Application and Comparison of Multiple Machine Learning Models in Finance

Yali Jiang

Guangxi Vocational College of Technology and Business, Guangxi 530015, China

Correspondence should be addressed to Yali Jiang; 1931110371@stmail.ntu.edu.cn

Received 27 January 2022; Revised 19 February 2022; Accepted 26 February 2022; Published 24 March 2022

Copyright © 2022 Yali Jiang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Accurate and effective financial data analysis is very important for investors to avoid risks and formulate profitable investment strategies. Therefore, the analysis of financial data has important research significance. However, the financial market is a complex nonlinear dynamic system affected by many factors. It is very challenging to analyze the financial data according to the obtained information. Among them, stock selection is the most typical financial data mining problem. The core of stock selection is to design a systematic scoring mechanism to quantitatively score stocks so as to more intuitively reflect the investment value of stocks. The scoring mechanism is based on the assumption that stocks with higher scores have higher investment value and stocks with lower scores have lower investment value. The stock selection model proposed in this paper mainly includes two steps: stock prediction and stock scoring. First, construct stock predictors and use machine learning forecasting methods to predict the future price of each stock. Second, construct a stock scoring mechanism to evaluate each stock through the predictive factors and financial factors in the previous step. Finally, select high-scoring stocks and make equal-weight investments. This paper applies the model to the empirical study of the A-share market, verifies its feasibility and effectiveness, and makes a systematic comparison with other benchmark models.

1. Introduction

With the development of the market economy, the issuance and trading of stocks occurred, and both promoted the development of the market economy. In recent years, stocks have shown their tenacious vitality, and the stock market has gradually become an indispensable part of the entire financial industry, especially the securities industry. In order to develop the national economy better and faster, as my country continues to enter the deep water zone of reforms, the reforms in the financial sector are gradually deepening [1]. Predicting the development of stocks has always been the direction of research and exploration by scholars from all walks of life. Predicting the development of stocks is a branch of prediction. The premise is to use scientific methods and methods, and it is necessary to know the development laws of financial markets, understand economics, and fully grasp the law of development of the financial market and the current stage of development [2].

Faced with confusing, high-dimensional, and huge amounts of historical stock information, it is a reasonable and efficient way to analyze and process with the help of computers, which frees stock analysts from tedious work and concentrates on analyzing key issues [3]. The high-speed processing capability of the computer makes it possible to provide customers with more comprehensive and accurate forecasts in a short period of time. In addition, algorithms can provide insights into the forecasting problem from novel perspectives to achieve better forecasting results, which are beyond the reach of human analysis. Model algorithms such as machine learning and deep learning are suitable for the analysis of high-dimensional big data, and the use of computers to customize auxiliary software for stock forecasting for different service objects has become a research hotspot [4]. This paper will combine neural networks and support vector machines, two of the most widely used machine learning models, and propose improvements to provide novel solutions to problems such as model mixing.
and stock forecasting [5]. According to the existing stock analysis and prediction methods, the participation of the whole people can be maximized, and it is not necessary to have sufficient financial expertise to participate in stock investment. For research topics with special properties such as time series and high-dimensional stock historical data, many forecasting methods can be used for reference [6]. Therefore, the research topic of stock forecasting has very important significance from both theoretical and application perspectives and meets the current needs.

2. Preliminary Study

2.1. Problem Definition. The total data set \( D = \{X, Y\} \) is composed of the input factor \( X \) and the real label \( Y \) corresponding to the input factor \( X \), and the prediction result \( Y \) is obtained by using the model \( F(.) \) to predict the data set \( D \). When designing a specific prediction scheme, the input factor \( X \) and label \( Y \) can be differentiated according to the prediction model \( F(.) \) and the purpose of prediction. Therefore, the input factor \( X \) can be time series data, text data, binary features, and so on, and the label \( Y \) can be a specific value, classification result, and so on.

Obviously, financial stock forecasting is a complex problem, which has the following five characteristics: (1) it is appropriate to use machine learning to analyze and predict financial stocks, and the input sources for stock prediction are huge, high-dimensional, and heterogeneous, and machine learning algorithms are particularly good at processing such data [7]; (2) indirect factors such as real-time data on the Internet and public sentiment will also affect the prediction results; (3) when considering numerical data as an input feature, the basic transaction data and the technical indicators calculated from the basic transaction data are huge and need to be selected, and in addition, the two features may affect each other; (4) each stock has its own specific development law, and the trained model algorithm is not necessarily applicable to the data sets of other stocks; and (5) for the prediction output, it can be divided into classification prediction and regression prediction according to the purpose of prediction. Classification and regression forecasts not only reflect trends and stock price forecasts but also provide new ideas for investment strategies and price reversal points [8].

2.2. BP Neural Network. The neural network model is currently the most widely used neural network. At the same time, the neural network has the ability to predict only after training and learning and can adapt to the prediction requirements, obtain the structure of learning knowledge, and map the nonlinearity to linear functions [9]. The main structure diagram and flowchart are shown in Figures 1 and 2.

When the BP neural network predicts stocks, the time series \( \{X\} \) is regarded as the input value of the input layer, and a part of it is selected for training. The price \( \{Y_i\} \) obtained from the training result is compared with the actual price and adjusted by feedback. After multiple iterations, the prediction model is obtained, and then the price of \( a + b + c \) (\( c \geq 1 \)) at a certain time in the future is predicted. The nonlinear relationship of the price \( Y_{a+b+c} \) time series at the predicted time can be expressed by the following formula:

\[
Y_{a+b+c} = f(X_a, X_{a+1}, \ldots, X_{a+b}).
\]

(1)

The fitting model \( f \) is obtained through the training set time series \( X_a + X_{b} + \ldots + X_{a+b} \), and the unvalued value is predicted from this.

2.3. Support Vector Machine Regression. Support vector machine (SVM) was first proposed by Vapnik et al. [10]. Based on the “structural risk minimization” criterion, the optimal hyperplane is constructed to separate the two types of samples to the greatest extent. Support vector machine regression (SVR) is an extension of SVM [11]. The basic principle of SVR is similar to SVM, but its purpose is different. It aims to construct an optimal classification plane to minimize the error of all samples from the classification plane [12].

There are \( n \) training samples, where \( x_i \) is the input variable and \( y_i \) is the output variable. SVR can be regarded as a quadratic programming problem, as shown in the following equation:

\[
\begin{align*}
\min & \quad \frac{1}{2} \|\mathbf{w}\|^2, \\
\text{subject to} & \quad y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) < \varepsilon, \quad i = 1, \ldots, n,
\end{align*}
\]

(2)

where \( \xi \) represents the boundary of the prediction error.

In order to ensure the generalization ability of the model, slack variables \( \xi_i \) and \( \xi_i^* \) are introduced to modify equation (2), namely,

\[
\begin{align*}
\min & \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \\
\text{subject to} & \quad y_i - \mathbf{w} \cdot \mathbf{x}_i - b - \varepsilon < \xi_i, \\
& \quad \mathbf{w} \cdot \mathbf{x}_i + b + y_i - \varepsilon < \xi_i^*, \\
& \quad \xi_i, \xi_i^* \geq 0,
\end{align*}
\]

(3)

where \( C \) is the penalty factor, which is the maximum tolerance error.

![BP neural network structure diagram.](image)
In order to solve equations (3) and (4), the Lagrange function is introduced and transformed into a dual problem as follows:

$$
L = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) - \sum_{i} \lambda_i (\varepsilon + \xi_i - y_i + w_i x_i + b)
- \sum_{i} \lambda_i^* (\varepsilon + \xi_i^* + y_i - w_i x_i - b)
- \sum_{i} (\eta_i \xi_i + \eta_i^* \xi_i^*), \in \eta_i, \eta_i^*, \xi_i, \xi_i^* \geq 0.
$$

(5)

Solve the Lagrange equation and get the optimal solution $w^*$ and $b^*$ as follows:

$$
w^* = \frac{\sum_{i} (\lambda_i - \lambda_i^*) x_i}{\sum_{i} \lambda_i^*},
$$

$$
b^* = y_i - w^* x_i + \varepsilon, 0 \leq \lambda_i \leq C, i = 1, \ldots, n,
$$

$$
b^* = y_i - w^* x_i + \varepsilon, 0 \leq \lambda_i^* \leq C, i = 1, \ldots, n.
$$

(6)

3. The Design of the Financial Stock Selection Model Based on Machine Learning Models

3.1. Stock Prediction Model Framework. The stock selection model proposed in this article mainly includes two steps: stock prediction (used to construct predictive factors) and stock scoring (used to evaluate stock value) [13–15]. The overall framework is shown in Figure 3.

3.2. Financial Stock Forecast and Scoring

3.2.1. Financial Stock Forecast. In order to construct predictors, we predict the returns of all stocks. In the period of $t$, use $z_{i,k,t} (k = 1, 2, \ldots, 17)$ transaction data to predict the stock price in period $t+1$ as follows [13]:

$$
\hat{P}_{i,t+1} = f(z_{i,1,t}, z_{i,2,t}, \ldots, z_{i,17,t}),
$$

(7)

where $\hat{P}_{i,t+1}$ is the predicted price of stock $i$ in period $t+1$ and $z_{i,k,t}$ is the $k$-th transaction data of stock $i$ in period $t$. A total of 17 trading factors are used to predict the future price of stocks. The 17 trading factors can be roughly divided into 4 categories. Volume and price factors include the previous day’s closing price, trading volume, opening price, average price, closing price, lowest price, turnover, and highest price [14]. Valuation factors include $P/B$ ratio, $P/I$ ratio, $P/E$ ratio, and $P/S$ ratio [15]. Risk factors include ups and downs of yuan, ups and downs, and turnover rate. Scale factors include total market capitalization and total equity. The predicted stock price $\hat{P}_{i,t+1}$ is transformed into the predicted stock return through equation (10):

$$
\hat{R}_{i,t+1} = \frac{\hat{P}_{i,t+1} - P_{i,t}}{P_{i,t}},
$$

(8)

where $\hat{R}_{i,t+1}$ represents the predicted return of stock $i$ in the period. Furthermore, normalize $\hat{R}_{i,t+1}$ to obtain $\hat{Y}_{i,t+1}$, and the process is shown in equation (11). $\hat{Y}_{i,t+1}$ will be used as a predictor for the stock scoring in the second step.

$$
\hat{Y}_{i,t+1} = \frac{\hat{R}_{i,t+1} - \bar{R}_{i,t+1}}{D_{i,t+1}},
$$

(9)

where $\bar{R}_{i} = \sum_{t=1}^{N} \hat{R}_{i,t}/N$ is the average value of the predictor and $D_{i,t} = \sqrt{\sum_{t=1}^{X} (\hat{R}_{i,t} - \bar{R}_{i})^2}/N$ is the standard deviation of the predictor, where $N$ is the number of stocks.

3.2.2. Financial Stock Scoring. In addition to predictive factors, financial factors $V_{i,j,t} (j = 1, \ldots, 12)$ have also been introduced into our stock selection model. These factors are widely used in existing stock selection models, and they usually reflect the company’s operating conditions. For example, ROE usually reflects the company’s profitability; debt to equity ratio (DE) usually reflects the company’s bar ratio; current ratio (CR) usually reflects the company’s liquidity; and inventory turnover rate (ITR) usually reflects the company’s operating efficiency and operating income [16]. The growth rate OIG usually reflects the company’s growth and so on. As with forecasted earnings, standardize them as follows [17]:

$$
\hat{Y}_{i,j,t} = \frac{V_{i,j,t} - \bar{V}_{j,t}}{D_{j,t}},
$$

(10)
where $\bar{V}_{jt} = \frac{1}{N} \sum_{i=1}^{N} V_{i,j,t}$ is the average value of financial factors and $D_{jt} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (V_{i,j,t} - \bar{V}_{jt})^2}$ is the standard deviation of the financial factor, where $N$ is the number of stocks.

The stock’s score in the period of $t$, $s_{i,t}$, can be described as a linear combination of predictive factors and financial factors as follows:

$$s_{i,t} = W_0 \bar{Y}_{i,t+1} + \sum_{j=1}^{J} W_j Y_{i,j,t}$$

(11)

where $W_0$ represents the weight of the predictive factor, $W_j$ represents the weight of the $j$ financial factor, and $s_{i,t}$ represents the comprehensive score of the stock in the period.

Each stock is sorted according to its comprehensive score from high to low, denoting $r_{i} \in \{1, 2, \ldots, N\}$, if $s_{i} \geq s_{k}$, then $r_{i} \leq r_{k}$, where $i, k \in \{1, 2, \ldots, N\}$ represents any two separate stocks. In each period, the top $m$ stocks are selected for equal-weight investment. The portfolio returns are as follows:

$$\bar{R}_{t+1} = \frac{1}{m} \sum_{j=1}^{m} R_{t+1}(r_{i,j}),$$

$$r_{i,j} = 1, 2, \ldots, m,$$

(12)

where $m$ is the number of stocks that the investor hopes to choose and $R_{t+1}(r_{i,j})$ is the $t$ return of the stocks selected during the period sorting in period $t+1$, $R_{t+1}$ is the average return of stocks selected according to the model in $t+1$.

This research introduces Spread as the objective function (fitness function) of the model. The objective function (fitness function) is defined as the long-short portfolio return of the top stocks in the $t$ period comprehensive score and the stocks after the $m$ period comprehensive score.

$$\min_{w_i} F = \frac{1}{T} \sum_{t=1}^{T} \text{Spread}_t,$$

$$\text{Spread}_t = \frac{\sum_{i=1}^{m} R_{t+1}(r_{i,t}) - \sum_{r_{i,t} \leq 0} R_{t+1}(r_{i,t})}{m}$$

(13)

3.3. Evaluation Indicator. Root mean square error (RMSE), mean absolute percentage error (MAPE), and directional statistics (D-STAT) are often used to evaluate the effectiveness of prediction models. This article also uses the above three indicators to evaluate the effectiveness of various prediction models in this article as follows [18]:

$$\text{RMSE} = \sqrt{\frac{1}{N \cdot T} \sum_{t=1}^{T} \sum_{i=1}^{N} (P_{i,t} - \bar{P}_{i,t})^2},$$

$$\text{MAPE} = \frac{1}{N \cdot T} \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{P_{i,t} - \bar{P}_{i,t}}{|P_{i,t}|},$$

$$D_{\text{stat}} = \frac{1}{N \cdot T} \sum_{t=1}^{T} \sum_{i=1}^{N} a_{i,t},$$

where $T$ is the number of test periods, $N$ is the number of stocks, $P_{i,t}$ is the price of stock $i$ in period $t$, $\bar{P}_{i,t}$ is the predicted price of the stock, and when $(P_{i,t} - P_{i,t-1}) (\bar{P}_{i,t} - P_{i,t-1}) > 0$, $a_{i,t} = 1$; otherwise it is 0. In
addition, the running time of the model is also an important indicator for evaluating the performance of the model. It will cover the entire time of the training period and the test period.

Choosing high-quality stocks for equal-weight investment means constructing an equal-weight investment portfolio. Portfolio income can most intuitively reflect the effectiveness of the stock selection model. The higher the portfolio return, the better the effect of the stock selection model. Therefore, this paper adopts the average return of all investment portfolios during the test period as one of the evaluation indicators of the stock selection model. In addition, portfolio risk is also a key concern. SharpeRatio is a commonly used indicator to describe portfolio risk. This article also uses SharpeRatio as one of the evaluation indicators of the stock selection model. The specific calculation formula is as follows [19]:

\[
AR = \frac{1}{T} \sum_{t=1}^{T} R_{t+1}, \quad \text{Sharpe Ratio} = \frac{E[R_t - R_f]}{\sigma}.
\]

where \( R_t \) is the average return of the portfolio in the \( t \) period, that is, the average value of the strategy return; \( R_f \) is the risk-free return in the \( t \) period; \( E[R_t - R_f] \) is the expectation; and \( \sigma \) is the portfolio volatility, that is, the strategy return.

3.4. Hyperparameter Setting. All parameters involved in the model and benchmark model proposed in this paper refer to previous studies, as shown in Table 1. Regardless of whether it is an optimization model or a prediction model, there are random variables or parameters, and different optimization values will be generated every time the model is run (the output result cannot be guaranteed to be the optimal result). Therefore, all models are run 30 times for each case, and their average value is used as the final output result.

4. Experimental Test and Result Comparison

4.1. Experimental Data Set. The trading volume and total market value of China’s A-share market in the global financial market are increasing day by day, and its status is increasing day by day. Therefore, the A-share band field is selected as the experimental sample of this article. There are a total of 2,473 stocks as candidate stocks for the investment portfolio, which excludes financial industry stocks with different balance sheet structures and specially processed stocks ST that may be unstable. In the existing stock forecasting models, this chapter will introduce the acquisition, processing, and cleaning of data sources from two aspects of stock index data and text data. Among them, stock index data refers to data related to stock trading. The text data mainly consists of news data. The trading data used for stock forecasting and the financial data of stock scoring are all quarterly data, and they all come from the Wind database.

4.2. Benchmark Model Comparison. As shown in Table 2, this article introduces a variety of benchmark models, including different prediction models (marked as M1), different factorial designs (marked as M2), different optimization algorithms (marked as M3), and different fitness functions (marked as M4). Among them, M0 is the stock selection model proposed in this article. Next, we will introduce the construction methods of various benchmark models in detail.

In each benchmark method, in order to ensure the fairness of comparison, each type of benchmark model only changes its unique variable. For type M1, BP neural network (BPNN) and support vector machine regression (SVR) are two classic forecasting models, which are widely used in stock forecasting, investment portfolio, and stock selection research. Therefore, this article will introduce these two prediction models as the benchmark model. For type M2, the model proposed in this paper contains two types of factors, namely, basic factors and predictive factors, which are denoted as A0. Two new designs are now introduced as the benchmark model of this article. Design A1 means that only financial factors are considered but not predictive factors, and design A2 means that only predictive factors are considered but not financial factors. For type M3, the more popular intelligent optimization algorithm in the stock selection model is introduced as the benchmark model of this article, namely, genetic algorithm GA and particle swarm algorithm PSO. For type M4, in the stock selection model based on the intelligent optimization algorithm, four commonly used optimization objective functions (fitness functions) are used as the benchmark model of this article, namely, information coefficient (IC), equity (CR), long-short combination winning rate (IFHR), and absolute win rate (WIN).

\[
IC = -\frac{1}{T} \sum_{t=1}^{T} \frac{\cos(r_{i,t}, r_{i,t}')} {\var{r_{i,t}} \var{r_{i,t}'}} ,
\]

\[
CR = -\frac{1}{T} \sum_{t=1}^{T} \left( \left( 1 + \tilde{R}_{t+1} \right) \prod_{t=1}^{M} \left( \sum_{r_{i,t}=1}^{m} \sgn \left( \tilde{R}_{t+1}(r_{i,t}) - M_{t+1} \right) \right) \right) \]

\[
IFHR = -\frac{1}{T} \sum_{t=1}^{T} \left( \sum_{r_{i,t}=1}^{m} \sgn \left( \tilde{R}_{t+1}(r_{i,t}) - M_{t+1} \right) \right) \]

\[
WIN = -\frac{1}{T} \sum_{t=1}^{T} \sgn, \quad (16)
\]

where \( T \) is the number of training periods, \( r_{i,t} \) is the ranking of stock \( i \) in the \( t \)-th period, \( r_{i,t} \) is the stock’s income ranking in \( t + 1 \) period, \( \cov(·) \) is the covariance, \( \var(·) \) is the variance, \( m \) is to select the number of stocks, and \( \tilde{R}_{t+1} \) is the average return of the stock selection strategy in the \( t + 1 \) period. It outputs 1 when it is \( x > 0 \); otherwise, it outputs 0. See Table 1 for benchmark model design.
4.2.1. Comparison of Different Prediction Models (M1).

When constructing predictors, this article uses the extreme learning machine ELM to predict stock prices. In order to prove the effectiveness of ELM in stock forecasting, two classic forecasting methods, BP neural network (BPNN) and support vector machine regression (SVR), are introduced as benchmark models. Table 3 shows the prediction results. Among them, Prob.(R1) is the probability that the stock selection model of this article will have a higher quarterly return than the quarterly earnings of R1; Prob.(R2) is the probability that the quarterly earnings of the stock selection model of this article will be higher than the quarterly earnings of R2; and Max. is the average value of the maximum return of the stock selection model of this article; Min. is the average value of the minimum return of the stock selection model of this article; and HitRate represents the probability of the model obtaining a positive return [20].

As shown in module A in Table 3, the stock selection model using ELM performs more prominently in AR, SharpeRatio, Prob.(R1), Prob.(R2), and HitRatio. At the same time, the SVR stock selection model has achieved a little advantage in Max. and Min. while the stock selection model of BPNN is at a disadvantage in any evaluation index. This shows that the predictive factors constructed by ELM can better assist stock selection decisions. In order to statistically prove that the stock selection model using ELM is significantly better than other benchmark models, this section tests each benchmark model. The process is as follows: first, the normality test is performed on each model, and the normality test is the prerequisite for the test. Second, construct a null hypothesis, H0: the AR of the stock selection model using ELM is significantly lower than other benchmark models. According to module B, all models pass the
normality test, and the values of all models are less than 5%. It shows that at the 95% confidence level, the stock selection model based on ELM is better than the benchmark model. This proves the predictive factors constructed by ELM can better assist stock selection decision-making.

Module C shows the evaluation indicators of all forecasting methods. Obviously, the method of using ELM to predict stock prices crushes all benchmark models in computing time, directional accuracy D-STAT, prediction accuracy MAPE, and RMSE. In the research of this article, ELM has more prominent predictive abilities. Similarly, in order to statistically prove that the prediction results based on the extreme learning machine ELM are significantly better than other benchmark models, this section conducts the Diebold–Mariano test (DM test) on each benchmark model. The process is as follows: construct a null hypothesis, \( H_0: \) the prediction result based on ELM is significantly lower than the prediction result based on the benchmark model. The DM test results are shown in module D in Table 3. All values are less than 5%, indicating that the prediction results based on ELM are better than the benchmark model at the 95% confidence level. This proves that ELM has a stronger predictive ability. It can be seen that the predictor based on ELM structure can better assist stock selection decision-making.

### 4.2.2. Comparison of Different Factorial Designs (M3)

The innovation of the stock selection model proposed in this paper is the use of predictive factors and financial factors to evaluate the value of stocks. The predictive factors contain the future information of the stock market, and the basic factors contain historical information of the stock. In order to prove the effectiveness of this innovative design, the proposed stock selection model A0 is compared with the benchmark model A1 and the benchmark model A2 (as shown in Table 4). Among them, the benchmark model A1 only uses basic factors, similar to the traditional stock selection model, for which refer to the research done by Huang, and the benchmark model A2 only uses predictive factors, for which refer to the research done by Huang and Quah.

Table 4 shows the relevant results. At the same time, the A0 stock selection model has achieved outstanding performance in AR, SharpeRatio, Min., and Prob.(R1). However, only the A1 stock selection model and the A2 stock selection model have no outstanding performance in the various evaluation indicators. In order to statistically prove that the stock selection model using A0 is significantly better than other benchmark models, this section tests each benchmark model. The process is as follows: first, the normality test is performed on each model, and the normality test is the prerequisite for the test. Secondly, construct a null hypothesis, \( H_0: \) the average return AR of the stock selection model A0 using predictive factors and financial factors is significantly lower than other benchmark models. According to module B, all models pass the normality test, and the values of all models are less than 5%. It shows that under the 95% confidence level, the stock selection model A0 using predictive factors and financial factors is better than other benchmark models. This proves that adding predictive factors can better assist stock selection decisions, and financial factors are equally important.

#### 4.2.3. Comparison of Different Optimization Algorithms (M3)

In this paper, the differential evolution algorithm is used to optimize the weights of various factors. In order to prove the effectiveness of DE, the classic genetic algorithm GA and particle swarm algorithm PSO are set as the benchmark model. As shown in Table 5, during the training and testing period, the DE stock selection model used in this paper is superior in average return AR and SharpeRatio, which not only shows that DE algorithm can better assist stock selection decision-making but also shows that the algorithm can achieve the best performance. The optimized factor weights have stronger generalization ability than the benchmark model. In order to statistically prove that the stock selection model adopted is significantly better than other benchmark models, this section tests each benchmark model. The process is as follows: first, the normality test is performed on each model, and the normality test is the prerequisite for the test. Secondly, construct a null hypothesis, \( H_0: \) the AR of the stock selection model using DE is significantly lower than other benchmark models. As shown in Table 6, all models passed the normality test, and the values of all models were less than 5%. It shows that under the 95% confidence level, the DE-based stock selection model is better than other benchmark models. Thus, it is statistically proved that the differential evolution algorithm can better assist the stock selection decision-making.

#### 4.2.4. Comparison of Different Fitness Functions (M4)

In this paper, the Spread fitness function (old standard han number) is used to make stock selection decisions, and four different fitness functions such as IC, CR, IFHR, and WIN are now used as benchmark models. The results are shown in Table 6. The stock selection model using the Spread fitness function has outstanding performance in average return AR and Sharpe Ratio. The stock selection model using the Spread fitness function can better assist stock selection decisions.

<table>
<thead>
<tr>
<th>Model</th>
<th>A0</th>
<th>A1</th>
<th>A2</th>
<th>R1</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module A: model performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>0.1102</td>
<td>0.0765</td>
<td>0.0799</td>
<td>0.8403</td>
<td>0.0398</td>
</tr>
<tr>
<td>SharpeRatio</td>
<td>0.4010</td>
<td>0.2979</td>
<td>0.3012</td>
<td>0.2988</td>
<td>0.1798</td>
</tr>
<tr>
<td>Max.</td>
<td>0.4587</td>
<td>0.4099</td>
<td>0.4010</td>
<td>0.3789</td>
<td>0.3594</td>
</tr>
<tr>
<td>Min.</td>
<td>0.2998</td>
<td>0.3201</td>
<td>0.3587</td>
<td>0.3099</td>
<td>0.3178</td>
</tr>
<tr>
<td>Prob.(R1)</td>
<td>0.6549</td>
<td>0.4378</td>
<td>0.5001</td>
<td>NA</td>
<td>0.0000</td>
</tr>
<tr>
<td>Prob.(R2)</td>
<td>0.8001</td>
<td>0.7603</td>
<td>0.7554</td>
<td>1.0000</td>
<td>NA</td>
</tr>
<tr>
<td>HitRatio</td>
<td>0.7602</td>
<td>0.7812</td>
<td>0.7399</td>
<td>0.7849</td>
<td>0.5645</td>
</tr>
<tr>
<td>Module B: t-test (( H_0: AR_{DE} \leq AR_{Benchmark} ))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normality test (( p ) value)</td>
<td>0.2998</td>
<td>0.0489</td>
<td>0.3379</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>t stat</td>
<td>NA</td>
<td>20.5879</td>
<td>14.1339</td>
<td>25.8999</td>
<td>57.0004</td>
</tr>
<tr>
<td>p value</td>
<td>NA</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
The selection model is the difference. Model A1 (considering only financial factors) and the This paper has a significant advantage over the benchmark different factors, the stock selection model proposed in introducing stock predictors can improve stock selection. This also supports the new idea that in-optimization, predictors were given higher weights, suggesting that stock predictors play an important role in stock selection. Thus, it is statistically proved that the stock selection model using the Spread fitness function can better assist stock selection decision-making.

5. Conclusion

This paper uses two machine learning models, neural network and support vector machine, to predict stock prices and conduct empirical research on China’s A-share market. The results show that the machine learning model adopted in this paper has obvious advantages in terms of ROI and model robustness. The returns of portfolios constructed using the stock selection model proposed in this paper are much higher than the average market performance (i.e., an equal-weighted portfolio of all stocks) and the China A-share index. Notably, in model optimization, predictors were given higher weights, suggesting that stock predictors play an important role in stock selection. This also supports the new idea that introducing stock predictors can improve stock selection decisions. Especially when comparing the designs of different factors, the stock selection model proposed in this paper has a significant advantage over the benchmark model A1 (considering only financial factors) and the benchmark model A2 (considering only predictive factors). This paper provides empirical experience and model design guidance for using machine learning for stock selection.

Data Availability

The data set can be accessed upon request to the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the 2020 Guangxi Vocational College of Technology and Business Party-Building Project “Research on Social Governance Improvement of the Information Construction of Primary-Level Party Organizations” (XY2020DJYB007) and the 2021 School-Level Scientific Research Project of Guangxi Police College, “Research on the Primary-Level Party’s Modernization of Social Governance Enabled by Information Technology.” 2022 Annual Basic Research Ability Improvement Project of Young and Middle-Aged Teachers in Guangxi Universities “Research on the innovation trend of quantum technology enabling smart bank in the Era of ABCDI technology” (number: 2022KY1350), and 2022 Annual Basic Research Ability Improvement Project of Young and middle-aged Teachers in Guangxi Universities “Research on the Innovation and Development Path of University Human Resource Management in the Era of Big Data” (number: 2022KY0860).

References


