Comparison of Stock price prediticon based on XGBoost and GARCH

Shikun Cui[1, *, †], Peiyang Zhao[2, †]

[1]Poole College of Management North Carolina State University Raleigh, U.S.

[2]International Education College Henan University Henan Province, China

*Corresponding author: scui4@ncsu.edu

†These authors contributed equally.

Abstract. For a long time, financial issues have been widely discussed by all sectors of society. There are many studies on the stock market, and the main purpose is mostly to predict stock prices and the overall trend of the stock market more efficiently. In this paper, the XGBoost Model and the GARCH Model are established in terms of the Shanghai Composite Index data. To be specific, the models are fitted and predicted by Python and Eviews, in order to find a better prediction mechanism. The XGBoost Model proposed in this paper is not satisfactory in terms of fitting and prediction effects, and has a certain degree of deviation. The GARCH Model shows a better performance in the short term. This research aims to find mathematical models that can effectively fit and predict the stock market through the combination of qualitative analysis and quantitative analysis. These results shed light on rationalizing the improvement of stock price prediction methods through simulation results.

Keywords: Shanghai Composite Index; XGBoost; GARCH.

1. Introduction

The analysis of financial market has always been one of the trendy topics around the world. The way an effective trading model to be built and accurately predict the trend of the financial market has always been the desire of investors and financial service industry, and one of the most practical approaches to achieve the goal is to assess the applicability of the latest technology in finance field. It is commonly known that the financial service industry is consistently the supporter and pioneer of emerging technology. From the 1960s and 1970s, when NASDAQ took the lead in introducing computer and remote communication technology for price quotation tracking system, risk assessment and intelligent trading, finance has been closely combined with new technology to present scenes where fin-tech companies use cutting-edge technologies, e.g., artificial intelligence for market analysis [1]. Meanwhile, big data analysis and machine learning are also widely used in financial service industry [2]. By using of computers and necessary coding to simulate human learning activities, in order to establish a self-learning model for analyzing the input data and predict the results [3]. Contemporarily, numerous finance firms use machine learning as a new direction for finance market prediction, as well as there are some positive outcomes follow by the machine-learning-derived new way of analysis. Some of the leading companies in financial service industry are generating profits from machine learning and big data analysis, and approximately three quarters of investment companies believe that the artificial intelligence technologies like machine learning will be the competencies of their business core [4]. It is clear that the combination between finance and technology is a solid force in the financial field, which is, to study the financial market tendency and facilitate market trading more efficiently based on the latest technologies. This paper will try to verify the applicability of the models assumed by the authors.

The references in this paper comes from Google Scholar. The paper of Arner et al. support the importance of technology in financial service industry [1]. Moreover, in the paper of Wittenbach et al., the authors show the popularity of machine learning in finance field [2]. After revealing the situation of machine learning in the financial industry, the authors find evidence from the paper of Sayavong et al. about the probable methods to predict the datasets based on historical statistics [3]. In the paper of Aziz et al., the authors claim that more of the finance corporations are attaching importance to machine learning in investment [4]. For XGBoost, the relevant interpretations of the model are mainly inspired by the paper ‘XGBoost: A scalable tree boosting system.’ [5]. The paper explains the basic principles and application scenarios of the model. Furthermore, in the paper of Kuhn et al., the authors illustrate the importance of feature engineering during the construct of XGBoost model.

The GARCH Model, known as the Generalized Arch Model, is an extension of the Arch Model, developed by Bollerslev in 1986 [6]. According to Engle and Robert, analysis and expansion of ARCH and GARCH Models has provided new ideas for many economic decisions, and many assets pricing and portfolio analysis theories have been tested and validated [7]. Klaassen found that the predictions of the GARCH Model during periods of volatility were significantly higher, and this was because in the GARCH Model's estimates, individual shocks persisted longer in volatility [8]. Hillebrand argues that it is unreasonable to use the GARCH Model to measure data volatility when parameter change-points are omitted [9].

In this paper, the research method adopts XGBoost model. According to Yun et al., the authors declare that the trading strategies based on XGBoost model performs the best back test statistics [10]. Firstly, XGBoost is used to extract the nonlinear relationship and partial linear relationship in the data. Secondly, if the arch effect exists, the GARCH model is constructed to eliminate the arch effect. After confirming the validity of the model, this paper chooses to add factors including Trading Amount, Close Price, Trading Volume, Highest Price, Lowest Price, RSI, Stochastic RSI, TRIX, ATR, MACD, WR, CCI and KDJ to predict the impact of such market and technical analysis indicators on the change of the Shanghai Composite Index. Then, we select the most effective and tested indicators combination as the input variables to train the model, and make model comparison interpretation after back testing the historical data.

2. Methodology

2.1 Data

This paper selects the most representative index of China's stock market, which is Shanghai Composite Index (Code: 000001. SH). The trend of Shanghai Composite Index can commendably reflect the emotions of the market and the overall performance of listed companies on Shanghai Stock Exchange. As the accuracy of the model is more reliable due to a long trading cycle; hence, the dataset of this paper includes 4 years trading statistics, which is the daily trading data of Shanghai Composite Index from January 2 of 2018 to April 8 of 2022. Among them, 744 sample points were selected from January 2 of 2018 to January 20 of 2021 as the training set; Select January 21, 2021 to April 8 of 2022 as the test set with total 291 sample points as shown in TABLE I.

Table I. The variable description.

Variables	Training Set	Test Set
Date	January 2, 2018	January 21, 2021
	January 20, 2021	April 8, 2022
Sample Points	744	291

According to Chen and Guestrin, the output of XGBoost is closely related to the input figures [5]. Therefore, the paper chooses 13 widely used indicators in market analysis, such as Relative Strength Index (RSI), Stochastic RSI, Triple Exponential Moving Average (TRIX), Average True Range (ATR), Moving Average Convergence Divergence (MACD), Williams Overbought/Oversold index (WR), Commodity Channel Index (CCI), Stochastic Oscillator (KDJ), amount, close price, volume, highest and lowest price during day trading.

Table II. Input and Output Variables.

Classification	Variables
Financial Technical Indicator	Relative Strength Index (RSI), Stochastic RSI, Triple Exponential Moving Average (TRIX), Average True Range (ATR), Moving Average Convergence Divergence (MACD), Williams Overbought/Oversold index (WR), Commodity Channel Index (CCI), Stochastic Oscillator (KDJ)
Price Indicator	Amount, Close Price, Volume, Highest Price, Lowest Price
Other Indicator	Date
Output Variable	SSEC (Close Price)

Creating a time series diagram is an efficient way to observe the change of data over time, which is suitable for observing the stock data. The time series statistics of Shanghai Composite Index is shown in Fig. 2. Subsequently, we calculate the basic statistical figures to have a basic understanding of the closing price of Shanghai Composite Index. The calculating results are shown in Table. II, where the mean value of the close price is 3149, the minimum and maximum values are 2464 and 3715, as well as the standard deviation is 322. The std value is high, which means the close price is highly volatile. Besides, the mean and 50% percentile value indicate that the close price during the selected period fluctuate around 3100.

In several cases, most of the raw data have missing values. The dataset of this paper collects from stock market, i.e., only trade days contain true value and the others are null. A module called pandas is applied and proceeded with python 3.10 to fill in with missing values. For this case, we use mean value of each column to replace the missing values. Furthermore, the authors reshape the origin dataset to fit the input data format of Stockstats for adding indicators to the dataset. Since the original data from Tushare is in descending order, we use pandas to form a new dataset in ascending order. After processing, the dataset has no missing values and is displayed in ascending order.

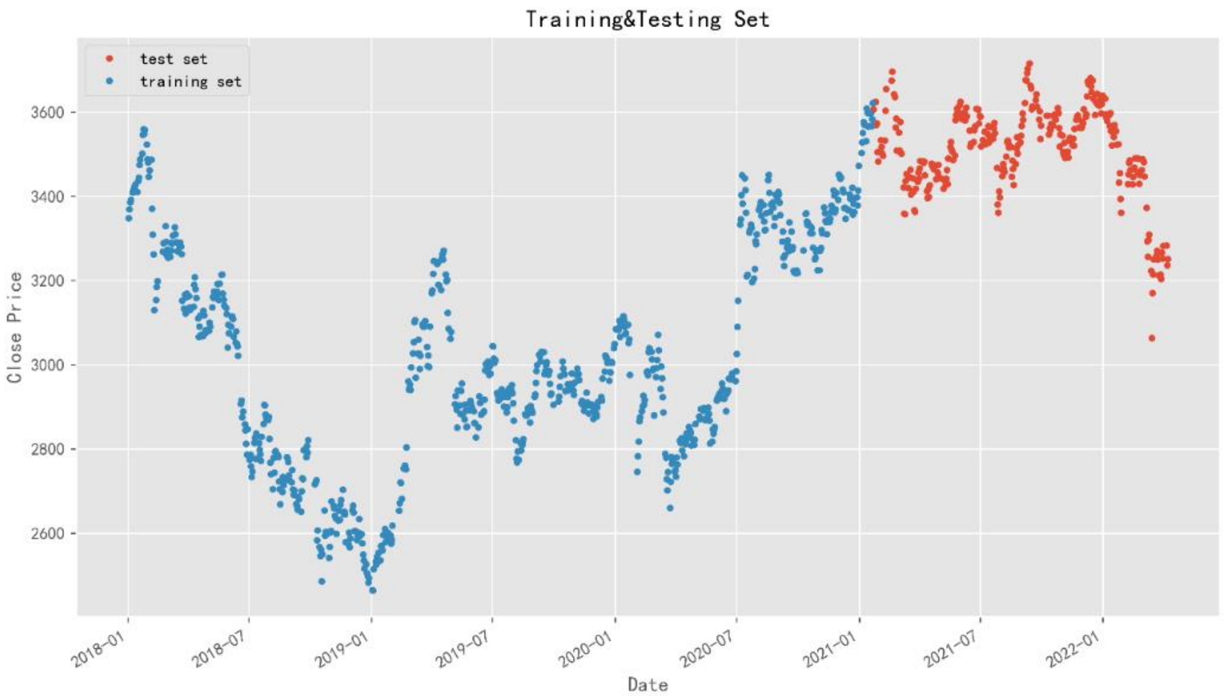

Figure 1. Training & Testing Set.

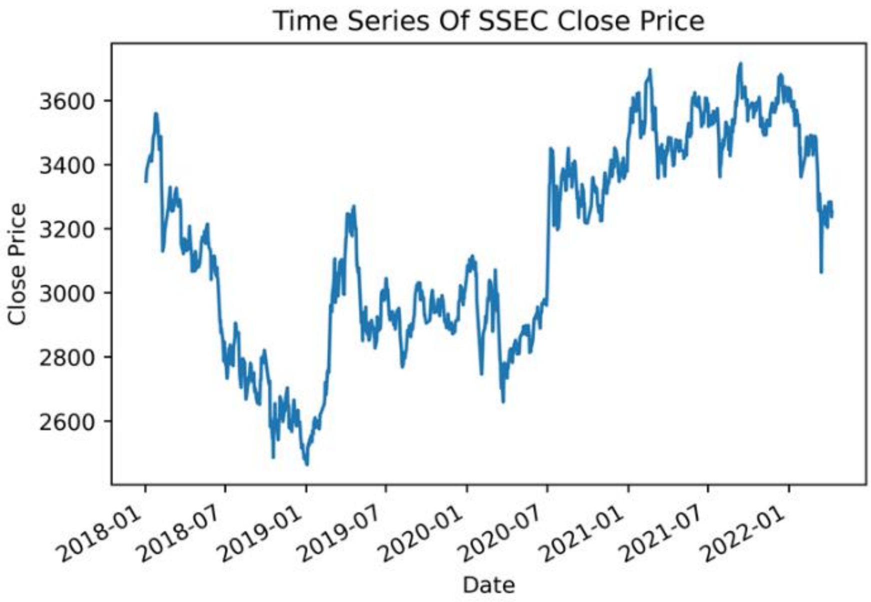

Figure 2. Time Series of SSEC Close Price.

Table III. Basic Statistical Figures of Close Price.

count	mean	std	min	max	25% percentile	50% percentile	75% percentile
1035	3149	322	2464	3715	2892	3153	3447

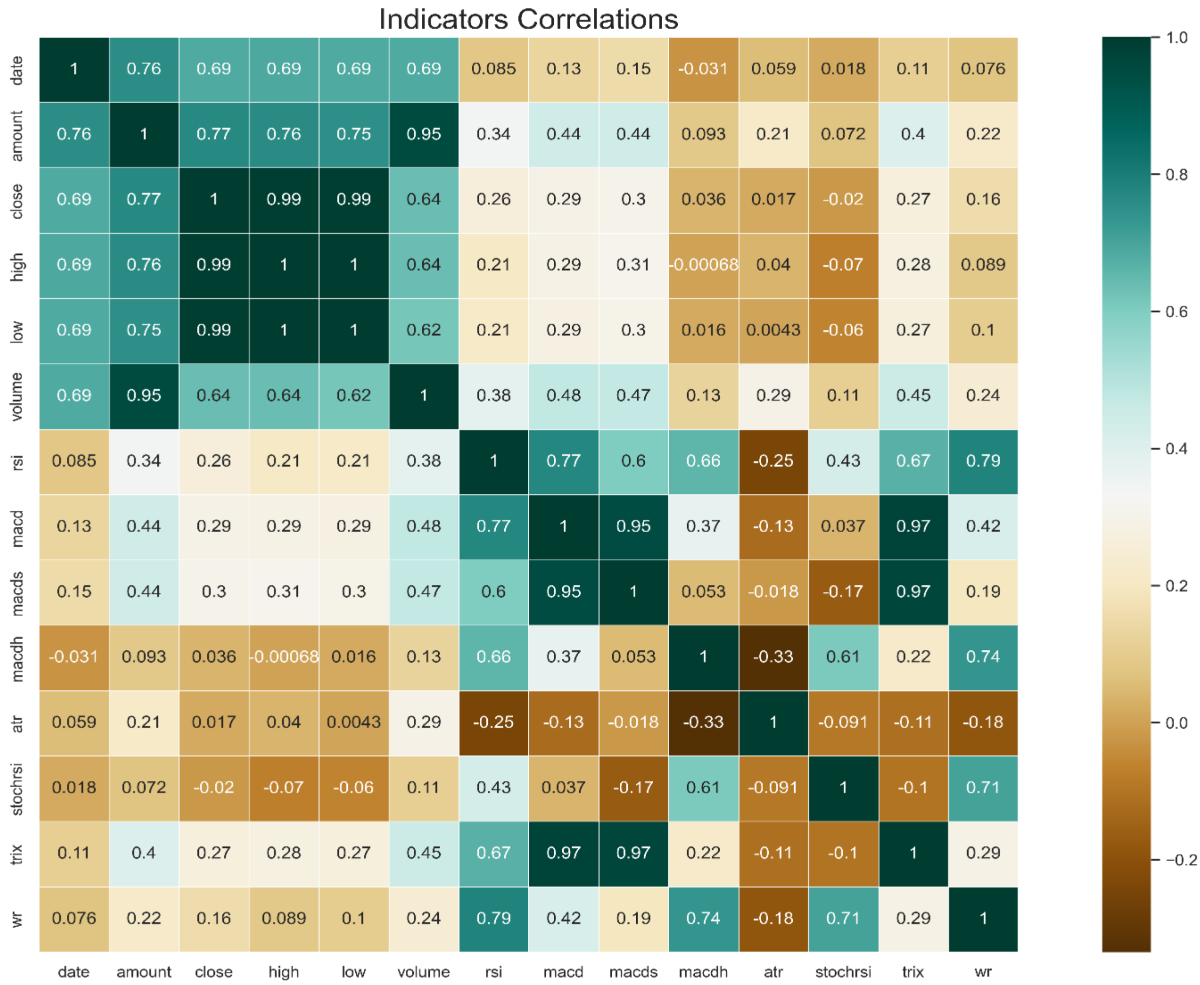

Figure 3. Correlation coefficients.

2.2 XGBoost

The XGBoost model is an open-source machine learning project proposed by Chen and Guestrin, which can fit complex nonlinear relations in data, efficiently implement GBDT algorithms, and make many improvements in algorithms and engineering. XGBoost is one of the boosting algorithms, which use raw data as a basis, train to get a minimized objective function, and use this to find the best set of parameters [5]. It uses multiple decision trees like other ensemble learning tree models, but it adds regular terms to the GBDT's objective function to help reduce the risk of overfitting. Here's how it's implemented.

The objective function of the model is:

$$obj^{(t)} = \sum_{i=1}^{n} l\left(\left(y_{i,}\hat{y}^{(t-1)}\right) + f_t(x_i)\right) + \Omega(f_t) + C \tag{1}$$

In the above equation, $\hat{y}^{(t-1)}$ represents the model prediction of the previous t-1 round, $f_t(x_i)$ is the new function, and C is the constant term. Applying the objective function to a Taylor second-order expansion and defining two new variables, one derives the translate function to:

$$obj^{(t)} \approx \sum_{i=1}^{n}\left[l\left(y_{i,}\hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) + C \tag{2}$$

When training a model, the objective function can be expressed as:

$$obj^{(t)} = \sum_{j=1}^{t}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T \tag{3}$$

Defining the function as follows:

$$G_j = \sum_{i\in I_j} g_i \,, H_j = \sum_{i\in I_j} h_i \tag{4}$$

2.3 GARCH

GARCH model, which full name is Generalized Autoregressive conditional heteroskedasticity model, proposed by Bollerslev as an extension of the ARCH model. It solves the problem posed by the second assumption of traditional econometrics on time series variables, the assumption that the variance is constant. Its model can be expressed as:

$$\begin{cases} y_t = \mu + e_t \\ e_t = \sigma_{t|t-1}\varepsilon_t \\ \sigma_{t|t-1}^2 = \omega + \sum_{i=1}^{p} \alpha_i e_{t-1}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j|t-j-1}^2 \end{cases} \tag{5}$$

In the above equation, μ is the conditional mean term, e_t is the residual term, p is the order of the ARCH model, and q is the order of the GARCH mode

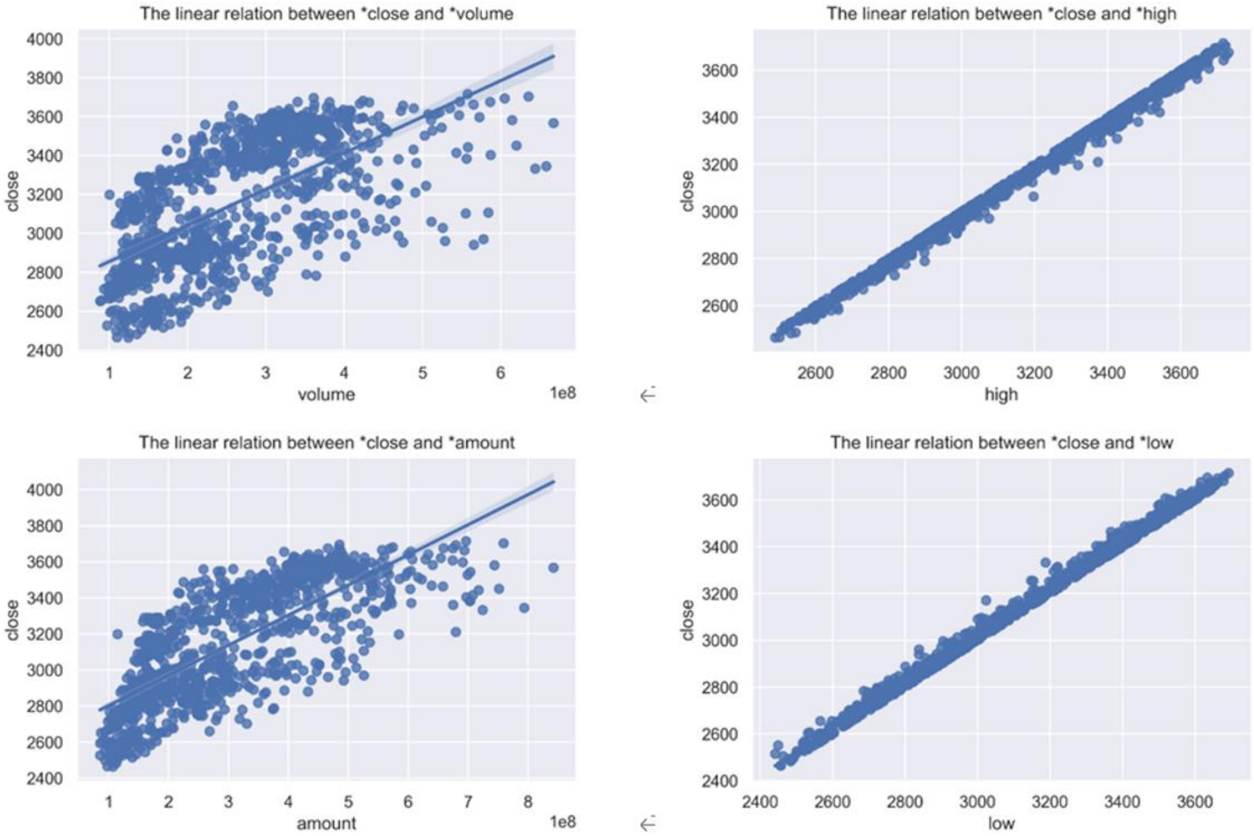

Figure 4. Scatter Plot of High Score Inputs.

3. Results & Discussion

3.1 Featured Engineering

Input variables are very important to the prediction outcome of the model, but too many variables will reduce the prediction efficiency of the model as the variables with low correlations often reduce the accuracy of the model. Therefore, it is necessary to find out the variables with high correlation coefficient when establishing the model. First, we deviated the two variables' high 'and' low 'in the original data from the original date, that is, move the data of those two columns back one day. Then, a heat map is depicted to illustrate the overall correlations between input variables, shown as Figure 3. Apparently, ‘amount’, ‘high’, ‘low’ and ‘volume’ have a relatively higher correlations with output variable ‘close’ compare to other input variables. Meanwhile, according to the correlation coefficient given in this chart, the author also makes a separate correlation analysis between several input variables with high correlation and output variables, shown as Fig. 4. Moreover, the ‘date’ has a significant score 0.69 correlate with close price, which means the close price and trade dates can be constructed to a time series structure.

Feature engineering is an indispensable part of machine learning and plays an essential role in the field of machine learning, claims by Kuhn, et al. [11]. Because too many features will reduce the training accuracy of the model, this paper selects several input variables with relatively higher correlation coefficient with the closing price as the features. In order to reduce the model defects caused by the homogenization of input variables, the authors have created new features based on the high correlation inputs, such as the difference between the highest price and the lowest price on the trading day, the average weekly closing price, the average monthly closing price, the average annual closing price, and the average highest price on the trading day, and the average price of the lowest price on the trading day.

3.2 XGBoost Model Training and Testing

When constructing the XGBoost regression model, the closing price is predicted according to the selected input features, which are ‘amount’, ‘high’, ‘low’, ‘high-low difference’, ‘dayofweek’, ‘dayofyear’, ‘dayofmonth’. Besides, the XGBregressor algorithm is used for target regression. In the stock price time series, the present daily stock price hinge on the stock price in the past several periods’ data [10]. On this basis, the fitting result of the training set can influence the prediction of testing set properly during the XGBoost construction. After fitting the training set using XGBregressor algorithm, the model output shows the MAE (mean absolute error) is 28.9459, the MSE (mean squared error) is 1633.3166, the RMSE (root mean squared error) is 40.4143, and the R2 score is 0.9832. Generally, the regression evaluation indexes that used in this model shows the model is mostly effective but not in a perfect condition. MSE value is larger than the theoretical zero deviation, which might be caused by the output variable ‘close’ because the value of stock index’s close price is relatively higher than the close price of any single stock. The prediction result shows in figure, it is clear to notice that the deviation of prediction data is insignificant when the volatility of input data is stable; which means, smoother training data could lead to the predictions more precisely. However, the stock market is not always steady as well as the volatility at some extreme moments will cause large deviation of the model.

3.3 Fitting and prediction of the GARCH Model

The GARCH Model is generalized from the Arch Model and is called the Generalized Arch Model. It eases the assumption on the model based on the Arch Model that it does not require constant variance. In this paper, the yield data is modeling and analyzed with Eviews software. Because the data used by the GARCH Model when forecasting is volatility, the yield is defined as:

$$r_t = \ln\left(\frac{y_t}{y_{t-1}}\right) \tag{6}$$

Where, r_t is the closing price variable, y_t is the time series of the yield can be calculated. In part, data from January 2,2018 to January 20,2021 are still used as historical data, and data from January 21,2021 to April 8,2022 as forecast data.

The data used in this paper is relatively simple, and the most common GARCH (1,1) Model can be directly selected for fitting. After determining the model, the data were first tested for stationarity. The Table. IV gives the results of the ADF unit root test.

Table IV. Adf unit root test.

Variable	Coefficient	Std.Error	z-Statistic	Prob
C	0.084206	0.020054	4.198896	0.0000
RESID $(-1)^2$	0.116425	0.014097	8.259033	0.0000
GARCH (-1)	0.836783	0.020381	41.05660	0.0000

Seen from Table. IV, the null hypothesis of having a unit root is rejected at the significance level of 0.01. So, the yield sequence is stable. Next, the authors fit the GARCH (1,1) model. According to the results shown in the figure above, the model fitting results are:

$$y_t = 0.7y_{t-1} + \mu_t \tag{7}$$

$$\sigma_t^2 = 0.084206 + 0.12\mu_{t-1}^2 + 0.837\sigma_{t-1}^2 \tag{8}$$

The model fit was valid since 0.12 + 0.837 <1. As illustrated in TABLE. V, the model results show that the predicted results fluctuate around the zero value and are close to the original data, so the model predicts better.

Table V. Variance equation.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RATE (-1)	-0.994255	0.031141	-31.92712	0.0000
C	0.003332	0.036193	0.092071	0.9267

Table VI. Forecast.

Forecast	RMSE	MAE
RATEF	1.010184	0.760069

According to the results of the model's predicted values, both the MAE and MSE values are low, so the model can predict the data more effectively in the short term. The outcome of the long-term prediction is unknown.

3.4 Limitation

In this research, we find that the predicted results from XGBoost deviate from the actual data when the market fluctuates significantly. In other words, large market volatility triggered by emergencies will impair the model's prediction accuracy. This flaw is unavoidable unless the training method or input data is changed from static to dynamic. However, adopting dynamic data will greatly increase the complexity and computation of the model, which would cause the model overfitting. Moreover, it is also found that uncertainty of predicting long period data in GARCH model.

4. Conclusion

In summary, this paper investigates the Shanghai Composite Index based on XGBoost and GARCH model. Precisely, to assess the applicability of model prediction. According to the test result, XGBoost performs well when the stock market is less volatile, while the GARCH Model has better fit and prediction of the price-related time series model in the short term. In the analysis, the XGBoost Model uses the closing price of the stock index as the prediction object, and the result is biased greatly when the market moves dramatically. The GARCH Model, on the other hand, takes the rate of return as the subject of study, and the predicted value is closer to the true value in the short term. In the future, it is hoped that relevant researchers will study the price trend of stock index with the method of composite model, which can better combine and engage the advantages of various models. In this paper, two of the more popular models are tested in order to illustrate the effectiveness and feasibility of the XGBoost and GARCH Models. Overall, these results offer a guideline for future research of machine learning in stock market.

References

[1] D. W. Arner, J. Barberis, and R. P. Buckley. "The evolution of Fintech: A new post-crisis paradigm." Geo. J. Int'l L. vol. 47, 2015, 1271.

[2] J. Wittenbach, B. d'Alessandro, and C. B. Bruss. "Machine Learning for Temporal Data in Finance: Challenges and Opportunities." arXiv preprint arXiv:2009.05636, 2020.

[3] L. Sayavong, Z. Wu, and S. Chalita. "Research on stock price prediction method based on convolutional neural network." 2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS). IEEE, 2019.

[4] S. Aziz, et al. "Machine learning in finance: A topic modeling approach." European Financial Management, 2019.

[5] T. Chen, and C. Guestrin. "XGBoost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016.

[6] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity." Eeri Research Paper vol. 31.3, 1986, pp. 307-327.

[7] R. Engle, "GARCH 101: The use of ARCH/GARCH models in applied econometrics." Journal of economic perspectives vol. 15.4, 2001, pp. 157-168.

[8] F. Klaassen, "Improving GARCH volatility forecasts with regime-switching GARCH." Physica-Verlag HD, 2002.

[9] Hillebrand, E. "Neglecting parameter changes in GARCH models." Journal of Econometrics vol. 129.1-2, 2005, pp. 121-138.

[10] K. K. Yun, S. W. Yoon, and D. Won. "Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process." Expert Systems with Applications vol. 186, 2021, pp. 115716.

[11] M. Kuhn, and K. Johnson. Feature engineering and selection: A practical approach for predictive models. CRC Press, 2019.