ADS2001

Stock Portfolio Optimisation

Prepared by:

Tashvin Ramesh 34675280

Regine Chong 34901167

Dayvin Rushil 33589070

Fatimah Ayub 34602291

Lim You Wei 34069518

Daniel Ong 34897887

Kalubowilage Niduli Sudewna Alwis 34144021

Table of contents

Executive Summary

The Problem

Stock prices are the share price of companies depending on how well they are doing. This fluctuation of prices allows investors to possibly gain or lose money over time. A stock portfolio represents the different amounts of stocks that a person owns. This means that a portfolio groups the various stock equity together. This is mainly done so that investors are able to spread out their capital over different companies which suits their investment goal or strategy. In this project, an optimized stock portfolio needed to be developed from US Equities. This tailor made portfolio would give potential investors a sense of understanding of which stocks and how much of each stock would be worth investing in. Hence,the four key questions that were planned to be answered in this project were:

1. What patterns and insights can be uncovered from historical price data?
2. How can investors use past performance to predict future trends and risks?
3. In what ways does analyzing past data help in building stronger, optimized portfolios?
4. How can different investor needs be addressed through data-driven strategies?

The Data

The dataset provided contained 9459 rows and 1200 columns where numerical and categorical data were recorded. The data types included strings, time series and floats. The data represented 1199 US Equity stocks and their recorded stock prices for each unique stock equity which spanned over 26 years (1993 - 2019). However, just by viewing the data set, there were quite a number of NaN values which were missing values. By getting the sum of the missing values, there were close to 4 million missing values. To avoid potential issues in the development of the project, this was addressed by using a forward and backward filling method which imputes the missing values based on the previous and next seen values respectively. This helped to fill all the missing values that were present in the dataset. Feature engineering was carried out so that annual average stock prices for each US Equity. This made the data more interpretable and suited for machine learning analysis .

Exploratory Data Analysis

This was the initial stage of data visualisation where trends and patterns were uncovered from the data. This stage helped to understand how stock prices changed in behaviour throughout the 26 years using different visualizations like bar charts, cumulative graphs and time-series plots. These visualizations included obtaining the top 10 most performing stocks, uncovering trends using cumulative growths, analyzing how the stock prices change over time using time-series data and observing stock volatilities to understand stock behaviour. This phase was crucial as it allowed the key characteristics of the data to be understood before applying any formal modelling techniques.

Findings

Eight different modelling techniques were used to forecast the stock prices for the top 10 strongest stocks for 2020 to 2025. Regression analysis using Linear, Ridge, Lasso and ARIMA regressions alongside Decision Tree, Random Forest and LSTM models were used for this analysis. All the regression models showed that only the Amazon stock outperformed the rest in terms of predicted price. However, the Decision tree and Random forest models failed at forecasting as there were stagnant lines shown in the outputs of these models. After the application of each model to the dataset, it was found that Ridge regression performed the best in terms of R^2 for eight of the stocks, whereas Lasso regression performed well for two stocks and Linear regression, however, performed poorly on all of them.

The process of building the portfolio included using Ridge Regression to calculate the predicted returns over the forecast period. Furthermore, a heavy part of the portfolio was based and built upon the EDA where stock volatility for the top ten stocks were converted into percentages. These stocks were then categorized into 3 different categories which were Low, Medium and High risk based on their volatility. 4 stocks fell in each High and Low risk categories with 2 falling under the Medium risk category. Based on this, Low-Risk and High-Risk Portfolios were made. The Low-Risk Portfolio showed the amount of allocation of capital that should be invested in the stocks for long-term growth where minimal capital was placed into high volatile stocks. The High-Risk Portfolio on the other hand prioritizes investment growth with higher exposure to stocks with higher volatility. This could potentially lead to a higher predicted yield of investment for investors seeking for this.

Introduction

What is a Stock Portfolio?

Stocks, also known as shares or equities, are a measure of ownership for an institution or an individual. When someone buys a stock, they essentially purchase a small share of a particular company. In today's dynamic financial landscape, making sound investment decisions are both a challenge and a necessity for investors looking to achieve financial security and long term growth. Investment portfolios on the other hand are a collection of different financial assets including stocks, bonds, commodities or cryptocurrency and are pooled together to guide potential investors to manage risk and increase returns. For this project, a stock portfolio optimisation was developed from many different US equities so that investors would have a general idea on the stocks that would be worth investing in and those that were not.

The Problem Statements

The central issue addressed in this project pertained to how stock prices could be leveraged in order to make informed forecasting strategies tailored to different skilled investors. The stock diversification is classified based on low-risk, medium-risk and high-risk investors. These three investor classes have different risk-reward trade-offs, so careful decisions would have to be made for these three classes. The four key questions that were discussed below formed the foundation of the analytical approach needed for the rest of the project:

- What patterns and insights can be uncovered from historical price data?
- How can investors use past performance to predict future trends and risks?
- In what ways does analyzing past data help in building stronger, optimized portfolios?
- How can different investor needs (high growth vs. low risk) be addressed through data-driven strategies?

By answering these questions, it would provide a solid foundation to the data analysis, which would certainly aid in decision-making strategies for investors who seek to align their risk tolerances with their financial goals. Furthermore, this project emphasizes the importance of model selection and evaluation in financial forecasting, ensuring that the portfolio recommendations for investors are as robust and clear-cut as possible.

Data Quality

Data Overview

The dataset provided in order to carry out this project was a comma separated values (CSV) file which was named “adjprice.csv”.

	Date	0111145D US Equity	0202445Q US Equity	0203524D US Equity	0226226D US Equity	0376152D US Equity	0440296D US Equity	0544749D US Equity	0574018D US Equity	0598884D US Equity	...	YNR US Equity	YRCW US Equity	YUM US Equity	YUMC US Equity	ZE Equ
0	19930907	13.2719	13.6829	8.4429	8.1042	11.000	57.3245	17.8887	6.8315	28.1246	...	NaN	144439.5121	NaN	NaN	Na
1	19930908	13.3263	13.5315	8.2147	7.9590	11.000	57.2096	17.8064	6.8315	27.5051	...	NaN	143691.1208	NaN	NaN	Na
2	19930909	13.7070	13.3800	8.7852	8.0627	11.125	59.1625	17.6831	6.8315	27.7529	...	NaN	143691.1208	NaN	NaN	Na
3	19930910	13.3807	13.4810	9.4127	8.0368	11.125	59.6220	17.6420	6.8773	27.5051	...	NaN	145187.9033	NaN	NaN	Na
4	19930911	13.3807	13.4810	9.4127	8.0368	11.125	59.6220	17.6420	6.8773	27.5051	...	NaN	145187.9033	NaN	NaN	Na
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9454	20190727	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	3.3500	114.02	45.31	134.
9455	20190728	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	3.3500	114.02	45.31	134.
9456	20190729	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	3.1800	114.10	45.43	134.
9457	20190730	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	3.1800	113.24	44.00	136.
9458	20190731	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	3.1800	113.24	44.00	136.

9459 rows × 1200 columns

Figure 1: Data frame of provided data when read in Python

As seen in Figure 1 above, the dataset contained 9459 rows and 1200 columns with categorical and numerical data included. The data types that are observed in the dataset included strings, time series and floats. The string data type could be seen on the top columns which listed the various 1199 US Equity stock names which categorised the data accordingly. The time series data could be seen on the first column in Figure 1 where it was formatted in YearMonthDate. This categorised the data collected according to the time for each US Equity. Next, the floats could be seen as the data collected from the US Equity at a certain period of time, where it ranges according to the price of each unique equity. Furthermore, Figure 1 had shown that the data contained many NaN (Not a Number) values throughout the bottom half of equities at a certain time.

Data Cleaning

In order to find the amount of missing values in the dataset, the code of isna().sum() was used. 3920447 missing values were present and needed to be filled as it could potentially cause issues in the development of machine learning. Emmanuel et al. (2021) further states in their journal article that it is important that handling of missing values is dealt with before using data for analysation as it could lead to a misinformed or biased analysis..

```
stock_data.fillna(method = "ffill", inplace = True)
stock_data.fillna(method = "bfill", inplace = True)
```

Figure 2: Forward and Backward Filling method code

Figure 2 above shows the code of forward and backward filling which was done on the dataset to fill in missing values where stock_data was the variable name given to the data frame. This method imputes values based on the last seen value and the next seen value respectively (Ribeiro & Castro, 2022). This method was used as it is a suitable method for data which has a time series category and it preserves the pattern seen in the data instead imputing the mode of the values as it could potentially disrupt the variability of the data. According to Alwateer et al. (2024), forward and backward filling helps in preserving stability in trends which then reflects observations better.

Feature Engineering

Since the initial dataset consisted of daily stock prices for 1199 companies over the years 1993 to 2019, interpretation and accurate forecasting were made challenging. The massive volume of data incurred risks of overfitting, high noise, volatility, and the influence of short-term market events. To address this, the daily stock price data was transformed into annual averages for each company. The dimensionality of the dataset was reduced, and irregular daily price fluctuations were smoothed out. Consistent long-term trends in stock performance were better captured as a result.

Date	0111145D US Equity	0202445Q US Equity	0203524D US Equity	0226226D US Equity	0376152D US Equity	0440296D US Equity	0544749D US Equity	0574018D US Equity	0598884D US Equity	0772031D US Equity	...	YNR US Equity	YRCW US Equity	YUM US Equity
1993	12.630372	14.232747	7.918696	8.426668	10.229526	56.872964	16.184505	7.209174	26.616612	4.348300	...	24.9203	143881.653562	4.03460(
1994	11.676695	20.149318	8.480758	5.899327	9.375685	70.788659	15.691691	7.059332	23.419443	4.348300	...	24.9203	133566.383317	4.03460(
1995	12.198784	26.912217	10.797010	5.900133	12.665068	92.839909	18.635049	7.287393	22.368739	4.348300	...	24.9203	107292.926265	4.03460(

Figure 3: Annual average stock prices for 1199 companies from 1993 - 1995

The resulting dataset now consisted of 26 annual average prices for each of the 1199 companies from the year 1993 - 2019 where Figure 3 above shows the example of the yearly average for the stocks from the year 1993-1995. Through this process, the data was made more interpretable and better suited for subsequent machine learning analysis.

Exploratory Data Analysis (EDA)

Choosing Top 10 Performing Stocks

Among the 1199 companies, the top 10 performing stocks were determined and chosen from them. There were two approaches to determine stock performance which were either determining the annual average stock prices or the average growth of the stocks over the 26 years. Both methods were used to come up with different insights and they were compared to see which would be the better of the two for further analysis.

Firstly, the annual average stock price for each company was calculated. The reason for including this in the analysis was that there would be a fairer comparison between the stocks of varying volatilities or market volume. This aided in analysing overall trends and figuring out which stocks reached their peak price and which stayed roughly stagnant. Figure 4 showed the top 10 stocks in terms of average price:

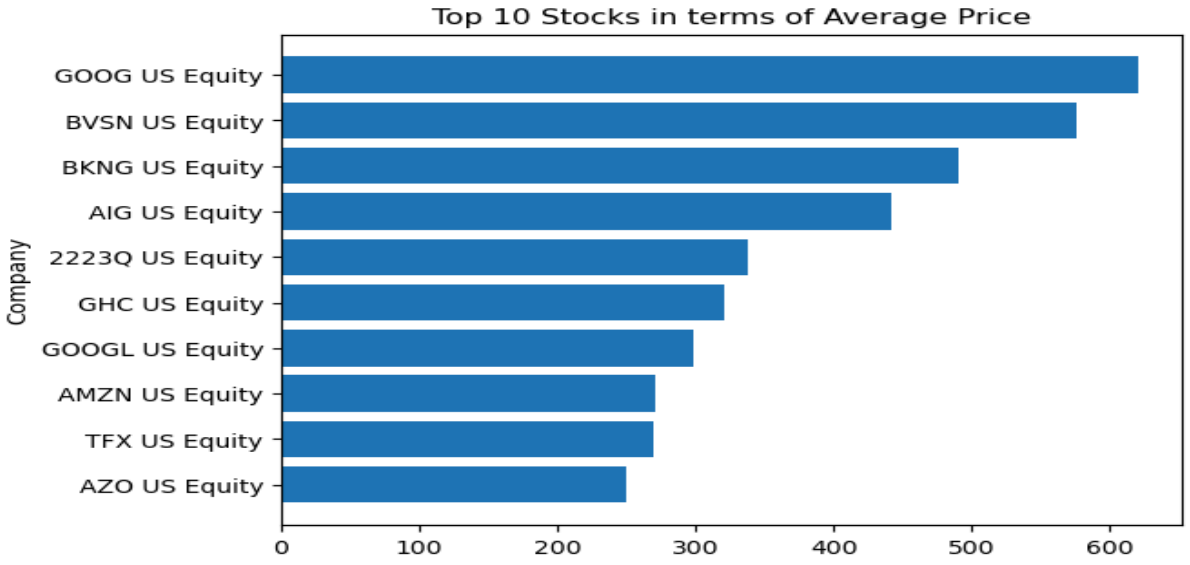

Figure 4: Top 10 stocks in terms of average price from 1993 to 2019

In comparison to the above, the growth of each stock was also analysed by calculating the average price in 2019, subtracting the average price in 1993, and then dividing this calculated value by the average price in 1993. Figure 5 showed the top 10 stocks in terms of growth:

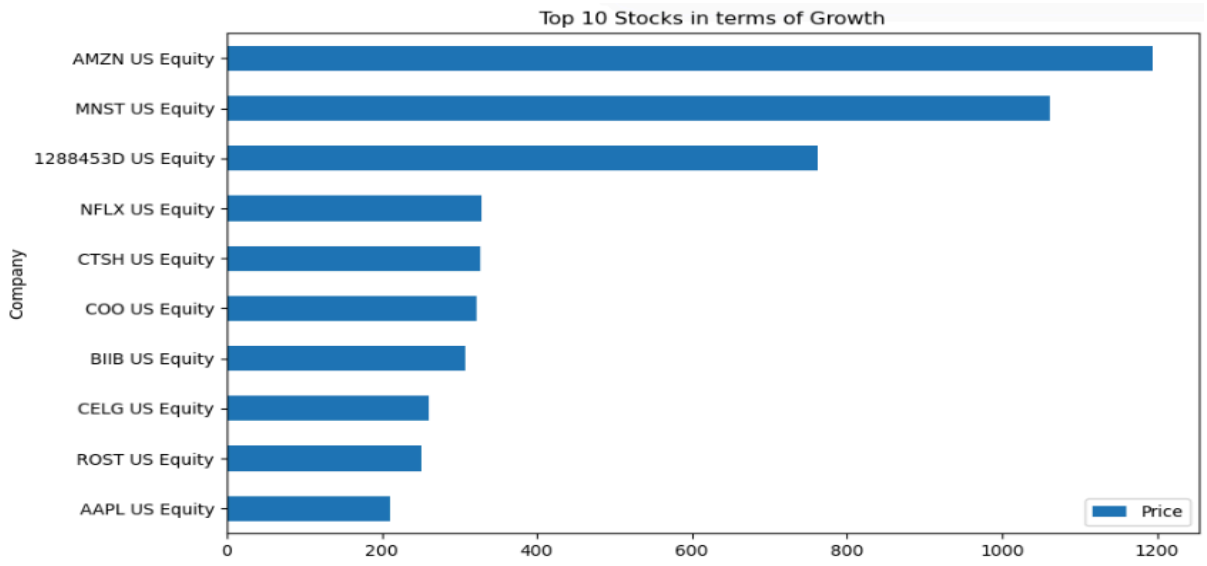

Figure 5: Top 10 stocks in terms of growth from 1993 to 2019

This graph gave insights into how much each stock appreciated in growth relative to their starting point. This was a considerably better visualisation to proceed with because this measured the actual performance of each stock over time, which benefits investors better as they wish to know how much a stock has grown instead of just the price at one particular point.

Performing Principal Component Analysis

Furthermore, the Principal Component Analysis (PCA) is an unsupervised machine learning technique and was used to select the top 10 features using dimensionality reduction. By applying PCA and selecting the first component, which captured the most information among all the principal components, the final top 10 stocks which were most important contributing to the principal component were determined (Figure 6):

```
USB US Equity    0.038674
PG US Equity     0.038669
UTX US Equity    0.038650
IRM US Equity    0.038626
PEP US Equity    0.038552
WFC US Equity    0.038435
AFL US Equity    0.038345
SO US Equity     0.038301
JNJ US Equity    0.038281
BMS US Equity    0.038261
```

Figure 6: The top 10 stocks contributed mostly to the variance captured by the first principal component

A larger absolute value indicated a stronger influence. The biggest absolute value in the principal component was only approximately 0.039, which was an extremely small value, indicating they were not a big influence on the principal component. Since there were over 1000 features in this dataset, most of the features would have a value of nearly 0 (Greenacre et al., 2022). Thus, this feature selection method was not suitable for use. Hence, the results from applying this algorithm did not yield the expected outcomes, so this method was eliminated from any further analysis that was done throughout the entirety of the project.

Cumulative Returns for Growth Over the 26 Years

Cumulative returns aid in illustrating long-term performance. It is assumed that for every increase in stock price over the years, all the dividends for each year were re-invested without any additional capital. Figure 7 illustrates the cumulative returns of 10 highest-growth U.S. stocks over a 26-year period. with key snapshots in 2000, 2005, 2010, and 2015:

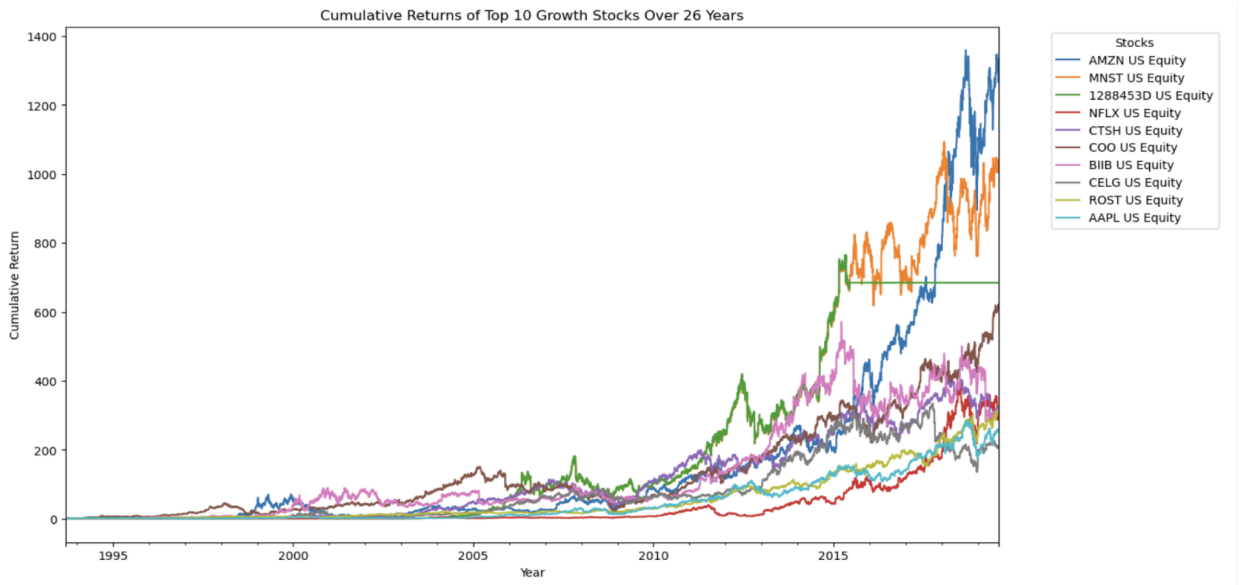

Figure 7: Top 10 stocks in terms of growth from 1993 to 2019

While the exact values are difficult to interpret, it is important to notice the trends of each stock. All of them generally move in an exponential manner, but Amazon (AMZN) out-performed the rest with the most cumulative growth over the 26 years. Prices started picking up from around the 2010 mark as this was when Amazon took control of the e-commerce business. They successfully attracted millions of customers to their platform and generated the most amount of revenue, compared to their competitors Alibaba and Ebay. It is

because of this that many investors decided to buy more shares of Amazon as they felt that they could maximize their returns confidently. The key takeaways from this graph is that it is important that investors seek long-term returns instead of short-term ones and also diversify their capital into different stocks.

Time Series Evaluation Plots

The time-series plots illustrate the growth trajectories of the top 10 companies by growth, with each graph representing a different company's performance over time. The y-axis displays the growth metric (likely stock price or market value), while the x-axis shows the timeline, though the specific dates are unclear due to the truncated labels. Companies such as AMZN US Equity and NFLX US Equity exhibit significant upward trends, with AMZN starting near 0 and peaking around 2000, while NFLX shows a rise from 0 to approximately 80. Other companies, like 1288453D US Equity and CTSH US Equity, display more moderate growth patterns. The plots collectively highlight the varying growth rates and market performances of these top companies, with some demonstrating exponential increases and others more stable, incremental progress. The lack of detailed axis labels and consistent scaling across plots, however, limits the ability to make precise comparisons.

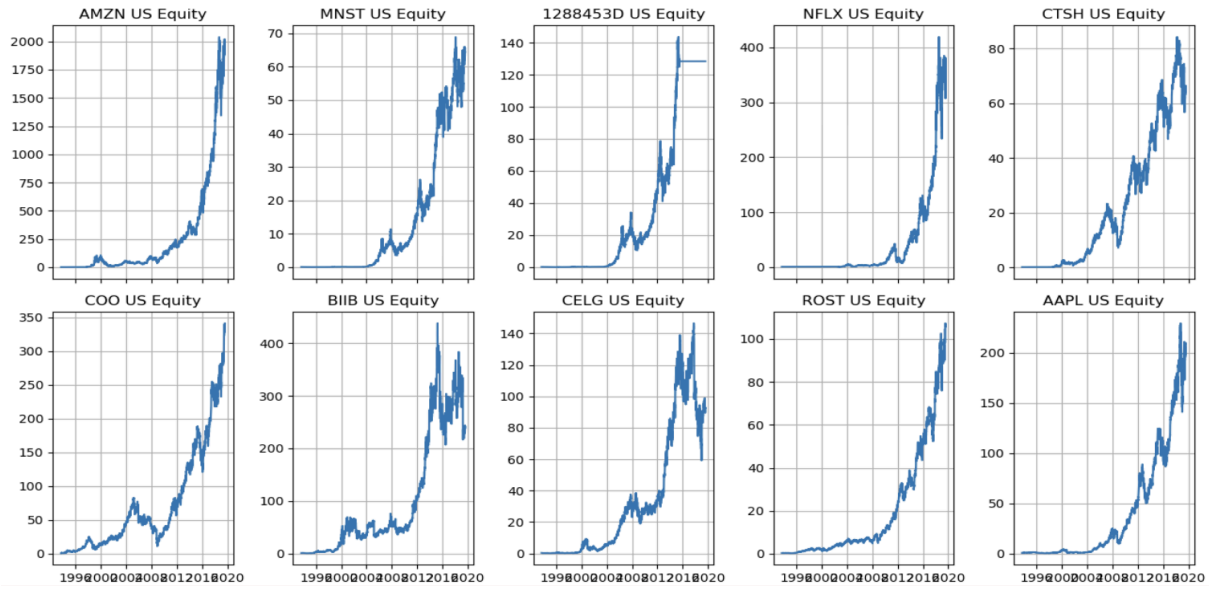

Figure 8: Time series plots for evaluating the stocks' performances over time

Stock Volatility

Volatility is a measure of how rapid one stock changes over a certain period of time. The more volatile a stock, the more inconsistent the stock price is and the more fluctuations the stock experiences. The bar graph below illustrates the volatility of the top 10 stocks in terms of growth, measured by the standard deviation of their returns:

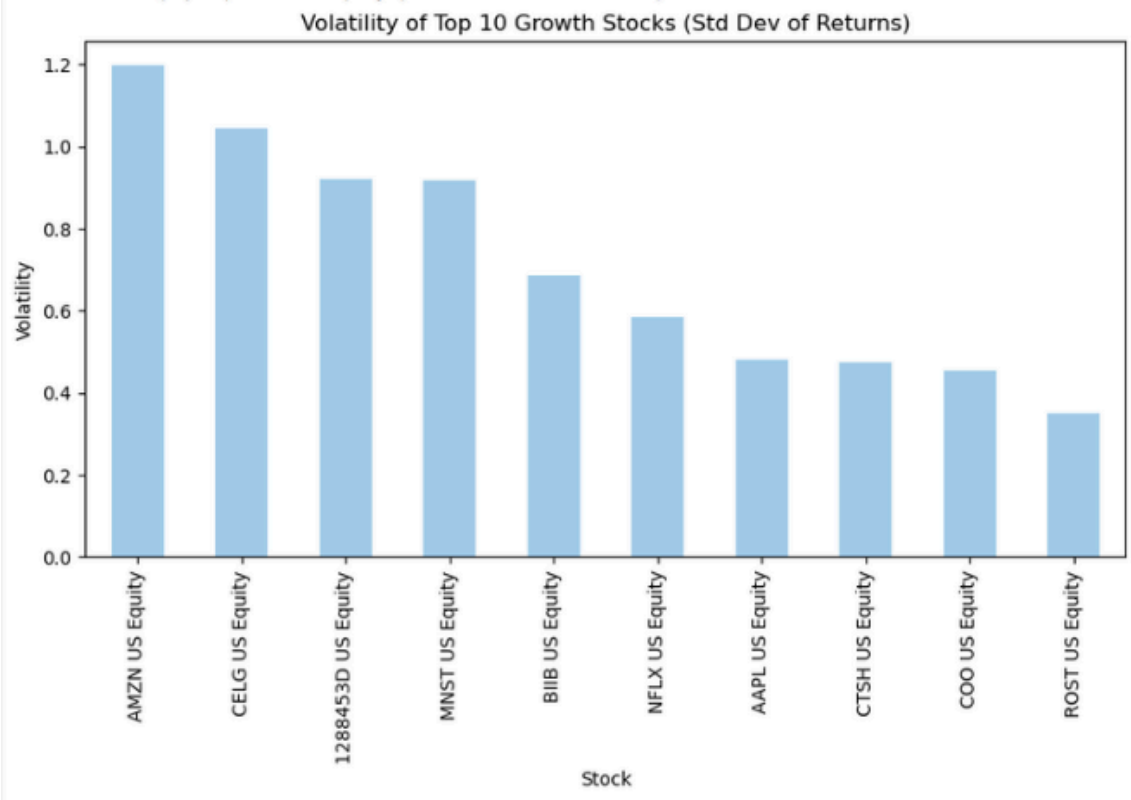

Figure 9: Barplot of the volatility of the Top 10 stocks in terms of growth

AMZN US Equity exhibited the highest volatility, while ROST US Equity was the most stable with the lowest volatility. It is worth noting that volatility doesn't mean poor performance, it just reflects how much a stock's price moves. Stocks with higher historical growth tend to experience greater price swings. Notably, companies in the tech and biotech industries show higher volatility, commonly due to their sensitivity to market speculation and regulatory developments. In contrast, stocks from more stable industries like ROST US Equity demonstrate lower volatility, reflecting more predictable business models. Investors seeking higher returns might gravitate towards volatile stocks but must be prepared for a high risk-to-reward outcome. Those with lower risk tolerance may prefer the steadier, lower-volatility options, even if it means slightly lower growth.

Model Development and Results

Eight different Machine Learning (ML) models were implemented to forecast the stock prices for the best top 10 growth companies over a five year period from 2020 to 2025. Each model was chosen to explore different strengths in capturing stock price behaviour.

Linear Regression

Linear regression was the first model implemented in the project and it operates by establishing a linear relationship between the predictor and the response variable (Antad et al., 2023). The year was used as the independent variable and the average annual stock price of each company was used as the dependent variable.

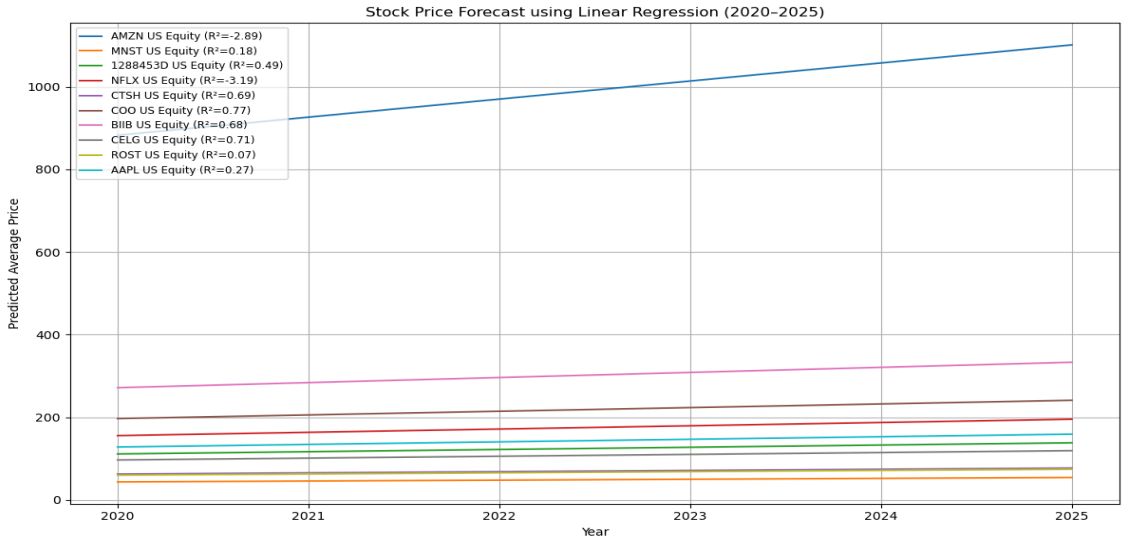

Figure 10: Forecasted Stock Prices using Linear Regression

In Figure 10, AMZN US Equity displayed a significant upward trend while other stocks displayed a moderate trend.

```
AMZN US Equity: MSE = 53025.14, R² = -2.892
MNST US Equity: MSE = 77.39, R² = 0.182
1288453D US Equity: MSE = 437.21, R² = 0.487
NFLX US Equity: MSE = 1793.59, R² = -3.193
CTSH US Equity: MSE = 86.87, R² = 0.692
COO US Equity: MSE = 487.08, R² = 0.772
BIIB US Equity: MSE = 2924.04, R² = 0.685
CELG US Equity: MSE = 279.40, R² = 0.709
ROST US Equity: MSE = 130.91, R² = 0.073
AAPL US Equity: MSE = 697.81, R² = 0.266
```

Figure 11 : MSE and R^2 Scores for Linear Regression

The model's performance was then evaluated using the Mean Squared Error (MSE) and R^2 score. MSE indicated the average squared difference between the actual and the predicted stock prices, while R^2 displayed the proportion of variance in stock prices explained by the model (Figure 11). The linear regression model performed poorly on highly volatile stocks like AMZN and NFLX, and this was seen by the negative R2 values. In contrast, stocks with steady growth like COO and CELG showed better fits with R^2 values above 0.7. Overall, the model was more effective for stocks with steady growth trends.

Ridge Regression

Ridge regression is a technique used in linear regression to address overfitting and multicollinearity among predictor variables. Data standardization was performed to ensure all features contributed equally to the model, regardless of their original scale.

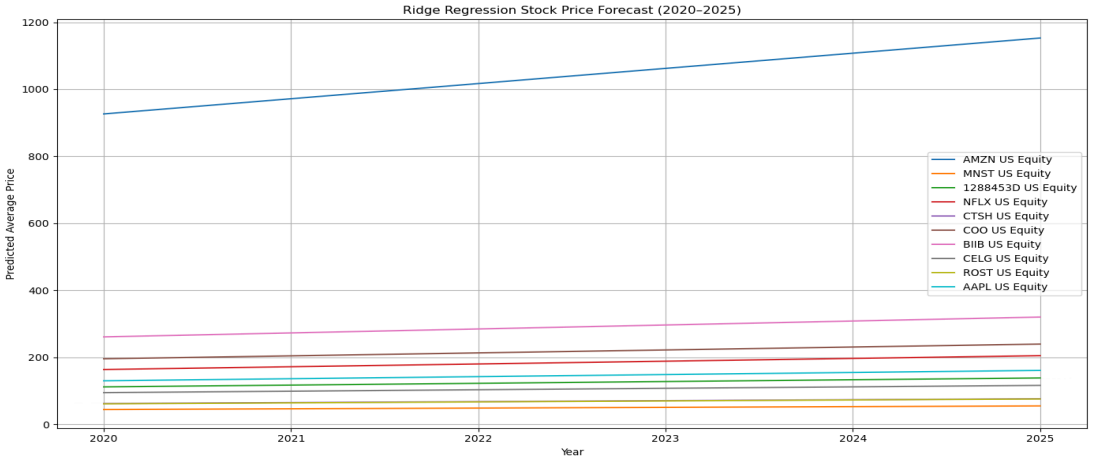

Figure 12: Forecasted Stock Prices using Ridge regression

Figure 12 above shows that AMZN US equity had a significant upward trend, while the other stocks displayed a moderate trend.

```
Processing: AMZN US Equity          Processing: COO US Equity
Ridge - MSE: 49169.68, R²: -2.609   Ridge - MSE: 416.00, R²: 0.805

Processing: MNST US Equity          Processing: BIIB US Equity
Ridge - MSE: 72.23, R²: 0.237       Ridge - MSE: 2899.48, R²: 0.687

Processing: 1288453D US Equity      Processing: CELG US Equity
Ridge - MSE: 416.74, R²: 0.511      Ridge - MSE: 278.59, R²: 0.709

Processing: NFLX US Equity          Processing: ROST US Equity
Ridge - MSE: 1651.06, R²: -2.859    Ridge - MSE: 119.80, R²: 0.152

Processing: CTSH US Equity          Processing: AAPL US Equity
Ridge - MSE: 83.38, R²: 0.705       Ridge - MSE: 661.34, R²: 0.304
```

Figure 13: MSE and R^2 Scores for Linear Regression

To evaluate the model's performance, the mean squared error (MSE) and R^2 value were calculated (Figure 13) above. The MSE value of the predictions was calculated, resulting in a range of 72.23 and 49169.68. There was only one model, which predicted AMZN US Equity, with a MSE of 49169.68, which was considered as it might not accurately capture the data or overfit the noise. Its MSE value was significantly larger than the other models, showing that it performed worse than them. The R^2 value of models had a range of 0.152 to 0.805. There were negative R2 values, which indicated that the model performed poorly in forecasting these stocks.

Overall, the model predicted CTSH US Equity was considered the best ridge model, as it had the second-lowest MSE value and the third-highest R^2 value, which was the most balanced compared to the other Ridge models.

Lasso Regression

Lasso Regression is a regularization technique that can prevent overfitting. Before applying the Lasso model, data standardization was done to ensure the process of comparing the size of coefficients would yield an accurate result, leading to a better model performance.

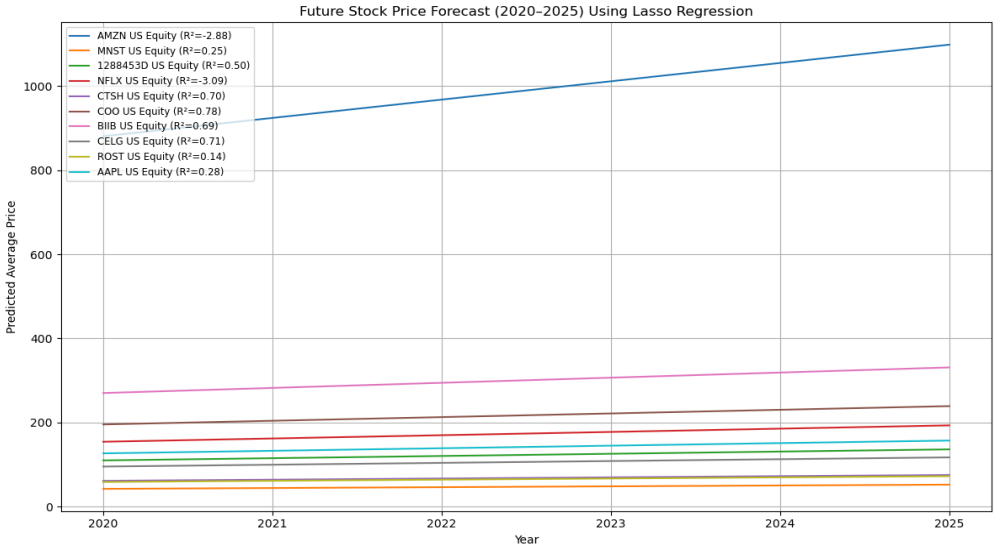

Figure 14: Forecasted Stock Prices using Lasso Regression

Figure 14 above showed that AMZN US equity once again had a significant upward trend compared to the rest.

```
AMZN US Equity: MSE = 52795.59,  R² = -2.875
MNST US Equity: MSE = 71.38,  R² = 0.246
1288453D US Equity: MSE = 426.58,  R² = 0.499
NFLX US Equity: MSE = 1747.96,  R² = -3.086
CTSH US Equity: MSE = 83.65,  R² = 0.704
COO US Equity: MSE = 464.58,  R² = 0.783
BIIB US Equity: MSE = 2916.02,  R² = 0.686
CELG US Equity: MSE = 278.40,  R² = 0.710
ROST US Equity: MSE = 121.07,  R² = 0.143
AAPL US Equity: MSE = 681.75,  R² = 0.282
```

Figure 15: MSE & R^2 Scores for Lasso Regression

The MSE of the predictions were calculated (Figure 15), resulting in a range of 71.38 and 52795.59. AMZN US Equity had a MSE of 49169.68, showing that it might not fully accurately capture the data which may show overfitting. The R^2 value of models had a range of 0.246 to 0.783. There were negative R2 values, indicating the model's poor performance once again.

Overall, the model predicted CTSH US Equity the best as it had the second-lowest MSE value and the third-highest R^2 value.

Decision Trees

The analysis was further expanded with the inclusion of Decision Trees, or Regression Trees. This is a supervised ML algorithm that splits the dataset into many sub-datasets and organises them in a tree-like structure for predictions. Since stock market prices are always fluctuating, the usage of decision trees would help to capture non-linear relationships in the data.

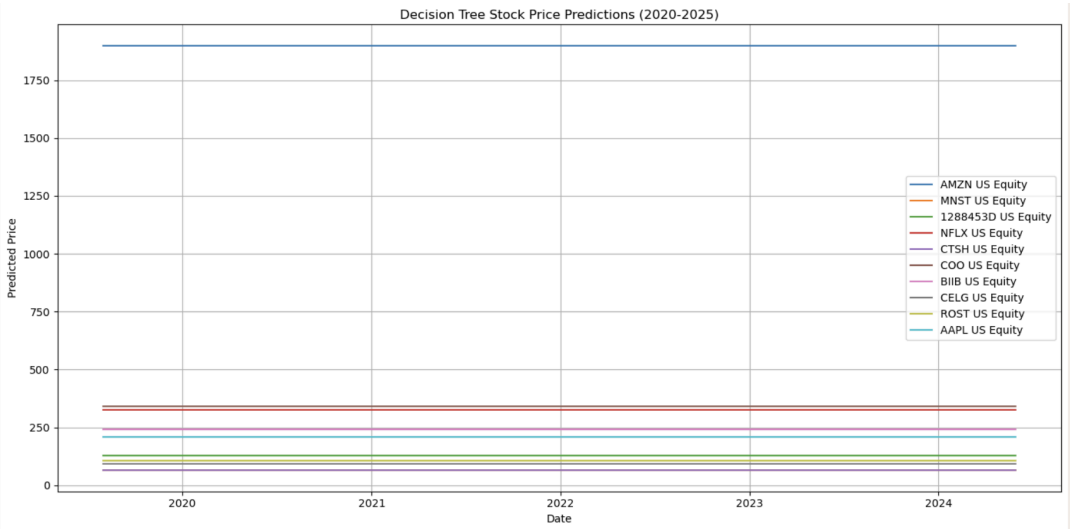

Figure 16: Forecasted Stock Prices using Decision Trees

According to Figure 16 above, it is obvious that all the forecasted prices stayed stagnant and did not show any change in price action. This highlighted the limitations of using this model as it was unable to capture time-series data efficiently. This led to the idea of using a more powerful model known as Random Forest shown next.

Random Forest

This is an ensemble supervised ML algorithm which essentially aggregates the outcomes based on the collection of multiple decision trees. This approach was thought to reduce overfitting in single trees and introduce variance to capture randomness in the data. The graph below illustrated the forecasted results using this model:

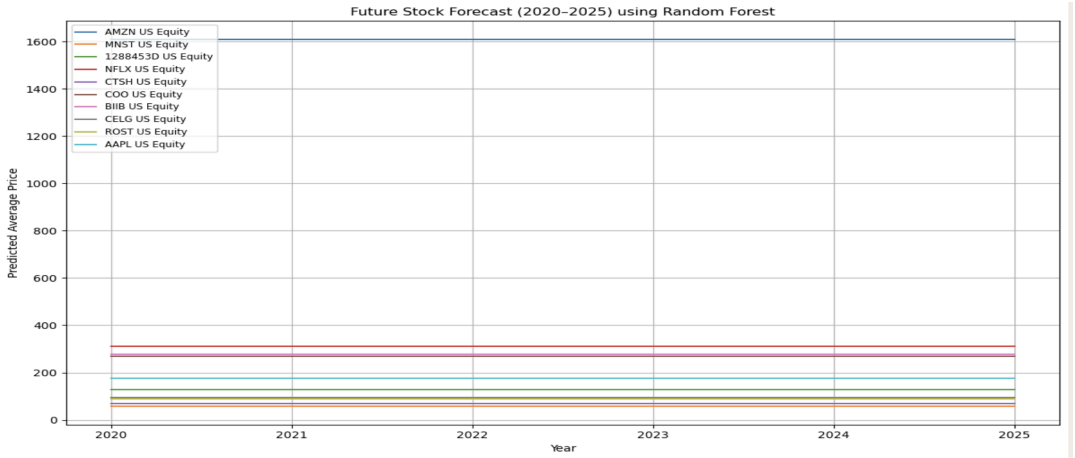

Figure 17: Forecasted Stock Prices using Random Forest

It can be observed in Figure 17 that the forecasted prices were stagnant once again. This is expected because since forecasting using Decision Trees failed at predicting trends, a collection of these trees will not make a difference to accuracy of the model.

Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a statistical model that utilizes historical time series data to predict future stock prices. Its other appeal comes from the fact that it often outperformed traditional methods in returns and risk efficiency (Chen, 2023). ARIMA, however, has notable limitations. Li (2024) found that ARIMA models tend to underperform when applied to broad market indices such as the S&P 500. Yang (2023) uncovered that it tends to struggle with volatile stocks. This was investigated as an additional method that has not been utilized before in previous projects.

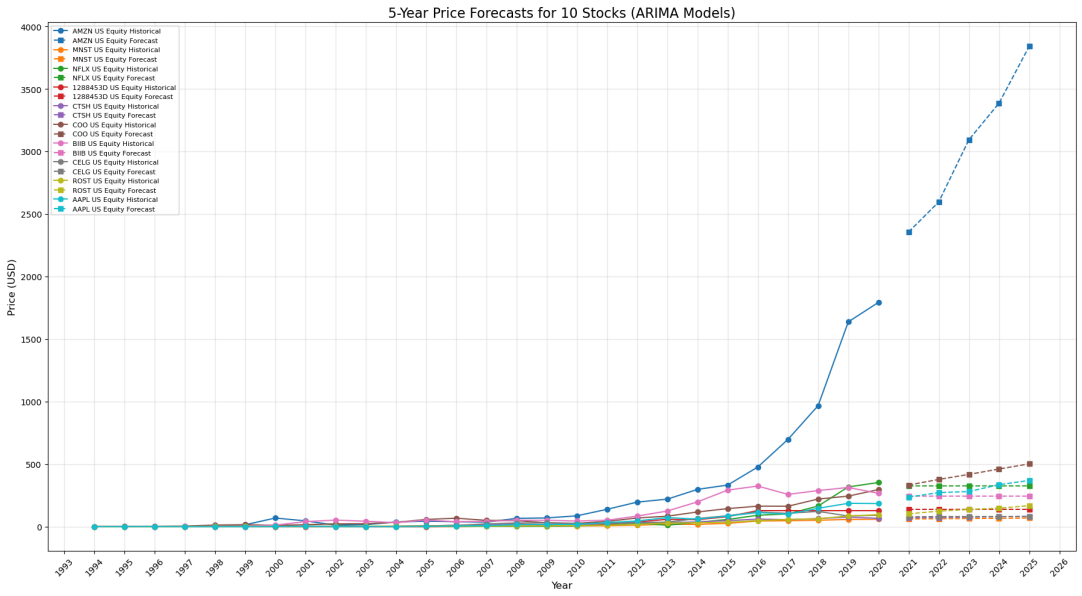

Figure 18: Forecasted Stock Prices using ARIMA

Based on Figure 18, AMZN US Equity had a significant exponential growth in relation to the other nine stocks especially from 2015 to 2019 due to the rise in e-commerce and cloud computing. The forecast continued this steep upward trend, projecting a future share price near $3900 in 2025. The others had a modest and steady forecasted growth post 2019, which reflected historically stable performance and moderate volatility.

Monte Carlo Simulation

Monte Carlo Simulation is a computational technique that models uncertain outcomes by accounting for randomness in systems where predictions are challenging. It is commonly used in quantitative analysis because it analyses probability distributions and risk for decision-making strategies (Pedersen, 2013).

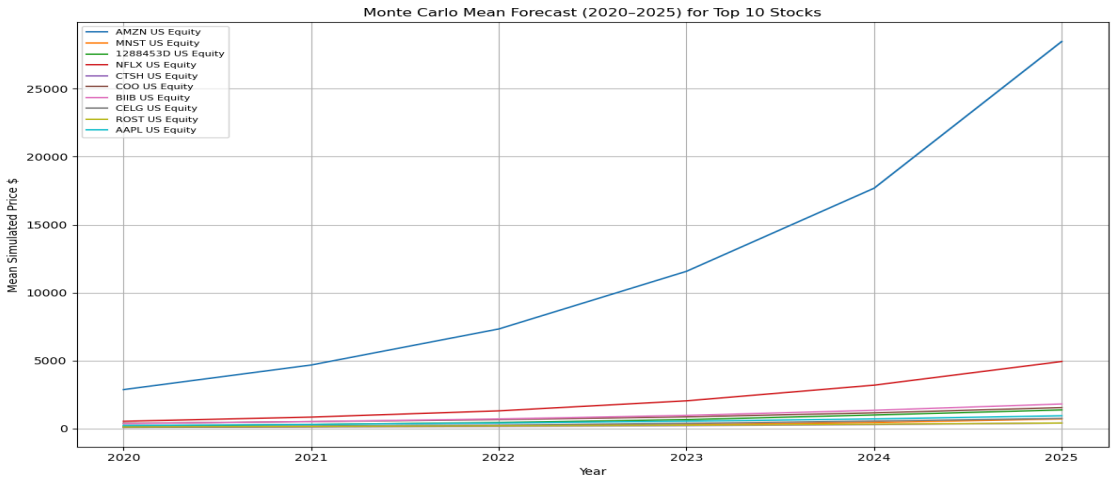

Figure 19: Forecasted Stock Price using Monte Carlo

Figure 19 shows that AMZN US Equity leads with a potential peak above $25,000 by 2025, which is expected for a stock with a very strong past performance and momentum. The forecast shows a general upward trend across the other stocks, though the scale varies considerably. However, a notable limitation is its assumption that stock prices have constant volatility and that accuracy depletes over long periods; studies did not find Monte Carlo simulation effective for multi-year forecasts (Prasad et al., 2022).

Long Short Term Memory (LSTM)

LSTM is a type of Recurrent Neural Network (RNN) which explores sequential data and learns patterns of trends over time. A study conducted by Bhandari et al. (2022) achieved desired results as they found that LSTM had the ability to accurately capture stock market trends especially during the 2008 market crash and the COVID-19 pandemic.

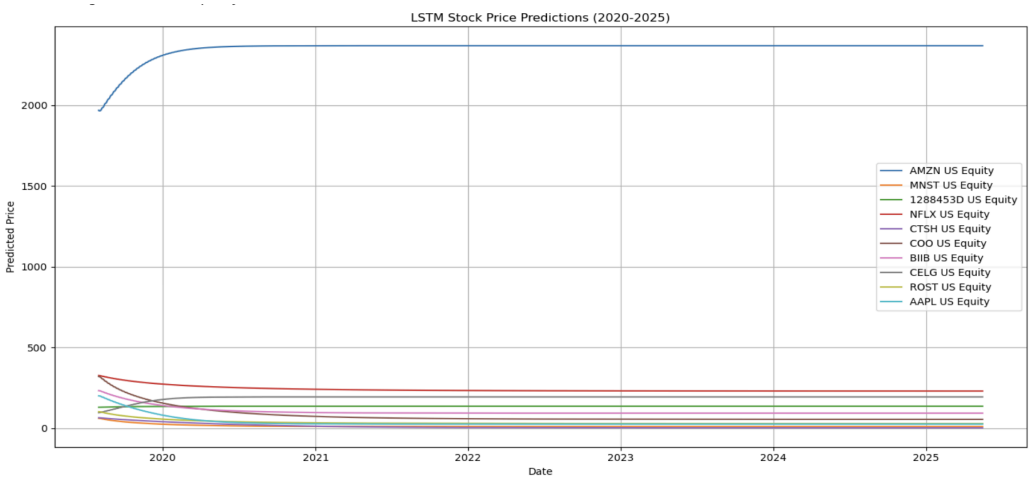

Figure 20: Future Stock Prices using LSTM

AMZN dominates the graph (Figure 20) with a higher predicted price and then plateauing near 2200. Contrastingly, most of the other stocks exhibit a sharp decline early on and then flattening out. In reality, stock prices can be extremely volatile, but the plots showcase flatlines. This suggested the model's inability in capturing natural trends due to overfitting. The inconsistency indicates that the model fits better for one stock and badly for others. This makes the model unsuitable for predicting stock prices in this project.

Overall Best Performing Model

After the ML process, the best model was needed to be picked in order to build the portfolio. Random forest, Decision tree, ARIMA and Monte Carlo had their limitations. Although the

regression models did not have fantastic R^2 values, they were the best ones in capturing the linear relationship reliably to predict stock prices.

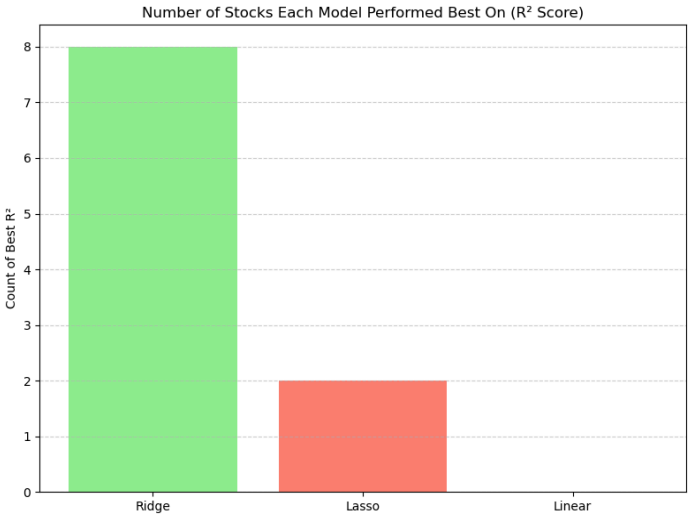

Figure 21 : Count of best R^2 values of different Regression models

Figure 21 shows that the Ridge regression model performed better than Lasso and Linear regression models in terms of predicting future stock prices as it had a majority of higher R^2 values. Hence, the Ridge regression model was the best in predicting future stock prices.

Building the Portfolio

Ridge Regression was used in order to calculate the potential returns that an investor could make by investing in the ten stocks. Furthermore, the volatility of the stocks will be considered in the building of the portfolio as this is able to separate the stocks into risk categories. Building from the EDA and Figure 9, the volatility of each stock will now be converted into a percentage basis for easier interpretation. A study by Rujivan et al. (2025), suggests that volatility is a good building block of a portfolio as it is able to capture dynamic and unstable trends in stock prices.

	Volatility (%)	Risk Category	Predicted Return (%)
AAPL US Equity	47.84	Low Risk	23.79
ROST US Equity	19.66	Low Risk	23.60
CTSH US Equity	45.68	Low Risk	22.66
COO US Equity	29.78	Low Risk	22.36
NFLX US Equity	66.04	Medium Risk	25.19
BIIB US Equity	58.56	Medium Risk	22.62
AMZN US Equity	110.05	High Risk	24.44
MNST US Equity	95.23	High Risk	24.04
1288453D US Equity	95.97	High Risk	23.69
CELG US Equity	111.67	High Risk	22.75

Figure 22: Investment summary of top 10 stocks

Figure 22 shows that stocks were categorized into low, medium and high risk based on their volatility percentages. This was done by separating them into 3 quantiles to see where the stock falls. High risk stocks exhibited volatility above 95% whereas low-risk stocks offered comparable predicted returns of around 23.6 to 23.8%. This suggested that higher volatility did not necessarily guarantee higher returns.

Low-Risk & High-Risk Portfolio

```
Low-Risk Portfolio Weights (%):
 AMZN US Equity          0.76
MNST US Equity           0.00
1288453D US Equity       0.00
NFLX US Equity          13.61
CTSH US Equity           0.00
COO US Equity           12.88
BIIB US Equity           0.00
CELG US Equity           4.08
ROST US Equity          42.11
AAPL US Equity          26.56
dtype: float64
```

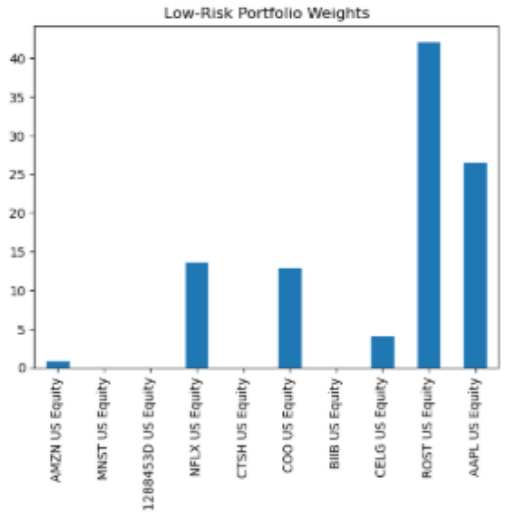

Figure 23: Weights of Low-Risk Portfolio Stocks as Percentages for investment

Figure 23 illustrates the percentage weights, which are the amount of allocation of capital that should be invested in for long-term growth. Minimal capital was needed for more volatile stocks like Amazon and Celgene Corp, whereas Ross Stores and Apple provided more opportunities. This limited diversification concentrated on a few reliable stocks.

```
High-Risk Portfolio Weights (%):
 AMZN US Equity          10.36
MNST US Equity           11.14
1288453D US Equity        0.00
NFLX US Equity           17.20
CTSH US Equity            0.00
COO US Equity            17.06
BIIB US Equity            0.00
CELG US Equity           12.27
ROST US Equity           31.97
AAPL US Equity            0.00
dtype: float64
```

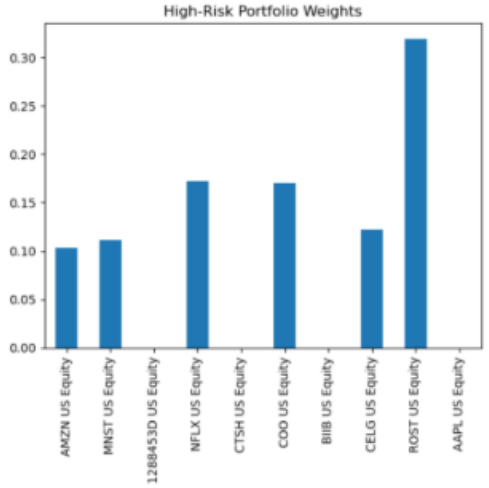

Figure 24: Weights of High-Risk Portfolio Stocks as Percentages for investment

Figure 24 displays the weight distribution in the high-risk portfolio, reflecting a more aggressive investment strategy that prioritizes investment growth. There is a higher rate of exposure given to the high risk stocks as this could potentially give a higher return investment. This risk-diverse approach helps investors who seek to invest in a portfolio which has a diverse range of volatility risks as well as potentially leading to a higher predicted yield of investment. Therefore, the weights in percentage represents the amount of allocation that a fund should be invested in each stock for a diversified and high-risk investment.

Conclusion

In conclusion, the aims of this project were answered through the various methodologies that were described above. Throughout the entirety of this project, the stock data was successfully leveraged to gain relevant insights with regards to investment schemes tailored to different skilled investors. Working with a large dataset imposed issues such as having more than 3,000,000 missing values. Successful data wrangling was done to fill in these missing values so that it would make the data analysis fairer. Multiple EDA techniques were used including bar and time series plots to evaluate the volatilities of each stock and identifying the top 10 most performing stocks. Moreover, further analysis was conducted by implementing ML models such as Linear, Ridge, and Lasso regressions, Decision trees, Random forest, ARIMA, Monte Carlo and LSTM to forecast the future stock prices from 2020 to 2025. Ridge regression performed the best in overall R^2 value compared to the rest of the ML models. Lastly, a diversified portfolio of high and low risks were successfully tailored to showcase the allocation of funds in stocks based on percentages to adhere to different investors with their own unique financial goals. This finally answered all the problem statements that were introduced at the beginning of this report. Ultimately, this project fulfilled all the objectives and through the combination of data analytics, key insights were unraveled and successful decision-making strategies were applied for the investors.

Reference List

Alwateer, M., Atlam, E., El-Raouf, M. M. A., Ghoneim, O. A., & Gad, I. (2024). Missing data Imputation: A comprehensive review. *Journal of Computer and Communications*, *12*(11), 53–75. https://doi.org/10.4236/jcc.2024.1211004

Antad, S., Khandelwal, S., Khandelwal, A., Khandare, R., Khandave, P., Khangar, D., & Khanke, R. (2023). Stock price prediction website using linear regression - a machine learning algorithm. *ITM Web of Conferences*, *56*, 05016. https://doi.org/10.1051/itmconf/20235605016

Bhandari, H. N., Rimal, B., Pokhrel, N. R., Rimal, R., Dahal, K. R., & Khatri, R. K. C. (2022). Predicting stock market index using LSTM. *Machine Learning with Applications*, *9*, 100320-. https://doi.org/10.1016/j.mlwa.2022.100320

Chen, Y. (2023). Application of ARIMA model in portfolio optimization. *Advances in Economics Management and Political Sciences*, *26*(1), 227–236. https://doi.org/10.54254/2754-1169/26/20230575

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data, 8*(1). https://doi.org/10.1186/s40537-021-00516-9

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers, 2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Li, W. (2024). Whether ARIMA-based portfolio can beat the market index. *Advances in Economics Management and Political Sciences*, *90*(1), 122–128. https://doi.org/10.54254/2754-1169/90/20242003

Pedersen, M. (2013). Monte Carlo simulation in financial valuation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2332539

Prasad, K., Prabhu, B., Pereira, L., Prabhu, N., & S, P. (2022). EFFECTIVENESS OF GEOMETRIC BROWNIAN MOTION METHOD IN PREDICTING STOCK PRICES: EVIDENCE FROM INDIA. *Asian Journal of Accounting and Governance*, *18*. https://doi.org/10.17576/ajag-2022-18-09

Ribeiro, S. M., & Castro, C. L. (2022). Missing Data in Time Series: A review of Imputation methods and case study. *Learning and Nonlinear Models*, *20*(1), 31–46. https://doi.org/10.21528/lnlm-vol20-no1-art3

Rujivan, S., Khuatongkeaw, T., & Sutchada, A. (2025). Optimal Portfolio Construction Using the Realized Volatility Concept: Empirical Evidence from the Stock Exchange of Thailand. *Journal of Risk and Financial Management*, *18*(5), 269. https://doi.org/10.3390/jrfm18050269

Yang, W. (2023). Application of ARIMA in Mean-Variance portfolio Optimization. *Advances in Economics Management and Political Sciences*, *36*(1), 11–16. https://doi.org/10.54254/2754-1169/36/20231777