Analysis of the S&P 500

Ronan Flannery

GitHub URL

https://github.com/rdflannery/UCDPA_rdflannery

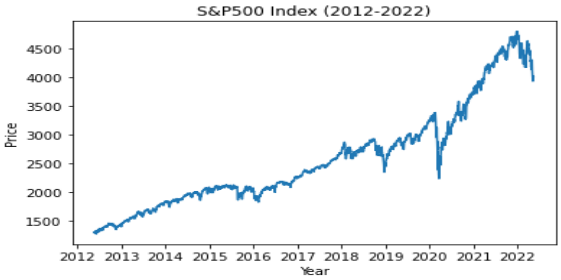
Abstract

This project investigates the performance of the Standard and Poor's 500 (S&P 500) index during the period 2012-2022. The S&P 500 is a stock market index, incorporating the 500 largest companies (by market capitalization) listed on the stock exchange in the United States. Stocks in the S&P 500 make up about 80% of total US equities by market capitalization, hence the S&P 500 is considered a good indicator for the overall US stock market.

This project aims to implement techniques learned throughout the course to analyze the performance of the S&P 500 over the time period 2012-2022. Focus is put on analysis of the index returns, the individual returns of the top 10 companies and sector returns. Finally, a deep dive into a single stock, Apple Inc., to investigate its return profile and the possibility of using machine learning techniques to predict Apple's future stock price.

Introduction

This project case was chosen due to recent employment in the investments team of a leading global life insurer. Equities have performed excellently for a considerable amount of time since the recovery from the global financial crisis of 2007-2008. Stocks have recovered from the considerable blip in March 2020 on the outbreak of the coronavirus pandemic and continued to press forwards right up until 2022. Many analysts consider this run to be technology driven due to the much publicized performance of large cap tech stocks such as Apple, Amazon, Facebook, Google etc.



However, since the beginning of 2022, there has been a dip in global financial markets for a host of reasons. Rising inflation rates are expected to be dampened by central banks increasing interest rates which is typically linked to poor equity returns. However, there are plenty of other factors at play such as global supply chains still being affected by coronavirus, most notably in China, the price of oil and war in Ukraine. The result of all this is great uncertainty in the equity market for the coming months ahead.

Data analytics can bring a deeper understanding into such complex situations and hopefully be useful in predicting future movements in equity markets.

Dataset

The datasets used were sourced from a number of different locations. The main datasets, consisting of the S&P 500 index prices, the S&P 500 constituent company prices and companies information was sourced from Kaggle at the following link:
https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks

Companies information was also web scraped from Wikipedia at the following url:
https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

Finally, the **yfinance** package for python was used to source data from the Yahoo finance API.

The reliability of the Kaggle datasets and that sourced from the **yfinance** were cross-checked against Yahoo finance to confirm the suitability of the datasets. All datasets were accurate.

Implementation Process

Data Import & Review

The datasets were loaded into variables in python as dataframes via csv files, web-scraping and pulling data directly from the Yahoo Finance API using the **yfinance** package. Summaries of the variables were reviewed, with null entries being omitted from the stocks dataframe where there were no stock prices for certain dates. There was no such occurrences in the index data.

A comparison of the companies information from the Kaggle dataset versus that scraped from Wikipedia showed that all the data in the Wikipedia table was also contained in the Kaggle data as well as further data. Hence, it was decided to proceed with the Kaggle dataset.

S&P 500 Index

The S&P 500 index daily returns were calculated and a histogram of returns produced. This histogram was plotted along with a generation of normal returns, produced with the mean and standard deviation calculated from the S&P 500 index dataset. Also, the mean and standard deviation of the index was plotted. The large spike in the standard deviation in March 2020 for the outbreak of the coronavirus global pandemic gave rise to another plot, featuring a comparison of the S&P with the VIX volatility index.

The VIX data was pulled from Yahoo finance. Initially, the plot consisted of both sets of data with two y-axis and a single x-axis. However, this was deemed unclear where the data overlapped and the solution being to use sub plots to better display the relationship between the movements of the index and the volatility index.

Constituents of S&P 500

The stocks data file consisted of daily prices from 2012-2022 for each of the constituents of the index. As such the file was quite large, and hence the work completed would not have been repeatable in MS Excel due to dataset size limitations.

A **for** and **if** loop was utilized to calculate the daily returns of each of the stocks. Using the companies table, the top 10 stocks by market capitalization were identified and merged with the stocks data using an inner join. A correlation matrix was computed of the correlation between the daily returns of these top 10 companies in the S&P 500 and plotted using the **seaborn** package.

Sector Analysis

The average daily return for each company within the various sectors of the S&P 500. The data was plotted in a bar chart to compare the performance of the sectors over the time frame.

Plotting stock prices

A custom function was defined called **stock_plot**. The aim of this function was that with one input, the string of the company ticker, the closing stock prices would be pulled for the company and plotted over time. This was tested for Apple, Google, Microsoft and Amazon.

Apple price moving averages

A single stock, Apple, was chosen to focus on at an individual level. The moving averages of the Apple close price for 50-day, 100-day and 200-day were plotted against time.

Price Prediction – LSTM Model

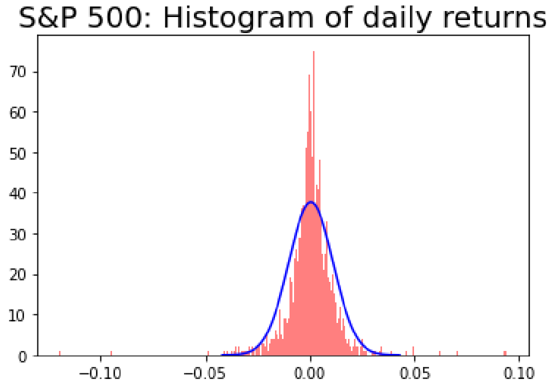
The inclusion of a machine learning element in the project is done using Long short-term memory (LSTM), which is a recurrent neural network. The code used is referenced in the ‘References’ section as it is not my own. This method is a popular choice generally for machine learning, but also particularly for stock market price prediction. Hence, it was decided to use other code to perform LSTM rather than perform a more simplistic modelling approach which may not be practically suitable, such as a linear regression model.

Further investigation into machine learning methodologies is required, specifically the LSTM approach used, and how they can be applied to financial data to predict future movements.

Results

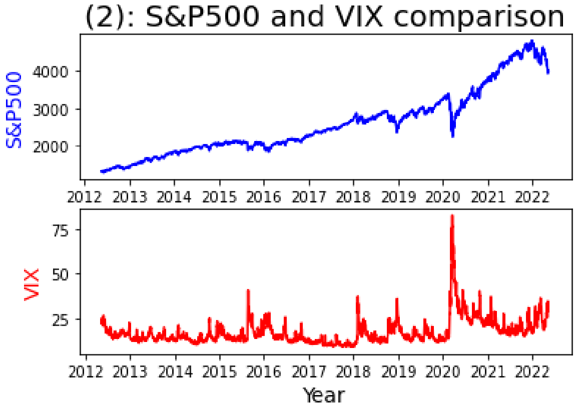
S&P 500 Index

The histogram of the S&P 500 returns confirms the literature that financial returns are not normal as the distribution tends to be leptokurtic and have fat tails. One can see from the histogram below the incidences of returns in the tails far beyond the four standard deviations of the normal distribution shown in blue.



The comparison of the S&P 500 prices and the VIX volatility index shows how they are linked over time.

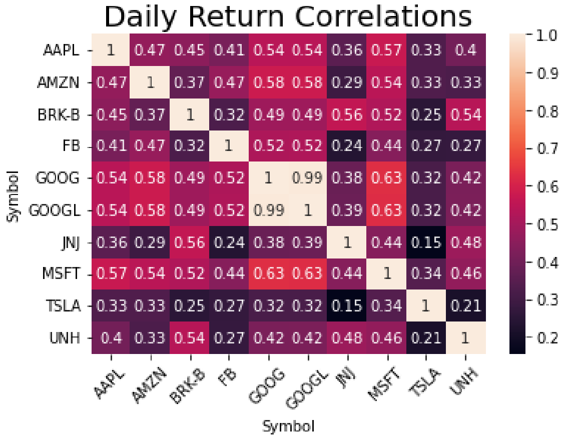
The S&P 500 grew steadily until 2020, with some minor spikes and drops along the way. In March 2020 there was a large downturn. This coincided with a large spike in volatility according to the VIX index. Volatility remained higher throughout the pandemic period which corresponds initially with strong growth in the S&P 500 and then followed with a drop from the start of 2022.



Top 10 constituents of the S&P 500

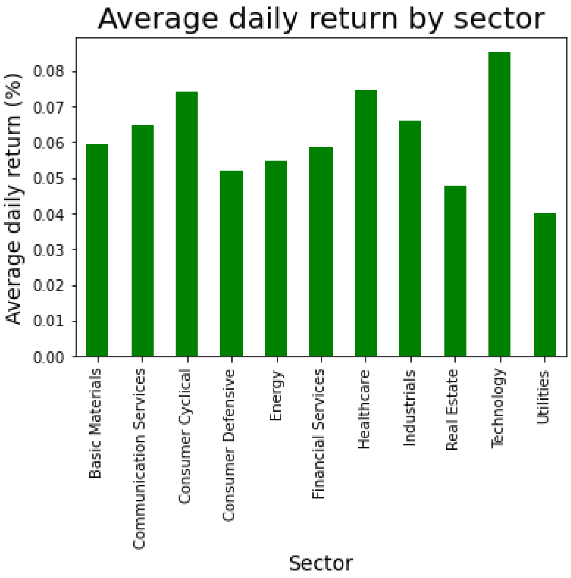
The correlation matrix of the top 10 companies (by market cap) of the S&P 500 shows a number of interesting points:

1. Some of the largest correlations occur between stocks of similar industries e.g. Google – Microsoft, Google – Amazon, Microsoft – Amazon.
2. The largest correlation is between “GOOG” and “GOOGL” which are actually share categories of Google stock. As such, a high correlation is expected.
3. The lowest correlations occur between EV maker Tesla and the health multinationals, Johnson & Johnson and United Health Group, which makes sense given the vastly different enterprises.



Sector Analysis of Returns

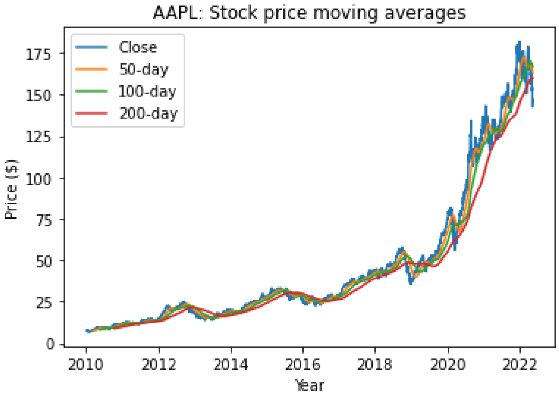
The top 10 constituents of the S&P 500 is dominated by technology stocks as per above. The sector returns graph below shows that technology was the highest performing sector. Other notable sectors of high performance were Healthcare (e.g. J&J, UHG) and Consumer cyclical (e.g. Amazon, Tesla, Home Depot). Healthcare stocks like Pfizer have performed higher during the pandemic years off the back of huge investment in vaccines. This effect would also be visible for stocks such as Amazon and Home Depot which soared during the pandemic throughout the various lockdowns.



The worst performer is the utilities sector. Utilities typically require large expenditure on infrastructure and so companies tend to carry a lot of debt which makes them less attractive to investors.

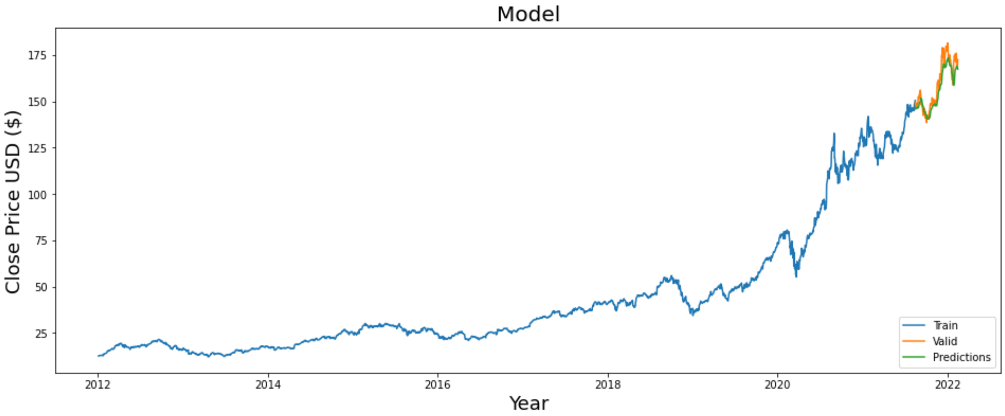
Apple Inc.

The Apple stock price has an exponential like trajectory from 2018-2022. Moving averages are used to identify the trend direction of a stock. The longer the time period for the moving average, the greater the lag. Traders typically eye the 50 and 200-day lags. The plot below shows the 200-day (red) trending upwards with less deviations than the 50-day.



The moving averages above show the general trend of the Apple stock price upwards. Use of moving averages might be more suitable for the buy and hold investor rather than those trying to more accurately predict price movements in the shorter term.

The LSTM neural network approach is one such example. Below, the train in blue relates to the training period used. The predictions in green are the predicted movements and the valid in orange is the actual stock price movements. It is clear the predictions are very accurate relative to the actual movements.



In saying this, this is just one instance of the model run. A Monte Carlo approach could be taken to run the model numerous times and calculate the average predictions.

Insights

A number of insights are as follows:

1. Financial returns are proven to not be exactly normal as the distribution is leptokurtic with fat tails.

2. Top 10 stocks within the same sector are moderately correlated (0.4 – 0.6). Lowest correlations between top 10 stocks occur between the technology and healthcare sectors.

3. Technology, healthcare and consumer cyclicals have been the best performing sectors in the S&P 500 over the past decade.

4. Moving averages of returns are useful in identifying the longer dated trends in stock prices proving suitable for buy and hold strategies.

5. Given the nature of financial markets, if price prediction methods such as the LSTM can be developed accurately, there is the potential for large financial gain!

Python is highly efficient and deals with large quantities of data with ease. The wide variety of available libraries make data analysis very user friendly.

References

Stock market price prediction using LSTM
https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm