Predicting the Price of SP500 Index Based on Machine Learning Methods

Xing Wei

Quantitative finance, Northeastern University, MA/02115/US

*Corresponding author. Email: wei.xing2@northeastern.edu

Abstract. This paper mainly introduces the machine learning algorithm to predict the rise and fall of SP500 stock return prediction. Data of stock trading in the past 12 years (opening price, highest price, lowest price, and closing price) were adopted and preprocessed as sample data. Finally, nine technical parameters were adopted in Support Vector Machine and Random Forest models to predict the rise and fall of stocks. For parameters, it was divided into discrete variables, and continuous variables then are used. In the discrete variables and continuous variable part, the F1 score result of Support Vector Machine were 0.90 and 0.89, and the F1 score result of Random Forest were 0.91 and 0.96. Therefore, it can be concluded that the Random Forest model is better than the Support Vector Machine model.

Keywords: Machine learning algorithm; Random Forest model; Support Vector Machine model

1 Introduction

Predicting the rise and fall of stocks is a dream of every investor, but big data has made it possible. The Efficient Market Hypothesis (EMH) published by Eugene Fama showed that in the strong form of efficiency current prices reflect all information which is both public and private and includes information known only to insiders [1]. The United States is a strong form efficiency market, so this article uses the SP500 in American stocks as the research object. The fluctuation of the stock price in the stock market corresponds to the change of information, so we can find nine factors more accurately.

Economic Factor Analysis is one of the models when predicting stock fluctuations. In this model, the excess return is equal to alpha (measures the risk-adjusted return on the stock) plus the return on factor at time t multiplied by the factor loads a measure the sensitivity of a stock's return to these factor shocks. Another model is Fundamental Factor Analysis. The return of stocks is equal to alpha plus observing stocks' exposures to the factors at a certain point in time multiplied by estimating the factor return for that period. The focus of the fundamental factor model is on identifying what actually gets rewarded in the stock market and measuring the magnitude of the reward.

Patel et al. used a single machine learning model [2]and hybrid method SVR-ANN,

H. Mallick et al. (Eds.): ICEMCI 2022, AEBMR 231, pp. 527–534, 2023.
https://doi.org/10.2991/978-94-6463-098-5_59

SVR-RF, and SVR-SVR fusion prediction model [3]. Then, 10 technical indicators were selected as the input of each prediction model. In the single machine learning model part, through comparison, it was concluded that the overall performance of the random forest was better than its artificial neural network (ANN), Support Vector Machine, and Naive Bayesian Prediction model. And the trend certainty of parameters was positively correlated with the effect of the prediction model. In the hybrid method part, the hybrid method is always better than SVR single model. Thomas Fischer et al. proposed the LSTM Network to predict the out-of-sample movement of SP 500 Index constituent stocks from 1992 to 2015 in 2018, and then compared Random Forest, Deep Neural Network, and Logistic Regression Classifier. LSTM return had low exposure to common system risk sources [4]. Eric et al. adopted the Winsorization method to deal with outliers in data in 2021 Outliers were the tails of the empirical distribution. Negative (positive) outliers were in the bottom(top) 1% of the sample. Negative (positive) outliers were reassigned to the 1st (99th) percentile taken from the sample [5]. Zheng et al. used the Random Forest model to test Chinese stocks and used two kinds of feature spaces: basic/technical feature space and pure momentum feature space to predict long-term and short-term price trends respectively. Finally, it is concluded that these two characteristics have brought excess returns [6]. Reza et al. adopted the stock selection step of index tracking in passive investment management and incorporated the continuous changes in market dynamics into the decision-making. Through the application of the proposed program tracking index by co-integration technology, the performance of the selection method used in the previous market state was analyzed, and the method to provide the best tracking portfolio in each period was determined. Finally, it showed that it provides better index tracking quality than the traditional method of using a single method/standard to select assets [7]. Huang et al. studied the predictability of SVM on the direction of financial movement by predicting the weekly movement direction of the Nikkei 225 index. They compared SVM with Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Elman Back Propagation Neural Network. Experimental results showed that SVM was superior to other classification methods [8]. The artificial Fish Swarm Algorithm (AFSA) was introduced by Shen et al. to train the Radial Basis Function Neural Network. Their experiments on the stock index of the Shanghai Stock Exchange showed that the RBF algorithm optimized by AFSA was an easy-to-use algorithm with considerable accuracy [9]. Tsai et al. studied the prediction performance of using the classifier ensemble method to analyze stock returns. The mixed method of majority voting and bagging was considered. In addition, the performance of using two classifier sets was compared with that of using a single baseline classifier (i.e., Neural Network, Decision Tree, and Logistic Regression). The results show that the performance of multiple classifiers is better than that of a single classifier in terms of prediction accuracy and return on investment [10].

This paper preprocessed MA, Bolling, Aroon, CCI, CMO, MACD, RSI, Stoch, and WILLR technical indicators, and then turned them into discrete variables and continuous variables. Support Vector Machine and Random Forest model were used to calculate accuracy, precision, recall, and F1 score, and then compared which model was more suitable for predicting the rise and fall of stock prices.

2 Data Research

According to Wikipedia, the SP 500 index was founded by Standard & Poor's which was the world's authoritative financial analysis institution. It records the stock index of 500 listed companies in the United States. The early SP500 index was composed of 425 industrial stocks, 15 railway stocks, and 60 utility stocks. Now it is composed of 400 industrial stocks, 20 transportation stocks, 40 utility stocks, and 40 financial stocks. It took the average market price of sampled stocks from 1941 to 1943 as the base period, took the number of listed stocks as the weight, and was weighted according to the base period. Here its base number was 10. The stock price index was equal to the stock market price multiplied by the number of shares issued on the stock market, then divided by the stock market price of the base period multiplied by the number of shares in the base period, and finally multiplied by 10. SP index has the characteristics of a wide sampling area, strong representativeness, high accuracy, and good continuity. It is generally considered to be an ideal target for stock index futures contracts. Compared with the Dow Jones industrial stock index, the SP index is more suitable for analyzing the long-term trend of stock prices.

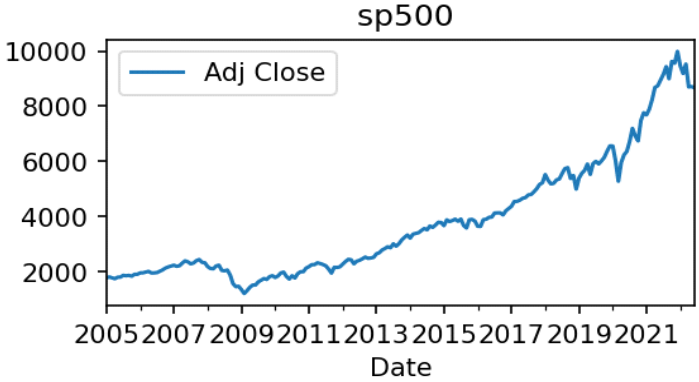

Fig. 1. SP500 Adj Close price (Picture source: Self-painted)

In Figure 1, we can find that there are huge declines in 2008, 2020, and 2021, respectively. Each decline was accompanied by the occurrence of the following events: in 2008, the United States broke out in the sub-prime crisis, the 2020 covid-19 pandemic in the United States, and the 2021 war and interest rate hike between Russia and Ukraine.

This article took the rise and fall of stocks as the Y variable and MA, Bolling, Aroon, CCI, CMO, MACD, RSI, Stoch, and WILLR as the independent variable X value. In discrete variables, all nine technical parameters are compared with SP500 one-day moving average index which is a short-term moving index parameter. When the output value is 0, it means that the stock price falls; when the output value is 1, it means that the stock price rises.

In the discrete variables part, the MA index is the moving average index. This paper took 10 days as the moving average index as the short-term moving index parameter. When the value of MA is less than the value of SP500, the output value is 0, otherwise, it is 1. Bolling index uses statistical principles to calculate the standard deviation and confidence interval of stock price. Under normal circumstances, the upper line of the Bolling index represents the highest price of the stock price, and the upper line will exert pressure on the stock price. The lower line represents the time when the stock price is the lowest, which plays a supporting role in the stock price. When the Bolling value is less than SP500, the output value is 0, otherwise, it is 1. Aroon index is to help investors predict trends and trend ranges. Here, the time period is 14. When the output values of Aroon up and Aroon down are less than 70, 0 is output, otherwise, it is 1. CCI index measures whether the stock price exceeds the normal distribution range. When the CCI index goes up beyond 200, the output is 1; when the CCI index goes down beyond -200, the output is 0. The numerator of the CMO index calculation formula adopts the data of rising and falling days. The CMO index is to find the conditions for extreme overbought and oversold. When the CMO index is less than 0, the output is 0, otherwise, it is 1. MACD is a technical index that uses the aggregation and separation between the short-term (usually 12 days) index moving average and the long-term (usually 26 days) index moving average of the closing price to study and judge the buying and selling opportunities. When the MACD index is less than 0, the output value is 0, otherwise, it is 1. RSI index predicts the strength of the market movement trend by calculating the range of stock price rise and fall and predicts the continuation or direction of the trend. The strength index mostly fluctuates between 70 and 30. When the six-day index rises to 80, it indicates that the stock market has been overbought. Once it continues to rise, it exceeds 90, it indicates that it has reached the warning zone of serious overbought, and the stock price has formed a head, which is likely to reverse in the short term. But in this article, I take 50 as the dividing line of RSI. When RSI is less than 50, the output value is 0, and vice versa The stop index is the KD index in our commonly used KDJ index. It consists of two lines, one is a fast confirmation line, and the other is a slow trunk line. When the K and D lines of Stoh are less than 50, the output value is 0, otherwise, it is 1. WillR mainly analyzes the relationship between the highest price, the lowest price, and the closing price of the stock price over a period of time to judge the overbought and oversold phenomenon of the stock market and predict the medium and short-term trend of the stock price In this article, taking 14 days as a cycle, when the value of Willr is greater than 1, the output is 0, otherwise, it is 1.

In the test of continuity variables, this article takes the direct output values of MA, Bolling, Aroon, CCI, CMO, MACD, RSI, Stoch, and WILLR as independent variables, and then takes the rise and fall of stock price (0, 1) as dependent variables.

3 Models

3.1 Random forest

Before introducing the Random Forest, let us briefly introduce the decision tree. The decision tree classifies all stocks according to a certain feature of stocks and then calculates the Information Entropy. For example, according to the specification, it can be divided into large, medium, and small.

$$i(p) = -\Sigma_j P(W_j) \log_2 P(W_j) \qquad (1)$$

Random forest is an integrated method that combines different decision tree structures, as shown in Figure 2. The Random Forest Algorithm builds each decision tree according to the following two-step method. The first step is called "row sampling". Sampling from all samples follows the sampling method of putting back, and then a bootstrap data set is obtained. The second step is called "column sampling", which randomly selects m features from all M features (original data) and trains a decision tree with m features of the bootstrap data set as a new training set. Finally, all n decision trees are combined by voting. Each tree in the random forest uses only one factor, but when combined, it is more stable and not easy to overfit. In addition to good stability, the Random Forest has the advantages of fast training and prediction, which is suitable for processing many high-dimensional data. At the same time, the algorithm can calculate the importance of each feature, so it has become a very popular machine learning algorithm.

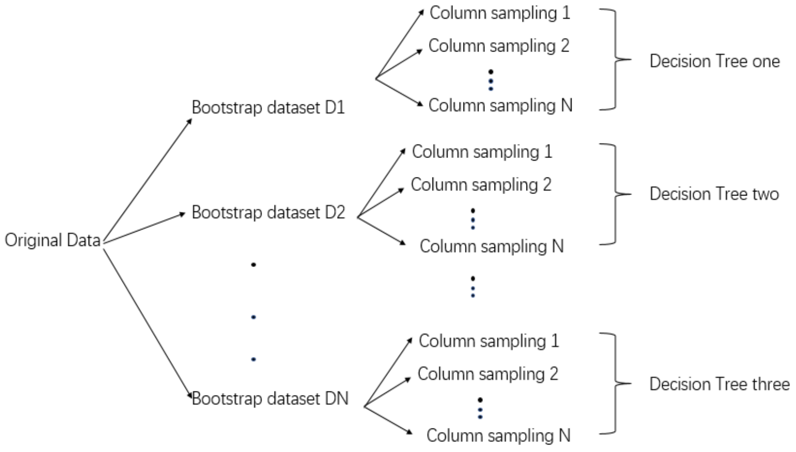

Fig. 2. Decision tree (Picture source: Self-painted)

3.2 Support vector machine

Support vector machine synthesizes the original features into new features through linear combination. Its core idea is to use a hyperplane to divide the sample space, so as to solve the classification problem, as shown in Figure 3. However, here we use the method of introducing kernel to calculate the hyperplane. Kernel functions are usually divided into three types: linear kernel, polynomial kernel, and Gaussian kernel.

- linear kernel: K(Xi,Xi')= $\sum_{j=1}^{p} XijXi'j$

- polynomial kernel: K(Xi,Xi')= $(1 + \sum_{j=1}^{p} \mathrm{XijXi'j})$^D, where D is the order of the polynomial
- Gaussian kernel: K(Xi,Xi')= $\exp(-\gamma(\sum_{j=1}^{p} \mathrm{Xij} - \mathrm{Xi'j})$^2

Its advantage is that it can analyze nonlinear classification through kernel function research, which breaks the phenomenon of overfitting small-scale data based on empirical risk minimization theory.

In some linear inseparable problems, it may be nonlinear separable, there is a hypersurface in the feature space to separate the positive and negative classes. Nonlinear separable problems can be mapped from the original feature space to a higher dimensional Hilbert space using nonlinear functions. In the support vector machine model, the kernel function of linear polynomial kernel is used.

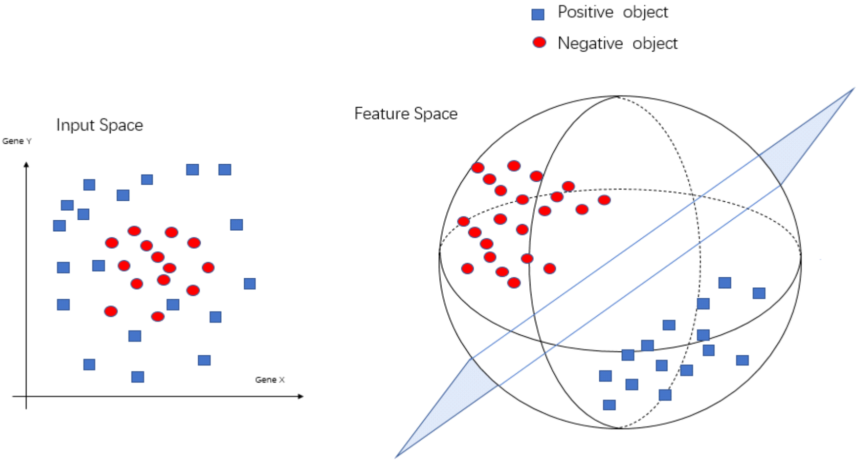

Fig. 3. Support vector machine (Picture source: Self-painted)

4 Result

In Support Vector Machine and Random Forest, both adopt kernel functions that use the same polynomial kernel. Four parameters: accuracy, precision, recall rate, and F1 score are adopted. The accuracy parameter is used to calculate the percentage of correct results. When the percentage is larger, it indicates that the output result is more accurate. The precision parameter is to refine the test results. For example, when the output result is 1, how many of these 1 will correspond to the rise of SP500. When the percentage of precision is larger, the accuracy of prediction results is higher. The recall parameter is the probability that SP500 will be guessed. When the percentage of recall is larger, the probability of guessing is higher. F1 score is an indicator of comprehensive accuracy, precision, and recall rate. F1 score is equal to two times of recall multiplied by precision and then divided by the sum of recall and precision. When the F1 score reaches 1, it indicates that the precision of recall and precision is very high, both of which are 1. In the test of random forest, accuracy, precision, recall rate, and F1 score are also used as indexes to test. Compare the Support Vector Machine and Random Forest by drawing table 1. In discrete variables, the accuracy, recall, and F1 score of the Random Forest are better than that of the Support Vector Machine. In continuous variables, the accuracy, precision, recall rate and F1 score of the random forest are higher than those of the support vector machine. Finally, it is concluded that the Random Forest model is better than the Support Vector Machine model.

Table 1. The output of SVM and RF (Table source: Self-painted)

Discrete variable	SVM	Rand Forest
accuracy	0.88	0.89
precision	0.92	0.89
recall	0.88	0.92
F_1 score	0.90	0.91

Table 2. The output of SVM and RF (Table source: Self-painted)

Discrete variable	SVM	Rand Forest
accuracy	0.88	0.96
precision	0.89	0.96
recall	0.89	0.96
F_1 score	0.89	0.96

5 Conclusion

This paper obtains the historical data of SP500 from 2001 to 2022, then processes MA, Bolling, Aroon, CCI, CMO, MACD, RSI, STOCH, and WILLR, turns them into discrete and continuous variables, and then the accuracy, precision, recall rate and F1 score measurement accuracy of these indicators are calculated by support vector machine and random forest. Finally, compare which model is more effective in predicting the rise and fall of stock price. From the above table, it can be concluded that the output of the four indicators of the Random Forest model is higher, and it is better to predict stock price. This article also needs to be improved. First, is the problem of the SVM model itself. Because Support Vector Machine uses quadratic programming to solve support vectors, when the number of M is large, the storage and calculation of the matrix will consume a lot of machine memory and calculation time. Support vector machine is difficult to solve multi-classification problems. Second, this article only uses a single model, not a hybrid model of support vector machine and Rand Forest.

References

1. Fama. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance (New York), 25(2), 383–. https://doi.org/10.2307/2325486
2. Patel, Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162–2172. https://doi.org/10.1016/j.eswa.2014.07.040
3. Patel, Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259–268. https://doi.org/10.1016/j.physa.2021.126810
4. Fischer, & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054
5. Sharma, & Chatterjee, S. (2021). Winsorization for Robust Bayesian Neural Networks. Entropy (Basel, Switzerland), 23(11), 1546–. https://doi.org/10.3390/e23111546
6. Tan, Yan, Z., & Zhu, G. (2019). Stock selection with random forest: An exploitation of excess return in the Chinese stock market. Heliyon, 5(8), e02310–e02310.https://doi.org/10.1016/j.heliyon.2019.e02310
7. Bradrania, Pirayesh Neghab, D., & Shafizadeh, M. (2021). State-dependent stock selection in index tracking: a machine learning approach. Financial Markets and Portfolio Management, 36(1), 1–28. https://doi.org/10.1007/s11408-021-00391-7
8. Huang, Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016
9. Shen, Guo, X., Wu, C., & Wu, D. (2011). Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm. Knowledge-Based Systems, 24(3), 378–385.https://doi.org/10.1016/j.knosys.2010.11.00110.
10. Tsai, Lin, Y.-C., Yen, D. C., & Chen, Y.-M. (2011). Predicting stock returns by classifier ensembles. Applied Soft Computing, 11(2), 2452–2459. https://doi.org/10.1016/j.asoc.2010.10.001