Heterogeneity in Stock Price Forecasting-Based on the ARIMA-GARCH Model and PCA-LSTM Model

Sirui Cheng *

Department of Economics, Nankai University, Tianjin 300071, China,

* Corresponding Author Email: 2013011@mail.nankai.edu.cn

Abstract. Financial theory suggests stock prices are mainly influenced by factors such as interest rates, market behavior, technical indicators, and firm value. Traditional approaches to stock price forecasting have been augmented by machine learning algorithms and time series models. Deep learning has experienced rapid development in the field of time series analysis and is becoming more mature. Therefore, this article delves into two prominent prediction methods: the ARIMA-GARCH model and Long Short-Term Memory network to compare their prediction performance and to analyze the heterogeneity in the effectiveness evaluation. The combination of the ARIMA and GARCH models makes it possible to account for both short-term fluctuations and long-term trends. And LSTM networks can capture the temporal relationships and diversified features of data, making it a popular choice for financial time series analysis. This study examines the forecasting of daily closing prices of representative individual stocks in Shanghai and Shenzhen A-shares and analyzes the factors that may cause heterogeneity in the prediction results. In addition to being affected by the chosen forecasting indicators, the differences in the forecasting results of individual stock prices can also be related to the company's economic information. Differences in the market structure of the industry and the effects of policy implementation can also lead to differences in forecasting results. These heterogeneity analyses provide references and suggestions for selecting models and variables in the financial field to improve prediction effectiveness.

Keywords: Stock forecasting; deep learning; heterogeneity; ARIMA model; LSTM.

1. Introduction

The securities exchange assumes a more significant part with the rising extent of money in the public economy. In this way, the interest for related monetary administrations encounters an increment, and stock cost forecasting has turned into an issue that financial backers join extraordinary significance to. Deviating from the original trend, unexpected events in the market often lead to huge fluctuations in the stock price, which makes the stock price a non-stationary and high-noise type of data. Therefore, it is challenging to make effective predictions of the stock market using relevant data.

Traditional stock price forecasting methods include time series models such as Autoregressive Integrated Moving Average (ARIMA) and BandH. With a relatively fixed process, these traditional methods are convenient to deal with general situations in the stock market. However, Hiemstra and Jones believe that simple linear regression models are unable to accurately predict the market under sudden shocks [1]. And the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model can better analyze the volatility of data. With the development of artificial intelligence technology, many neural network models have been applied to the study of nonlinear and non-stationary financial time series [2]. Based on the basic information of listed companies, Atiya et al. applied the neural network model to make predictions and concluded that these methods could expand the scope of objects and analyze special situations specifically [3]. Xie et al. applied the Recurrent Neural Network model (RNN) to make an empirical study of the China stock market [4]. It is proved that the RNN model can provide valuable information for investors to react and make stock transactions more rationally. A sequence-to-sequence (Seq2Seq) machine learning model was proposed by Sutskever et al. to address the issue of the input and output value's unpredictable length. This approach allows for complete coverage of the variables influencing stock price variations [5]. Based on the neural network, the method of deep learning is added to the research and analysis. Nelson used Long Short-Term Memory (LSTM) neural networks to forecast stock prices with

historical prices and technical analysis indicators. He concluded that this model's predictive effectiveness was noticeably higher than that of the prior artificial intelligence model [6]. DiPersio et al. compared the differences between three different models (multi-layer perceptron, long short-term memory network, and convolutional neural network model) when studying the same data set [7]. Fischer et al. used the S&P500 data set to conduct analysis and research on the basis of the LSTM network. The conclusion is that the prediction errors of these models are smaller with more accurate parameters [8].

However, in addition to the different prediction models mentioned above, many other indicators also have impacts on the prediction results. For example, the different industries of companies may make sense. Therefore, the differences in these macro and micro economic factors (such as financial statements, market news and public opinions, and changing trading behavior) may lead to a significant difference in the same model. In different models, these economic factors may also lead to different results. The discussion of these heterogeneity problems may be helpful to improve the economic significance of stock price forecasting models.

When choosing the indicators to describe the stock data, there can be a certain correlation between different factors, leading to an overlap. Principal Components Analysis (PCA) can transform multiple indicators into several unrelated comprehensive indicators with little information reduced, thus avoiding overfitting and improving the efficiency of analysis [9]. The daily stock price of nine stocks in five industries on the China A-share market will be predicted in this study, using the ARIMA-GARCH technique and PCA-LSTM method and the historical data from January 2022 to July 2023. On this basis, this study will compare the prediction accuracy of different industries and different methods. Five stocks in the pharmaceutical industry will be selected to further analyze whether economic factors (the company's market value and stock price are selected in this paper) have impacts on the accuracy. In this study, the historical stock data will be standardized and differentiated and scores of each principal component will be obtained. After the stationarity test, the ARIMA-GARCH method is used to predict. After setting other parameters of the model, the principal component scores will be input into the LSTM model for training and prediction [10]. Finally, this study will compare and analyze the heterogeneity of the results to draw conclusions and give suggestions.

2. Methods

2.1. Source of Data

This study focuses on forecasting and comparative analysis of China's A-share market. The dataset ranges from January 4, 2022, to July 28, 2023. All data are sourced from the wind database.

Table 1. The stock code and industry of each company

Stock Name	Stock Code	Industry Type
Dong-E-E-Jiao	000423	Biopharmaceutical
Changchun High-Tech	000661	Biopharmaceutical
Renhe Pharmaceuticals	000650	Biopharmaceutical
Zhifei Biology	300122	Biopharmaceutical
Tianxin Pharmaceuticals	603235	Biopharmaceutical
Lion Electronics	605358	Semiconductor
Bojun Technology	300926	Automotive
Huafa Holding	600325	Real Estate
Yingjia Gongjiu	603198	Wine

To ensure the objectivity of the study, the representative stocks studied in this paper cover a relatively diversified range of industry types, with other characteristics such as large size and good liquidity. The selection of stock sectors includes emerging industries linked to technological development, national industries that are highly influenced by macro policies, and traditional Chinese

industries that are more susceptible to the economic environment and brand reputation. As a result, five stocks in the Biopharmaceutical sector were chosen, four of which have a market capitalization of more than RMB 200 million, and three of which have a share price of about RMB 50 per share at the cutoff time of the dataset. A total of five sectors and nine stocks were selected for the dataset, and the details are shown in Table 1.

2.2. Indicator Selection

The predictor variables for this study include date, open, close, high, low, volume, amount, and range (change and percent-change). The trading date of the stock serves as the time series' index. Other attributes are included into the database as model independent variables. he closing price is output as the only prediction.

When selecting evaluation indicators that can objectively reflect the strengths and weaknesses of the model, this study considers three indicators: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the accuracy of up/down study. The error between the projected and actual values of the closing price is calculated using the first two evaluation indicators. The third indication assesses if the actual rise or decrease is compatible with the projected one by comparing the closing price to the day's beginning price. By computing the proportion of projected consistent results in all data, one may assess the validity and accuracy of the model's qualitative predictions of volatility patterns.

2.3. Introduction to Methods

The ARIMA-GARCH model combines two basic techniques. The ARIMA is a linear model for smooth series, capable of identifying and capturing underlying patterns like trends and seasonality in time-series data. For financial data with an exponential growth trend, it is usually necessary to differentiate the series to ensure its smoothness before modeling. While the GARCH is a moving average model based on the square of residual, it focuses on modeling volatility and considering conditional heteroskedasticity. It can model the noise term of the ARIMA model, combining new information and analyzing the series according to conditional variance.

PCA utilizes orthogonal transformations to transform multiple variables into fewer composite variables that are linearly uncorrelated. It is a variable dimensionality reduction method, which mainly includes the following steps.

The first step is to standardize the raw indicator data:

$$y_{standard} = \frac{y_i - \min(y)}{\max(y) - \min(y)} \quad (1)$$

Then it calculates the sample correlation coefficient matrix and solves the characteristic equation to get the eigenvalues and eigenvectors. After obtaining the number of principal components, the corresponding eigenvectors are formed into a matrix.

Long Short-Term Memory (LSTM) network is a variant of a Recurrent Neural Network (RNN), intended to resolve the issue of slope disappearing in conventional RNNs. LSTM adds an input layer, an output layer, and a forgetting layer to the traditional recurrent neural network. In this project, the LSTM model will be built based on the deep learning framework in Matlab.

Results and Discussion

2.4. ARIMA-GARCH Method

Rstudio is used in this paper to model and analyze the ARIMA-GARCH method. Based on the autocorrelation plots and partial autocorrelation plots of the time series, utilizing a unit root test to determine the smoothness of the series, the conduct of difference operation is found to be necessary. All the time series pass the smoothness test after one or two differencing operations and none of them are white noise, which means that the ARIMA model is suitable for modeling these series. After ordering and parameterizing the model using the minimum AIC criterion, the residuals of the model are further subjected to the white noise test and ARCH test, and the GARCH model is found necessary

for forecasting volatility. Ultimately, the GARCH (1,1) model is chosen, combined with the ARIMA model obtained previously to predict the stock price. The closing prices of the stocks in the dataset until July 2023 are used as the training set, and Figure 1 reports the volatility of the dataset.

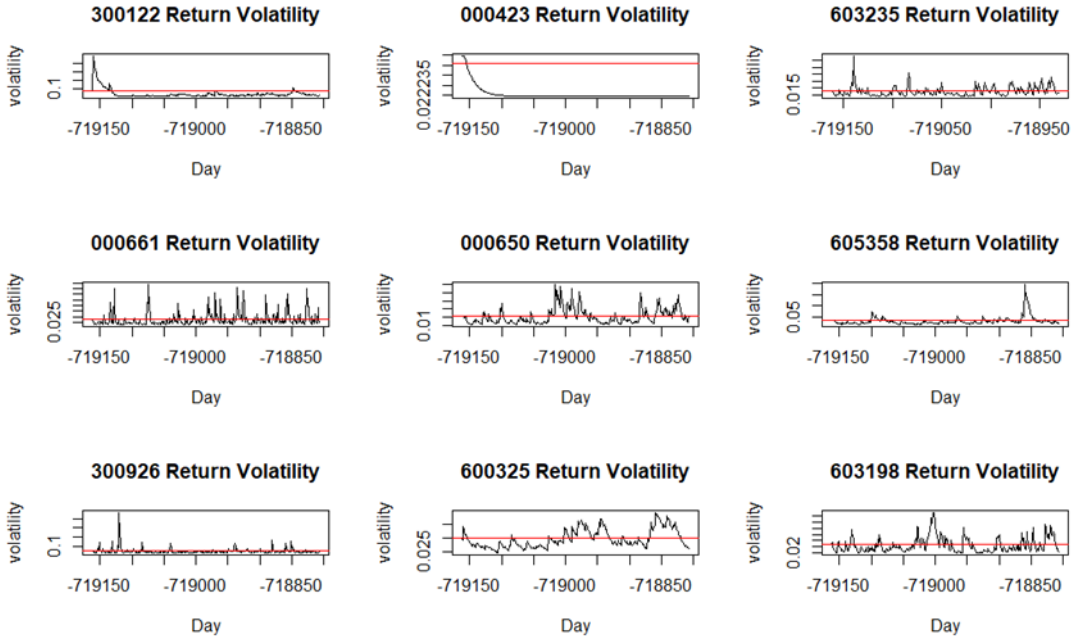

Fig. 1 Trends in volatility for nine stocks over 18 months. (Note: The solid red line depicts the Standard deviation of residuals).

Comparing the trends of the volatilities of the nine stocks, it can be seen that all volatilities show a distribution of sharp peaks and thick tails, except for one stock (000423) with a smooth curve. There are differences in the frequency and intensity of the spikes. Some stocks (300122 and 605358) have spikes that occur only once or twice, which also suggests that these stocks experienced an extremely violent shock. And the volatility at this moment is much higher than the values of the rest of the volatility, making those relatively moderate fluctuations visually unremarkable in the graph. Most of the graphs depict multiple peaks that are extremely different from the values of most of the volatilities, making the standard deviation of the residuals represented by the red line locates away from the peaks. However, there are exceptions. Some stocks (e.g., 600325) do not experience violent changes over the period studied, whose volatility data are all closer together. Also, there is no distinct peak described in these graphs, and the standard deviation distribution of these stocks is distributed close to the middle of the range of data values.

The fitting of volatility not only improves the accuracy of the forecasts, but also reveals differences in the policy shocks and intra-firm shocks to which each stock is exposed, providing more possible reasons for the model to explain the existence of heterogeneity. The last 20 days of data for each of the nine datasets are used as a test set and are compared with the predictions from the fitted model to obtain three indicators to describe the prediction errors.

The results reported in Table 2 intuitively show that the predictions are relatively accurate. The differences between each value of the accuracy of the up/down study do not distinguish the nine stocks well and are less capable of describing the magnitude of the error than the other two indicators. Therefore, this paper mainly uses the RMSE and the MAE to evaluate the prediction effect when performing the heterogeneity analysis.

Table 2. Errors in ARIMA-GARCH forecasting results.

Stock Code	Model	RMSE	MAE	Accuracy (up/down)	Market Capitalization (billion yuan)	Share Price (yuan/share)
300122	ARIMA(0,1,0)	0.112	0.109	50%	0.478	47.00
000423	ARIMA(3,1,5)	0.029	0.026	75%	0.437	50.00
603235	ARIMA(0,2,2)	0.054	0.043	80%	0.024	30.50
000661	ARIMA(1,1,0)	0.124	0.076	60%	0.488	146.00
000650	ARIMA(4,1,4)	0.070	0.057	55%	0.382	6.80
605358	ARIMA(1,1,5)	0.013	0.011	55%	0.448	37.00
300926	ARIMA(3,1,3)	0.033	0.025	55%	0.405	25.00
600325	ARIMA(1,1,0)	0.021	0.016	60%	0.391	67.00
603198	ARIMA(0,1,0)	0.057	0.042	45%	0.393	10.00

Note: This table only reflects the order of the different ARIMA models used for different stocks, omitting the specific GARCH model. Volatility is obtained by fitting GARCH (1,1) for all datasets. Market capitalization and stock prices are actual data after the close of trading on July 25th.

When the model is utilized to predict the stock prices of organizations that are in the same industry (the first five rows in the table), stocks with lower stock prices have higher predictive accuracy, and stocks with the highest stock prices possess the largest errors. After taking differences in volatility into account, among those companies with close stock prices, the predictions have lower errors for stocks with lower company market capitalization. There are also differences between companies with similar volumes in different industries, which may be attributed to industry structure. Specifically, the company in the semiconductor industry (605358) is more affected by market demand and has a more continuous change, while the company in the real estate industry (600325) is more regulated by policies and is more predictable in terms of the extent of changes, except for obvious shocks. Companies with competitive advantages in the pharmaceutical industries are instead predicted to deviate more from the true value.

2.5. PCA-LSTM Method

Due to the significant correlations between some of the original indicators (correlation coefficients are greater than 0.5), PCA is used in this paper to perform dimensionality reduction, aiming to reduce data redundancy. By calculating the principal component scores for each of the nine stocks, the first two principal components are found to contribute more than 95% of the cumulative variance.

Table 3. Composition of the first two principal components

Principal Component	Open X1	High X2	Low X3	Close X4	Change X6	Vol X8	Amount X9
C1	0.403	0.377	0.383	0.361	-0.345	-0.025	0.112
C2	0.104	0.100	0.165	0.150	0.305	-0.683	-5.34

The specific composition of the two principal components is reported in Table 3. It is observed that principal component C1 shows the historical price drivers and C2 represents the market drivers. Therefore, in this study, the LSTM model is built in MATLAB.

After adding the LSTM layer to the established sequential model, the parameters are set. The optimization algorithm uses Adam, and the model uses 50 epochs with each batch size of 50. The number of feature variables is 5, the time step is 1, and in the Dropout layer, 20% of the input units are set to 0. The model is trained using 50% of the previously divided training set, which takes the values of the three features (change, close, amount) of the previous day as the variables. After the training is completed, the predicted values can be compared with the actual close price to evaluate the effectiveness of the model, as reported in Figure 2.

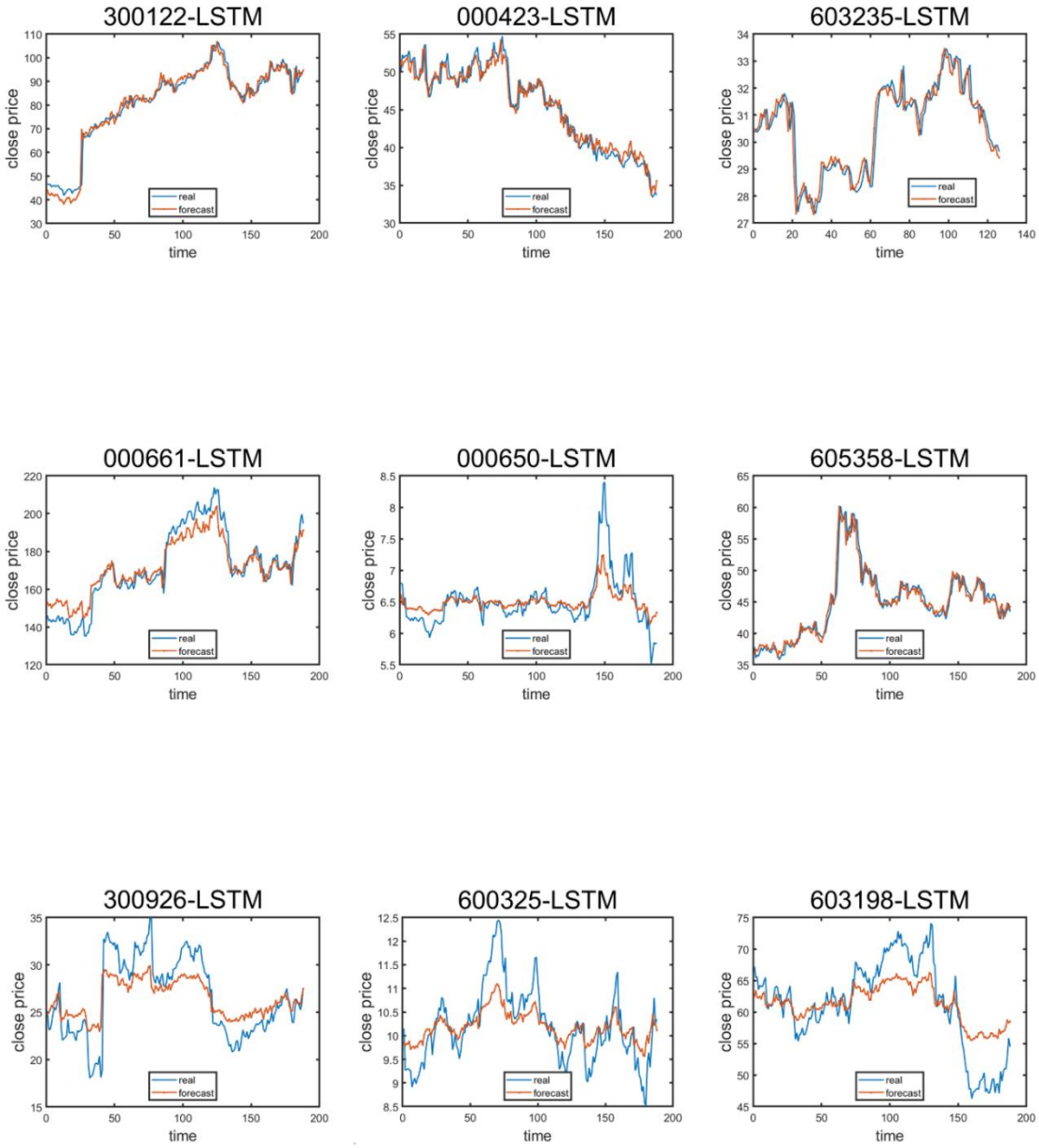

Fig. 2 Results of LSTM model fitting

Based on the reported price prediction plots of 188 days of the nine stocks, it can be seen that the lines of predicted and actual values coincide in most of the plots. Although the values at some transitions are not exactly equal, it can still be concluded that the LSTM network can more accurately predict stock price movements over long periods, capturing complicated relationships among features better. The prediction errors of the model of different stocks are compared in Table 4, using RMSE and MAE as criteria.

Table 4. Errors in LSTM forecasting results

Stock Code	RMSE	MAE	Market Capitalization (billion yuan)	Share Price (yuan/share)
300122	80.91	79.55	0.478	47.00
000423	47.71	47.37	0.437	50.00
603235	29.55	29.44	0.024	30.50
000661	172.1	171.8	0.488	146.00
000650	6.558	6.210	0.382	6.80
605358	39.72	39.23	0.448	37.00
300926	27.38	26.45	0.405	25.00
600325	13.74	13.62	0.391	67.00
603198	61.94	61.22	0.393	10.00

The results show that the heterogeneity in prediction accuracy still exists, and the pattern is similar to the prediction results of the ARIMA-GARCH method, especially the differences in the errors between companies in different industries. Generally speaking, the larger the stock price and market capitalization of the company, the larger the prediction error of the model.

3. Conclusion

Based on the theoretical study of stock prices, the author uses two methods to predict the daily closing prices of nine stocks with good liquidity distributed across multiple industries and to make comparisons. This study finds that the ARIMA-GARCH method, more suitable for short-term forecasting, can combine recent information and analyze volatility, while the PCA-LSTM network, more advantageous in forecasting long-term trends, can integrate and utilize more stock features.

Several factors contribute to the differences in the effectiveness of individual stock price forecasts. In addition to being influenced by the chosen forecasting features, such differences may be related to factors such as the market structure of the industry, the position of the company, and the ability of policies to influence the industry. On the one hand, industries that are more affected by market demand have more persistent stock price changes and higher forecasting accuracy. On the other hand, for those industries that are more subject to policy regulation, the developed models need more perspectives on policy to work on the exactness of the expectation of cost variances.

Generally speaking, the stocks with the most accurate forecasts are with low share prices, large market capitalization, low valuations, and low levels of competition. This conclusion is also reliable with the consequences of the correlation analysis and the principal component analysis of the raw metrics conducted in this paper. Adding these variables to the model may further improve the precision of the model's predictions and maintain the stability of the financial market.

References

[1] Hiemstra C, Jones J D. Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation. Journal of Finance, 1994, 49(5): 1639-1664.

[2] Vapnik V. The nature of statistical learning theory. Springer Science & Business Media, 2013.

[3] Atiya A, Talat N, Shaheen S. An Efficient Stock Market Forecasting Model Using Neural Networks. International Conference on Neural Networks.1997, 4: 2112-2115.

[4] Xie X K, Wang H. Recurrent Neural Network for Forecasting Stock Market Trend. International Conference on Computer Science, 2016.

[5] Sutskever I, Vinyals O, Le Q. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, 2014, 3104-3112.

[6] Nelson D, Pereira A, Oliveira R. Stock Market's Price Movement Prediction with LSTM Neural Networks. International Joint Conference on Neural Networks, 2017, (4): 1419-1426.

[7] Di Persio L, Honchar O. Artificial neural networks approach to the forecast of stock market price movements. International Journal of Economics and Management Systems, 2016, 1: 158-162.

[8] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 2018, 270(2): 654-669.

[9] Hotelling H. Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 1933, 417-441.

[10] Kim H Y, Won C H, Forecasting the Volatility of Stock Price Index: A Hybrid Model Integrating LSTM with Multiple GARCH-Type Models. Expert Systems with Applications, 2018, 103: 25-37.