Stock Volatility Prediction Analysis Based on Different Models

Boao Yang *

Shanghai Pinghe Bilingual School, Shanghai, China

* Corresponding Author Email: yangboao@shphschool.com

Abstract. Stock volatility prediction is an important method for making the correct financial decision in modern stock market. Additionally, traditional method of stock indices prediction has been proved by modern scholars to be ineffective. This review focuses on the volatility prediction performance of various models, aiming to figure out the most effective volatility prediction model for modern stock market. The four models discussed in this review are ARIMA, XGBOOST, GARCH, and LSTM. Concluding the effectiveness of different models in volatility prediction can be helpful for practical applications of these models in modern stock markets, which can improve the profitability of stock investors under certain situations. Among all four models discussed in this review, LSTM model shows the best performance among all four models in stock volatility prediction. The unique machine learning ability of LSTM has great potential value in future science field research. Since this study is only a general review of stock volatility prediction models, additional research and analysis are essential in order to further conduct this topic for better accuracy and credibility.

Keywords: machine learning; stock volatility prediction; ARIMA; XGBOOST; GARCH.

1. Introduction

In stock markets, volatility fluctuation is a common phenomenon where high volatility patterns are usually followed by periods of low volatility and vice versa. Forecasting stock volatility of modern stock markets is essential for the right financial decision-making. This review investigates the performance of the Auto Regressive Integrated Moving Average (ARIMA) model, Extreme Gradient Boosting (XGBOOST) model, Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model, and Long-Short Term Memory (LSTM) model in the predicting process of modern stock volatility. Evaluating the effectiveness of different models in volatility prediction can be helpful for practical applications of these models in modern stock markets, which can improve the profitability of stock investors under certain situations. As a classic time, series forecasting model, ARIMA is constituted by three components: autoregressive (AR), moving average (MA) and differencing(I); moreover, this model shows excellent ability on capturing linear relationship between the variables [1]. XGBoost is a strong machine learning algorithm designed for achieving successful regression and classification. Gradient boosting is an ensemble learning technique that combines a number of weak learners (often decision trees) to produce a powerful prediction model. This model has advanced from gradient boosting [2]. GARCH is a statistical model used for forecasting the volatility of financial time series data. This model specifically addresses the presence of heteroskedasticity, which refers to the changing variability of a variable over time [3]. Therefore, the GARCH model can be assigned to situations where stock volatility has been fluctuating over time. An RNN architecture known as the LSTM model was created to solve the problems associated with learning long-range dependencies in sequential data. Unlike traditional RNN models, LSTM is able to capture and learn information over extended time series, making the predictions more reliable for tasks that involve context and temporal patterns [4]. The review's subsequent sections are organized as follows: Section two is divided into four parts and each part introduces the basic principles of a model and suggests empirical analysis based on experiments. Section 2.1 evaluates ARIMA; section 2.2 evaluates XGBOOST; section 2.3 evaluates GARCH; section 2.4 evaluates LSTM. Section 3 concludes the discoveries in this review and proposes suggestions for this review's further developments. Section 4 lists all the articles that have become references in this study.

2. Stock Index Prediction Analysis Based on Different Models

2.1. Autoregressive Integrated Moving Average Model

2.1.1 Introduction to ARIMA

ARIMA is a widely used forecasting model that is suitable for analyzing and forecasting data with trend or seasonality. In ARIMA, AR is an auto-regressive component, meaning the number of lagged values comprised in the mode, 'I' am integrated component, implying the procedure of making the data stationary, MA is the moving average component, indicating the number of errors in the calculation process. The mathematical nomenclature for the ARIMA model is ARIMA (p, d, q), where p is the auto-regressive component, d denotes integration, and q denotes the moving average component [5]. Here is the mathematical expression of the ARIMA model [1].

$$Y_t = \emptyset_0 + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

Where Y_t represents the true observations, ε_t represents random error at t, $\emptyset_i$ represents autoregressive coefficients, and θ_j represents moving average coefficients.

2.1.2 Application of ARIMA

The study of Wang investigates multiple methods of forecasting market volatility and stock prices based on different models [6]. This study mainly compares the performance of hybrid models and single models, based on the predicting accuracy. The researcher first introduces the data of ten stocks that have been used in this research and decomposes the data into two parts: approximate part and error part, which can also be represented by the basic formula $Y_t = L_t + N_t$ where Y_t is the raw data, L_t is the approximate, and N_t is the error [6]. In the experiment, the researcher proposes to use ARIMA model for the prediction of L_t and the XGBOOST model for the prediction of N_t. Moreover, the researcher adds Discrete Wavelet Transform(DWT) with ARIMA and XGBOOST together as a new hybrid model in order to achieve better performance on stock volatility prediction. The study employs the mean absolute percent error (RMSE), mean absolute error (MAE), determination coefficient, the area under the curve (AUC), and percentage accuracy to evaluate the prediction findings in order to demonstrate the viability of the hypothesis. Here is a table of the volatility prediction comparison based on several remarkable models discussed in the study [6].

Table 1. ARIMA Model Stock Volatility Prediction Performance

	RMSE	MAE	P SQUARE	AUC	ACCURACY
ARIMA(0, 1, 1) model					
CNPC	0.67501	0.7871	-2.0283	0.54582	54.55%
PING AN	471.001	21.2465	-54.547	0.48506	48.48%
KMCL	134707	365.869	-45.662	0.46481	46.46%
SAIC MOTOR	108.213	10.3056	-29.141	0.53518	53.54%
HT-FOOD	1027.71	31.2866	-71.497	0.43428	43.43%
CSCEC	6.39467	2.12402	-3.6265	0.56591	56.57%
S.F	190.09	13.5408	-21.072	0.46481	46.46%
Fuyao Group	128.236	10.5442	-18.771	0.54551	54.55%
CASTECHINC	5.39963	2.0152	-5.7946	0.52512	52.53%
CITIC	1.6336	0.92211	-1.9490	0.57591	57.58%

Table 2. XGBOOST Model Stock Volatility Prediction Performance

	RMSE	MAE	P SQUARE	AUC	ACCURACY
XGBoost model					
CNPC	0.00539	0.053	0.97581	0.73261	73.74%
PING AN	0.54991	0.57517	0.93514	0.72258	71.72%
KMCL	81.3232	7.08542	0.97182	0.64814	63.64%
SAIC MOTOR	0.13253	0.29011	0.96308	0.7549	75.76%
HT-FOOD	2.27602	1.1452	0.83944	0.56408	56.57%
CSCEC	0.0522	0.18098	0.96222	0.75714	75.76%
S.F	0.3215	0.41549	0.96266	0.56362	55.56%
Fuyao Group	0.2236	0.37932	0.96552	0.65714	65.66%
CASTECHINC	0.02889	0.13483	0.96363	0.71384	71.72%
CITIC	0.03054	0.13427	0.9449	0.72775	72.73%

Table 3. Hybrid Model Stock Volatility Prediction Performance

	RMSE	MAE	P SQUARE	AUC	ACCURACY
DWT+ ARIMA(0, 1, 1)+ XGBoost model					
CNPC	0.00237	0.0374	0.98932	0.8572	85.86%
PING AN	0.3565	0.47561	0.95795	0.82835	82.83%
KMCL	36.3691	4.89102	0.9874	0.77777	77.78%
SAIC MOTOR	0.05886	0.19827	0.9836	0.80053	79.80%
HT-FOOD	0.48033	0.56755	0.96611	0.80795	80.81%
CSCEC	0.02662	0.13392	0.98073	0.77734	77.78%
S.F	0.14146	0.26035	0.98357	0.83694	83.84%
Fuyao Group	0.23138	0.40342	0.96432	0.73734	73.74%
CASTECHINC	0.0266	0.13404	0.96652	0.82843	82.83%
CITIC	0.02483	0.1265	0.95516	0.80775	80.81%

The information in the table leads to the conclusion that the hybrid model DWT-ARIMA-XGBOOST has the best performance on stock volatility prediction among three models because of its high accuracy percentage value. Among the other two single models, the XGBOOST model has a better ability on stock volatility prediction and ARIMA model shows a relatively weaker ability on prediction. Therefore, it can be concluded that ARIMA model needs further improvements in the field of stock volatility prediction.

2.2. Extreme Gradient Boosting Model

2.2.1 Introduction to XGBOOST

XGBOOST is defined as a supervised machine learning algorithm, functioning for analyzing regression and classification tasks. This method was first published by Tianqi Chen and Carlos Guestrin in 2016, and it has become commonly used due to its efficiency and accuracy in various prediction fields [7]. This model enhances multi-threading by focusing on processing features individually, employing prefetching technology to independently initiate a thread for data reading, and proactively loading the subsequent data into memory [8]. XGBOOST can be expressed mathematically in the following formula. The number of trees is illustrated by k, the space of each Regression Tree is represented by F, and the i_{th} predicted rate is indicated by $\hat{y_i}$ [2].

$$\hat{y_i} = \sum_{k=1}^{n} f_k(x_i)\,, f_k \in F \tag{2}$$

Additionally, the following formula can be used to define the goal function of XGBOOST. In this case, $L(\theta)$ represents the loss function and $\Omega(\theta)$ represents the regularization parameter [2].

$$\mathrm{Obj}(\theta) = L(\theta) + \Omega(\theta) \tag{3}$$

The $L(\theta)$ and $\Omega(\theta)$ can be expanded as the following [3].

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i)\ , \Omega(\theta) = \sum_{k=1}^{k} \Omega(f_k) \tag{4}$$

To optimize the objective function swiftly, the second Taylor expansion is enlisted on the loss function [2].

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)})] + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_k) \tag{5}$$

Here g_i and h_i can be expanded:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}), h_i = \partial^2_{y_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{6}$$

2.2.2 Application of XGBOOST

The research conducted by Doctor Zhu investigates the effects of machine learning models on forecasting stock volatility based on the data for 112 stocks offered by KAGGLE [10]. This study combines the Light Gradient Boosting Machine model (LightGBM) and Extreme Gradient Boosting model (XGBOOST) into a new predicting strategy, comparing with predicting results from other single models. According to the experiment, the calculated results of the fusion strategy that combines LightGBM and XGBOOST is the weighted average of the two models with 40% from XGBOOST and 60% from LightGBM [8]. To train the fusion model, the researchers use the K-means algorithm in order to help the memory of volatility in a multidimensional way. Then the study implements Naive Bayes, neural network (NN), support vector machine (SVM), and XGBOOST on same set of data for further discussion of the fusion model's performance. In the assessment of the performance of volatility prediction, the study uses Root Mean Square Percentage Error (RMSPE) as the evaluation index, which is proved to be more susceptible to the magnitude of errors. Here is the result of the evaluation.

Table 4. Volatility Prediction Accuracy of Different Models

Model	RMSPE
Naive Bayes	0.410
NN	0.275
SVM	0.298
XGBOOST	0.235
XGBOOST+LightGBM	0.230

The graphic leads to the conclusion that, compared to other single method-based algorithms, the fusion algorithm based on XGBOOST and LightGBM exhibits greater ability on stock volatility prediction. Furthermore, among all the single method strategies, the XGBOOST model has the lowest value of RMSPE, and this method's RMSPE value is 14.5% less than NN's value. Thus, the XGBOOST model is a well-performing machine learning algorithm as a single indicator.

2.3. Generalized AutoRegressive Conditional Heteroskedasticity Model

2.3.1 Introduction to the GARCH model

The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model was invented by Bollerslev in 1986 for volatility prediction, and variants of this model arekept updated by scholars who aim to achieve better volatility prediction performance.[3] The GARCH model consists of autoregressive and conditional variance equations. The values of parameters are estimated through maximum similarity estimation, and the model is used to predict future volatility based on historical data. This is the mathematical expression of GARCH [9].

$$y_t = \sigma_t \varepsilon_t \tag{7}$$

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{8}$$

Where y_t is the observed return at time t, ε_t is a white noise error term with mean zero and normally distributed variance, σ_t^2 is the squared volatility at time t, ω is the constant baseline volatility, α is the coefficient of lagged squared returns, and β is the coefficient of the lagged squared volatility. GARCH model shows great potential in economics for risk management, portfolio optimization, and option pricing because it helps users predict accurate changes in future stock volatility.

2.3.2 Application of GARCH model

The study proposed by RONI uses various types of GARCH models to address the relationship between stock returns and stock index volatility in the Bangladesh stock market [9]. The study looks at the data of daily closing prices from the Dhaka Stock Exchange (DSEX). The DSEX index is processed to index returns, tested for heteroscedasticity, and checked for stationarity before implementing the model. In the experiment, the study proposes four different GARCH-based models: GARCH model, GARCH-M model (risk premium parameter included), TARCH model (GARCH includes threshold parameter), and EGARCH model (allow positive and negative shocks). In this experiment, the evaluation procedure is split into four time periods: Full, Pre-Crisis, Crisis, and post-crisis. Theil inequality coefficient (TIC), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE) are used to evaluate the error statistics. The model accuracy is evaluated by minimum Akaike Information Criteria (AIC), minimum Schwarz Information Criteria (SIC), and maximum Log-likelihood (LL). Here shows the two tables from researchers that present the models' performance from forecasting error statistics and fitted model accuracy directly in the form of rankings [9].

Table 5. Error Statistics-based Forecasting of Predicted Performance of Several Models

Time	Model	Rank
Full	GARCH	1
	GARCH-M	4
	TGARCH	2
	EGARCH	3
Pre-Crisis	GARCH	1
	GARCH-M	4
	TGARCH	2
	EGARCH	3
Crisis	GARCH	4
	GARCH-M	1
	TGARCH	3
	EGARCH	2
Post Crisis	GARCH	3
	GARCH-M	4
	TGARCH	2
	EGARCH	1

Table 6. Prediction Results of Several Models Using Model Accuracy

Time	Model	Rank
Full	GARCH	4
	GARCH-M	2
	TGARCH	1
	EGARCH	3
Pre-Crisis	GARCH	3
	GARCH-M	1
	TGARCH	4
	EGARCH	2
Crisis	GARCH	3
	GARCH-M	4
	TGARCH	1
	EGARCH	2
Post Crisis	GARCH	3
	GARCH-M	4
	TGARCH	2
	EGARCH	1

From the diagram, it can be concluded that there exists no single model that can accurately fit each and every series of volatility in the Bangladesh stock market. Although the calculated data shows little difference between the four models, the ranking can evaluate these models' performance in a more objective way. The rankings demonstrate that the GARCH model performs best in terms of estimating stock future returns, which are represented by error statistics, indicating that GARCH is the most effective algorithm for predicting stock volatility and returns in the DSEX index among all four discussed models. In the measurement of model fitness, which is represented by model accuracy, the rankings show that TGARCH fits the data series best, illustrating its accuracy on the data description of DSEX. In sum, among the four GARCH models studied in the research, each model has its advantages and disadvantages. For future developments, the combination of multiple algorithms may produce better GARCH family models that can both fit historical data and predict future indices volatility well, especially GARCH and TGARCH models.

2.4. Long-Short Term Memory Model

2.4.1 Introduction to LSTM

LSTM was invented to solve the unprecedented problems traced in neural networks, such as error backflow, gradient descent, and data memory [10]. This model can be applied to various cases and is widely used in index prediction and volatility prediction. The structure of LSTM model contains three parts: input gate, forgetting gate, and output gate [4]. The mathematical expression of the final output from LSTM model can be expressed by the following formulas [4].

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

Here o_t indicates the forgetting rate, W_o represents the weight of the output gate, b_o shows the bias of the output gate, h_t indicates the hidden layer's output, and C_t indicates the current unit state value in the hidden layer.

2.4.2 Application of LSTM

Using raw data from the Shanghai Composite Index (SSEC), Shenzhen Component Index (SZSE), and Industrial and Commercial Bank of China stock price index (ICBC), the study from ShuiLing Yu and Zhe Li compares the efficiency of the GARCH model with the LSTM model. The scholars' research in data processing focuses on optimizing volatility series. [10]. The common method to determine stock volatility is to calculate the square of return rate; however, this traditional approach may not yield results with ideal accuracy. Therefore, the researchers substitute the volatility series with the real volatility theory proposed by Parkinson and normalize the whole data set to (0,1) [10]. In the experiment, the data set is divided into two sections: 70% for model training and 30% for model testing. The volatility prediction of the two models is plotted by R and examined by loss functions. From the plotted graph, observation indicates that both LSTM and GARCH perform well in volatility prediction; nonetheless, the LSTM model performs better fitness to the true volatility where the data fluctuates to a great extent. In the evaluation stage, four functions were given in the study: the normalized mean squared error (NMSE), the normalized mean absolute error (NMAE), the logarithmic loss function (LL), and the exponential loss function (LINEX). Here is a thorough table of the evaluation's findings.

Table 7. Volatility Predicted Errors of LSTM and GARCH based on Different Indices

Stock index	Model	NMSE	NMAE	LL	LINEX
SSEC	LSTM	0.6257	0.7525	0.8631	5.1346
	GARCH	0.9377	1.0647	1.1879	7.9324
SZSE	LSTM	0.7251	0.7968	0.8094	7.1968
	GARCH	1.0635	1.2053	1.3492	8.1872
ICBC	LSTM	0.8357	0.9345	0.9763	7.8592
	GARCH	1.1523	1.0549	1.2621	8.5837

According to the results, LSTM model has a better performance than GARCH model on every category of data sets; however, the LSTM model calculates slightly larger volatility discrepancy in ICBC than in SSEC and SZSE. In sum, it can be concluded from this experiment that LSTM model is a practical model that has well performance in the field of stock volatility prediction.For future developments, since the predictive ability of LSTM model on individual stocks is slightly worse than that of the indices, future research can focus on improving LSTM mode’s predicting performance on single stocks.

3. Conclusion

This review discusses the volatility prediction performance of ARIMA model, XGBOOST model, GARCH model, and LSTM model in modern stock market. ARIMA model is compared with XGBOOST model. Among these two single models, XGBOOST model shows better accuracy on stock volatility prediction and ARIMA model shows relatively weaker ability on prediction. The reason for this result may be the assumption in ARIMA that all the variables are linear factors, which ignores the possibility of nonlinear relationships. To improve ARIMA’s performance, designing hybrid model based on ARIMA may be an optimal approach. XGBOOST model is compared with LightGBM model and hybrid models. The hybrid model based on XGBOOST and LightGBM shows better performance on stock volatility prediction than other single method-based algorithms; however, among all the single method strategies, XGBOOST model has the least value of errors, indicating that this model is a well-performance machine learning algorithm as a single indicator. The GARCH model is compared with the other three ARCH family models (GARCH-M, TGARCH, EGARCH). When calculating stock volatility and returns, GARCH model has best performance, indicating that GARCH is the most effective algorithm for predicting stock volatility and returns among all four GARCH family models. In the measurement of model fitness, TGARCH fits the data series best. For further improvements in ARCH family models, a hybrid model based on GARCH and TARCH may be a possible direction. Based on the GARCH model, the LSTM model is compared. The outcome demonstrates that LSTM has significantly increased the accuracy of volatility prediction; however, LSTM has a weaker ability to predict single stock volatility than indices volatility, which may need to be solved by developing hybrid models based on LSTM. In conclusion, this review clearly points out the characteristics of the four models and discusses their performances. LSTM model shows the best performance among all four models in stock volatility prediction. LSTM's unique ability of machine learning has great potential value in future research. This review only discusses the performance of discussed models under particular stock indices, which may not provide exactly accurate judgement on the previous models. To further conduct this review, research and experiment on the same stock indices are necessary, so that more credible assessments can be reached.

References

[1] Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model[C]//2014 UKSim-AMSS 16th international conference on computer modelling and simulation. IEEE, 2014, 106-112.

[2] Eslami Ghayour A, Yousefi A. Performance analysis of gradient Boosting Models for forex market prediction. Available at SSRN 4450688.

[3] Forecasting Bitcoin Volatility Using Hybrid GARCH Models with Machine Learning.

[4] Liwei T, Li F, Yu S, et al. Forecast of lstm-xgboost in stock price based on bayesian optimization. Intell. Autom. Soft Comput, 2021, 29(3): 855-868.

[5] Idrees S M, Alam M A, Agarwal P. A prediction approach for stock market volatility based on time series data. IEEE Access, 2019, 7: 17287-17298.

[6] Wang Y, Guo Y. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. China Communications, 2020, 17(3): 205-221.

[7] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, 785-794.

[8] Zhu D. Predicting stock volatility based on weighted fusion model of XGBoost and LightGBM[C]//2022 3rd International conference on computer vision, image and deep learning & international conference on computer engineering and applications (CVIDL & ICCEA). IEEE, 2022: 65-67.

[9] Roni B, Wu C, Jewel R K, et al. A study on the volatility of the Bangladesh stock market—based on GARCH type models. Journal of Systems Science and Information, 2017, 5(3): 193-215.

[10] Yu S L, Li Z. Forecasting stock price index volatility with LSTM deep neural network[C]//Recent Developments in Data Science and Business Analytics: Proceedings of the International Conference on Data Science and Business Analytics (ICDSBA-2017). Springer International Publishing, 2018, 265-272.