Stock Price Prediction based on Multiple Regression Models

Yuhan Li

Beijing University of Technology, Beijing 100083, China

liyuhan@emails.bjut.edu.cn

Abstract. Over the past two years, global stock markets have gradually recovered and new investors have entered the market. While there are many factors affecting stock prices and the stock market is changing rapidly, the way to accurately predict stock prices has become the focus of investors. This paper will use the concept of machine learning to predict the stock prices of three listed companies based on three different regression models (i.e., OLS, Ridge and XGBoost). According to the analysis, the OLS model and the Ridge model are very accurate in predicting stock prices, especially in the low and middle price ranges. In contrast to these typical linear regression models, the XGBoost model is not as accurate in predicting stock prices and even has a significant prediction bias in the high price range. These results will enable subsequent research to make better choices when selecting models for forecasting, especially for data sets with different characteristics.

Keywords: Stock Price Prediction; Linear Regression; XGBoost; Machine Learning.

1. Introduction

Since the outbreak of COVID-19 in 2020, the global economy has been hit hard, especially the stock market. Globally, all kinds of investment activities were significantly restricted in the early stage of the pandemic, with a sharp decline in the number of investment transactions and the amount. However, as online offices matured, outbreak control systems became more sophisticated and relevant vaccines were developed, global investment activities have recovered significantly [1, 2]. In countries where the epidemic was more rapidly controlled, equity investment activity has shown signs of recovery since mid-2020. As there is a significant causal relationship between the development index of stocks and economic growth, the recovery degree of the stock market largely depends on the economic freedom, governance, and social productivity of the country [1, 2].

In a tough market environment, forecasting stock prices becomes crucial. Using the basics of the stock market and technical indicators, financial markets can be examined from different perspectives. This information can be used to make reliable forecasts and achieve higher profits [3]. The lower the forecast error, the lower the investment risk and the higher the profitability [4]. The formation and fluctuation of stock prices are not only subject to various economic and political factors but are also influenced by investment psychology and trading techniques [5]. For investors, there is a direct link between the prediction of stock market trends and the acquisition of profits, and the more accurate the prediction, the more certain the prevention of risks [5]. For listed companies, the stock index reflects the company's operations and future trends and is the main technical indicator for analyzing the company [5, 6].

Hossain et al. proposed a new stock price forecasting technique by optimizing the Belief Rule-Based Expert System (BRBES) with its model-related parameters through the fmincon function. It was found that the optimized BRBES has better performance and better potential in stock price prediction [4]. Bazrkar et al. proposed a support vector machine (SVM) trained on existing stock data and used the particle swarm optimization (PSO) algorithm to improve the prediction accuracy [7]. Studies have found that the prediction accuracy of the optimized SVM model is above 90% [7]. At present, when using machine models for stock price prediction, researchers generally prefer to use SVM and pay more attention to technical indicators, while Harikrishnan et al. found that the usage of hybrid indicators and Long-Short Term Memory (LSTM) model is becoming a new trend [8]. Mohanty et al. adopt the LSTM model to predict market stock prices by using the sequential structure of time series data. According to the results of StockBot, this model has achieved relatively accurate

prediction results, and can even obtain 15 times higher returns than the most aggressive ETF in the market according to the recommended trading time of StockBot [9]. By combining LSTM and Gated Recurrent Units (GRU) models with four neural network block architectures, Lawi et al. proposed a set of methods to test the accuracy of the model from three dimensions: low, medium, and high [10].

This research aims to predict stock prices by using knowledge related to machine learning. Firstly, this author conducts data screening and selects Tesla Inc. from the US market and two biopharmaceutical companies from the Chinese market. Subsequently, three representative linear regression models were chosen to predict the stock prices of these three listed companies. In this paper, the trained model is tested, and the results are analyzed by visualizing the results and comparing the model correlation coefficients. Finally, the paper develops the analysis to address the shortcomings of the three models and the limitations of the data set selection in this research.

2. Data & Method

Three stocks were selected for this essay, one of them is Tesla from the United States, and the other two are from China's medical sector, namely Vcanbio and Sinocare. Historical data relating to the stocks are downloaded from investing.com and money.163. Tesla is a new energy company specializing in electric vehicles located in the United States, involving electric motors, electric cells, automatic driving, new energy, and artificial intelligence. It has the world's leading development speed and several technology patents in the field of high-performance electric vehicles. Tesla not only dominated the electric vehicle market in the United States for a long time, but even after entering the Chinese market, it quickly became the first seller in the new energy market. Due to its excellent market performance, the company's stock has also received a great deal of attention. In this study, the stock of Tesla Inc. will be selected as a representative of the US market using machine-learning knowledge to forecast it. As a company related to cell genetic engineering, Vcanbio represents the current level of cell medical treatment in China to some extent. Sinocare is a company dedicated to the development of rapid detection of chronic infectious diseases using biosensors. Both Chinese companies are among the more representative companies in the medical sector.

The opening price, the highest price, the lowest price, the volume, and the transaction amount of the stock are selected as independent variables for analysis in an attempt to predict the closing price. This study selected all the data of the opening day from January 1, 2018, to October 13, 2022, and randomly divided them into the training set and the test set in the ratio of 8:2. Since the time series has a very clear trend, so it is necessary to shuffle the samples.

Three models were selected for the study, the Ordinary Least Square (OLS) model and the Ridge model, which are linear regressions, and the XGBoost model. The OLS model minimizes the sum of variances between the true and predicted values of the sample to calculate the parameters w_0, w_1, w_2, etc. The regression equation can be written as:

$$y = w_0 + \sum_{i=1}^{n} w_i x_i \tag{1}$$

The optimization target for the OLS model can be expressed as:

$$\arg\min\left(\Sigma(y_i - y)^2\right) \tag{2}$$

The Ridge model was proposed by Tikhonov, which is an improvement on OLS. It aims to control the crazy growth of regression parameter values along with the dimension by changing the regression objective function. The model can be expressed as:

$$\arg\min\left(\Sigma(y_i - y)^2 + \alpha\Sigma w_i^2\right) \tag{3}$$

The XGBoost model is a machine learning algorithm based on the decision tree, with gradient boost as its framework. The model can be expressed as:

$$F_T(x) = \sum_{m=0}^{T} f_m(X) \tag{4}$$

Table 1. OLS regression results for three companies.

	Vcanbio	Sinocare	Tesla
R-squared	0.994	0.998	1
Adj. R-squared	0.994	0.998	1
F-statistic	2.39E+04	7.75E+04	5.36E+05
Prob(F-statistic)	0	0	0
Log-LikeLIhood	-71.983	-480.06	-2267.5
AIC	158	974.1	4545
BIC	191.7	1008	4569

3. Results & Discussion

Firstly, the OLS model is used to predict the stock closing price. The prediction results by the OLS model for three companies are shown in Fig. 1 and Table. 1. The regression model shows that the OLS model is generally accurate in the estimation of Vcanbio, especially in the lower price ranges. However, in the few cases where the closing price is higher, the prediction deviation occurs and even differs significantly from the real data. For Sinocare, the regression model image and $R^2 > 0.8$ show that the OLS model is relatively accurate for the estimation of Sinocare, especially in the low and mid-range of the stock price. However, when the stock price exceeds RMB 43, the OLS model will be out of calibration but does not produce a large difference from the real data. As for Tesla, the regression model image and the fact that R^2 is close to 1 show that the OLS model predicts Tesla stock very accurately and rarely misses the mark.

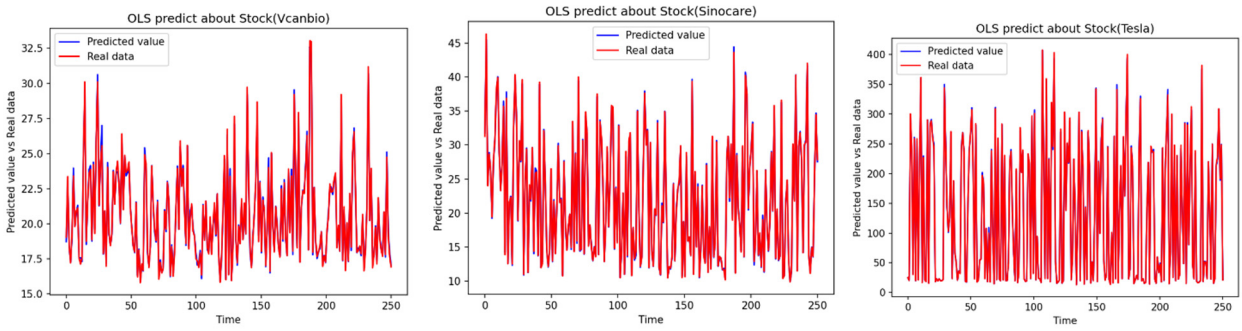


Fig 1. OLS prediction of Vcanbio, Sinocare and Tesla.

Secondly, the Ridge model is used to predict the stock closing price. The prediction results by the Ridge model for three companies are shown in Fig. 2. A comparison of the predicted and real data shows that the Ridge model is generally accurate for CPY stock, especially in the RMB 17.5 - 22.5 region. However, when the closing price exceeds RMB 22.5, the probability of the Ridge model's prediction deviation will gradually increase, but there will be no big difference with the real data. For Sinocare, by comparing the predicted values with the real data, the study found that the Ridge model was generally accurate in estimating the closing price of Sinocare, especially in the area below 30 RMB, but when the closing price exceeded 35 RMB, the Ridge model began to show frequent prediction errors, and even one large error when the price exceeded 45 RMB. Regarding to Tesla, a plot of the real data against the predicted values shows that the OLS model is very accurate in predicting Tesla stock, and there will be an inaccurate phenomenon when the price is very high.

Finally, the XGBoost model is used to predict the stock closing price. The prediction results by the XGBoost model for three companies are illustrated in Fig. 3. Seen from the results, the XGBoost prediction of Vcanbio is significantly less accurate than the previous two models. The model frequently misses the mark and can even have very large forecast deviations (> 5 RMB). For Sinocare, XGBoost forecasts for Sinocare have improved compared with Vcanbio, especially in the low to mid-range, but there are still frequent misses in the above 30RMB region. In addition, XGBoost closing

price predictions for Tesla are generally very accurate, but on rare occasions, they are significantly off the mark.

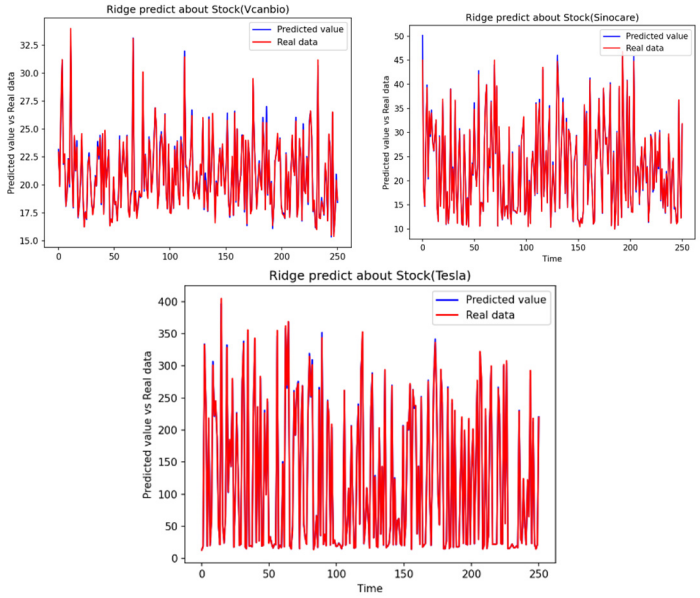


Fig 2. Ridge prediction for Vcanbio, Sinocare and Tesla.

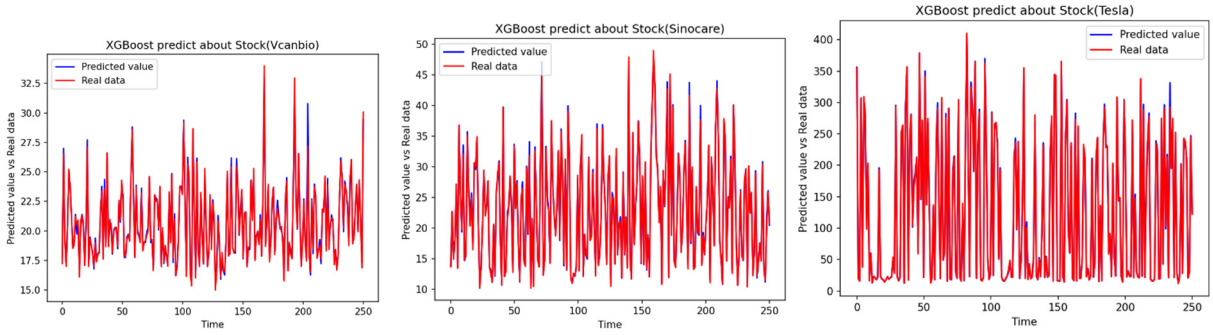


Fig 3. XGBoost prediction for Vcanbio, Sinocare and Tesla.

Table 2. MSE, MAE, and R2 about OLS, Ridge, and XGBoost.

	OLS			Ridge			XGBoost		
	Vcanbio	Sinocare	Tesla	Vcanbio	Sinocare	Tesla	Vcanbio	Sinocare	Tesla
MSE	0.07	0.11	6.85	0.07	0.09	5.88	0.12	0.30	15.80
MAE	0.18	0.23	1.45	0.18	0.22	1.34	0.26	0.36	2.32
R2	0.99	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00

The OLS, Ridge, and XGBoost models are trained and tested based on the data of three stocks, namely, Vcanbio, Sinocare, and Tesla over the past five years. As shown in Table. 2. The results show that the OLS and Ridge models are more accurate in predicting the closing price than XGBoost, especially in the low to mid-region. This may be due to the OLS model is suitable for situations where the relationship to be modeled is not very complex and the amount of data is not too large, which just fits the characteristics of the data set obtained in this study. The Ridge model, on the other hand, by better controlling the variance and bias of the model, significantly improves the selection of variables and the robustness of prediction compared to OLS regression, solves the phenomenon of overfitting of OLS in data with high feature dimensions, and makes the prediction results more accurate. Due to a large number of parameters and the complexity of parameters, the model does not work well when the principle of XGBoost is not well understood by the user, resulting in a larger deviation of stock price prediction results compared to the first two models. When there is not a lot of stock data and

the variables are not complicated, OLS and Ridge models can be selected, which are simple to use and more accurate.

4. Limitations & Prospects

The OLS and Ridge models used in this research are linear regression models, which are simple and easy to use and require a high level of data set selection. It can only be used when the relationships modeled are not very complex and the amount of data is not too large. As shown in Table. 2, compared to Vcanbio and Sinocare, the prediction accuracy of the OLS model and Ridge model is significantly lower for Tesla, which has complex stock price fluctuations. Whereas the XGBoost model has good processing speed and accuracy for low and medium-dimensional data, it is difficult to handle datasets with high-dimensional features. In addition, the dataset of this study also has some limitations. During the training of the model, the multicollinearity problem of the dataset is existed, for there is no better solution except to eliminate the relevant factors. In further study regarding to stock price prediction, if the dataset is large and complex, it is not recommended that researchers choose a linear regression model. When faced with high-dimensional datasets, deep learning models are more effective in prediction compared to XGBoost models. Moreover, when using machine learning related knowledge to predict prices, the selection of a dataset is extremely demanding and it is necessary to have a deep understanding of the characteristics of the model and the correlation between the data in order to train the model correctly. When predicting stock prices, it will be better to choose independent variables that are uncorrelated or less correlated to greatly improve model accuracy.

5. Conclusion

In summary, this paper investigates stock prices based on multiple regression models, such as the OLS model, ridge model, and XGBoost model. The stock data of three listed companies (i.e., Vcanbio, Sinocare, and Tesla) in the recent five years is used as the dataset with a ratio of 8:2 to train and test these models. According to the analysis, the OLS and Ridge models are very accurate in predicting stock prices, especially in the low and middle-price regions, but both models show small deviations from prediction accuracy when the stock prices are high. Compared with these two typical linear regression models, XGBoost, a machine learning algorithm based on the decision tree, has a lower accuracy in predicting the stock prices of these three companies, especially in the area of high stock prices, and even has a significant bias in prediction. As the regression model cannot cope well with the prediction of a large amount of data, the dataset with a large amount is not chosen. In addition, the factors in the dataset used are correlated, which leads to the multicollinearity problem in prediction. In future research on stock price prediction, more care should be taken to conform to the requirements of the model-related characteristics when selecting data for the independent variables. These results indicate that the subsequent selection of the linear regression model and XGBoost more suitable and accurate. Overall, these results offer a guideline for stock price prediction in the concept of machine learning.

References

[1] Uddin, M., Chowdhury, A., Anderson, K., Chaudhuri, K.: The effect of COVID – 19 pandemics on global stock market volatility: Can economic strength help to manage the uncertainty? Journal of Business Research, 128(2021), 31- 44 (2021).

[2] Bhattarai, J. K., Gautam, R., Chettri, K. K.: Stock Market Development and Economic Growth: Empirical Evidence from Nepal. Global Business Review, 0(0), 09721509211016818 (2021).

[3] Thakkar, A., Chaudhari, K.: A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. Expert Systems with Applications, 177, 114800-114817 (2021).

[4] Hossain, E. et al.: Machine learning with Belief Rule-Based Expert Systems to predict stock price movements. Expert Systems with Applications, 206, 117706-117719 (2022).

[5] Yang, Y. J., Yang, Y. M., Zhou, W.: Research on a hybrid prediction model for stock price based on long short-term memory and variational mode decomposition. Soft Comput 25, 13513-13531 (2021).

[6] Bousoño-Calzón, C., Bustarviejo-Muñoz, J., Aceituno-Aceituno, P., Escudero-Garzás, J. J.: On the Economic Significance of Stock Market Prediction and the No Free Lunch Theorem. IEEE Access, 7, 75177-75188 (2019).

[7] Bazrkar, M. J., Hosseini, S.: Predict Stock Prices Using Supervised Learning Algorithms and Particle Swarm Optimization Algorithm. Computational Economics 1-22, (2022).

[8] Harikrishnan, R., Gupta, A., Tadanki, N., Berry, N., Bardae, R: Machine Learning Based Model to Predict Stock Prices: A Survey. IOP Conference Series. Materials Science and Engineering, 1084(1), (2021).

[9] Mohanty, S., Vijay, A., Gopakumar, N.: StockBot: Using LSTMs to Predict Stock Prices. Statistical Finance, arXiv:2207.06605v2 [q-fin.ST] (2022).

[10] Lawi, A., Mesra, H., Amir, S.: Implementation of Long Short-Term Memory and Gated Recurrent Units on grouped time-series data to predict stock prices accurately. Journal of Big Data 9, 89 (2022).