Comparison of Stock Price Prediction in Context of ARIMA and Random Forest Models

Yuxin Zhao[1, *]

[1]School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China

*Corresponding author: Yuxin.Zhao2002@student.xjtlu.edu.cn

Abstract. With the gradual improvement of the stock market system, coupled with the increase of per capita personal income, speculating in the stock market will become the main way for residents to invest in the future. Being able to predict the return on stocks is a highly desirable thing for investors. Predicting the future trend of stocks is an essential part when investing in stocks. There are many categories of stock prediction. The purpose of this paper is to compare the ability of the ARIMA and the random forest model to predict the stock market by introducing these two models' theoretical knowledge. The results turn out that the ARIMA model is more suitable for short-term forecasting although it's essentially only valid for linear relationships. As for the random forest model, it has higher accuracy but it's more complex and computationally expensive. Overall, these results shed light on guiding further exploration of stock price prediction based on the state-of-art machine learning scenarios.

Keywords: Stock Prediction; ARIMA; Random Forest Model; Comparison.

1. Introduction

As the most important part of the capital market, the stock market is playing an increasingly important role in the social and economic development. The research of the stock trend prediction model has always been a hot topic for experts and scholars in the fields of economics and finance as well as the analysis of data [1]. With the hot stock market in recent years, people's enthusiasm for investment has risen. As the result, the demand for stocks is increasing. The stock market is like a giant prize wheel. Some investors lost a lot on this investment who didn't seize the moment, while some others devise strategies within a command tent to make a huge profit. It is the high liquidity and volatility of the stock market that leads to the high risk. Nevertheless, the stock market is not entirely down to the luck. People always want a higher yield and a lower risk. Although there is no such thing as a completely accurate prediction, it is also necessary to make the short and long forecasts before operation on the stock market which helps to understand the trends in the stock. The stock prediction is designed to maximize returns for the investors by creating the optimal portfolio. Therefore, it is significant to master several methods of stock prediction.

In the long history of finance development, scholars have put forward many models to predict stocks. There are many options for people to choose. The LSTM RNN, which can realize the a relatively accurate prediction of stock price after the accuracy test, makes the speculators have a macro concept of the trend of the stock market [2]. The momentum strategy is a tactics which suggests to buy stocks when they rise and sell them at the first sign of a decline to look for market volatility [3]. In addition, the MARS model is an algorithm can adapt to the nonlinearity of data set whose principle of operation is to decompose the input variables into several step functions [4]. Apparently, the stock forecasting models are much more than these mentioned above. In this article, autoregressive integrated moving average model ARIMA and multiple regression model will be compared in detail.

This article aims to introduce the ARIMA model and random forest model and then make a comparison between them. Not only introduce the flow of the prediction means, it also shows some results of the prediction application. The analysis process of ARIMA model will be presented first. And then the following is the diagrams of how this method work. The analysis of random forest models follows a similar procedure.

2. ARIMA

ARIMA model is one of the most commonly used time series forecasting models in statistical models [1]. In the first time, it is vital to understand the specific meaning of this model. ARIMA is proposed by Box and Jenkins in the 1970s. It is a well-known and widely applied time series prediction method and sometimes it's noted as ARIMA(p, d, q). The implication of ARIMA is divided into three parts which are AR, I, MA. 'AR' is expressed as auto regression which is the autoregression model. Describes the relationship between current and historical values. Use the historical time data of the variable itself to predict itself. A p order linear process excluding the mean and deterministic components can be expressed as:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t \tag{1}$$

Here, y_t is the current value; μ is the constant term; P is the order; γ_i is the autocorrelation coefficient and ε_t is error. There are some limitations of such a model [5]:

- It uses their own data to make predictions.
- It has to be stationary.
- It must have autocorrelation.

Usually, one uses a stationary process of first order, that is, the mean Mt of the stochastic process Xt does not change with time. 'I' is expressed as integration which means the order of integration is single. The time series model must be stationary series to establish the econometric model. It means that in the non-stationary time series, the forecasts may be inaccurate on account of the influence of seasonality. Therefore, in order to avoid this situation, the time series first become stationary by difference which is between the observations and the most recent observations.

'MA' is moving average. A q order stochastic process with the mean and deterministic components removed can be expressed as follows:

$$y_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{2}$$

The auto regressive moving average model ARMA consists of the autoregressive model which is for 'AR' and moving average which is for 'MA' [6]. In the arithmetic, an ARMA model of p and q orders can be expressed as follows:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{3}$$

Sometimes, differences value will be used, for example, first order difference:

$$\Delta x_t = x_t - x_{t-1} \tag{4}$$

and Second order difference:

$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1} \tag{5}$$

The principle of this model is to convert the non-stationary time series into stationary time series. Then here comes a new concept ACF, where the formula is

$$ACF(k) = \rho_k = Cov(y_t, y_{t-k})/Var(y_t) \tag{6}$$

Here, ρ_k ranges from -1 to 1.

3. Random Forest

Subsequently, it's turn to lead to another main character of this article which is the random forest model. Random Forest (RF) is a relatively new machine learning model (nonlinear, tree-based model) ensemble learning method. In various machine learning methods, random forest is the most popular prediction tool using an integrated approach. The reason why it can be an ideal prediction tool is that RF can handle the complex interrelationships between input parameters. In simple terms, RF is a combination of bagging and decision trees. The random forest regression model is composed of multiple regression trees. Each decision tree in the forest has no correlation with each other. The characteristic of decision trees is that they start from the root node and work their way up to the leaf node. Afterwards, all data ends up in the leaves. The random forest model is a machine learning tool proposed several years ago for classification and regression. It has a process which generating the importance of the variables [7].

The following is the main process of the random forest algorithm:

- Select the sample set. In the extraction process of the original sample set, there may not be samples that are repeatedly drawn, but also be samples which are not drawn at all. There are k rounds of drawing in total. Thus the training sets extracted in each round are respectively T_1, T_2 ... T_k.
- Generating the decision trees. Then, a decision tree is generated by using a new set of features. A total of k decision trees is generated in k rounds. These k decision trees are independent of each other on account of the randomness of decision trees in the selection of training sets and features.
- Composite the model. Each decision tree is of equal importance. Therefore, their weights are not needed to be considered. For the classification problem, the final classification result is determined by using all decision tree voting.
- Test the model. To test and verify the model, usually a validation set ought to be prepared. However, there is no need to obtain additional validation sets. It is only necessary to select samples that have not been used from the original sample set. During selection, some samples are not selected once, and some features may not be used in feature selection. Just take the unused data and use it to validate the final model.

The commonly used type of decision trees in random forest model are the CART trees. The CART regression tree is using the principle of minimum mean square deviation (MSE). CART starts by growing a very large tree and then pruning it.

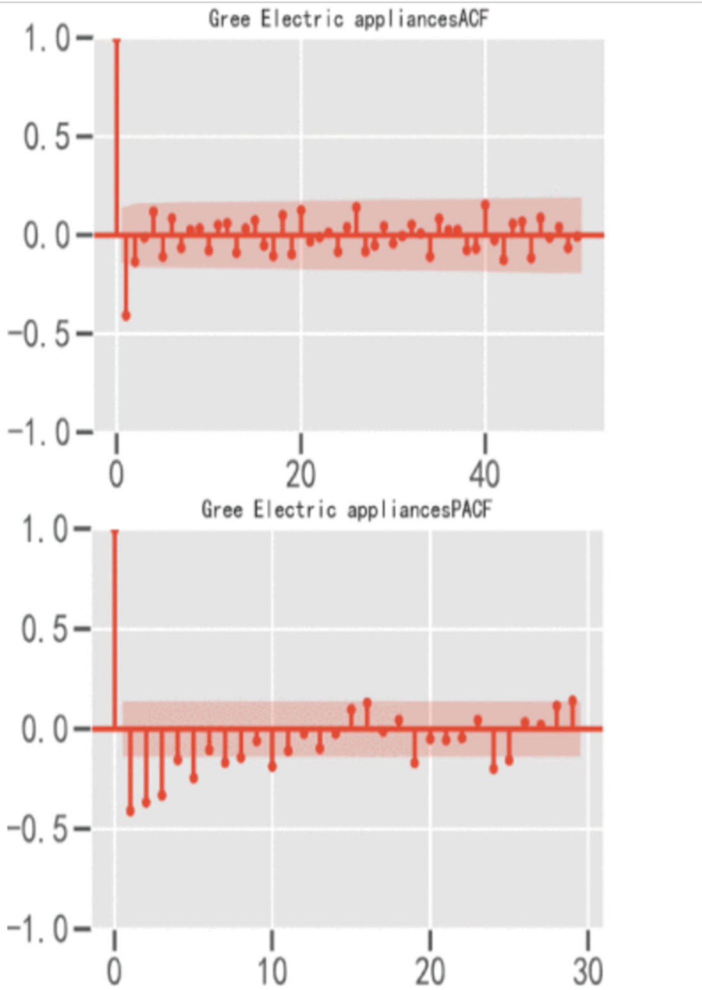

Fig. 1 The ACF and PACF of ARIMA.

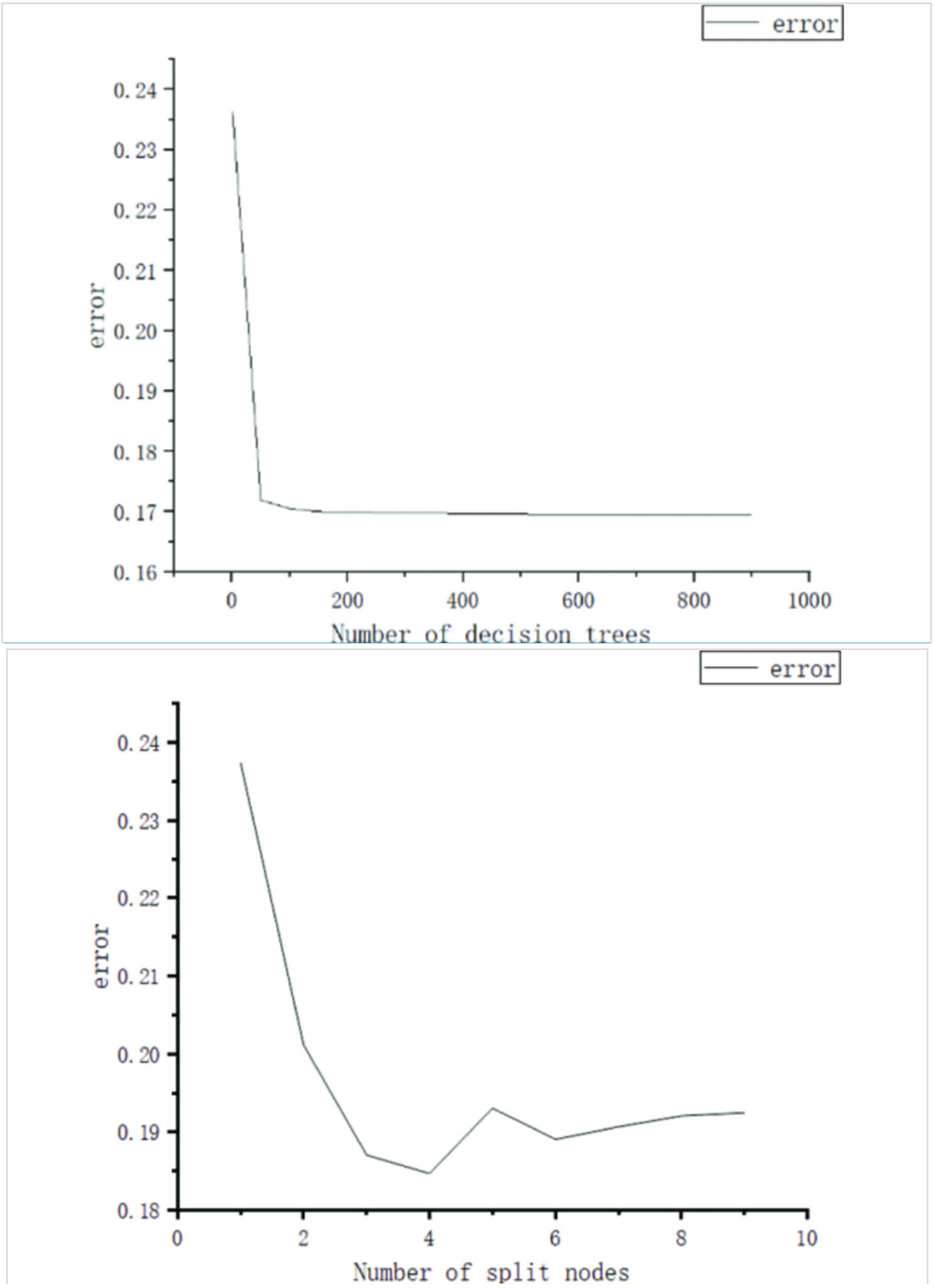

Fig. 2 The error of random forest as a function of tree numbers (upper panel) and split nodes (lower panel).

4. Implementations & Applications

Fig. 1 and Fig. 2 present the ARIMA model and random forest model performances. Through the comparation and analysis of the ARIMA model, it is found that the linear modal does not have a high fitting degree in general [8]. In fact, it is not easy to overfit.

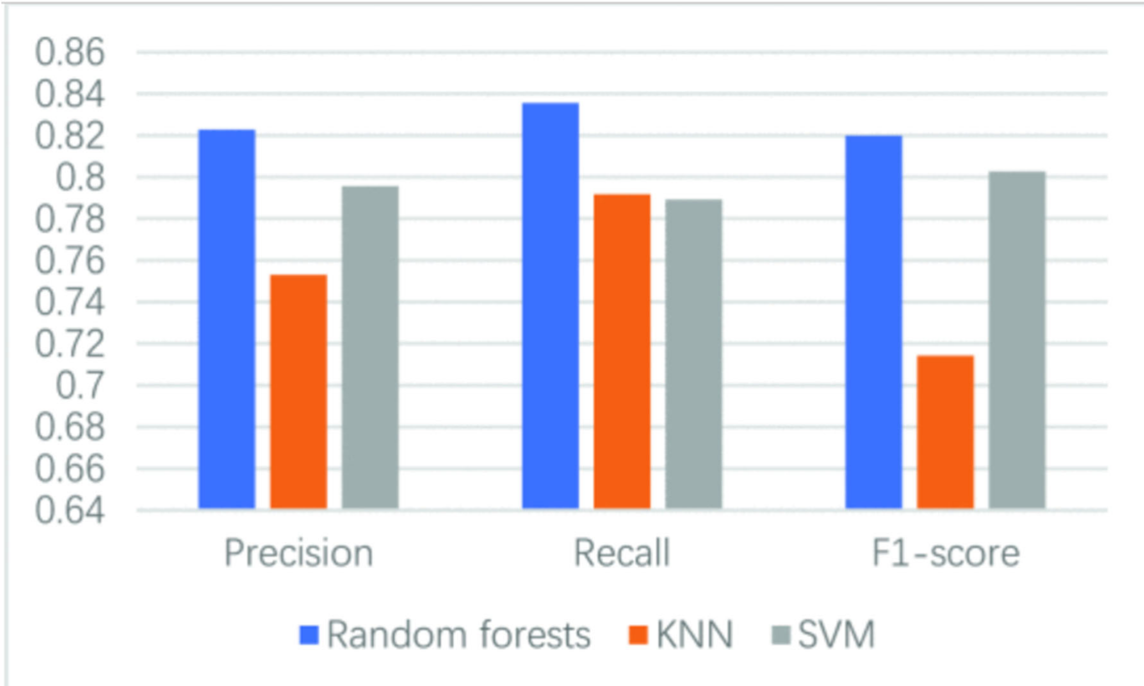

Fig. 3 Comparison of the performance of random forest, KNN and SVM.

In addition, this model also requires only endogenous variables, not any exogenous variables. A comparison of random forest with other models is given in Fig. 3, where one sees clearly that it has the best performances among the three models. Analyzing the results of random forest model, it is found that the random forest model is faster in training speed and easy to make parallel method. In addition, it's relatively simple to implement. The random forest regression algorithm is used in the scenario with relatively low requirement for data dimension and relatively high requirement for accuracy. Besides, the RF can process not only the discrete data, but also continuous data.

5. Limitations & Prospects

Stock prediction is a very magnetic research subject. An accurate prediction of stock market is able to acquire high returns [9]. Recent research shows that the stock returns can be partly predictable in the long run. Nevertheless, it is like the weather forecast, stock prediction cannot foresee the stock trends too far ahead. The best predictor of future returns is not the historical average [10]. Sometimes use only one stock forecasting method is prone to errors [11]. Although the future changes based on the past frequently, there is nothing wrong with empirically predicting the future. Nevertheless, history does not repeat itself. As a consequence, it is also not advisable to focus only on the past. Furthermore, the handling of data values is also important. A small deviation from the initial value can lead to a huge difference in the results.

It is more reliably to use a combination of several prediction means. It believes that with the current academic development and the degree of the future knowledge's perfection, the stock forecasts will be more common and reliable in the future. Stock price adjustments based on information should be prompt [12].

6. Conclusion

In summary, this paper compares the stock prediction's process of ARIMA and the random forest model. To be specific, this article described the methods and steps of building the two models in detail and compare the applications. According to the analysis, the ARIMA model is relatively simple to implement in the short-term stock prediction and the random forest model has higher accuracy. Moreover, it is able to run efficiently on large data sets. However, there should not be too many decision trees in the random forest. Nevertheless, stock prediction is not good at making the long-term stocks trend forecasts. Additionally, minor errors in the processing of data can lead to vastly different results. In the future, with the establishment of more stock prediction models, it is potentially that the longer-term stock tendencies will be more accurately predicted. Overall, these results offer a guideline for stock price prediction based on machine learning scenarios.

References

[1] Zheng X, Cai J, Zhang G. Stock Trend Prediction Based on ARIMA-LightGBM Hybrid Model. 2022 3rd Information Communication Technologies Conference (ICTC). IEEE, 2022: 227-231.

[2] Yao Y. Data Analysis on the Computer Intelligent Stock Prediction Model Based on LSTM RNN and Algorithm Optimization. 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA). IEEE, 2022: 480-485.

[3] Gao X, Qiu M, He Z. Big Data Analysis with Momentum Strategy on Data-driven Trading. 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). IEEE, 2021: 1328-1335.

[4] Lu C J, Chang C H, Chen C Y, et al. Stock index prediction: A comparison of MARS, BPN and SVR in an emerging market. 2009 IEEE International conference on Industrial Engineering and Engineering Management. IEEE, 2009: 2343-2347.

[5] Youness J, Driss M. An ARIMA Model for Modeling and Forecasting the Dynamic of Univariate Time Series: The case of Moroccan Inflation Rate. 2022 International Conference on Intelligent Systems and Computer Vision (ISCV). IEEE, 2022: 1-5.

[6] Chatterjee A, Bhowmick H, Sen J. Stock price prediction using time series, econometric, machine learning, and deep learning models. 2021 IEEE Mysore Sub Section International Conference (MysuruCon). IEEE, 2021: 289-296.

[7] Grömping U. Variable importance assessment in regression: linear regression versus random forest. The American Statistician, 2009, 63(4): 308-319.

[8] Lai S, Ye C, Zhou H J H. Chinese stock trend prediction based on multi-feature learning and model fusion. 2021 IEEE International Conference on Smart Data Services (SMDS). IEEE, 2021: 18-23.

[9] Reichenstein W, Rich S P. Predicting long-horizon stock returns: Evidence and implications. Financial Analysts Journal, 1994, 50(1): 73-76.

[10] Thakkar A, Chaudhari K. A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. Expert Systems with Applications, 2021, 177: 114800.

[11] Ball R. The theory of stock market efficiency: accomplishments and limitations. Journal of Applied Corporate Finance, 1995, 8(1): 4-18.

[12] Suha S A, Sanam T F. A Machine Learning Approach for Predicting Patient's Length of Hospital Stay with Random Forest Regression. 2022 IEEE Region 10 Symposium (TENSYMP). IEEE, 2022: 1-6.