An Effective Composite Investment Quantitative Strategy with Machine Learning Method

Jun Yu *

Department of Kyoto University of Foreign Studies, Kyoto, Japan

* Corresponding Author Email: 22gs5602@@kufs.ac.jp

Abstract. In recent years, significant breakthroughs in AI and machine learning have garnered international attention. One such advancement is the development of key algorithms, including "machine self-learning and self-coding algorithms." These algorithms enable machines to continually learn from data and adjust parameters autonomously, without human intervention. Additionally, machines engage in "reinforcement learning," where algorithms achieve set goals through constant trial and error in specific environments. This has allowed current AI systems to adapt to more complex settings and learn more intricate content. In terms of data processing, AI has also made significant strides, particularly with "Recurrent Neural Networks" enabling current AI to handle sequential data more effectively like neural networks, natural language texts, and speech. Consequently, applying these technologies in the financial investment sector has become a hot topic in the industry. This paper outlines the limitations of traditional methods and proposes a new quantitative composite investment strategy. Our main approach involves designing an intelligent investment advisory model capable of self-learning, continuous self-improvement, and market adaptation. Our process includes analyzing industries, collecting extensive stock and securities data, and eliminating noise data. We then construct a model and assess its stability. Finally, accuracy is determined through an indicator. The results of our method show a high degree of similarity between the predicted stock prices and actual stock prices, indicating the effectiveness of our approach. Thus, our method offers a valuable reference for future applications.

Keywords: Machine Self-Learning, Reinforcement Learning, Recurrent Neural Networks, Financial Investment, Investment Advisory Model.

1. Introduction

Quantitative investment is a relatively novel concept. It differs significantly from qualitative or human investment, which relies on investors conducting market research and making financial investment decisions based on personal judgment and experience. In contrast, quantitative investment leverages mathematical models to realize trading ideas, placing a stronger emphasis on data.

This comparison highlights the limitations of traditional methods, which solely rely on personal experience and subjective judgment. Experience usually requires long-term accumulation, which may not be friendly to novice investors. Additionally, subjective judgments are prone to various influences, such as the investor's mindset and pressure.

The current research and methodological approach in this theme involve machine learning, a technology that allows computer systems to self-learn and improve using data. This differs fundamentally from previous programming models, which automatically perform operations when parameters reach certain values. Machine learning utilizes algorithms and data for identification, self-learning, self-analysis, and ultimately, decision-making. The main research methods and approaches in academia and industry include supervised learning, semi-supervised learning, unsupervised learning, reinforcement learning, and deep learning.

Our method and approach in this paper employ the most direct methodology by establishing a framework that combines data with machine learning. This framework is then applied to different experimental scenarios, leading to conclusions and a summary of results. For instance, we evaluate the effectiveness of our approach in scenarios involving risky securities portfolios and risk-free securities portfolios.

2. Related work

Qlib is an AI-oriented Quantitative Investment Platform [1], which is designed combination of traditional quantitative investment methods still face challenges in terms of flexibility and data-driven aspects of AI technology. To address this issue, it has developed Qlib, specifically designed for quantitative investment scenarios, supporting high-performance data processing and machine learning integration. Optimization of investment strategies through machine learning proposes a quantitative investment model that integrates machine learning with economic value-added techniques for optimizing stock selection and algorithmic trading [2]. Through empirical research, the authors demonstrate that the model can achieve returns exceeding the market average in the US stock market, proving its effectiveness and practicality. Fundamental Quantitative investment research based on Machine learning, which introduces machine learning into the field of fundamental quantitative investment, employing eight different machine learning algorithms to construct stock return prediction models and finds that linear machine learning algorithms generally outperform non-linear ones [3].

Quantitative investment decisions based on machine learning and investor attention analysis explores how to utilize machine learning and investor attention data to construct quantitative investment decision models [4]. By combining signal decomposition techniques and neural networks, the study proposes a new quantitative trading strategy aimed at enhancing prediction accuracy, controlling risk, and meeting the risk preferences of various investors [5]. A review of machine learning experiments in equity investment decision-making: why most published research findings do not live up to their promise in real life, and this paper discusses why despite numerous academic studies claiming that machine learning can make accurate predictions and profitable strategies in financial market [6], AI-driven investments lack high-profile success stories in the real world [7]. It offers suggestions to make future experimental results clearer and more useful to the investment industry through a retrospective and critical assessment of 27 academic experiments over the past two decades and discusses the feasibility and profitability of current machine learning algorithms in the stock market [8].

3. Methodology Design

The literature provided in this paper primarily revolves around recent developments in the field of quantitative investment. Each piece presents different methods and strategies but also exhibits certain limitations. For instance, the first point concerns change in structural markets. Financial markets constantly evolve over time, rendering even complex models potentially obsolete. Machine learning models need to adapt effectively to market changes, presenting a significant challenge for designers. The second point is the lack of interpretability in models. Many advanced machine learning models, such as deep learning, are 'black boxes,' making it difficult to decipher their research processes. However, clear understanding of investment logic is crucial in the field of financial investment. The third point relates to the application in the real world, as mentioned in the last article. Many theoretical research claims effectiveness in real markets but fail to achieve anticipated success. My model design aims to avoid these three flaws, ensuring adaptability to the market and using interpretable research models. Finally, an indicator will be established to test its accuracy.

3.1. Architecture Design

The initial work involves randomly selecting a few sets of data, removing noise, and then inputting them into a machine learning-based quantitative model, while incorporating different composite investment scenarios. The first step involves data selection, where we choose representative market types, such as stocks (randomly, but could be any type), selecting 50 in total. Then, the time range is determined to ensure that the data covers a sufficient period for the model to learn behavior under different market conditions. Considering the chosen data type is stocks, which are more suited for medium to long-term returns as opposed to highly volatile types like cryptocurrencies or futures, the

data selection is based on daily units, providing daily prices and various information over five years. This approach focuses on medium to long-term trends and models, ignoring short-term fluctuations. Importantly, data is chosen from relevant industries, such as the real estate sector. Selecting industry-specific data is vital because market behavior and influencing factors can vary significantly between different industries. This choice is important for several reasons: 1. Industry-specific factors: Each industry has unique influencing factors and cycles. For example, the technology sector may be affected by new inventions and regulatory changes, while the consumer goods sector may be more influenced by economic conditions and seasonal factors. 2. Reducing noise: Selecting data related to a specific industry can help reduce noise from fluctuations in other irrelevant industries, thereby improving the model's predictive accuracy and stability. 3. Better understanding of market dynamics: Focusing on a specific industry allows for a deeper understanding of that industry's specific trends and dynamics, enabling the development of more effective investment strategies. 4. Data manageability: Concentrating on a specific industry reduces the volume of data that needs to be analyzed, making data processing and model training more efficient and manageable.

The following are 10 real estate industry-related stocks selected as per requirement:

1. Prologis, Inc. (PLD)
2. American Tower Corporation (AMT)
3. Equinix, Inc. (EQIX)
4. Public Storage (PSA)
5. Crown Castle Inc. (CCI)
6. Welltower Inc. (WELL)
7. Simon Property Group, Inc. (SPG)
8. Realty Income Corporation (O)
9. Digital Realty Trust, Inc. (DLR)
10. Extra Space Storage Inc. (EXR)

Data from the past year for Prologis, Inc. (PLD) was selected and inputted into the formula. The following shows data for five days, but data for the entire year was collected and inputted. The data table includes the stock of date, opening price, highest price, lowest price, closing price, adjust closing price, Stock trading volume. (Table1)

Table 1: The PLD stock price sample

	Open	High	Low	Close	Adj Close	Volume
2023-01-09	117.500000	118.589996	116.059998	116.059998	112.877533	2504400
2023-01-10	115.599998	116.160004	114.300003	116.050003	112.867805	3097800
2023-01-11	117.349998	121.010002	117.300003	120.949997	117.633446	3084500
2023-01-12	121.279999	122.680000	120.239998	122.120003	118.771370	2608500
2023-01-13	120.400002	122.510002	120.160004	121.900002	118.557411	2283300

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta n x_n + \varepsilon \quad (1)$$

Where y is the dependent variable (predicted value). x_1, x_2, ..., x_n are the independent variables (predictors). β_0 is the intercept. β_1, β_2, ..., β_n are the coefficients of the respective predictors. ε is the error term, representing the random fluctuations not explained by the model.

This type of regression is used to establish a linear relationship between one or more independent variables and a dependent variable. The results of the regression model analysis are as follows:

- R-squared: 0.990. This indicates that the model explains a significant portion of the variation in the target variable.
- Coefficients:
- Open: -0.6204, indicating a negative correlation between the opening price and closing price.
- High: 0.7420, indicating a positive correlation between the highest price and closing price.
- Low: 0.8823, indicating a positive correlation between the lowest price and closing price.

- Volume: The coefficient is very small, almost zero, suggesting that trading volume has a minimal impact on the closing price.

- P>|t|: This column shows the statistical significance of each coefficient. The P-values for the opening price, highest price, and lowest price are very low, meaning these variables are statistically significantly related to the closing price.

- Durbin-Watson: 2.053, close to 2, indicating no autocorrelation in the data.

R-squared (Coefficient of Determination):

- A value of 0.990, very close to 1, indicates that the model performs very well in explaining the variation in the closing price. This value means that the independent variables (opening price, highest price, lowest price, and trading volume) can explain 99% of the variation in the dependent variable (closing price).

Coefficients:

Open: A coefficient of -0.6204 indicates a negative correlation between the opening price and closing price. This means that if all other factors remain constant, a 1-unit increase in the opening price will lead to a 0.6204-unit decrease in the expected closing price.

High: A coefficient of 0.7420 indicates a positive correlation between the highest price and closing price. For every 1-unit increase in the highest price, the closing price will increase by 0.7420 units.

Low: A coefficient of 0.8823 also shows a positive correlation with the closing price. For every 1-unit increase in the lowest price, the closing price will increase by 0.8823 units.

Volume: The coefficient is very small, almost zero, indicating that trading volume has a negligible impact on the closing price in this model.

P-value (P>|t|): Apart from trading volume, the P-values for other variables are very low (close to 0), indicating a statistically significant relationship between these variables and the closing price. A low P-value suggests that the probability of observing such a relationship by chance, if our null hypothesis is that the variable is unrelated to the closing price, is very low.

Durbin-Watson Statistic: A value of 2.053, close to 2, indicates no autocorrelation in the residuals (the differences between the actual observed values and the predicted values). This is a positive sign, as time series data, such as stock prices, are often prone to time-related correlations.

In conclusion, the regression model shows a significant linear relationship between the opening price, highest price, and lowest price with the closing price. However, due to the complexity of the stock market, this model should be applied cautiously. The model is continuously applied to the aforementioned 50 randomly selected stocks to compare their actual closing prices with the predicted closing prices. The predicted values are found to be very close to the actual values. Below are randomly selected predictions for five days of Prologis, Inc. (PLD) stock prices. (Table2)

Table 2: The PLD stock forecast price

Date	Actual Closing Price	Actual Predicted Price
2023-01-09	116.06	117.27
2023-01-10	116.05	115.08
2023-01-11	120.95	120.24
2023-01-12	122.12	121.64
2023-01-13	121.90	121.99

Similarly, we apply the regression model to the real estate industry stock mentioned earlier, American Tower Corporation (AMT). In this model, the adjusted closing price is used as the dependent variable, while the opening price, highest price, lowest price, and trading volume are set as independent variables. I will present five sets of data here to demonstrate this.

Table 3: The AMT stock price sample

	Open	High	Low	Close	Adj Close	Volume
2024-01-05	213.699997	216.250000	213.039993	214.279999	214.279999	1884000
2024-01-08	213.419998	216.220001	212.520004	216.080002	216.080002	1703200
2024-01-09	214.350006	214.419998	210.270004	211.860001	211.860001	1578800
2024-01-10	211.610001	212.089996	208.369995	208.970001	208.970001	2088900
2024-01-11	208.289993	208.740005	206.089996	207.649994	207.649994	2271600

The linear regression model has been constructed and evaluated based on the data. The model has a coefficient of determination (R^2 value) of 0.9783 on the test dataset. This value, being close to 1, indicates that the model performs well in explaining the data in the test set.

Table 4: The AMT forecast price

Date	Actual Closing Price	Actual Predicted Price
2023-01-09	116.06	117.27
2023-01-10	116.05	115.08
2023-01-11	120.95	120.24
2023-01-12	122.12	121.64
2023-01-13	121.90	121.99

The above compares the predicted values and actual values over a random five-day period.

4. Experimental Results

Based on the regression analysis and prediction results, we can draw the following conclusions:

1. Model Effectiveness: The linear regression model shows good statistical performance, with a coefficient of determination (R-squared) of 0.990, indicating that the model can effectively explain changes in closing prices.

2. Impact of Variables: The opening price, highest price, and lowest price are significant factors affecting the closing price, particularly showing a significant positive correlation between the highest and lowest prices and the closing price.

3. Minor Impact of Trading Volume: Trading volume has a less significant impact on stock prices in this model.

4. Prediction Accuracy: The model's predicted closing prices are very close to the actual closing prices, but due to the complexity and unpredictability of the stock market, there are some differences between predicted and actual values.

5. Caution in Application: Although the model's statistical indicators show good fit and predictive capacity, caution should be maintained in practical applications. Stock prices are influenced by a variety of complex factors, including market sentiment, economic data, political events, etc., which can all affect the actual performance of stock prices.

Table 1: Data for Prologis, Inc. (PLD) from January 9 to January 13, 2023. Although data for an entire year was collected and applied, only five days are shown here.

Table 2: Comparison of actual and predicted closing prices for Prologis, Inc. (PLD) over a random five-day period.

Table 3: Stock data for American Tower Corporation (AMT) from the 5th to the 11th of 2024. Again, data for an entire year was collected and applied, but only five days are shown here.

Table 4: Comparison of actual and predicted closing prices for American Tower Corporation (AMT) over a random five-day period.

5. Conclusion

By selecting ten sets of data related to real estate and inputting them into a machine learning-based quantitative model, along with various composite investment scenarios, we assessed their yield and risk rates. The results demonstrate that both risk and return rates are within the expected values, proving the feasibility of our approach. We have constructed a framework aimed at optimizing profit and risk indicators. As technology advances and the market continues to evolve, there remains significant research space and potential in this field. We look forward to machine learning and quantitative strategies bringing more accurate and efficient investment decisions in practice.

References

[1] Qlib: An AI-oriented Quantitative Investment Platform Xiao Yang Weiqing Liu Dong Zhou. (2020). Papers 2009.11189, arXiv.org.

[2] Optimization of investment strategies through machine learning Jiaqi li Xiaoyan Wang Saleem Ahmad Xiaobing Huang Yousaf Ali Khan (2023) 2405-8440/© 2023 The Authors. Published by Elsevier Ltd.

[3] Fundamental Quantitative investment research based on Machine learning Jiao Xu. (2023). published by EDP Sciences.

[4] Quantitative investment decisions based on machine learning and investor attention analysis Jie GAO, Yunshu MAO, Zeshui XU, Qianlin LUO (2023) Technological and Economic Development of Economy. (Aug. 2023), 1-35.

[5] A review of machine learning experiments in equity investment decision-making: why most published research findings do not live up to their promise in real life Wojtek Buczynski, Fabio Cuzzolin, Barbara Sahakian. 2021;11(3):221-242. doi: 10.1007/s41060-021-00245-5. Epub 2021 Apr 5. PMID: 33842690; PMCID: PMC8019690.

[6] Social Computing Empowered Cloud Service on Quantitative Investment Li Zhang, Peng Chen, Qian Li, 2014 International Conference on Computer, Communications and Information Technology. Published by Atlantis Press

[7] The Analysis and Forecast of Bank Index and Its Constituent Stocks, Xiao Han, Xinghu Teng, Ting Dou. Published by Hans Publishers Inc. (2020)

[8] Attention-BiLSTM stock price trend prediction model based on empirical mode decomposition and investor sentiment, Shuaibin ZHAO, Xudong LIN, Xiaojian WENG Journal of Computer Applications 43 (S1), 112, 2023