Portfolio Construction Based on XGBoost-CAPM Model: Evidence from the Cryptocurrency Market

Jintong Yang

School of Finance and Trade, Zhuhai College of Science and Technology, Zhuhai 519041, China

YangJT16@stu.zcst.edu.cn

Abstract. This study investigates the construction of an optimal investment portfolio in the cryptocurrency market using the XGBoost algorithm and the CAPM model. To achieve this objective, one selected Bitcoin, Ethereum, Litecoin, and Tether as representative assets in the cryptocurrency market and used the CMC200 as the market index. The study primarily employed the XGBoost algorithm to predict the returns of the market index, while using the CAPM model and OLS regression analysis to calculate the alpha and beta of each asset. Based on the predicted market index returns and the alpha and beta values of each asset, this paper further calculated the expected returns of each asset. For portfolio optimization strategies, one used the maximum Sharpe ratio and minimum volatility as optimization goals to determine the weights of the optimal investment portfolio. The results indicate that by combining the XGBoost prediction model, CAPM theory, and portfolio optimization strategies, one can build investment portfolios in the cryptocurrency market with higher returns and lower risk, providing valuable guidance for cryptocurrency investors.

Keywords: Portfolio Construction, Cryptocurrency, CAPM Model, XGBoost, Time series.

1 INTRODUCTION

Portfolio construction is one of the important problems in the field of finance. Investors allocate funds to different assets to achieve the goal of risk diversification and maximizing returns. The XGBoost-CAPM model combines the ideas of the XGBoost algorithm in machine learning and the CAPM model, which can better estimate the volatility of asset returns and predict the expected future returns, thus constructing investment portfolios more effectively. This study aims to explore the portfolio construction method based on the XGBoost-CAPM model and evaluate its application effect in real investment. Since its inception, the cryptocurrency market has become a significant direction in the investment field. As the investment market becomes more sophisticated and increasingly complex, more research is devoted to exploring how to effectively construct a variety of investment portfolios for maximizing returns and minimizing risks.

R. Magdalena et al. (eds.), *Proceedings of the 2024 9th International Conference on Social Sciences and Economic Development (ICSSED 2024)*, Advances in Economics, Business and Management Research 289,
https://doi.org/10.2991/978-94-6463-459-4_16

The founder of modern portfolio theory, Markowitz, was the first to propose the famous mean-variance theory, laying the foundation for portfolio research [1]. American economist Sharpe et al. proposed the CAPM model (Capital Asset Pricing Model) based on the mean-variance theory, reducing the computational process required for portfolio construction optimization [2]. To enhance the practical applicability of this model, Jensen and other scholars added a constant α (Jensen's alpha) to the formula of the CAPM model, which has continued to be used in portfolio management research [3]. Stanisław used the classic CAPM model to calculate the cost of equity capital in a portfolio of stocks listed on the Warsaw Stock Exchange, demonstrating the diversity of the CAPM model in quantitative financial research [4]. Using the CAPM model alone to construct a portfolio, due to the limitations of using historical data, cannot satisfy the increasingly rich and complex financial market and investors' growing demand for quantitative precision. Researchers have started combining the CAPM model with other prediction models for the quantification of investment portfolio returns and risks. Scholars combined the ARMA time series estimation model and the CAPM model along with a Nadaraya-Watson transform to obtain the regression equation for individual stocks and conducted empirical research on portfolio optimization [5]. Chen researched the prediction models combined with the CAPM model and concluded that the equation fitted by the GARCH model and the CAPM model gave the best results [6].

As machine learning is increasingly applied to the field of financial quantification, it has also started to probe into portfolio construction optimization. Tianqi Chen and Carlos proposed a scalable tree boosting system, the XGBoost algorithm in machine learning, which has improved the accuracy of time series forecasting to a certain extent [7]. Yin proposed a hybrid prediction method using three machine learning models to improve traditional portfolio models, aiming to reduce forecasting errors [8]. Min combined robust optimization and other machine learning algorithms with portfolio models, and found that the portfolio's return and risk performance were significantly improved to a certain extent [9]. Li used various machine learning algorithms to conduct strategy research on China's stock portfolio, and found that the prediction effect of XGBoost was better as compared to other algorithms [10]. In the existing literature, there is a wealth of research on portfolio construction for the stock market, with a wide range of methods. The Capital Asset Pricing Model (CAPM) and methods for historical data time series prediction of relevant indicators are relatively well developed. Researchers are also committed to quantifying the optimization of portfolio construction using more accurate methods. In contrast, there is not as much research on portfolio construction and optimization in the cryptocurrency market. While most scholars focus on portfolio research in fields such as stocks and bonds, Rudys et al. have been studying the life cycle of cryptocurrency portfolios [11].

The rapid development of the cryptocurrency market and the gradual rise in trading volume and high interest rates and risks have attracted significant attention from investors around the world. However, the inherent volatility and unpredictability of the cryptocurrency market has raised concerns about the best investment strategies to mitigate risk and maximize wealth. Traditional investment models such as the CAPM alone are often unable to cope well with the uniqueness of the cryptocurrency market. In this

context, there is a need to explore and develop innovative approaches to constructing optimal portfolios for the cryptocurrency market to ensure that investment results are more reliable, efficient, and more in line with actual market conditions. Firstly, the time series prediction of the market index return is performed by the XGBoost algorithm, which is fitted to the relevant parameters calculated by the CAPM model to obtain the expected return of each asset in the constructed portfolio. Using the optimization method based on the scipy package in python, the weight size of each asset is constrained with the sum size to obtain the optimal asset weight ratio. Define whether the optimal portfolio is reached by maximizing the Sharpe ratio or minimizing the volatility to determine the final optimal portfolio. And draw the efficient frontier.

2 DATA AND METHOD

2.1 Model Selection and Principles

An important indicator in the CAPM model, the Sharpe ratio, is calculated by the formula

$$Sharpe\ Ratio = \frac{R_i - R_f}{\sigma_r} \tag{1}$$

Assuming the asset is i,

$$E(r_i) = r_f + \beta_{im}[E(r_m) - r_f] \tag{2}$$

In the equation, $E(r_i)$ represents the expected return of asset i; r_f represents the risk-free rate; β_{im} represents the systematic risk coefficient of asset i, Its calculation formula is$\beta = \frac{\mathrm{Cov}(r_i, r_m)}{Var(r_m)}$;$E(r_m)$ is the expected return of the market portfolio; and the difference between $E(r_m)$ and r_f is called the market risk premium. For the value of the risk-free rate of return r_f, many domestic and foreign scholars approximate the value of the interest rate on short-term treasury bonds instead, because the risk-free rate is usually subject to the condition that there is no credit risk of default, and the default rate on treasury bonds is close to zero in the investment market. In order to reduce the errors caused by price volatility, the daily closing prices of various assets in the investment portfolio will be processed by calculating the logarithmic returns. Logarithmic returns are continuous compounded returns, which make the analysis more meaningful in practice. Its calculation formula of the rate of return at time t is:

$$r_t = ln(1 + R_t) = ln(\frac{P_t}{P_{t-1}}) = ln(P_t) - ln(P_{t-1}) \tag{3}$$

In 1968, Jensen introduced the parameter α to construct an empirical model that is more commonly used in practical applications:

$$R_{it} - R_{ft} = \alpha_i + \beta_i(R_{mt} - R_{ft}) + \varepsilon_{\mathrm{it}} \tag{4}$$

The portfolio is the main focus of this study. As such, one introduces asset weights denoted as w. To obtain the return of the entire portfolio, one simply needs to sum the product of each asset's return and its weight. Let i be the portfolio of assets. The formula for calculating the portfolio is:

$$r_i = \sum_{x=1}^{n} w_x r_x \tag{5}$$

All investors typically hope to achieve maximum returns with minimum risk when investing their funds in the market. In such situations, the Capital Asset Pricing Model (CAPM) can be used to estimate the expected return rate and systematic risk coefficient of asset portfolios. Ultimately, an optimal investment portfolio can be obtained through optimized asset weights. Two essential concepts in the CAPM model are the efficient frontier, the capital market line(CML) and securities markets line(SML). The Fig. 1 features the Efficient Frontier, which represents the set of portfolios that offer the highest expected return for every level of risk. The curve of the Efficient Frontier shows the optimal combination of different assets that an investor can choose from. The Capital Market Line (CML) is a straight line that connects the risk-free asset and the tangency point of the Efficient Frontier. It represents the optimal portfolio that balances risk and return based on the investor's risk preference.

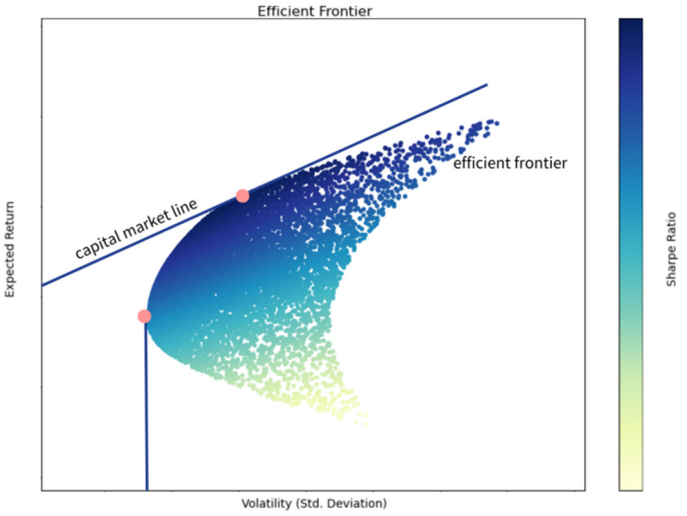

Fig. 1. An example graph depicting the Efficient Frontier and Capital Market Line (CML) (Photo/Picture credit: Original).

In this paper, the XGBoost machine learning algorithm will be used for time series data forecasting of expected returns on assets and their corresponding predicted returns on market indices. Therefore, the main focus is on the process of applying the XGBoost algorithm to univariate time series historical data forecasting. XGBoost is an efficient integrated learning algorithm that continuously improves the prediction accuracy of the model in an iterative manner by integrating multiple weak learners. In this case, this

study will use the XGBoost algorithm to construct a model to predict the future return of an asset. Firstly, one prepared time series data for training and testing the model and smoothed the data to ensure that the time series are stable. Next, useful features were extracted from the raw time series data as input variables for the XGBoost model. The dataset was then divided into a training set and a test set, with earlier data used for model training and newer data used for validation and testing. Then, the algorithm was used to train the model on the training set and the hyperparameters were optimised to achieve better model performance. Finally, the trained and validated models are used to predict the future returns of the assets.

2.2 Data Processing

By analyzing the current cryptocurrency market and its historical development, this study has selected Bitcoin(BTC), Ethereum(ETH), Tether(USDT), and Litecoin(LTC) as representatives in the cryptocurrency asset portfolio. This study has downloaded the trading data for these four cryptocurrencies from November 16, 2020, to November 16, 2023, from Investing.com. This dataset adequately reflects the sufficiency of the observation period and the representativeness of asset selection in the cryptocurrency market, aligning with the asset selection mindset commonly used by most investors. In the cryptocurrency market, there are no standard market indices that are as widely accepted as in traditional financial markets. However, several organizations and platforms offer their own cryptocurrency market indices to track and measure the performance of the cryptocurrency market as a whole. This paper will choose the CMC Crypto 200 Index, launched by CoinMarketCap, which is already listed on the Nasdaq exchange, as a benchmark to measure the market performance of the portfolio. The risk-free rate is considered to be the level of interest rate on an asset that has no risk. It represents the expected rate of return on a risk-free asset available to investors. In many studies, the risk-free rate is usually measured by using government bonds or bank time deposits as risk-free assets. This is because such assets are usually considered to have low risk. However, due to the recent behavior of the Federal Reserve's frantic interest rate hikes, both its treasury bond and time deposit interest rates are inflated, so this paper decides to formulate the risk-free rate as 0. Besides, this study will calculate the daily logarithmic returns between November 17, 2020, and November 16, 2023, using the closing prices of four cryptocurrencies on trading days from November 16, 2020, to November 16, 2023.

2.3 Model Assumption

CAPM modeling needs to be based on the following five most basic assumptions. Investors' utility is a function of their wealth investors seek to maximize their wealth and their decisions are influenced by investors' preferences for different levels of wealth. Investors are able to know in advance the probability distribution of investment returns and that the distribution is the same between different investment opportunities. Investment risk can be measured by the variance or standard deviation of the rate of return on investment. The greater the variance or standard deviation, the higher the volatility of

the rate of return and the greater the investment risk. Investors mainly consider the two aspects of expected rate of return and risk. The expected rate of return is the investor's expectation of future returns, while risk measures the investor's concern about potential losses. At the same level of risk, an investor will choose a security with a higher yield; at the same level of yield, an investor will choose a security with a lower risk.

3 RESULTS AND DISCUSSION

The prediction is done based on the chosen market index, CMC200. Firstly, one processes the weekly closing prices of CMC200 during the sample period, calculate the logarithmic return and remove all NaN values. The obtained processed data can be presented as a time series line graph as given in Fig. 2. As one can see from the Fig. 2 it is the data that is smoother and does not have any particular trend. The last year's weekly data is used as the test set for model evaluation. Assuming there are 52 weeks in a year, the value of MAE (Mean Absolute Error) is obtained as 0.0710. The line graph of expected and predicted CMC200 returns is as shown in the Fig. 3.

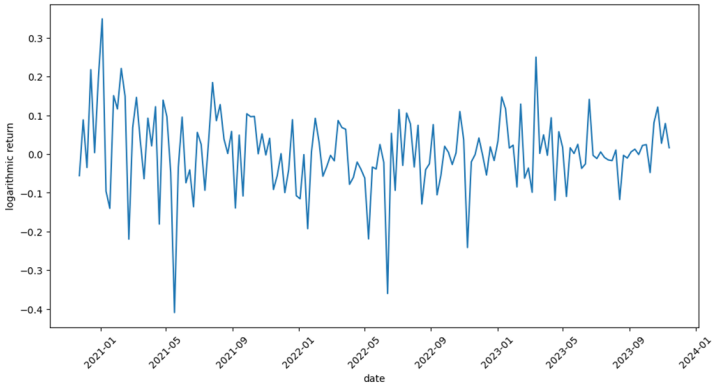

Fig. 2. Log return daily evolution (Photo/Picture credit: Original).

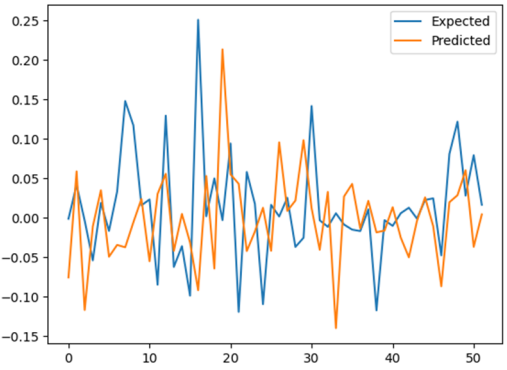

Fig. 3. Expected and predicted CMC200 returns (Photo/Picture credit: Original).

In XGBoost's univariate time series forecasting, by training and learning from the historical data, one finally gets the predicted market index return of 0.0179. this means that based on the algorithmic analysis and the current situation, it is expected that there will be a small increase in the return of market index. The weekly closing price data of BTC, ETH, LTC, and USDT in the sample range are exported from Yahoo Finance, and the corresponding log returns for each week are calculated. Construct an empirical model with a constant term α

$$R_i - R_t = \alpha_i + \beta_i(R_m - R_f) \tag{6}$$

Since the risk-free rate is formulated as 0, the final model is:

$$R_i = \alpha_i + \beta_i R_m \tag{7}$$

For each of the four assets, BTC, ETH, LTC and USDT, an OLS regression is performed against the market index CMC200 to obtain β, the intercept term α, and the decidability coefficient R2. The market returns predicted by the XGBoost algorithm are then substituted into the empirical model to obtain the expected return for each asset presented in Table 1.

Table 1. The coefficients for different underlying assets

	β	α	R^2
BTC	0.9434	3.691e-05	0.852
ETH	1.0919	0.0175	0.822
LTC	0.9634	-0.0132	0.579
USDT	0.0010	-7.217e-05	0.055

Table 2. The weight for 2 models.

	BTC	ETH	LTC	USDT
largest Sharpe ratio	0.0000%	71.9924%	0.0000%	28.0076%
minimum volatility	0.0000%	0.0000%	0.1549%	99.8451%

From the determination coefficients of each asset in the table above, one can see that the fit between the three assets (except USDT, which is a stable cryptocurrency) and the market index is quite good. Therefore, in the calculation of expected returns for USDT, the constant term α will be ignored, and the calculation will be done directly with β and the expected market index returns. The final calculated expected returns for BTC, ETH, LTC, and USDT are 0.0169, 0.0370, 0.0040, and 1.79e-07, respectively. Based on the expected returns, covariance matrix, and other metrics calculated, optimization of the optimal investment portfolio is confirmed using the scipy package, with the constraint that the sum of the weights of the four assets must equal 1, and each individual asset's weight must be within the [0,1] range. The weight of each asset in the portfolio with the largest Sharpe ratio and with minimum volatility are given in Table 2. The maximum Sharpe ratio value obtained is 0.1824. The minimum volatility value obtained is 0.0017. The optimized results show that portfolios with the largest Sharpe ratios have a higher weighting of ETH (71.9924%) and USDT (28.0076%), while BTCn or LTC both have an approximate share of 0. This implies that the optimal trade-off of high return versus risk does not require diversification across all the cryptocurrencies available to choose from but rather a concentration of investments based on the

expected return and risk characteristics of the specific assets. The composition of the Minimum Volatility Portfolio is almost entirely dominated by the stable cryptocurrency USDT (99.8451%), with a very small amount of LTC (0.1549%). This signals a conservative strategy with minimal risk, valuing portfolio stability over potentially high returns, resulting in a very low volatility value of 0.0017. One draws the efficient frontier and mark the points of maximum Sharpe ratio and minimum volatility as presented in Fig. 4.

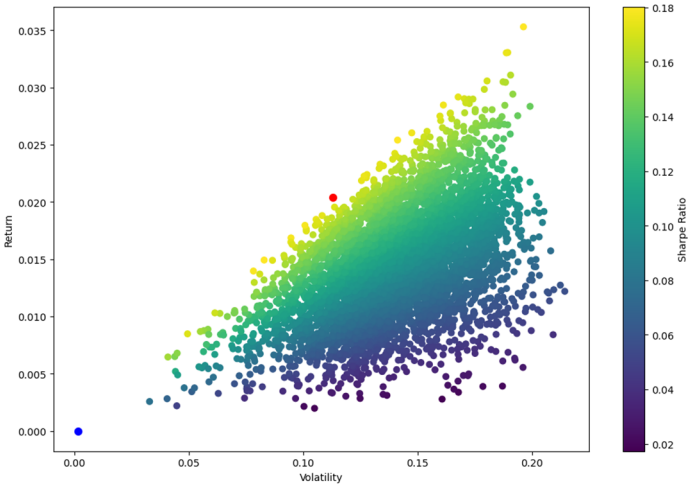

Fig. 4. The efficient frontier, the points of maximum Sharpe ratio and minimum volatility (Photo/Picture credit: Original).

4 LIMITATIONS AND PROSPECTS

Constructing portfolios based on the CAPM (Capital Asset Pricing Model) requires several assumptions and simplifications, such as the assumption that all investors have the same investment horizon and that the market is characterized by frictionless rationality. However, these assumptions may cause the model predictions to be inconsistent with reality. The cryptocurrency market, on the other hand, is usually characterized by high volatility and uncertainty, and price fluctuations can be significant. This makes portfolios difficult to evaluate and can generate extreme fluctuations in value over short periods of time. This leads to the possibility that various model predictions may not be valid. In this paper, the use of the XGBoost algorithm as a reliable predictor of cryptocurrency returns combined with the CAPM model for portfolio construction optimisation is promising. Future research could include refining the model and enriching the feature engineering by incorporating more market metrics, wider datasets, and real-time data to improve the accuracy of the predictions.

5 CONCLUSION

Through the research in this paper, it is shown that in the cryptocurrency market, which is known to be extremely volatile, a combination of traditional CAPM models with the modern XGBoost algorithm is able to compute portfolios that achieve either maximum return or minimum risk. With the rapid development of the cryptocurrency space and increased regulatory levels, more accurate and objective models that combine machine learning, and traditional financial investment models have the potential to provide investors with clearer and more useful guidelines for investment decisions in the future.

REFERENCES

1. Markowitz, H. M.: Foundations of portfolio theory. The journal of finance, 46(2), 469-477 (1991).
2. Sharpe, W. F.: Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 19(3), 425-442 (1964).
3. Michael, C. J.: The Performance of Mutual Funds in the Period 1945-1964[J].Journal of Finance, 23 (1968).
4. Urbański, S.: The Cost of Equity Capital in Stock Portfolios Listed on the Warsaw Stock Exchange Using the Classic CAPM. E-Finanse, 15(2), 48-62 (2019)
5. Chen, Z.: Application of CAPM-ARIMA Model and CAPM-GARCH Model in Investment Portfolio. Nanjing University of Finance and Economics (2021).
6. Parmikanti, K., Gw, S. H., Saputra, J.: Mean-Var investment portfolio optimization under capital asset pricing model (CAPM) with Nerlove transformation: An empirical study using time series approach. Industrial Engineering & Management Systems, 19(3), 498-509 (2020).
7. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining pp. 785-794 (2016).
8. Yin, X.: Research on Investment Portfolio Selection Based on Machine Learning. Information System Engineering, 12, 128-131 (2021).
9. Min, L.: Research on Investment Portfolio Model Assisted by Machine Learning Algorithm. Shanghai University of Finance and Economics (2021).
10. Li, L.: Research on Stock Investment Portfolio Strategy Based on Supervised Learning and Deep Reinforcement Learning. Central South University (2022).
11. Valentinas, R., Daniel, S.: Optimal cryptocurrency portfolio allocation over the life cycle. Applied Economics, 55(46), 5419-5433 (2023).