# The Random Forest Model for analyzing and Forecasting the US Stock Market under the background of smart finance

Jiajian Zheng[1,a,&*] , Duan Xin[2,b,&] , Qishuo Cheng[3,c], Miao Tian[4,d], Le Yang[5,e]

[1]Bachelor of Engineering, Guangdong University of Technology, ShenZhen, China
[2]Accounting, Sun Yat-Sen University, HongKong, China
[3]Department of Economics, University of Chicago, Chicago, IL, USA
[4]Master of Science in Computer Science, San Fransisco Bay University , Fremont CA, USA
[5]Master Science in Computer Information Science, Sam Houston State University, Huntsville，TX，USA

[a]im.jiajianzheng@gmail.com, [b]duanxin12314057@gmail.com, [c]qishuoc@uchicago.edu, [d]miao.hnlk@gmail.com, [e]wesleyyang96@gmail.com

**Abstract.** As an important part of the financial market, The stock market plays a crucial role in wealth accumulation for investors, financing costs for listed companies, and the stable development of the national macroeconomy. Consequently, significant fluctuations in the stock market will not only damage the interests of stock investors, but also cause the imbalance of the industrial structure, which will interfere with the development of the national economy on the macro level. As a result, the prediction of stock price trend has become a hot research topic in the academic circles. Therefore, the prediction of three movement trends of stock price trend, namely, rising, sideways and falling, is more helpful for stock investors to make choices among all decision-making behaviors, namely, buying, holding and selling stocks. Given this context, establishing an effective forecasting model for these three stock price trends is of substantial practical importance to establish an effective forecasting model for the prediction of the three movement trends of stock prices. In this paper, the stock price trend of the financial market under the background of smart finance is predicted by model, and the stock price trend of the United States is predicted by random forest model through the combination of artificial intelligence, deep learning and other fields. [1]Moreover, the test set of three stocks is used to test the prediction effect of the model under the optimal parameters of the random forest models combined with artificial intelligence. Based on the modeling and forecasting process, the corresponding time consumption is recorded. Therefore, the prediction performance of the model is evaluated comprehensively by using the prediction effect and time consuming of the model.

[&] *These authors contributed equally to this work and should be considered co-first authors.*


K. Subramaniyam et al. (eds.), *Proceedings of the 3rd International Academic Conference on Blockchain, Information Technology and Smart Finance (ICBIS 2024)*, Atlantis Highlights in Computer Sciences 21,
https://doi.org/10.2991/978-94-6463-419-8_11

**Keywords:** Prediction of stock price trend; Random forest; Artificial intelligence; Smart finance

## 1 Introduction

Stocks and stock markets have been around for hundreds of years. Predicting stock price movements has long been an investor's dream, but given that it is a dream, you can only imagine the difficulty of the task. In the past, when people analyzed listed companies, they mainly focused on the production and operation conditions, financial conditions, stock trading technical indicators, and even studied the psychology and behavior of investors. [2]But no matter what kind of analytical thinking, it is strongly dependent on the subjective experience of the analyst. In the face of the listed companies in many industries and the changing and complex stock market, it is always difficult for people to predict the stock price trend.

With the continuous progress of computer technology, artificial intelligence technology has achieved breakthroughs such as smart finance and began to develop rapidly. Among them, Machine learning is an important practice direction of artificial intelligence technology. Machine learning focuses on creating algorithms that can learn from data and constantly improve their accuracy. [3]Machine learning "trains" algorithms to discover patterns and feature laws in massive amounts of data in order to make decisions and predictions about new data. As the amount of data processed increases, the algorithm's decisions and predictions become more accurate. Machine learning has already had a full impact on human production and life. The Random Forest is a powerful and versatile machine learning algorithm. The algorithm can be implemented based on R language or Python language, and can be applied to classification problems and regression problems by forming many decision trees to establish random forests. The breakthrough of AI technology has opened up new ideas for people to analyze and study stock price trends. Applying machine learning technology such as random forest to stock price analysis to realize the auxiliary or even leading of algorithms for trading decisions is a direction worthy of our research.

## 2 Related work

Quantitative analysis and stock price prediction technology based on machine learning started earlier in foreign countries and developed rapidly. At present, although there have been a lot of research in [4]China, it is still in the exploration stage, and the application results are not extensive enough. Through the study of the existing application, it is found that most of the ways to forecast the rise and fall of stock price are to treat the rise and fall of stock price as a binary problem to carry out research analysis. In order to deepen the application of machine learning in stock price trend prediction, this paper analyzes the connotation of deep learning, random forest model and other related research and development.

### 2.1 Decision tree method

decision tree is a classification and regression method. This paper mainly discusses the decision tree used for classification. The structure of decision tree is a tree-like structure. In classification problems, decision tree represents the process of classifying data based on features, which can generally be regarded as a set of if-then rules. It can also be considered as a conditional probability distribution defined on the feature space and class space. The main advantages of the model are good readability and fast classification speed. [5]During the training, the decision tree model is built by using the training data according to the principle of minimization of loss function. The decision tree is used to classify the new data. The learning of decision tree usually includes three steps: feature selection, decision tree generation and decision tree pruning. The idea of these decision trees is mainly derived from the ID3 algorithm proposed by Quinlan in 1986 and the C4.5 algorithm proposed in 1993, and the CART algorithm proposed by Breiman et al., in 1984.

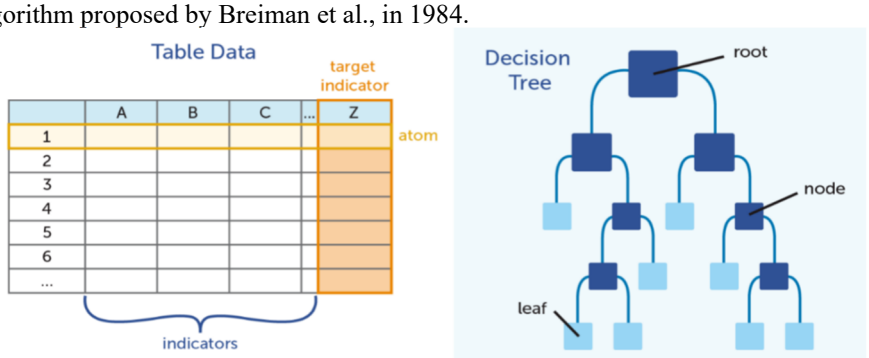


**Fig. 1.** Structure of decision tree model

The structure of a decision tree model, as depicted in Figure 1, is pivotal in the classification of data. Comprised of nodes and directed edges, it delineates features and attributes through internal nodes, while assigning classes via leaf nodes. Despite its intuitive construction, decision trees are susceptible to overfitting, even with pre-pruning measures, leading to suboptimal generalization. Consequently, their application in various contexts is often limited. [6]To mitigate this challenge, an integrated approach is frequently adopted, favoring ensemble methods over standalone decision trees. Notably, random forests emerge as a potent solution to the overfitting conundrum. By aggregating multiple decision trees, each with slight variances, random forests bolster predictive performance and enhance generalization capabilities.

### 2.2 Random forest

Random forest algorithm is a nonlinear model that integrates multiple decision trees into a forest. There are two key points to understanding random forests: random sampling and majority voting. First, for each decision tree, the training set is randomly selected from the whole sample set. In this paper, the standard of decision tree classification is the technical index in the feature matrix X, and the technical index classifi-

cation is used until the [7]Gini impurity is small enough to meet the requirements. These decision trees predict independently, and then vote on the outcome of each decision tree prediction, with the most votes becoming the prediction of the random forest. This avoids overfitting of a single decision tree. Due to random sampling, each decision tree does not use the full Sample (only about 2/3 of the sample is drawn), [8]and the samples that are not drawn are the non-sample set of this decision tree (Out of Bag Sample). For all out-of-pocket samples generated by decision trees, for each sample, the classification of the tree it is used as an oob sample (about 1/3 of the trees) is calculated, and then a simple majority vote is taken as the classification result of the sample, and finally the ratio of the number of mismarks to the total number of samples is taken as the OOB mismarks rate of the random forest. Therefore, the smaller the OOB deviation, the lower the proportion of misclassification and the more accurate the random forest classification.

To build a Random Forest model, first need to set the number of trees, specified by the `n_estimators` parameter, such as `n_estimators=10`. In Random Forest, each tree is constructed independently to ensure diversity. This is achieved through a process called bootstrap sampling.

Bootstrap sampling involves creating a new dataset for each tree. From the original dataset, which has 'n' samples, some data points are randomly selected with replacement. This means a single sample can be chosen multiple times, leading to a new dataset of the same size as the original, but with some data points repeated and others possibly omitted.For example, consider an original list of samples ['a', 'b', 'c', 'd', 'f']. An independent bootstrap sample could be[9] ['a', 'a', 'c', 'f', 'b'], and another might be ['d', 'a', 'f', 'c', 'd']. Each of these samples is used to build a separate decision tree, contributing to the ensemble of trees that form the Random Forest.

In the process of realizing the stock price trend prediction of random forest model, it is necessary to average the prediction probability of all trees, and then take the category with the greatest probability as the prediction result.A random forest of five or five trees will be applied to the two_moons data set:

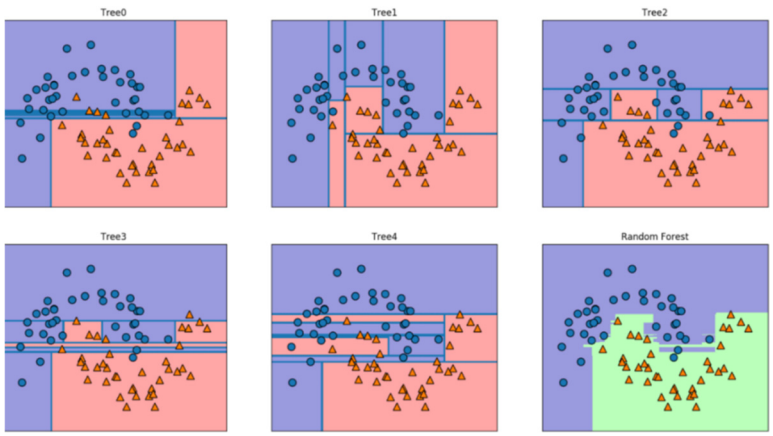


**Fig. 2.** Random forest model with five decision variables

Figure 2 showcases a random forest model incorporating five decision variables, illustrating distinct decision boundaries for each tree. It is evident that these boundaries vary significantly, reflecting the inherent diversity within the ensemble. Some misclassifications are apparent, attributed to the inclusion of training points not sampled by individual trees—a consequence of the deliberate subsampling inherent to random forests. Despite these discrepancies, the collective output yields more interpretable decision boundaries compared to single decision trees prone to overfitting. Notably, while this depiction involves a modest number of trees, real-world applications often deploy a multitude, numbering in the hundreds of thousands. This abundance ensures a more refined and smoother interface, enhancing the model's overall predictive prowess.

# 3 Methodology

## 3.1 Data collection and preprocessing

This US stock trend prediction model selected and collected the stock price trend data of about 7,000 trading days of Apple, Samsung and GE, and predicted the stock trend in 30 days, 60 days and 90 days, respectively. In order to remove the noise in the historical data and show the actual law of the historical data, the author adopts the exponential smoothing method to preprocess the stock price data:

$$S_0=Y_0$$

$$\text{For } t>0,\ S_t= \alpha *Y_t+(1-a)*S_{t-1} \tag{1}$$

Where alpha, ranging between 0 and 1 and typically closer to 1, assigns greater weight to recent data.Because recent movements are somewhat more consistent, greater weight is placed on the most recent data.

## 3.2 Data feature extraction

$$target\_i = Sign(close\_i+d - close\_i) \tag{2}$$

In the formula, d is the predicted time window and Sign is the symbolic function. When the value of $targe_{t-i}$ is 1, it means that the closing price after d is higher than the closing price today at the time of i, that is, the stock will rise after d; Vice versa. So target is also the target that the model needs to predict.

Technical indicators are important signals used to judge bear and bull in stock analysis. In this paper, six technical indicators are used as classification criteria, so that random forest model can learn these characteristics. Indicators are listed below(Table 1):

**Table 1.** Stock price trend forecast in six technical indicators as classification

| | | |
|---|---|---|
| RSI strength index | Stochastic Oscillator stochastic index | Williams %R William Index |
| MACD | Price Rate of Change. Price rate of change | On Balance Volume energy tide indicator |

### 3.3 Stock price trend forecast in six technical indicators as classification

Before establishing the model, the first conducted linear divisible tests on the two types of data, rise or fall, and found that the stock trend prediction problem was not linearly divisible (convex hull was found to have a large amount of overlap when projected into two-dimensional space), so all the algorithms related to linear discriminant analysis, such as SVM, were not applicable. Random forest, as a nonlinear algorithm, can avoid this situation and has important application significance in the following stock trend prediction research.

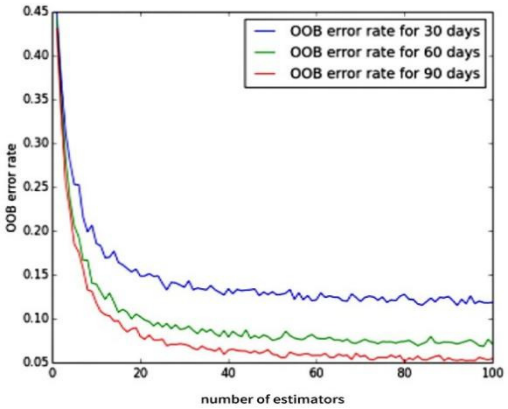


**Fig. 3.** Linear separable test results

Figure 3 presents the outcomes of a linear separable test, pivotal in establishing a predictive model using random forest methodology. Leveraging Apple's stock price data, the random forest model is deployed to forecast stock price trends over 30, 60, and 90-day intervals. Remarkably, as the number of decision trees within the model escalates, the model's accuracy demonstrates a notable uptick, eventually stabilizing as it converges. Notably, the efficacy of the model improves with the elongation of the forecast time window, signifying heightened predictive capabilities over longer horizons.

**Table 1.** OOB error Specific result

| Trading Period(Days) | No.of Trees | Sample Size | OOB |
|---|---|---|---|
| 30 | 5 | 6590 | 0.215372 |
| 30 | 25 | 6590 | 0.156348 |
| 30 | 45 | 6590 | 0.236276 |

| | | | |
|---|---|---|---|
| 60 | 65 | 6590 | 0.215372 |
| 60 | 5 | 6590 | 0.156348 |
| 60 | 25 | 6545 | 0.236276 |
| 90 | 45 | 6545 | 0.215372 |
| 90 | 65 | 6545 | 0.236276 |
| 90 | 5 | 6545 | 0.215372 |

According to Table 2, OOB errors gradually decrease and converge, and reach a steady state when the number of decision trees is greater than 45. The larger the forecast time window, the smaller the OOB error, but this reduction is marginal

### 3.4 Model comparison

To demonstrate the superiority of the Random Forest algorithm the study compares it with SVM, logistic regression, Gaussian discriminant analysis, quadratic discriminant analysis and other models. Based on a comparison of four monitoring algorithms used by Dai and Zhang in their 2013 paper, using historical 3M stock data from September 1, 2008 to August 11, 2013, it was found that compared with SVM, logistic regression, Gaussian discriminant analysis, and quadratic discriminant analysis models, the accuracy rate was 30%-80%. The random forest algorithm can achieve 80-99% accuracy, and the accuracy can be stable at about 98% when the time window of prediction is greater than 60 days.

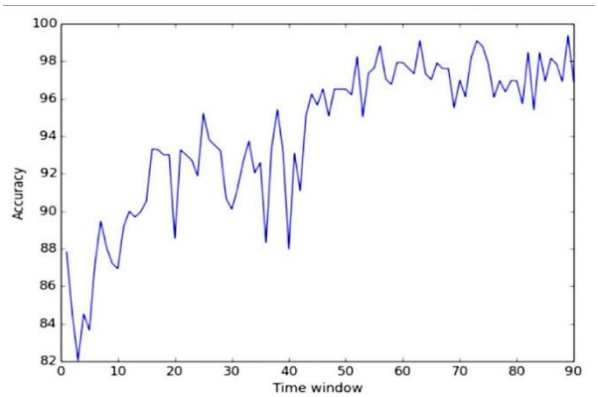


**Fig. 4.** Random forest algorithm results

Figure 4 encapsulates the outcomes of a comprehensive analysis on the integration of the random forest algorithm with artificial intelligence for stock price trend prediction in the realm of smart finance. The results underscore the remarkable efficacy of employing the random forest classifier, particularly in forecasting the long-term trajectories of stocks. Notably, across diverse datasets including those of Apple, Samsung, and General Electric, the predictive accuracy of the random forest model consistently registers between 85 to 95 percent—an impressive testament to its robust-

ness. Furthermore, as the number of decision trees within the random forest escalates, the model's performance tends towards stability, augmenting its reliability for strategic investment decision-making. This research presents a pivotal avenue for designing effective stock investment strategies, with the potential for adaptation to shorter prediction horizons through training the model with granular transaction data at hourly or even minute intervals.

## 4 Conclusion

This paper combines the methods of artificial intelligence, deep learning and other fields to predict the trend of the US stock market under the background of intelligent finance by using the random forest model. [10-12]Random forest is a powerful and flexible machine learning algorithm that provides efficient predictions of stock price trend changes (up, sideways, down) by integrating multiple decision trees. This paper firstly collects and preprocesses the stock market data, then extracts the key features, and uses the random forest model for classification and prediction. The experimental results show that the stochastic forest model has high accuracy and stability in predicting stock price trend, especially in predicting long-term trend.

In addition, this paper also compares the prediction effect of random forest with other common machine learning algorithms (such as SVM, logistic regression, etc.), and finds that random forest is superior in predicting stock price trend. These findings not only demonstrate the practical application potential of random forest in stock price prediction, but also provide new ideas for future financial market analysis and decision-making. As AI and machine learning technologies continue to advance, their application in the financial field will become more extensive and in-depth, providing investors and market analysts with more accurate and efficient tools. This will not only help improve investment strategies and decision-making processes, but could also revolutionize the entire financial industry.

## Acknowledgment

The research in this paper is largely informed by the paper "Integration and Performance Analysis of Artificial Intelligence and Computer "published by Liu Bo and his collaborators in 2023 Vision Based on Deep Learning Algorithms" (arXiv preprint arXiv:2312.12872). Their in-depth research on the application of artificial intelligence and deep learning algorithms to computer vision provides a valuable frame of reference for our analysis and methodology. We would like to thank Liu Bo and his team for their contributions, whose advanced research results have had an important impact on the writing and research direction of this paper.

link: Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

## Reference


1. Markowitz, H. M. Portfolio Selection [J]. The Journal of Finance, 1952, 7(1):77-91.
2. Franco Modigliani, Merton H. Miller. The Cost of Capital, Corporation Finance and the Theory of Investment[J]. The American Economic Review, Vol.48, No.3. (Jun..1958), pp.261-297.
3. Sharpe, W. F. Capital asset prices: a theory of market equilibrium under condition so f risk*[J]. The Journal of Finance, 1964,19(3):425-442.
4. Tianbo, Song, Hu Weijun, Cai Jiangfeng, Liu Weijia, Yuan Quan, and He Kun. "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition." In 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 834-837. IEEE, 2023.DOI: 10.1109/mce.2022.3206678
5. Che, C., Liu, B., Li, S., Huang, J., & Hu, H. (2023). Deep learning for precise robot position prediction in logistics. Journal of Theory and Practice of Engineering Science, 3(10), 36-41.
6. Liu, B., Zhao, X., Hu, H., Lin, Q., & Huang, J. (2023). Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. *Journal of Theory and Practice of Engineering Science*, *3*(12), 36–42.
7. Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." *arXiv preprint arXiv:2312.12872* (2023).
8. Li, Linxiao, et al. "Zero-resource knowledge-grounded dialogue generation." Advances in Neural Information Processing Systems 33 (2020): 8475-8485.
9. Liu, Yuxiang, et al. "Grasp and Inspection of Mechanical Parts based on Visual Image Recognition Technology." Journal of Theory and Practice of Engineering Science 3.12 (2023): 22-28.
10. Zong, Yanqi, et al. “Improvements and Challenges in StarCraft II Macro-Management A Study on the MSC Dataset”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 29-35, doi:10.53469/jtpes.2023.03(12).05.
11. Tian, Miao, et al. "The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier." Academic Journal of Science and Technology 8.2 (2023): 57-61.
12. Wan, Weixiang, et al. "Development and Evaluation of Intelligent Medical Decision Support Systems." Academic Journal of Science and Technology 8.2 (2023): 22-25.