Algorithmic Application to Stock Trading Methods

Edward Celella
Supervisor: Dr Shan He

School of Computer Science, University of Birmingham,
Birmingham B15 2TT, UK
emc918@student.bham.ac.uk

Abstract. The purpose of this paper is to provide reasoning as to why machine learning is an effective tool in the field of stock trading. This is achieved by an examination of the traditional methods used by active traders, namely fundamental and technical analysis. The review of these techniques includes the models utilised, as well as the underlying economic theories each strategy rests on. The conclusion reached by this paper is that machine learning is applicable to the field, as many of the underlying structures used by machine learning models (e.g. linear and non-linear), are already utilised within the field. Furthermore, the task is shown to be easily formulated into other types of problem structures (e.g. Bayesian and ensemble). However, although machine learning is shown to be an effective tool, it is noted that many studies in the field train models using technical indicators (e.g. past price values). This paper instead advocates for the use of fundamental indicators over technical, due to the underlying economic theories technical analysis rests on, producing inherent noise in the data. In comparison, fundamental data does not suffer from this problem. Thus, the case is made that it can be used to produce more accurate, and universally applicable models.

Keywords: Stock Trading · Fundamental Analysis · Technical Analysis · Algorithmic Trading · Machine Learning Application

⋆ *Preprint submitted for Individual Study 2 (2020)*

Table of Contents

List of Figures

List of Tables

List of Equations

1 Introduction

Every day businesses require capital in order to operate. One such method of obtaining funding is through investors. Investors provide funding, in return for shares of a business. These shares recuperate investments via dividends payed to the holder, or through selling them when they are valued at a higher price. In order for investors to mitigate the risk of losing money (in the event a business shares lose value and/or profits), they require techniques which can predict whether an investment is good or bad.

Investment funds are traditionally managed using one of two techniques, passive or active. Both of which use widely different approaches.

Passive investors utilise indexes in order to make decisions. Indexes are a collection of companies that fall within a specific sector, each companies stock price is taken in order to produce an overall price for the index (either as an average or weighted calculation). This provides a snapshot of the economic sector. An example of an index is the FTSE 100, which is a collection of the top 100 companies with the largest market capitalisation (market value of outstanding stocks) on the London Stock Exchange. Passive investors merely invest in an index, the goal being not to beat the market, merely match its growth (O'Shaughnessy 1997, p.1).

Active investors on the other hand attempt to beat the market. They achieve this by analysing a companies history, using past prices, growth rates and a multitude of other factors in order to reach a decision. These variables are then used to forecast the future trend of a companies price, and thus deducing whether to buy or sell shares in that particular company (O'Shaughnessy 1997, p.1). Active investment strategies generally fall into one of two camps. The first being technical analysis, which solely uses the historical data of a companies stock price, in order to forecast future prices. The second is fundamental analysis, which looks at the business itself in order to make decisions.

Algorithmic trading is a relatively new field of study, and is the application of computer programs to the world of trading. This study will outline how the field of machine learning is being used to develop new strategies that help investors in trading world. Furthermore it will delve into the world of stock trading, outlining the traditional methods used, in order to develop an understanding of how the data can be used with established machine learning models, and using this knowledge, provide evidence as to why they are applicable.

2 Technical Analysis

2.1 Introduction

Technical analysis solely uses the past share prices in order to generate a forecast for the future(Becket & Essen 2010, p.70). It is based upon the economic theory castle-in-the-air which was proposed by John M. Keynes in 1936. The theory is based on the idea that movements in the market are not governed by the value of a company, but instead the psychological behaviour of investors. The general idea is that investors "follow the crowd", thus when a subset begin to invest others do so as well. This herd mentality builds up hope on investment opportunities, hence producing "castles in the air". Therefore, in order to forecast the movement of stock prices, one must identify which opportunities are "susceptible to public castle-building and then buying before the crowd". (de Prado 2018, p.100)Malkiel1973

The methodology is also based heavily on the efficient market theory. An efficient market is any capital market in which security prices at any time "fully reflect all available information" (Fama 1970). This is formally defined in equation 1 where: E is the expected value operator, $p_{j,t}$ is the price of security j at time t, $r_{j,t+1}$ is the one period percentage return $(p_{j,t+1} - p_{j,t})/p_{j,t}$, and Φ_t is the set of information available. Variables with the tilde indicate random variables (Fama 1970).

$$E(\tilde{p}_{j,t+1}|\Phi_t) = [1 + E(\tilde{r}_{j,t+1}|\Phi_t)]p_{j,t} \tag{1}$$

However, the efficient market hypothesis also works against the idea of technical analysis. This is due to two factors. The first being that because security prices fully reflect all available information, each change in price is therefore independent to the previous (as the information between each step changes). The second factor is that each change is identically distributed. This means that although the current market information is the same across all price changes, each price is a different representation of said information. These two factors form the random walk hypothesis, which states that the movement of prices is merely a series of random steps. Therefore, using only the previous price data will give no insight into future movements. (Fama 1970)

But proponents of technical analysis refute this claim. Lo and MacKinley in their book "A non-random walk down wall street" (Lo & MacKinlay 1999), state that prices are not random but move in trends, and thus can be predicted. Furthermore, the idea that the market as a whole is random does not necessarily mean technical analysis is a redundant system. As the market "may witness

extended periods of random fluctuation, interspersed with shorter periods of nonrandom behaviour" (Schwager 1999, p.1).

Combining these two theories forms the core of technical analysis. Each price movement is a causation of all available information. This information is available to investors, and so technical analysts (chartists) can use past data to predict behavioural patterns.

2.2 Models and Techniques

The field of technical analysis utilises a variety of techniques in order to forecast future price movements. Jack Schwager, in his book "Getting Starting in Technical Analysis" (Schwager 1999, p.1), lays out the key methods utilised, and how to form models using these methods. This section will provide an overview of said methods.

2.2.1 Trends

The ability to identify price trends is one of the fundamental building blocks of technical analysis. Trends are displayed visually using trend lines. Uptrend lines connect a series of significant higher low values, and downtrend lines connect a series of significant lower high values. There is no consensus on how to draw trend lines due to the process being fundamentally subjective. However, objective methods have been developed such as only connecting the two most recent relative high or low values (DeMark 1994). Generally, trend lines can be used to identify profit taking zones. These zones occur when a trend line changes direction, which indicates to buy if a downward trend is penetrated, and to sell if an upwards trend is penetrated. Internal trend lines are a different form of trend analysis, which simply draws a line which approximates the relative high and low points, without any consideration to extremes. These internal lines are analysed in the same way as the standard.

Another method of trend analysis is through the use of moving averages. A moving average is simply the average price over the past N time steps (equation 2). Normally, a moving average is calculated by taking the closing price over the past N days, but any price metric (i.e. high, low, open) and time step (i.e week, month, year) can be used. These points can be plotted, which not only accurately reflects trends in the price, but also smoothes out unimportant fluctuations (with higher values of N producing smoother graphs). However, these smoothing properties come at the expense of time-lag, which means sudden changes in trend take longer to catch.

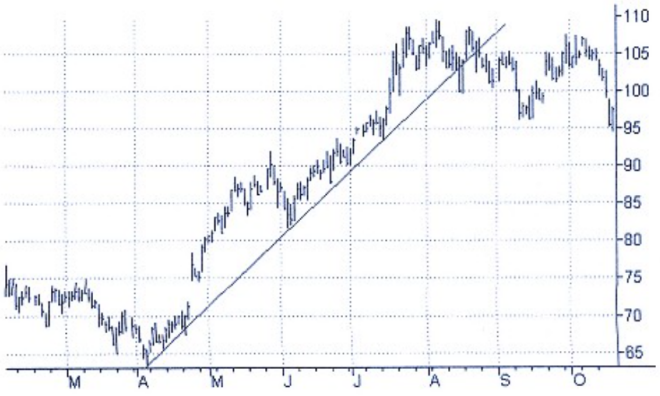

Fig. 1: Graph showing an uptrend line. (Schwager 1999)

$$MA = \frac{\sum_{t=1}^{N} p_t}{N} \tag{2}$$

where:

t = The current time step.
N = Amount of time steps.
p_t = The price at time step t.

Variations of moving averages seek to improve responsiveness to market changes by weighting values. The two most common variations are linear weighted moving average (LWMA), and exponentially weighted moving average (EWMA). Both these techniques weight recent prices higher than older prices, in order to catch sudden movements in the market more quickly. LWMA weights each day by the time step (equation 3). Whereas EWMA, defined in equation 4, uses a smoothing constant (α), which weights all previous prices with an exponentially decreasing drop-off.

$$LWMA = \frac{\sum_{t=1}^{N} p_t}{\sum_{t=1}^{N} t} \tag{3}$$

$$EWMA = \alpha.p_t + (1 - \alpha).EWMA_{t-1} \qquad (0 \leq \alpha \leq 1) \tag{4}$$

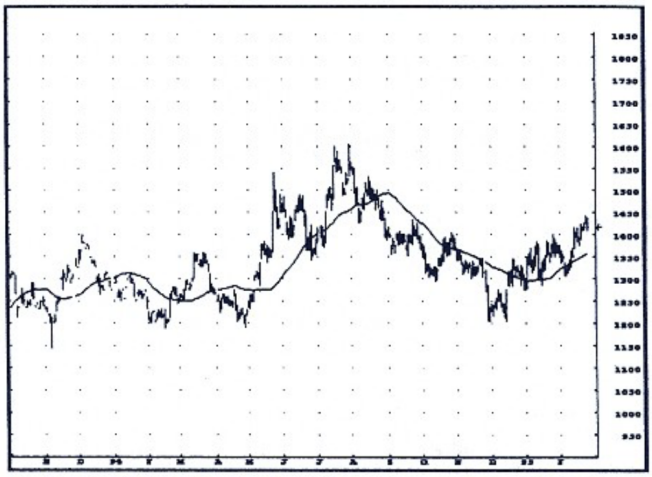

Fig. 2: Moving average graph plotted against stock price. (Schwager 1999)

These techniques form the basis of trend following models, which identify trends and invest assuming that the trend will continue. The most common form of model in this category are moving average models. In general these models use two moving average trend lines, one with a high N (slow trend line), and the other with a low N (fast trend line). If the slow trend line rises above the fast trend line, this is an indicator to sell. The reverse is true to buy. Breakout systems are another common trend model. These models simply track the markets ability to reach new high and low points over the past N days, and assume the market will continue along this trend.

2.2.2 Chart Patterns

Chart patterns are predefined patterns which occur in price charts. Each pattern serves as an indicator to the future direction of market movement. A few common patterns are described briefly in appendix A . These patterns on their own do not provide solid indication of a price movement, however pattern recognition models can combine multiple patterns in order to produce accurate forecasts (e.g. if 5 specific patterns occur then initiate transaction).

2.2.3 Oscillators

Oscillators are mathematical models which calculate the rate of change in order to describe the momentum of a market. A strong momentum indicates that the

trend will continue, and vice versa. Usually all oscillators use the close price as the daily value in calculations, with the most basic oscillator being that of momentum which is simply the current price subtracted by the price N days ago (equation 5). The rate of change (ROC) is another common oscillator which simply divides instead of subtracting (equation 6). These oscillators use a signal line, which is simply the median value, and initiate buy or sell signals if the momentum line crosses above or below said signal line respectively.

$$Momentum = p_t - p_{t-N} \tag{5}$$

$$ROC = \frac{p_t}{p_{t-N}} \tag{6}$$

Moving averages, as those described in equations 2-4, can also be used to form oscillators. This can be achieved my simply subtracting the moving average value from the corresponding price value. Price oscillators work similarly be subtracting a slow moving average from a fast moving average. An example of a price oscillator is the moving average convergence-distance (MACD) indicator. This model simply uses 12 and 26 day EMWA, and then forms a 9 day moving average by calculating the distance. This 9 day moving average is used as a signal line, which indicates to buy or sell shares if the primary MACD line crosses above or below the signal line respectively.

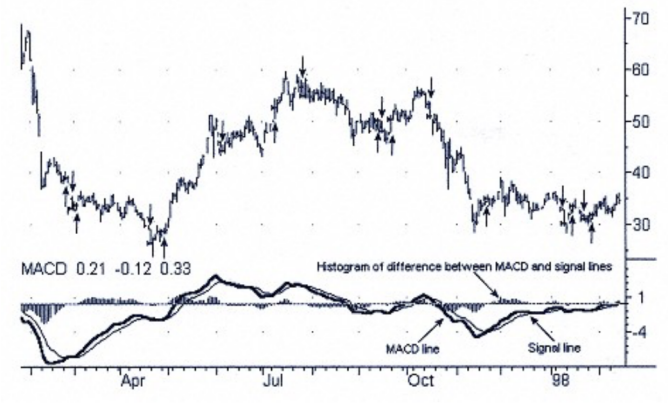

Fig. 3: MACD and signal line shown against price graph. (Schwager 1999)

Oscillators can also be normalised. These models have the advantage of setting predefined overbought and oversold lines, which provide greater detail into when

to buy or sell shares. One such example of a normalised oscillator is the relative strength index (RSI) (Wilder 1978), which is formally defined in equation 7.

$$RSI = 100 - \frac{100}{1 + \text{average price increase/average price decrease}} \tag{7}$$

where:

$$\text{average price increase} = \frac{\sum_{t=1}^{N} \overbrace{(p_t - p_{t-1})}^{\text{0 if less than 0}}}{N}$$

$$\text{average price decrease} = \frac{\sum_{t=1}^{N} \overbrace{(p_t - p_{t-1})}^{\text{0 if greater than 0}}}{N}$$

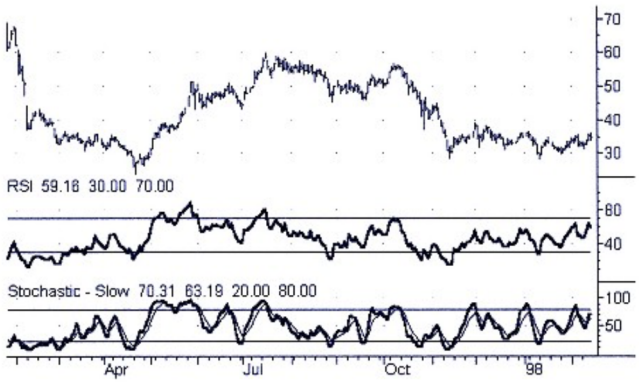

Fig. 4: RSI against stock price. (Schwager 1999)

Oscillators are the basis of countertrend models, which use these mathematical models and indicators for trend reversal. It is at these points an investor should buy or sell shares depending on the direction of change.

3 Fundamental Analysis

3.1 Introduction

Fundamental analysis is the study of a businesses economic strength. The goal of fundamental analysis is to provide a measure of companies value, which can then in turn be used to guide investment strategies. It is based on the Firm-Foundation economic theory, which hypothesises that each investment opportunity has an intrinsic value. This value can be determined through analysis of the past, present, and future conditions (Malkiel 1973, p.29).

This idea was formalised by John B. Williams in his book "The Theory of Investment Value" (Williams 1938), which relates the value of a stock to the value of future dividends. In other words, Williams directly relates the value of a stock to the claim on future goods. The model proposed is built upon the time value of money, which simply takes into account that money over time will increase due to interest (equation 8) (Chen 2020).

$$FV = PV \times (1 + i) \tag{8}$$

where:

FV = future value
PV = present value
i = interest rate

Williams expands upon this concept, by using the idea that the future value of a stock is solely the dividends expected the next year. Therefore, the present intrinsic value of a stock can be determined by discounting the interest rates on future dividends (equation 9) (Williams 1938). This calculated intrinsic value can then be compared to the actual current stock price, to evaluate whether the investment is over or under priced. This idea of formulating an intrinsic value, is the core idea behind all fundamental models, and many of the techniques used aim to forecast a companies future dividends. This makes logical sense, as for investors to make money they must locate opportunities that are undervalued, with the expectation they will grow.

$$V_0 = \sum_{t=1}^{N} \pi_t . \left(\frac{1}{1 - i} \right)^t \tag{9}$$

However, like technical analysis, the random walk hypothesis apposes this theory. This is due to the idea that the markets efficiency in capturing and applying information to current prices (equation 1), cannot be matched. In other words, the market uses information at a size, and applies this information at a rate, which cannot be replicated by any investors models. Furthermore, the techniques employed by fundamental analysis relies on the assumption that the information, and application of this information, can correctly estimate the intrinsic value of a company. And even if these methods are correct, the market may never reflect the true intrinsic value of a company, due to its inherent volatility (Malkiel 1973, p.132). As Malkeil describes, "the security analyst must be a prophet without the benefit of divine inspiration" (Malkiel 1973, p.129).

But supporters of fundamental analysis disagree with the efficient market hypothesis (and by extension the random walk). Starting that the theory treats the market as a separate entity from the actual businesses it represents, when in reality the market is merely built upon these businesses (Shostak 1997). Therefore, by definition, analysis of the business allows prediction of its share price. In addition to this, even if the information used is incorrect, as all investors have access to the same information, it is merely the interpretation of the information which guides price. Further to this point, in James O'Shaughnessy book "What Works on Wall Street", the idea is proposed that it is not the models which are ineffective, but merely the humans which implement them (O'Shaughnessy 1997, p.1). The theory that human judgement is limited has been a well known idea within the scientific world. This was first cemented by David Faust, who found that numerical models outperformed human judges over a variety of fields consistently (Faust 1984). In relation to the stock trading, O'Shaughnessy found this to be also true, with even simple models such as the Dogs of Dow, seeing a compound return of of 12.4% per year(O'Shaughnessy 1997, p.10).

3.2 Models and Techniques

Fundamental analysis takes two-forms, quantitative and qualitative. Qualitative techniques involve an investor making a judgement of a businesses value through methods such as interviews with the CEO and employees. These methods are useful, however as previously discussed it is human judgement which causes the failure of models. As O'Shaughnessy states "In almost every instance, from stock analysts to doctors, we naturally prefer qualitative, intuitive methods. In most instances, we're wrong"(O'Shaughnessy 1997, p.14).

In contrast the quantitative method use a companies current financial situation, as well as its financial history in order to make predictions. The models developed for this task use this data to produce indicator values, which measure a wide

variety of factors such as risk, volatility, and return on investment. Appendix B describes some of the common indicators used, detailed in the book "How the Stock Market Works" (Becket & Essen 2010, p.58). This section evaluates the two main strategies in fundamental analysis, value and growth, as detailed in James P. O'Shaughnessy book "What Works On Wall Street" (O'Shaughnessy 1997, p.1).

3.2.1 Value Factors

Value investing is the process of identifying undervalued stocks, with the idea that these prices will rise to their intrinsic values. There are five main indicators utilised by value investors, which provide a ratio between the price of a share and a certain aspect of the companies financial information:

1. The price-to-earning factor provides a ratio between the earnings of a company per share, and the price of a share (equation 10). Investors interpret price-to-earnings ratio differently, with value investors believing the right time to invest is when the ratio is a low value, as this shows the price of the share is at a discount.
2. Price-to-Book provides another value ratio, but instead measures the price of a share against the book value per share. The book value per share is simply the equity available to a shareholder divided by the number of outstanding shares (Hayes 2020). Low values indicate that the share price is close to the liquidated value of the company, thus meaning investors will be paying a low price for the companies assets.
3. Price-to-Cashflow measures the price of a share against a businesses cashflow. Lower values indicate a stock that is being sold below the intrinsic value of a company and so should be bought.
4. Price-to-Sales is another measure of the reasonability of a shares current price. It is a ratio between the price of a share and the amount of sales it has made in a year. Again, lower ratios indicate the company is undervalued, and thus is an indication to buy.
5. Measuring a companies dividend yield is another straightforward value indicator of a stock. It simply divides the annual dividend rate paid to investors per share, by the price of the share. This value is then converted to a percentage. Stocks with high dividend yields are therefore intrinsically worth more, due to the fact "dividends have historically accounted for more than half a stock's total return"(O'Shaughnessy 1997, p.143).

Each of these indicators, although simple, can be used to identify investment opportunities to a surprisingly accurate effect. With, O'Shaughnessy finding that the price-to-sales ratio performed the best, returning a compound return of 15.95% between 1951 and 2003 (O'Shaughnessy 1997, p.128).

3.2.2 Growth Factors

Growth investing techniques aim to identify stocks which will outperform the market, due to their intrinsic value and future potential. As with value investing, there are a variety of indicators used to measure the growth of a company.

One common strategy used by growth analysts is the increase in a companies earnings over time. This method uses the same price-to-earnings ratio (equation 10) as discussed with value investing. However, growth investors believing higher ratio values are better investment opportunities, as the value indicates future growth. This strategy can taken further by evaluating the percentage change in the price-to-earnings over a multi-year period (usually 5 years). This mitigates the chance a company has had a single good year, providing a more accurate picture of a companies growth rate.

$$\mathrm{P/E} = \frac{\text{Share Price}}{\text{Earnings per Share}} \tag{10}$$

Another method used to evaluate a companies growth is the evaluation of its profits. Two indicators which measure this are Profit Margin (equation 11) and Return on Equity (equation 12). These are given by the ratio of the company's net income to net sales and shareholder equity respectively. Higher profit margins show that a company has a greater operating efficiency, and are more competitive within its respective market. Whilst return on equity indicates how effective a company is investing its assets.

$$\text{Profit Margin} = \frac{\text{Income} - \text{Expenses}}{\text{Net Sales}} \times 100\ (\%) \tag{11}$$

$$\text{Return on Equity } = \frac{\text{Income} - \text{Expenses}}{\text{Shareholder Equity}} \tag{12}$$

In addition to these models, a simpler strategy is often deployed by growth investors, which is that stocks that are increasing value will continue to increase. This is idea is formalised as the relative price strength of a company, and simply divides the price share of one company by another in the same industry. This model is directly related to technical analysis, as it involves identifying the stocks which are undergoing an upwards trend (which is achieved by comparing the current price strength to its past price strength). However, different investors interpret the result of this indicator differently, with some preferring declining company strength in order to identify stocks which are cheap, and are ready to rise back to their intrinsic value.

Unlike the value factors, O'Shaughnessys' research models showed that the growth indicators failed to beat the market a majority of the time (O'Shaughnessy 1997, p.194 and p.219). The best results came from the use of the relative strength index, which when utilised as a short-term indicator achieved a compound return of 12.6% between 1951 and 2003. But this was a high risk strategy, having a standard deviation of 37.8% (O'Shaughnessy 1997, p.222)

3.2.3 Multi-Factor Models

Although each of the indicators discussed in this section can be used independently to make investment decisions, the predictive ability of these factors increases when they are used together. These multi-factor models simply take a certain number of factors, and indicate to investors stocks which meet all the criteria. Both growth and value factors can be used in multi-factor models, with each strategy seeing an increase in its predictive capability.

A common two factor model is the combination of a price ratio factor along side the relative price strength of a company over a one year period. This strategy simply involves evaluating all companies using a price ratio, and then selecting N companies with the best relative price strength that achieve a price ratio below a certain value. These two factor models can themselves be further extended, by combining multiple two factor models. This strategy works by evaluating companies using each model independently, and then weighting the investment by the amount of times it is indicated by the models. This results in the same return on investment as using the models independently, but reduces the overall risk. (O'Shaughnessy 1997, p.243)

4 Algorithmic Trading

4.1 Introduction

The application of computer algorithms to the field of trading is an expansive and fastly developing field. The focus of this study will be on how machine learning algorithms are applied to the field, and why they work. However, it should be noted that this isn't the only form of algorithmic trading. For example, the simplest form of this field is the direct implementation of the previously described models (sections 2 and 3), as computer programs. Although this is a relatively simple approach, it does come with a few advantages over manually using the models. Firstly, fully automated systems are faster than humans, and are less error prone. This advantage in time provides a competitive edge over traders who do not use computerised systems, as it allows investment in viable businesses before the crowd. This time advantage is especially prevalent when applied to technical analysis, which requires a trend to start before investing (Chan 2009, p.80).

Furthermore, the removal of human judgement from the decision making process allows the models to perform at their best. As previously discussed O'Shaughnessy found that, when investing took a consistent approach, many of the models used were accurate predictors. It was merely the lack of discipline, and human bias, which caused deviation from these strategies and thus losses (O'Shaughnessy 1997, p.14).

The remainder of this chapter focuses on how financial data can be applied to machine learning models. Each section explains why certain models are applicable to the domain problem, and provides examples of research applying the models.

4.2 Training and Testing using Financial Data

The first task of developing any machine learning algorithm is an understanding of the data being used, as well as appropriate preparation for training and testing. As discussed throughout this report, financial data falls into one of two categories, fundamental or technical. Both types of data can be applied to machine learning models with varying effect.

One of the main issues that plagues the financial application of machine learning is backtesting, which is the evaluation of a trading model using historic data. One of the main approaches in determining the fitness of a model is cross validation, however due to the nature of financial data, this method can often lead

to misleading results. The problem is due to leakage, which occurs as financial data has a serialised structure.

Supervised learning models require an input vector, $\boldsymbol{x}$, and a corresponding label, y. The models goal is to develop a system which can approximate y from $\boldsymbol{x}$. As financial data is serially correlated, this means that the labels correspond to overlapping data points (equation 13) (de Prado 2018, p.103-111). As cross validation partitions the data, this means that t and $t+1$ will occur in different sets. Thus meaning some of the training set occurs in the test set. This leads to inaccurate results, as the model is more likely to predict y_{t+1}, even if $\boldsymbol{x}$ contains features irrelevant to the prediction making process. The problem compounds, resulting in an overfitted model. (de Prado 2018, p.103-111)

$$\begin{aligned} \boldsymbol{x}_t &\approx \boldsymbol{x}_{t+1} \\ \therefore y_t &\approx y_{t+1} \end{aligned} \tag{13}$$

A solution to this problem is simply to remove overlapping data from the training set, which occurs in the test set. Methods which reduce overfitting such as bagging, can also mitigate the impact of leaked data.(de Prado 2018, p.103-111) In addition to this, there are other challenges which must be contended with when preparing financial data such as (Jansen 2018, p.130):

1. Survivorship Bias - Ensuring past data includes securities which are no longer listed.
2. Look-Ahead Bias - Ensuring the past data used for testing, was only available at that time (e.g. corrections on financial reports).
3. Outlier Control - Outliers in financial data can sometimes be the most informative data points, therefore any outliers must be carefully considered before removing.

4.3 Linear Models

Linear regression is a widely used technique, in formulating a model which describes the relationship between a set of features and an output, by redistributing weights between features (equation 14). Formulating stock prediction as a linear regression task can work in two ways. The first method is to use the raw financial data as the input. The second option is to use the numeric results produced from the fundamental factors (see section 2 and appendix B). This option forms the basis for factor models, which are widely used to evaluate to relationship between risk and return of an investment. (Jansen 2018, p.175)

$$y = \beta_0 + \sum_{i=1}^{N} \beta_i . x_i + \epsilon \tag{14}$$

One of the most prominently used linear factor models is the fama-french three-factor model, which was proposed by Eugene Fama and Kenneth French, in their paper "Multifactor Explanations of Asset Pricing Anomalies" (FAMA & FRENCH 1996). In their study, they discovered that the 95% of the return on a stock trading portfolio is determined by the three factors:

1. Market risk.
2. The performance of small capitalization companies relative to high capitalization companies (SMB).
3. The out performance of companies with a low book-to-market ratio relative to those with a high ratio (HML).

These three factors can be used to predict expected return of an investment portfolio (equation 15), and their research shows that investments in small capitalization companies, and value stocks outperform investments in high capitalization companies and growth stocks(FAMA & FRENCH 1996). In 2015 this model was extended to include two other factors, profitability and investment. (Fama & French 2015)

$$R_i - R_f = \alpha_i + b_i(R_m - R_f) + s_i . SMB + h_i . HML + \epsilon_i \tag{15}$$

Where:

$R_i - R_f$ = Expected Excess Return
$(R_m - R_f)$ = Total Market Portfolio Return - Risk Free Rate of Return
SMB = Size Premium (Small Minus Big)
HML = Value Premium (High Minus Low)
b_i, s_i, h_i = Coefficients

As shown by equation 15, linear models have a solid foundation within the world of stock prediction. Noting this precedent, shows a clear use for linear regression within this industry. It is therefore unsurprising that this a field of active research, with many studies obtaining promising results.

4.4 Non-Linear Models

As with linear, non-linear models also have a precedent within traditional stock market techniques. For example the moving average models, such as EMWA and MACD (section 2.2.1 and 2.2.3), are all examples of non-linear models. These techniques formulate the process of forecasting stocks as a time series problem. This method of structuring the data is natural in this field, as financial data is already organised in this way (e.g. daily share price data).

Neural networks are one of the most commonly used non-linear machine learning algorithms. It is therefore of no surprise that these algorithms have been applied to this topic in many formats. One particular study of note by Min Qi(Qi 1999)compared linear models to non-linear neural networks, in predicting the movement of the S&P 500 index. The study found that the non-linear neural networks produced more accurate predictions. In addition to this when comparing recursive linear and non-linear models, the non-linear models had a higher risk-adjustment returns.

4.5 Bayesian Models

The traditional process of stock market forecasting required investors to make decisions on securities, using indicators to guide judgement. This process can be formalised as a Bayesian model. For example, equation 16 shows how the use of price-to-earnings ratio (equation 10) can be expressed in this manner.

$$P(\text{Stock Increasing} \mid P/E\ ) = \frac{P(\text{ Stock Increasing }) \times P(\ P/E \mid \text{Stock Increasing }\ )}{P(\ P/E\ )} \tag{16}$$

Formulating the problem as a Bayesian probability model provides many advantages, as it allows for prior beliefs to be updated given new information. This method can be applied to a both fundamental and technical indicators, and even allows the combination of multiple factors. Furthermore the use of a Bayesian framework allows model parameters to be inferred from given financial data (e.g. income of a company, stock price movements etc.). This can further enhance linear and non-linear models, as well as used independently as its own model. (Jansen 2018, p.268)

One such example of Bayesian models being applied to this field, was a study conducted in 2015 which used 9 economic factors in order to produce a dynamic Bayesian factor graph. This model was then used to forecast the movement

of the ShenZhen and S&P 500 indexes. This solution managed to capture all major changes in trend for both indexes, which were indicated through structural changes to the graph.(Wang et al. 2015)

4.6 Ensemble Methods

As with any application of machine learning, bias, variance, and noise are problems which must be contended with. The choice of which to reduce is dependent on the type of data used. For example, fundamental data is considered to have relatively low noise, whereas technical data does. This means that for fundamental techniques, boosting is a more useful tool. However, bagging provides a reduction in overfitting, which tackles the main problem when working with financial data as discussed in section 4.2 .(de Prado 2018, p.100)

One specific ensemble method which is of high use in this field are random forests. This is because decision trees can formalise the decision making process of investors, removing human judgement. Figure 5 demonstrates this, by modelling a two-factor model described in section 3.2.3. The use of decision trees are an effective tool, due to the fact they can clearly show relationships between data, however they are prone to overfitting. This problem is overcome through the use of random forests, which reduces the variance of a a group of models, whilst preventing overfitting due to its additional random element. Random forests can not only be applied to machine learning algorithms, but also the traditional methods discussed in sections 2 and 3.

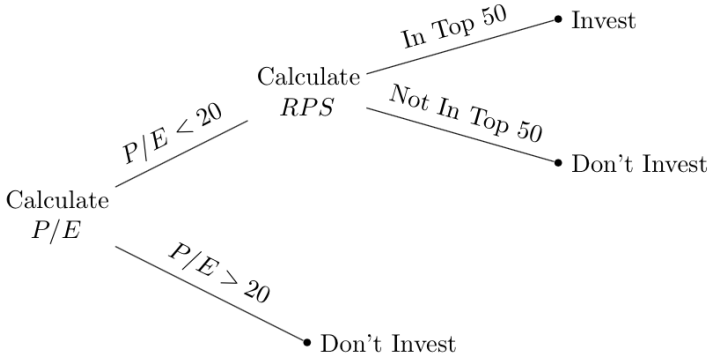

Fig. 5: Decision tree of P/E and RPS two factor model.

The application of random forests in this way has been studied and produced promising results. One such study, used a variety of traditional technical indicators and models, such as trend signals, oscillators and volume indicators. These metrics were then applied to a random forest algorithm, and the resulting model was then tested by simulating trading using businesses in the S&P 500. In this test, the resulting algorithm did not beat the benchmark, and thus failed as a profitable solution. However, when used with a non-stationary time series the results improved, indicating that the model could work given live data.(Ładyżyński et al. 2013)

4.7 Evolutionary Models

The use of evolutionary models within this subject is also an extensive point of research. A majority of studies use evolutionary techniques conjunction with other models, in order to improve their performance. One such example of this used particle swarm optimisation, in order to improve the fitness of single layer neural networks trained on past share prices (technical indicators). The resulting networks are then used to calculated recommendations on 350 S&P 500 stocks, with each output then passed into a hyperbolic tangent function, which produces a value of -1, 0 or +1 used to rank stocks. The strategy managed to consistently beat the market (with a low commission fee of 0.15%), shown in figure 6. (Nenortaite & Simutis 2004)

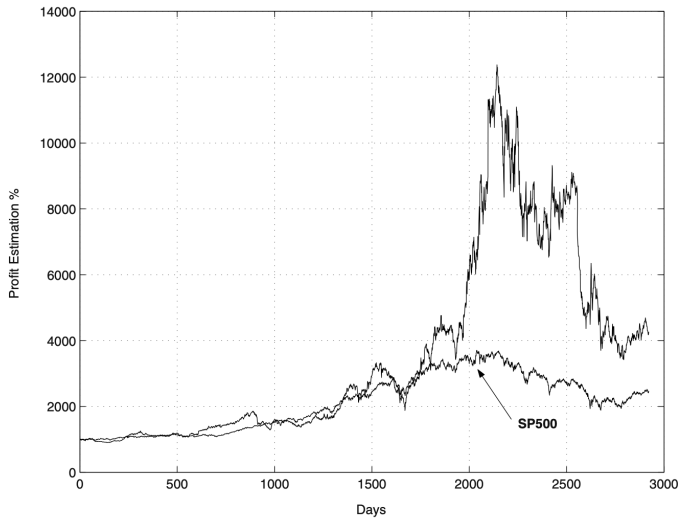

Fig. 6: Profit estimations obtained from partcle swarm and neural network model. (Nenortaite & Simutis 2004)

However, with the relatively recently developments in evolutionary programming, some researchers have generated forecasting models directly from evolutionary techniques. A study conducted in 2000, used genetic programming to construct a regression models for this purpose, utilising technical indicators. The resulting model produced an effective single-day trading strategy, showing that the application of genetic programming could work in this field. However, the model being limited forecast to a single day shows that work in this field still requires development. (Kaboudan 2000)

5 Concluding Remarks and Discussion

As shown in this paper, the field of stock trading is a diverse subject, with a multitude of techniques, models and strategies. Machine learning enhances this field, providing a new avenue for predictions to be made. However, after researching the current state of the field, it is clear that there are some problems, many of which I believe stem from a lack of domain knowledge on the side of computer scientists. One of the main points of contention I have with the current application of machine learning models, is the use of technical analysis as the input, which many studies do. I believe there are factors intrinsic to the data that lead to sub-optimal, or misleading results.

One problem machine learning models encounter, is noise and overfitting. The theory technical analysis is built upon the efficient market hypothesis, meaning that a stock price is influenced by all information within the market. This by definition means that price data has a large amount of built in noise, as although the price may represent all necessary information, this does not necessarily mean that all information is equally impactful.

Furthermore, as proponents of the efficient market point out, each stock is a different representation of the market information, thus any model which is trained on past price will be overfitted for that specific stock. This means for each stock that requires forecasting, the model will need to be trained on that data. This is a waste of computational resources, and in the world of finance, the speed of a system can be the cause of majour gains or losses.

In addition to this, even technical proponents accept that for the efficient market to be a true explanation, this must include random points of stock price movement. This only compounds the problem of noise within the data. This is a specific problem which cannot be overcome, as there is no way of knowing if a price move is random, without introducing fundamental factors to explain movements. Therefore, the conclusion can be made of just using the fundamental factors to begin with.

The use of fundamental factors solves the problems outlined above. This is due to financial data of a company being less up for interpretation. For example, the amount of sales which a company has made, reflects exactly what it describes. Although noise will still be prevalent in the data, this will be a much lower rate when compared with technical indicators, and unlike with technical indicators can be removed with thorough preprocessing. Using fundamental factors also allows for models to be developed, which can be deployed at a more general level. This is because if a relationship is found between factors and a movement, this can be applied to a wider range of stocks.

Another aspect of algorithmic trading which I've noted, is that some researchers in the field are aiming to develop one global model which can be solely used to identify stocks. It is my belief that this is the wrong approach. When looking at the traditional trading methods, it was never the case that one model or indicator was used, as the combination of multiple factors provided a better method of forecasting. Because of this, the use of ensemble and evolutionary models (through genetic programming) are of particular interest, as they can be used to simulate the decision making process of investors. This allows for multiple machine learning models to be combined in the same way investors combine traditional methods.

In my future work on this topic, I would like to explore the use of genetic programming in building decision trees. These will include machine learning models trained using fundamental data, as well as traditional models. In addition to this, the use of genetic algorithms to build factor models is an interesting avenue of research. For example, I would be interested in building new factor models, using the traditional factor models as the initial population, in order to discover if any new relationships can be found.

Bibliography

Becket, M. & Essen, Y. (2010), *How the Stock Market Works: A Beginner's Guide to Investment*, 3rd edn, Kogan Page.

Chan, E. (2009), *Quantitative Trading: How to Build Your Own Algorithmic Trading Business*, The Wiley trading series, John Wiley & Sons.

Chen, J. (2020), 'Time value of money (tvm) definition'.
URL: *https://www.investopedia.com/terms/t/timevalueofmoney.asp*

de Prado, M. (2018), *Advances in Financial Machine Learning*, Wiley.

DeMark, T. (1994), *The New Science of Technical Analysis*, Wiley Finance, Wiley.

Fama, E. F. (1970), 'Efficient capital markets: A review of theory and empirical work', *The Journal of Finance* **25**(2), 383–417.

FAMA, E. F. & FRENCH, K. R. (1996), 'Multifactor explanations of asset pricing anomalies', *The Journal of Finance* **51**(1), 55–84.

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

Faust, D. (1984), *The Limits of Scientific Reasoning*, ned - new edition edn, University of Minnesota Press.

Hayes, A. (2020), 'Book value of equity per share (bvps) definition'.
URL: *https://www.investopedia.com/terms/b/bvps.asp*

Jansen, S. (2018), *Hands-On Machine Learning for Algorithmic Trading: Design and implement investment strategies based on smart algorithms that learn from data using Python*, Packt Publishing.

Kaboudan, M. A. (2000), 'Genetic programming prediction of stock prices', *Computational Economics* **16**(3), 207–236.

Ładyżyński, P., Żbikowski, K. & Grzegorzewski, P. (2013), Stock trading with random forests, trend detection tests and force index volume indicators, *in* L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh & J. M. Zurada, eds, 'Artificial Intelligence and Soft Computing', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 441–452.

Lo, A. W. & MacKinlay, A. C. (1999), *A Non-Random Walk Down Wall Street*, Princeton University Press.

Malkiel, B. G. (1973), *A Random Walk Down Wall Street*, Norton, New York.

Nenortaite, J. & Simutis, R. (2004), Stocks' trading system based on the particle swarm optimization algorithm, *in* M. Bubak, G. D. van Albada, P. M. A. Sloot & J. Dongarra, eds, 'Computational Science - ICCS 2004', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 843–850.

O'Shaughnessy, J. (1997), *What Works on Wall Street: A Guide to the Best-performing Investment Strategies of All Time*, What Works on Wall Street, McGraw-Hill.

Qi, M. (1999), 'Nonlinear predictability of stock returns using financial and economic variables', *Journal of Business Economic Statistics* **17**(4), 419–429.

Schwager, J. (1999), *Getting Started in Technical Analysis*, Getting Started In..., Wiley.

Shostak, F. (1997), 'In defense of fundamental analysis: A critique of the efficient market hypothesis', *The Review of Austrian Economics* **10**(2), 27–45.

Wang, L., Wang, Z., Zhao, S. & Tan, S. (2015), 'Stock market trend prediction using dynamical bayesian factor graph', *Expert Systems with Applications* **42**.

Wilder, J. (1978), *New Concepts in Technical Trading Systems*, Trend Research.

Williams, J. (1938), *The Theory of Investment Value*, Investment value, Harvard University Press.

A Technical Analysis Chart Patterns

Table 1: Patterns observed on price charts. (Schwager 1999)

Pattern	**Description**
Gaps	Gaps are days in which the daily low value of a day is above the previous days high value, or vice versa.
Spike High	A spike high is a day whose high value is drastically above the proceeding and succeeding days high value, whilst the days close is the near the lower end of the days range. This indicates that the market is about to downtrend.
Spike Low	A spike low is a day whose low value is drastically below the proceeding and succeeding days low value, whilst the days close is the near the higher end of the days range. This indicates that the market is about to uptrend.
Reversal Days	A reversal day is a day in which the market reaches a new high or low and then reverses direction at the close price. Formally, a reversal high day is a day which beats the previous days high, but closes below the previous days low. A reversal low day is a day whose low beats the previous days low, but then closes above the previous days high. These are interpreted as market trend reversals (much like spikes).
Thrust Days	Thrust days are categorised by their direction of movement. Upthrust days are days which close above the previous days high, and down thrust days are days which close below the previous days low. Series of upthrust days indicate a strong market, whereas series of down thrust days indicate the opposite.
Wide Ranging Days	A wide ranging day is any day whose range (difference between high and low values) is significantly larger than the preceding days. Wide ranges days can be formally defined as any day whose range is above the average range of the past N days. Wide ranging days can also indicate a reversal in market trend, depending on if the close price is near the higher or lower end, and the current market trend.
Continuation Patterns	Continuation patterns are merely phases of side ways movement that occur with long-term trends. These patterns are generally broken by the same trend that preceded them.
Triangles	Triangles are patterns which show the trend of a market. They are formed by drawing two trend lines, one connecting the relative highs, and the other connecting the relative lows. If the triangle is symmetrical, this is seen as an indication the market will carry on the current trend. Whereas in any other case, the direction of the hypotenuse shows the trend.

Continued on next page

Continued from previous page

Pattern	**Description**
Flags and Pennants	Flags and pennants are shorter congestion patterns within trends. They are interpreted as mere pauses in the current trend, and formed by drawing two trend lines in the same fashion as triangles. If the lines are parallel this is a flag, else it is a pennant.
V Tops and Bottoms	A V-pattern is a turn around point of a down or uptrend. These can generally only be defined through a combination of other patterns.
Double Tops and Bottoms	Double tops and bottoms are two peaks or troughs that occur within the same price vicinity sequentially. They occur after large price moves and are indications for market reversal. (Note that any number of tops and bottoms can occur, for example triple tops, but these are extremely rare)
Head and Shoulders	Head and shoulder patterns consist of one large peak or trough nested between two smaller peaks or troughs sequentially. The two low points between the three formations form a neckline, if this neckline is penetrated it is consider an indicator of trend reversal.
Rounded Tops and Bottoms	Rounded tops and bottoms are defined as the name implies. They indicate trend reversal.
Wedges	Wedges are either rising or declining. They are formed by using two lines which connect the relative highs and lows respectively. In rising wedges, prices increase in a converging pattern. Declining wedges show prices decreasing in a converging pattern. If the price breaks a wedge line, this can be interpreted as a sell or buy signal (depending on the direction of the break).
Island Reversals	Island reversals are similar to spike and reversal days. They occur when two price gaps isolate a small cluster of days, and imply trend reversal.

B Fundamental Analysis Indicators

Table 2: Fundamental analysis indicators. (Becket & Essen 2010)

Indicators	Description	Formula
Acid Test (Quick Ratio)	Measures a companies readily liquidable assets in order to pay short-term debts. A score lower than 1 means that the companies assets are less in value than the outstanding debts. Whereas a score higher than 2 indicates that a companies asset value is double that of the debts, indicating the company is financially secure.	$\frac{(\text{Current Assets} - \text{Net Monetary Assets})}{\text{Current Liabilities}}$
Altman Z-Score	Predicts the likelihood of a company becoming insolvent within the next two years. It uses 5 ratios, each weighted differently to determine a companies ability to pay its outstanding debts on their due dates. The ratios used are: Working Capital (a), Retained Earnings (b), Operating Income (c), Sales (d), Total Assets (e), Net Worth (f), Total Debt (g). The score produced ranges from a value of -4 and 8, with scores lower that 1.8 indicating insolvency is likely, and above 3 indicating insolvency is unlikely.	$\frac{(1.2a+1.4b+3.3c+d+0.6(f/g))}{e}$

Continued on next page

Continued from previous page

Indicators	Description	Formula
Beta	Beta measures the volatility of the share price of a company relative to the rest of the industry. This is ratio of the movement of an individual share relative to the market. A high positive beta value indicates that a share will move in line with the market, at a more volatile rate. A negative beta value indicates the share will move in the opposite direction to other shares.	$\frac{\text{Share Price Movement}}{\text{Market Movement}}$
Current Ratio	Assesses a company's ability to pay its bills by comparing current assets to its current debts.	$\frac{\text{Current Assets}}{\text{Current Liabilities}}$
Dividend Cover (Cover)	The proportion of a company's earnings that are paid to shareholders.	
Dividend Yield (Yield)	This measures a company's return on investment at the current share price and the rate of payments by the company.	$\frac{\text{Dividend per Share}}{\text{Share Price}}$ (%)
Employee Efficiency	The proportion of sales that are paid out in employee wages.	$\frac{\text{Wages}}{\text{Sales}}$ (%)
Gearing (Debt/equity ratio)	Gearing is an indication of a company's risk. It is calculated by comparing the amount of money a company has borrowed with the amount that has been invested by the shareholders (equity).	$\frac{\text{Total Borrowings}}{\text{Total Amount of Shareholders' Funds}}$
Net Asset Value (Asset backing)	Provides a ratio between the net value of all a company's assets after any costs, and the number of shares issued. This provides the shareholders' equity in the company per share. (CA represents current assets, L represents liabilities, and CC represents capital charges)	$\frac{CA-L-CC}{\text{Number of Shares}}$

Continued on next page

Continued from previous page

Indicators	Description	Formula
Net Current Assets	The total current assets of the company minus the total current liabilities. This gives and indication of its solvency in the short term.	Current Assets − Current Liabilities
Price / Earnings Ratio	A measure of how long it will take for a company's profits to reach the total price of its shares. Calculated by comparing the company's profits to its share price.	$\frac{\text{Share Price}}{\text{Earnings per Share}}$
Profit Margin	Ratio of a company's operating profit and turnover to indicate its underlying profitability.	$\frac{\text{Trading Profit}}{\text{Turnover}}$ (%)
Return on Capital Employed	Measures how efficiently a company is using its capital in the long term. A company could have a low return on capital, even if profit margins are high.	$\frac{\text{Trading Profit}}{\text{Average Capital Employed}}$ (%)
Return on Sales	Measures the ratio of pre-tax profit and sales, giving as an indication of profit margins.	$\frac{\text{Pre-tax Profit Before Interest}}{\text{Total Sales}} \times 10$
Return per Employee	Indicates how efficiently a business uses its employees.	$\frac{\text{Operating Profit}}{\text{Number of Employees}}$
Return to Shareholders	Measures the total performance of an equity over a given period. (CSP represents change in share price, D represents dividends, and ID represents interest on dividends)	$\frac{CSP+D+ID}{\text{Starting Price}}$ (%)
Stock Turnover	The proportion of sales cost to the end of year stock level.	$\frac{\text{Cost of Sales}}{\text{Stock Level at the End of the Year}}$
Value Added	Indicates how a business has increased the value of shareholder investment.	