EasyChair Preprint

№ 9856

Comprehensive Analysis and Research on the Mainstream Algorithm of Machine Learning in Stock Trend Prediction

Xiuyan Zheng, Jiajing Cai, Jiabin Wang, Shangyu Meng, Guangfu Zhang and Jinling Wei

March 9, 2023

Comprehensive analysis and research on the mainstream algorithm of machine learning in stock trend prediction

Xiuyan Zheng
College of Information Engineering
Hainan Vocational University of Science and Technology
Haikou, Hainan, 571126, China
2787157086@qq.com

Jiajing Cai
College of Information Engineering
Hainan Vocational University of Science and Technology
Haikou, Hainan, 571126, China
3473135707@qq.com

Jiabin Wang
College of Information Engineering
Hainan Vocational University of Scienceand Technology
Haikou, Hainan, 571126, China

Shangyu Meng
College of Information Engineering
Hainan Vocational University of Science and Technology
Haikou, Hainan, 571126, China

Guangfu Zhang
College of Information Engineering
Hainan Vocational University of Science and Technology
Haikou, Hainan, 571126, China

Jinling Wei
College of Information Engineering
Hainan Vocational University of Science and Technology
Haikou, Hainan, 571126, China

***Abstract*—Stock market is an important part of financial market and closely related to economic development. Various problems of stock price analysis and prediction have always existed with the establishment of financial market. Therefore, this paper uses the historical transaction data of Fenghua Hi-Tech As the research object to forecast and analyze the trend of its rise and fall. Through Support Vector Machine(SVM) model, LGBM model, Random Forest(FM) model to predict the stock trend. Through the empirical study, combined with the prediction chart of each model and the evaluation indexes of stock rise and fall prediction, such as RMSE, MAE, MAPE and R2, the prediction effect and prediction accuracy of the model were demonstrated. Finally, it was concluded that the model based on Random Forest had better prediction accuracy.**

Keywords: SVM model, LGBM model, Stock prediction, FM model

I. INTRODUCTION

With the gradual improvement of China's stock market and stock market system, stock investment has become one of the main investment channels for investors. The stock market has long been regarded as a barometer of the economic development of a country or region. Predicting the trend of a company's stock and grasping the rule of change of the stock market has always been a hot topic in our research[1]. Therefore, it is of great significance to accurately predict the impact of a company's stock trend and formulate investment strategies.

At present, there are a lot of research on stock prediction at home and abroad. Hu Di and Huang Wei[2] conducted empirical analysis on stock correlation based on SVM combination algorithm and clustering stock prediction algorithm, and verified that AP algorithm combined with other algorithms improved the accuracy of stock prediction. Zhang Jinghua and Gan Yujian[3] proposed that the deep learning support vector machine was used to optimize the

configuration of model parameters, and the model was used for simulation experiments. The results showed that the prediction accuracy of deep learning SVM was significantly improved compared with the existing SVM. Liu Daowen et al.[4] adopted the stock selection model based on support vector machine and determined the optimal regression parameters by cross verification method, and established a prediction model based on which the prediction effect of Shanghai Stock Exchange stock price index was ideal, but there was room for improvement in the selection of kernel function and optimal parameters. Because support vector machine is sensitive to missing data[5], it will greatly affect the output result and cannot meet the actual needs of the current stock prediction model. Therefore, many scholars suggest using model combination to improve the accuracy of prediction. Random forest algorithm is a model combination, which has achieved remarkable results in different fields [6].

Based on the advantages of random forest algorithm, the algorithm can be applied to the prediction of stock rise and fall, so as to avoid the shortcomings of the above prediction model. Therefore, this paper first uses the simple SVM model to predict the rise and fall of the stock, and by observing the comparison graph, it is found that the SVM model is extremely insensitive to the prediction of the stock trend, so the LGBM model with the integration algorithm is used, and the prediction result is significantly better than that of the SVM model. Finally can draw LGBM model has better fitting for stock prediction, but for the future stock forecast still unable to obtain accurately, as much as possible in order to remedy this defect, this article also uses the random forest model to predict stock movements, found that the minimum error, fitting degree is higher, its forecast effect is best, the highest prediction accuracy.

II. METHODOLOGY

A. SVM model

Support Vector Machine (SVM) is a new generation of machine learning technology, which can effectively process small sample data, nonlinear data and high-dimensional pattern recognition data. It solves the "dimension disaster" and "overlearning" and "under-learning" problems faced by traditional machine learning methods[7]. In addition, SVM has perfect mathematical model and clear geometric interpretation principle, with better generalization ability, so SVM can better predict the stock trend.

1. Data preprocessing

Because the actual data may contain a large number of missing values and noise data, the training of the algorithm model is very unfavorable. In order to obtain clean, standard and continuous data, this paper first performs data cleaning on dirty data. The quality of data preprocessing often determines the accuracy of subsequent data analysis and mining and modeling.

2. SVM nonlinear regression model

Nonlinear regression[8] uses mapping φ to map number set X to high-dimensional space H, and then carries out nonlinear regression in high-dimensional space. Its specific implementation is achieved through the kernel function, the detailed structure is as follows:

$$f(x) = w.\varphi(x) + b \qquad (1)$$

Where W and B represent model parameters.

In order to make Formula (1) approximate to the original function, relaxation variables ξ_i^* and ξ_i are introduced to minimize the risk function of the loss function and obtain:

$$Min\frac{1}{2}||w||^2 + c\sum_{i=1}^{1}(\xi_i + \xi_i^*)$$
$$s.t.\begin{cases} y_i - w.\varphi(x_i) - b \le \varepsilon + \xi_i \\ w.\varphi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i^* \ge 0, \xi_i \ge 0, i = 1,\ldots,1 \end{cases} \qquad (2)$$

In addition, we need to define the kernel function to be $K(X_i,X_j) = \varphi(X_{i,}).(X_j)$, where

A is function that maps X to higher dimensional space H.

Then, the Lagrange multiplier α_i and α_i^* are introduced, then the dual function of Equation (2) can be expressed as:

$$Max\begin{cases} W = \sum_{i=1}^{1} y_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^{1} \alpha + \alpha_i^* \\ -\frac{1}{2}\sum_{i=1}^{1}\sum_{j=1}^{1}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(X_i.X_j) \end{cases}$$

$$s.t.\begin{cases} \sum_{i=1}^{1} \alpha_i - \alpha_i^* - 0 \\ 0 \le \alpha_i \le C \\ 0 \le \alpha_i^* \le C, i = 1,\ldots,1 \end{cases} \quad (3)$$

The formula above can be converted into:

$W = \sum_{i=1}^{1}(\alpha_i - \alpha_i^*)\varphi(X_i)$,

Finally, the nonlinear regression function is:

$$f(X) = \sum_{i=1}^{1}(\alpha_i - \alpha_i^*)K(X_i, X_j) + b \quad (4)$$

B. LGBM model

LightGBM refers to a lightweight Gradient lifting Machine[9]. As a distributed Gradient lifting framework based on histogram decision tree algorithm, LGBM aims to reduce model calculation time. Its main design concepts are as follows:

First, reduce the dependence of data on memory, and use as much data as possible under the condition of ensuring speed.

Secondly, LightGBM is designed to reduce the cost of communication, improve the efficiency of multilevel parallelism, and achieve linear acceleration in computing. Therefore, it can be seen that LightGBM is designed to be a fast, efficient, highly accurate data science tool that supports parallel and large-scale data processing.

Based on LightGBM algorithm, this paper builds a prediction model of stock increase rate, and the specific steps are as follows:

Step1: divide training set and test set. 70% of the data were taken as the training model parameters of the training set, and 30% of the model's generalization ability was tested as the test set.

Step2: Use GridSearch algorithm to determine model parameters.

Step3: the LightGBM model calls the feathure_importances_ method in sklearn to make statistics on the splitting gain of each node in the decision tree splitting process, and obtain the important indicators of each feature.

Step4: Use LightGBM algorithm to predict the stock increase rate.

Step5: Model performance evaluation.

C. Random Forest model

Random forest is composed of several decision trees {t(Y,Xn),n=1,2... N} constitute the integrated classification model. Where, N represents the number of decision trees in the random forest. X_n is a random vector that is independent of the same distribution. The specific process is as follows:

Step1 randomness of samples

The primary data set is constructed by sampling the original data set with a put back method. The total amount of data in the primary data set can be the same as the total amount of the original data set. Elements of different primary data sets can be repeated, as can elements of the same primary data set.

Step2 randomness of features

When constructing a decision tree, the information gain (ID3) of all stock features is firstly calculated on a node of the tree, and then the feature with the maximum gain is selected as the trend of the next child node.

However, in the random forest, instead of calculating the gain of all stock features, y features are randomly selected from the total X feature vector, where y can be equal to SQRT (X), and then the gain of X features is calculated to select the optimal stock feature attribute.

Step3 build a decision tree

The sub-decision tree is constructed by using the sub-data set to obtain a classification or prediction decision tree. The operation of Step2 above is repeated to produce a single decision tree

with better effect.

Step4 random forest voting classification

After the above three steps are calculated, a decision tree can be obtained, and then the corresponding decision tree can be obtained according to the number of decision trees set artificially. Given a test sample, each decision tree can be used to classify it, and K classification results can be obtained. Then a simple voting mechanism can be used to obtain the final classification result of the test sample.

D. Evaluation index of the model

In order to better evaluate the prediction effect of each model, the mean absolute percentage error (MAPE), square mean error (RMSE), mean absolute error (MAE) and correlation coefficient (R^2) were used to measure the performance of the model before and after optimization.

MAPE is the Mean Absolute Percentage Error, which can be calculated as follows:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (4)$$

The value range of MAPE is $[0,+\infty)$. When MAPE is closer to 0, the model fitting is better; when MAPE is greater than 100%, the model fitting effect is poor.

RMSE is Root Mean Square Error. For example, RMSE=10, it can be considered that the regression effect differs by 10 on average from the true value. Its calculation formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (5)$$

The value range of RMSE is $[0,+\infty)$. When the predicted value and the real value are closer to 0, the model fitting effect is better; the larger the RMSE value is, the worse the model fitting effect is.

MAE is the mean absolute error, which can be calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (6)$$

The MAE value range is $[0,+\infty)$. When the predicted value and the true value are equal to 0, the model is a perfect model. When the error is larger, the MAE value is larger, and the MAE value is smaller, indicating that the prediction model has good accuracy.

R^2 stands for correlation coefficient, and its calculation formula is as follows:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (7)$$

The value range of R^2 is [0,1]. When the result is closer to 0, the simulation effect is poor; when the result is closer to 1, the simulation effect is very good.

III. Data and results analysis

A. Data

The data in this paper are mainly obtained through the Wind database terminal index data of Fenghua Hi-tech stock. The data from August 2020 to November 2021 are selected, and the data from August 2020 to August 2021 are used as the training set and the data from September 2021 to November 2021 are used as the test set to complete the fitting of the final model and evaluate the merits of the model.

B. SVM model prediction

The established SVM model was trained and tested from August 2020 to November 2021, in which the model training was conducted from August 2020 to August 2021, and the model test was conducted from September 2021 to November 2021. The actual and predicted values are shown in Figure 1. The red line represents the actual situation of the stock, and the blue line represents the forecast situation of the stock trend. It can be seen that the stock trend can be roughly predicted. The fluctuation trend of the predicted value is basically the same as that of the real value, but the deviation is large, so the prediction effect

is relatively not very good.

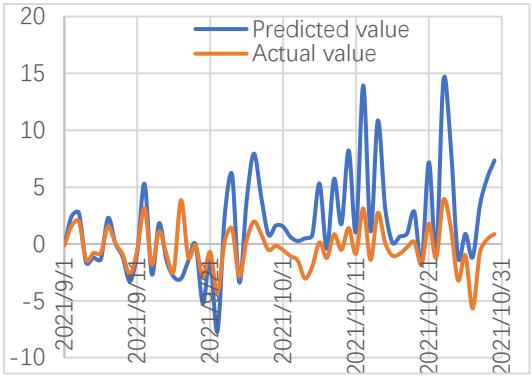

Fig. 1.FIG. 1 Prediction diagram of SVM model

C. LGBM model prediction

The established LGBM model was trained and tested from August 2020 to November 2021, in which model training was conducted from August 2020 to August 2021, and model testing was conducted from September 2021 to November 2021. The actual value and predicted value are shown in Figure 2. The red line represents the actual situation of the stock, and the blue line represents the forecast situation of the stock trend. It can be seen that the stock trend can be roughly predicted, and the prediction effect is relatively good, but the trend can not be accurately predicted at the inflection point.

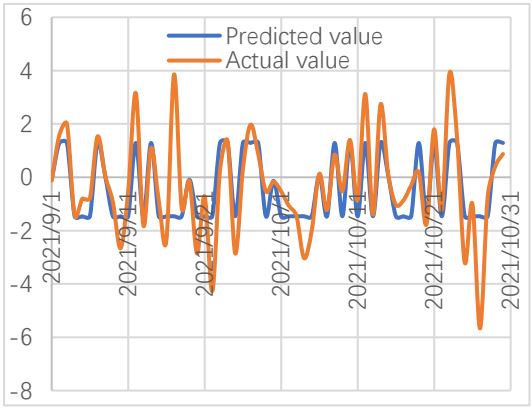

Fig. 2.FIG. 2 Prediction diagram of LGBM model

D. Random forest model prediction

The established random forest model was trained and tested from August 2020 to November 2021, in which model training was conducted from August 2020 to August 2021, and model testing was conducted from September 2021 to November 2021. The actual value and predicted value are shown in Figure 3. The red line represents the actual situation of the stock, and the blue line represents the forecast situation of the stock trend. It can be seen from the figure that the predicted value is very close to the real value, and the prediction result has high accuracy and the best prediction effect.

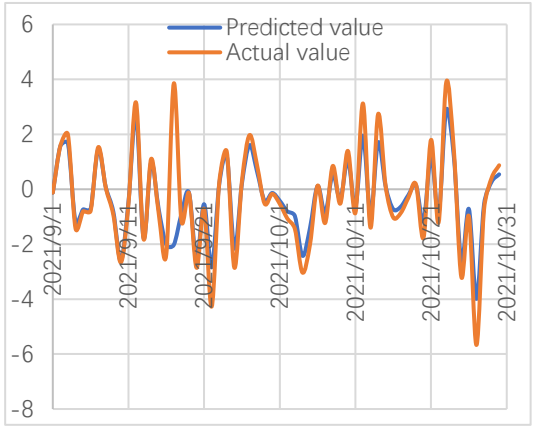

Fig. 3.FIG. 3 Prediction diagram of random forest model

E. Results analysis

FIG. 4 shows the prediction results of the increase rate of Fenghua High-tech By each model. The blue line represents the true value, the red line represents the predicted value of the support vector machine, the gray line represents the predicted value of the LGBM model, and the yellow line represents the predicted value of the random forest model. As can be seen from the figure, the deviation between the real value and the predicted value of the SVM model in the first 10 days is small and the fitting degree is high, while the deviation between the predicted value and the real value in the last 50 days is large, so the overall fitting degree is poor. The deviation between the real value and the predicted value of LGBM model is small, so the fitting degree is good on the whole. The random forest model has the smallest deviation between the real value and the predicted value, and the predicted value curve is closer to the real value curve, which has the highest fitting degree and the best prediction effect.

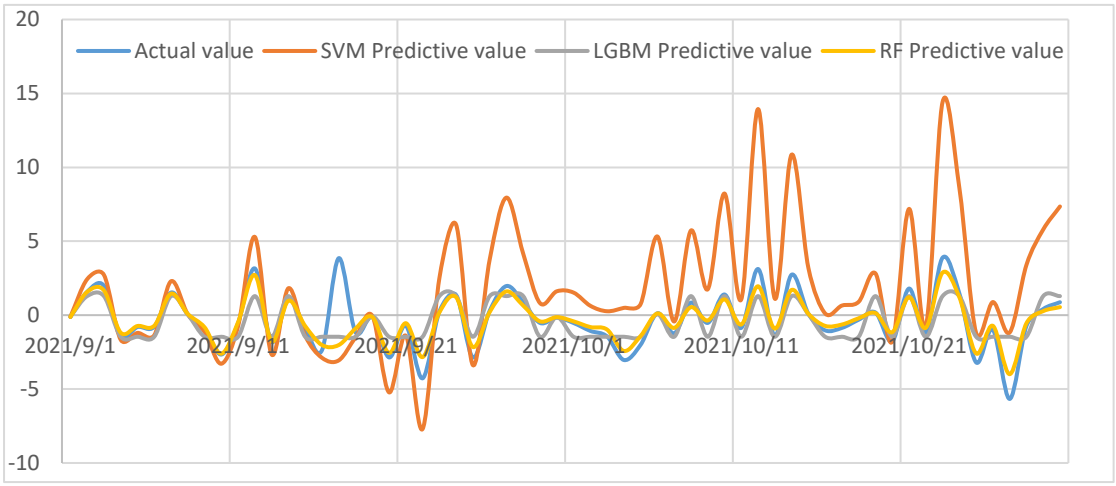

Fig. 4.FIG. 4 Model comparison

In order to verify the prediction effect of each model more intuitively and accurately and reflect the performance advantages of the model in this paper, Table 1 shows the evaluation results of each model.

Table 1 Evaluation results of each model

MODEL	RMSE	MAE	MAPE	R^2
RF model	0.89	0.41	23.6	0.76
LGBM model	1.21	0.83	81.7	0.53
SVM model	3.86	2.76	373	-2.96

Table 1 shows that the RMSE, MAE and MAPE values corresponding to the random forest model are 0.89, 0.41 and 23.6, respectively. Compared with other models, the random forest model has the lowest evaluation index and the highest accuracy. The R^2 of this model is 0.76, which is closer to and 1. The greater the goodness of fit, the better the fitting effect of the model is. The comprehensive prediction results show that the stochastic forest model proposed in this paper is more effective in processing time series data, and the prediction accuracy is higher, which has practical universal value.

IV. Conclusion

In order to improve the prediction effect of stock trend prediction model, the random forest model is proposed in this paper to reduce the influence of human factors and improve the prediction accuracy of the model. This paper also compares the prediction accuracy of SVM model, LGBM model, and the results show that the random forest model has a high prediction accuracy for stock trend prediction, which verifies the effectiveness of the model.

In the face of complex stock market, stochastic forest model can achieve more rapid and accurate prediction, reduce investment risk to a certain extent, and obtain the maximum return. This model can deal with time series problems efficiently and has transferability. It also has some practical reference value for other time series problems.

V. Reference

[1] Kang Ruixue, NIU Baoning, Li Xian, MIAO Yuxin. LSTM stock price prediction with self-attention mechanism based on multi-source data input [J]. Small microcomputer systems :1-9[2022-01-12].

[2] Hu Di, HUANG Wei. Stock Price Prediction Based on AP_SVM Combination Model [J]. Journal of Wuhan Institute of Technology, 2019 (6) : 297-301.

[3] Zhang Jinghua, GAN Yujian. Shanghai Stock Index Prediction based on deep learning support vector Machine [J]. Statistics and Decision, 2019, (2) : 178.

[4] Liu Daowen, FAN Mingzhi. Modeling and prediction of stock price index based on support vector machine [J]. Statistics and Decision, 2013(2): 76 -- 78.

[5] PANDEY A K, BISWAS M. Damage detection in-structures using changes in flexibility[J]. Journal of Sound and Vibration, 1994, 169(1): 3-17.

[6] SHI Z Y, LAW S S, ZHANG L M. Structural damage-detection from modal strain energy change[J]. Journal ofEngineering Mechanics, 2000, 126 (5) : 1216-1223.

[7] Ma Shengyu. Stock price prediction based on support vector machine (SVM) research [D]. Hebei university of technology, 2018. The DOI: 10.27105 /, dc nki. Ghbgu. 2018.000586.

[8] Zhang Xiangyu, Wang Fusen, Yang Lingjie, Liu Haifei. Research on Shanghai Stock Index Prediction based on Support Vector Machine [J]. Commercial Economics,2011(03):104-106.

[9] Yu X . Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data[J]. Agricultural Water Management, 2019, 225:105758.