A Portfolio Strategy Based on XGBoost Regression and Monte Carlo Method

Mingxuan Wang[1,*]

[1]*Department of Mathematics, University College London, UK*
[*]*Corresponding author. Email: mingxuan.wang.20@ucl.ac.uk*

ABSTRACT

In this research, XGBoost algorithm was used to choose stocks. The stock data was downloaded from Yahoo Finance. The volumes, the differences of open price and close price, the differences of high price and low price, the adjusted close prices of the previous three days were considered as factors. Based on XGBoost, the data were segmented and trained to obtain the importance of each factor for each stock. The price of the previous three days is the most important factor for most stocks. In addition, RMSE and MAPE were calculated. After selecting the stocks with the minimum MAPE, the mean variance portfolio optimization model and the Monte Carlo method were used to find a range of portfolio weights of each stock in the stock pool. The return was calculated under the condition of reducing the risk. When the weights of the stock portfolio with the maximum Sharpe ratio are applied to the next year, the portfolio will achieve higher returns. Therefore, the model can be considered as a suitable tool to help investors implement better portfolio strategies.

Keywords: *XGBoost, Monte Carlo, Portfolio Strategy, ETF, machine learning, Optimization*

1. INTRODUCTION

The prediction of stock market price is a topic of concern to stock investors. The change trend of stock price is a very complex nonlinear system. It is not only affected by many factors, but also affected by random events to a large extent, which makes the prediction of price more difficult. With the development of machine learning, researchers try to use different machine learning algorithms to explore the prediction of stock market prices. XGBoost is a new machine learning algorithm developed by Tianqi Chen [1]. However, there is little research on stock portfolio based on XGBoost. This research attempts to propose a portfolio strategy by studying the leading stocks of the 30 large companies and ETF industry stocks in the United States through XGBoost algorithm and Monte Carlo. A number of risk indicators are calculated to test the rationality of the model.

At a certain risk level, investors expect the maximum return. Correspondingly, at a certain income level, investors expect the minimum risk. The Mean-Variance (MV) model is a risk investment model, which was proposed by Markowitz. In addition, the efficient frontier theory was proposed in the Markowitz model. By calculating the efficient frontier, risk and expected return can be reflected in the same figure [2]. Sharpe Ratio is another important indicator when measuring the risk and return. It can be used to calculate how much excess return a portfolio will generate for each unit of total risk [3].

Li and Zhang studied the dynamic weighting based on the XGBoost model, which proved this model can improve the performance of a multi-factor stock selection strategy. They considered IC coefficients as their factors and calculated the correlations between different parameters [4]. Kim used the XGBoost and mean-variance portfolio model to show the results of return good performance compared to another traditional method [5]. Uras and Ortu used SVM, XGBoost, CNN, and LSTM to predict the price of bitcoin. From this research, the XGBoost model is very good in the prediction results of bitcoin [6]. Fan et al. showed that the Monte Carlo method is the simplest way to compute risk measures [7].

They predicted stock prices through an integrated method of the genetic fuzzy system (GFS) and artificial neural network (ANN). In addition, they used stepwise regression analysis (SRA) to determine the factors that have the greatest impact on stock prices and used MAPE as an indicator to evaluate the model. The method in the

Y. Jiang et al. (Eds.): ICEDBC 2022, AEBMR 225, pp. 896–902, 2022.
https://doi.org/10.2991/978-94-6463-036-7_132

research was superior to all previous methods in results [8].

In the research of asset allocation portfolios, there are few studies using machine learning and deep learning to study the difference between investment return and optimal asset allocation results in the past. Day and Lin have developed a robot consultant to help investors optimize their portfolios to achieve higher returns. This optimization model can help investors make better decisions when the market environment is more complex [9]. Hill studied Bogle's description of ETF trading strategy. In addition, for institutional investors and retail investors, he also proposed that ETF is of great significance in promoting strategic portfolios [10].

This research will use XGBoost to train and test the historical data for five consecutive years. Improving returns and minimizing risks are the core objectives of our stock portfolio strategy. By comparing with the actual stock market data, the value of MAPE is calculated. The smallest ones are selected to enter the final stock pool. Then, Monte Carlo is used to calculating the five-year optimal weight when considering volatility and Sharpe ratio The weight is applied to the next year to measure the return of the portfolio.

2. METHOD

For studying a portfolio strategy, the research plan is divided into two steps. The first step is to select the appropriate stocks in the stock pool. The second step is to optimize the portfolio of the selected stocks.

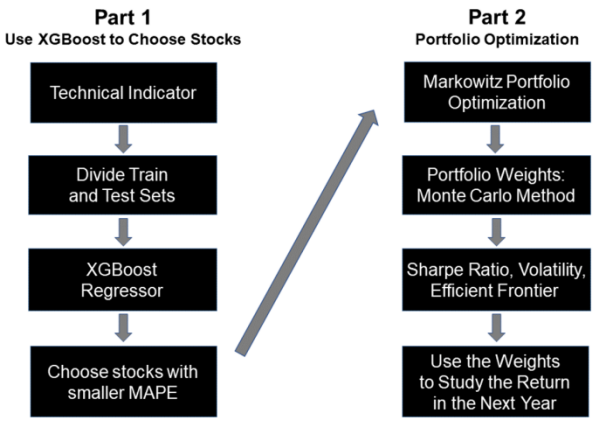

Figure 1 The outline of how to choose a suitable portfolio in this research.

Two different initial stock pools are considered in this research. One of the stock pools consists of the stocks of the 30 large-cap companies in the United States. The stocks of leading enterprises can better reflect the changing trend of the stock price of the whole market, compared with small and medium-sized enterprises. The rise and fall of leading stocks in different industries can reflect a trend in the whole stock market. That is the reason why the stocks of the leading cap companies are chosen as research objects. According to the Global Industry Classification Standard, the universe of stocks is split up into 11 sectors. Another stock pool consists of 11 different ETF industry factor stocks, which include energy, materials, industrials, consumer discretionary, consumer staples, health care, financials, information technology, real estate, communication services and utilities.

Table 1. The ETF Industry stocks in this research

ETF Industry Classification	ETF Stocks
Vanguard Information Technology ETF	VGT
Vanguard Real Estate ETF	VNQ
Financial Select Sector SPDR ETF	XLF
Health Care Select Sector SPDR ETF	XLV
Energy Select Sector SPDR ETF	XLE
Industrial Select Sector SPDR ETF	XLI
Consumer Discretionary Select Sector SPDR ETF	XLY
Consumer Staples Select Sector SPDR ETF	XLP
Utilities Select Sector SPDR ETF	XLU
Materials Select Sector SPDR ETF	XLB
Vanguard Communication Services ETF	VOX

** SPDR ETF focuses on market capitalization and industry sectors in the S&P 500.*

Next, several risk indicators to test the rationality of the model were calculated. In the end, the calculation results were analyzed and a portfolio strategy was proposed.

Suppose there are m kinds of risk assets in the market, and the return rates of assets are $r_1, r_2, \ldots, r_m$. The allocation weights of investors on each risk asset are $\omega_1, \omega_2, \ldots, \omega_m$ respectively.

Then, the total return r_{total} can be calculated as

$$r_{total} = \sum_{k=1}^{m} r_k \omega_k, \quad \text{where} \sum_{k=1}^{m} \omega_k = 1 \tag{1}$$

Then, the expected return of the portfolio is

$$E(r_{total}) = \sum_{k=1}^{m} \omega_k E(r_k) \tag{2}$$

Applying the formula to this research, if the weight and daily return of each stock are known, the accumulated annual return in the portfolio can be calculated.

Assuming that the initial asset of the investor is W_0, the future asset is $W_0(1 + r_{total})$. If the utility level of the investor is only related to the asset level, the utility level $U(r_{total})$ is also a random variable. From the perspective of maximizing expected utility, the decision-

making process of investors is shown by the following formula.

$$\max_{\omega_k} E[U(W_0 r_{total})], \quad where \sum_{k=1}^{m} \omega_k = 1 \tag{3}$$

Since W_0 is a fixed number, the aim can be transformed to maximize the value of $E[U(r_{total})]$. What is more, if $r_1, r_2, \ldots, r_m$ follow a normal distribution, the total expected utility only depends on the mean and the variance of the total portfolio returns.

All of the stock data was imported from Yahoo Finance. It concluded all the open prices, close prices, high prices, low prices, adjusted close prices, and volumes from 2014 to 2021 in the stock pool. Then, the differences between open prices and closed prices were calculated every day. The differences between high prices and low prices were also calculated. There are many factors affecting the stock market and the fluctuation of the stock price is greatly affected by news. This includes many uncontrollable factors. Although there are few factors selected in our model, it is good to predict the characteristics of the stock market by using fewer data. This is another advantage of XGBoost, which means classifying and predicting with the least data possible. Data standardization is an important step before adopting XGBoost. We calculate the standardized data and segment it. It is proved that $Var(r_{total})$ is a quadratic function of ω. The stocks with the top four minimum MAPE was chosen to consist the portfolio. The efficient frontier and optimal weights for a given portfolio can be obtained in the five past years. Then, the weights were used to test the portfolio return in the next year.

2.1. XGBoost

This research adopts XGBoost to select the appropriate stocks in the stock pool. XGBoost (extreme gradient boosting) is an algorithm or engineering implementation based on GBDT. XGBoost can solve real-world problems with a minimal number of resources [2]. In addition, this machine-learning algorithm has been widely used in major competition platforms in recent years. That is the reason why XGBoost was chosen as the algorithm to study the strategy of the portfolio in this research. Moreover, XGBoost explicitly adds regular terms to control the complexity of the model, which can effectively prevent overfitting. The objective function of XGBoost consists of a loss function and a regularization term. This research will not prove the mathematical principle of XGBoost repeatedly. The XGBoost database in python was directly used. Another advantage of XGBoost is that it can be processed in parallel, which can greatly reduce the amount of computation. The volumes, the differences between the open price and close price, the differences between high price and low price, and the adjusted close prices of the previous three days were considered as factors. The standardized data was uniformly calculated and segmented. Next, the training data and test data were generated. Due to the different standardization methods of training data and test data, it is necessary to segment training and test data. In the process of standardization, the information should not be accompanied by test sets to avoid information leakage.

2.2. Monte Carlo

The Monte Carlo method can calculate a large number of random results and compare them to obtain the most appropriate solution. After determining the stocks in the portfolio, the data of the past five years are used to simulate the weight of each stock under the given indicators by Monte Carlo method. There are two indicators in this step. The first indicator is the weights of each stock with the maximum Sharpe ratio. The second indicator is the weights of each stock with the minimum volatility in the portfolio. Since the Sharpe ratio measures the rate of return per unit risk, it is good if the Sharpe ratio will be greater under a given risk. Generally, when the value of the Sharpe ratio is greater than one, it indicates that the stability of the model is relatively good. Based on the Monte Carlo simulation of the above two indicators, two weights portfolios of each stock can be obtained. Then, the return can be calculated by a certain weight of each stock. To maximize returns, the weights with the maximum Sharpe ratio are chosen as the final weights to study the return in the next year. To make the results accurate enough, a large number of simulations need to be calculated. The number of simulations was set as 100000 in this research.

3. RESULTS

The stocks of the 30 largest cap companies and ETF industry stocks are studied in this part respectively. The stock of head enterprises can reflect the trend of the whole market to a certain extent. Considering the diversity of stock selection, ETF stocks are also studied in the same way to avoid the limitations of stock selection.

In the prediction process, the stability of the model is determined by MAPE. The data of the first five years are used for training and test sets. Among them, the proportion of the test set is set to 0.2. Thus, the test results can be compared with the actual situation and MAPE can be calculated. The stocks with the top 15% minimum MAPE are selected into the stock pool of the final portfolio. Next, the Monte Carlo method is used to calculate the weight of each stock when the Sharpe ratio is the largest in five years. As a result, the weight of each stock is applied to the next year to predict the new return.

3.1. Leading stocks

The selected leading stocks are listed in Table 2. These stocks also mean the stocks with large companies by market cap. In addition, the stocks of large companies will affect the environment of the whole industry and

even the market. That is the reason why the leading stocks were chosen as the research objects.

Table 2. The selected leading stocks

AAPL	MSFT	GOOG	AMZN	TSLA
NVPA	FB	UNH	JNJ	V
JPM	WMT	PG	XOM	BAC
HD	MA	CVX	BABA	KO
PFE	ABBV	LLY	COST	AVGO
DIS	PEP	TMO	CSCO	VZ

Firstly, the stock data from 2016-2020 were studied by XGBoost regression. The adjusted close values of apple stock in the test set were shown as an example in Figure 2. The same things we have done 30 times.

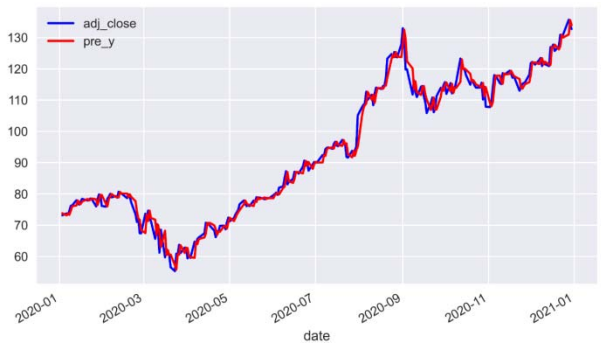

Figure 2 The predicted adjusted close values about AAPL by XGBoost

In terms of importance analysis, the prices of the previous days are the most important variable affecting the stock price of the day based on the XGBoost regression of each stock. In addition, the adjusted close prices of the previous three days are more important than other factors to affect the stock price on the next day. For the different 30 stocks, the root means square error (RMSE) and the mean absolute percentage error (MAPE) were calculated. If the values of RMSE and MAPE are smaller, it also means that the error is smaller and the model is better. The top 15% smallest RMSE is 0.702, 0.830, 0.980, and 0.998, which present the stocks of KO, BAC, CSCO and PFE. However, the top 15% smallest MAPE is 1.183%, 1.193%, 1.214%, 1.235%, which presents the stocks of COST, JNJ, PG and WMT. Finally, the stocks with the top 15% smallest MAPE were selected into the portfolio stock pool.

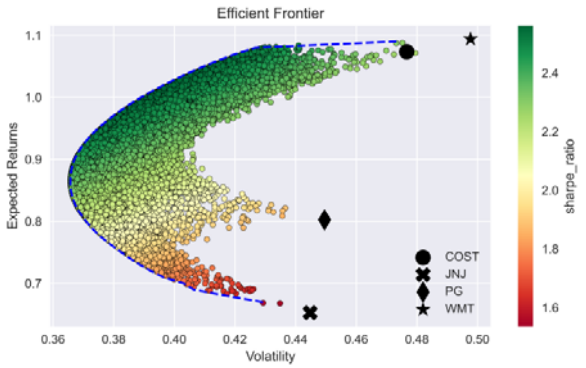

Figure 3 Efficient Frontier for COST, JNJ, PG and WMT

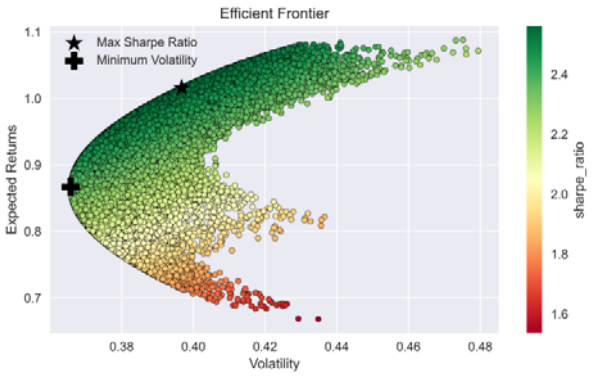

Figure 4 Efficient Frontier of the portfolio with the max Sharpe ratio and min volatility

Then the return, volatility and Sharpe ratio were studied when the portfolio got the minimum volatility and maximum Sharpe ratio. When the portfolio got the maximum Sharpe ratio, the weight of COST, JNJ, PG and WMT are 41.16%, 4.33%, 17.80% and 36.72%. Correspondingly, the returns, volatility and Sharpe ratio are 101.66%, 39.61% and 256.62% in the five years.

When the portfolio got the minimum volatility, the weight of COST, JNJ, PG and WMT are 21.47%, 33.62%, 25.64% and 19.27%. The returns, volatility and Sharpe ratio are calculated as 86.78%, 36.58% and 237.25% in the five years.

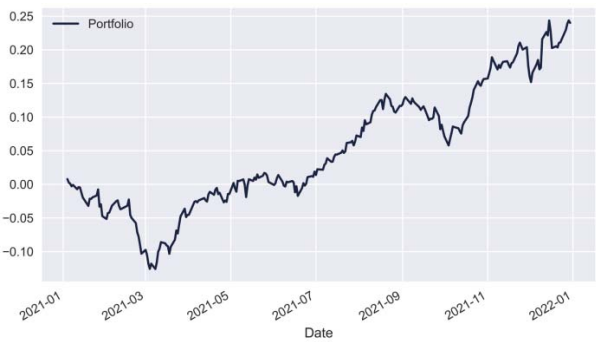

Figure 5 The return of the leading stocks portfolio in 2021

From above, there are two different portfolio weights. Since there is little difference between the volatilities of the two portfolios, the weights with the maximum Sharpe ratio are chosen as the final weights to predict the stock prices in the next year. According to the stock prices, the daily stock returns can be calculated. Finally, when the weights of each stock are used in the portfolio, the cumulative return is 24.01% in 2021.

When studying the stock data for 2015-2019, the MAPE between the predicted prices and the actual prices in 2019 can be calculated by training the data for 2015-2018. Similarly, the top 15% of stocks can be obtained, which are JNJ, KO, PEP and WMT respectively. The final cumulative return is 11.76% in 2020.

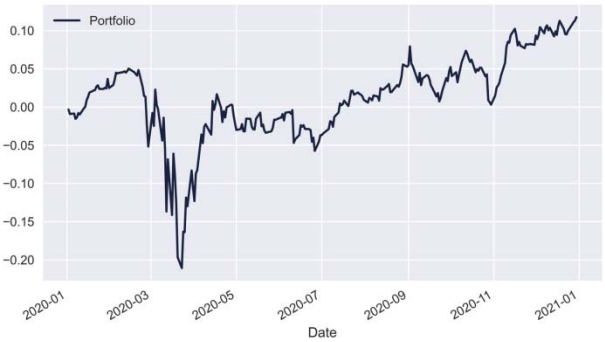

Figure 6 The return of the leading stocks portfolio in 2020

Next, the stock data for 2014-2018 were used to study the portfolio strategy in 2019. In 2019, the portfolio return is 11.24%, which is not very large. The results will be compared and analyzed later.

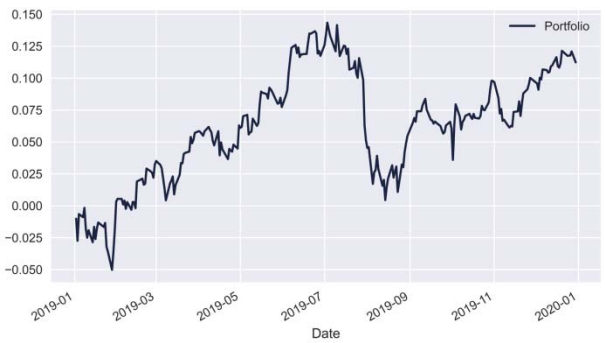

Figure 7 The return of the leading stocks portfolio in 2019

3.2. ETF industry stocks

A similar method was used to predict the ETF stock prices. The proportion of the test set is 0.2 and the ETF stock prices are studied in the past five years. By XGBoost regression, the top four smallest MAPE are VOX, XLP, XLV and XLY when comparing the predicted and the actual stock prices. In the ETF stock pool, the stock prices with the top four smallest MAPE are VOX, XLP, XLV and XLY. Thus, the weights of the above stock are 0.30%, 3.23%, 2.76% and 93.70% when the portfolio gets the maximum Sharpe ratio. The weights when the portfolio gets the minimum volatility are also calculated and the final portfolio return is 18.93%. Because the Sharpe ratio is 199.24%, it shows that the return brought by the model is larger than the loss brought by risk in the past five years. To obtain greater returns, the weights with the maximum Sharpe ratio are chosen as the final portfolio weights in 2021. Taking the weights into the markets in 2021, the cumulative return of the portfolio is 27.86%.

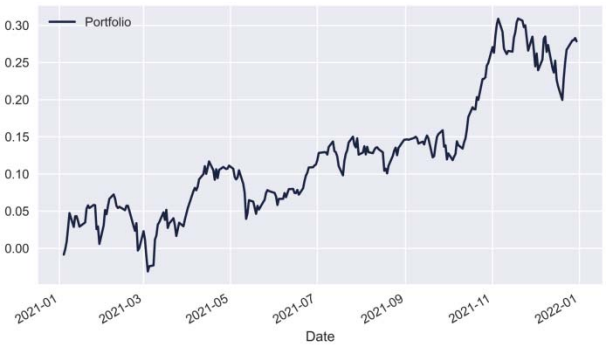

Figure 8 The return of the ETF portfolio in 2021

By using the XGBoost algorithm, the stock data from 2015 to 2018 were trained and the data in 2019 were tested. The ETF stocks with the top four smallest MAPE are VNQ, XLP, XLU and XLV. In addition, the weights of each stock are 0.03%, 13.88%, 46.13% and 39.96%. The returns, volatility and Sharpe ratio are 51.87%, 25.39% and 204.28% in the five years (2015-2019). After taking the weight into the daily adjusted close prices in 2020, the final cumulative portfolio return is 6.20%.

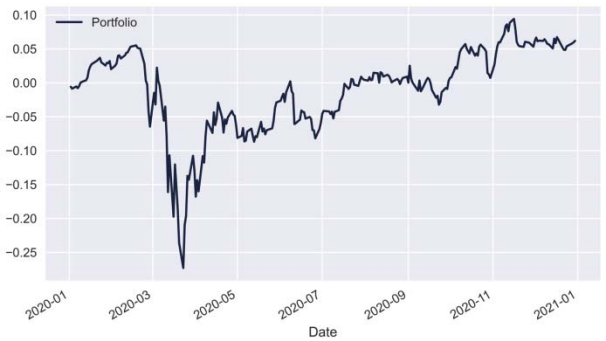

Figure 9 The return of the ETF portfolio in 2020

In the end, the portfolio strategy was studied in 2019. The ETF stocks with the top four smallest MAPE are XLP, VNQ, XLU and XLV. The weights are 0.80%, 0.19%, 51.37% and 47.64% when the portfolio got the maximum Sharpe ratio between 2014 to 2018. By calculating the cumulative returns, the portfolio return is 23.26% in 2019.

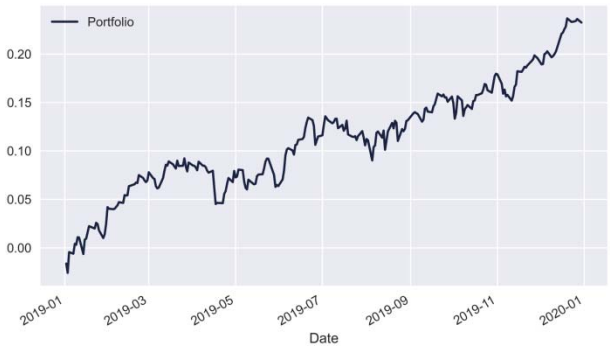

Figure 10 The return of the ETF portfolio in 2019

4. DISCUSSION

In this part, the different results of returns will be compared and analyzed under the market environment. In 2019 and 2021, the return of the ETF portfolio is larger than that of the leading stocks portfolio. Especially in 2019, the return of ETF portfolio reaches 23.26% and that of leading stocks portfolio is 11.24%. However, the return of the leading stocks portfolio is larger than the return of the ETF portfolio in 2020.

From Figures 6 and 9, the portfolio prices have decreased significantly since February 2020. Besides, the return of the ETF portfolio is less than the return of a leading stock portfolio. The ETF industry sectors, which have been relatively stable in the past 5 years, suffered heavy losses in 2020. The market trend for the whole year of 2020, especially in the first half of the year, can be summarized by "structural market imbalance". A structured market is related to the cyclical development of enterprises. In addition, the structural market is also related to the prosperity of the industry. Historically, the sharp rise in US stocks in 2019 has led to a high price at the beginning of 2020. In addition, the United States does not pay attention to the COVID-19 epidemic at the beginning of 2020 and the data is not public, which also leads to a negative effect on the U.S. stock market. A series of policies of the Federal Reserve also caused investors to panic. Since April, the stock price has rebounded rapidly and is close to the level at the beginning of the year. The stock price gradually rose in the third and fourth quarters. The model is based on historical data to improve the stability of the model while reducing risks and calculating the optimal weight of the stock selected by XGBoost in the past five years. Although the return is relatively small compared with other years and some stocks with better performance, it is very good to achieve a certain amount of positive return at the end of the year under the impact of the epidemic.

In 2019, a large number of funds poured into the ETF market. The whole market shows that the performance of passively managed stocks is better than that of actively managed stocks. This also led to a higher stock price in 2019, which is one of the reasons for the sharp decline of the stock price in early 2020.

In 2021, the return of the ETF portfolio is larger than the return of the leading stocks portfolio. With the development of vaccines and the adjustment of policies, the whole market environment began to recover. The return of the leading stocks portfolio fell first and gradually increased after March. In addition, the highest return for an ETF portfolio is more than 30% from Figure 8.

From the above results, the results of the model have achieved great returns in practice. The fluctuation of the stock price is inseparable from the actual situation of the market. The superiority of the results also proves the feasibility of the XGBoost stock selection strategy.

5. CONCLUSION

This research imported the past stock data of Yahoo Finance and standardized the data. Then, XGBoost was used to train and test stock data for five consecutive years. By comparing with the actual stock data, the four stocks with the smallest MAPE were selected for an investment portfolio. Next, the Monte Carlo method was used to study the volatility, Sharpe ratio and return of the selected stocks. After comparison, the weight when the Sharpe ratio reached the maximum value was used to calculate the return for the next year. The stock data for five consecutive years was used to calculate the return of the portfolio in 2021, 2020 and 2019. The returns of the leading stocks portfolio are 24.01%, 11.76% and 11.24% in 2021, 2020 and 2019. The returns of the ETF portfolio are 27.86%, 6.20% and 23.26% in 2021, 2020 and 2019. The model research objects, whether ETF stocks or leading stocks, eventually have a relatively high return. Therefore, the model can be considered a suitable tool to help investors implement better portfolio strategies.

REFERENCES

[1] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785-794. DOI: https://doi.org/10.1145/2939672.2939785

[2] H. Markowitz, Portfolio Selection, Journal of finance, no. 7, 1952, pp. 77–91. DOI: https://doi.org/10.2307/2975974

[3] A. W. Lo, The statistics of Sharpe Ratios. Financial Analysts Journal, 2002, vol.58, no.4, pp.36–52. DOI: https://doi.org/10.2469/faj.v58.n4.2453

[4] J. Li, R. Zhang, Dynamic weighting multi factor stock selection strategy based on XGboost machine learning algorithm. In 2018 IEEE International Conference of Safety Produce Informatization

(IICSPI), IEEE, 2018, pp. 868-872. DOI: 10.1109/IICSPI.2018.8690416

[5] H. Kim, MEAN-VARIANCE PORTFOLIO OPTIMIZATION WITH STOCK RETURN PREDICTION USING XGBOOST. Economic Computation Economic Cybernetics Studies & Research, 2021, vol.55, no.4. DOI: 10.24818/18423264/55.4.21.01

[6] N. Uras, M. Ortu, Investigation of Blockchain Cryptocurrencies' Price Movements Through Deep Learning: A Comparative Analysis. In 2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), IEEE, 2021, pp. 715-722. DOI:10.1109/SANER50967.2021.00091

[7] G. Fan, Y. Zeng, W. K. Wong, Decomposition of portfolio VaR and expected shortfall based on multivariate Copula simulation. International Journal of Management Science and Engineering Management, 2012 vol.7, no.2, pp.153-160. DOI: https://doi.org/10.1080/17509653.2012.10671219

[8] E. Hadavandi, H. Shavandi, A. Ghanbari, Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. Knowledge-Based Systems, 2010 vol.23, no.8, pp.800-808.DOI: https://doi.org/10.1016/j.knosys.2010.05.004

[9] M. Y. Day, J. T. Lin, Artificial intelligence for ETF market prediction and portfolio optimization. In Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, 2019, pp. 1026-1033. DOI: https://doi.org/10.1145/3341161.3344822

[10] J. M. Hill, The evolution and success of index strategies in ETFs. Financial Analysts Journal, vol.72, no.5, 2016, pp.8-13. DOI: https://doi.org/10.2469/faj.v72.n5.2