Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/30/20241086

Research on the selection of stock prediction models

Renjun Huang

School of Electronics and Information, South China University of Technology, Guangzhou, China

202130240404@mail.scut.edu.cn

Abstract. Against the backdrop of increasing attention to the integration of machine learning and stock analysis, stock prediction models are a hot topic. The question this paper is studying in this study is which stock prediction model is more accurate in predicting stocks. The method of this study is based on the stock prices of new energy vehicle leader Tesla Motors in the past three years as a data set, using a random forest model and an SVR model to predict the stock prices over the next 10 days. Based on the parameter MSE values of the training models of two stock prediction models, compare their sizes to determine the accuracy and stability of the models. This study found that the stock prediction results of the SVR model are more accurate and stable than those of the random forest model. Therefore, it is believed that the stock prediction model using the SVR method will have more market value and occupy an important position in the integration of machine learning and stock trading analysis.

Keyword: Machine learning, stock forecast, SVR, random forest.

1. Introduction

The data set for this study uses the stock price data set of Tesla Motors for the past three years. Why use Tesla car company, new energy vehicles are no doubt the trend of future development, whether in the world more and more attention to the perspective of clean energy or in the data in recent years of new energy vehicles accounted for more and more, in the new energy vehicles is the core direction of the future automobile development, and tesla automobile company as a leader of new energy vehicles, leading the global new energy automobile market direction, so tesla car company stock price will have great reference significance. It can not only reflect the stock market price of Tesla Motors Company itself but also reflect the core technology and development potential of new energy vehicles. So studying the stock price of Tesla Motors is a very meaningful process. This study has a very important impact on the integration of machine learning and the stock market. In the context of the Internet era, the stock market is also gradually becoming intelligent. Specifically, adding is to use machine learning methods to predict stocks, and people can adjust stock trading according to the forecast results. People can use machine learning to improve the success rate of stock trading and reduce the risk of stock failure. So choosing an accurate and stable stock model is a very important thing. This study will compare the prediction outcome parameters to determine which model is more accurate. Therefore, this study has a great reference significance. Among the many stock prediction models, each stock prediction model uses different methods. The methods used in the stock prediction model are svr, lstm, random forest, XGBoost, and lgbm. So which method of stock prediction model will be more accurate,

DOI: 10.54254/2753-8818/30/20241086

this is an inconclusive question, and also a question worth studying. This study will explore which stock prediction model is more accurate under the stock prediction model using random forest methods and SVR methods, respectively. This paper will use these two models to analyze the data for a stock over three years and predict the results for the next 10 days. This research will compare the model parameters of the data results to determine which stock model is more accurate. Because the model parameters of the data results can objectively reflect the accuracy and stability of the results, the model parameters can only be compared to obtain which stock prediction model is more accurate.

2. Literature review

Summer change based on abu quantitative system, build a specific trading environment, using sklearn linear regression module simulation analysis, relatively ideal prediction results, according to the prediction results from a certain extent demonstrated the feasibility of quantitative trading, also proved the machine learning algorithm in the actual investment operation advantages [1, 2].

Qian used the actual stock trading data to model the random forest algorithm and the gradient lifting tree (GBDT) algorithm to predict the returns of a single stock. The experimental results show that the interpretation of the random forest is high, but the accuracy of the GBDT model is slightly inferior. After comparison, the GBDT model has a little adjustment, and the model also achieves a better prediction effect [3].

Yang Yi mainly uses LSTM, support vector machine, and XGBoost model to predict liquor stock price. LSTM can mine time information and can well fit the forecast stock price. XGBoost is an emerging Boosting integrated algorithm model, efficient and accurate, support vector machine algorithm has a simple principle, and has good promotion ability [4].

In order to improve the stability and separation efficiency of the traditional Fast ICA algorithm, a new nonlinear function based on Tukey M estimation was constructed and the MTICA algorithm was proposed; On this basis, a new MTICA-AEO-SVR stock price prediction model was established by combining the SVR algorithm. Use the MTICA algorithm to decompose the original stock data into independent components for sorting and denoising and select different SVR models to predict each independent component and stock price separately. The artificial ecosystem optimization algorithm (AEO) was introduced into the SVR algorithm to select parameters, which improved the prediction accuracy of the model [5].

3. Methods

The dataset for this study was selected from the Kaggle. The selected data set is Tesla Motors' stock from 2019 to 2022. For the data of these three years, the data was cleaned and collated, and the specific data of close price in the next 10 days was predicted according to the data. Among the models, two models were used, the first a random forest model and the second a SVR model. Random forest has many advantages. It avoids overfitting, can process high-dimensional data, can get feature importance after training without making feature selection, and can effectively run on large datasets [6, 7]. The random forest model has good noise resistance and can handle over fitting with high data noise. The SVR model also has great advantages, it is a learning machine for finite sample cases, achieving structural risk minimization: seeking a compromise between the accuracy of the given data approximation and the complexity of the approximation function, in order to obtain the best generalization ability; and the SVR model ultimately solves a convex quadratic planning problem, theoretically, the global optimal solution, solving the unavoidable local extremal problem in the neural network method [8-10]. In this study, two models were used to predict the price of the next 10 days. This paper compared whether the data results of the next 10 days were reasonable and calculated the mse value of the two models. If the mse value is smaller, the stock prediction model is more accurate. Based on the model with a smaller mse, if the model is more accurate, the model will be more profitable.

DOI: 10.54254/2753-8818/30/20241086

4. Data analysis

4.1. Data cleaning

Table 1. The result graph of data cleaning

Missing Data Count	
Data	0
Open	0
High	0
Low	0
Close	0
Adj Close	0
Volume	0
dtype	int64

As shown in Table 1, this is the result table of data cleaning. The result shows that in the data of the 'TSLA. csv' file, the number of missing values in each column is 0, indicating that the data is complete and there are no missing values. Due to the absence of missing values in the original data, the cleaned data remains the same as the original data. The code prints out the first five lines of the cleaned data. These five lines of data include the opening price, highest price, lowest price, closing price, adjusted closing price, and trading volume of Tesla stock from January 2, 2019 to January 8, 2019. Overall, this code performs preliminary data cleaning and processing on a CSV file containing Tesla stock data, checks for missing values, and processes it if necessary. In this example, the data is complete without missing values, so no additional data cleaning steps were performed.

4.2. Training results

Table 2. Training Model Results

	SVR model training results	Random forest model training results
MSE	484.85	20601.78
R2	0.9777	0.0520
Best Parameters	C:982.17 gamma:0.0117	number of trees:80 minimum sample segmentation:2 minimum sample leaf count:2 maximum feature count: sqrt maximum depth:40 Bootstrapping: false

As shown in table 2, The data on the left is the training model results of the SVR model, with the optimal parameters of C being approximately 982.17, gamma being approximately 0.0117, and kernel being linear. The MSE on the test set is 484.85, and the R ^ 2 is 0.978. MSE measures the error size of model prediction, the smaller the better; R ^ 2 measures the model's ability to explain variation, ranging from 0 to 1, with closer proximity to 1 indicating better model performance. The data on the right shows the training model results of the random forest model. The optimal hyperparameter combination obtained through random search is: 80 trees, with a minimum sample segmentation of 2, a minimum sample leaf count of 2, a maximum feature count of 'sqrt', and a maximum depth of 40, without the use of bootstrapping. The MSE on the test set is 20601.78, and R ^ 2 is 0.052.

DOI: 10.54254/2753-8818/30/20241086

4.3. Comparison of the mse-values

Table 3. Comparison of the mse-values

	SVR model training results	Random forest model training results
MSE	484.85	20601.78

As shown in table 3, it can be found that the mse value of the SVR model is 484.85, and the mse value of the random forest model is 20601.78. The main model parameter examined in this article is the mse value, which refers to Mean Squared Error and is a widely used evaluation indicator in regression problems. It is used to measure the square of the average error between the predicted value and the actual value. The smaller the mean square error, the better the top shape effect.

4.4. Model Prediction

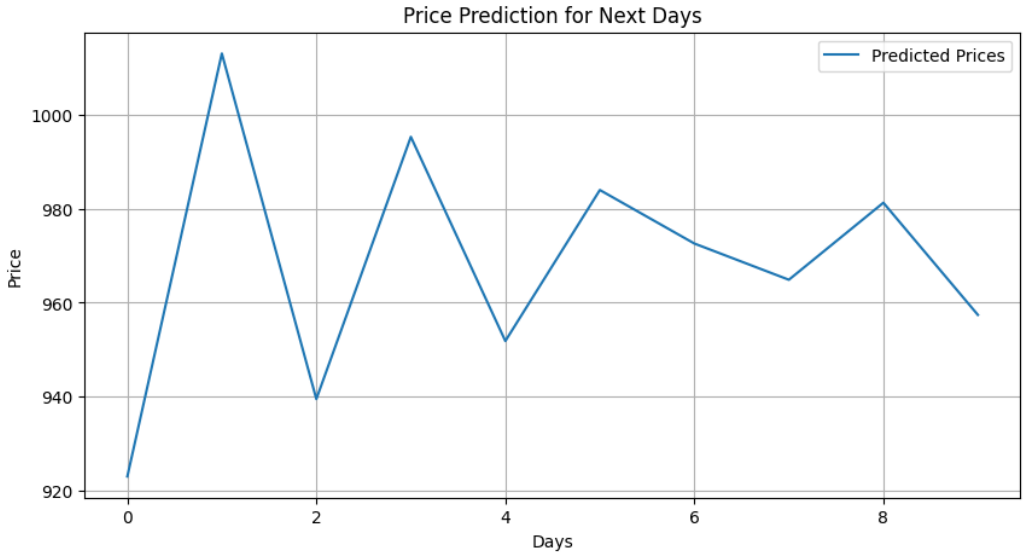

Figure 1. Line graph of prediction results of SVR model

As shown in Figure 1, this is a line graph of the predicted results of the SVR model. From the graph, it can be seen that in the predicted prices for the next 10 days, the prices vary from day to day. Among them, the second day was the highest value of the price, reaching a value of 1013.18. The first day is the lowest price, which is a value of 922.95. The fastest rising rate occurred from the first day to the second day, rising from the value of 922.95 to the value of 1013.18, an increase of 90.23. The fastest period of decline is from the second day to the third day. The value decreased from 1013.18 to 939.46, a decrease of 73.72. In the predicted results for these 10 days, there will be either an increase or a decrease compared to the previous day. But overall, it is showing a stable development trend. Among them, the price on the tenth day is higher than that on the first day, showing an upward trend.

DOI: 10.54254/2753-8818/30/20241086

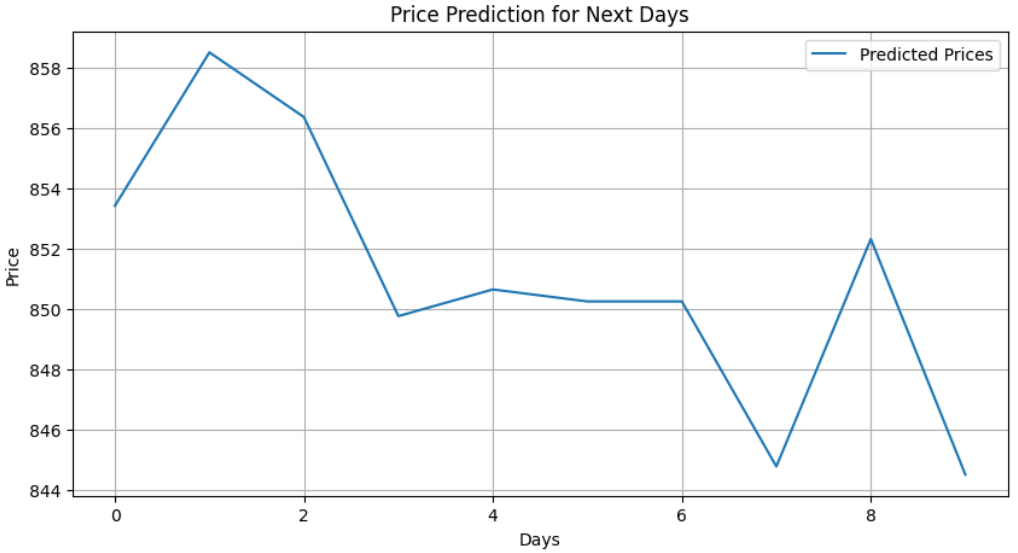

Figure 2. Line graph of the predicted results of the random forest model

As shown in Figure 2, this is a line graph of the predicted results of the random forest model. From the graph, it can be seen that in the predicted prices for the next 10 days, the prices vary from day to day. Among them, the second day was the highest value of the price, reaching a value of 858.52. The 10th day is the lowest value of the price, which is 844.51. The fastest rate of increase occurred from the 8th to the 9th day, rising from the price of 844.78 to the price of 852.32, an increase of 7.54. The period with the fastest decline rate is from the 9th to the 10th day. The price decreased from 852.32 to 844.51, a decrease of 7.81. In the predicted results for these 10 days, there will be either an increase or a decrease compared to the previous day. But overall, it is showing a stable development trend. Among them, the price on the tenth day is lower than the price on the first day, showing a downward trend.

5. Discussions

This study first checked the data set for missing values and processed them if necessary. In this data set, the data set is complete without missing values, so no additional data cleaning steps are required. In terms of prediction results, the two models predicted the specific stock prices over the next 10 days. It can see that the specific values predicted by the two models are different every day, indicating that the predicted results of different models are different. However, the numerical difference between these two models is not particularly significant overall, and further analysis and judgment are needed. In the training process of the SVR model, the optimal parameters are C approximately 982.17, gamma approximately 0.0117, and kernel linearly. This indicates that in the search parameter space, this set of parameters makes the SVR model perform best in cross validation. The MSE on the test set is 484.85, and the R ^ 2 is 0.978. MSE measures the error size of model prediction, the smaller the better; R ^ 2 measures the model's ability to explain variation, ranging from 0 to 1, with closer proximity to 1 indicating better model performance. In this example, a R ^ 2 value of 0.978 means that the model can explain 98% of data changes, which is a quite good result. During the training process of the random forest model, the optimal hyperparameter combination obtained through random search is: 80 trees, with a minimum sample segmentation of 2, a minimum sample leaf count of 2, a maximum feature count of 'sqrt', and a maximum depth of 40, without the use of bootstrapping. In terms of model performance, the MSE on the test set is 20601.78, and the R ^ 2 is 0.052. it can be seen from the above that the mse value of the SVR model is much smaller than that of the random forest model. The smaller the mse value, the higher the accuracy and stability of the model. It concluded that the SVR model is more accurate, and the SVR model has higher accuracy and stability compared to the random forest model.

DOI: 10.54254/2753-8818/30/20241086

6. Conclusion

This is a very valuable discovery to refer to. Among many stock prediction models, choosing an accurate stock prediction model is a very important thing. Many times, upgrading and renovating a model may result in significant improvements, but the upper limit of this improvement is determined by the selection of stock prediction model methods. So it is more important to choose a stock prediction model with high upper limit and strong stability. Therefore, this study found that the SVR model is superior. If future generations upgrade and transform its model more, it will be more efficient and have more potential to improve the accuracy of stock prediction results. This study has directional guidance significance for future generations to use machine learning to analyze and predict the stock market. This study promotes the integration of machine learning and the stock market, and will guide more scholars to join the competition of stock market trading based on machine learning. The results of this study will guide them to conduct more in-depth analysis and research on stock prediction models using SVR methods. In summary, machine learning is playing an increasingly important role in the stock market. It needs to strengthen the efficiency and functionality of machine learning, improve the accuracy and stability of stock prediction models, in order to be in an advantageous position in a highly competitive stock trading market and become the ultimate winner.

References

[1] Li, X. J., Xia, H. 2023, Research on Stock Price Regression Prediction Based on Machine Learning Algorithms. (Science and Technology Information, vol. 21), no. 14, pp. 227-231.

[2] Li, F. M. 2023, Research on Stock Prediction Based on Machine Learning Fusion Model. (Lanzhou University).

[3] Q, Q. M. Zhang, D. Wang, Y. Y. et al. 2022, The Application of Machine Learning in Stock Price Prediction. (China Market, vol. 21), pp. 7-10.

[4] Yang, Y. 2022, Research on Baijiu stock prediction based on machine learning algorithm model. (Shandong University).

[5] Deng, J. L., Zhao, F. Q. Wang, X. X.. 2022, MTICA-AEO-SVR stock price prediction model. (Computer Engineering and Applications, vol. 58), no. 8, pp. 257-263

[6] Cheng, H. 2023, Research on Stock Price Prediction Method Based on Deep Learning. (Shandong University of Business and Economics).

[7] Li, Q. 2023, Research on Stock Price Prediction Based on RCI and GC-GAN. (Jiangxi University of Finance and Economics).

[8] Zhang, D. 2023, Construction and Prediction Method of Stock Evaluation Index System Based on Artificial Intelligence. (Nanjing University of Information Engineering).

[9] Gao, X. H. 2023, The Application of Long and Short Term Memory Neural Networks in Stock Trend Prediction. (Harbin University of Technology).

[10] Pan, X. D. 2021, Research Group of New Era Securities Co., Ltd Research on the Development Law and Reference of Global Stock Markets. (China Securities Industry Association). Zhang, C. H. 2014, Development and Reform of the OTC Stock Market. (China Finance, vol. 7), pp. 43-46