The Stock Price Prediction Based on Time Series Model, Multifactorial Regression, Machine Learnings

Xiangyu Jin[1, †], Luya Wei[2, *, †], Qihua Zhang[3, †]

[1]Department of Statistics, University of Illinois Urbana-Champaign, Champaign, the United States

[2]Faculty of Business Administration, Yorkville University, Toronto, Canada

[3]Department of Mathematics, University of California, Santa Barbara, Santa Barbara, the United States

*Corresponding author: Luya.Wei@yorkvilleu.ca

[†]These authors contributed equally.

Abstract: In general, it is hard to forecast the prices the stock prices due to the stochastic fluctuations. This research aims to describe the process to use time series models, multifactorial regression, and machine learning to predict stock prices. ARIMA and EGARCH models are frequently used time series models to predict stock prices. Least-squares linear regression model, Lasso, and Polynomial Linear Regression model predict well in statistical regression methods. RNN and LSTM have higher prediction accuracy. Overall, time series models, statistical regression, and machine learning all can predict stock prices. Summarizing the different methods or models to forecast stock market trending can help investors to prepare relevant investing strategies. These results shed light on guiding further exploration of

Keywords: stock price prediction, time series model, multifactorial regression, machine learning.

1. Introduction

In general, stocks refer to partial ownership of an underling assets by an individual or group of people. Stock market predictions attempt to determine the value of stocks and to give individuals an accurate idea to understand the market. Stock forecasting can help companies improve economics, interest rates, and so on to influence the market [1]. However, predicting stocks is a very difficult task because almost all investors do not always correctly predict these hyper-parameters. For private individuals, promotions are regarded as games of chance. Because of its uncertainty, it attracts the attention of many scholars [2]. Many factors determine stock trends. In particular, some significant changes in society may affect trends in stocks seriously. The rapid spread of COVID-19 and subsequent lockdowns have contributed to the crash since 2020. The Stock Market Crash in 2020 has also been referred to as the Great Coronavirus Crash [3]. Therefore, it is very important to offer an appropriate approach to predict the trend of the market, which offers a tool to investors to prepare corresponding investment strategies after evaluating the behavior of the market. This research aims to summarize stock price predictions based on time series models, multivariate regression, and machine learning over the past five years.

On acount of difficulties in predicting behaviors of stock markets, multiple advanced ways are developed for stock share price prediction. In time series models for stock prediction, the ARIMA models are the most widely used to predict future price and the EGARCH model is used to predict volatility [4]. Multifactorial regression is also combined used in time series model. Hybrid models have become the most popular models in stock prediction. Other than that, deep learning neural networks (DLNNs) have powerful statistic probabilities for image modeling [5].

In this paper, we will demonstrate different methods and models to analyze stock markets. The rest part of the paper is organized as follows. The Sec. 2 will introduce forecasting models used in stock markets. The Sec. 3 will be applications of time series models. The Sec. 4 will be applications of statistical regression. The Sec. 5 will be applications of machine learning.

2. Forecasting Models

2.1 ARIMA(p,d,q)

The ARMA are statistical models that predict the future value of the variable based on the value during the previous periods [6], which can be mathematically described as [7]

$$Y_t = i + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + e_t - \grave{e}_1 e_{t-1} - \grave{e}_2 e_{t-2} \ldots - \grave{e}_q e_{t-q} \quad (1)$$

Where p is the autoregressive part (AR), q means the moving part (MA) and $\{e_t\}$ is the white noise. ARIMA(p,d,q) is an extension of the ordinary ARMA model:

$$\varphi_p(B)(1-B)^d Y_t = \grave{e}_0 + \grave{e}_q(B) e_t \quad (2)$$

Here, d is integration (differencing). Typically, the ARIMA model is equivalent to ARMA model when d=0.

2.2 GARCH

GARCH describes a method to evaluate the fluctuations of the variable, which is widely implemented to measure the future volatility of the underlying assets [4]. According to Ref. [4], following exponential GARCH can be used to describe the leverage effects

$$h_t = a_0 + \sum_{i=1}^{p} a_i \frac{|\varepsilon_{t-i}| + \gamma_i \varepsilon_{t-i}}{\sigma_{t-i}} + \sum_{j=1}^{q} b_j h_{t-j} \quad (3)$$

Here, $h_t = log\sigma_t^2$ and $\sigma_t^2 = \sigma_t^2 = e^{ht}$.

2.3 OLS

Ordinary Least Squares regression is the most common regression scenario to analyze the importance of the features in terms of the estimating coefficients [8]. Here, least square stands for the minimum squares error, where the OLS regression model writes [8]:

$$Y = \beta_0 + \sum_{j=1..p} \beta_j X_j + \varepsilon \quad (4)$$

2.4 Ridge Regression

Different from OLS, Ridge regression penalizes the size of the regression coefficients based on their L2 (sum of squared coefficients) norm [9]:

$$L_{ridge}(\hat{\beta}) argmin_\beta \sum_i (y_i - \beta' x_i)^2 + \lambda \sum_{k=1}^{K} \beta_k^2 \quad (5)$$

where $argmin_\beta \sum_i (y_i - \beta' x_i)^2$ is the residual sum of squares, and $\lambda \sum_{k=1}^{K} \beta_k^2$, where β_k^2 is the squared coefficients, λ is a controllable parameter can be adjusted manually.

2.5 LASSO regression

Lasso is similar to ridge regression, except for the penalty term, which is L1 (sum of absolute values of the coefficient estimates) in LASSO [9]. The penalty term formula is [9]:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \beta' x_i)^2 + \lambda \sum_{k=1}^{K} |\hat{\beta}_j| \quad (6)$$

2.6 Tree Algorithm (Random forests, Gradient boosting-Xgboost lightgbm)

Typically, a decision tree structure has three components, i.e., a root node, test nodes, and decision nodes (leaves) [10]. It uses different conditions to filter data. After the root node, each test node splits the data into further parts according to some criteria [10]. The final node is the leaf node. An example is sketch in Fig. 1. With the proposal of the gradient boosting, novel scenarios are also presented (e.g., Xgboost, lightgbm), which possess better performance and cost smaller training time, i.e., is capable of implantation in big data analysis.

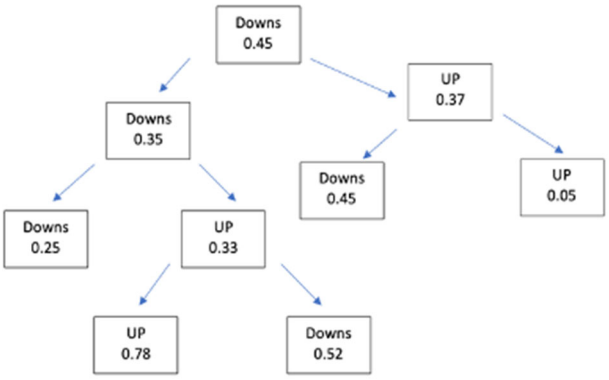

Figure 1. Up/Down Decision Tree [10]

2.7 Neural Network

A neural network is a series of algorithms that ensembles to simulate decision procedures of human being [11]. As shown in the Fig. 2, there are three layers in this simple neural network. Within several hidden layers, the algorithm takes two inputs and produces the output layer.

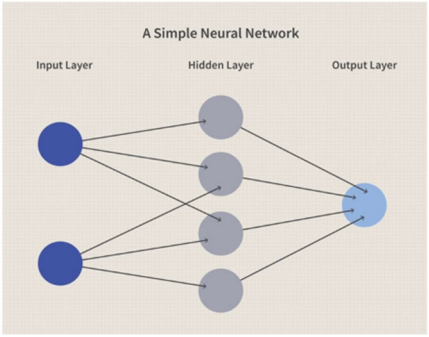

Figure 2. Simple Neural Network [11].

3. Applications of Time series models

Contemporarily, time series model is widely adopted in the stock price prediction as the price itself is clearly an evolution time series. Among various approaches, the ARIMA model is one of the most utilized approaches for linear cases [12]. As a matter of fact, it possesses plenty of advantages, e.g., easily implementation, reliability, capability for different time periods. Hayes attempts to use ARIMA model to predict some of underlying assets in India, listed under NIFTY100 [12]. In this analytical study, the secondary data involves the daily closing prices of shares of pharmaceutical companies listed in NIFTY 100 of India from January 1, 2017 to December 31, 2019 is used to make predictions [12]. However, the model developed for Dr. Reddy Laboratories is less reliable because of slightly more deviations.

Other scholars utilize ARIMA to mine stock data of banking, where Amman stock market (ASE) in Jordan from 1993 until 2017 is used [6]. To find the optimal parameters, the metric RMSE is chosen for judgement. After analysis based on MINTAB, ARIMA (1,1,2) was found as the best performance model in terms of RMSE. In this paper, the authors collect sufficient real data to build the ARIMA models for the sake of improving short-term prediction [6]. It should be noted that the approach is limited to a short-term forecasting, while might be meaning less for long periods [6].

Moreover, other researches apply the GARCH framework to predict the volatility, which is crucial for risk analysis and management [4]. To be specifica, Mohamed F and Ahmed J used several volatility models based on GARCH framework for NSE of India [4]. The author used the python programming language version 3.7.4 for implementing the GARCH models [4]. According to the analysis, the models are fine-tunned and then back tested on the out-of-sample data to estimate their

accuracy in the prediction of future volatility of the stocks [4]. It is concluded that EGARCH yielded the most accurate results [4].

4. Applications of Statistical regression.

Regression is a statistical method to determine the relationship between variables [13]. In ref. [1], the authors use the least-squares linear regression model to predict stock prices. The equations of least squares method are:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \quad (8)$$

where N is total sample, y is real sample values, and $\hat{y}$ is model estimated values. The scholars collect the dataset from yahoo finance from the Bank of America stock for the last 7 years. The authors split the dataset into testing, validation, and training sets, calculate the lowest mean absolute percent error and root mean squared error, and try to adjust parameters to lower errors and to predict the testing dataset. N is previous number days which determine the present day's stock. Since N=4 is the lowest mean absolute percent error, it is the optimal prediction in the testing dataset. As shown in Fig. 4, MAPE is 1.367%, and the RMSE is 0.512, which means that the results were accurate. Hence, this figure proves the model can help to predict stock prices [1].

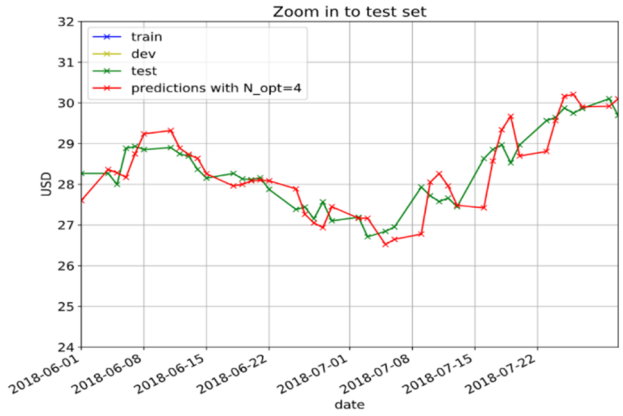

Figure 3. Trained Model Result [1].

Bose et al. compare the linear, Ridge, Bayesian Ridge, and Lasso regression models' performance. The dataset is recorded for each day for about the last 8 years, from December 8, 2008 to August 30, 2017. According to Figure 4, Linear regression's score is about 0.93674, which is the lowest. Lasso regression is the most accurate. Its performance score is about 0.9568 [14].

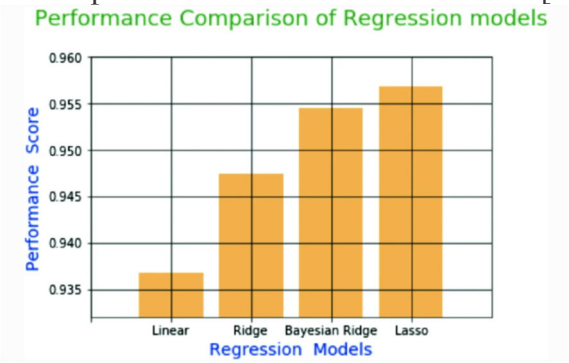

Figure 4. Different Regression Models Performance [14].

Amrutphale et al. compare simple linear regression model (SLR) with polynomial linear regression model (PLR) to determine which is more accurate to predict stock prices. Below are formulae of SLR and PLR equations:

$$Y = b_0 + b_1 X_1 \quad (9)$$

$$Y = b_0 + b_1 X_1 + b_2 {X_1}^2 + \cdots + b_d {X_1}^d \quad (10)$$

where Y is a dependent variable; X_1 is an independent variable; $b_1$is a coefficient for X_1. The dataset is recorded by the official website of NSE India from Dec. 04, 2017 to Dec. 3, 2018. As shown in Fig. 5, the PLR model is more accurate than the SLR model. Hence, the PLR model is a better option to predict stock prices, and it is hard to predict it by using the SLR model.

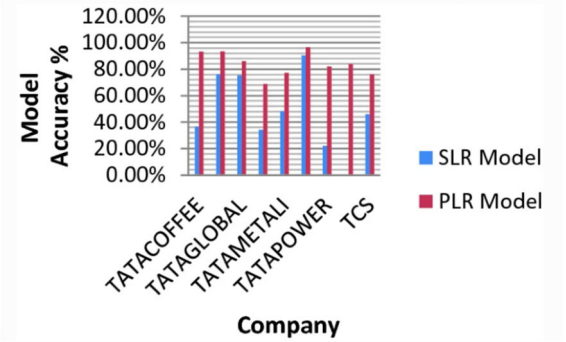

Figure 5. SLR and PLR Model Accuracy [15].

5. Applications of machine learning

As a matter of fact, Machine Learning is the core of artificial intelligence, where Deep Learning is one of the crucial subsets based on ANN. RNN is a popular deep learning model with a specific memory function that can save the previously calculated information for a long time [16, 17]. Among various types, LSTM is a particular type of RNN that uses both standard and special units. LSTM can process single data points and complete data sequences due to its feedback connections. Furthermore, the unique design structure of LSTM makes it more suitable for processing and predicting important events with long intervals and delays in time series [16]. Figure 6 describes the composition of LSTM nodes [18].

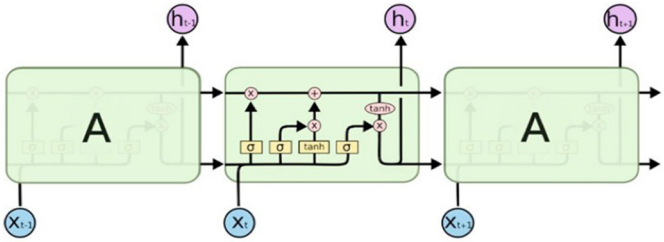

Figure 6. The composition of Long and short-term memory nodes [18]

In a related report, deep learning methods (RNN and LSTM) are explored to evaluate stock price trends more accurately. According to the results, the first method uses continuous feature data, and it shows that RNN and LSTM have the highest prediction accuracy of about 86%, which is quite different from other models. The second method uses binary data for features, and though the prediction performance is significantly improved, RNN and LSTM are better prediction methods with an F1- score is about 90% [17]. Therefore, both RNN and LSTM show skills in predicting stock movements.

6. Limitations & Future prospect

The stock market is non-linear, so the prediction of stock prices can be influenced by other factors. On the one hand, macro and micro factors. For example, politics, the state of the global economy, corporate financial performance and other unexpected events make stock prices dynamic and volatile. On the other hand, consumer physiological and psychological factors, rational behavior and irrational behavior will cause the change in stock price. Therefore, it is challenging to predict prices accurately because these factors are difficult to predict accurately.

Linear models and machine learning approaches covered in the report suffer from a lack of baseline data sets and performance metrics. Without a unique data set and the definition of appropriate performance indicators, the results of a complete comparison of the studies presented cannot be shown in the report to select an appropriate solution for a particular problem. In addition, the accuracy of the algorithm decreases as the number of technical indicators decreases, so better accuracy can be obtained by using a more extensive data set.

The stock market includes companies in different industries, and stocks in each industry will show different trends. Therefore, stock selection is part of the investment decision process and requires a model. Picking an arbitrary stock to start with would distort the character of the portfolio of investment opportunities. What is more, in an environment where stock prices and underlying data are accessible and free, the indiscriminate application of preprocessing techniques and machine learning algorithms will produce arbitrary results.

Although the past performance of a stock price is not necessarily indicative of its future development, it is still necessary to monitor its past performance. With the introduction of machine learning and its powerful algorithms, the stock's opening and closing value as well as the highest and lowest value on the same day, are displayed on each date. In addition, the total volume of shares in the market is indicated, too. Therefore, data scientists can use this information to look at the data and develop different algorithms to help find the appropriate stock value. Moreover, with machine learning in stock market forecasting, stock market forecasting procedures have become much more straightforward. Machine learning is based on facts, figures and data, without regard to emotions or biases. On this basis, machine learning saves time and resources and outperforms humans in performance [19].

7. Conclusion

In summary, this paper discusses how to predict stock prices from the perspective of time series models, multifactorial regression and machine learning. Specifically, the ARIMA model, EGARCH model, least-square linear regression model, Lasso model, polynomial linear regression model, RNN and LSTM model were demonstrated. According to the analysis, the ARIMA and EGARCH models have higher prediction usage; the least-square linear regression, Lasso, and polynomial linear regression models have better prediction effects; and the RNN and LSTM models have higher prediction accuracy. Nevertheless, the market economy and consumer behavior will affect the stock price prediction. Besides, the selection of models and the number of samples will affect the accuracy of stock price prediction. In the future, stock price prediction will become simpler and more accurate with technology development. Overall, these results offer a guideline for stock price forecasting.

References

[1] Emioma C C, Edeki S O. Stock price prediction using machine learning on least-squares linear regression basis[C]. Journal of Physics: Conference Series. IOP Publishing, 2021, 1734(1): 012058.

[2] Sudhakar K, Naganjaneyulu S. STOCK PRICE PREDICTION BASED ON FINANCE RELATED NEWS USING NLP, LASSO AND ARIMAX[J]. Journal on Software Engineering, 2020, 14(4).

[3] Shu M, et al. The ‘COVID’crash of the 2020 US Stock market[J]. The North American Journal of Economics and Finance, 2021, 58: 101497.

[4] Fakhfekh M, Jeribi A. Volatility dynamics of crypto-currencies’ returns: Evidence from asymmetric and long memory GARCH models[J]. Research in International Business and Finance, 2020, 51: 101075.

[5] Liu Q, et al. Stock market prediction with deep learning: The case of China[J]. Finance Research Letters, 2022, 46: 102209.

[6] Hayes A, Autoregressive Integrated moving average (ARIMA)[EB/OL]. Investopedia, Investopedia, 2021-10-12. Information on: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-

arima.asp#:~:text=An%20autoregressive%20integrated%20moving%20average%2C%20or%20ARIMA%2C%20is%20a%20statistical,values%20based%20on%20past%20values.

[7] Almasarweh M, Alwadi S. ARIMA model in predicting banking stock market data[J]. Modern Applied Science, 2018, 12(11): 309.

[8] Ordinary least squares regression (OLS)[EB/OL]. XLSTAT, Your data analysis solution, Addinsoft, 2022. Information on: https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols#:~:text=Ordinary%20Least%20Squares%20regression%20(OLS)%20is%20a%20common%20technique%20for,simple%20or%20multiple%20linear%20regression).

[9] Melkumova L E, Shatskikh S Y. Comparing Ridge and LASSO estimators for data analysis[J]. Procedia engineering, 2017, 201: 746-755.

[10] RATNAPARKHI S, PARADKAR M, Machine learning: An introduction to decision trees[EB/OL]. Quantitative Finance & Algo Trading Blog by QuantInsti, Quantitative Finance & Algo Trading Blog by QuantInsti, 2019-08-27. Information on: https://blog.quantinsti.com/use-decision-trees-machine-learning-predict-stock-movements/.

[11] Chen J, Neural network definition[EB/OL]. Investopedia, Investopedia, 2021-12-08. Information on: https://www.investopedia.com/terms/n/neuralnetwork.asp.

[12] Meher B K, et al. Forecasting stock market prices using mixed ARIMA model: A case study of Indian pharmaceutical companies[J]. Investment Management and Financial Innovations, 2021, 18(1): 42-54.

[13] Beers, Brian. “What Regression Measures.” Investopedia, Investopedia, 8 Feb. 2022, Information on: https://www.investopedia.com/terms/r/regression.asp.

[14] Bose R, et al. Risk analysis for long-term stock market trend prediction[C]. International Conference on Advances in Computing and Data Sciences. Springer, Singapore, 2019: 381-391.

[15] Amrutphale J, et al. A Novel Approach for Stock Market Price Prediction Based on Polynomial Linear Regression[M]. Social Networking and Computational Intelligence. Springer, Singapore, 2020: 161-171.

[16] Rundo F, et al. Machine learning for quantitative finance applications: A survey[J]. Applied Sciences, 2019, 9(24): 5574.

[17] Nabipour M, et al. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis[J]. IEEE Access, 2020, 8: 150199-150212.

[18] Moghar A, Hamiche M. Stock market prediction using LSTM recurrent neural network[J]. Procedia Computer Science, 2020, 170: 1168-1173.

[19] Soni P, et al. Machine Learning Approaches in Stock Price Prediction: A Systematic Review[C]. Journal of Physics: Conference Series. IOP Publishing, 2022, 2161(1): 012065.