# Volatility Prediton based on Multiple-factor Model

Jingdan Yang*

School of Business Zhejiang University of Finance and Economics Hangzhou, China

* Corresponding author: 180120430220@zufe.edu.cn

**Abstract.** Generally, due to unpredictable and vulnerable features of the stock market, traditional statistical methods with single factor cannot explain the movements of stock prices well. In this paper, multi-factor model of quantitative stock selection is applied to predict stock price index to perceive the fluctuation of stock in multiple aspects and improve the accuracy of prediction. This study also combines machine learning measurements with historical data fitting effects. The data used for analysis is all collected from Tushare website, and the stock data are from the 2015 to 2022 data of Shanghai Pudong Development Bank and China Merchants Bank, and the selected factors include price-to-earnings ratio, price-to-book ratio, volume ratio, total equity and so forth. Ordinary least square linear regression and random forest nonlinear approaches are utilized to predict the stock price. According to the analysis, the accuracy of random forest is higher than OLS in stock price prediction. Among all the factors, the opening price, the highest price and the lowest price have relatively large impacts on the closing price of the next day. However, when selecting different factors, stocks and train-testing periods, the obtained regression models are different. Therefore, the determination of the correlation coefficient is not invariable and needs to be analyzed on a case-by-case basis. Overall, these results shed light on the superiority of machine learning model and the significant contribution of some factors on stock price prediction.

**Keywords:** Stock price index, Multi-factor model, OLS, Random forest

## 1. Introduction

The origins of stocks can be traced back to the 16th century, when the Dutch East India Company developed its overseas trading business. However, the company ran short of cash as it expanded. To raise money, the Dutch East India Company divided and sold ownership of the company, which was originally in the form of stocks. With the improvement of social economy and productivity, an increasingly large number of people are keen on stock trading, and a great deal of theoretical models and empirical conclusions are pouring out. The stock market is highly sensitive and widely affected by economic indicators, leaders' statements, national policies, and even rumors [1]. Besides, some factors are long term and have a big impact, while others are short term and have a small impact. In an era of explosive growth in information, it is also significant to confirm the authenticity of information.

Obviously, because of the existence of these problems, stocks have become a hot topic in the financial field, and the prediction of stock price has attracted more and more attention from all circles of the scholars, e.g., education, sociology, and economics. Meanwhile, the rival of globalization and big data have provided copious data materials, which is conducive to experts' collections, and made it possible to combine machine learning and researches of stock price prediction. From Capital Asset Pricing Model (CAPM model), to Fama-French Three-Factor Model, and then to Fama-French Five-Factor Model, factor pricing models are constantly updated, stimulating researchers to dedicate to optimizing the selection and combination of stock factors, and aiming at producing more accurate predictions of stock market trends.

At present, reports that combine multiple factors theory and machine learning predicting trends of stock price are scarce, although there are many motifs on stock. Therefore, based on existing literature, this study introduces multiple factors model of quantitative strategies for selecting stocks, and reflecting the economic characteristics of stock from various aspects. Besides, this paper also achieves factor acquisition, screening, standardization, validity testing and so forth by combing python with other scientific strategies. Then, the trends of the two sample stocks' prices are predicted, and a valuable model is successfully built, providing some inspiration for the stock market.

Based on existing theories, investors are trying to come up with innovative ideas to better fit historical stock data and predict future stock prices. Pei and Zhu used multi-factor theory and long and short-term memory network model to predict stock prices, proving that the introduction of multi-factor model improves the accuracy of prediction by conducting an experiment [2]. They described the multi-factor model as a linear regression problem:

$$P_{stock} = \sum_{i=1}^{N} w_i F_i + \epsilon \quad (1)$$

Where $P_{stock}$ is the stock price, $N$ represents the number of factors, $F_i$ is a set of factor values, $w_i$ is a set of the factors' weight, and $\varepsilon$ is a random term with mean zero.

Compared with the results by using deep learning and multi-factor model [2], other scholars used a very simple way to predict the price of a domestic stock [3]. Although it is only chosen a small number of trading days and factors, the construction and improvement of the model is clarified clearly and excluded the multicollinearity problem in multiple linear regression, which is suitable for beginners.

Levin used multilayer feedforward neutral networks and various technical and fundamental factors to predict a stock's excess return. The author pointed out that while linear factor models have proven to be very useful tools for studying stocks, the assumption of linear relationship between factor values and expected return was quite restrictive. The linear model ignored the possible interactions between factors. Therefore, a nonlinear relational model was considered and its validity was verified by experiments [4].

Huang and Liu proposed hypotheses that 7 risk factors (market premium, size premium, book-to-market premium, profitability premium, investment growth premium, momentum premium, asset turnover premium) have a positive relationship with excess return. This research used some statistical methods (e.g., stability test, OLS regression, ridge regression, robustness test, Chi-Square test) to verify the hypotheses, providing some theoretical guidance for others [5].

With the rapid development of artificial intelligence, scholars are actively exploring new methods to predict stock prices. Pei proposed a method to screen stock factors based on the combination of GBDT (Gradient Boosting Decision Tree) and FFM (Field-aware Factorization Machine) [6]. It was a very successful attempt to improve the LSTM model (Long Short-Term Memory) and increase the accuracy of stock price prediction by 10.12%.

Stock volatility prediction based on the multi-factor theory in quantitative stock selection can indeed improve the fit with the real situation to a certain extent, but it will lead to multicollinearity problems caused by multi-factors. Yang, et al. point out that multicollinearity refers to the fact that there is better correlation between two or more variables in a linear regression model, which makes the model difficult to estimate accurately. When the variables are highly correlated or the number of sample data is very small, it is easy to produce multicollinearity problems [7].

Besides, another study explains the theory and implementation method of multi-factor model very clearly [8], which points out that the factor model is based on a fundamental principle of financial theory: no reward without risk. In this article, there are three methods to estimate multi-factor models: time series analysis, cross-section analysis, and statistical factor analysis, and it is be proved that the last one has the highest explanatory power. Multi-factor model can explain more than 30% of variance on average by using cross section method, and standardizing factors such as yield, roe, peg.

Each scholar cannot put forward a variety of innovative points without predecessors' exploration and efforts. Stock price prediction methods can be roughly divided into two categories, mathematical statistics, and machine learning methods. In the early days, people generally used mathematical statistics methods, including Auto-Regressive Integrated Moving Average (ARIMA), Generalized Auto-Regressive Conditional Heteroskedasticity Model (GARCH), and Auto-Regressive Conditional Heteroscedasticity Model (ARCH) [9, 10]. However, these methods can only linearly fit the stock price with other influencing factors, and it is difficult to reflect the nonlinear characteristics of the stock. With the development and popularization of computers, scholars have begun to explore the use of Support Vector Machines (SVM), Artificial Neural Networks (ANN) and other models to predict stock prices, improving the prediction accuracy of models. Nevertheless, traditional machine learning

algorithms have no "memory", triggering further exploration by scholars. The Multi-Layer Perceptron model (MLP) proposed by Professor G. Hinton is the earliest depth model [11]. Later, scholars have proposed Back Propagation algorithm (BP), Recurrent Neural Network (RNN), Long- and Short-Term Memory Network (LSTM), and a combination of various models, which solve the complex computing problems of neural network, reduce the risk of gradient disappearance and gradient explosion. These studies make the model mature and improve the performance and application of the model.

The multi-factor model is used to predict the volatility of stock price from the quantitative stock selection problem, and two stocks in the CSI 300 stock market are predicted. This study will screen out the most representative 4-7 factors, using multiple regression method, combined with machine learning to predict the stock price trend. It is hoped that the model and method of this study can predict the price well and provide some guidance for investors. The rest part of the paper is organized as follows. The Sec. II will select apt models and clean data. The Sec. III will represent and discuss the results of prediction, The Sec. IV will summarize the whole research and indicate future research direction.

## 2. Methodology

### 2.1. Data

In this study, Shanghai Pudong Development Bank (stock code: 600000.SH) and China Merchants Bank (stock code: 600036.SH) are selected as the research objects, and the time range is confirmed as September 15, 2015 to March 1, 2022. 20 indicators including price/earnings ratio, price-to-book ratio, turnover rate, total equity, and total market value are taken as candidate factors, and next-day closing price is taken as dependent variable. All data will be retrieved from the Tushare platform.

### 2.2. Model

Factor can be regarded as a mathematical expression affecting stock price fluctuation, which comes from economic laws and market experience. The establishment of multi-factor model requires the following three steps: factor selection, factor screening, model construction and update. In this study, 20 candidate factors were selected by referring to relevant literature and combining with the real situation of China's stock market and previous experience. The factors with relatively high correlation degree and high differentiation degree to the annualized compound return of stock are screened by validity test and redundancy factor elimination. Then using regression method to fit the selected stocks and factors, and testing the effect of the model.

Ordinary least square method is the most common method to estimate unknown parameters of multiple linear regression model. According to the principle of OLS, the parameter estimates should minimize Q.

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i^2\right) \quad (2)$$

Random forest is an integration algorithm based on decision tree proposed by Breiman. It belongs to bagging algorithm, and its idea adopts a method similar to democratic voting, which means each base model has one vote, and the final result is voted by all the base models, using the principle of majority rule to produce the predicted result. Random forest is widely used to solve regression problems. The model averages the results to avoid over-fitting and improve the prediction performance. Besides, the benefit of random forest model is that it can estimate missing data effectively and deal with high-dimensional data.

**Table 1.** Partial data presentation of turnover rate factor of 600000.SH stock after standardized processing.

| Trade_date | Turnover_rate | Turnover_rate_3mad | Turnover_rate_stand |
|---|---|---|---|
| 2015-09-15 | 0.721 | 0.178 | 0.565 |
| 2015-09-16 | 0.497 | 0.118 | -0.235 |
| 2015-09-17 | 0.341 | 0.185 | 0.647 |
| 2015-09-18 | 0.274 | 0.172 | 0.481 |
| 2015-09-21 | 0.271 | 0.159 | 0.313 |
| 2015-09-22 | 0.259 | 0.139 | 0.042 |
| 2015-09-23 | 0.366 | 0.130 | -0.066 |
| 2015-09-24 | 0.529 | 0.089 | -0.605 |

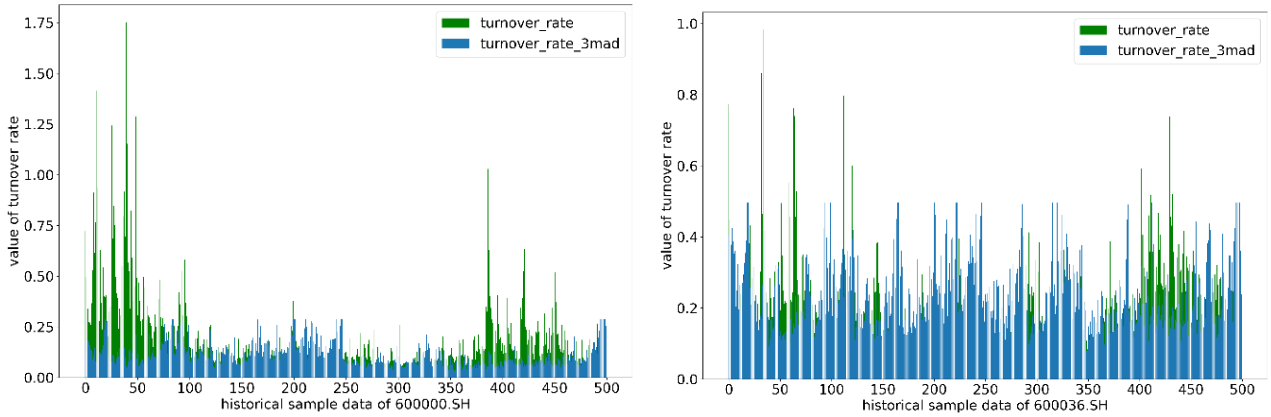



**Figure 1.** The comparison of turnover rate and turnover rate after extreme value removal of 600000.SH and 600036.SH stocks (top 500 data).

### 2.3. Processing

In the first place, this study selects some general alternative factors through the references, and deals with the factors to the extreme. Because factors with extreme values may lead to large standard deviations, affecting the accuracy of conclusions. Instead of deleting the abnormal data, this approach constrains them within normal values. There are three common methods for factor minimization: quartile minimization, median absolute deviation minimization and normal distribution minimization. In this study, the second method is selected and the results of the turnover rate factor are presented. The results are shown in Figure 1.

When the elimination of extreme factors is complete, the factors need to be standardized. Since the units of factors are different, some are ratios, some are numbers, data standardization can eliminate the inaccuracy and unreliability caused by such factors. There are three methods of data standardization: linear method, folded method, and curve method. This study adopts the first method and uses turnover rate factor of 600000.SH as demonstration.

## 3. Results & Discussion

Correlation coefficient is a statistical indicator used to reflect the closeness of correlation between variables, and is usually defined by the following formula:

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{3}$$

where Cov(X, Y) is the covariance of X and Y, Var(X) represents the variance of X, and Var(Y) represents the variance of Y. When the correlation coefficient between each factor is calculated, the factor with strong explanatory strength is retained. The correlation between factors and dependent

variable is shown in Figure 2.

According to the diagrams, both 600000.SH and 600036.SH keep price earning ratio, price book value ratio, price-to-sales ratio, dividend ratio, total market capitalization, opening price, high price, low price, volume of transaction, average true amplitude, Bollinger band and super trend as factors, because these factors have a high correlation coefficient with the dependent variable.

OLS model is used to make linear regression of stock price index in this study. As for 600000.SH stock, pe, pb and so forth fail to pass the T test, indicating that these factors have no significant impact on the stock price trend. The regression equation can be written in the following form:

$$y_{pred} = 13.5192 + 0.1297ps - 1.3889open + 1.9422high + 1.9890low \quad (4)$$

As for 600036.SH stock, pe, ps, atr and so forth fail to pass the T test, and the regression equation is as follows:

$$y_{pred} = 25.7859 + 0.1771pb + 0.1972dv_{ratio} - 0.2076total_{mv} - 4.0358open + 5.8293high + 4.9217low + 0.3536boll \quad (5)$$

The model is used in the train-testing of historical data from 2015 to 2020, and the results are shown in the Figure 3. It is found that this model has good fitting effect and can be used to predict stock price index. However, the limitation of this model lies in the existence of multicollinearity among factors, which will have a bad effect on the estimated parameters.

Random forest model (RF) is used to make nonlinear regression of stock price index in this study, and the results are shown in Figure 4. This model is not sensitive to multiple collinearity problems, and its results are robust to missing data and unbalanced data. It is a very friendly regression algorithm.

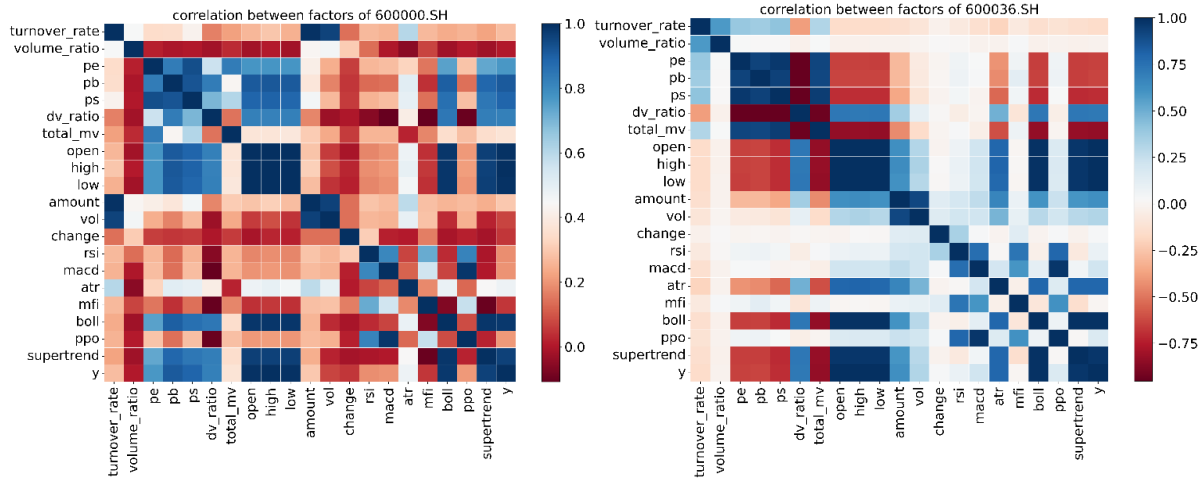



**Figure 2.** Factor correlation analysis of 600000.SH and 600036.SH stocks.

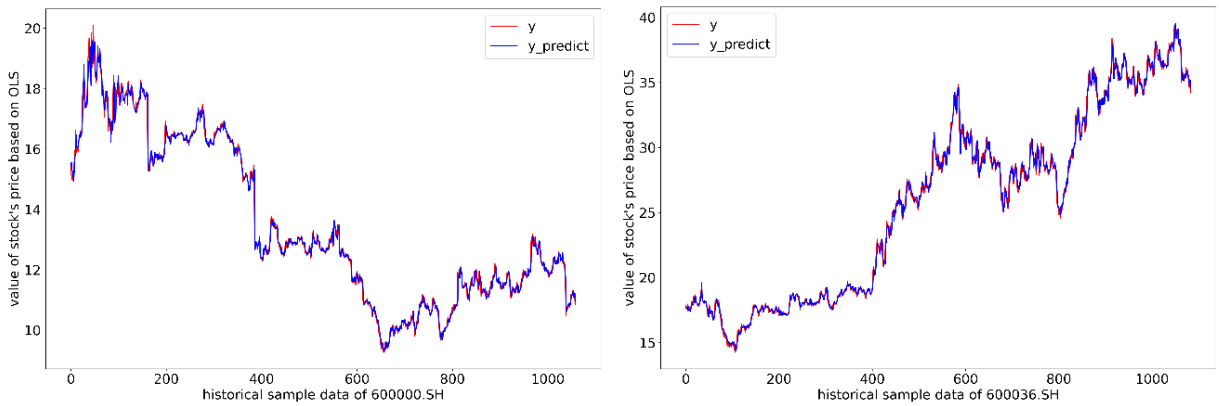



**Figure 3.** OLS linear regression analysis of 600000.SH and 600036.SH stocks (Top 1000 data).

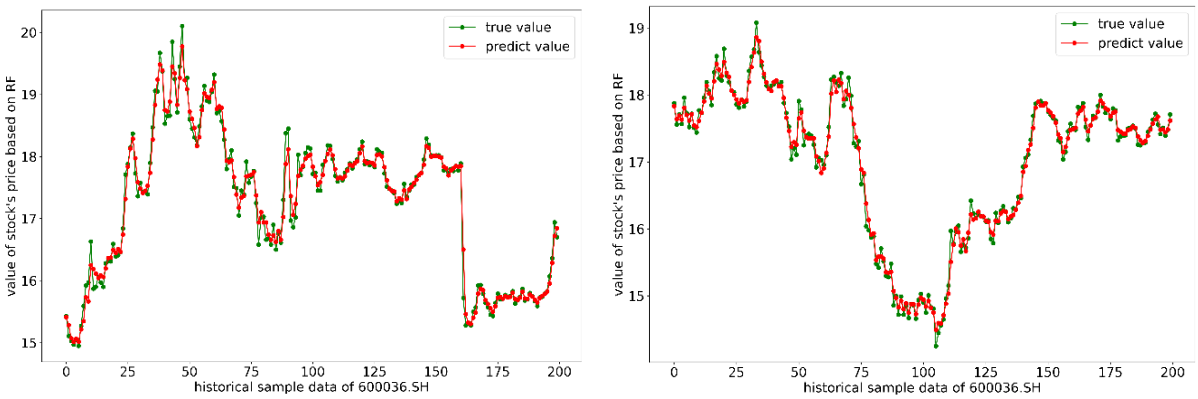



**Figure 4.** Nonlinear regression analysis of 600000.SH and 600036.SH stocks (Top 200 data).

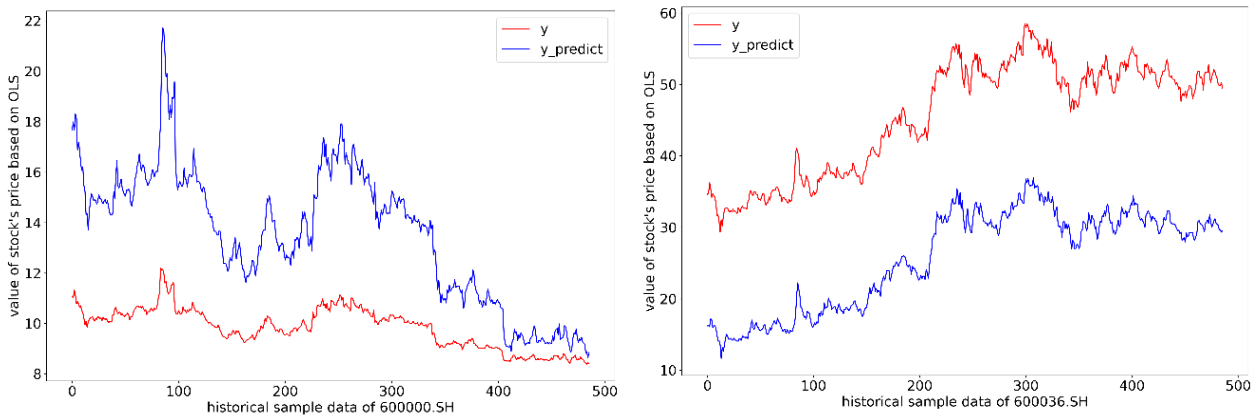



**Figure 5.** 600000.SH and 600036.SH stocks' prediction results based on OLS.

The previous subsection found through practical operation that OLS model and random forest model both performed well and could accurately fit the trend of stock prices. Therefore, this study will use these two models to predict the stock price index from March 1, 2020 to March 1, 2022.

According to Figure 5, OLS was not very effective in predicting stock prices, and the fluctuation range was relatively large. This traditional regression method can only show the linear relationship between independent variables and dependent variables, and there are many restrictions in order to obtain a better fitting effect, such as eliminating the multicollinearity problem between factors, controlling the value of DW around the number 2, avoiding the sequential correlation problems with the first order autoregressive form of random error term. As shown in Figure 6, the prediction result is pretty good.

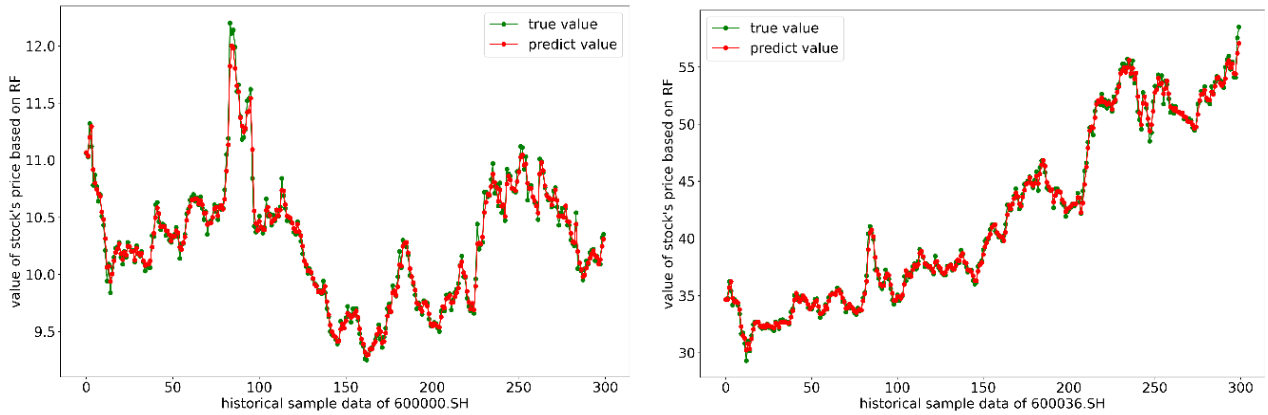



**Figure 6.** 600000.SH and 600036.SH stocks' prediction results based on RF(Top 300 data).

According to all the results, the prediction ability of random forest on stock price trend in this study is stronger than that of OLS. It reflects the trend and the number is close to the actual price of

the stock. The model has high explanatory power and fitting level.

Compared the results of 600000.SH stock and 600036.SH stock, ordinary least squares model and random forest model, the prediction ability of RF is better than that of OLS method, which reflects the inevitable trend that stock prediction will gradually expand from traditional method to machine learning. The correlation and effectiveness of factors for different stocks are disparate. Therefore, one should not directly copy the original theoretical model when constructing prediction models, but needs to find appropriate factors and regression coefficients in terms of practice.

The limitations of this study are as follows. Firstly, fundamental factors and technical factors are selected, but macro factors are ignored in the option of factors. Therefore, it is not considered comprehensively in this part. Besides, this study does not perform first-order exponential smoothing on the price series in the training set when predicting stock prices, making data lack generalization ability. These are the reasons for the large gap between the predicted value and the real value based on OLS analysis.

## 4. Conclusion

In summary, this paper finds that OLS model and random forest model developed in recent years have certain explanatory and forecasting ability for stock price index. However, predictive results of machine learning are more reliable and convincing, and the model is advanced enough that researchers can ignore the effects of multicollinearity, helping investment enthusiasts save a lot of time. Therefore, the usage of machine learning to predict stock prices has become a development trend. The study also found that the opening price, the highest price, and the lowest price of the day among the 20 factors had the most significant impact on the closing price of the next day. Each factor has a different contribution and needs to be assigned different weights in order to better explain price movements. The purpose of this study is to provide some theoretical guidance for novice investors by expounding the practical operation of multi-factor prediction of stock price volatility, which is helpful for individual investors to have a rough estimate of stock price trend. However, there are still some shortcomings in this study (e.g., a few factors selected), which reflects the lack of explanatory power. When OLS model was used, the multicollinearity between factors was not eliminated, and the influence of time period selection on prediction was not considered. These issues still need to be addressed in further research. Overall, these results offer a guideline for stock price prediction based on multifactorial linear models and the state-of-art machine learning approaches.

## References

[1] Shah, P. M. A., and C. Bhavsar. “Predicting Stock Market using Regression Technique.” Res. J. Financ. Account, vol. 6.3, 2015, pp. 27-34.

[2] D. Pei, and M. Zhu, “Stock Price Prediction Based on Multi - Factor and Multi - Variable Long and Short-Term Memory Networks.” Computer System Application, vol. 28.08, 2019, pp. 30-38.

[3] J. Li, “Stock Price Prediction Model Based on Improved Multiple Linear Regression.” Scientific and Technological Economic Market, vol. 08, 2019, pp. 61-62+64.

[4] A. Levin, “Stock selection via nonlinear multi-factor models.” Advances in Neural Information Processing Systems, vol. 8, 1995.

[5] J. Huang, and H. Liu, “Examination and modification of multi-factor model in explaining stock excess return with hybrid approach in empirical study of Chinese stock market.” Journal of Risk and Financial Management, vol. 12.2, 2019, 91.

[6] D. Pei, “Research on Multi-Factor Stock Price Prediction Method Based on Deep Learning,” 2019. University of Science and Technology of China, MA thesis.

[7] M. Yang, J. Xiao, and H. Cai, “Multicollinearity and Treatment in Multivariate Analysis.” Chinese Journal of Health Statistics, vol. 29.04, 2012, pp. 620-624.

[8] T. G. Stephan, R. Maurer, and M. Dürr. "A multiple factor model for European stocks." No. 57. working paper series: finance & accounting, 2000.

[9] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, vol. 50, 2003, pp. 159-175.

[10] H. Herwartz, "Stock return prediction under GARCH—An empirical assessment," International Journal of Forecasting, vol. 33.3, 2017, pp. 569-580.

[11] G. E. Hinton, "Inferring the meaning of direct perception." Behavioral and Brain Sciences, vol. 3.3, 1980, pp. 387-388.