STAT3612 Statistical Machine Learning
Final Project
Prediction of Index Price of S&P 500 Index

Group members:

Name	UID
Chan Hong Ching	3035574696
Cheung Ka Fai	3035688801
Lo Ming Kwan	3035606463
Sit Wai Tang, Vento	3035761603

Table of Contents

1. Introduction

This project aims to identify a set of predictors to predict the index price of the S&P 500 Index. The S&P 500 Index is a stock market index in the United States, which consists of 500 listed companies. It is a market-value-weighted index, where its components are weighted according to their relative total market capitalization (Hayes, 2021). In this project, we want to explore the possible factors that would affect the performance of the S&P 500 Index. We first identify 82 possible predictors, as suggested by Hoseinzade and Haratizadeh (2019). Then, we will use different machine learning techniques to determine the suitable predictors that should be included in our model. 80% of the data will be used to train the model, while the other 20% are reserved as test data. Finally, we use the model to predict the price of the S&P 500 Index, and evaluate its performance by comparing the predicted price with the actual price.

2. Data description and preparation

The data is obtained from the UCI machine learning repository. The observations are from 31/12/2009 to 15/11/2017, which include 1984 trading days. The dataset includes 82 predictors. Those 82 predictors can be categorized into 7 different groups, which are technical indicators, world stock market indices, the exchange rate of U.S. dollar, commodities, stock performance of big companies in the U.S. markets, future contracts, and other useful variables (Hoseinzade and Haratizadeh, 2019).

We explain how these 7 groups of predictors may affect the performance of the S&P 500 Index. Firstly, technical indicators are obtained from historical data of the S&P 500 index prices and trading information. Analysts often use technical indicators to analyze the price movement. Secondly, due to globalization of the economy and difference in time zones, stock markets in different countries often interact with each other. Thirdly, the change in exchange rate of U.S. dollar to other currencies would affect the profit of multinational companies, which may affect their stock prices and therefore affect the performance of the S&P 500 Index. Besides, the price of commodities like gold, oil and wheat can reflect the situation of the global market, which may be related to the stock market. In addition, since S&P 500 is a market-value-weighted index, the performance of the stocks of big companies would have a larger influence on the performance of the index. Meanwhile, future contracts reflect the expected value of the underlying assets in the future, hence investors would prefer to buy stocks with higher expected value than their current value. Eventually, some scholars suggest that other variables, such as Treasury bill rates, may also be useful in stock market prediction (Zhong & Enke, 2017).

From the dataset, we notice that not all predictors contain data in every trading day. However, if any one of the predictors has a missing value on a particular day, it means that we cannot train the model using the data on that day. There are several reasons for missing values. Using the predictor Hang Seng Index as an example, its missing values may be due to some special public holidays in Hong Kong, such as Chinese New Year holidays and the National Day, which do not exist in the United States. Therefore, when we train the model using all the 82 predictors, some trading days will not be included. Figure 1 shows the amount of consecutive trading days missed due to the missing data. In total, 870 trading days are missed.

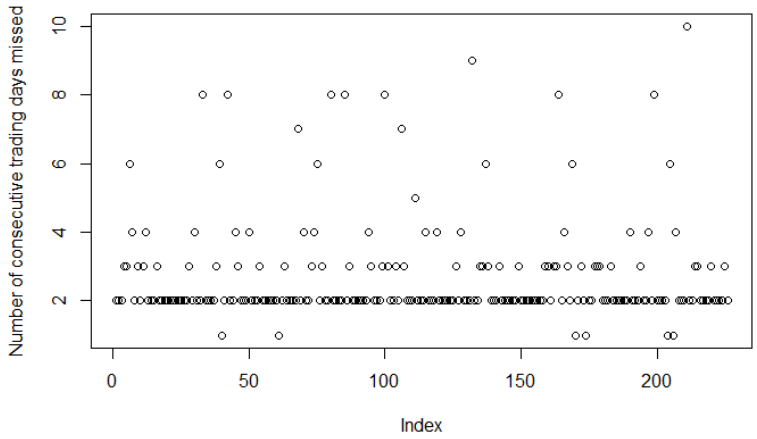

Figure 1

As a result, we need to perform data cleansing. We decide to eliminate the predictors with more than 50 missing values. We observe the result in Figure 2:

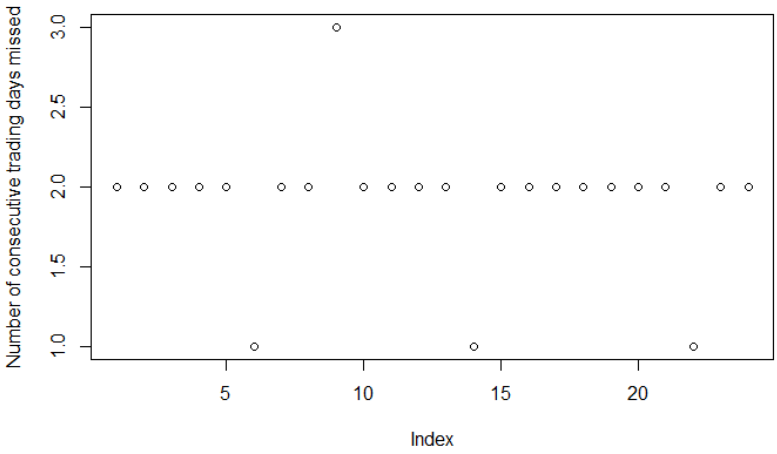

Figure 2

After data cleansing, 11 predictors are eliminated. The data is much better now, since the number of consecutive trading days missed is now limited to only 1 to 3 consecutive trading days. Therefore, we will then use the remaining 71 predictors to train the model. Since the index price of S&P 500 is a time series, we will use the data after cleansing in the first 80% trading days as training data, and reserve the data in the last 20% trading days as test data.

3a. Shrinkage Methods

By shrinkage methods, we first fit a model including all the 71 predictors. Then, the estimated coefficients are shrunken towards zero relative to the least squares estimates. The shrinkage methods can reduce variance of the model.

The first shrinkage method we will use is ridge regression. Ridge regression can shrink the estimates of the predictors towards zero, and hence we can identify the important predictors. The result of the ridge regression is shown in Figure 3:

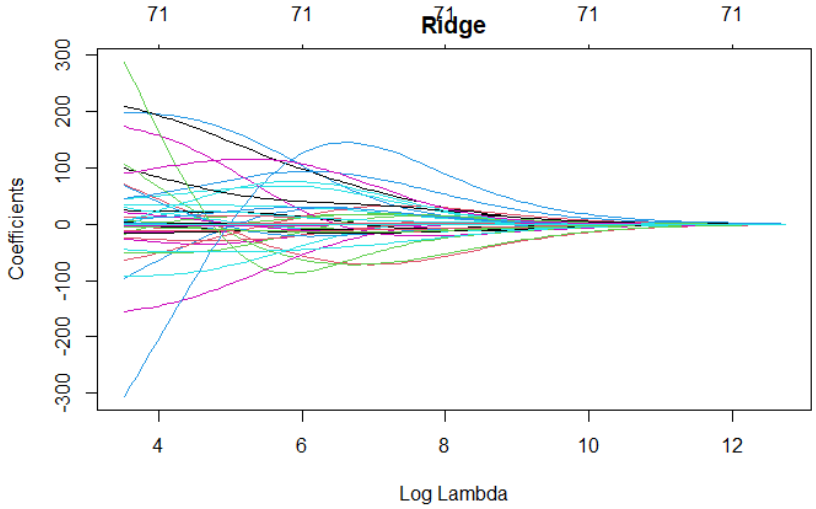

Figure 3

To select the optimal λ for the ridge regression, we use ten-fold cross validation. From the validation result in Figure 4.1 and Figure 4.2, we notice that the mean-squared error is minimized at λ = 33.72179. Using the test data to test the trained model, we find that the mean-squared error of test data is 3000.203.

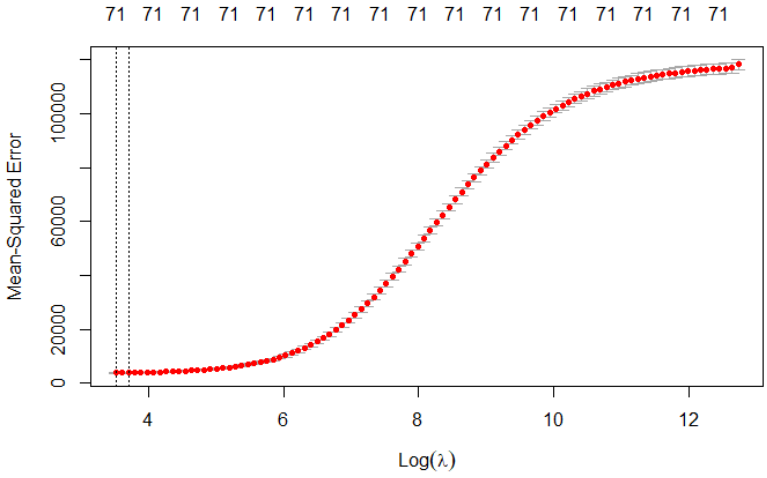

Figure 4.1

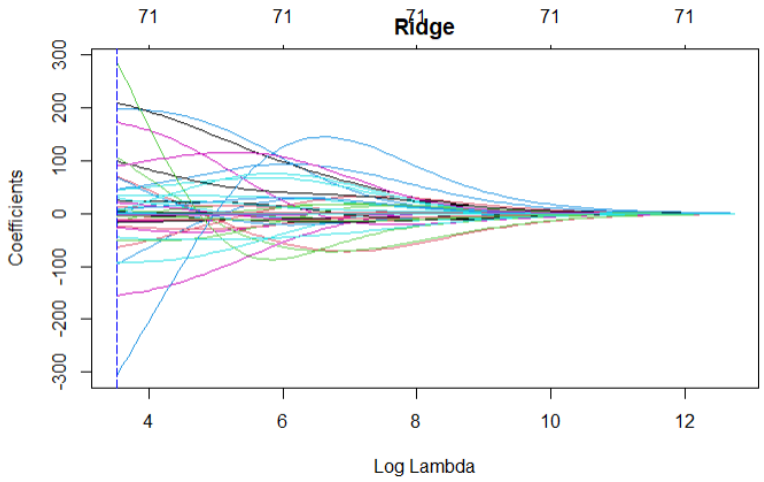

Figure 4.2

However, the disadvantage of ridge regression is that all the 71 predictors will be included in the final model, even though some of the coefficients of the predictors are shrinked toward zero. Meanwhile, lasso regression not only can shrink the estimates of the predictors towards zero like ridge regression, but it also has the ability to force some of the coefficient estimates to be exactly equal to zero, and therefore performs variable selection. Hence, we will next use lasso regression. We observe the result of the lasso regression in Figure 5:

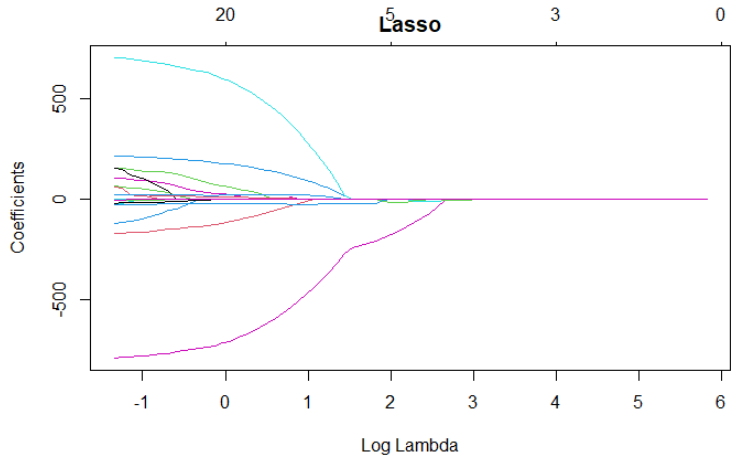

Figure 5

To select the optimal λ, we use ten-fold cross validation. From the validation result in Figure 6.1 and Figure 6.2, we observe that the mean-squared error is minimized at λ = 0.6619724.

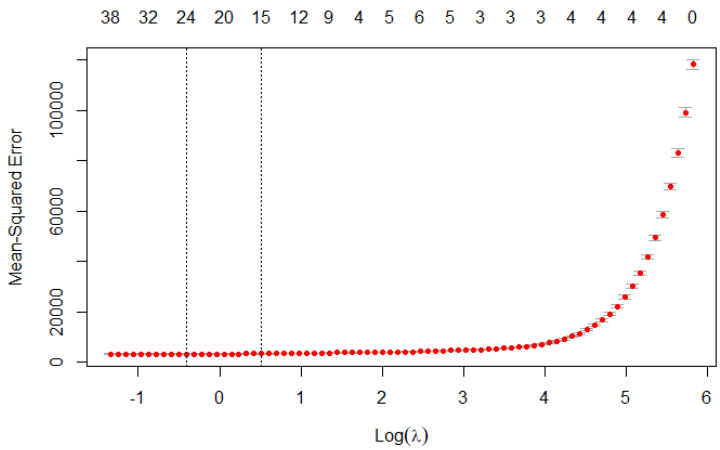

Figure 6.1

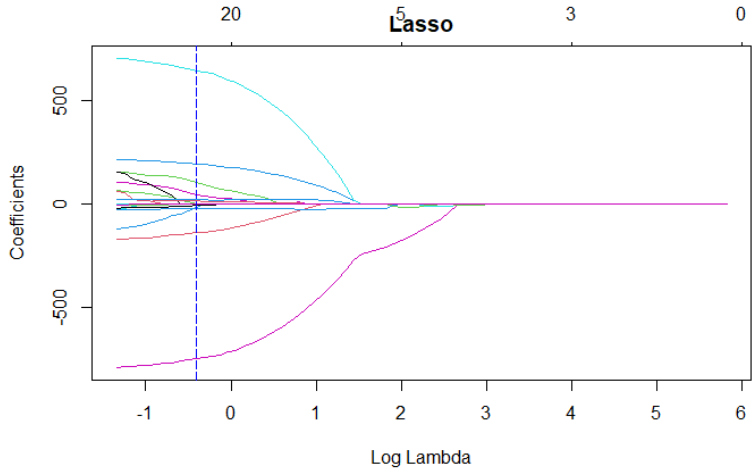

Figure 6.2

With λ = 0.6619724, the coefficients of 46 predictors become 0. Therefore, only 25 predictors are remaining in the lasso model. Using the test data to test the trained model, we find that the mean-squared error of test data is 8154.627.

3b. Subset selection

By subset selection, we identify a subset of the 71 predictors which are believed to be related to the response. Then, we fit a model using least squares on the reduced set of variables.

The first subset selection method we use is forward stepwise selection. We begin with a model with no predictor, then we add predictors to the model one-at-a-time. At each step, the predictor that gives the greatest additional improvement to the fit is added. To decide the amount of predictors needed in the model, we use the mean error of 10-fold cross-validation as our criteria to identify the best model. The mean error of 10-fold cross-validation provides us an estimate of the test error. The results are shown in Figure 7.

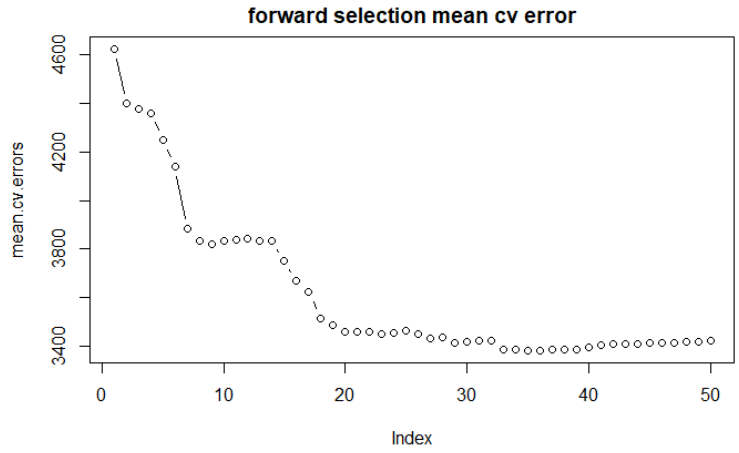

Figure 7

We observe that the mean error of 10-fold cross-validation is minimized when there are 35 variables in the model. The value of error is 3383.7962.

Next, we use backward stepwise selection. It begins with the full least squares model containing all the 71 predictors, and then iteratively removes the least useful predictor one-at-a-time. To decide the amount of predictors needed in the model, we again use the mean error of 10-fold cross-validation as our criteria to identify the best model. The result is shown in Figure 8.

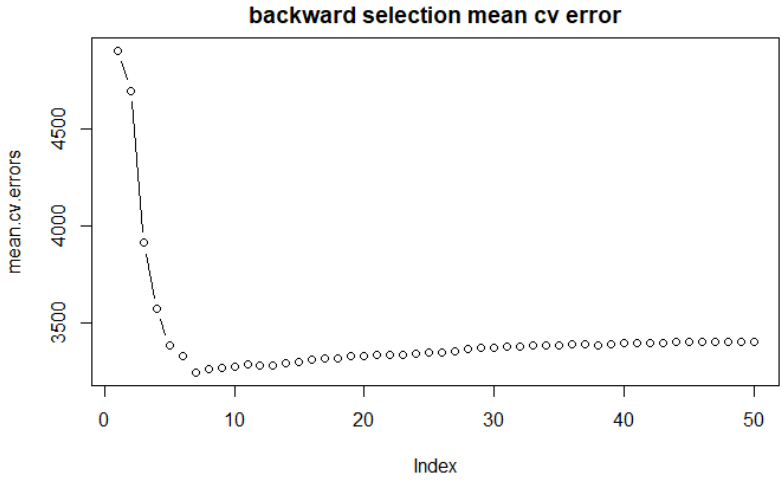

Figure 8

We observe that the mean error of 10-fold cross-validation is minimized when the model includes 7 variables, where the value of error is 3246.2148.

3c. Dimension Reduction

We project the 71 predictors into a M-dimensional subspace, where M < 71, by computing M different linear combinations of the variables. Then, these M projections are used as predictors to fit a linear regression model through least squares.

Principal components analysis (PCA) is applied to define the linear combinations of the predictors in our regression. The first principal component is the normalized linear combination of the variables with the largest variance. Then, the second principal component has the largest variance, subject to being uncorrelated with the first, and so on.

The main weakness of PCA is that it tends to be highly affected by outliers in the data. Hence, several robust variants of PCA have been developed to iteratively discard the data points that are described poorly by the initial components (VanderPlas, 2016).

The result of the dimension reduction method can be seen in Figure 9.1 and Figure 9.2:

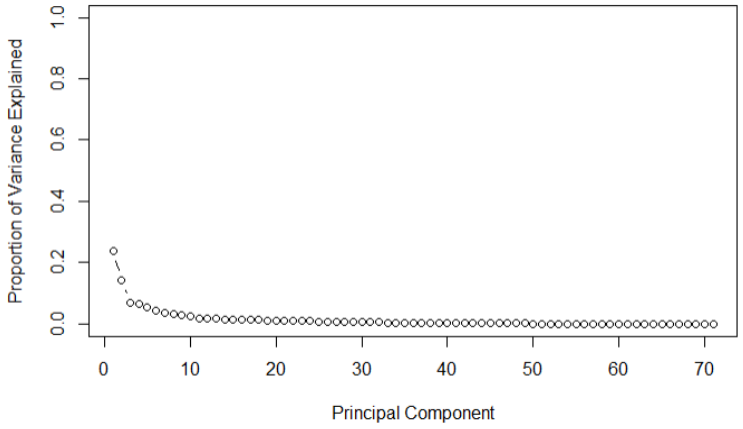

Figure 9.1

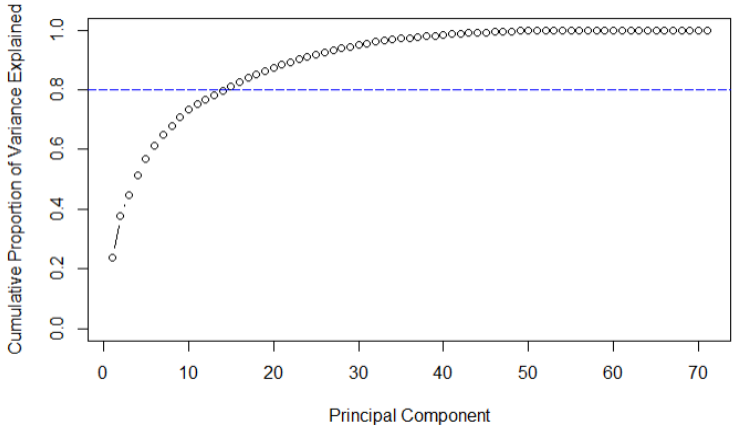

Figure 9.2

To decide the amount of principal components needed, we set a threshold of 80% for the cumulative proportion of variance explained. It is found that at least 14 principal components are required to explain 80% of the variance.

4. Final model selection

From the above machine learning techniques, we obtained 5 models from lasso regression, ridge regression, forward stepwise selection, backward stepwise selection and dimension reduction respectively.

The test error of the lasso regression model is too large compared to the others. Meanwhile, the best model from forward stepwise selection requires 35 variables, which is too complicated. It is also ineffective to use dimension reduction, as at least 14 principal components are needed to explain 80% of the variance.

The ridge regression model has the lowest mean-squared error of 3000.203. However, the model with 7 variables identified by backward stepwise selection also performs well, with a slightly higher mean error of 10-fold cross-validation of 3246.2148. Since compared to ridge regression, the backward stepwise selection performs variable selection, leading to a simpler model, therefore finally we decide to select the regression model identified by backward stepwise selection. We then use backward stepwise selection to identify the 7 best predictors. The regression model is as below:

$$\hat{Y} = -3.685882 + 3.077501 \times ROC\_15 + 3.173849 \times ROC\_20 + 1.034894 \times EMA\_50 + 1082.598939 \times DTB3 - 831.579277 \times DTB6 - 54.474219 \times DGS10 + 31.765676 \times DAAA$$

Where $\hat{Y}$ is the predicted index price of S&P 500, ROC_15 is the 15 days rate of change of the index, ROC_20 is the 20 days rate of change of the index, EMA_50 is the 50 days exponential moving average of the index, $DTB3$ is the secondary market rate of the 3-month Treasury Bill, $DTB6$ is the secondary market rate of the 6-month Treasury Bill, $DGS10$ is the 10-Year Treasury constant maturity rate, $DAAA$ is the Moody's seasoned Aaa corporate bond yield.

5. Discussion

In the beginning, we mentioned that the original 82 predictors can be classified into 7 different groups. With reference to those 7 groups, ROC_15, ROC_20, EMA_50 are technical indicators, while $DTB3$, $DTB6$, $DGS10$ and $DAAA$ belong to the group of "other useful variables".

It is not surprising to find that ROC_15, ROC_20 and EMA_50 are important predictors on the index price. These three technical indicators take into account a relatively longer trend of the index performance. Therefore, they are more stable and are able to analyze the future movement of price accurately.

The model also implies that some of the bond yields can influence the index price. From a financial point of view, bond yield should have an inverse relationship with index price (IG,

2021). It is because investors will dump low-yielding bonds and invest in stocks with potentially higher returns. With more investors buying the stocks, the higher the stock prices could rise, and vice versa. From the model, we can see the index price have an inverse relationship with $DTB6$ and $DGS10$. However, surprisingly, the index price have a direct relationship with $DTB3$ and $DAAA$ under this model. This relationship was not what we would expect to see from a theoretical point of view.

6. Limitations

Firstly, the dataset that we obtained contains some missing values. As a result, we have to perform data cleansing by removing the variables with more than 50 missing values. In total, 11 predictors were removed. However, it might be possible that any of those 11 predictors are in fact significant.

Secondly, we find that EMA_50 is a predictor in our final model. However, since EMA_50 is a 50-day exponential moving average, the first value of EMA_50 can only be obtained on the 50th trading day in our dataset. Therefore, the first 49 trading days cannot be used to train the model mainly due to EMA_50, and it is not sure whether removing this predictor can actually give us a better model.

In addition, ROC_15 and ROC_20 in our model are terms of rate of change. If Δx tends to zero, then the rate of change is close to a linear function. However, when the sample size tends to infinity, Δx does not tend to 0, meaning the rate of change is no longer close to a linear function. Therefore, with ROC_15 and ROC_20 as predictors, this model may not be a good model in the long run.

Eventually, we are developing linear models in this project by implementing shrinkage methods, subset selections and dimension reduction. However, although linear models are easy to interpret, they may not be complex and accurate enough. Some more complex models such as support vector machines, bagging and random forests may also be tried.

7. Conclusion

To sum up, after applying the shrinkage methods, subset selections and dimension reduction, we selected the model from backward stepwise selection, which contains 7 predictors. Under the model, ROC_15, ROC_20, EMA_50, $DTB3$, $DTB6$, $DGS10$ and $DAAA$ serve as the predictors for the index price of the S&P 500 Index. Therefore, apart from the historical performance of the index, the S&P 500 Index investors should also focus on the performances of other financial instruments, such as Treasury Bills and corporate bonds.

References

Hayes, A. (2021, September 30). How a capitalization-weighted index works and stocks impact it. Retrieved from https://www.investopedia.com/terms/c/capitalizationweightedindex.asp

Hoseinzade, E., & Haratizadeh, S. (2019). CNNpred: CNN-based stock market prediction using a diverse set of variables. *Expert Systems with Applications, 129*, 273-285.

IG. (2021, July 5). What's the relationship between stocks and bonds? Retrieved from https://www.ig.com/en/trading-strategies/what_s-the-relationship-between-stocks-and-bonds--210702

VanderPlas, J. T. (2017). *Python Data Science Handbook: Essential Tools for working with data*. Sebastopol, CA: O'Reilly Media.

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications, 67*, 126-139.