Stock Portfolio Optimisation

ADS2001 Final Report

Group Members	Contribution
Joshua Hudson	Modelling, Results
Nithik Vijayanand	Report, Preprocessing
Joshua Gonzales	Preprocessing, EDA
Ashlyn Tiwari	Report, Conclusion

Executive Summary

Project Brief:

Stock portfolio optimisation is the process of strategically choosing and allocating investments over a range of different stocks in order to maximise returns as well as simultaneously minimise the risks involved. The portfolio in this case, refers to the collection of stocks selected to achieve the highest gains in investing. In order for good stock portfolio optimisation, we must try to diversify our investments across a range of stocks. We can also incorporate mathematical models and algorithms to systematically allocate investments in a way that maximises the portfolio's expected return while minimising its volatility or other measures of risk. Although the return on investments can be measured by the increase in the value of our portfolio, it is not as quite clear on how to compute the risk of our overall portfolio. As a result we investigate some common measures of risks including the variance/standard deviation, value at risk and the expected shortfalls.

Our Data:

The S&P 500 is a stock market index that measures the performance of 500 large-cap U.S. companies listed on the stock exchanges. The data we are dealing with is historical stock price data from the S&P 500 index spanning a period of 26 years from 1993 to 2019, meaning that there are many different companies that have entered, left, and re-entered our dataset over the 26 year time period. We have data for around 1200 companies entering and exiting the index over this time. As a result of the large quantities of companies coming in and out of the top 500 index, we are left with many null values as once a company decreases in performance or ceases to exist in the top 500 index, all remaining dates will have been unaccounted for in terms of the stock price in our dataset. However despite the large amount of null data we were able to subset the data to ensure that we have an accurate set of data to wrangle and model with.

Main Objective:

Our Primary objective is to construct a portfolio that achieves the highest possible return for a given level of risk, or conversely, the lowest possible risk for a desired level of return. We must investigate selecting optimal weights of each asset in the portfolio which achieves our main objective. We need to assess various risk management strategies such as the Sharpe ratio (measures how well an investment performs relative to the risk taken to achieve those returns (Baldridge, 2024)) to determine the optimal portfolio weights.

Findings:

Our findings and analysis conducted on various investment models reveals key insights into the dynamics of stock market investing. It demonstrates that investing is not just about picking stocks, but also about understanding the principles of exponential growth and the importance of time in investing. In this project, we utilised both linear and non-linear models to find the best way to maximise our gains whilst minimising our risks. We found that each of these models have their respective strengths and weaknesses. While the linear regression model showed potential for high returns, it also exhibited high risk due to its volatility. The decision tree model, on the other hand, offered a lower risk option with steady growth. The hybrid model attempted to capitalise on the strengths of both models but showed only slight improvements. The analysis underscored the importance of diversification in managing risk and the impact of market conditions on investment performance. Despite the challenges in predicting stock performance, the models outperformed the S&P 500, highlighting the potential of these models in guiding investment decisions. We found however, that further investigation is needed in order to develop a model that better balances return and risk.

Main Body

Introduction:

Investing is the process of combining multiple assets with the overall objective of generating long term returns. In order to create successful investment portfolios, we must select the combination of the stocks in a way which can both hold minimum amounts of risk as well as reap high returns. The Data that we are given to work with consists of historical stock data in the S&P 500 from the time period 1993 to 2019. Many companies and businesses have entered into and have left the index over this time period leaving us with many columns containing NA values as these companies would have left the index. However once we managed to focus on the time period from 1999 to 2019, most of the companies listed on the S&P 500 remained within the index. Although the stock market is highly volatile and often difficult to predict as there numerous factors affecting the price levels on a daily basis, we are generally able to forecast an overall trend in the stock based on its historical performance. To overcome these challenges, we have tested and compared the performance of linear regression, decision tree, and a hybrid model. Additionally, we have also considered the number of companies to invest in and how it impacts the performance of the model. The main relationships expected to be explored in this project are the correlations between different stocks, the impact of time on investment growth due to the principle of compounding, and the performance of different investment models. We also expect that with generating high returns over a short period of time, comes higher rates of risk associated.

Data Quality & Preprocessing:

The dataset we were provided with was a table comprising all the companies that have ever been in the S&P 500 from September 1993 to July 2019. As a result, instead of having 500 companies to choose from, instead there were 1200 companies present. Furthermore, out of these companies, it would be unclear on whether they were comprised of industry leaders, or other legacy companies that have since gone bust or have fallen out of the index.

As a result, we decided to narrow the dataset we were provided with. Along with the issue of having too many companies in the dataset, there was another factor.

A lot has changed in 36 years. In the time period in which the dataset is in, there have been 2 major recessions; the dot-com bubble, the 2008 global financial crisis. Consequently, the companies comprising the S&P 500 have greatly changed since. For instance, in 1993 the leading sectors in the S&P 500 were Industrial, Consumer Discretionary, and Finance, comprising 94, 73, and 66 companies in the S&P 500 respectively. Whereas, in 2019, the leading sectors were Industrial, Information Technology, and Finance, each comprising 70, 68, and 67 companies respectively (Finch, 2019). Although the 1st and 3rd leading industries in the S&P 500 remain the same, the number of companies in these sectors have decreased. This implies a greater diversity of sectors that now have more companies listed in the S&P 500, as the distribution of companies for each sector in the S&P 500 is more even, rather than being top-heavy.

Therefore, after considering the changes in the S&P 500 from 1993 to 2019, we decided to choose 2000 as our starting point, with the dataset from 1999 up to then serving as the training data. This is due to these two years serving as a great transition period for the changes previously mentioned. During these time periods, the number of companies belonging to the Industrial sector was declining, sitting at 67 companies in the index. On the contrary, the Information Technology sector was booming, sharply rising to 68 companies in the S&P 500 (Finch, 2019).

From there, the actual processing of the dataset was rather simple. We converted the Date column from an integer to pandas' DateTime variable. Subsequently, we set the starting point for the dataset to be from 1 Jan 1999, up to the original endpoint of 31 July 2019. Finally, we removed the companies in this filtered dataset that had any missing values in them, implying that either they went bankrupt, or had shrunk in size to no longer qualify to be in the S&P 500.

This left us with a final dataset that contained 484 companies that have remained in the S&P 500 continuously from 1999 to 2019. This value is much closer to the 500 companies one would expect in the index.

However, a possible limitation of our methodology to clean the dataset is that it might have removed companies that had always been in the S&P 500, but their stock prices have fallen to below the qualifying criteria for a short period of

time. This could be due to regular fluctuations in the stock market, or short-term slumps in the company's stock price.

Despite this shortcoming, our procedure maintains 96.8% of the companies listed in the S&P 500 across 20 years, where it is expected that 20 companies fall out of the S&P 500 annually (Daily, 2023). Thus, the final dataset can be appropriately used to create our portfolio, as it captures nearly all of the companies in the S&P 500 over 20 years that can be chosen in the portfolio.

Exploratory Data Analysis:

Before creating our portfolio, it would be due diligence to first analyse some of the individual stocks present in the dataset, as well as a high-level view across all the companies present. We decided to focus our individual stock analysis on Apple, as it was the second largest company in the S&P 500, comprising 6.19% of the index's portfolio (*S&P 500 ETF Components*, n.d.). Microsoft is the largest company in the S&P 500, comprising 7.02% of the portfolio. However, Apple has had greater growth over the course of the timespan of the dataset.

In 1999, Apple had a stock price of $1.29 a share, and by 2019 had a value of $208.78 each share. Apple's rate of return of approximately 160 can be considered an outlier, as only 3 other companies have returns greater than it. This exponential growth can be primarily seen in 2 time periods. From 2000 to 2007, Apple stock grew from $3.24 to just under $23, and after the 2008 recession, from 2009 up to 2018 Apple stock grew from $10.76 in 2009 to around $200 by the end of 2018.

However, across all 484 companies in the dataset, 430 of them had a positive rate of return, and 117 had rates of return greater than 10. Thus, when it came to selecting companies for our portfolio, it would be fair to expect that their predicted returns would be highly unlikely to be considered neither an outlier nor anomalous.

Subsequently, we began to create a model that would aim to predict stock prices for the companies in the dataset across 20 years. This would then be used to determine which companies' stocks will be used in our portfolio. However, even after considering the acceptance of otherwise anomalous return values, there were some concerns on how this modelling process would be conducted.

Firstly, across the years from 1999 to 2019, there were a couple of time periods where the stock prices for all the companies had extremely high volatility, which would make it difficult to accurately predict them. For instance, Apple stock had extreme fluctuations from 2007 to 2009, and from the start of 2019.

Secondly, we were worried about the runtime of any machine learning algorithm being used on this dataset. This is due to the dataset containing daily stock prices for each company across 20 years, leaving approximately 7500 entries per company.

Lastly, selecting the training data to be used for the model. We originally planned to use 1999's stock prices as the training dataset for our future model, however due to the exponential nature of the stock prices' growth, we were concerned that the model would consistently underestimate the predicted stock prices for the companies in the dataset.

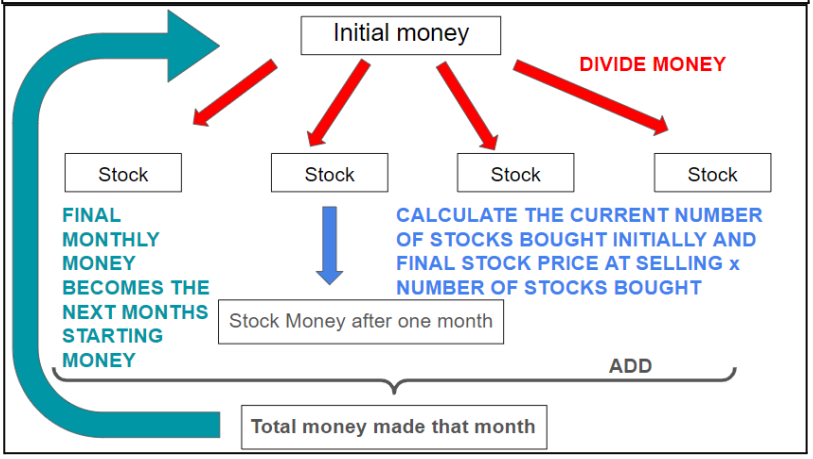
After factoring these issues, we decided to initially use a Linear Regression model to predict the stock prices for each company in the dataset. The relative simplicity of this model allows it to be quicker in running predictions for each company. We then decided to amplify the runtime of the model, by choosing to predict a company's stock prices monthly, instead of daily. Consequently, this reduces the number of predictions to be made from 7500 to nearly 200. However, we chose to maintain the training dataset's daily entries, to maximise the accuracy of the model. Furthermore, we chose to reset the training dataset after each iteration of predictions, by shifting its starting point by 1 month, and setting its endpoint to the model's starting period, while maintaining its daily entries. This helped reduce runtime even more, as it ensured that the size of the training dataset remained constant instead of being cumulative, along with weeding out older entries that might cause the model to underestimate its predictions, thus improving accuracy and reducing the impact of fluctuations in the data.

These considerations made in the model achieved notable results. For Apple, this model returned a correlation coefficient (R^2) of 0.98 across 20 years, even despite the mentioned fluctuation periods. And across all 484 companies present in the dataset, over 300 of them had R^2 scores greater than 0.95, and the average R^2 score being 0.92. Thus, we proceeded to use this model to choose our portfolio.

Modelling:

The goal of the modelling process is to create a model that will predict a stock's opening stock price for the next month, based on the previous opening data. Using the time period of 2000-2019, a training and testing period, we can evaluate how well different models are performing, and hypothetically, if these models were used from the start of the 2000's how would they perform. To do so, this involved a three step process: creating a function that will invest money and keep track of the current money made, a function that will choose a number of stocks based on the predicted output stock price for the next month, and the stock prediction function. Figure 1 is a diagram showing the process of the portfolio.

Figure 1: Diagram of investing cycle



For our modelling we initially started with a total of $5000, and the goal was to find a model that would make the most money within this period. Initially started by splitting this $5000 evenly between stocks that had chosen to invest in. After this, the money for each stock will be invested in the company by dividing the amount of money by the current share price. The value obtained from this is the number of shares has now been bought. This value multiplied by the actual stock price in 1 month's time gives the total money, if you selled, for the next month. Since for our portfolio we want to change the stocks chosen each month, the model takes the money out for each stock at this time point and then adds all the stocks money together to give a new total initial money. This investing process is then repeated for the next month, but instead of a starting total of $5000 it would be the total money after the last month. This aims to simulate what real life investing would look like. However, when investing in the S&P 500, buying stocks would usually pay an investment fee to a brokerage company (the middleman who buys and sells the shares for you). One of Australia's most popular brokers is Commsec, who charges 0.12% for investments into the American share market, the S&P 500 included (Commsec, *Rates & fees*). Implementing this the investing process for the portfolio has been created.

The next function is a function to choose the stocks that have been predicted to increase the most over the next month. This is done by taking the stocks that have changed the most based on percentage of growth from their initial stock price to their predicted stock price. This must be used instead of the maximum value, as different stocks have a different starting price. For example, a stock changing from $1 to $1.5 is far more significant than $50 to $50.5 even though the rise in their stock price was the same. Ranking the stocks based on their percentage change now allows for the stocks to be chosen and added to the portfolio to be invested in.

The most important part of this process is how to predict the future stock prices. Based on the exploratory data analysis, the stock price appears to be non-linear, however multiple methods were explored. For all the models, the models will train off the past 1 year of the individual stocks prices and then output a predicted stock value for 1 month after the investment date. This is then repeated for all of the stocks and then the predictions can be used to choose the best stocks to invest in for that month. An initial problem was found when using more than one year's worth of data. This caused overfitting and overall led to a volatile and high riskportfolio as there were many inaccurate predictions, thus the 1 years worth of data restriction was used. Overall, we used 3 models total, and each model explored the influence of time and the number of companies chosen to invest in. The 3 models were: Linear regression, decision tree regression, and a hybrid of the two. The performance of all the models will be explored in the Results section.

Results:

In order to measure the performance of each model it is necessary to understand how investing works. The investing that this portfolio relies on to grow is ‘exponential growth’ or ‘compounding’ . This is because of the formula, total_money = starting_money * (monthly_growth) ^months. This formula shows that the longer you hold your investment the more money you will make. This is highlighted in all the graphs highlighting the amount of total money, based on the number of months. However this method of measurement does not tell you how the model has performed each month. Taking inspiration from the Sharpe Ratio, a commonly used measurement for determining the performance of a portfolio, the method of measuring the monthly portfolio growth as a percentage and the monthly portfolio standard deviation can be used to counter

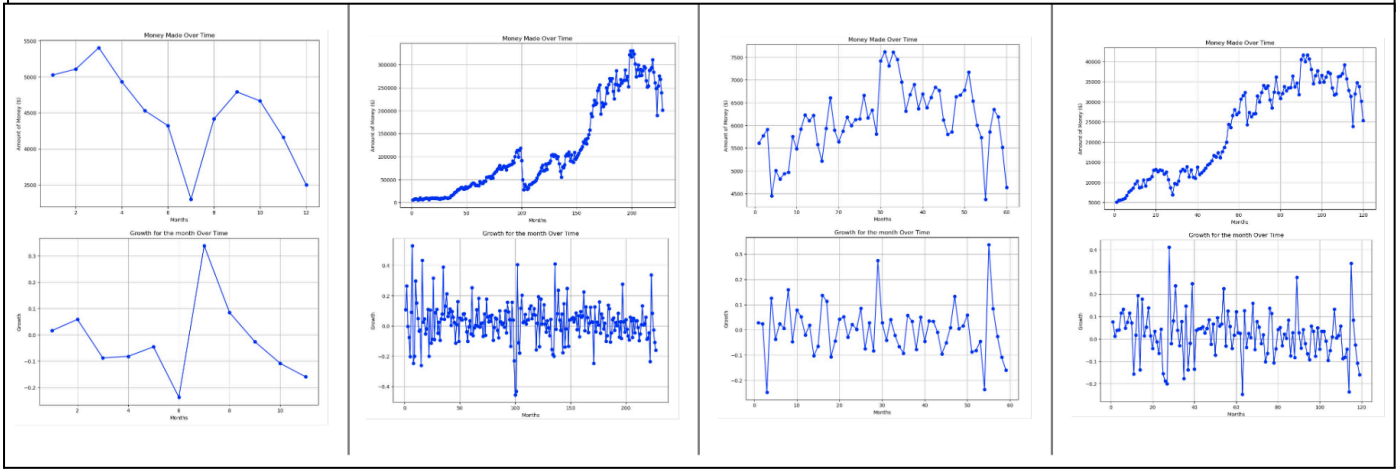
the exponential growth and analyse whether the model is truly performing well. The average monthly growth of the portfolio can help indicate how much money on average you expect to gain each month if you use this portfolio, and the standard deviation of the monthly growth indicates the volatility and risk of the model. A high standard deviation can indicate that the model may not perform very well consistently, or it may be because of market crashes. The main goal is to have a portfolio that can outperform the S&P 500 over the 19 years chosen. As a baseline between January 2000 and June 2019 the S&P has increased from 1469.25 to 2766.149902. Figure 2 shows the S&P 500 over the last 19 years. The data set was from Kaggle and can be found in the Bibliography. Meaning you would have 1.88 times your money during this period. The S&P 500 during this period of time had an average growth rate of 0.2% of growth.

Figure 2: S&P 500 Index over time

For the Linear Regression model, it was initially predicted that it will not perform well. This is because of the unpredictability of stocks and their non-linear nature. This model was tested over a number of circumstances. It was tested over a period of 1, 5, 10 and 19 years, to see how it would perform. It was also tested to see if the number of companies chosen to invest in would also impact the performance of the model. The number of companies chosen to test was 3,5,10, and 20. Figure 3 displays the stock price and growth for different time periods. The first thing that is noticeable from all the graphs for the period of 1 year is that they do not perform well. All the models had negative growth, and a high standard deviation. This is likely due to the overall stock market crash in 2018-2019 because of COVID-19. This meant that the majority of companies stock prices ended up decreasing and this Linear Model is unable to pick stocks that would not decrease. However when looking at the longer term results of the portfolio's average growth for 10 and 19 years, it performs a lot better in nearly all the years except 2019. Looking at the results, having 5 companies chosen seems to be the optimal amount, due to having the highest average monthly growth of 2.39%. However the one downside of this model is the standard deviation of the growth. The Standard deviation being 12.9% is massive, and can be seen in the accompanying graph. This means that this model is very high risk and has the potential to make you a lot of money some months but also loose.

The next best number of companies was 10, however, as seen in figure 4, it has much lower growth of 1.7%, but a bit better standard deviation of 10.8%. The overall performance of this model is much better than the S&P 500 in terms of growth, however its high volatility leaves room to improve. If someone was to use this portfolio it would be classified as a very high risk portfolio.

Figure 3: Total money made, and monthly growth for 5 companies, over different time periods



The next model that was run was a decision tree model, with the aim to reduce the variability and volatility of the previous linear model. Once again the optimal number of companies chosen was 5, with 10 being slightly inferior. The first significant difference between the two models was the growth of the portfolio over the 1 year time period, as seen in figure 5. Although the model still had an average negative growth rate, the standard deviation was much lower than the linear counterpart. This is further highlighted by the graph showing the growth of the portfolio from month to month. Comparing the two, we see that the growth is a lot less volatile, but also the magnitude of the growth is a lot less. This is both good and bad for the model. It means that unlike the linear model

Figure 5: 5 companies, decision tree total money and monthly growth over 1 year

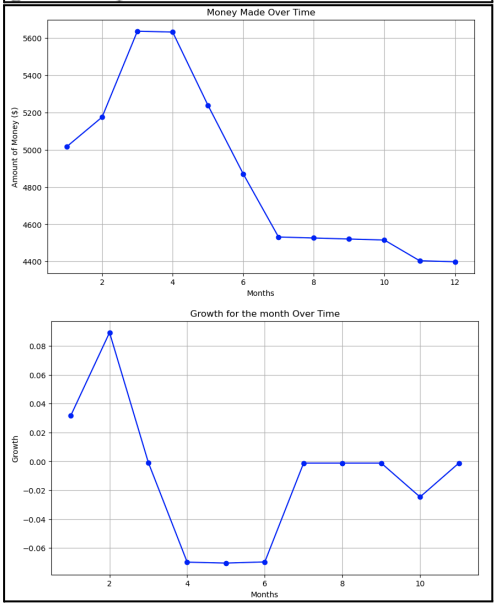
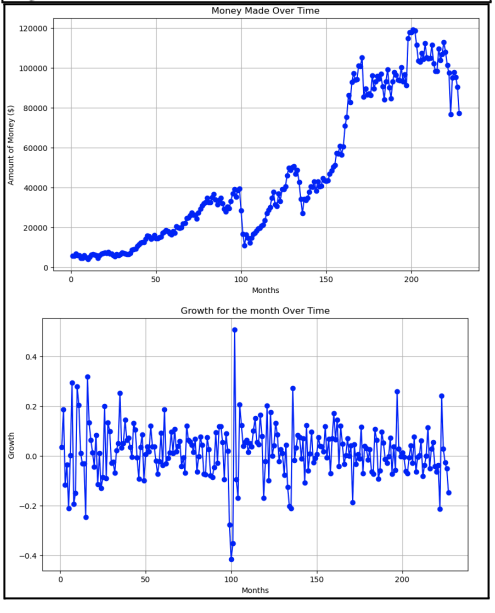
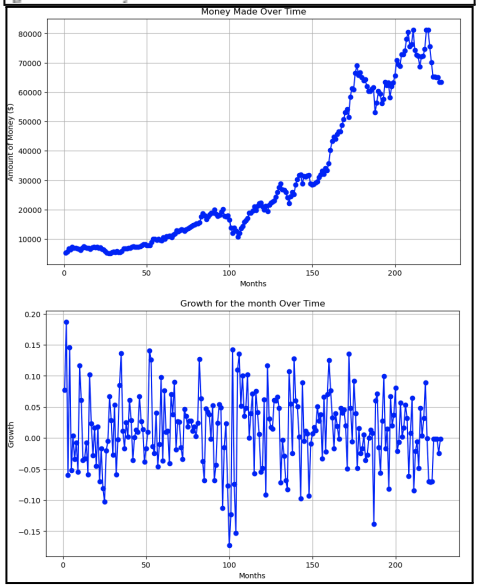


Figure 4: 10 companies total money and monthly growth over 19 years



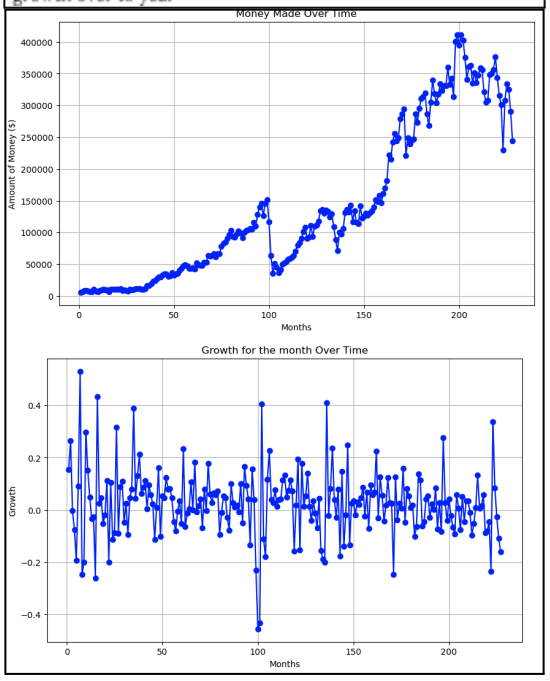
there was no significantly high growth, but there is a lower possibility that you will lose all your money in a month when the market crashes. This pattern continues for the whole 19 years, as shown in figure 6. This is highlighted in the total money made over time steadily increasing exponentially, as expected. The model over the 19 years produced an average growth of 1.275% and a 5.9% standard deviation. Although there is a significant drop in the average growth, the average standard deviation has also fallen significantly. The performance of this model still outperforms the S&P 500 over this time period, however the standard deviation is still too high to call it a low risk option. However relative to the Linear model it is a much lower risk option.

Figure 6: 5 companies, decision tree total money and monthly growth over 19 year



The next model that was built was a hybrid of the two models, that tries to capitalise on the previous models strengths. This was done by using the linear model as a base for the prediction of the stocks, however if the non-linear model predicted the same stock the portfolio would double down and invest twice the money into this stock. The aim of this was to help mitigate the standard deviation and capitalise on stocks that had a higher likelihood to increase, as both models predicted that stock. The downside to this model was seen with only a slight improvement to the models performance, with an average monthly growth of 2.48% and standard deviation of 12.9%, as visualised in figure 7. The main issue with this model was that both models rarely selected the same stocks, and even if the number of stocks to select was raised to 20, they would still only select 3 common companies every few months.

Figure 7: 5 companies, hybrid model total money and monthly growth over 19 year



Overall, these models have highlighted that by investing over a long period of time and reinvesting your money consistently over that time period you can exponentially grow your money. The models also highlight that predicting the right stocks can be very difficult. Obviously, if it was an easy task it would be a get rich quick scheme and many companies such as Vanguard and Commsec have invested millions in trying to find the best

way to build a portfolio and make consistent returns at minimal risk. The volatility of the stock market is also highlighted, and with just the stock prices of companies it proves very challenging to predict how they will change. A company's stock price is influenced by hundreds of factors, and with just the stock prices it limits the ability to predict a stock's future value. The models have proven to outperform the S&P 500, however they are far from perfect. Out of the portfolios presented we would recommend the non-linear model because of its lower risk and higher chance of performing well during market crashes.

In the future to improve this modelling, it is recommended to explore other methods of predicting the stock prices, such as neural networks. The expansion of the data set to not only include the opening stock prices, but also other information such as market cap, can allow for the models to build a greater understanding of the current economic conditions of the company and hopefully be able to predict the changes in stock prices more accurately. Many big financial companies also use web scraping to find news, information and public perception of a company in the hopes of being able to predict a stock's value. On the contrary a different approach to creating a portfolio of stocks could be to choose a range of diversified stocks based on certain characteristics such as how their stock prices have deviated and what type of business they are. This approach could be approached in the future and see how it performs against the portfolios previously presented.

Conclusion:

In conclusion, this project aimed to optimise a stock portfolio by strategically selecting and allocating investments across different stocks within the S&P 500 index, utilising various predictive models. By analysing historical stock price data from 1993 to 2019 and focusing on the period from 2000 to 2019, we sought to construct a portfolio that could generate the highest returns with an acceptable level of risk. We began with a dataset of historical stock prices for companies that have been part of the S&P 500 at various points between 1993 and 2019. To ensure data quality and relevance, we filtered the data to include only companies with complete records from January 2000 to July 2019.

Our modelling process involved three main steps: creating an investment function to track and reinvest money, developing a stock selection function

based on predicted stock prices, and employing predictive models to forecast future stock prices. We experimented with three models: linear regression, decision tree regression, and a hybrid model combining both approaches. From this we were able to assess their performance based on average monthly growth and volatility. The linear regression model provided higher returns but with significant risk, while the decision tree model offered a lower average monthly growth rate, but significantly reduced volatility. The hybrid model attempted to combine the strengths of both but showed only marginal improvement.

Our results indicate that the concept of exponential growth or compounding is fundamental to investing. Thus, the longer you hold your investment, the more money you will make (Brock, 2024). Furthermore, the models have shown that there's a trade-off between risk and reward. Higher potential returns often come with higher risk. Therefore, understanding risk tolerance is crucial in making informed investment decisions. The importance of diversification across various sectors or products is important in reducing the different types of risk. This involves investing in a variety of stocks to minimise the risk associated with any single investment. A well balanced mix of stocks can enable a portfolio to grow with significantly less risk and volatility (Cussen, 2024). Hence shown through our analysis of the linear regression model, decision tree model, and the hybrid model, revealing how the number of companies chosen to invest in impacts the performance of the model.

Additionally, the performance of the models can be influenced by market conditions. For example, the models may not perform well during market crashes, as seen in 2018-2019 due to COVID-19. This highlights the importance of considering external factors and market conditions when making investment decisions.

Overall, the analysis has shown that predicting the right stocks and the volatility of the stock market is challenging. The models have proven to outperform the S&P 500, however there is still room for improvement. Out of the portfolios presented, the non-linear model was recommended because of its lower risk and higher chance of performing well during market crashes. Our findings thus highlight the importance of understanding the principles of investing, assessing risk and reward, diversifying investments, considering market conditions, and continuously learning and adjusting investment strategies.

Bibliography

Baldridge, R. (2024, February 27). *Understanding The Sharpe Ratio*. Forbes Advisor. https://www.forbes.com/advisor/investing/sharpe-ratio/

Brock. (2024, January 30). *Risk versus reward // The Motley Fool Australia*. The Motley Fool Australia. https://www.fool.com.au/investing-education/introduction/risk-reward/

Cussen, M. P. (2024, January 12). 7 *Simple Strategies for Growing Your Portfolio*. Investopedia. https://www.investopedia.com/articles/basics/13/portfolio-growth-strategies.asp#toc-3-diversification

Daily, I. B. (2023, September 6). *More than a third of s&p 500 stocks get kicked out in nine years*. Investor's Business Daily. https://www.investors.com/etfs-and-funds/sectors/sp-500-stocks-more-than-a-third-get-kicked-out-in-nine-years/

Finch, C. (2019, October 3). *History of companies and industries listed on the s&p 500*. *QAD Blog*. https://www.qad.com/blog/2019/10/sp-500-companies-over-time

Han, H. (2020, November 5). *S&P 500 historical data*. Kaggle. https://www.kaggle.com/datasets/henryhan117/sp-500-historical-data

Mtrucc. (2024, April 2). *Trading*. Kaggle. https://www.kaggle.com/code/mtrucc/trading

Rates & fees. CommSec. (n.d.). https://www.commsec.com.au/support/rates-and-fees.html