Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3589198, 13 pages
https://doi.org/10.1155/2020/3589198

Research Article

Financial Trading Strategy System Based on Machine Learning

Yanjun Chen,[1] Kun Liu,[1] Yuantao Xie ,[2] and Mingyu Hu[2]

[1]*School of Information and Management, University of International Business and Economics, Beijing, China*
[2]*School of Insurance and Economics, University of International Business and Economics, Beijing, China*

Correspondence should be addressed to Yuantao Xie; xieyuantao@uibe.edu.cn

Received 6 May 2020; Accepted 7 July 2020; Published 28 July 2020

Guest Editor: Qian Zhang

The long-term and short-term volatilities of financial market, combined with the complex influence of linear and nonlinear information, make the prediction of stock price extremely difficult. This paper breaks away from the traditional research framework of increasing the number of explanatory variables to improve the explanatory ability of multifactor model and provides a new financial trading strategy system by introducing Light Gradient Boosting Machine (LightGBM) algorithm into stock price prediction and by constructing the minimum variance portfolio of mean-variance model with Conditional Value at Risk (CVaR) constraint. The new system can capture the nonlinear relationship between pricing factors without specific distributions. The system uses Exclusive Feature Bundling to solve the problem of sparse high-dimensional feature matrix in financial data, so as to improve the ability of predicting stock price, and it can also intuitively screen variables with high impact through the factor importance score. Furthermore, the risk assessment based on CVaR in the system is more sufficient and consistent than the traditional portfolio theory. The experiments on China's stock market from 2008 to 2018 show that the trading strategy system provides a strong logical basis and practical effect for China's financial market decision.

1. Introduction

With the development of stock market, the efficiency of artificial subjective investment mode is gradually reduced due to the complex and diverse investment targets. Benefiting from the advancement of data science and statistical method, the former subjective investment mode has been gradually replaced by quantitative investment strategy, which uses data and models to construct investment strategies. New investment model, selecting stocks with investment value by combining the open information in the market with statistical methods, avoids the subjective impact of human to some extent.

As the most widely used quantitative stock selection model at present, multifactor model is based on finding out factors with the highest correlation with the stock return rate, which can predict the stock return to some extent. However, in the empirical test, scholars have found that it could not bring sustained returns to investors due to the low prediction accuracy and the lack of stability of the prediction results.

At the same time, through the empirical study with financial market data, scholars found that the financial market is a dynamic system with high complexity, including long-term and short-term fluctuations and linear and nonlinear information. The formation and change of stock price involve various uncertain factors, and there are complex relationships among them. To further study and analyze financial data and for more accurate prediction, the application of machine learning algorithms in the research of financial time series has been widely concerned by scholars.

Compared with the linear model, machine learning algorithm takes the nonlinear relationship between variables into account. It does not need to be based on the assumption of independence and specific distribution and has higher flexibility and efficiency, making it excel at dealing with big data, especially the huge amount of financial data.

The innovations of this paper are as follows: Firstly, it breaks away from the traditional research framework of improving the explanatory power of multifactor model by increasing the number of explanatory variables and provides a new trading strategy system by introducing one of the

latest machine learning algorithms, LightGBM, into the field of portfolio. LightGBM does not need to consider the specific distribution form of financial data, and it can capture the nonlinear relationship between pricing factors, and the Exclusive Feature Bundling method can solve the problem of sparsity of high-dimensional characteristic matrix and improve the prediction accuracy of stock returns. Secondly, the new system can be used to generate importance score of factors. It directly shows the impact variables have on stock return, which has important practical significance for stock selection. Thirdly, for stock position allocation module, we construct minimum-variance weight method of mean-variance model with CVaR constraint and test the trading strategy system on the real data of China's A-share market. The experiment result shows that the system can bring stable excess return to investors and provides a logical basis and practical effect for China's stock market.

2. Literature Review

Since Markowitz proposed mean-variance model to quantify the risk in 1952, scholars have been trying to find ideal models for stock pricing. In 1960s, the CAPM model proposed by Sharpe, John Lintner, and Jan Mossin expresses the relationship between expected return and expected risk as a simple linear relationship [1, 2]; Ross [3] put forward the Arbitrage Pricing Theory on this basis, revealing that stock return is affected by multiple factors rather than a single factor; Fama and French [4, 5] proposed a three-factor model, which includes market value, book value ratio, and P/E ratio of listed companies as compensation for the risk factors that β cannot reflect. However, the classic models above were lacking explanation in the empirical test. At the beginning, scholars attributed that to the omission of explanatory variables, so they successively put forward factors that may lead to excess return: For example, Datar et al. [6] used the data of NYSE for empirical analysis and found that there was a significant negative correlation between stock return and stock turnover. Piotroski [7] selected nine indicators from the perspective of the company's financial indicators, including profitability, robustness, and growth. By comprehensively scoring each indicator, he established a stock pool and selected the stocks according to the scores. Novy-Marx [8] proved that there is a strong correlation between the company's profitability and stock return. Aharoni et al. [9] also found that the company's investment level and stock return are remarkably correlated. In 2014, Fama and French added Robust Minus Weak (RMW) as profit factor and Conservative Minus Aggressive (CMA) as investment factor based on the three-factor model and put forward five-factor model to further explain the stock return [10].

However, none of the models above deviated from the research framework of linear asset pricing under the background of small sample data, that is, extract excess return factors from historical data and then use these factors as independent variables to construct a linear model to evaluate the investment value of stocks.

Since the 1970s, with the increasing availability of empirical data in the financial market and the improvement of computer technology, scholars have found several abnormal phenomena in the financial market through empirical research. These abnormal phenomena are contrary to the basic assumption of CAPM: financial market data obey normal distribution, have no long memory, and satisfy the linear model, which challenges the traditional model [11]. For the stock market, Greene and Fielitz [12] performed a test and confirmed that the American stock market has the feature of long memory, which showed that even if the time interval is very long, it still had significant autocorrelation; that is, the historical events would affect the future for a long time. On the one hand, it proved the importance of historical information and the predictability of return; on the other hand, it also reflected the nature of nonlinear structure of stock market. As for the distribution of financial data, scholars pointed out that, in reality, the distribution of financial data is usually characterized by thick tail and asymmetry [13]. Therefore, the traditional use of normal distribution to fit the actual financial data has limitations. For example, in VaR calculation, due to the thick tail of financial data distribution, the calculation under the assumption of normal distribution will lead to huge errors [14]. In order to find the most reasonable distribution hypothesis, Mandelbrot [15] proposed replacing the normal distribution of financial data with the stable distribution. However, because the tail of the stable distribution is usually thicker than the actual distribution, some scholars proposed using the truncated stable distribution as the distribution of securities returns [16], but where to cut off had become another question.

In recent years, in order to analyze and predict financial data more accurately, machine learning has received wide attention from scholars. Compared with the traditional model, machine learning has a unique advantage in dealing with financial data. First of all, it can automatically identify the hidden features behind the financial data, reducing human intervention. Secondly, the traditional linear asset pricing model is based on the assumption that the financial system is linear. However, scholars' research on the nonlinear characteristics of financial time series, such as long memory and nonpairing distribution, indicates that the stock market system is actually a dynamic system with linear and nonlinear information. Machine learning models can deal with high-dimensional and collinear factors and are not limited to the probability distribution of investment income. Machine learning models do not need to calculate high-dimensional covariance matrix [17].

Mukhejee (1997) firstly proposed the application of support vector machine in nonlinear chaotic time series, which provided the basis for the application of stock series. In the empirical study, Fan and Palaniswami [18] first applied the support vector machine model to stock selection. Based on the data of Australian stock exchange, the model they constructed could identify the stocks that outperform the market and the five-year yield of the equal weight portfolio constructed was 208%, which was far higher than the benchmark return of the large market. Kim [19] used

support vector machine (SVM) and artificial neural network (ANN) to predict the market index and the results showed that SVM had more advantages than ANN in stability. There are also some literatures that focus on the differences between different models in variable selection and modeling characteristics; for example, Xie et al. [20] and Huang et al. [21] set the rise and fall of stock market as dichotomous variables and used linear model, BP neural network, and support vector machine model to predict them. It was found that SVM had better classification performance than other methods, and the combined model performed best in all prediction methods when SVM is combined with other models. Nair et al. [22] used C4.5 decision tree algorithm to extract the characteristics of stock data and then applied it to the prediction of stock trend. They found that the prediction effect of C4.5 decision tree was better than neural network and naive Bayesian model. Zhu et al. [23] applied Classification and Regression Tree (CART) algorithm and traditional linear multifactor model in North American market during the outbreak of financial crisis and found that the stock selection model based on CART algorithm had a significant effect on risk dispersion. Kumar and Thenmozhi [24] used random forest model to predict the up and down direction of Standard and Poor's and found that the result was better than that of SVM. Bogle and Potter [25] used decision tree, artificial neural network, support vector machine, and other machine learning models to predict the stock price of Jamaica stock exchange market and found that, in this market, the prediction accuracy of stock price could reach 90%.

In recent years, the first mock exam has also been made in some areas, such as the sequence dependence of financial time series data and the local association characteristics of different financial market time series data. For example, Xie and Li [26] discussed the joint pricing models and Yan [27] constructed a CNN-GRU neural network, which combines the advantages of convolutional neural network (CNN) and gating loop unit (GRU) neural network. There are also some papers that study the computing power, time consumption, and even hardware layout of various algorithms. For the discussion of machine learning related hardware, refer to Tang et al. [28, 29].

3. System Introduction

Based on machine learning algorithm, this paper constructs an optimal trading strategy system, which aims to bring stable excess return to investors. According to Figure 1, this system is divided into four modules: data preprocessing, stock pool selection, position allocation, and risk measurement. The details are as follows.

3.1. Data Preprocessing. Because of the noise and format asymmetry in financial data, preprocessing plays an important role in getting accurate prediction results. According to Figure 2, we preprocess the financial data according to the following steps:

Step 1(financial data processing): due to the differences in the format of financial statements of different companies, the data sets have sparse data spaces. The financial accounts with over 20% missing values are discarded directly. The remaining default values are filled with the average values of the previous and next three quarters. Since the income statement and cash flow statement are process quantities, representing the accumulation of quarterly values, the data of these two financial statements are differentially processed to obtain quarterly data.

Step 2 (market data processing): since the stock market data set is monthly, in order to match the data in the financial statements, it is processed on a quarterly average basis.

Step 3 (data screening): due to the differences in the format of financial statements in different industries, financial data are divided into banking, securities industry, insurance industry, and genera business. As the banking, insurance, and security industries are all subindustries under the financial industry, their businesses are complicated and are greatly affected by the macro impact, resulting in the uncertainty and volatility of their stock prices far greater compared to the general business. Therefore, prediction from their financial data alone is difficult. Moreover, after preprocessing, it is found that the data of these industries are too insufficient to make a prediction. In view of the lack of reference in the forecast results, these three industries are deleted from the data, and only the data of general business are retained.

Step 4 (data splicing): we take the company's stock code as the primary key and combine the financial data with market data of the same quarter into a wide table to prepare for feature engineering.

Step 5 (feature engineering): this paper applied machine learning models into stock return forecasting. The reason why machine learning can achieve high prediction accuracy is that it can deal with the nonlinear relationship between variables. Unlike simple linear model, the complexity of the model leads to machine learning being regard as a "black box." Due to the lack of interpretation, it is contradicted by the traditional financial industry. In order to improve the credibility of our model, not only do we use the importance of variables to analyze the important influencing factors of stock price return, but also we accord to the previous literature and construct the characteristics that have been proved to have strong significance in the previous multifactor model. Table 1 shows the calculation method and index implication.

Step 6 (missing value processing and standardization): due to the high requirements of data integrity in stock forecast, we delete the data with the missing rate more than 10%. For the remaining missing value, based on the idea of moving average, we take the same fields of the three records before and after the missing record to

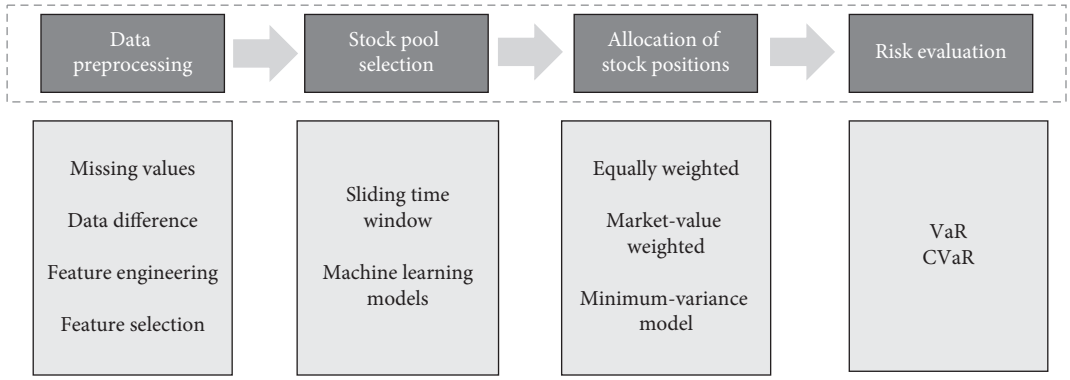

Figure 1: Framework of the trading strategy system.

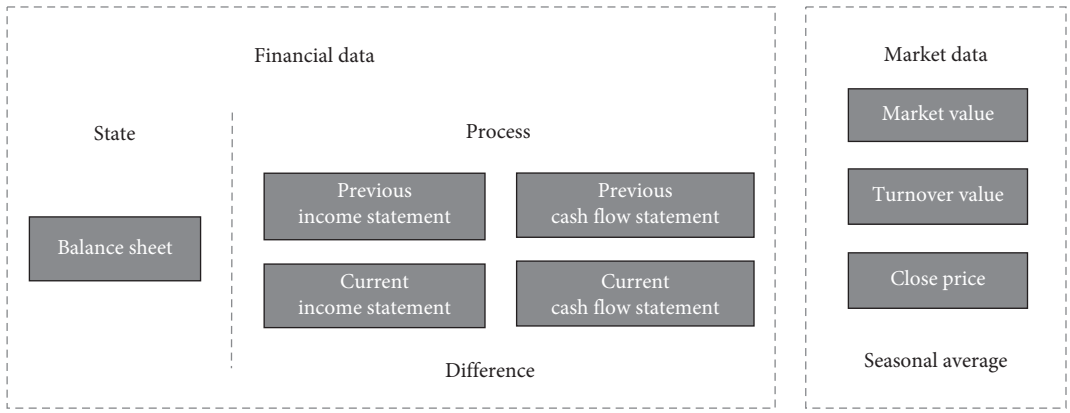

Figure 2: Data and data processing.

calculate the moving average value, thus retaining the trend information of the stock as much as possible. Finally, in order to improve the convergence speed and prediction accuracy of the model, the data are standardized.

3.2. Stock Pool Generation. This paper uses LightGBM algorithm under the sliding time window training method to predict the quarterly earnings of stocks and compares it with traditional linear model, support vector machine, artificial neural network, random forest, and other machine learning models. The evaluation criteria of models are R2 and RMSE. Based on the prediction results, a stock pool composed of a limited number of stocks is generated, and then the portfolio is constructed according to different position allocation methods.

3.2.1. Algorithm Principle: LightGBM. Compared with the generalized linear model, machine learning, as a new model, does not need to be based on the assumption that the variables are independent and obey the distribution of specific functions and has greater flexibility and efficiency. Machine learning algorithms have unique advantages in dealing with a large amount of data, such as financial market data.

Among many machine learning algorithms, LightGBM algorithm has the characteristics of high speed, high accuracy, high stability, and low memory space. It has wide application space in the financial field with large amount of data and high requirements for prediction accuracy and stability. LightGBM is essentially an enhanced gradient lifting tree that can be used for regression and classification. Compared with the previous gradient lifting tree (GBT), it has the following advantages: LightGBM uses the leaf-wise (best-first) strategy to grow a tree: find the leaf with the largest gain each time and split it while other nodes without the maximum gain do not continue to split. LightGBM does not need artificial trim, so the result is relatively objective and stable. At the same time, LightGBM adopts histogram algorithm and accumulates statistics in the histogram according to the value after discretization as the index to calculate the information gain. LightGBM also adopts two new calculation methods: Gradient-Based One-Side Sampling and Exclusive Feature Bundling, greatly improving the accuracy and efficiency of calculation.

(1) Given the supervised training set $X = \{(x_i, y_i)\}_{i=1}^{n}$, the goal of LightGBM is to minimize the objective function, which is

Table 1: Features and implication.

	Index	Calculation method	Implication
Profitability	ROE	Net profit/owner's equity	Return on equity
	ROA	Net profit/total assets	Return on assets
	ROS	Total profit/operating income	Return on sales
Fluidity	D/E	Liabilities/owner's equity	Debt to equity ratio
	Cash ratio	(Current assets – inventory)/current liabilities	Ratio of quick assets to current liabilities
	Current ratio	Current assets/current liabilities	Ratio of current assets to current liabilities
Operating efficiency	Equity turnover	Sales income/average shareholders' equity	The efficiency of the company in using the owner's assets
	Asset turnover	Sales income/average total assets Average balance of total assets = opening balance + closing balance/2	An important financial ratio to measure the efficiency of enterprise asset management
Valuation index	B/M	Outstanding stock ∗ closing price/shareholder equity	Book-to-market ratio High B/M, considered to be undervalued by the market, resulting in high yield
	P/E	Market price per common share/earnings per common share per year	Price-to-earnings ratio The lower the price earnings ratio is, the lower the profitability of the market price relative to the stock is
	P/B	Share price/net asset per share	Price-to-book ratio The higher the investment value of stocks with low market to net ratio is, the lower the investment value is
	Market value of listed company	Market price per share ∗ total number of shares issued	The total value of shares issued by a listed company at market price
	Turnover	Trading volume/total issued shares	The higher the turnover of a stock is, the more active the stock is in trading
β Coefficient	Systematic risk coefficient	Regression of the historical rate of return of a single stock asset to the index rate of return of the same period	β describes the systemic risk of a fully diversified portfolio

$$J(\phi) = \sum_i l(\hat{y}_i, y_i), \tag{1}$$

where i denotes the i – th sample and $l(\hat{y}_i, y_i)$ denotes the prediction error of the i – th sample.

(2) For optimizing the objective function, LightGBM uses gradient boosting method to train rather than using bagging method to directly optimize the whole objective function. The gradient boosting training method optimizes the objective function step by step. Firstly optimize the first tree and then optimize the second one and so on until the K tree is completed.

(3) When generating a new optimal tree, LightGBM uses the leaf-wise algorithm with depth limitation to grow vertically. Therefore, the leaf-wise algorithm is more accurate compared with the level-wise algorithm when they have the same number of splitting times.

When the T – th tree is generated, every time a newly generated split node is added, and then the objective function can be obtained as follows:

$$J = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right], \tag{2}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$, $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$; I_L and I_R are the sample sets of the left and right branches, respectively.

3.2.2. A Method for High-Dimensional Feature Matrix Sparsity: Exclusive Feature Bundling. In order to get the data of each quarter, we carried out differential processing on the financial statements of process volume during data preprocessing. However, due to the lack of mandatory regulations on the quarterly reports of listed companies in China, some companies have the problem of time lag in the quarterly reports, and there are a large number of zero values in the data of nonupdated adjacent quarterly reports after differential processing. At the same time, due to business differences between companies, nonshared business accounts in financial statements sometimes are also zero.

In conclusion, the financial data of listed companies with large subjects and different businesses eventually form a high-dimensional feature matrix, and a large number of zero values in the matrix lead to the problem of data sparsity. Traditional statistical methods are often unable to extract enough effective information when dealing with high-dimensional and sparse data, which makes the prediction results inaccurate or even wrong. In order to solve this problem and improve the prediction accuracy, we introduce Exclusive Feature Bundling (EFF) method of LightGBM into the stock pool selection module.

Exclusive Feature Bundling aims to solve the problem of data sparsity by merging mutually exclusive features to reduce the dimension of feature matrix [30]. Since LightGBM stores the features divided into discrete values by constructing histogram instead of storing continuous values directly, we can combine mutually exclusive features by assigning them to different intervals of the same histogram.

For example, X_a and X_b are two mutually exclusive features, where $X_a \in [0, a)$ and $X_b \in [0, b)$.

The new bundling feature X_c can be obtained by adding the value range of X_a as offset and value range of X_b, where $X_c \in [0, a + b)$:

$$\begin{cases} X_a = X_c, X_b = 0 & \text{when } 0 \le X_c < a, \\ X_a = 0, X_b = X_c - a & \text{when } a \le X_c < a + b. \end{cases} \tag{3}$$

3.2.3. Training Method: Sliding Time Window. Since the data of the stock market is a time series, the historical information of the company has a great influence on the future stock price. Considering that the sequence has a great influence on the values of the sequence nodes, the prediction model in stock pool selection should consider the key element, "time," in the prediction process, instead of treating the stock prices of all times as the same data and randomly selecting the training set and the test set. Therefore this system does not use cross-validation but uses sliding time window to randomly simulate the prediction process.

In the experiment, suppose that the size of sliding time window was N quarters. In a unit time window, the first N-1 quarter is the training set, and the last quarter is the test set. The size of the time window should take into account the characteristics of the data set. If it is too short, the natural time period of the test set may be outside the training set, and the time information brought by the time window will be greatly reduced; if it is too long, some unnecessary noise may be introduced.

In essence, the model in each time window is a new model and they are all independent of each other. According to Figure 3, the stock price of each stock in the next quarter in each time window is predicted according to all the information in the latest quarter.

3.2.4. Stock Selection. The goal of this paper is to apply a new machine learning model, LightGBM, to the prediction of stock return and to construct a low-risk and high-yield portfolio compared with the stock price prediction models used in previous studies. In order to highlight the risk of different portfolio construction methods, the first N stocks are selected as the stock pool from the stock list sorted by the yield, only the long purchase rule is allowed in this part to ensure that the yield equals the required value. Then adjust the position of each stock according to different weights to find the optimal portfolio.

3.3. Stock Position Allocation Method. This paper uses three methods of stock position allocation: (1) equal-weight method, (2) market-value weighting method, and (3) minimum-variance method of mean-variance model with CVaR constraint. Combined with the sliding time window training method to predict the quarterly earnings of the stock, we use R-squared and RMSE of the model as the evaluation criteria, while in the traditional linear model, support vector machine, artificial neural network, random forest, and other machine learning models are used. Based on the prediction results, a stock pool composed of a limited number of stocks is generated, and then the portfolio is constructed according to different position allocation methods.

3.3.1. Method 1: Equal-Weight Method. Each stock is assigned the same weight. If there are n stocks in the stock pool, the weight of each stock is $w_i = 1/n$.

3.3.2. Method 2: Market-Value Weighting Method. The ratio of market value of stock i to the market value of all n stocks in the stock pool is taken as the weight of this stock, and the calculation formula is as follows:

$$w_i = \frac{\text{Market value}_i}{\sum_{i=1}^{n} \text{Market value}_i}, \tag{4}$$

where Market value_i is the market value of the company at the closing of the previous period and it is calculated by multiplying the market price of each share by the total number of shares issued.

3.3.3. Method 3: Minimum-Variance Weight Method of the Mean-Variance Model with CVaR Constraint. In 1952, Markowitz published the beginning article of modern portfolio theory "portfolio selection" in the financial magazine, which studies how to allocate risk assets effectively. Markowitz believed that investors only consider two factors of expected return and standard deviation of forecast when making portfolio decision, so portfolio decision was mainly based on the following two points: (1) when the investment return is the same, investors want to minimize the risk; (2) when the risk is the same, investors want to maximize the income. According to the principle of mean-variance efficiency, the optimal portfolio can be expressed by mathematical programming in the process of investing in assets.

Assuming that the return of risk assets obeys normal distribution, consider CVaR constraint in Markowitz mean-variance model, and then the portfolio optimization model based on CVaR constraint is

$$\begin{cases} \min & \sigma_p^2 = \min X^T \sum X, \\ \text{s.t.} & \text{CVaR}_\beta = C_2(\beta)\sigma_p - E\left(r_p\right) \le L, \\ & E\left(r_p\right) = X^T R, \\ & X^T I = 1, \quad I = (1, 1, \ldots, 1)^T, \\ & x_i \in [0, 1], \end{cases} \tag{5}$$

where $C_2(\beta) = \phi(\Phi^{-1}(\beta))$, $R = (R_1, R_2, \ldots, R_n)$, $R_i = E(r_i)$is the expected return of i – th stock, and $X = (x_1, x_2, \ldots, x_n)^T$ is the weight of each portfolio. We have the following:

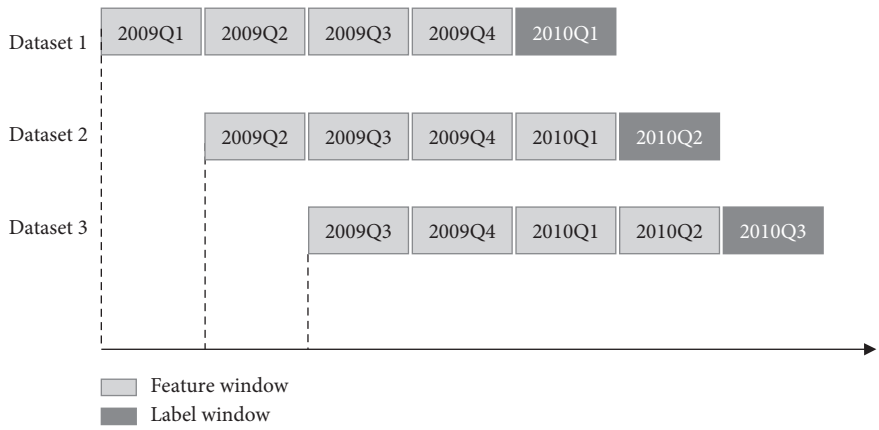

Figure 3: Sliding time window model.

(1) According to the constraint equation, $\mathrm{CVaR}_\beta = C_2(\beta)\sigma_P - E(r_P) \le L$, and we obtain

$$\mathrm{CVaR}_\beta = C_2(\beta)\sigma_P - E(r_P) = L. \tag{6}$$

(2) After transformation, we obtain

$$\sigma_P^2 = \frac{1}{(C_2(\beta))^2}(E(r_P) + L)^2 = \frac{1}{(C_2(\beta))^2} \cdot \left(E^2(r_P) + 2LE(r_P) + L^2\right). \tag{7}$$

(3) Define $R_P = E(r_P) = X^T R$ and $\sigma_P^2 = X^T \sum X$; equation (7) can be

$$X^T \sum X = \frac{1}{(C_2(\beta))^2}\left(\left(X^T R\right)^2 + 2L\left(X^T R\right) + L^2\right). \tag{8}$$

(4) Since the rank of the system of linear equation (8) is $n-2$, the number of its basic solutions is 1. That is to say, $x_2, x_3, \ldots, x_{n-1}$, can be linearly expressed by x_1 (obtained by elimination method). Since $\sum_{i=1}^{n} x_i = 1$, x_n can also obviously be expressed by x_1, by substituting $x_2, x_3, \ldots, x_{n-1}$ into equation (8), we can get a quadratic equation of x_1, and then we can obtain $x_2, x_3, \ldots, x_{n-1}$ by calculating x_1.

(5) Therefore, if x_1 has no solution, the constraint line does not intersect the mean-variance effective front, which means CVaR does not play a constraint role; if x_1 has two multiple roots, there is only one intersection A between the constraint line and the mean-variance effective front, and the weight of A is x_1; if there are two different roots, the constraint line and the mean-variance effective front have two intersections, A and B. The expected return and variance of portfolio at A and B are R_A, R_B, σ_A^2, and σ_B^2.

3.4. Risk Evaluation. This paper mainly studies how to reduce the risk of portfolio. At present, the most common measurement methods of portfolio risk are sensitivity method, volatility method, VaR method, and CVaR method. They are introduced, respectively, as follows.

3.4.1. Method 1: Sensitivity Method. Sensitivity method is a method to measure the risk of financial assets by using the sensitivity of the value of financial assets to market factors. Sensitivity refers to the percentage change in the value of financial assets when market factors change by a percentage unit. The greater the sensitivity of financial assets is, the greater the impact of market factors is and the greater the risk is. The sensitivity of different types of financial assets has different names and forms, such as the duration and convexity of bonds, beta of stocks, Delta, Gamma, and Vega of derivative assets. The problems of sensitivity method lie in the following: (1) The sensitivity is only valid when the range of market factors changes is very small. (2) A certain sensitivity concept is only applicable to a certain class of specific assets or a certain class of specific market factors, which makes it difficult to compare the risks among different kinds of assets. (3) The sensitivity is only a relative proportion concept, which cannot determine the risk loss of a certain portfolio body value.

3.4.2. Method 2: Volatility Method. Volatility method is a statistical method, which is usually described by standard deviation or covariance. Variance represents the volatility of the actual rate of return deviating from the average rate of return. The greater the volatility, the greater the uncertainty of the actual rate of return, regardless of whether the actual rate of return is higher than the average rate of return or lower than the actual rate of return. One of the main defects of variance measurement of investment risk is that variance represents positive and negative deviation, generally speaking, investors do not want that the actual return is less than the expected return, but they do not refuse when the actual return is higher than the expected return. Therefore, Markowitz put forward the semi variance method in 1959; that is, the part of the actual income higher than the expected income is not included in the risk, and only the loss is included. There are some problems in both variance method and semi variance method: (1) The method is based on the

assumption that variance exists, but whether the variance of return rate exists is still questionable. (2) Variance implies the assumption of normality, so only the linear correlation structure between risks can be analyzed. In reality, the risk dependence structure may be a nonlinear complex structure. (3) The square deviation does not specifically indicate how much the loss of the portfolio is. (4) This method is not suitable for comparing the risk of assets with different expected return.

3.4.3. Method 3: VaR Method. VaR is Value at Risk [31]. VaR is the maximum possible loss expected in the holding period of an investment within a certain confidence level. Its mathematical expression is as follows:

$$\text{Prob}\left(\Delta p \leq \text{VaR}\right) = \alpha, \tag{9}$$

where Δp is the loss amount of the portfolio during the holding period Δt, which is the Value at Risk under the given fixed credit level α, that is, the upper limit of possible loss. The meaning of the expression of the above formula is that the risk loss of the portfolio is not less than the VaR at the level of probability. In the research, the above expression is regarded as a function of VaR on α, and the probability distribution function of the portfolio return is expressed with $F(\alpha)$, which means

$$\text{VaR}_{\alpha} = F^{-1}(\alpha). \tag{10}$$

The advantages of VaR method are: the following (1) The measurement of risk is simple and clear, the risk measurement standard is unified, and it is easy for managers and investors to understand and grasp. (2) VaR can also be used to compare the risks of different types of financial assets, but its disadvantage lies in the inconsistency.

3.4.4. Method 4: CVaR Method. CVaR is a new risk measurement method proposed by Roekafellor and Uryasev (1999), also known as Conditional Value at Risk method, which means that, under a certain confidence level, the loss of portfolio exceeds the mean value of a given VaR, reflecting the average level of excess loss. Its mathematical expression is as follows:

$$\text{CVaR}_{\alpha} = E\left(-X \mid -X \geq \text{VaR}_{\alpha}\right), \tag{11}$$

where $-X$ represents the random loss of the portfolio and VaR_{α} is the Value at Risk under the confidence level.

The advantages of CVaR are as follows: (1) It solves the problem of inconsistency measurement, satisfies the additivity of risk, and improves the defect of VaR. (2) It does not need to realize the form of assumed distribution, and, in any case, its calculation can be realized by simulation. (3) It fully measures tail loss and calculates the average value of tail loss, which considers all tail information larger than VaR rather than based on a single quantile to calculate.

Therefore, this paper mainly measures the risk of stock portfolio based on the VaR and CVaR. Monte Carlo simulation method is used in the specific calculation method, which is the most effective method to calculate VaR and CVaR as it can solve the nonlinear relationship of various targets well without making assumptions on the distribution of portfolio income.

4. Empirical Research

4.1. Experimental Results and Analysis. The data used in this paper are the market data of 3676 A-share listed companies in China from 2008 to 2018, as well as the financial data disclosed by the company on a quarterly basis (including the company's balance sheet, profit statement, and cash flow statement).

After data preprocessing and feature engineering, the processed data from the fourth quarter of 2008 to the first quarter of 2018 are selected as the final data set.

Based on the total split times of features, the top 20 variables in the variable importance score obtained by our trading strategy system are as follows.

According to Figure 4, the factor affecting the next stock price most significantly is the closing price of the current stock (CLOSE_PRICE). According to Charles Dow's technical analysis theory, the historical price of the stock contains a lot of information. The price will evolve in the way of trend, and the history will always be repeated because of human psychology and market behavior. By studying the historical price of the stock, investors can find out the current market and the trend, so as to better detect the stock selection target and the opportunity to build a position. Therefore, the closing price of the current stock is highly related to the next stock price, which is the most important variable to predict the stock price. The second in the list is accounts payment (AP). In order to expand sales and increase market share, enterprises often buy materials and accept services first and then pay service fees and commissions. The time difference between sales and payment also reflects the risk that enterprises may be short of funds and cannot pay in time. Therefore, AP is an important reference index for investment. The third index, the growth rate of the previous period's stock price (Price_rate), has a great contribution to the prediction of the stock price. It has a positive correlation with the growth of the current period's stock price. Generally, the higher the growth rate of the previous period's stock price is, the higher the growth of the stock is and the greater the possibility of continuous growth in the current period is.

Other financial indicators, such as the balance of cash and cash equivalents at the beginning of balance (N_CE_BEG_BAL), turnover (TURNOVER_VALUE), cash paid for goods purchased and services received (C_PAID_G_S), surplus reserve (SURPLUS_RESER), and return on assets (ROA), also have impact on the stock price. The balance of cash and cash equivalents at the beginning of the balance refers to the amount of cash and cash equivalents carried over from the previous year to the current year for current turnover. It reflects the cash stock of the enterprise. Turnover rate measures the frequency of stock turnover in the market within a certain period of time, reflecting the activity of market trading investment. Its calculation formula is as follows: turnover rate = trading volume/total number of shares issued. The higher the turnover of a stock

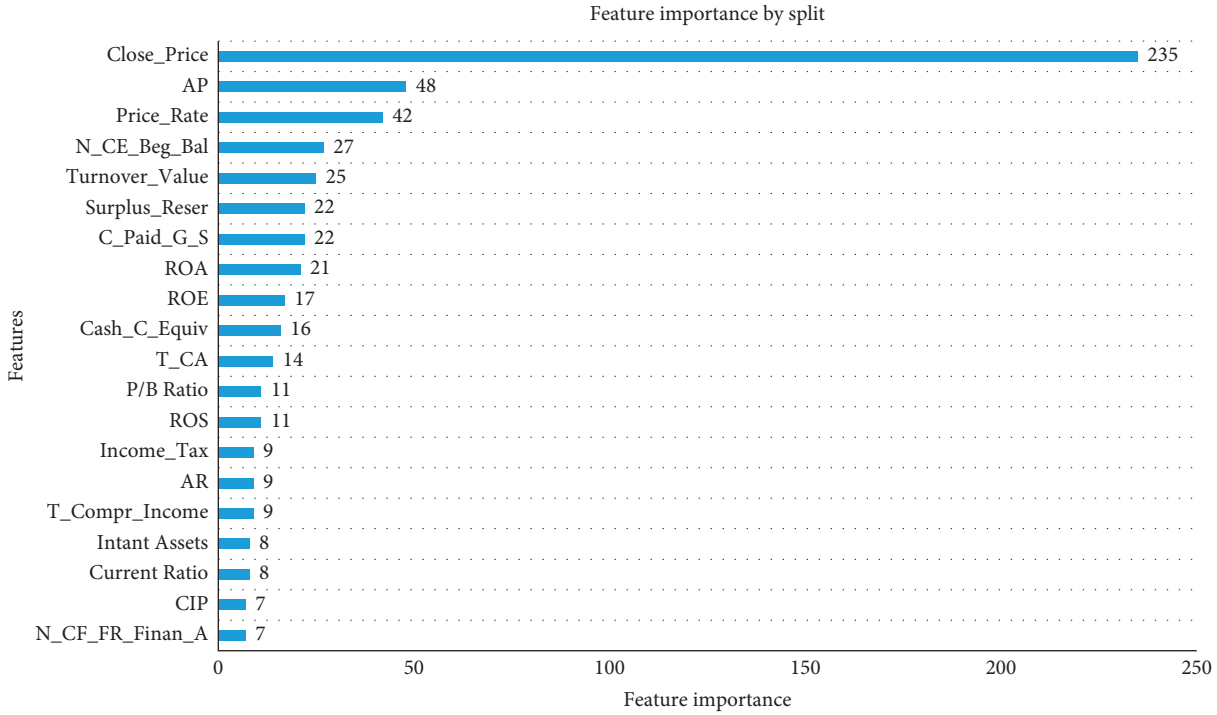

Figure 4: Variable importance score.

is, the more active the stock is. The cash paid for purchasing goods and receiving services is the main cash flow generated in the business activities of industrial and commercial enterprises, which reflects the status of the main business. The surplus reserve is the accumulation of earnings that the enterprise keeps in the enterprise from the after tax profits. It can be used to expand production and operation, increase capital (or share capital), or distribute dividends, which has a direct impact on the stock price. The return on assets is an important index to measure the profitability of a company relative to its total assets. The calculation method is as follows: return on assets = net profit/total assets. The higher the index is, the better the asset utilization effect of the enterprise is, indicating that the enterprise has achieved good results in increasing revenue and saving capital use. It can be seen that LightGBM considers a variety of financial indicators comprehensively, and the importance score also provides a reference for the research and analysis of long-term investment in enterprise value.

It is noteworthy that among the top 20 factors obtained by the model, 40% of the factors are constructed by us according to the fundamental theory. On the one hand, it reflects the scientific nature of combining the fundamental aspect theory with the pure machine learning method. On the other hand, it can also be seen that, in the previous literature, the prediction of stocks completely depending on individual fundamental factors may cause inaccuracy.

4.1.1. Comparisons of Models. In order to compare the accuracy of LightGBM and other algorithms, this paper uses GLM (generalized linear model), DNN (deep neural network), RF (random forest), SVM (support vector machine), and LightGBM to predict the next stock price of each quarter. The results are shown in Figure 5. R-squared measures the goodness of fit, which is equal to the ratio of the sum of squares of regression to the total sum of squares. The closer R-squared is to 1, the better the fitting degree of regression model is. RMSE represents the root mean square error. The smaller RMSE is, the more accurate the prediction is. It can be seen from the left that the R-squared under LightGBM model is 0.798; that is, this model can solve the variation of 79.8% of the stock price, which is higher than the other four methods and indicates that LightGBM is the best to fit the stock price. It is noteworthy that R-squared of the linear model is almost 0.443, and the correlation is very weak. It is speculated that the reason is that the stock price and a large number of factors do not satisfy the linear relationship at the same time. The linear model is only applicable to the model composed of a few fundamental factors. Even if it contains the most influencing factors as much as possible, such a model still lacks explanation for excess return. It can be seen from (b) that the RMSE of LightGBM model is 6.1829, which is lower than the other four methods, also showing the high accuracy of LightGBM.

4.1.2. Risk Assessment of Models. In order to compare the risk of the portfolio of GLM, DNN, RF, SVM, and LightGBM under equal-weight allocation method, market-value weighting method, and minimum-variance weight method of mean-variance model with CVaR constraint, the top N stocks with the highest investment income in stock pool are calculated under these 15 conditions, respectively, and the VaR and CVaR of portfolio investment are calculated at the confidence level of $\alpha = 5\%$, as shown in Figure 6.

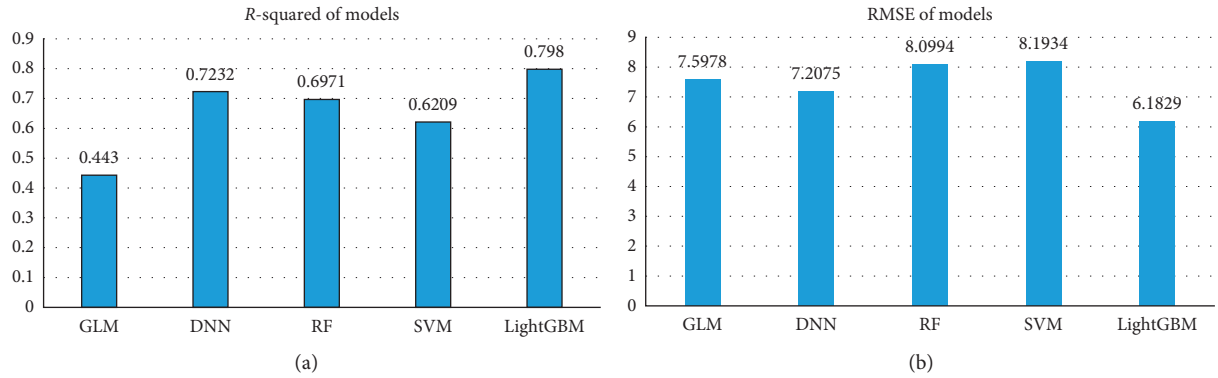

Figure 5: *R*-squared and RMSE comparisons of five models.

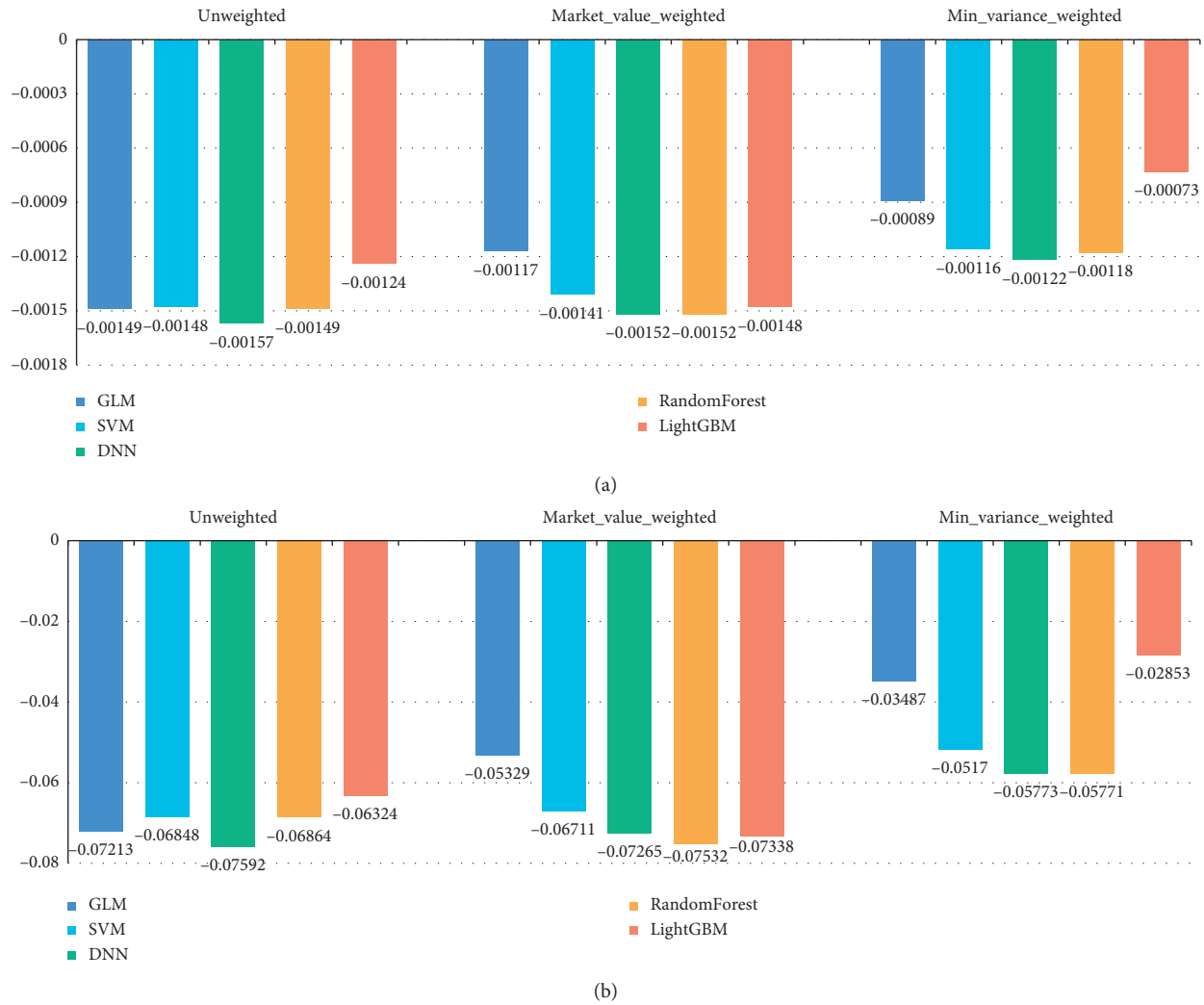

Figure 6: Risk assessment of 5 algorithms under 3 position allocation methods. (a) Var. (b) CVaR.

In the selection of position allocation methods, it can be seen from Figure 6 that, whether the risk is measured by VaR or CVaR, minimum-variance weight method is more able to minimize the risk of the portfolio and reduce the loss compared to the other two allocation methods. At the same time, the overall ranking of algorithms under VaR and CVaR is basically the same. It is because CVaR is based on VaR and CVaR is the optimization of VaR in risk measurement. The two are highly correlated and meet the expectations; thus the model results are reasonable.

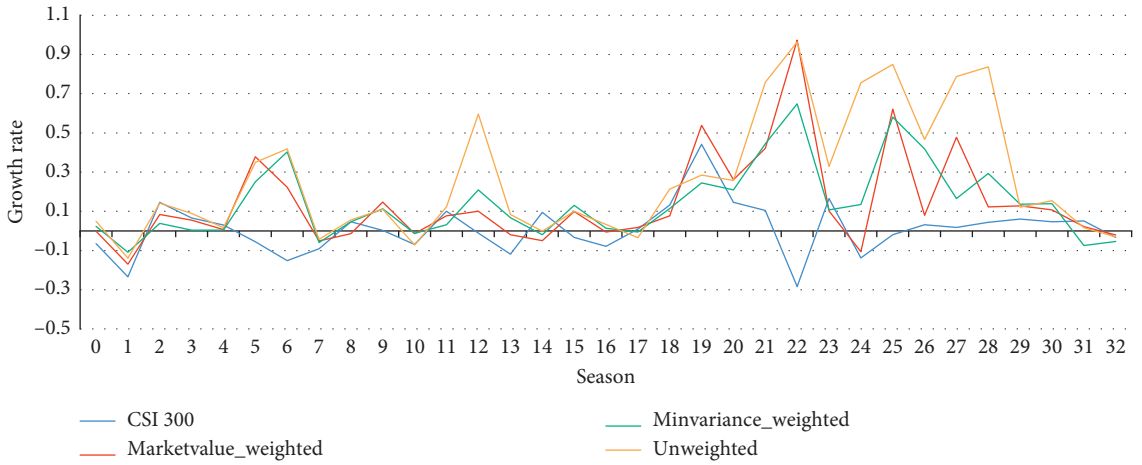

Figure 7: Market index tracking results.

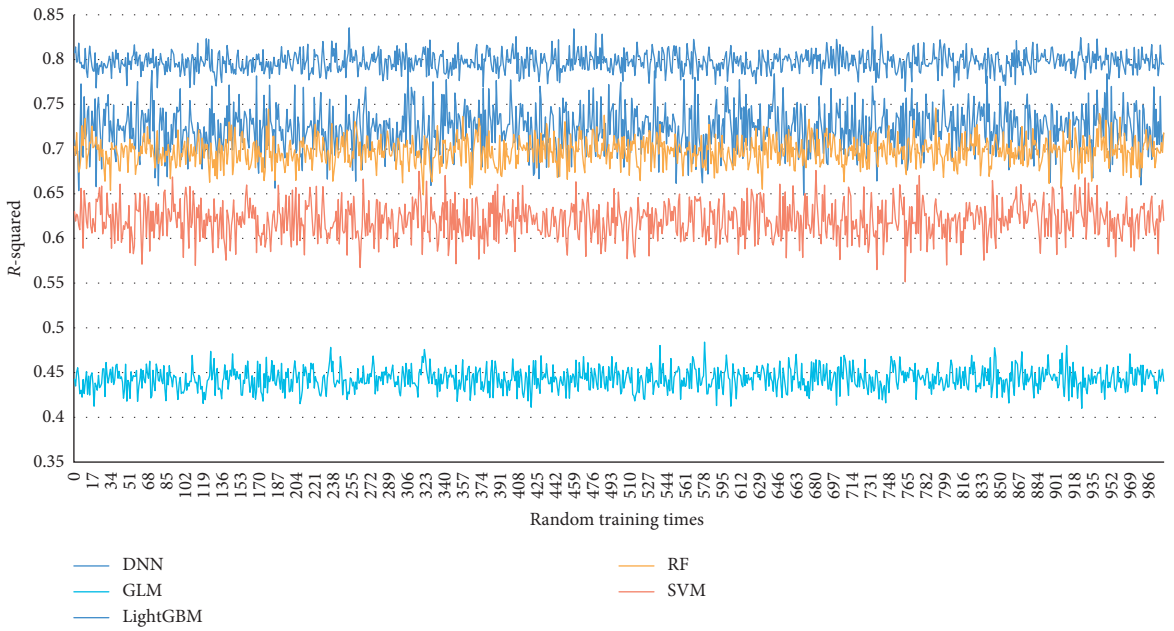

Figure 8: 1000 repetitive tests of 5 models.

In algorithm selection, LightGBM has lower risk compared to the other four models under minimum-variance weight method of mean-variance model with CVaR constraint. Its VaR is −0.0073 and CVaR is −0.02853, which means that, under the normal fluctuation of the stock market, the probability that the return of the optimal portfolio declines by more than 0.73% due to market price change is 5%, and the expected loss of the whole stock portfolio is 2.835%. Under equal-weight allocation method, LightGBM also has lower risk than the other four models, VaR is −0.00124, and CVaR is −0.06324, which shows that it can significantly reduce the risk of portfolio and make the expected return of investors more stable than linear models and traditional machine learning algorithms. It is noteworthy that LightGBM is not the best under market-value weighting method. The reason may be that market-value weighting method pays too much attention to the stocks with higher market value and does not consider the investment value of small- and medium-sized stocks in the stock market well. However, investors in real market often tend to invest in potential stocks with small market value at the early stage of growth, so the reference value of the result of market-value weighting method is limited. On balance, LightGBM has a higher risk reduction effect in a more practical situation.

4.1.3. Yield Analysis. In this paper, we calculate the quarterly yield of stocks in the stock pool based on LightGBM under three position allocation methods and use it to track CSI 300 index in China's stock market from 2009 Q4 to 2018 Q1. From Figure 7, it is obvious that the yield obtained by three position allocation methods can outperform the market, and it is more likely to ensure a considerable yield in a bear

market. Moreover, compared with market-value weighted method and equal-weight method, the optimal portfolio constructed by minimum-variance weight method has a stable performance in the actual market. The combination tracking error constructed by market-value weighting method is the largest, and its performance is not as good as the other two methods in the bear market. Therefore, the portfolio constructed by minimum-variance weight method of mean-variance model with CVaR constraints can continuously obtain relatively stable excess income.

4.1.4. Model Robustness Comparison. The data structure is different for different stock markets. In order to reflect the influence of different data structures on the model proposed in this paper, we use stochastic simulation technology to verify the robustness. In order to keep these dependency structures [17], this paper uses repeated sampling technology to do random simulation. In order to compare the robustness and generalization ability of models, we randomly select 1 year's original data each time and conduct repeated regression predictions1000 times, and the results are shown in Figure 8. According to the comparison of curve volatility, LightGBM can still maintain a lower volatility while maintaining a higher goodness of fit; regardless of the accuracy of prediction or the robustness of the algorithm, LightGBM's regression effect on this type of data set is significantly better compared to the other models.

5. Conclusion

This paper takes the financial risks and returns of the stock market as the research object and uses the method of machine learning and data mining to build a financial trading strategy system based on LightGBM. During data preprocessing and feature engineering, we construct multiple variables that have been proved to have strong significant features in previous multifactor models. Experimented on the market data of China's A-share listed companies, the prediction model in this system is trained to predict the stock price of next quarter. The method will perform in other stock markets. The variable importance score affecting the next stock price is also generated, and the accuracy of our systems is compared with GLM, DNN, RF, and SVM model. At the same time, VaR, CVaR, and quarterly return of the portfolio based on LightGBM are calculated and compared with the market index.

The results show the following:

(1) Compared with the traditional linear model, machine learning models do not need to be based on the assumption that the variables are independent and obey the distribution of specific functions, and they have greater advantages in dealing with big data in financial market. The result of LightGBM is 0.798 for R-squared and 6.1829 for RMSE, which is much better than GLM. The prediction error of LightGBM is also significantly smaller than that of the other machine learning models, which shows that LightGBM has high accuracy.

(2) Compared with equal-weight method and market-value weight method, the portfolio under minimum-variance weight method of mean-variance model with CVaR constraint has the best risk aversion effect. At the same time, the three position allocation methods can outperform the market and are more likely to ensure a considerable yield in a bear market. In general, the portfolio, constructed by minimum-variance weight method of mean-variance model with CVaR constraint, has the best stability and yield, followed by market-value weighting method and equal-weight method.

(3) In this paper, we generate feature importance score to find the most important factor affecting the next stock price. The three most influencing factors are the closing price of the current stock, accounts payable, and the growth rate of stock price. Other financial indicators, like the balance of cash and cash equivalents at the beginning of schedule, turnover rate, and cash paid for goods and services, etc., also have great impact on the stock price.

However, there is still room for improvement in this paper: (1) The experimental data in this paper is the quarterly data of stocks, which is of great value for the long-term strategy. In the future, we can try to use the monthly data of stocks or even the daily data. (2) When we allocate the position of stock, we do not think about shorting, and the weights are limited between 0 and 1, so we can try to add the short strategy in the later research. (3) We choose the mean-variance portfolio model with CVaR constraint for position allocation, but the variance itself is inconsistent. In the future, we can consider using the mean-CVaR model to calculate the weight of the portfolio.

Data Availability

All stock return data are available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] W. F. Sharpe and F. William, "Capital asset prices: a theory of market equilibrium under conditions of risk," *The Journal of Finance*, vol. 19, no. 3, pp. 425–442, 1964.

[2] J. Mossin, "Equilibrium in a capital asset market," *Econometrica*, vol. 34, no. 4, pp. 768–783, 1966.

[3] S. A. Ross, "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, vol. 13, no. 3, pp. 341–360, 1976.

[4] E. F. Fama and K. R. French, "The cross-section of expected stock returns," *The Journal of Finance*, vol. 47, no. 2, pp. 427–465, 1992.

[5] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993.

[6] V. T. Datar, N. Y. Naik, and R. Radcliffe, "Liquidity and stock returns: an alternative test," *Journal of Financial Markets*, vol. 1, no. 2, pp. 203–219, 1998.

[7] J. D. Piotroski, "Value investing: the use of historical financial statement information to separate winners from losers," *Journal of Accounting Research*, vol. 38, pp. 1–41, 2000.
[8] R. Novy-Marx, "The other side of value: the gross profitability premium," *Journal of Financial Economics*, vol. 108, no. 1, pp. 1–28, 2013.
[9] G. Aharoni, B. Grundy, and Q. Zeng, "Stock returns and the miller modigliani valuation formula: revisiting the Fama French analysis," *Journal of Financial Economics*, vol. 110, no. 2, pp. 347–357, 2013.
[10] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of Financial Economics*, vol. 116, no. 1, 2014.
[11] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.
[12] M. T. Greene and B. D. Fielitz, "Long-term dependence in common stock returns," *Journal of Financial Economics*, vol. 4, no. 3, pp. 339–349, 1977.
[13] S.-R. Yang and B. W. Brorsen, "Nonlinear dynamics of daily futures prices: conditional heteroskedasticity or chaos?," *Journal of Futures Markets*, vol. 13, 1993.
[14] J. Fajardo, A. R. Farias, and J. R. H. Ornelas, "Goodness-of-fit tests focus on value-at-risk estimation," *Brazilian Review of Econometrics*, vol. 26, no. 2, pp. 309–326, 2006.
[15] B. Mandelbrot, "The variation of certain speculative prices," *The Journal of Business*, vol. 36, no. 4, pp. 394–419, 1963.
[16] M. Y. Romanovsky, "Truncated levy distribution of sp500 stock index fluctuations. Distribution of one-share fluctuations in a model space," *Physica A*, vol. 287, no. 3-4, pp. 450–460, 2000.
[17] J. Zhang, Y. T. Xie, and J. Yang, "Risk dependence, consistensy risk measurement and portfolio: based on mean-copula-CVaR models," *Journal of Financial Research*, vol. 10, pp. 159–173, 2016.
[18] A. Fan and M. Palaniswami, "Stock selection using support vector machines," in *Proceedings of the International Joint Conference on Neural Networks. (Cat. No.01CH37222)*, vol. 3, IEEE, Washington, DC, USA, pp. 1793–1798, 2001.
[19] K. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1, 2003.
[20] Y. T. Xie, J. Yang, and W. Wang, "Comparison analysis on logistic regression and tree models: based on response ratio of credit mail advertising," *Statistics & Information Forum*, vol. 26, no. 6, pp. 96–101, 2011, in Chinese.
[21] W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & Operations Research*, vol. 32, no. 10, pp. 2513–2522, 2005.
[22] B. B. Nair, V. P. Mohandas, and N. R. Sakthivel, "A decision tree- rough set hybrid system for stock market trend prediction," *International Journal of Computer Applications*, vol. 6, no. 9, pp. 1–6, 2010.
[23] M. Zhu, D. Philpotts, and M. J. Stevenson, "The benefits of tree-based models for stock selection," *Journal of Asset Management*, vol. 13, no. 6, pp. 437–448, 2012.
[24] M. Kumar and M. Thenmozhi, "Forecasting stock index returns using arima-SVM, arima-ANN, and arima-random forest hybrid models," *International Journal of Banking, Accounting and Finance*, vol. 5, no. 3, p. 284, 2014.
[25] S. Bogle and W. Potter, "Using hurst exponent and machine learning to build a predictive model for the Jamaica frontier market," in *Transactions on Engineering Technologies*, pp. 397–411, Springer, Singapore, 2016.
[26] Y. T. Xie and Z. X. Li, "Extension of bonus-malus factor based on joint pricing models," *Statistics & Information Forum*, vol. 30, no. 6, pp. 33–39, 2015.
[27] H. Yan, "Integrated prediction of financial time series data based on deep learning," *Statistics & Information Forum*, vol. 35, no. 4, pp. 33–41, 2020.
[28] Z. Tang, R. Zhu, P. Lin et al., "A hardware friendly unsupervised memristive neural network with weight sharing mechanism," *Neurocomputing*, vol. 332, pp. 193–202, 2019.
[29] Z. Chang, Y. Chen, S. Ye et al., "Fully memristive spiking-neuron learning framework and its applications on pattern recognition and edge detection," *Neurocomputing*, vol. 403, pp. 80–87, 2020.
[30] G. L. Ke, Q. Meng, T. Finley et al., "LightGBM: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3149–3157, 2017.
[31] P. Jorion, "Risk 2: measuring the risk in value at risk," *Financial Analysts Journal*, vol. 52, no. 6, pp. 47–56, 1996.