The Comparison of Stock Price Prediction Based on Linear Regression Model and Machine Learning Scenarios

Xiwen Jin[1(✉)] and Chaoran Yi[2]

[1] Intelligence Science and Technology, University of Shanghai, Shanghai, China
febawa@i.shu.edu.cn
[2] School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia
chaorany1@student.unimelb.edu.au

Abstract. Financial price prediction always plays a vital role for investment decision. This paper investigates the prediction of the close price of LONGi based on linear models and machine learning approaches, including ordinary least square (OLS), Lightgbm, XGBoost, random forest, LSTM and GRU models. Specifically, according to our result, the LSTM and the GRU perform relatively better results and the random forest is the worst. Based on the analysis, all the models can predict the trend of the close price. These results offer a guideline for investors that desires to forecast the price trend of a specific underlying assets. These results shed light on comprehending the characteristics of different regression models.

Keywords: Stock Prediction · Machine Learning · OLS · Lightgbm · XGBoost · Random Forest · LSTM · GRU

1 Introduction

The stock price prediction is realized by analyzing the past data to predict the future event or events. On this basis, it is able to help investors to make decisions in stock market about weather to realization the assets by selling raised stocks cause prediction shows a high drop in price as well as to short selling that stock, or to buy some promising stocks cause prediction illustrate that there is a great chance that this stock will rice in the future, etc. With the help of the stock price prediction, investors will have less probabilities to make mistakes, this action will lead to an increase in the cash flow of the market and continues to boost the national economy. As the disposable income increases of people due to extra gain from the stock market, more money can be spent, i.e., Gross Domestic Product (GDP) and National Happiness Index will increase. The prediction of stock price can be categorized into three types according to time period. Specifically, short term forecasting represents for prediction within one year, medium term forecasting for shorter than five years longer than one year, while long term forecasting for beyond five years.

X. Jin and C. Yi—Contributed equally.

D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 837–842, 2023.
https://doi.org/10.2991/978-94-6463-030-5_82

Currently, there are three popular models for stock price prediction, which are LSTM, RNN and Random Forest. According to previous paper, it is feasible to boost the forecasting accuracy in a large extent compared to the middle and low frequency grey Markov models [3]. Besides, other scholars also state that the incremental of factor quantities is capable of improving the performance of the prediction results i.e., the forecasting will become more accurate [1, 4, 5, 9]. In addition, plenty of previous studies offer some effective factors and demonstrate the mechanism for the impact of these significant factors on price prediction [2, 6–8, 10].

In this paper, OLS, Random Forest, Xgboost, Lightgbm and LSTM will be applied with official published stock market data of one typical listed company collected from official database. For the sake of determining the best performance model, the prediction value errors and R-square score between the real price and prediction price are calculated. Thereby, the best model for predicting stock price is obtained according to the metrics mentioned above.

The rest part of the paper is organized as follows. The Sect. 2 will introduce the data origination as well as the regression models. Subsequently, the results explanations and comparisons of the models will be presented in Sect. 3. Eventually, a brief summary will be given in Sect. 4.

2 Data and Method

The data selects LONGi Green Energy Technology Co., Ltd. (601012.SS) stock information from August 1, 2021 to December 31, 2021. To predict and evaluate the models, the daily closing price is collected. To realize the prediction, the single-step forecasting method is implemented, i.e., predicting with the historical data base and updating the data base with the newly known conditions.

The data is obtained by the interface of tushare.pro library. To train the data and process the mode, both conventional linear models and machine learning approaches are carried out, including ordinary least square (OLS), Lightgbm, Xgboost, random forest, LSTM and GRU.

As a matter fact, it is necessary to give a brief introduction to these models. The OLS model is one of the most common and simple models with a long history. The core of it is to find the coefficients of a model whose prediction values with minimum sum of squared error. It is used in regression, fitting and optimization in the early stage. However, this method is unable to find the best training parameters due to the existence of the saddle and it often faces the problem of overfitting. Different from such a traditional prediction model, there are plenty of state-of-art machine learning approaches. One of the early proposals is the random forest model which predict the values based on different decision trees. To increase the calculation speed, Xgboost and LightGBM are developed, which both increase the model accuracy and decrease the training speed based on resampling. Despite the tree algorithms, the neural networks models are also taken into consideration. Specifically, the LSTM and GRU (the variation of RNN) are selected to process the data. Thereinto, the structure of GRU is simpler and more concise than LSTM. To save the time for training, all the training parameters of above models are tightly controlled in a small range that close to the default values.

3 Results and Discussion

To achieve the goals of this paper, 6 different models mentioned above are utilized to forecast the close price of a chosen stock called Longi Green Energy Technology Co., Ltd. (601012). The time span is from August 2021 to January 2022, we used the real data that collected on the first day to predict tomorrow's data. When we had the real data of the second day, we then use that data to predict the day after tomorrow, and it has been going back and forth for five months.

As shown in Fig. 1, all the prediction errors of the different models are illustrated. Based on the results, it is not difficult to find out that the larger the value, or the farther the value is from zero, the greater the error will be. This is because every value of every point on this chart is calculated by y_ Predict minus y_ True.

To have a closer look of Fig. 1, we can clearly find out that during August 2021 to January 2022 among the six different models represented by different colors, the most prominent color is the purple which represents Random Forest Regressor. On this basis, it indicates that this model has the largest close price error and is obviously the worst model among all stock price prediction models.

The errors of the remaining five models are relatively small. The scattered points on the graph are dense, and the connecting lines are staggered with one and another. It is difficult to distinguish which model has the highest accuracy from the graph. Although some green points (LGBM Regressor) appear further compare to other models from time to time, but still more advanced data and tables are needed for further analysis in order to keep our result accurate and professional. Besides, LGBM Regressor cannot be ruled out easily, just in case, if LGBM Regressor only performed poorly in those days.

Subsequently, let's look at the histogram of R2 score and MSE as exhibited in Figs. 2 and 3.

When $0 < \text{R2 score} \leq 1$, the maximum value is reached, i.e., the numerator is 0 and $R2 = 1$, which means that the predicted value of the stock price data is completely equal to the real value without any error. In other words, the effect of our model is the best at present. But usually, there will always be errors in the model. When the error becomes larger, the value of R2 score will close to zero. With regard to R2 Score < 0, it means

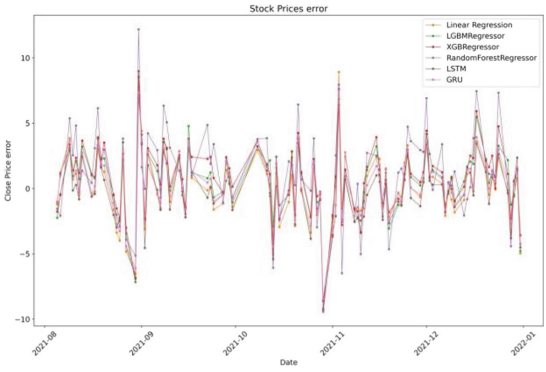

Fig. 1. The price prediction error of different models.

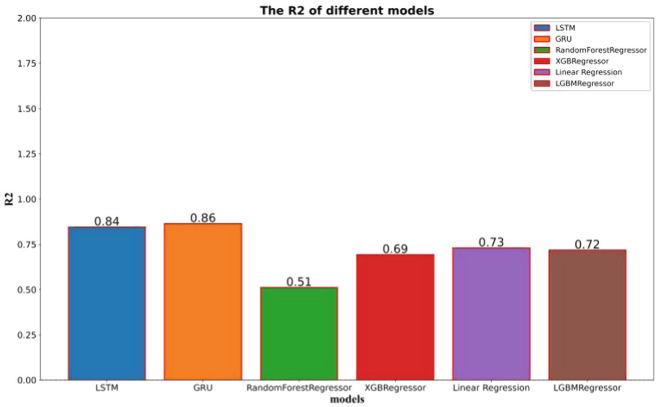

Fig. 2. The R^2 scores of different models.

that the error generated by the stock price prediction model is greater than that generated by using the mean value.

As given in Fig. 1, we obtain the R2 score for different models: LSTM 0.84, GRU 0.86, Random Forest Regressor 0.51, XGB Regressor 0.69, Linear Regression 0.73 and LGBM Regressor 0.72.

It is easy to see that LSTM and GRU have the best results among other models, because they are closest to the value of 1, and Random Forest Regressor is the worst model, since it has the lowest R2 score.

Before reporting the results of Mean Square Error (MSE), the definition of it is given first. In parameter estimation, the mean square error refers to the expected value of the square of the difference between the estimated value of the parameter and the true value of the parameter. Therefore, the closer the MSE value of the model is to zero, the more accurate the model is.

For MSE (seen from Fig. 3), the values of different models are given as follows: LSTM 7.06, GRU 6.26, Random Forest Regressor 12.02, XGB Regressor 7.55, Linear Regression 6.64 and LGBM Regressor 6.94. It's not difficult to witness that the GRU model is the most accurate (the lowest MSE) one and again Random Forest Regressor is the worst one (the highest MSE).

Although this paper gives an intuitive picture of price prediction for a certain underlying asset in terms of different forecasting models, the results obtained in this paper still contain a lot of limitations. Primarily, this paper only covers one year of a stock and forecasts it, so the sample size and data set are still too small to compare different models, i.e., some of the functions are not able to achieve. In addition, the factors included in the model are too simple and multi-factorial models should be used if the factor quantities surpass certain value. At the same time, in the process of parameter adjustment, the optimal solution may not be covered, i.e., some models may not play the maximum effect. Besides, the cross validation training is also not taken into consideration since the data size is too small.

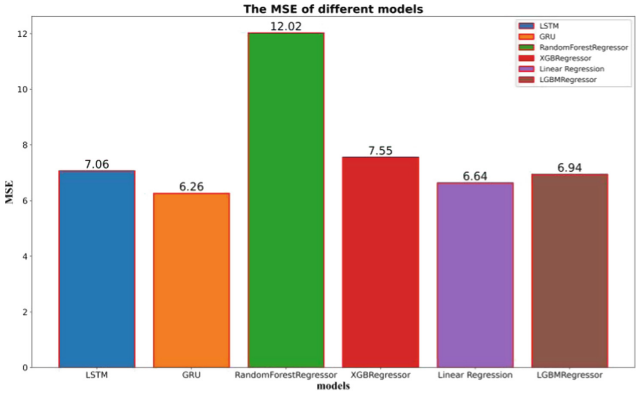

Fig. 3. The MSE of different models.

4 Conclusion

In summary, this paper investigates the differences of price prediction of a certain underlying assets based on 6 different models. According to the experimental results, LSTM and Gru have the best results while Random Forest Regressor has the worst result. Nevertheless, in general, all models can correctly predict the trend of the close price, and successfully predicted the close price of LONGI Green Energy Technology Co., Ltd. (601012.SS) without great error.

However, prediction stock price based on above models still has some defects. Because the stock market is complex, and there are many factors that will affect the stock market (e.g., policy changes, main force buy at the bottom or sell at the top, and the capital volume and information of different companies), which may also have different effects on the accuracy of the model.

Generally speaking, using model to predict the stock price has certain practical use, which can provide great help for investors and investment institutions in short-term stock trading. Nevertheless, further studies should be carried out to systematically verify the feasibility and increase the accuracy with the support of statistical techniques. Overall, these results offer a guideline for price prediction based on linear model and machine learning approaches.

References

1. Baoshi Wen, Qisheng Yan "Data multidimensional processing LSTM stock price prediction model." Jiangxi science 38.4 (2020): 8.
2. Dai Chengjun Research and implementation of stock price prediction based on text emotion analysis [D] Chongqing University, 2016
3. Dong Li, Xiaohong Su, Shuangyu Ma "Stock price prediction algorithm based on new dimensional Grey Markov model." Journal of Harbin Institute of technology 35.2 (2003): 5.
4. Hao Li. "Stock price forecast based on multi input LSTM Diss. Shanghai Jiaotong University, 2019.

5. Jie Zhang "Empirical analysis of stock forecasting based on LSTM." Shandong University (2020). Junhao Li "Stock price prediction model based on improved multiple linear regression." Science and technology economy market 8 (2019): 3.
6. Jing Zhang "Research and development of stock price prediction system based on technical analysis." Computer development and application 23.12 (2010): 3.
7. Qin Lu Research on investment value analysis and prediction modeling of stock information [D] Jilin University of Finance and economics, 2019. https://doi.org/10.26979/d.cnki.gccsc.2019.000282.
8. Nan Jiang "On the prediction model of stock price changes." Wuhan finance 5 (1993): 55–56.
9. Yan Peng, Yuhong Liu, and Rongfen Zhang "Modeling and analysis of stock price prediction based on LSTM." Computer engineering and application 055.011 (2019): 209–212.
10. Ning Yihe Research on stock price forecasting model based on correlation [D] Beijing University of Posts and telecommunications, 2018