The 2022 International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/6/2022138

Stock Price Prediction with Traditional Statistics and Machine Learning

Zheng Xu[1,a,*]

[1]Wuhan University, Wuhan, Hubei Province, China
a. 2020302011113@whu.edu.cn
**corresponding author*

Abstract: Stock forecasting is a very difficult and challenging task. There are so many subjective factors that are difficult to quantify and it is almost impossible to take all of them into consideration. The stock market changes so fast that forecasting models need to have forecast results in a relatively short period of time. For ordinary investors, it is necessary to predict relatively accurate results with less data. This paper compares the traditional statistical model with the machine learning forecasting models, linear regression, Backpropagation neural networks, time series ARIMA and LSTM. After predicting and comparing the four stocks with significantly different changes, it is found that the models with better prediction effects are different in different situations. In view of the fact that the neural network method requires a larger amount of data as a training set and needs to be retrained each time, so linear regression may be more suitable for ordinary investors.

Keywords: Stock price movement prediction, machine learning, neural network.

1. Introduction

The stock market is a platform for investors to trade stocks to earn money. Nowadays, there are more than 200 million investors in China investing in stock. Almost every investor in this large market is eager to earn as much money as possible. In this situation, a precise way to forecast stock price movement can efficiently help investors avoid losing significant amount of money and reap more benefits. Stock market is not only popular among normal investors, but also concerned by many financial researchers. With the development of computer science, people hope to make a highly accurate prediction with computer programs and big data. Quantitative investment plays an increasingly important role in financial trading markets. However, stock movement prediction is a complicated and challenging task. There are two mainly barriers if we try to establish a math model to forecast stock price movement: the stock market varies so quickly, resulting in the mismatch between training and test data and meaningless prediction due to too much time wasted on data collection and program calculations; The stock market is a noisy system which influenced by many uncontrollable and irrational factors.

Until now, a large number of scholars have put forward a lot of theories and methods about stock prediction. In 1937, John Maynard Keynes developed Keynesian beauty contest to explain price fluctuations in equity markets [1]. It describes that judges are more likely to select the most popular faces rather than those they may personally think the most attractive. Eugene Fama, a well-known American economist, first advanced the Efficient Markets Hypothesis (EMH) in 1970 [2]. It claims

DOI: 10.54254/2754-1169/6/2022138

that absent market manipulation, it is difficult for investors to make extra gains over the market average by looking at previous prices. Since the stock can be regarded as a series varying with time, many researchers hope to forecast the stock by establishing the time series Autoregressive Moving Average (ARMA) models, such as Stock Return Analysis Based on ARMA(2,2) Model [3], Stock Prices Prediction Based on ARMA Model [4] and Prediction of SSE 50 index based on ARMA model [5].

In recent years, with the development of technologies in machine learning fields, using neural network and SVM to predict stock price movement has become very popular. In fact, as early as 2003, a scholar Kim began to try to use technical analysis (TA) index as the predictor variable, proposed to support the use of Support Vector Machine (SVM) to classify the trend of the South Korea's KOSPI index (KOSPI), and compared the outcomes to those produced by neural networks or case-based reasoning (CBR) [6]. With the emergence of XGBoost [7], the investigation of the use of XGBoost to stock prediction has attracted more and more attention, such as Research on the Application of Machine Learning in Stock Index Futures Forecast—Comparison and analysis based on BP neural network, SVM and XGBoost [8] and Using Investor and News Sentiment in Tourism Stock Price Prediction based on XGBoost Model [9].

However, in all past forecasting models, only past data and indexes were taken into account. The subjective factors that may affect the stock are not considered, resulting in a low accuracy of the prediction model. Now, it`s very convenient for investors to share opinion or have a discuss with others via the web, which creates massive unstructured data. Therefore, in the last a few years, researchers have introduced sentiment analysis techniques into models. By collecting news from mainstream media and analyzing keywords on welcomed social platforms, these models can assess the public's sentiment toward stocks, and then have a more accurate prediction of the trend of stock prices. An extremely famous instance is that Jabeen Ayesha and other researchers improved stock prediction by combining the LSTM prediction model with coronavirus event sentiments [10].

After using sentiment analysis, the stock prediction has been greatly improved. However, it is almost impossible for normal investors to waste so much money on collecting and analyzing society sentiment and most of them have no access to these technologies. What`s more, they don`t need to forecast very accurately because they just need to know when to buy or sell. Thus, for them, what they really want is a model that requires less data from the past and takes less time to compute but can predict a rough trend in the next few days.

Therefore, a truly meaningful prediction for ordinary people should be both accurate and time-saving. Prediction should be made before the market change and the error should be as little as possible. After comparing the traditional statistical analysis methods and machine learning methods for stock prediction in each situation, the method with high accuracy and short time consumption will be considered as the optimal method.

2. Methodology

2.1. Selection of Indexes

In the first part, the research object is the closing price of Tongrentang(TRT) from May to October 2022. A total of 120 days of daily data are collected, with 115 days of data as the training set and the last 5 days of data as the test set. In order to more comprehensively find useful indexes that correlate with closing prices and make contribution to fit curves, 71 technical indexes and stock price indexes are collected for research.

After calculating the correlation coefficient matrix, we only select one of the indices with strong correlation. If the two indexes have a Pearson correlation coefficient that is larger than 0.8 in absolute terms, one of the indexes should be eliminated because they are highly connected.

DOI: 10.54254/2754-1169/6/2022138

Through this round of filter, 41 indexes are removed. A learning technique called random forest builds a lot of decision trees during training to do classification, regression, and other tasks. A number of weak classifiers are combined and averaged to get the final results, which have great generalization performance and accuracy. It is also highly stable. An important characteristic of random forest is that it has the capability to quantificationally evaluate the importance of individual feature variables. This can be applied in the situation that the model has many features and many of them have a lot of noise. It is very useful to evaluate the importance of each feature, sort these features, and then select the most important features from all of them.

The construction of the random forest

Step 1: Establish a few sub-dataset. Samples are randomly selected from the original data set by sampling with replacement (each time a sample is drawn and then it is returned to the original data set).

Step 2: Build decision trees from sub data sets. Assuming that a sub data set has X features. Y features are randomly chosen from these features when a decision tree node has to be split. Then select one of the Y of them as the splitting features of the node in some way. Repeat this step until it can no longer be split.

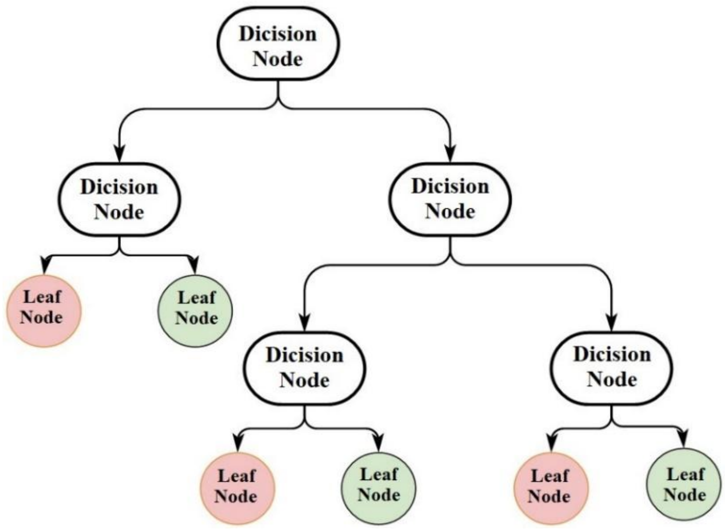

Figure 1: Schematic Diagram of the Decision Tree Model.

Step 3: Repetition of the first two phases will result in a vast number of decision trees, which together will form a random forest.

Feature selection with random forest:

Step 1: Rank the feature variables in the random forest by decreasing variable relevance. The importance to feature X can be represented by the out-of-bag error.

$$I_X \sim \Delta OOB.err = \sum \frac{OOB.err_2 - OOB.err_1}{n_{tree}} \tag{1}$$

Note that $OOB.err_1$ means out-of-bag data error for each class decision tree in random forest. $OOB.err_2$ means out-of-bag data error after adding noise interference. n_{tree} means there are n_{tree} trees in the random forest.

Step 2: If a feature's accuracy out of bag is significantly decreased when random noise is added to it, this indicates that the feature has a significant impact on the sample's outcome. Because of this, this method can be used to gauge how important a particular feature is.

Step 3: Select only a small number of features with high variable importance and adequate prediction of the dependent variables.

A random forest model can help to select the features of a dataset. The number of leaf nodes in the random forest is 5, and the number of decision trees in the forest is $n_{tree} = 300$. The training model of the entire random forest is established based on these two key parameters. In this random forest model, the input variable of training set is the index, and the output variable of the random forest model is the TRT closing price. 5 most important indexes are left due to their highest contribution. They are Williams Accumulation Distribution(WAD), Price Volume Trend(PVT), Bollinger Bands(Boll), Standard Deviation(STD) and CSI Smallcap 500 Index(CSI 500). All these 5 indexes are not highly linear correlated.

Figure 2: Heatmap of the New Correlation Coefficient Matrix.

2.2. Stock Price Prediction

2.2.1. Linear Regression

If the output is only regressed by simple linear regression, which is a linear approach for modelling the relationship between 5 indexes and the target variable, the curve fits very poor and the predicted results are very inaccurate.

$$\hat{y} = -264.6958 + 0.4455x_{WAD} + 1.4033 \times 10^{-6}x_{PVT} - 0.2198x_{Boll} - 1.5987x_{STD} - 0.0045x_{CSI500} \quad (2)$$

$\hat{\boldsymbol{y}}$: the prediction of the closing price.

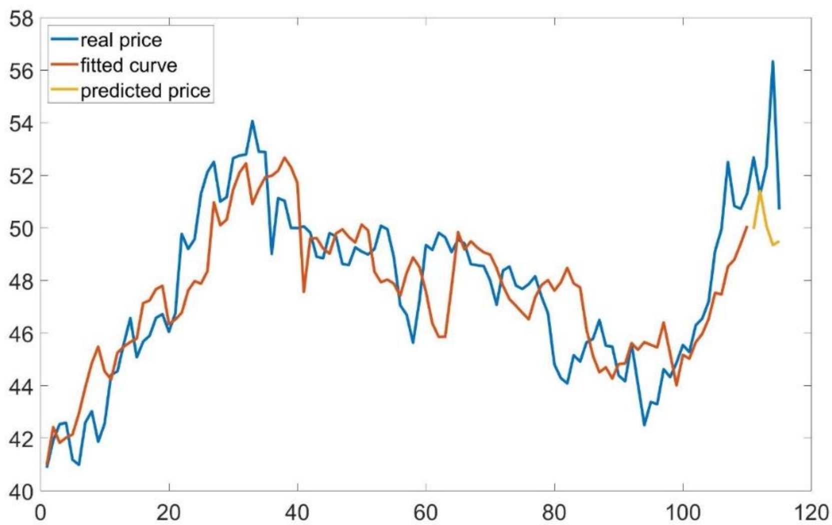

Figure 3: Graph of the Prediction Function of Linear Regression.

2.2.2. Backpropagation Neural Networks(BP)

Teuvo Kohonen has given the definition of artificial neural networks in Neural Networks [11]. Artificial neural networks are massively parallel interconnected networks of simple (usually adaptive) elements and their hierarchical organizations which are intended to interact with the objects of the real world in the same way as biological nervous systems do.

BP is a popular approach in machine learning for training feedforward neural networks. The multi-layer feedforward neural network consists of an input layer, several hidden layers, and an output layer. The multi-layer feedforward neural network can be viewed as a nonlinear composite function. Before delivering the back propagation algorithm (BP) to the output layer, its learning process moves layer by layer from the input to the output layer in the direction of the network structure. Updates are made to the weights and bias terms. Repetition of the entire procedure is required to get the loss function to stop decreasing or to reach its maximum number of rounds.

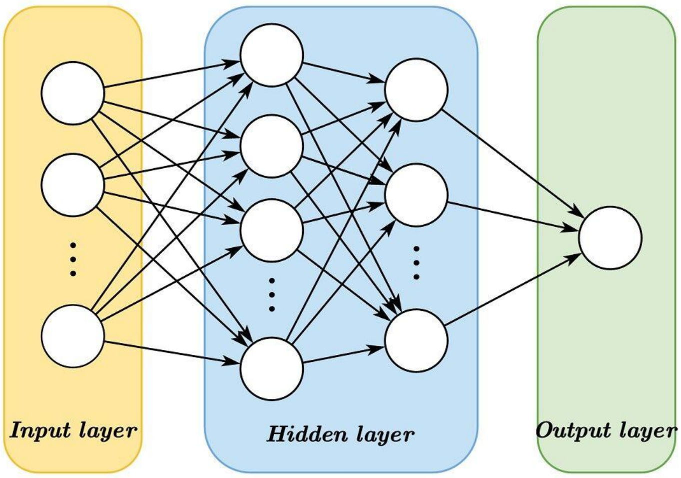

Figure 4: Schematic diagram of BP structure.

When BP is introduced into the regression process, the accuracy can be greatly improved. Through Figure 6, the forecast for the next two days is very accurate, but the forecast for the next three days is very biased. However, neural networks are random and make different predictions each time. Therefore, the prediction of this model is not very well.

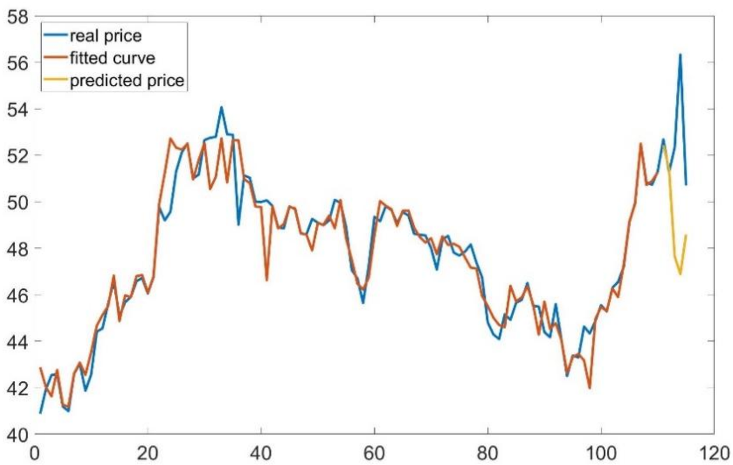

Figure 5: Graph of the Prediction Function of BP.

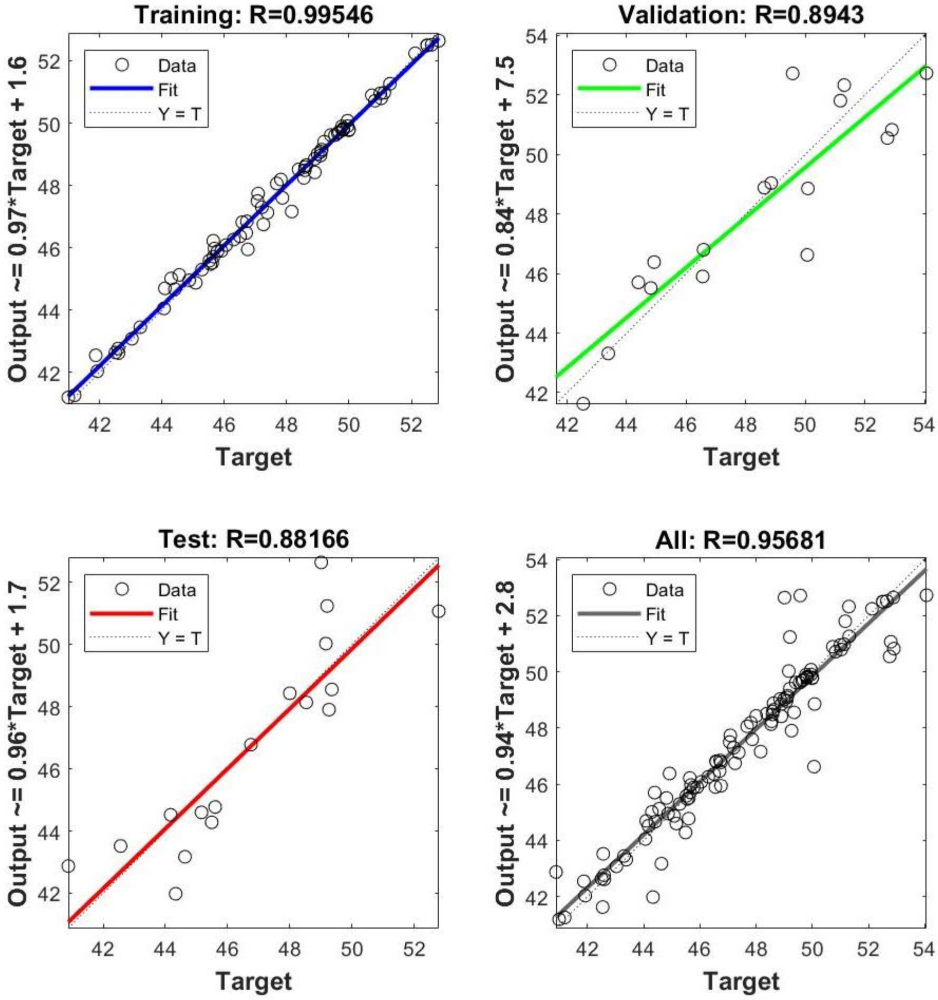

Figure 6: Diagram of Neural Network Analysis.

2.2.3. Autoregressive Integrated Moving Average(ARIMA)

The ARIMA model is a very well-liked and commonly applied statistical technique for time series forecasting. An ARIMA model is a generalization of the Autoregressive Moving Average(ARMA) model.

AR is a linear time series analysis model, which establishes a regression equation through the correlation (autocorrelation) between its current data and previous historical data. In the time series, the current observation value can be obtained through the historical observation value combined with the correlation, the formula is as follows:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \dots + \varphi_p Y_{t-p} + e_t \tag{3}$$

The moving average equation can be obtained by the weighted sum of the white noise sequences in the time series. In a time series, the current observation is a linear combination of historical errors, which are distributed independently of each other and obey a normal distribution. The formula is as follows:

$$Y_t = c + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q} \tag{4}$$

An ARIMA model means that the d-th difference of a time series $W_t = \nabla^d Y_t$ is a stationary ARMA process. Therefore, the ARIMA model can be expressed by the following formula:

$$W_t = \varphi_1 W_{t-1} + \varphi_2 W_{t-2} + \dots + \varphi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \dots - \theta_q e_{t-q} \tag{5}$$

DOI: 10.54254/2754-1169/6/2022138

2.2.4. Long Short-Term Memory(LSTM)

Recurrent neural networks (RNNs) have different subtypes, but the LSTM learns long-term dependencies. Hochreiter & Schmidhuber (1997) made the initial LSTM proposal [12]. Recently, researchers have enhanced and pushed LSTM. In many application problems like stock price movement prediction, LSTM has been considerably successful and widely used.

All RNN take the shape of a series of neural network repeating modules. Although LSTM shares similar chain structure with RNNs, its repeating unit is distinct. There is only one network layer, and inside it there are four other network layers. Figure 7 depicts the LSTM's internal structure.

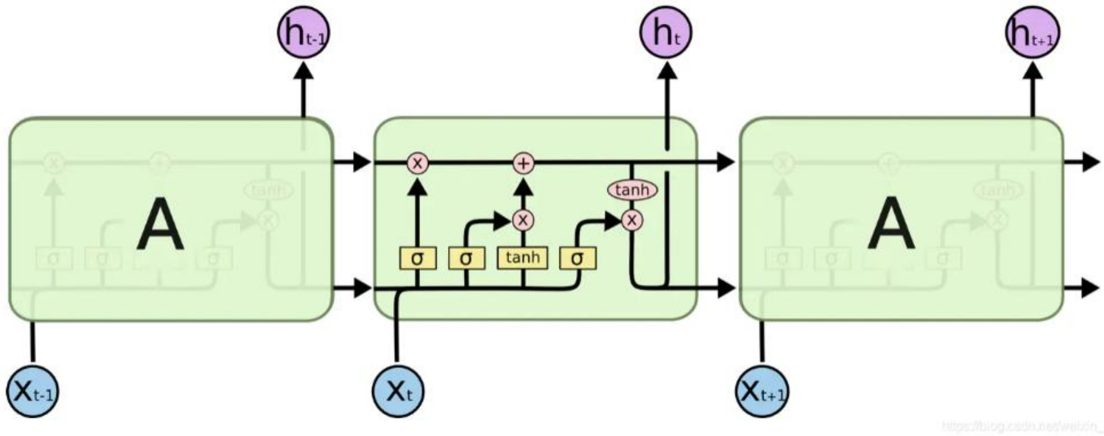

Figure 7: Schematic Diagram of LSTM Structure.
(https://blog.csdn.net/m0_37758063/article/details/117995469)

This model is trained on the data for the 115 days in the training set described above. Finally, the prediction made by ARIMA (8, 1, 8) and LSTM model and actual closing price of the stock market in the next 5 days are shown in the figure8. It can be seen that the forecast fluctuations made by the traditional time series model are very small and do not reflect the change law. Although the forecast results of the LSTM model do not completely overlap, the general trend of rising and falling on the last 5 days has been successfully predicted. Compared with the previous linear regression, BP and ARIMA, LSTM is the most accurate and successful prediction.

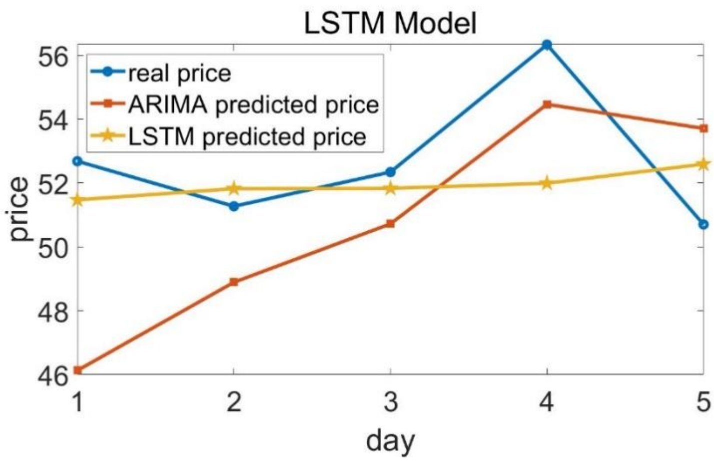

Figure 8: Graph of the Prediction Function of Time Series.

DOI: 10.54254/2754-1169/6/2022138

3. Results and Discussion

3.1. Larger Data

Due to the small amount of 120-day daily data, the results of the prediction model are not representative. In order to increase the amount of data, the 5-minute closing prices of TRT stock from January to October 2022 are collected. Using the above 4 forecasting models, make a training set of 9264 pieces of data in the first 193 days, and make predictions on 240 pieces of data in the next 5 days. In order to visually see the prediction effects of all models, the actual values and the prediction values of the four models are shown in the figure below.

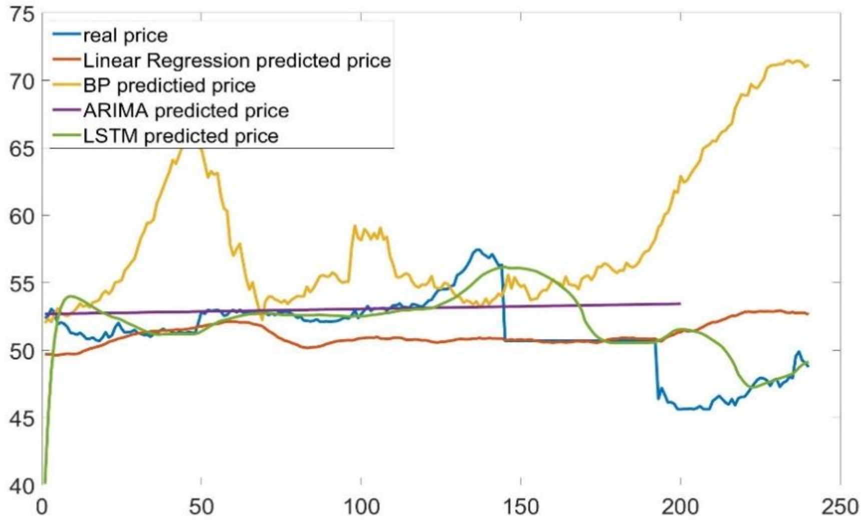

Figure 9: Graph of the Prediction Function with 5-Minute data.

3.2. Parallel Experiments

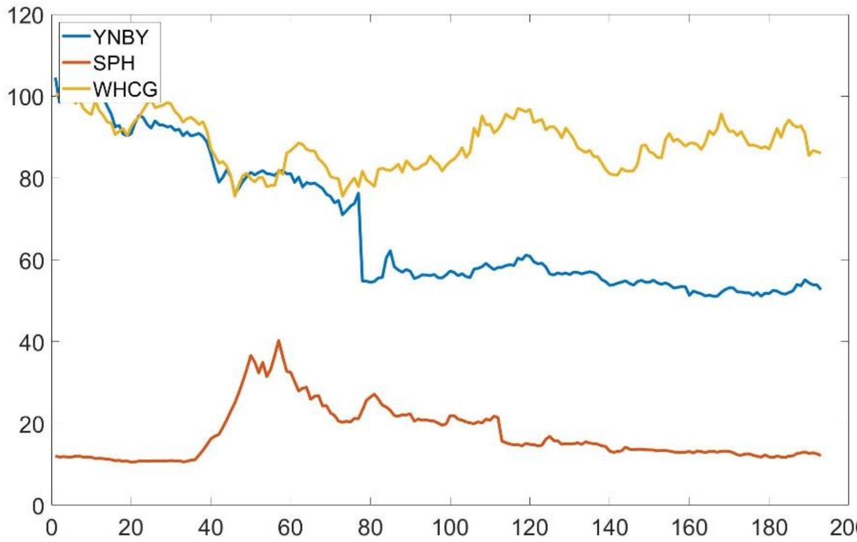

Figure 10: Graph of the Price Movement of YNBY, SPH and WHCG.

Some stocks fluctuate more, while others are relatively stable. For stocks with different characteristics, the prediction effect of the same model will also be different. Moreover, the cost of different forecasting models is different, and the cost of a more accurate forecasting model may be higher, which may not be the most suitable for ordinary investors. In order to test the generality of several forecasting models and find a model with lower cost and relatively better forecasting effect under different circumstances, this paper also made forecasts on several other stocks whose changes were completely different from TRT. The daily closing prices of Yunnan Baiyao(YNBY), Sinopharm(SPH) and Wanhua Chemical Group Corporation(WHCG) are very different. Compared with TRT, these three stocks are relatively stable and there have been no ups and downs. From figure 11, figure 12 and figure 13, we can see the forecasting effects of these three stocks under the four forecasting models. It can be seen that machine learning is not the best predictor for every stock. Among these three stocks, the mort accurate prediction model is the linear regression model. Therefore, the simplest traditional statistical model can make better predictions.

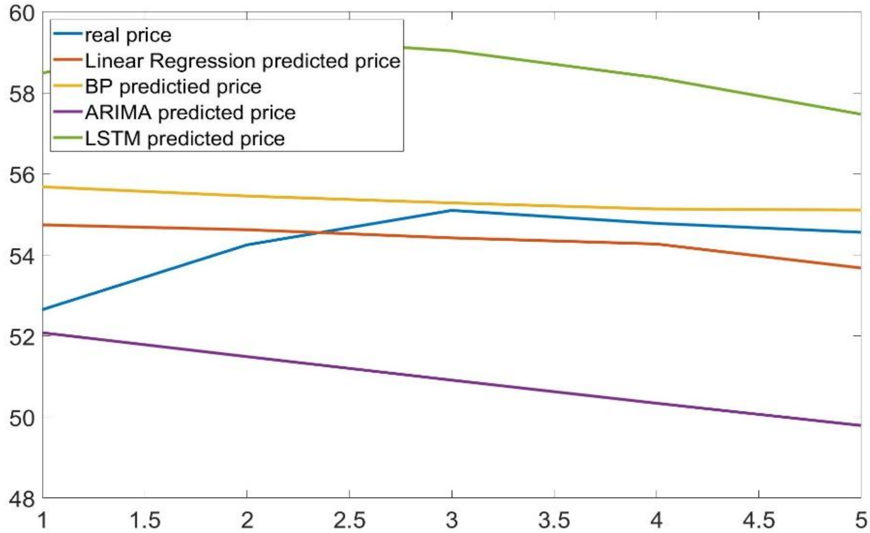

Figure 11: Graph of the Price Movement of YNBY.

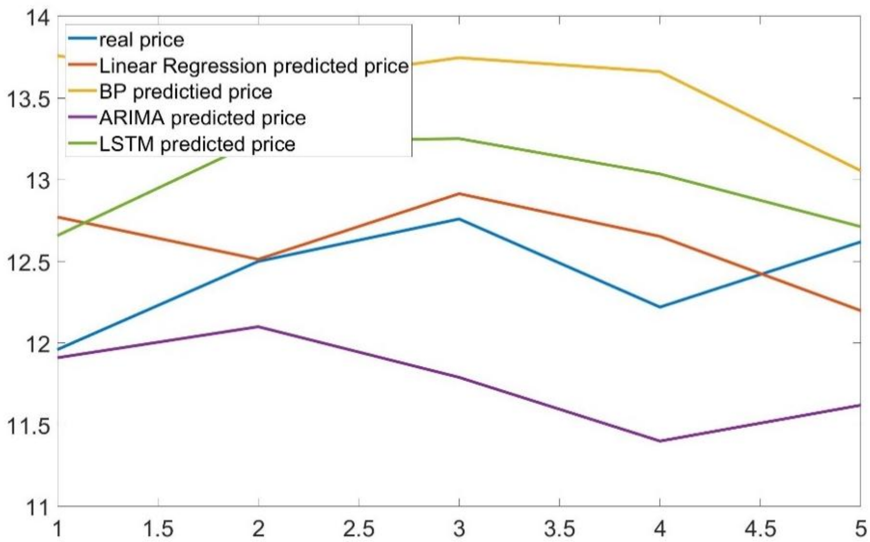

Figure 12: Graph of the Price Movement of SPH.

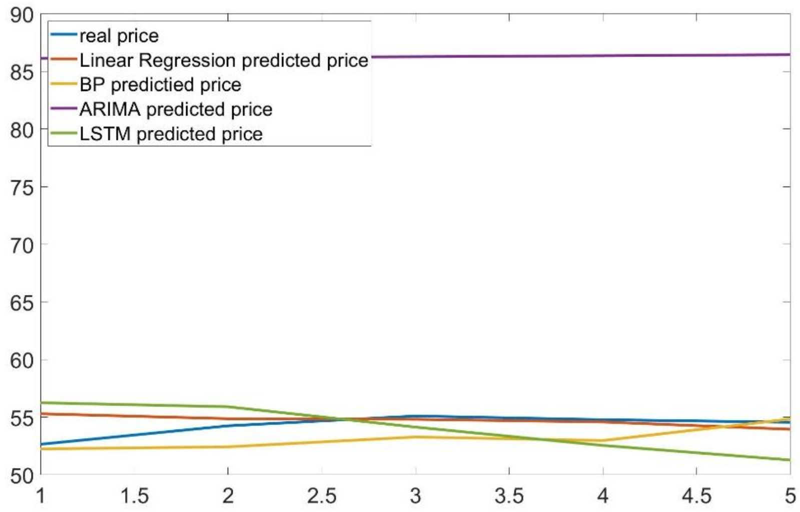

Figure 13: Graph of the Price Movement of WHCG.

4. Conclusion

Forecasting stock price movement has always been a very difficult challenge. The fundamental reason is that the stock market changes too fast and there are too many factors. It is impossible for people to consider all subjective factors that are difficult to quantify. Although the recently emerging sentiment analysis can quantify the sentiment effectors, it can make better predictions through a large number of keyword analysis, data analysis and neural network training. However, this paper believes that such a forecast is not accessible for ordinary investors. The cost of collecting data and neural network training is extremely high. It is difficult for ordinary investors to obtain so much data and build extremely complex neural networks in this way.

Emerging models do not necessarily mean better, while traditional models do not necessarily mean that they must be very bad. If a traditional model can achieve fairly good forecasting results at a very small cost, then it must be more suitable for ordinary investors. If a model can combine the respective advantages of traditional models and machine learning methods, then such a model must be more popular.

The first model in this paper combines the machine learning method random forest to find several more influential indexes. It predicts the data for the next 5 days through a simple linear regression method. Although the effect is very poor for TRT, the prediction effect of the 3 stocks in further research is very good. It can be seen from the figure 9 that since TRT plummets quite quickly in the last 5 days, which is a very special situation. Moreover, the other three models cannot predict such a sudden drop. However, LSTM can forecast a slow decline.

To sum up, for different stocks, there are also different models with the best predictive effect on predicting stock prices. When the price movement has a large fluctuation, the prediction effect of LSTM is slightly better than other models, but it is also not perfect. However, if its fluctuations are relatively small, the prediction effect of machine learning methods, such as BP and LSTM, is not ideal. The model of linear regression is better in this situation.

References

[1] *Keynes J M 1937 The general theory of employment. The Quarterly Journal of Economics 51(2) 209-223*
[2] *Fama E 1970 Efficient Capital Markets: A Review of Theory and Empirical Work. Journal of Finance 25(2) 383-417*
[3] *Yan H R 2022 Stock Return Analysis Based on ARMA(2,2) Model. Lecture Notes on Data Engineering and Communications Technologies 129 213-219*

[4] Tang H Z 2021 Stock Prices Prediction Based on ARMA Model. 2021 International Conference on Computer, Blockchain and Financial Development, CBFD 2021 201-204
[5] Sun Y Y 2022 Prediction of SSE 50 index based on ARMA model. The International Society for Optical Engineering, vol 12345
[6] Kim K 2003 Financial time series forecasting using support vector machines. Neurocomputing 55(1-2) 307-319
[7] Chen T, Guistrin C 2016 Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference On Knowledge Discovery and Data Mining. San Francisco, USA: ACM Press 785-7946
[8] Huang Q, Xie H L 2018 Research on the Application of Machine Learning in Stock Index Futures Forecast—Comparison and analysis based on BP neural network, SVM and XGBoost. Mathematics in Practice and Theory 48(08) 297-307
[9] Hu Y J, Shao L W, La L, Hua H L 2021 Using Investor and News Sentiment in Tourism Stock Price Prediction based on XGBoost Model. 2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science, BCD 2021, 20-24, September 13
[10] Ayesha J et al. 2021 An LSTM based forecasting for major stock sectors using COVID sentiment. Journal of Finance 67(1): 1191-1206
[11] Teuvo K 2003 An Introduction to Neural Computing. Neural Networks 1(1) 3-16
[12] Hochreiter S and Schmidhuber J 1997 Long Short-Term Memory. Neural Computation 9 1735-1780