Electronics, Communications and Networks
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA241350

Research on Prediction of ChiNext and Blue-Chip Stocks in the Chinese Stock Market Based on Random Forest Algorithm

Guoli XIE[a], Lei YANG[a,b,1], Liwei TIAN[a], Sijin LI[a], and Jiacheng LAI[a]
[a] *Guangdong University of Science and Technology, Dongguan, China*
[b] *Dongguan Key Laboratory of Industrial Software for Intelligent Manufacturing*
ORCiD ID: Lei Yang https://orcid.org/0000-0001-6526-809X

Abstract. This research focuses on the predictive analysis of ChiNext and blue-chip stocks in the Chinese stock market, employing the random forest model to conduct in-depth research on the stock trends of these two sectors. In the experimental process, we randomly selected six stocks from each sector as research samples, aiming to explore whether there are significant differences in prediction results between the two. Through rigorous experimental verification and data analysis, we found no significant difference in the predictive effect of random forest models between ChiNext and blue-chip stocks.

Keywords. ChiNext stocks, Blue-chip stocks, Random forest model, Data analysis

1. Introduction

In the current field of machine learning, random forest, as an outstanding representative of ensemble learning techniques, is attracting significant attention due to its superior predictive performance and generalization ability. The random forest model significantly improves prediction accuracy and stability by constructing multiple decision trees and integrating their prediction results. Its core idea lies in the use of bootstrap sampling and random feature selection, which not only ensures the independence and diversity of each decision tree but also reduces the variance of the model, thereby enhancing the overall predictive capability.

In the construction of the random forest, each decision tree is trained based on randomly sampled subsets of samples and features. This mechanism enables each decision tree to capture different information in the data and integrates their prediction results through the final majority voting (for classification problems) or average prediction (for regression problems) mechanism to form the final prediction.

This paper applies the random forest model to the predictive analysis of ChiNext and blue-chip stocks in the Chinese stock market. We carefully selected six representative stocks from each sector as research samples and designed comparative experiments to explore the differences in prediction results between the two [1].

[1] Corresponding Author: Lei Yang, yanglei@gdust.edu.cn

The experimental results show that after using the random forest model to predict and analyze the stock trends of these two sectors, there is no significant difference in the predictive effect between ChiNext and blue-chip stocks [2].

2. Dataset and method

2.1. Dataset

This research focuses on the Chinese stock market, selecting 2021 data as the research sample. Specifically, we randomly selected historical trading data for six blue-chip stocks and six ChiNext stocks. The blue-chip stocks include Sinopec (600028), Agricultural Bank of China (601288), CATL (300750), China Merchants Bank (600036), Industrial and Commercial Bank of China (601398), and China Life Insurance (601628). The ChiNext stocks include East Money (300059), Sungrow Power Supply (300274), Wens Foodstuffs (300498), Aier Eye Hospital Group (300015), Inovance Technology (300124), and AIMEIKE (300896).

2.2. Method

To deeply analyze the market performance of these stocks, we extracted key data features such as the opening date, opening price, daily high, daily low, closing price, and trading volume for each stock. These features not only reflect the fluctuations of stocks on trading days but also provide important clues for analyzing market trends [3].
In the data preprocessing step, we first standardize our data to ensure that the features follow a normal distribution with a mean of 0 and a variance of 1. Subsequently, we employ the random forest algorithm to predict stock closing prices. To build the prediction model, we divided the original dataset of each stock into a training set and a test set. Specifically, we used the first 220 samples as the training set to train the random forest model, while the remaining 23 samples served as the test set to evaluate the model's predictive performance [4].

During the construction of the random forest model, we set 200 trees as the base classifiers for ensemble learning. At the same time, to control the complexity of the model and avoid overfitting, we set a stopping condition for splitting when the number of samples in a subset is less than 2. After the model training was completed, we output the prediction results and visualized the trees in the random forest to better understand the model's decision-making process.

Through comparative validation, the experimental results show that the stocks of these two sectors exhibit similar market behavior and predictive response when using the random forest model for prediction.

3. Experiment and analysis

3.1. Experiment

We selected 200 trees as the number of base classifiers for the random forest. This choice aims to reduce the variance of the model by increasing the diversity of the base classifiers,

thereby improving the overall prediction accuracy. Additionally, we set an important splitting stop criterion that halts further splitting when the sample size of a node is less than 2. This strategy helps prevent model overfitting and enhances computational efficiency. To visually demonstrate the characteristics of this random forest model, we have visualized it. As shown in Figure 1, it represents the forest model for predicting the stock price of Sinopec, a selected blue-chip stock. The random forest models for the other eleven stocks are similar [5].

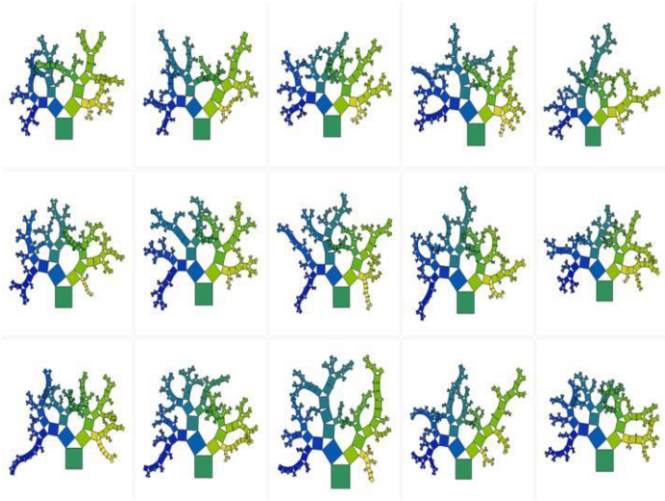

Figure 1. Random Forest Model for Sinopec.

3.2. Analysis

In this paper, we divided the historical trading data of 12 stocks into a training set and a test set, totaling 2,916 data entries. Each stock contributed 243 data points. To better compare the insignificant differences in prediction results between the two market segments, we used data from the first 11 months of 2021 for the training set and data from December 2021 for the test set. The training set consisted of 220 samples, while the test set comprised 23 samples. To conserve space, only three out of the selected six blue-chip stocks are presented, and the prediction results are shown in Table 1. The other three stocks exhibited similar outcomes.

As can be seen from the table 1, the deviations between the predicted values and the actual values for the three stocks are mostly within 1.5%. The comparison curves of the predicted values and the actual values for the six stocks are shown in Figure 2.

In Figure 2, the blue solid line represents the actual stock values, while the yellow solid line represents the predicted stock values. The horizontal axis represents the 23 test sets, and the vertical axis represents the value ranges of the actual and predicted values for each stock. From this line chart, it can be observed that after using the random forest model to predict and analyze the trends of blue-chip stocks, there is no significant difference between the predicted values and the actual values. Similarly, to conserve space, only three out of the selected six ChiNext stocks are presented, and the prediction results are shown in Table 2. The other three stocks exhibited similar outcomes.

Table 1. Blue-Chip Stock Predictions

Sample	Sinopec	Sinopec pred	Sinopec	ABC real	ABC pred	ABC devi	CATL real	CATL pred	CATL devi
1	4.13	4.1204	0.23%	2.93	2.92585	0.14%	672	675.36	0.50%
2	4.19	4.1838	0.15%	2.93	2.92895	0.04%	688	685.18	0.41%
3	4.2	4.1918	0.20%	2.93	2.9286	0.05%	687	685.11	0.27%
4	4.18	4.1891	0.22%	2.93	2.9301	0.00%	657.28	665.56	1.26%
5	4.23	4.2191	0.26%	2.94	2.94025	0.01%	629	631.43	0.39%
6	4.22	4.2225	0.06%	2.93	2.933	0.10%	639	638.20	0.12%
7	4.27	4.25875	0.26%	2.95	2.94535	0.16%	633.8	635.10	0.21%
8	4.22	4.22155	0.04%	2.94	2.9423	0.08%	639.95	639.84	0.02%
9	4.22	4.22285	0.07%	2.93	2.9337	0.13%	658.36	653.18	0.79%
10	4.16	4.168	0.19%	2.93	2.931	0.03%	651.01	651.12	0.02%
11	4.15	4.1565	0.16%	2.93	2.9315	0.05%	649	650.90	0.29%
12	4.22	4.20695	0.31%	2.94	2.93635	0.12%	655.06	653.61	0.22%
13	4.2	4.21035	0.25%	2.93	2.934	0.14%	651.19	649.03	0.33%
14	4.18	4.17785	0.05%	2.94	2.9397	0.01%	610.21	616.98	1.11%
15	4.21	4.2059	0.10%	2.94	2.9391	0.03%	604.07	604.66	0.10%
16	4.2	4.2083	0.20%	2.94	2.9436	0.12%	610.24	611.73	0.24%
17	4.25	4.24055	0.22%	2.94	2.93995	0.00%	622.07	622.28	0.03%
18	4.21	4.21595	0.14%	2.94	2.94	0.00%	576.8	583.76	1.21%
19	4.25	4.2444	0.13%	2.94	2.94	0.00%	578.6	582.32	0.64%
20	4.24	4.2362	0.09%	2.95	2.94875	0.04%	585	583.85	0.20%
21	4.24	4.2351	0.12%	2.94	2.9401	0.00%	575.55	579.36	0.66%
22	4.22	4.2235	0.08%	2.94	2.9394	0.02%	590	589.04	0.16%
23	4.23	4.2286	0.03%	2.94	2.93915	0.03%	588	589.34	0.23%

Figure 2. Comparison Curves of Predicted and Actual Values for Six Blue-Chip Stocks.

Table 2. ChiNext Stock Predictions

Sample	EMI real	EMI pred	EMI devi	SPS real	SPS pred	SPS devi	WFG real	WFG pred	WFG devi
1	34.74	34.7243	0.05%	160	160.079	0.05%	16.2	16.1174	0.51%
2	34.97	34.8935	0.22%	157.98	157.988	0.01%	16.77	16.7288	0.25%
3	34.87	34.7935	0.22%	152.99	153.111	0.08%	17.5	17.2448	1.46%
4	35.29	35.3511	0.17%	150.25	150.858	0.40%	17.18	17.2825	0.60%
5	35.45	35.4236	0.07%	148.7	148.961	0.18%	17.54	17.426	0.65%
6	36.13	36.0025	0.35%	150.2	149.65	0.37%	17.49	17.5142	0.14%
7	37.23	37.2318	0.00%	148.03	147.488	0.37%	17.45	17.427	0.13%
8	36.9	36.7908	0.30%	149.99	149.527	0.31%	17.27	17.3071	0.21%
9	38.13	38.2764	0.38%	148.09	148.925	0.56%	17.46	17.4669	0.04%
10	38.31	38.2675	0.11%	148.86	148.348	0.34%	17.69	17.623	0.38%
11	38.67	38.5999	0.18%	149.1	148.881	0.15%	17.59	17.6724	0.47%
12	38.78	38.5128	0.69%	152.63	151.646	0.64%	17.48	17.4487	0.18%
13	37.93	38.0863	0.41%	146.5	147.785	0.88%	17.75	17.6965	0.30%
14	37.57	37.6441	0.20%	137.6	138.038	0.32%	17.72	17.7695	0.28%
15	37.69	37.5682	0.32%	135.32	135.996	0.50%	18	17.9335	0.37%
16	37.35	37.4177	0.18%	137.4	136.762	0.46%	18.7	18.5703	0.69%
17	37.47	37.3442	0.34%	140.98	140.583	0.28%	19.36	19.2772	0.43%
18	37.31	37.3642	0.15%	133.21	135.465	1.69%	19.46	19.3949	0.33%
19	36.97	37.0359	0.18%	135.7	135.753	0.04%	19.4	19.3822	0.09%
20	37.66	37.5547	0.28%	138.85	138.261	0.42%	19.13	19.1966	0.35%
21	36.31	36.6188	0.85%	138.49	139.035	0.39%	19.05	19.1077	0.30%
22	37.04	36.9264	0.31%	140.97	140.553	0.30%	19.11	18.9232	0.98%
23	37.11	37.0488	0.16%	145.8	146.761	0.66%	19.26	19.2171	0.22%

As can be seen from Table 2, the deviations between the predicted values and the actual values for the three ChiNext stocks are also mostly within 1.5%. The comparison curves of the predicted values and the actual values for the six ChiNext stocks are shown in Figure 3.

In this research, we employed the random forest model to predict and analyze the ChiNext and blue-chip stocks in the Chinese stock market. By randomly selecting six stocks from each segment as samples, we found that these two market segments exhibited similar market behaviors and prediction responses when using the random forest model for prediction. Furthermore, there was no significant difference in the prediction

performance. This result indicates that the random forest model can provide effective predictions for both growth-oriented ChiNext stocks and stable blue-chip stocks, and the predicted trends of these stocks are highly similar. This finding provides important reference for investors to formulate investment strategies and demonstrates the effectiveness and reliability of the random forest model in handling complex financial data.

From the line chart, it can be observed that after using the random forest model to predict and analyze the trends of blue-chip stocks, there is no significant difference between the predicted values and the actual values.

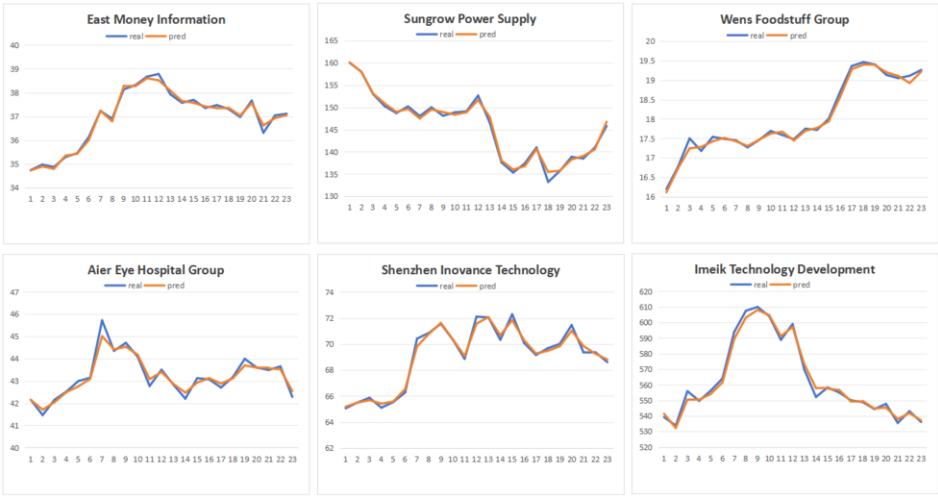

Figure 3. Comparison Curves of Predicted and Actual Values for Six ChiNext Stocks.

4. Conclusion

This research conducts a predictive analysis of ChiNext and blue-chip stocks in the Chinese stock market, leveraging the Random Forest algorithm. Through a meticulously designed experiment, we selected six stocks from each sector as research samples, aiming to explore whether there exist significant differences in the predictive outcomes generated by the Random Forest model when forecasting the trends of these two types of stocks. Following rigorous experimental validation and data analysis, we discovered that the Random Forest model demonstrates comparable predictive performance in forecasting both ChiNext and blue-chip stocks, revealing no notable discrepancy between the two. This conclusion not only underscores the extensive applicability of the Random Forest algorithm in stock market prediction but also affirms its stability and reliability in handling diverse types of stock data.

Acknowledgments

This research was funded by the following programs: Natural Science Project of Guangdong University of Science and Technology (GKY-2022KYZDK-12); Innovation and School Strengthening Project of Guangdong University of Science and Technology (GKY-2022CQTD-4, CQ2020062).

References

[1] Kumar B M, Manohar S, Ramona B, et al. Forecasting stock prices of fintech companies of India using random forest with high-frequency data[J]. Journal of Open Innovation: Technology, Market, and Complexity,2024,10(1).

[2] Kumar P I, Balakesavareddy P, Kumar A S. Stock price prediction methodology using random forest algorithm and support vector machine[J]. Materials Today: Proceedings,2022,56(P4).

[3] Zi R, Jun Y, Yicheng Y, et al. Stock price prediction based on optimized random forest model[C]//2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML). IEEE, 2022: 777-783.

[4] Sun Z, Tao W, Gao M, et al. Broiler health monitoring technology based on sound features and random forest[J]. Engineering Applications of Artificial Intelligence,2024,135108849-108849.

[5] Wei Z, Wang X, Lu L, et al. Construction of an Early Risk Prediction Model for Type 2 Diabetic Peripheral Neuropathy Based on Random Forest. [J].Computers, informatics, nursing : CIN,2024.