Items	0.011
SGI	−0.011
DLIQUID	0.010
LGVI	−0.010
CLCA	−0.010
GMI	−0.009
VARSGR	−0.009
TATA	0.009
X1	−0.008
X2	0.008
DGMAR	−0.008
Gross Margin	0.008
DMARGIN	−0.007
DROE	−0.007
current_ratio	−0.005
NITA	0.005
DGPOA	−0.004
LEVER	−0.004
D$FLO	0.004
ROE	−0.003
Total Interest Expenses	0.000
Principal Payments from Securities	0.000

Table A.1: Full differences between class averages

A.1.2 Variable Importance by XGBoost

Variable	Gain
Gross Profit	20.212
Net Income Incl Extra Before Distributions	14.038

Variable	Gain
EBITDA	7.461
Accounts Receivable - Trade, Net	7.386
Total Current Liabilities	6.481
Property/Plant/Equipment, Total - Net	6.276
Selling/General/Administrative Expense, Total	6.240
Total Inventory	6.070
Accounts Receivable (CF)	6.018
Long Term Debt	3.705
Property, Plant And Equipment, Total - Gross	3.557
Total Assets, Reported	3.495
Interest Expense	3.409
Depreciation And Amortization	3.401
Research And Development	3.390
Advertising Expense	3.389
Cash and Short Term Investments	3.339
Cost of Revenue, Total	2.918
Total Equity	2.772
Revenue	2.578
Short Term Investments	2.295
Capex, Discrete	2.238
Total Current Assets	2.126
Dividends Payable	2.039
Accounts Payable	2.014
Total Long Term Debt	2.002
Total Revenue	1.777
Market Value for Company	1.674
EBIT	1.657
ISSUES	1.500
DMARGIN	1.414
Accounts Receivable - Trade, Gross	1.390
Total Debt	1.240
Total Liabilities	1.240
DLEVER	1.153
Operating Income	1.106
Retained Earnings (Accumulated Deficit)	1.023
Basic EPS Including Extraordinary Items	0.993
$flo	0.895
Common Stock, Total	0.893
GSCORE	0.739
Notes Payable/Short Term Debt	0.729
Net Income Before Extraordinary Items	0.690
FSCORE	0.653
X2	0.571

Variable	Gain
Operating Expenses	0.506
CHIN	0.394
X1	0.369
WCTA	0.361
Size	0.223
DROA	0.213
GMI	0.090
VARROA	0.082
TLTA	0.020
SGAI	0.004
All other variables	0.000

Table A.2: Variable Importance as inferred by XGBoost

A.1.3 Implied Variable Importance

Variable Group	Accuracy
Industries	0.811 8
Scaled	0.843 2
Revenue	0.903 3
Operating Expenses	0.907 7
Gross Profit	0.907 7
Net Income Incl Extra Before Distributions	0.907 8
Retained Earnings (Accumulated Deficit)	0.907 9
Operating Income	0.907 9
EBIT	0.907 9
Selling/General/Administrative Expense, Total	0.907 9
Net Income Before Extraordinary Items	0.908 0
Research And Development	0.908 0
TIE	0.908 1
O-score	0.908 1
EBITDA	0.908 1
Cost of Revenue, Total	0.908 1
Market Value for Company	0.908 1
Accounts Payable	0.908 2
Total Current Liabilities	0.908 2
Total Inventory	0.908 3
Advertising Expense	0.908 3
MSCORE	0.908 3
Dividends Payable	0.908 3
Total Assets, Reported	0.908 3
Cash	0.908 3
Accounts Receivable	0.908 4

Variable Group	Accuracy
Gross Margin	0.908 4
LEVER	0.908 4
current_ratio	0.908 4
debt_to_assets	0.908 4
ROE	0.908 4
Principal Payments from Securities	0.908 4
QMJ	0.908 4
Notes Payable/Short Term Debt	0.908 4
Baseline	**0.9084**
ZSCORE	0.908 4
Total Liabilities	0.908 4
Accounts Receivable - Trade, Net	0.908 4
Short Term Investments	0.908 4
Capex, Discrete	0.908 4
Basic EPS Including Extraordinary Items	0.908 5
Non-Current Marketable Securities	0.908 5
acidity	0.908 5
GSCORE	0.908 5
Total Current Assets	0.908 5
Total Debt	0.908 5
fixed_asset_turnover	0.908 5
Total Equity	0.908 6
Interest Expenses	0.908 6
Common Stock, Total	0.908 6
Depreciation And Amortization	0.908 6
invtry turnover	0.908 7
Long Term Debt	0.909 0
Property, Plant And Equipment	0.909 6
FSCORE	0.910 0

Table A.3: Implied Variable Importance

A.2 Selected Companies

A.2.1 Linear Model Predictions

Prediction	Company Name	Picked	Year
1 349 584 572 284.99	Orion Engineered Carbons SA	0	2017
622 542 663 680.41	Orion Group Holdings Inc	0	2016
622 542 663 680.41	Orion Group Holdings Inc	0	2009
565 676 183 497.35	Darling Ingredients Inc	0	2006
565 676 183 497.34	Darling Ingredients Inc	0	2001
492 282 492 654.40	ParkerVision Inc	0	1997

Prediction	Company Name	Picked	Year
431 972 651 595.09	Liquidmetal Technologies Inc	0	2003
407 944 639 139.33	Myers Industries Inc	0	2005
407 925 828 517.27	Simpson Manufacturing Co Inc	0	2005
373 315 986 120.83	Molson Coors Beverage Co	0	2001
368 850 611 190.67	Ecosciences Inc	0	2017
352 659 021 375.48	Enviva Partners LP	0	2018
327 301 560 775.23	Gaming and Leisure Properties Inc	0	2018
324 505 381 841.86	Centric Brands Inc	0	2000
315 735 707 227.08	Xenetic Biosciences Inc	0	2018
295 024 280 257.15	TORtec Group Corp	0	2017
287 783 593 043.68	Childrens Place Inc	0	2012
279 756 004 222.00	Jewett-Cameron Trading Company Ltd	0	2001
253 644 526 016.78	Forward Industries Inc	0	2011
253 644 526 016.77	Forward Industries Inc	0	2017
251 610 009 214.22	International Flavors & Fragrances Inc	0	2000
246 013 443 974.36	Cemtrex Inc	0	2015
245 553 594 511.60	BioLargo Inc	0	2000
245 553 594 511.36	Commodore Applied Technologies Inc	0	1997
236 750 422 027.27	Tofutti Brands Inc	0	2001
221 542 958 249.54	Timken Co	0	1997
218 903 614 259.72	CPI Card Group Inc	0	2017
218 106 586 572.85	Lincoln Electric Holdings Inc	0	1998
191 759 105 894.52	S&W Seed Co	0	2016
186 879 162 712.88	Yulong Eco-Materials Ltd	0	2016
174 521 816 336.74	Costar Technologies Inc	0	2001
162 497 971 321.98	Lentuo International Inc	0	2012
161 752 165 371.56	Hubbell Inc	0	2000
157 130 552 561.92	American International Industries Inc	0	2001
152 491 167 990.95	Sonic Automotive Inc	0	2018
152 491 167 990.90	Sonic Automotive Inc	0	1998
141 693 445 732.73	KROMI Logistik AG	1	2008
131 522 832 255.00	Ethan Allen Interiors Inc	0	2003
129 044 643 441.13	Ormet Corp	0	2011
126 337 763 380.29	Mansfelder Metals Ltd	0	2001
124 432 364 163.88	Blue Apron Holdings, Inc.	0	2018
118 814 234 017.29	Megatech Corp	0	2003
116 316 887 563.34	Aptargroup Inc	0	1997
106 601 384 539.47	Gratitude Health Inc	0	2017
105 950 256 702.33	Alkaline Water Company Inc	0	2018
103 611 154 395.17	Hyster-Yale Materials Handling Inc	0	2017
102 977 606 112.29	Cyclopss Corp	0	2001
98 277 388 420.10	Simply Good Foods Co	0	2017
94 340 980 435.87	EquiFin Inc	0	1998

Prediction	Company Name	Picked	Year
89 654 580 573.43	New Age Beverages Corp	0	2017
77 579 983 159.78	Trane Technologies PLC	1	2005
76 402 148 465.33	Tile Shop Holdings Inc	0	2017
75 815 957 357.05	Carmax Inc	1	2006
68 327 666 798.80	Avitar Inc	0	1997
66 458 216 823.09	United Parcel Service Inc	1	2005
52 777 370 662.19	Indoor Harvest Corp	0	2018
52 652 531 951.27	Kohls Corp	0	1999
52 652 531 951.24	Kohls Corp	0	2004
52 004 354 084.69	Ageagle Aerial Systems Inc	0	2018
50 936 320 445.89	Carrier Alliance Holdings Inc	0	2009
47 607 462 632.74	Lighting Science Group Corp	0	2000
46 302 529 329.07	Carpenter Technology Corp	0	1999
43 717 205 166.16	Cummins Inc	0	2007
43 717 205 166.09	Cummins Inc	0	1996
43 717 205 162.99	Thermon Group Holdings Inc	0	2016
32 405 395 861.36	Huntwicke Capital Group Inc	0	2016
28 060 507 408.88	Titan Machinery Inc	0	2009
24 610 461 636.08	Hebron Technology Co Ltd	0	2015
20 385 891 605.43	BGI Inc	0	2003
19 358 925 120.83	Ryder System Inc	0	2017
19 358 925 120.65	Ryder System Inc	0	1996
15 800 602 595.68	Freeport-McMoRan Inc	0	2004
15 763 965 552.37	Sunvault Energy Inc	0	2015
909 532 968.30	Lamb Weston Holdings Inc	0	2018
91 910 343.94	California Style Palms Inc	0	2006
46 182 734.22	Dougherty's Pharmacy Inc	0	2007
42 896 046.06	Newpark Resources Inc	0	2003
37 193 071.78	MPM Technologies Inc	0	1997
132 468.56	SEACOR Holdings Inc	0	2005
3.60	Emerald Holding Inc	0	2018
3.41	Graphene & Solar Technologies Ltd	0	2015
2.15	Meritor Inc	0	2007
2.10	Dell Technologies Inc	0	2018
1.99	Tenneco Inc	0	1996
1.93	Vertex Energy Inc	0	2004
1.67	Kroger Co	1	2018
1.65	Viavi Solutions Inc	0	2003
1.59	Green Dot Corp	0	2018
1.56	Ford Motor Co	0	2017
1.43	Hewlett Packard Enterprise Co	0	2016
1.20	Southwest Airlines Co	1	2015
1.05	Halliburton Co	0	2018

Prediction	Company Name	Picked	Year
0.96	Intel Corp	1	2010
0.95	Visteon Corp	0	2011
0.92	Express Scripts Holding Co	1	2013
0.92	Dow Jones & Company Inc	1	2006
0.91	Precision Castparts Corp	1	2011
0.89	Sealed Air Corp	1	1999
0.88	Servicemaster Company LLC	1	2003
0.88	USG Corp	1	1999
0.88	Justin Industries Inc	1	1999
0.88	Benjamin Moore and Co	1	1999
0.88	Xtra Corp	1	2000
0.88	KEMET Corp	0	2004
0.87	Stemline Therapeutics Inc	0	2018
0.86	Great Lakes Chemical Corp	1	1998
0.79	Wiltel Communications Group Inc	1	2002
0.75	Visteon Corp	0	2004
0.73	Crawford & Co	0	2008
0.70	HCA Inc	1	2002
0.69	GlaxoSmithKline PLC	1	2006
0.68	Marathon Oil Corp	0	2009
0.67	EVO Transportation & Energy Services Inc	0	2018
0.67	Greenbox Pos	0	2017
0.64	Alibaba Group Holding Ltd	0	2018
0.63	Innospec Inc	0	2008
0.62	Innospec Inc	0	2012
0.57	Centurylink Inc	0	2013
0.47	United Rentals Inc	0	2015
0.46	Alliance Data Systems Corp	0	2004
0.46	NXP Semiconductors NV	0	2018
0.41	United Airlines Holdings Inc	1	2015
0.37	Bunge Ltd	0	2008
0.35	eBay Inc	0	2018
0.35	Drive Shack Inc	0	2012
0.34	NortonLifeLock Inc	0	2011
0.33	Forward Air Corp	0	2005
0.33	Drive Shack Inc	0	2008
0.33	Wall Street Media Co Inc	0	2016
0.31	Booking Holdings Inc	0	2015
0.29	Ashland Global Holdings Inc.	0	2010
0.27	Chesapeake Energy Corp	0	2006
0.26	Circle Entertainment Inc	0	2013
0.25	Flex Ltd	0	2011
0.25	Black Ridge Oil & Gas Inc	0	2017

Prediction	Company Name	Picked	Year
0.24	Sportsmans Warehouse Holdings Inc	0	2017
0.24	Endo International PLC	0	2018
0.24	No Fire Technologies Inc	0	2014
0.23	Summit Environmental Corporation Inc	0	2004
0.22	Lee Pharmaceuticals	0	1999
0.22	Accenture PLC	0	2018
0.22	Tech Central Inc	0	2018
0.22	Planet Fitness Inc	0	2015
0.22	AutoNation Inc	0	2006
0.22	VirnetX Holding Corp	0	2000
0.21	RC-1 Inc	0	2018
0.20	GDS Holdings Ltd	0	2018
0.20	Ford Motor Co	0	2004
0.19	Strat Petroleum Ltd	0	1996
0.19	Redwood Green Corp	0	2015
0.19	Charles Schwab Corp	0	1997
0.18	Web Blockchain Media Inc	0	2001
0.18	W. R. Grace & Co	0	2002
0.17	R1 RCM Inc	0	2015
0.17	AMC Entertainment Holdings Inc	0	2018
0.17	W. R. Grace & Co	0	2010
0.16	Willdan Group Inc	0	2007
0.16	CSW Industrials Inc	0	2017
0.16	Rent-A-Center Inc	0	2015
0.15	Conagra Brands Inc	0	2002
0.15	TriNet Group Inc	0	2018
0.15	Vistra Energy Corp	0	2018
0.15	Claxson Interactive Group Inc	0	2004
0.15	Franklin Resources Inc	0	2010
0.15	Yum! Brands Inc	1	1999
0.14	Gilead Sciences Inc	0	2008
0.14	eBay Inc	0	2007
0.14	Baidu Inc	0	2016
0.14	Tuniu Corp	0	2017
0.14	Ultra Petroleum Corp	0	2018
0.14	S&P Global Inc	0	2006
0.13	Avaya Holdings Corp	0	2015
0.13	Merck & Co Inc	0	1999
0.13	Eastman Kodak Co	0	2002
0.13	Bridgeway National Corp.	0	2018
0.13	Concorde Gaming Corp	0	2000
0.13	Sunoco LP	0	2014
0.13	Raymond James Financial Inc	0	1997

Prediction	Company Name	Picked	Year
0.13	Oblong Inc	0	2003
0.12	Piedmont Mining Co Inc	0	1996
0.12	Comstock Mining Inc	0	2016
0.11	CIM Commercial Trust Corp	0	2017
0.11	Red Cat Holdings Inc	0	2000

Table A.4: Companies picked by the linear model

A.2.2 SVM Predictions

Prediction	Company Name	Picked	Year
6.01	Express Scripts Holding Co	1	2013
5.97	Comcast Corp	1	2003
5.68	HCA Inc	1	2002
5.38	Mondelez International Inc	1	2006
5.27	Merck & Co Inc	0	1999
5.18	Walt Disney Co	1	1999
4.99	HCA Healthcare Inc	0	2007
4.98	United Parcel Service Inc	1	2005
4.95	GlaxoSmithKline PLC	1	2006
4.95	Eli Lilly and Co	0	2006
4.95	Halliburton Co	0	2018
4.83	United Airlines Holdings Inc	1	2015
4.82	Hewlett Packard Enterprise Co	0	2016
4.80	Linde PLC	0	2018
4.65	Accenture PLC	0	2018
4.61	Marathon Oil Corp	0	2009
4.59	Centurylink Inc	0	2013
4.52	Marathon Petroleum Corp	0	2018
4.47	Kroger Co	1	2018
4.47	eBay Inc	0	2018
4.42	Intel Corp	1	2010
4.06	Southwest Airlines Co	1	2015
3.98	Dell Technologies Inc	0	2018
3.88	Bunge Ltd	0	2008
3.84	Conagra Brands Inc	0	2002
3.81	FedEx Corp	0	2002
3.72	Eastman Kodak Co	0	2002
3.59	Oracle Corp	1	2017
3.55	Abbott Laboratories	0	1997
3.41	PG&E Corp	0	1998
3.24	Baidu Inc	0	2016
3.18	Marathon Oil Corp	0	1998
3.08	Aon PLC	0	2011

Prediction	Company Name	Picked	Year
3.07	NXP Semiconductors NV	0	2018
3.01	Flex Ltd	0	2011
3.01	Aptiv PLC	0	2016
2.84	Conagra Brands Inc	0	2007
2.67	Precision Castparts Corp	1	2011
2.61	iHeartMedia Inc	0	2015
2.46	Vistra Energy Corp	0	2018
2.40	Westrock Co	0	2016
2.34	PPL Corp	0	2008
2.31	Booking Holdings Inc	0	2015
2.30	Mylan NV	0	2014
2.30	Tenneco Inc	0	1996
2.29	AutoNation Inc	0	2006
2.20	US Foods Holding Corp	0	2016
2.19	JOYY Inc	0	2018
2.18	Qurate Retail Inc	0	2016
2.17	Masco Corp	0	2009
2.09	Ameren Corp	0	2010
2.06	News Corp	0	2016
1.93	Visteon Corp	0	2004
1.92	Ryder System Inc	0	2017
1.89	United Rentals Inc	0	2015
1.83	Ford Motor Co	0	2004
1.80	Chesapeake Energy Corp	0	2006
1.69	Halliburton Co	0	1999
1.64	eBay Inc	0	2007
1.64	Franklin Resources Inc	0	2010
1.62	Trane Technologies PLC	1	2005
1.58	Ford Motor Co	0	2017
1.56	Kohls Corp	0	2004
1.55	Autoliv Inc	0	2014
1.50	Enable Midstream Partners LP	0	2016
1.46	Yum! Brands Inc	1	1999
1.35	Discovery Inc	0	2012
1.33	Coty Inc	0	2018
1.32	Cummins Inc	0	2007
1.31	S&P Global Inc	0	2006
1.28	NetApp Inc	0	2015
1.27	NortonLifeLock Inc	0	2011
1.26	Taylor Morrison Home Corp	0	2018
1.16	Ameren Corp	0	2006
1.14	WPX Energy Inc	0	2014
1.14	California Resources Corp	0	2017

Prediction	Company Name	Picked	Year
1.09	Reliance Steel & Aluminum Co	0	2014
1.08	Huntsman Corp	0	2006
1.06	Ashland Global Holdings Inc.	0	2010
1.04	Brixmor Property Group Inc	0	2017
1.01	Centurylink Inc	0	2009
1.00	Servicemaster Company LLC	1	2003
0.99	USG Corp	1	1999
0.97	Avnet Inc	0	2007
0.91	Sirius XM Holdings Inc	1	2015
0.91	Interactive Brokers Group Inc	0	2013
0.89	Wiltel Communications Group Inc	1	2002
0.88	Park Hotels & Resorts Inc	0	2003
0.87	Office Depot Inc	1	2000
0.85	Expedia Group Inc	0	2015
0.84	Sealed Air Corp	1	1999
0.83	C.H. Robinson Worldwide Inc	0	2016
0.81	Great Lakes Chemical Corp	1	1998
0.80	Gilead Sciences Inc	0	2008
0.73	Crown Holdings Inc	0	2006
0.71	Dillard's Inc	0	1997
0.70	Genuine Parts Co	0	1999
0.67	Apartment Investment and Management Co	0	2009
0.66	Fiserv Inc	1	2009
0.65	Dow Jones & Company Inc	1	2006
0.65	Eastman Chemical Co	0	1997
0.60	Western Midstream Partners LP	0	2017
0.58	Xtra Corp	1	2000
0.56	Ryder System Inc	0	1996
0.56	CME Group Inc	0	2008
0.51	RR Donnelley & Sons Co	0	1997
0.50	Meritor Inc	0	2007
0.49	Avaya Holdings Corp	0	2015
0.49	QEP Resources Inc	0	2018
0.45	Rent-A-Center Inc	0	2015
0.45	Eversource Energy	0	2007
0.42	Illumina Inc	0	2018
0.41	E. W. Scripps Co	0	2006
0.40	Visteon Corp	0	2011
0.38	Sonic Automotive Inc	0	2018
0.37	KEMET Corp	0	2004
0.36	Tyson Foods Inc	0	1998
0.36	Endo International PLC	0	2018
0.36	AMC Entertainment Holdings Inc	0	2018

Prediction	Company Name	Picked	Year
0.35	CBL & Associates Properties Inc	0	2008
0.30	Benjamin Moore and Co	1	1999
0.29	Alibaba Group Holding Ltd	0	2018
0.25	Range Resources Corp	0	2012

Table A.5: Companies picked by the SVM

A.2.3 XGBoost Predictions

Prediction	Company Name	Picked	Year
0.98	Sealed Air Corp	1	1999
0.98	Servicemaster Company LLC	1	2003
0.98	Dow Jones & Company Inc	1	2006
0.98	Express Scripts Holding Co	1	2013
0.98	HCA Inc	1	2002
0.97	Great Lakes Chemical Corp	1	1998
0.97	USG Corp	1	1999
0.97	Xtra Corp	1	2000
0.97	KEMET Corp	0	2004
0.97	United Airlines Holdings Inc	1	2015
0.96	Comcast Corp	1	2003
0.96	Marathon Petroleum Corp	0	2018
0.94	Wiltel Communications Group Inc	1	2002
0.94	Mondelez International Inc	1	2006
0.90	Ford Motor Co	0	2017
0.90	Kroger Co	1	2018
0.90	Centurylink Inc	0	2013
0.89	Justin Industries Inc	1	1999
0.89	Precision Castparts Corp	1	2011
0.85	Southwest Airlines Co	1	2015
0.82	Benjamin Moore and Co	1	1999
0.80	Merck & Co Inc	0	1999
0.78	Eli Lilly and Co	0	2006
0.78	Marathon Oil Corp	0	2009
0.78	United Parcel Service Inc	1	2005
0.76	Alibaba Group Holding Ltd	0	2018
0.76	United Rentals Inc	0	2015
0.74	NXP Semiconductors NV	0	2018
0.71	Hewlett Packard Enterprise Co	0	2016
0.70	GlaxoSmithKline PLC	1	2006
0.64	Linde PLC	0	2018
0.61	Abbott Laboratories	0	1997
0.57	Walt Disney Co	1	1999
0.55	S&P Global Inc	0	2006

Pr*ed*iction	Company Name	Picked	Year
0.50	eBay Inc	0	2018
0.48	Lamb Weston Holdings Inc	0	2018
0.47	Mylan NV	0	2014
0.46	Stemline Therapeutics Inc	0	2018
0.42	HCA Healthcare Inc	0	2007
0.42	Yum! Brands Inc	1	1999
0.41	PPL Corp	0	2008
0.40	Aon PLC	0	2011
0.39	Halliburton Co	0	2018
0.39	Fiserv Inc	1	2009
0.38	Bunge Ltd	0	2008
0.38	Kohls Corp	0	2004
0.38	FedEx Corp	0	2002
0.36	Halliburton Co	0	1999
0.36	Ford Motor Co	0	2004
0.36	Chesapeake Energy Corp	0	2006
0.31	Conagra Brands Inc	0	2002
0.30	eBay Inc	0	2007
0.29	AutoNation Inc	0	2006
0.29	Franklin Resources Inc	0	2010
0.29	Baidu Inc	0	2016
0.28	Westrock Co	0	2016
0.27	Cummins Inc	0	2007
0.27	Qurate Retail Inc	0	2016
0.26	Accenture PLC	0	2018
0.24	Conagra Brands Inc	0	2007
0.23	Autozone Inc	0	1999
0.21	Autoliv Inc	0	2014
0.21	Trane Technologies PLC	1	2005
0.21	WPX Energy Inc	0	2014
0.21	Aptiv PLC	0	2016
0.21	Genuine Parts Co	0	1999
0.21	JOYY Inc	0	2018
0.20	Intel Corp	1	2010
0.20	Marathon Oil Corp	0	1998
0.20	Park Hotels & Resorts Inc	0	2003
0.19	Brown-Forman Corp	0	2010
0.19	Cummins Inc	0	1996
0.18	Dell Technologies Inc	0	2018
0.17	Vistra Energy Corp	0	2018
0.17	Cooper Tire & Rubber Co	0	1997
0.17	Oracle Corp	1	2017
0.17	Sirius XM Holdings Inc	1	2015

Prediction	Company Name	Picked	Year
0.16	Eastman Kodak Co	0	2002
0.16	Ameren Corp	0	2010

Table A.6: Companies picked by XGBoost

A.2.4 Neural Network Predictions

Prediction	Company Name	Picked	Year
1.000	Comcast Corp	1	2003
1.000	Southwest Airlines Co	1	2015
1.000	Express Scripts Holding Co	1	2013
0.999	United Airlines Holdings Inc	1	2015
0.999	Marathon Petroleum Corp	0	2018
0.999	Green Dot Corp	0	2018
0.997	HCA Inc	1	2002
0.995	Mondelez International Inc	1	2006
0.994	Precision Castparts Corp	1	2011
0.991	USG Corp	1	1999
0.990	Bunge Ltd	0	2008
0.988	Sealed Air Corp	1	1999
0.986	Dow Jones & Company Inc	1	2006
0.981	Wiltel Communications Group Inc	1	2002
0.980	Servicemaster Company LLC	1	2003
0.980	Xtra Corp	1	2000
0.974	Great Lakes Chemical Corp	1	1998
0.971	Benjamin Moore and Co	1	1999
0.965	Kroger Co	1	2018
0.948	Justin Industries Inc	1	1999
0.931	Halliburton Co	0	2018
0.911	Walt Disney Co	1	1999
0.906	KEMET Corp	0	2004
0.839	Intel Corp	1	2010
0.832	Gilead Sciences Inc	0	2008
0.818	Aptiv PLC	0	2016
0.786	Linde PLC	0	2018
0.784	NXP Semiconductors NV	0	2018
0.764	Ryder System Inc	0	2017
0.700	Harte Hanks Inc	0	1998
0.686	Tile Shop Holdings Inc	0	2017
0.573	US Foods Holding Corp	0	2016
0.483	Oracle Corp	1	2017
0.364	Tenneco Inc	0	1996
0.354	United Rentals Inc	0	2015
0.187	Marathon Oil Corp	0	2009

Prediction	Company Name	Picked	Year
0.155	Mylan NV	0	2014
0.065	Stemline Therapeutics Inc	0	2018
0.025	Meritor Inc	0	2007
0.020	Westrock Co	0	2016
0.016	AMC Entertainment Holdings Inc	0	2018
0.016	Merck & Co Inc	0	1999
0.012	Coty Inc	0	2018
0.011	Rent-A-Center Inc	0	2015
0.008	Sirius XM Holdings Inc	1	2015
0.007	Hibbett Sports Inc	0	2009
0.007	Chesapeake Energy Corp	0	2006
0.007	Hibbett Sports Inc	0	2005
0.006	Freeport-McMoRan Inc	0	2004
0.006	Illumina Inc	0	2018
0.006	Tofutti Brands Inc	0	2001
0.006	Ferroglobe PLC	0	2018
0.005	PPL Corp	0	2008
0.005	E. W. Scripps Co	0	2006
0.004	Centurylink Inc	0	2013
0.004	Office Depot Inc	1	2000
0.004	S&P Global Inc	0	2006
0.004	Eversource Energy	0	2007
0.004	United Parcel Service Inc	1	2005
0.004	Sequential Brands Group Inc	0	2018
0.003	Mesa Air Group Inc	0	2008

Table A.7: Companies picked by the Neural Network

A.2.5 Hybrid Model Predictions

Prediction	Company Name	Picked	Year
0.973 3	Express Scripts Holding Co	1	2013
0.969 1	HCA Inc	1	2002
0.959 6	Comcast Corp	1	2003
0.958 3	United Airlines Holdings Inc	1	2015
0.946 7	Marathon Petroleum Corp	0	2018
0.928 1	Mondelez International Inc	1	2006
0.860 0	Kroger Co	1	2018
0.834 0	Southwest Airlines Co	1	2015
0.824 9	Precision Castparts Corp	1	2011
0.703 2	USG Corp	1	1999
0.701 3	Servicemaster Company LLC	1	2003
0.675 2	Sealed Air Corp	1	1999

Pr*ed*iction	Company Name	Picked	Year
0.655 9	Great Lakes Chemical Corp	1	1998
0.655 5	Wiltel Communications Group Inc	1	2002
0.633 7	Dow Jones & Company Inc	1	2006
0.610 4	Xtra Corp	1	2000
0.556 2	NXP Semiconductors NV	0	2018
0.519 6	KEMET Corp	0	2004
0.517 4	Walt Disney Co	1	1999
0.499 1	Linde PLC	0	2018
0.465 6	Justin Industries Inc	1	1999
0.459 4	Benjamin Moore and Co	1	1999
0.367 6	Bunge Ltd	0	2008
0.359 0	Halliburton Co	0	2018
0.234 0	United Rentals Inc	0	2015
0.169 2	Intel Corp	1	2010
0.163 5	Aptiv PLC	0	2016
0.143 5	Marathon Oil Corp	0	2009
0.094 4	Ryder System Inc	0	2017
0.080 6	Oracle Corp	1	2017
0.069 4	Gilead Sciences Inc	0	2008
0.066 8	Mylan NV	0	2014
0.047 5	Tenneco Inc	0	1996
0.032 4	US Foods Holding Corp	0	2016
0.015 2	Green Dot Corp	0	2018
0.014 7	Stemline Therapeutics Inc	0	2018
0.012 6	Harte Hanks Inc	0	1998
0.012 4	Merck & Co Inc	0	1999
0.005 1	Westrock Co	0	2016
0.004 1	Tile Shop Holdings Inc	0	2017
0.003 9	Centurylink Inc	0	2013
0.002 8	United Parcel Service Inc	1	2005
0.002 0	Chesapeake Energy Corp	0	2006
0.001 9	Eli Lilly and Co	0	2006
0.001 9	PPL Corp	0	2008
0.001 6	S&P Global Inc	0	2006
0.001 5	eBay Inc	0	2018
0.001 3	GlaxoSmithKline PLC	1	2006
0.001 1	Aon PLC	0	2011
0.000 9	Sirius XM Holdings Inc	1	2015
0.000 9	Coty Inc	0	2018
0.000 7	AutoNation Inc	0	2006
0.000 7	Kohls Corp	0	2004
0.000 7	Hewlett Packard Enterprise Co	0	2016
0.000 6	Trane Technologies PLC	1	2005

Prediction	Company Name	Picked	Year
0.000 6	AMC Entertainment Holdings Inc	0	2018
0.000 6	Halliburton Co	0	1999
0.000 5	Fiserv Inc	1	2009
0.000 5	Abbott Laboratories	0	1997
0.000 5	HCA Healthcare Inc	0	2007
0.000 5	Illumina Inc	0	2018
0.000 5	Yum! Brands Inc	1	1999

Table A.8: Companies picked by the Hybrid Model

A.3 Code

The entire code can be found on https://gitlab.com/Spatiality/PORTFOLIO2020/-/tree/master/03_thesis, the author's GitLab page.

Bibliography

[1] Thomson reuters eikon.

[2] *Rules and regulations under the securities exchange act of 1934*. Securities and Exchange Commission, Washington, D.C, 1935.

[3] Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4):589–609, 1968.

[4] Clifford Asness, Andrea Frazzini, and Lasse Pedersen. Quality minus junk. *Review of Accounting Studies*, 24(1):34–112, 2019.

[5] Messod D. Beneish. The detection of earnings manipulation. *Financial Analysts Journal*, 55(5):24–36, 1999.

[6] Warren Buffett. *Warren Buffett on business : principles from the sage of Omaha*. John Wiley & Sons, Hoboken NJ, 2010.

[7] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[8] Francois Chollet et al. Keras, 2015.

[9] S. Christoffersen, E. Danesh, and D. Musto. Why do institutions delay reporting their shareholdings? evidence from form 13f. 2015.

[10] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[11] Simon Cross, Robert Harrison, and R Kennedy. Introduction to neural networks. *The Lancet*, 346(8982):1075–9, 1995.

[12] Timothy Dozat. Incorporating nesterov momentum into adam. 2016.

[13] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[14] Ehsan Habib Feroz, Taek Mu Kwon, Victor S. Pastena, and Kyungjoo Park. The efficacy of red flags in predicting the sec's targets: an artificial neural networks approach. *Intelligent Systems in Accounting, Finance and Management*, 9(3):145–157, 2000.

[15] V. Filimonov and D. Sornette. A stable and robust calibration scheme of the log-periodic power law model. *Physica A: Statistical Mechanics and its Applications*, 392(17):3698 – 3707, 2013.

[16] Andrea Frazzini, David Kabiller, and Lasse Heje Pedersen. Buffett's alpha. *Financial Analysts Journal*, 74(4):35–55, 2018.

[17] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[18] Trevor Hastie. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York NY, second edition, corrected at 12th printing 2017 edition, 2017.

[19] J. Heaton. An empirical analysis of feature engineering for predictive modeling. In *SoutheastCon 2016*, pages 1–6, 2016.

[20] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[21] Sotiris Kotsiantis, Euaggelos Koumanakos, Dimitris Tzelepis, and Vasilis Tampakas. Predicting fraudulent financial statements with machine learning techniques. In *Advances in Artificial Intelligence: 4th Helenic Conference on AI, SETN 2006, Heraklion, Crete, Greece, May 18-20, 2006. Proceedings*, volume 3955 of *Lecture Notes in Computer Science*, pages 538–542. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[22] Jerry W Lin, Mark I Hwang, and Jack D Becker. A fuzzy neural network for assessing the risk of fraudulent financial reporting. *Managerial Auditing Journal*, 18(8):657–665, 2003.

[23] Warren S. McCulloch and Walter Pitts. *A Logical Calculus of the Ideas Immanent in Nervous Activity*, page 15–27. MIT Press, Cambridge, MA, USA, 1988.

[24] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[25] Partha Mohanram. Separating winners from losers among lowbook-to-market stocks using financial statement analysis. *Review of Accounting Studies*, 10(2-3):133–170, 2005.

[26] James Montier. Cooking the books, or, more sailing under the black flag. *Mind Matters*, June 2008.

[27] Benjamin Moritz and Tom Zimmermann. Tree-based conditional portfolio sorts: The relation between past and future stock returns. *SSRN Electronic Journal*, 2016.

[28] James A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131, 1980.

[29] Travis Oliphant. NumPy: A guide to NumPy. USA: Trelgol Publishing, 2006–. [Online; accessed ¡today¿].

[30] Jane A Ou and Stephen H Penman. Financial statement analysis and the prediction of stock returns. *Journal of accounting & economics*, 11(4):295–329, 1989.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Johan Perols. Financial statement fraud detection: an analysis of statistical and machine learning algorithms.(report). *Auditing: A Journal of Practice & Theory*, 30(2):19, 2011.

[33] Joseph D. Piotroski. Value investing: the use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, 38(SUPP):1–51, 2000.

[34] Joseph D. Piotroski and Eric C. So. Identifying expectation errors in value/glamour strategies: A fundamental analysis approach. *The Review of Financial Studies*, 25(9):2841–2875, 2012.

[35] Richard Sloan. Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 71(3):289, 1996.

[36] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[37] Jan Szilagyi, Jens Hilscher, and John Campbell. In search of distress risk, 2008.

[38] Luo Y. Wang S. Signal processing: The rise of the machines. *Deutsche Bank Quantitative Strategy*, Jun 2012.

[39] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

THE ROBO-INVESTORS FROM GRAHAM-DODDSVILLE -- APPLYING MACHINE LEARNING TO THE INVESTMENT CHOICES OF WARREN BUFFETT

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):
SCHWEIZER - GAMBORINO

First name(s):
ERNST FLORIAN

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date
UERIKON, 22. 12. 2020

Signature(s)
f.

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.