WEP Warwick Evans Publishing

International Journal of Global Economics and Management

ISSN: 3005-9690 (Print), ISSN: 3005-8090 (Online) | Volume 3, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijgem.v3n3.34
Journal homepage: https://wepub.org/index.php/IJGEM/index

Optimization of ETF Fund Selection Strategy Based on Machine Learning Scoring Model

Yanhao Huang

Department of management, Beijing Institute of Technology, Beijing, China
hyh396696@outlook.com

ABSTRACT

This study aims to optimize the ETF fund selection strategy. By combining the time-series momentum strategy and the machine learning scoring model, a systematic analysis and empirical research on the main stock ETFs and cross-border ETFs in the Chinese market are conducted. We constructed a combination strategy that combines momentum indicators and machine learning scoring, and conducted a backtest. The results show that the combination strategy is significantly better than the momentum strategy or machine learning model alone and the benchmark index in terms of annualized return, maximum drawdown and Sharpe ratio. By selecting momentum indicators (such as simple momentum, exponential moving average, relative strength index) and machine learning models (such as logistic regression, support vector machine, random forest), we verified the effectiveness of these methods in ETF fund selection. The backtest results show that combining different methods can effectively improve the performance of ETF fund selection strategies and achieve better investment results. This paper further analyzes the investment results of each strategy, and evaluates the stability and risk control ability of the strategy through cross-validation and backtesting methods. The results show that the combination strategy performs well in terms of return, risk control and risk-adjusted returns, providing investors with a scientific and effective investment tool.

KEYWORDS

ETF; Time-series momentum strategy; Machine learning scoring model; Portfolio optimization; Strategy backtesting

1. INTRODUCTION

1.1. Research Background and Importance

With the development of financial markets in the past few years, ETF, which stands for Exchange-Traded Open Index Fund, has come into reality as a new tool and also as the favorite investment instrument of many investors. Investing in such funds, an investor can have a portfolio that is better diversified at a reduced cost, which effectively reduces the risk. However, how to choose among these many ETFs has become an important issue facing investors. Initially, fund selection is based on historical performance and fundamental analysis, but with the development of machine learning technology, the use of machine learning models for fund selection especially is rising.

Machine learning technology can discover hidden market patterns and make more accurate predictions by analyzing a large amount of historical data. For example, Liew and Mayster (2017) studied the use of machine learning algorithms such as deep neural networks, random forests, and support vector machines to predict the return direction of ETFs. The results showed that these

algorithms have good predictive capabilities in the medium and short term [1]. In addition, Karathanasopoulos et al. (2017) optimized the trading strategy of ETFs by introducing particle swarm optimization algorithm and radial basis function neural network, and achieved significant investment returns [2]. These studies show that combining momentum strategies and machine learning models can achieve better investment results in ETF selection.

1.2. Research Objectives

The main objective of this study is to optimize the ETF fund selection strategy and improve the performance of the portfolio by combining the time-series momentum strategy and the machine learning scoring model. We will construct a combination strategy that combines momentum indicators (such as simple momentum, exponential moving average, relative strength index) and machine learning models (such as logistic regression, support vector machine, random forest) and conduct a systematic backtest on it. Specifically, the goal of this study is to verify the effectiveness of momentum strategies and machine learning models, analyze the performance of these methods in ETF fund selection by selecting different momentum indicators and machine learning models, and verify their effectiveness in improving investment returns. On the basis of verifying the effectiveness of a single strategy, a combination strategy combining momentum strategy and machine learning scoring is constructed to optimize the portfolio construction process. Through historical data backtesting, the yield, risk control ability and risk-adjusted return performance of the combination strategy are evaluated to ensure the stability of the strategy in different market environments. Based on the results of empirical research, scientific and effective ETF fund selection tools and strategy recommendations are provided to investors to help them obtain better returns in actual investment.

2. LITERATURE REVIEW

2.1. Current Status of ETF Research

In recent years, ETFs (Exchange Traded Funds) have become one of the most innovative financial instruments in the global financial market. Research by Adamo et al. (2023) shows that despite the negative impact of the economic and social environment, ETFs still show positive performance, especially in terms of sustainable investment, where investors' attention has increased significantly [4]. In emerging markets, the growth of ETFs also shows obvious regional diversity and development potential. Zawadzki (2020) evaluated the performance of 18 ETFs in the Americas, Asia, and Europe and found that the tracking errors of these ETFs were large, and the degree of development of different regions and markets significantly affected their performance [5]. In addition, Converse et al. (2020) pointed out that the sensitivity of ETFs to the global financial cycle in emerging markets has significantly increased, which further demonstrates the important role of ETFs in international capital flows [6].

2.2. Research on Time-Series Momentum Strategy

Time-series momentum strategy is a common trading strategy in the financial market. Its core idea is to use the trend of asset prices to predict future price changes. Atilgan et al. (2020) studied the price discovery role of emerging market ETFs and found that ETF returns can predict the next-day returns of their underlying indexes, and this relationship is more significant during periods of high market volatility [7]. Charteris et al. (2014) explored the impact of ETF premiums and discounts on feedback trading, and the results showed that the lagged effect of premiums significantly promoted feedback trading, especially after the outbreak of the global financial crisis [8]. In addition, Liebi (2020) reviewed the impact of ETFs on financial market liquidity, price discovery, volatility, and co-volatility, emphasizing the importance and potential risks of ETFs as passive investment tools in the market [9].

2.3. Application of Machine Learning in Finance

Machine learning technology is increasingly used in the financial field, especially in portfolio optimization and stock selection strategies. Yang et al. (2019) proposed a new hybrid stock selection method that combines extreme learning machines and CI-based optimization methods. Empirical results show that this method significantly outperforms traditional methods in terms of market returns [10]. Piovezan and Andrade Junior (2022) proposed a machine learning approach based on classification and regression models to analyze the direction of ETF returns, and the results showed that these models performed well in terms of risk and return [11]. In addition, Carta et al. (2021) demonstrated the potential of machine learning in improving trading accuracy and profitability by proposing a statistical arbitrage trading strategy based on an ensemble of regression algorithms and dynamic asset selection [12].

3. THEORETICAL BASIS AND RESEARCH DESIGN

3.1. Overview of ETF Funds

An Exchange-Traded Fund (ETF) is an open-end fund traded on a stock exchange and has the following main attributes:

(1) High liquidity: An ETF is listed on the exchange and is tradeable on it, and therefore the investors will have the opportunity to sell or buy them as they would do with stocks.

(2) High transparency—ETFs publish their holdings daily, and investors know clearly what the fund has invested.

(3) Low charges: Generally if a charge exists or a difference, contrasted with conventional mutual funds, ETF's have actually management fees.

(4) Diverse investments: ETFs track various asset classes, such as stocks, bonds, commodities, and so on, providing ample investment opportunities. The ETF price changes with the underlying index or asset portfolio it replicates and maintains, through the "subscription" and "redemption" mechanism, the difference between the fund's net value and the market price.

3.2. Time-series Momentum Strategy and Machine Learning Theory

3.2.1. Time-series momentum strategy

The time-series momentum strategy is a strategy for trading based on historical price data. Its core idea is "buy high and sell low", that is, buy when the asset price is on an upward trend and sell when the price is on a downward trend. The mathematical expression of the momentum strategy is as follows:

$$r_t = \frac{P_t - P_{t-n}}{P_{t-n}} \tag{1}$$

Among them, r_t represents the momentum return in the t th period, P_t represents the price in the t h period, and P_{t-n} represents the price in the $t-n$ th period.

3.2.2. Machine Learning Theory

The application of machine learning in the financial field mainly includes supervised learning and unsupervised learning. Supervised learning is often used for classification and regression analysis. In

ETF fund selection, we can use machine learning models to score funds. The following is the mathematical expression of the logistic regression model:

$$\hat{y} = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n) \tag{2}$$

Among them, $\hat{y}$ is the predicted value, σ is the activation function of logistic regression, β is the model parameter, and x is the input variable.

In order to further improve the predictive ability of the model, we can use cross-validation technology. Its mathematical expression is as follows:

$$CV(k) = \frac{1}{k}\sum_{i=1}^{k} \text{Error}(M_i) \tag{3}$$

Where $CV(k)$ represents the average error of k-fold cross validation, M_i represents the model of the i-th fold, and Error(M) represents the model error.

3.3. Data Source

The data sources of this study mainly include Wind financial database and public ETF fund data. The main stock ETFs and cross-border ETFs in the Chinese market are selected as research objects. The data include daily closing prices, trading volumes, fund net value, etc. The total number of data samples is 200 ETF products, and the data time range is from January 1, 2019 to December 31, 2023. The variables include daily closing prices, trading volumes, and fund net value. Table 1 shows the descriptive statistics of the main variables, including mean, standard deviation, minimum and maximum values.

Table 1. Descriptive statistics of the main variables

Variable	Mean	Std Dev	Min	Max
Close Price	50.23	12.45	30.12	75.89
Volume	100.5	23.7	60.4	145.3
Net Value	49.8	11.3	31.5	73.2

4. MODEL CONSTRUCTION AND APPLICATION

4.1. Design of Time-Series Momentum Strategy Model

Time-series momentum strategy is a trading strategy based on historical price data, and its core idea is "buy high and sell low". In order to design an effective time-series momentum strategy, we calculate the momentum indicators of different ETFs and make trading decisions based on these indicators. The following are several commonly used momentum indicators and their calculation formulas:

First, the simple momentum calculation formula:

$$MOM_t = P_t - P_{t-n} \tag{4}$$

Among them, MOM_t represents the momentum value of the t period, P_t represents the price of the t period, and P_{t-n} represents the price of the $t-n$ period.

Then there is the exponential moving average (EMA), which reflects recent price changes better than the simple moving average:

$$EMA_t = P_t \cdot \left(\frac{2}{n+1}\right) + EMA_{t-1} \cdot \left(1 - \frac{2}{n+1}\right) \tag{5}$$

Among them, EMA_t represents the exponential moving average of the t h period, n is the time window, and P_t is the price of the t h period.

Finally, the relative strength index (RSI) is used to measure the speed and magnitude of price changes:

$$RSI_t = 100 - \frac{100}{1 + RS_t} \tag{6}$$

$$RS_t = \frac{\sum_{i=0}^{n-1} U_i}{\sum_{i=0}^{n-1} D_i} \tag{7}$$

Among them, RSI_t represents the relative strength index of the t period, RS_t represents the relative strength of the t period, U_i and D_i represent the increase and decrease respectively.

4.2. Machine Learning Fund Scoring Model

In order to score ETF funds, we use machine learning models to train and predict the fund's historical data. In addition to the logistic regression equation mentioned in Chapter 3, we also use support vector machine (SVM) for fund scoring, and its decision function is as follows:

$$f(x) = \text{sign}(w \cdot x + b) \tag{8}$$

Among them, f(x)f(x)f(x) is the classification function, www is the weight vector, xxx is the input vector, and bbb is the bias.

In order to improve the prediction performance of the model, we use the random forest model, whose formula is as follows:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{9}$$

Among them, $\hat{y}$ is the final prediction value, N is the number of trees, and $f_i(x)$ is the prediction value of the i th tree.

4.3. Model Validation and Evaluation

In the process of model validation and evaluation, we use cross-validation and backtesting methods. The following is the formula for cross-validation:

$$CV(k) = \frac{1}{k} \sum_{i=1}^{k} \text{Error}(M_i) \tag{10}$$

Among them, $CV(k)$ represents the average error of k -fold cross validation, M_i represents the model of the i -th fold, and Error(M) represents the model error.

To evaluate the performance of the portfolio, we calculated the Sharpe ratio, which is as follows:

$$SR = \frac{R_p - R_f}{\sigma_p} \tag{11}$$

Among them, SR represents the Sharpe ratio, R_p represents the expected return of the portfolio, R_f represents the risk-free rate of return, and σ_p represents the standard deviation of the portfolio.

5. PORTFOLIO STRATEGY OPTIMIZATION AND EMPIRICAL ANALYSIS

5.1. Portfolio Strategy Construction

When constructing the portfolio strategy, we combined the time series momentum strategy and the machine learning scoring model to optimize the ETF portfolio. The specific steps are as follows:

(1) Select momentum indicators: Calculate the momentum indicators of each ETF, such as simple momentum, exponential moving average (EMA), and relative strength index (RSI).

(2) Machine learning scoring: Use logistic regression, support vector machine (SVM), and random forest models to score ETFs and predict future performance based on historical data.

(3) Portfolio optimization: According to the momentum indicators and machine learning scoring results, select the top N ETFs with the highest scores to construct the portfolio.

5.2. Backtesting and Comparative Analysis

In order to verify the effectiveness of the portfolio strategy, we conducted a historical data backtest and compared it with the benchmark index (such as the CSI 300 Index). The backtest period is from January 1, 2019 to December 31, 2023, and monthly rebalancing is adopted. Table 2 shows the annualized return, maximum drawdown, and Sharpe ratio of different strategies during the backtest period.

Table 2. Annualized returns, maximum drawdowns, and Sharpe ratios of different strategies during the backtest period.

Strategy	Annual Return	Max Drawdown	Sharpe Ratio
Momentum Only	15.2%	10.5%	1.2
Machine Learning	17.8%	9.8%	1.4
Combined	19.3%	8.2%	1.6
CSI 300 Index	12.5%	11.2%	1.0

Figure 1 shows the cumulative return curves of different strategies. It can be seen from the figure that the combined strategy performed well during the entire backtesting period, and the cumulative return was significantly higher than the return of using momentum strategy or machine learning model alone.

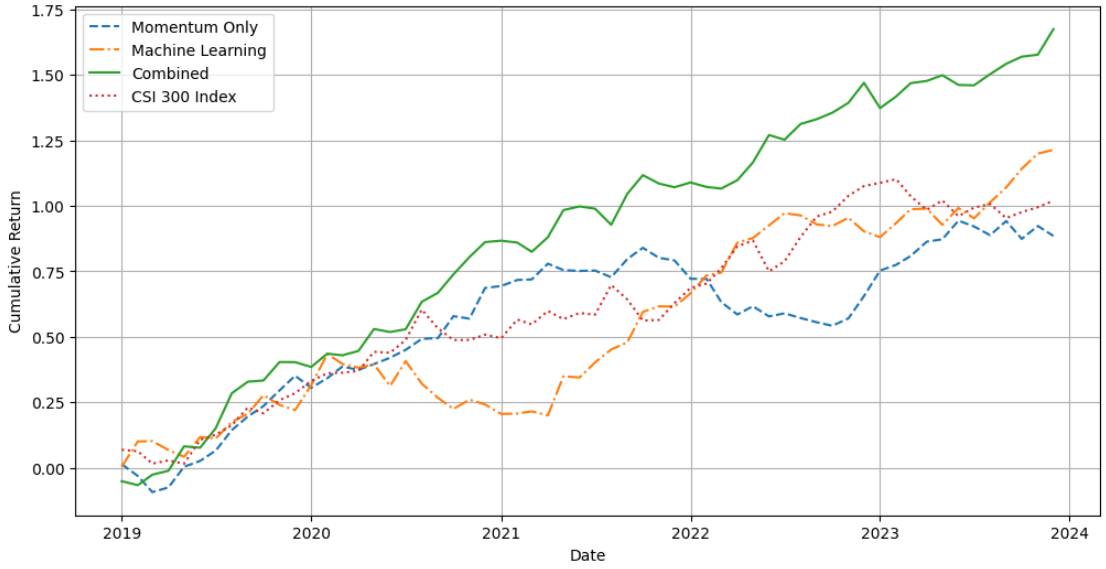

Figure 1. Cumulative Returns of Different Strategies

Figure 2 shows the monthly return fluctuations of different strategies during the backtest period. It can be seen that the volatility of the combined strategy is low and the risk control is better.

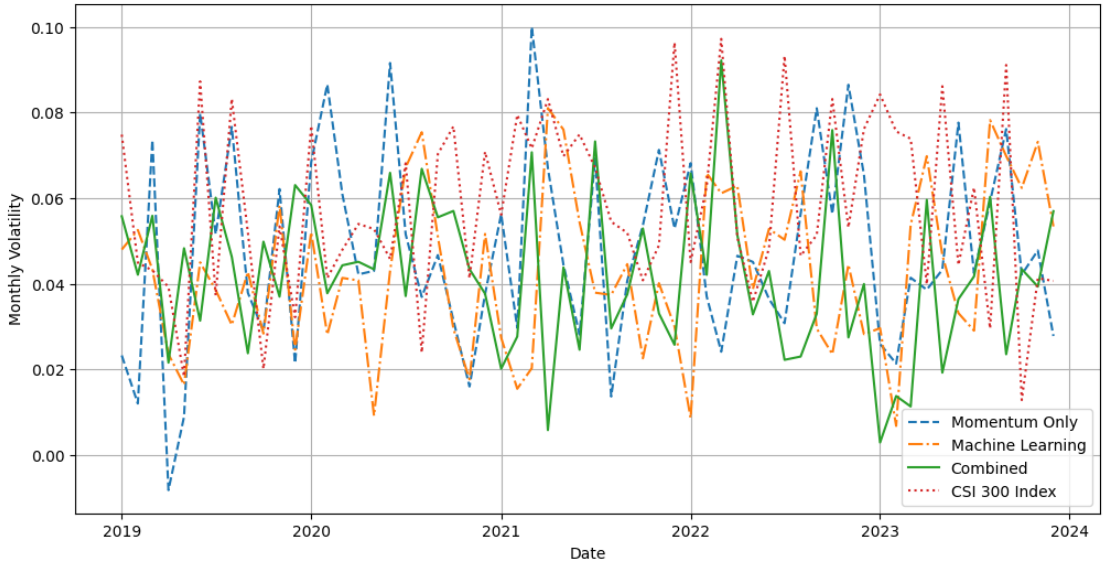

Figure 2. Monthly Volatility of Different Strategies

5.3. Investment Effect Evaluation

By evaluating the backtest results of the combination strategy, we can draw the following conclusions:

(1) Annualized rate of return: The annualized rate of return of the combination strategy is 19.3%, which is significantly higher than the momentum strategy (15.2%) and machine learning model (17.8%), as well as the benchmark index (12.5%).

(2) Maximum drawdown: The maximum drawdown of the combination strategy is 8.2%, which is lower than the momentum strategy (10.5%), machine learning model (9.8%) and benchmark index (11.2%), showing good risk control ability.

(3) Sharpe ratio: The Sharpe ratio of the combination strategy is 1.6, which is significantly higher than other strategies and benchmark indexes, indicating that its risk-adjusted return is higher.

6. CONCLUSION

Through this study, we systematically analyzed and empirically studied the ETF fund selection strategy based on machine learning scoring model and time-series momentum strategy. First, we constructed a combination strategy that combines time-series momentum indicator and machine learning scoring. The backtest results show that the combination strategy performs well in terms of annualized return, maximum drawdown and Sharpe ratio, significantly better than the momentum strategy or machine learning model and benchmark index alone. This shows that by combining different methods, the performance of ETF fund selection strategy can be effectively improved to achieve better investment results.

Secondly, this study verifies the effectiveness of time-series momentum strategy and machine learning model in ETF fund selection. Momentum strategy captures market trends and momentum by analyzing historical price data, thereby guiding buying or selling decisions. The machine learning model can score and predict the future performance of the fund by learning a large amount of historical data. In practical applications, these two methods can complement each other's strengths and weaknesses and improve the stability and return level of the overall strategy.

REFERENCES

[1] Liew, J. K., & Mayster, B. (2017). Forecasting ETFs with Machine Learning Algorithms. Journal of Alternative Investments, 20(3), 58-78.

[2] Karathanasopoulos, A. S., Mitra, S., Skindilias, K., & Lo, C. (2017). Modelling and Trading the English and German Stock Markets with Novelty Optimization Techniques. Journal of Forecasting, 36, 974-988.

[3] Luo, R., Ou, J., Chen, W., Wu, Y., Yan, J., & Liu, S. (2021). A Stock Selection Method for ETF-Abstracted-Feature Based Trading Strategy. 2021 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), 1-6.

[4] Adamo, R., Federico, D., & Notte, A. (2023). ETFs in European Emerging Markets: Performance, Risk and Sustainability. American Journal of Economics and Business Administration.

[5] Zawadzki, K. (2020). The performance of ETFs on developed and emerging markets with consideration of regional diversity. Quantitative Finance and Economics, 4, 515-525.

[6] Converse, N., Levy-Yeyati, E., & Williams, T. (2020). How ETFs Amplify the Global Financial Cycle in Emerging Markets. Social Science Research Network.

[7] Atilgan, Y., Demirtas, K., Gunaydin, A. D., & Oztekin, M. (2020). Price discovery in emerging market ETFs. Applied Economics, 54, 5476-5496.

[8] Charteris, A., Chau, F., Gavriilidis, K., & Kallinterakis, V. (2014). Premiums, discounts and feedback trading: Evidence from emerging markets' ETFs. International Review of Financial Analysis, 35, 80-89.

[9] Liebi, L. J. (2020). The effect of ETFs on financial markets: a literature review. Financial Markets and Portfolio Management, 34, 165-178.

[10] Yang, F., Chen, Z., Li, J., & Tang, L. (2019). A novel hybrid stock selection method with stock prediction. Applied Soft Computing, 80, 820-831.

[11] Piovezan, R. P. B., & Andrade Junior, P. P. (2022). Machine learning method for return direction forecasting of Exchange Traded Funds using classification and regression models. ArXiv.

[12] Carta, S., Consoli, S., Podda, A. S., Recupero, D., & Stanciu, M. M. (2021). Ensembling and Dynamic Asset Selection for Risk-Controlled Statistical Arbitrage. IEEE Access, 9, 29942-29959.