Analysis of Stock Price and Price Movement Prediction based on Machine Learning Models for E-Hualu

Linyuan Han*

School of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China

*Corresponding author: 2020040095@mail.buct.edu.cn

Abstract. At present, many multi-information input models are used in various stock forecasts, but for most Chinese investors, it is difficult to obtain effective stock information, and most of them buy and sell stocks only based on stock trends and stock prices. Therefore, in order to get better prediction results based on basic information, the author tries to use some machine learning methods to predict E-Hualu in the Chinese stock market, and the aim is to select the best prediction model. In this study, five machine learning models were applied to predict stock prices, namely linear regression, SVR, random forest, XGBoost, and LSTM, then the close price of the previous 14 days were converted into the input of each model to estimate the close price of the next day. After the price prediction results were obtained, secondary processing was carried out, transforming previous outputs into trend prediction results. The research results indicate that the LSTM model has outstanding performance in both price prediction and trend prediction. Hopefully, the results of this research can provide some suggestions for investing in E-Hualu.

Keywords: Machine learning; stock price; price movement; E-Hualu.

1. Introduction

Compared with computers, brain capacity has a certain limiting effect on thinking. Human beings can only focus on 3-4 things at the same time at most, while the processing power of computers is thousands of times that of human beings. In addition to speed, computers will also outperform humans in other aspects of finance, having higher reliability, faster calculation speed, and better security.

In the financial industry, computers are also widely used, especially using machine learning models to solve practical financial problems, such as credit scoring, fraud detection, financial resource management, bank location selection, intelligent personalized service, and stock investment [1-3].

The stock market is a market in which issued shares are transferred, traded and distributed at current prices, including the OTC market and the exchange market. Furthermore, it is also called the secondary market because it is based on the issue market (the primary market). Generally, a stock market, also known as equity market, is a public market where traders exchange or deal in electronic or physical forms to buy and sell company shares and derivatives [4]. Equity securities are among the most traded securities because they offer attractive returns and are a relatively liquid asset since they can be resold and bought back through stock exchanges [3]. However, due to the uncertainty and volatility of the stock market, equity investment has a higher risk [5]. For effective price estimation, the size and category of data as well as the screening and processing of data are particularly important in equity investment. Therefore, investors need to obtain more comprehensive information, more accurate data analysis and appropriate investment models to realize profits, where machine learning models can be used.

In the stock market, the price of a stock is affected by many factors. The company's operating conditions and net profit affect stock price changes, and high profits with good development prospects often lead to stock price increases [6]. Tax news also affects the corporate stock price to some extent. The study found that legality is a significant factor in the impact of tax information on company stock prices, so tax evasion often has an adverse impact on stock prices, while tax avoidance has relatively no impact [7]. Moreover, the rises in Gross Domestic Product and the falls in the Consumer Price Index also have a significant impact on stock prices [8]. Additionally, the outbreak of global pandemic and the associated lockdown measures exert influence on both the liquidity of the stock market and the movement of stock prices. Researchers found that stocks in transport, mining, power, environment industries and the heating enterprises were negatively affected by the pandemic, while stocks in manufacturing, information technology industries, healthcare sectors and education-related fields have been pandemic-resistant [9]. At last, the historical price of the stock itself also has a certain effect on the future stock price change [10]. However, for most investors, information is always limited and difficult to obtain. The author believes that even with limited information, people can still use better machine learning models to make reasonable predictions on stock prices.

To make the best stock price prediction, many machine learning models and feature engineering methods are used and compared, namely linear regression, logistic regression, SVM, random forest, boosting, and other models rely on neural networks [3]. Although the price and trend of the stock are difficult to be precisely predicted, many studies have shown that LSTM is the best model to operate time series data for stock price prediction, having high forecasting accuracy under the appropriate number of neural network layers and parameters [11].

This study tries to use machine learning models to predict the stock price of Beijing E-Hualu Information Technology Co., Ltd., evaluates the prediction results of different models on this stock through error analysis, and selects the most suitable prediction model for Beijing E-Hualu Information Technology Co., Ltd., and compared with the results of other studies. The machine learning methods used in this research include linear regression, random forest regressor, SVR, extreme gradient boosting and LSTM. The study uses the close price of the past 14 days to forecast the close price of the following day, measuring and analyzing the error of the results by using the mean absolute error, mean square error, root mean square error and r-squared. Afterwards, the predicted price is converted into a forecast of the stock trend, and its total accuracy is calculated, in order to double check the result.

Beijing E-Hualu Information Technology Co., Ltd. was listed on May 5, 2011. It belongs to the A-shares of the Growth Enterprise Market of Shenzhen Stock Exchange, and the symbol of the stock is 300212. In addition, according to the "Special Regulations on Transactions on the ChiNext Board of Shenzhen Stock Exchange", the auction trading of ChiNext stocks is subject to price rise and fall restrictions, and the price limit ratio is 20%. The stock data of Beijing E-Hualu Information Technology Co., Ltd. is obtained from Yahoo Finance (finance.yahoo.com).

2. Data Analysis

2.1 Data

The stock data of Beijing E-Hualu Information Technology Co., Ltd. was downloaded from the official website of Yahoo Finance, and the stock data from January 2012 to December 2022 were used in the research. The data set contains a total of 2658 rows and 5 columns. The index is the date of each trading day, and each row includes the trading date, open price, high price, low price, close price, and total volume of the corresponding trading day. Regarding the trading volume, during the ten years, the maximum volume in a single day is 86.6 million, with the minimum volume is 0, and the average volume is 10.3 million. As for the closing price, the maximum value is 51.96 yuan, the minimum value is 3.07 yuan, and the average value is 20.40 yuan.

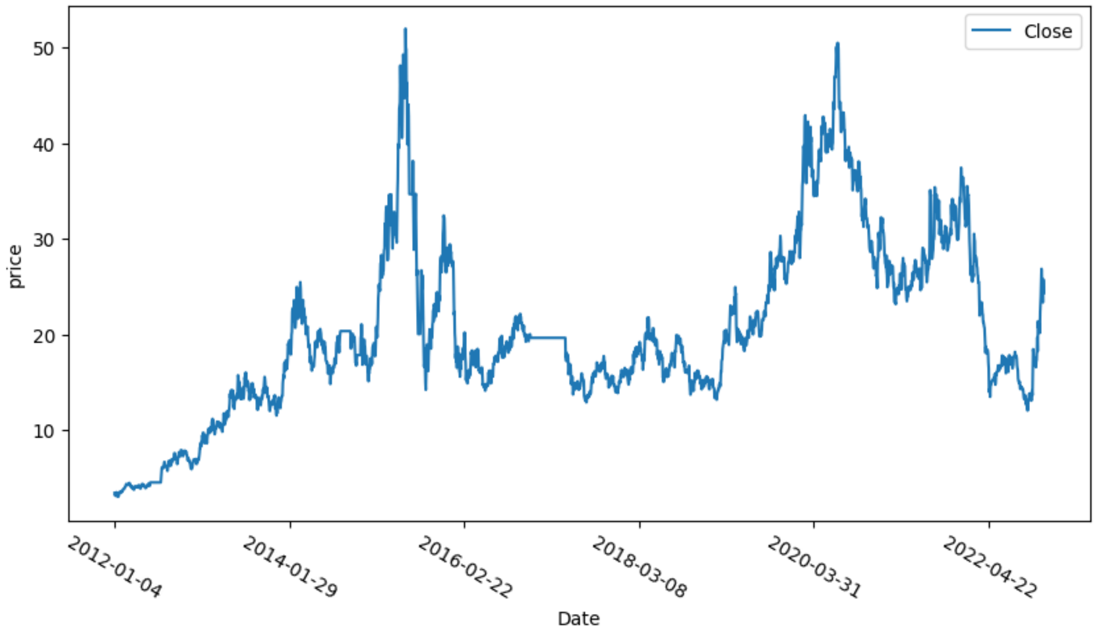

Figure 1. Close price of E-Hualu

Figure 1 shows the closing prices of Beijing E-Hualu Information Technology Co., Ltd. over the past ten years since its listing, which has undergone considerable changes. After a period of stable growth, its stock price rose from 15 yuan to the highest point of 50 yuan in a substantial increase in 2015. From 2016 to 2018, its stock price fluctuated between 15 yuan and 35 yuan; From 2019 to the middle of 2020, the stock price of E-Hualu continued to rise from 10 yuan to 50 yuan; in 2021, even if it fell back, its price remained stable at 20 yuan to 40 yuan; Nonetheless the stock price of Hualu continued to fall from 35 yuan to less than 15 yuan, and failed to break through 20 yuan until November 2022, but at the end of November 2022, its stock price continued to rise and broke through 25 yuan.

2.2 Feature Engineering

In the research, the index and closing price in the original data are selected as the experimental data. First, divide the whole data into a train set and a test set, 80% of which are training sets and 20% are test sets. Since it is time series data, the order of dividing the data sets is not disturbed. Then, feature engineering is performed on the data, and the closing price of the 14 days before each date is used as the input corresponding to the close price of the current day, and the close price of the current day is used as the target. Finally, the close price of the previous 14 days is obtained as the input X of each machine learning model, and the close price of the current day is used as the target value Y of each machine learning model. When using models other than LSTM for prediction, certain adjustments must be made to the input, which needs to be dimensionally reduced to meet the requirements of the model. Before entering the prediction model, all data needs to be normalized by MinMaxScaler in the sklearn library to convert the data between 0 and 1, and then convert the result back to the real size after the prediction is completed.

2.3 Methodology

2.3.1 Linear Regression

The linear regression method chooses the linear regression model in the sklearn library for prediction without adjusting parameters.

2.3.2 SVR

Support Vector Machine is an algorithm normally used for classification, but it can also support regression, and different models can be implemented under different parameter adjustment. The SVR method also uses the SVR model in the sklearn library for prediction. The predictive accuracy of the SVR model is very dependent on the selection and adjustment of parameters [11]. After testing and screening, the parameters kernel, C, and gamma of the SVR model are rbf, 1e3, and 0.14, respectively. The SVR model is based on the SVM algorithm for regression.

2.3.3 Random Forest

The model, known as random forest regression, constructs multiple independent decision trees by randomly sampling samples and characteristics, and provides parallel predictive results. Each decision tree can obtain a predicted result through the selected samples and features, and obtain the entire forest regression prediction result by integrating the outputs from all trees and taking the mean value. In the study, the random forest regression model in sklearn was used, and the random_state was set to 43 to ensure that the results were repeatable. When the best results were obtained, the parameter n_estimators were 200, meaning that the number of decision trees is 200. In addition, the parameter criterion should be set to squared_error, otherwise an error will be reported.

2.3.4 XGBoost

The XGBoost model is referred to as eXtreme Gradient Boosting. It retains the practice of Gradient Boosting, where each tree is related to each other, and the aim is that the mistakes made by the previous tree can be corrected by the tree generated later. The XGBRegressor used is included in the xgboost library, and the parameters n_estimators, max_depth, gamma, and learning_rate is 400, 8, 0.01, and 0.05, respectively. In addition, to ensure repeatable results, random_state is set to a fixed integer of 43.

2.3.5 LSTM

LSTM is a distinctive kind of RNN that is capable of learning long-term dependencies [12]. When designing the LSTM model, the author first considered the loss and validation loss changes during model training, but good loss changes during training did not yield good test results. Therefore, in order to obtain better prediction results, the author abandoned the requirement for historical loss and validation loss, and then designed the LSTM model in Table 1. The batch size and epochs in LSTM1 are set to 64 and 50 respectively. The batch size and epochs in LSTM2 are set to 32 and 50 respectively.

Table 1. LSTM model

Layer	Output shape	Param
Input layer	(None, 14, 1)	0
LSTM	(None, 14, 50)	10400
LSTM	(None, 14, 100)	60400
LSTM	(None, 14, 200)	240800
LSTM	(None, 300)	601200
Dropout	(None, 300)	0
Dense	(None, 100)	30100
Dense	(None, 1)	101

2.3.6 Analysis Methods

For the prediction results of each model, four kinds of errors are used for evaluation, namely MSE, RMSE, MAE, and R-Squared. The mathematical formulas of the four errors are as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_j - y_i\right)^2} \tag{2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^1 - y_i\right| \tag{3}$$

$$\text{R2} = 1 - \frac{SSE\ (\text{error sum of squares})}{SST\ (total\ sum\ of\ squares)} \tag{4}$$

Based on the difference between the closing prices of every two days, if today's price is greater than yesterday's real price, it will be recorded as 1, otherwise it will be recorded as 0. The original data and the predictive results of each model are processed above to obtain the fluctuation information of the stock closing price. 1 means that it is higher than yesterday, and 0 is lower than yesterday. The results of each model are compared with the original data fluctuations to obtain True Negatives, False Positives, False Negatives and True Positives, then use these four factors to calculate three kinds of accuracy rates for each machine learning model.

3. Results and Discussion

The different close prices predicted by the five models were plotted, while the errors and trend prediction results of each model were recorded in multiple tables. The horizontal coordinate of each line chart is the date, and the vertical coordinate is the closing price, with red lines represents the actual value and blue lines represents the predicted value. In each table, each row contains multiple test results for a model. The charts, tables, and related analysis are as follows.

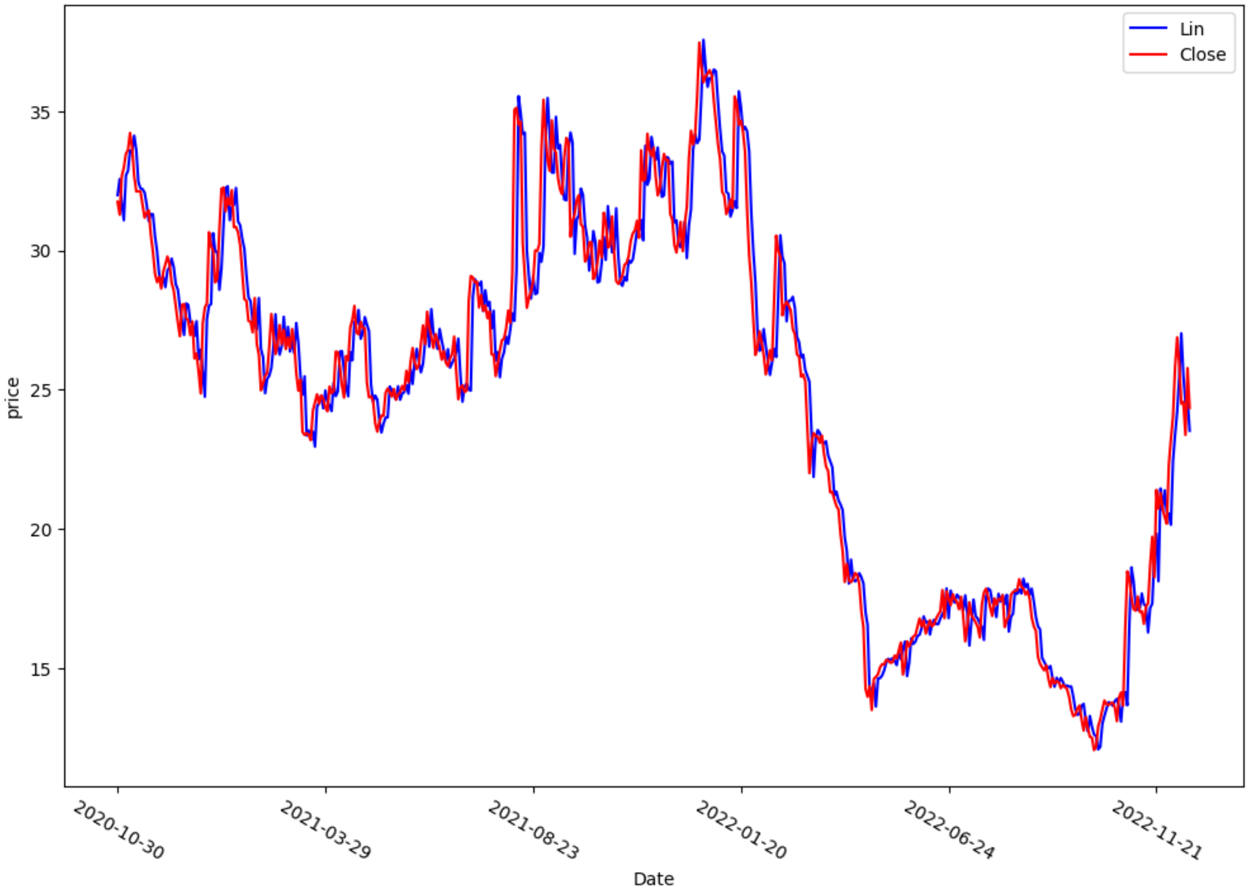

Figure 2. Result of linear regression

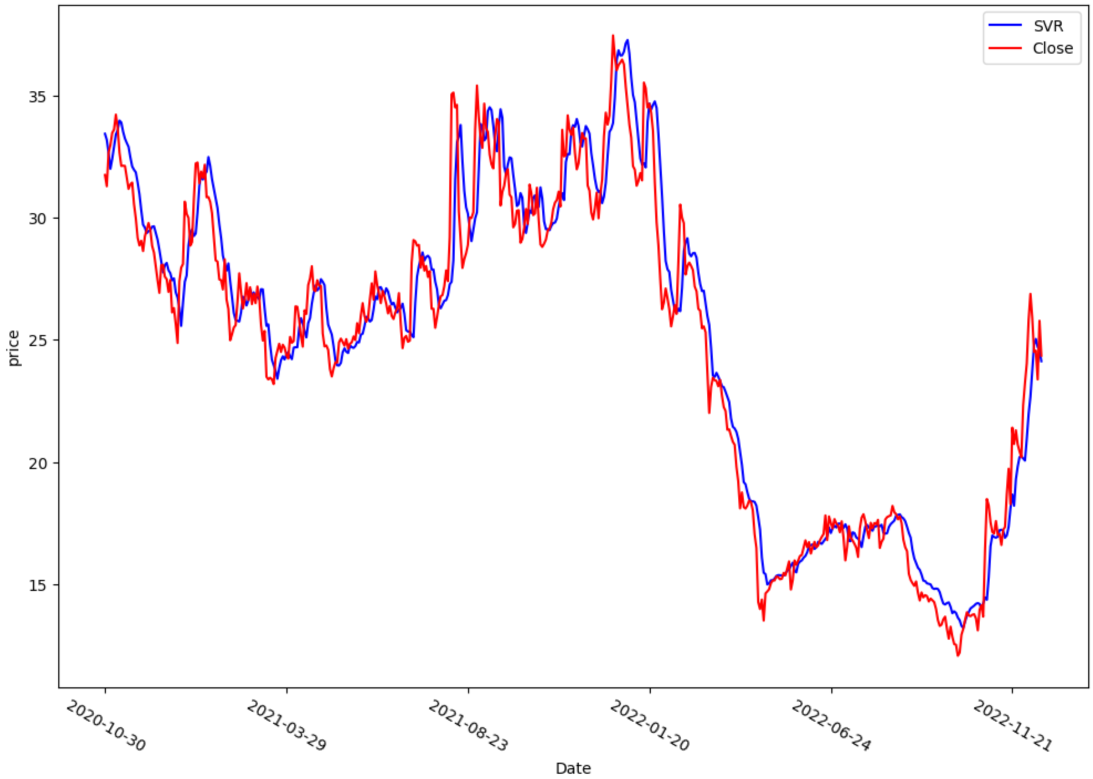

Figure 3. Result of SVR

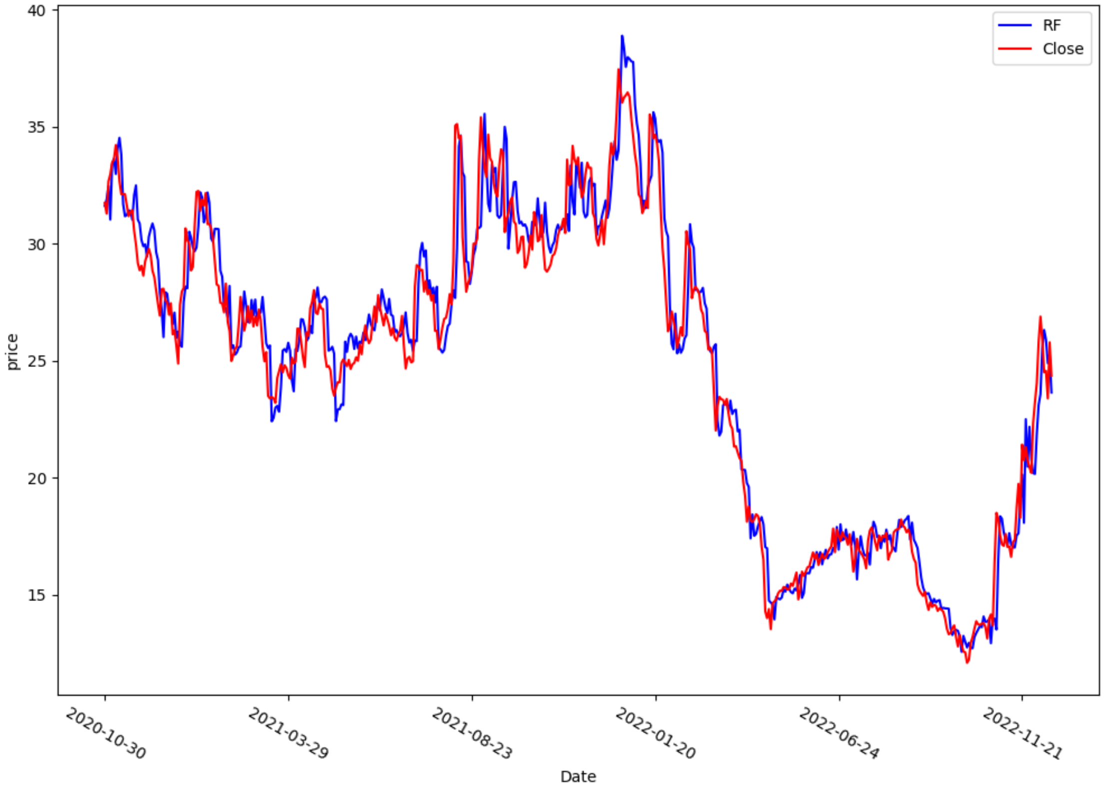

Figure 4. Result of random forest

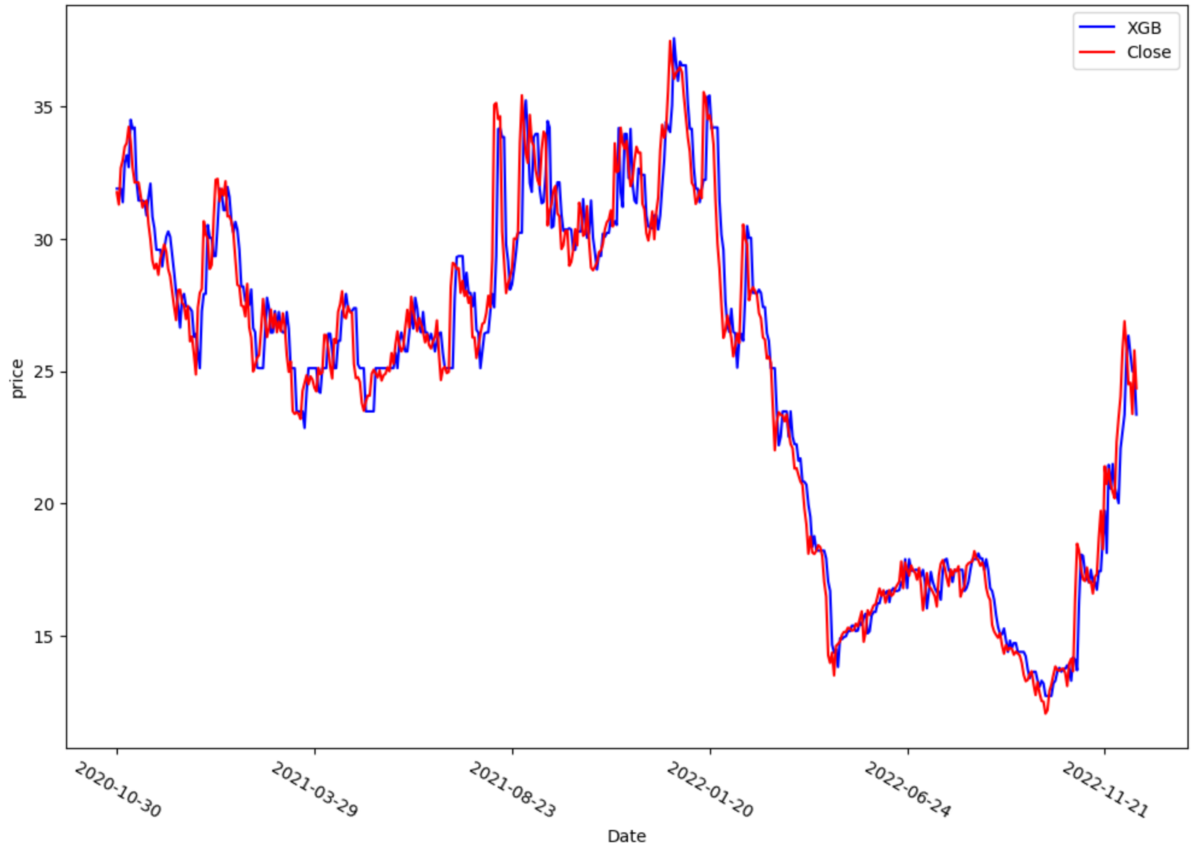

Figure 5. Result of XGBoost

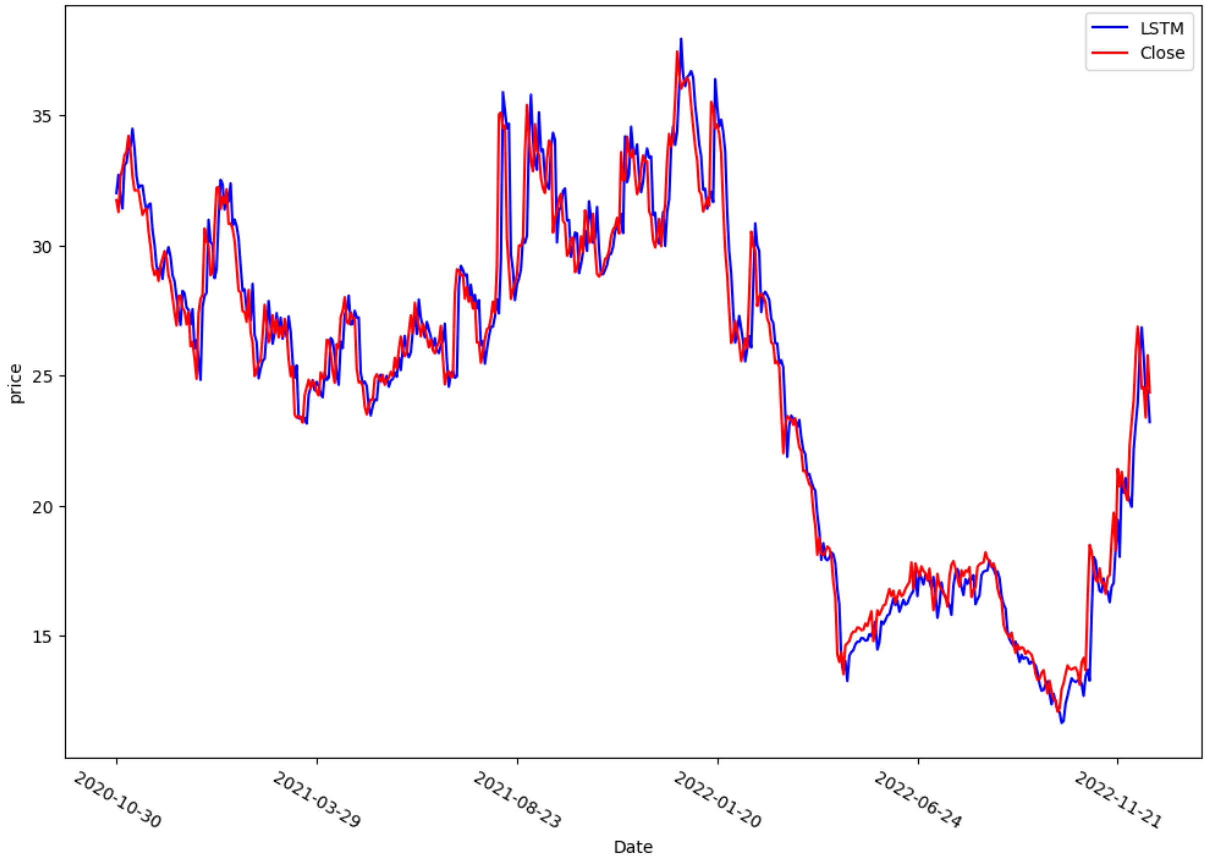

Figure 6. Result of LSTM 1

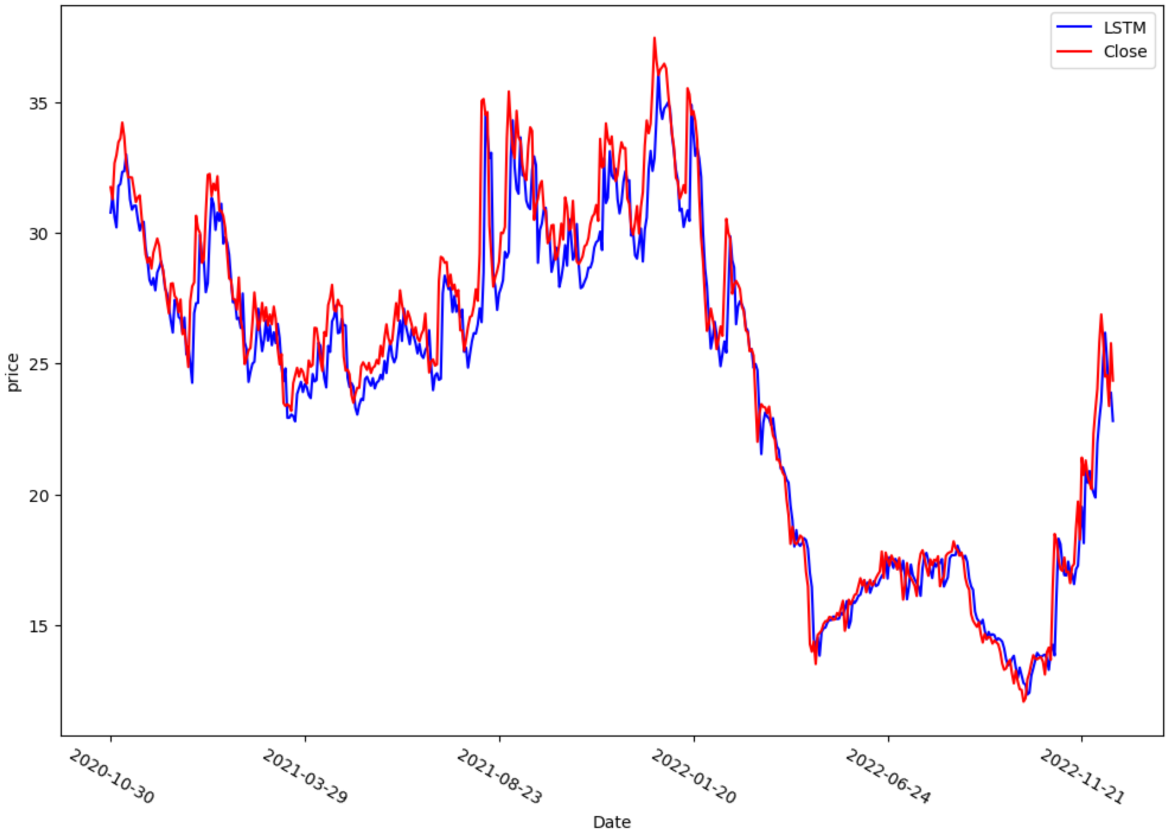

Figure 7. Result of LSTM 2

It can be seen from Figure 2 and Figure 6 that the linear regression model and LSTM1 model predict the closing price of the stock of Beijing E-Hualu Information Technology Co., Ltd. more accurately, which are closer to the real data and has more overlapping parts, and the prediction results of the random forest regression model (Figure 4) and the XGBoost model (Figure 5) are also relatively successful. However, the results of SVR model and LSTM2 model, from Figure 3 and Figure 7, are not ideal.

Table 2. Model errors

Model	MSE	RMSE	MAE	R-Squared
Linear regression	0.758	0.871	0.592	0.982
SVR	1.361	1.166	0.842	0.968
Random forest	1.024	1.012	0.760	0.976
XGBoost	0.838	0.915	0.627	0.980
LSTM 1	0.744	0.863	0.590	0.982
LSTM 2	1.224	1.106	0.790	0.966

In Table 2, the information of MSE, RMSE, MAE, and R-Squared corresponding to each model is recorded. The values of the three errors of MSE, RMSE, and MAE are related to the size of the data itself, but normally the smaller they are, the more accurate the prediction is. R-Squared should be a number between -1 and 1. If the result is closer to 1, the model is more accurate. If it is negative, the model is poor.

Overall, the MSE, RMSE, and MAE of LSTM1 and linear regression are all less than 0.9, and the values of R-Squared are both 0.982, indicating that the predictions of these two models are more accurate. However, LSTM2 produced the largest MSE, RMSE, MAE, and the smallest R-Squared, which is considered to be the worst predictive model.

Table 3. TN FP FN TP

Model	TN	FP	FN	TP
Linear regression	112	160	126	119
SVR	108	164	88	157
Random forest	114	158	98	147
XGBoost	114	158	106	139
LSTM 2	206	66	180	65

Table 4. Accuracy

Model	Total accuracy (%)	True positive rate (%)	True negative rate (%)
Linear regression	44.68	48.57	41.17
SVR	51.25	64.08	39.70
Random forest	50.48	60.00	41.91
XGBoost	48.93	56.73	41.91
LSTM 2	52.41	26.53	75.73

In Table 3 and Table 4, the prediction results and accuracy of each predictive model for stock movement are recorded respectively. The three accuracy rates in Table 4 are calculated based on the data in Table 3. The mathematical formula is as follows:

$$\text{Total accuracy} = \frac{(TP + TN)}{(TN + FP + FN + TP)} \tag{5}$$

$$\text{True positive rate} = \frac{TP}{(TP + FN)} \tag{6}$$

$$\text{True negative rate} = \frac{TN}{(TN + FP)} \tag{7}$$

In Table 4, the LSTM2 model and the SVR model, which previously performed poorly in predicting stock prices, performed well in predicting stock trends, with an overall accuracy of 52.41% and 51.25%, respectively. Moreover, XGBoost and linear regression, which performed well before, performed poorly in predicting trends. The accuracy of XGBoost was less than 50%, and the linear regression model was even lower than 45%. In particular, LSTM2 has an excellent prediction effect on True negative, with an accuracy rate of 75.73%, while SVR has a good prediction result on True positive, with an accuracy rate of 64.08%.

To sum up, for the stock of Beijing E-Hualu Information Technology Co., Ltd., whether it is predicting stock prices or stock trends, the LSTM model has good accuracy with the right parameters. Both linear regression and eXtreme Gradient Boosting are good model choices when predicting stock prices. When it comes to stock trend forecasting, try random forests and support vector regressor. On top of that, combining the prediction features of the LSTM model and the SVR model, the SVR model is used to predict the upward trend, and the LSTM model is used to predict the downward trend, in order to obtain the highest prediction accuracy. The results of this study verified that LSTM is one of the best models in financial forecasting like stock-related forecasting as other studies have shown.

4. Conclusion

In this research, with only closing price input, different prediction results are obtained through different machine learning models. One of the LSTM models and the linear regression model performed well for close price forecasting, while the other LSTM model and support vector regression model performed well for trend forecasting. Based on the results, LSTM model can satisfy price prediction and trend prediction under different parameter adjustments. In particular, when predicting stock trends, two or more models can be combined to obtain the best results according to the prediction characteristics of different models.

In future studies, the design of the LSTM model will be improved to achieve the same or even better prediction results with a lower number of layers. Meanwhile, the author will also try to integrate a variety of machine learning models, study the advantages and disadvantages of the models, and predict more stocks, in order to find the best and more generally applicable stock prediction method.

References

[1] Thennakoon A, Bhagyani C, Premadasa S, et al. Real-time credit card fraud detection using machine learning. IEEE 2019 9th International Conference on Cloud Computing, Data Science & Engineering, 2019: 488 - 493.

[2] Kaur I, Kaur J. Customer churn analysis and prediction in banking industry using machine learning. IEEE 2020 Sixth International Conference on Parallel, Distributed and Grid Computing, 2020: 434 - 437.

[3] Obthong M, Tantisantiwong N, Jeamwatthanachai W, et al. A survey on machine learning for stock price prediction: algorithms and techniques, 2020.

[4] Preethi G, Santhi B. Stock Market forecasting techniques: a survey. Journal of Theoretical & Applied Information Technology, 2012, 46 (1).

[5] Hyndman R J, Athanasopoulos G. Forecasting: principles and practice. OTexts, 2018.

[6] SUKESTI F, GHOZALI I, Fuad F, et al. Factors affecting the stock price: The role of firm performance. The Journal of Asian Finance, Economics and Business, 2021, 8 (2): 165 - 173.

[7] Blaufus K, Möhlmann A, Schwäbe A N. Stock price reactions to news about corporate tax avoidance and evasion. Journal of Economic Psychology, 2019, 72: 278 - 292.

[8] Sun Jingyu. The influence of macroeconomic factors on stock price fluctuations. Quality and market, 2021.

[9] Zaremba A, Aharon D Y, Demir E, et al. COVID-19, government policy responses, and stock market liquidity around the world: A note. Research in International Business and Finance, 2021, 56: 101359.

[10] Long W, Lu Z, Cui L. Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Systems, 2019, 164: 163 - 173.

[11] Bathla G. Stock Price prediction using LSTM and SVR. IEEE 2020 Sixth International Conference on Parallel, Distributed and Grid Computing, 2020: 211 - 214.

[12] Staudemeyer R C, Morris E R. Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586, 2019.