Check for updates

Comparison Analysis of ARIMA and Machine Learning Methods for Predicting Trend of US Semiconductor Stocks

Mingtao Jia[1](✉), Haichen Xu[2](✉), and Sicong Zhang[3](✉)

[1] Business School, University of Sydney, Sydney 00026A, Australia
mjia6176@uni.sydeny.edu.au
[2] College of Art and Science, Boston University, Boston 02215, USA
xhcbu@bu.edu
[3] Warren College, University of California (Sandiego), Sandiego 92092, USA
s3zhang@ucsd.edu

Abstract. The stock price trend prediction has some challenges for the investors because there are many unknown risks and great variation in the stock market. Some researchers have studied how to give the prediction of the stock price trend with high accuracy. However, the systematic analysis of the comparisons for this field is still insufficient. In this paper, the Arima and machine learning methods are applied to predict the trend of US semi-conductor stocks. The comparison analysis of the Arima-based method and machine learning-based methods are given to evaluate their performances. The comparison results indicate that the Arima-based method has a better performance than that of machine learning methods in the application of fitting the variation of the stock prices. Our research has great significance in the application of stock price trend prediction.

Keywords: component · Stock Price Trend Prediction · Arima · Machine Learning · Linear Regression · Random Forest · Decision Tree · Gradient Boosting

1 Introduction

Trend prediction of the stock market is regarded as a financial time-series task that has a great challenge [1]. Practically, plenty of features related to stock trend classification make the problem being hard to solve. Especially, some of the features are not relevant, while some are redundant from the viewpoint of the field of machine learning. Stock market trends prediction with high accuracy has a great significance for the investors because the precise prediction is very useful for the investors to obtain the best return [2]. Unfortunately, the irrelevant and redundant information may cause the false result of some machine learning algorithms [1, 3].

Many researches have been applied to model and predict the time-series data. The early researchers focused on predicting the stock market using ANN, which achieved a

M. Jia, H. Xu and S. Zhang—These authors contributed equally.

© The Author(s) 2022
F. Balli et al. (Eds.): ESFCT 2022, AEBMR 226, pp. 1607–1614, 2022.
https://doi.org/10.2991/978-94-6463-052-7_178

good result. For example, Kim and Han [4] proposed a genetic algorithms approach, an approach they claimed to decrease the dimensionality of the feature space and enhance the performance of the forecasting, to feature discretization and the determination of connection weights for ANN to predict the stock price index. However, some of these studies indicated that ANN had some drawbacks in learning the patterns because stock market data has enormous noise and complex dimensionality. Support vector machine (SVM) proposed by Vapnik [5] implements the structural risk minimization principle. However, the traditional neural networks implement the empirical risk minimization principle. Practically, the neural networks-based method is more easily occur over-fitting than the SVM-based. This is because the SVM may be the global best choice. However, the neural network models tend to fall into local optimal solutions. For example, Kim [6] proposed the SVM-based method to predict the stock price trend, which achieves 57% accuracy. Hassan et al. [7] proposed a fusion model which combines the Hidden Markov Model (HMM), ANN, and Genetic Algorithms (GA) to predict the trend of stock price, which achieves a good result. Their proposed method uses the features extracted by ARIMA analyses [8].

Although there are so much researches about the trend prediction of stock prices, the research about US Semiconductor Stocks trend prediction is still insufficient. The US Semiconductor Stocks have the special characteristics that they are more stable to resist the risk. Therefore, we compare the performance of ARIMA and Machine Learning Methods for predicting the Trend of US Semiconductor Stocks in this paper. Firstly, we use time series model to represent the five selected stocks; secondly, we use the Arima model to forecast the stock price (or trend) in the next time stamp; finally, we compare and analyze the performances of different methods in the application of stock price prediction, such as Arima and machine learning-based method.

The most contributions of our work include 1) comparison of Arima and machine learning-based methods for predicting the price trend of US Semiconductor Stocks market;2) a novel stock selection is proposed based time series.

Section 2 introduces the methods; Sect. 3 introduces the results and discussion; Sect. 4 introduces the conclusion.

2 Methods

In this paper, the Arima-based method and machine learning-based method are compared to predict the trend of US semiconductor stock price.

2.1 Data Preparation

In this paper, the US semiconductor stocks are used as the object because the semiconductor has a good characteristic that they all have good stability in resisting the risk. And their characteristics are very useful to decrease the difficulty of the problem. The five selected stocks include "Intel Corporation", "Micron Technology", "Qualcomm Incorporated", "Nvidia Corporation," and "Texas Instrument Incorporated". They are all the typical corporations in the US semiconductor market. Please note that all of the data set derives from the Yahoo Finance website [9].

2.2 Arima-Based Trend Prediction

The Arima model has been proposed by Box and Jenkins [10], which is also called the Box-Jeckins methodology. It mainly consists of identifying, estimating, and diagnosing of time series data. The Arima model has been frequently used for financial forecasting [11–13]. The Arima model is a useful tool for generating forecasts in a short period. In many occasions, it performs better than those much complicated structural models in short-period predictions.

ARIMA is used for prediction in this section. By programming the code, we drew the graph and calculated the value of the p, d, q based off result of the graph. The result would be obtained by looking at the graph, predictions on future movement of five companies' stock prices could in turn be made.

The following formula (1) could be utilized to select the best model,

$$Y_t = \varnothing_t Y_{t-1} + \theta_0 + \epsilon_t \tag{1}$$

where $\epsilon_t = Y_t - \widehat{Y_t}$ is real value minus forecasted value.

2.3 Machine Learning-Based Trend Prediction

In this paper, three machine learning methods (random forest, decision tree, and gradient boosting) and linear regression are tested and compared.

2.3.1 Linear Regression Model for Trend Prediction

Assume there are N number of distinguished classes with p_i number of training stocks from the ith class (i = 1,2,…, N). Each training stock is of an order $a \times b$ and is presented as $\overbrace{u_i}^{(m)} \in R^{a\times b}$, i = 1,2, …, N. The sequence of stocks can be represent as X = {x1, x2, …, xN}. And the linear combination of the training stocks from the same class can be calculated by the formula (2),

$$y = x_i\,\beta_i,\ i = 1, 2, 3, \ldots N. \tag{2}$$

where y means the ith class; the $\beta_i \in R^{p_i\times 1}$. In fact, the trend prediction using linear regression in this paper is binary classification.

2.3.2 Random Forest Model for Trend Prediction

As a supervised learning method, the random forest is proposed by Breiman [15] and utilizes ensemble regression learning method. Ensemble regression learning method aggregates predictions from many machine learning to achieve more accurate forecasts than a single model.

Assuming classifier m_t to be consistent if probability error,

$$L(m_n) = \mathbb{P}[m_n(X) \neq Y]_{x\to\infty} \to L^* \tag{3}$$

Random Forest classifier is obtained by a majority vote among classification trees, such that

$$\begin{cases} m_{M,n}(X; \Theta_1, \cdots, \Theta_M, D_n) = 1, \ \textit{if} \ \frac{1}{M} \sum_{j=1}^{M} m_n\big(X; \Theta_j, D_n\big) > \frac{1}{2} \\ m_{M,n}(X; \Theta_1, \cdots, \Theta_M, D_n) = 0, \ \textit{otherwise} \end{cases} \tag{4}$$

For each random tree classifier,

$$\begin{cases} m_n\big(X; \Theta_j, D_n\big) = 1, \ \textit{if} \ \sum_{i \in D_n^*(\Theta_j)} 1_{X_i \in A, Y_i = 1} > \sum_{i \in D_n^*(\Theta_j)} 1_{X_i \in A, Y_i = 0} \\ m_n\big(X; \Theta_j, D_n\big) = 0, \ \textit{otherwise} \end{cases} \tag{5}$$

2.3.3 Decision Tree Model for Trend Prediction

Another widely employed machine learning method in classifications is decision tree. Decision tree consist of internal nodes which indicate tests on features, and branches which illustrate the results. After follow-up processing, decisions are made and are expressed as leaves.

Subsets with information of similar feature attributes are made. Every subset in the tree will reach a leaf before the repeating process stops. From the root of the tree, class label prediction for a record in the decision tree starts. Values between root attributes and next record attributes are compared. Next node's value will depend on the result from the comparison. Specifically in this paper, prediction of future movements of stock prices are treated as binary classification problems and decision tree method will be used.

2.3.4 Gradient Boosting Machine for Trend Prediction

Gradient boosting is another machine learning method suitable for both classification and regression problems. It also uses ensemble learning technique and produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Similar to other boosting methods, the model is built in step-wise fashion and generalization is done by optimizing an arbitrary differentiable loss function.

3 Results and Discussion

In this paper, the results include two parts: The Arima results for stock price future movement forecasts and its comparisons with multiple machine learning methods for future movement forecasts.

3.1 Results of Arima-Based Trend Prediction

The results of Arima-based trend prediction can be found in Fig. 1.

In Fig. 1, subfigure (a) represents the result of TXN; subfigure (b) represents the result of QCOM; subfigure (c) represents the result of NVDA; subfigure (d) represents the result of MIRCON; subfigure (e) represents the result of INTC.

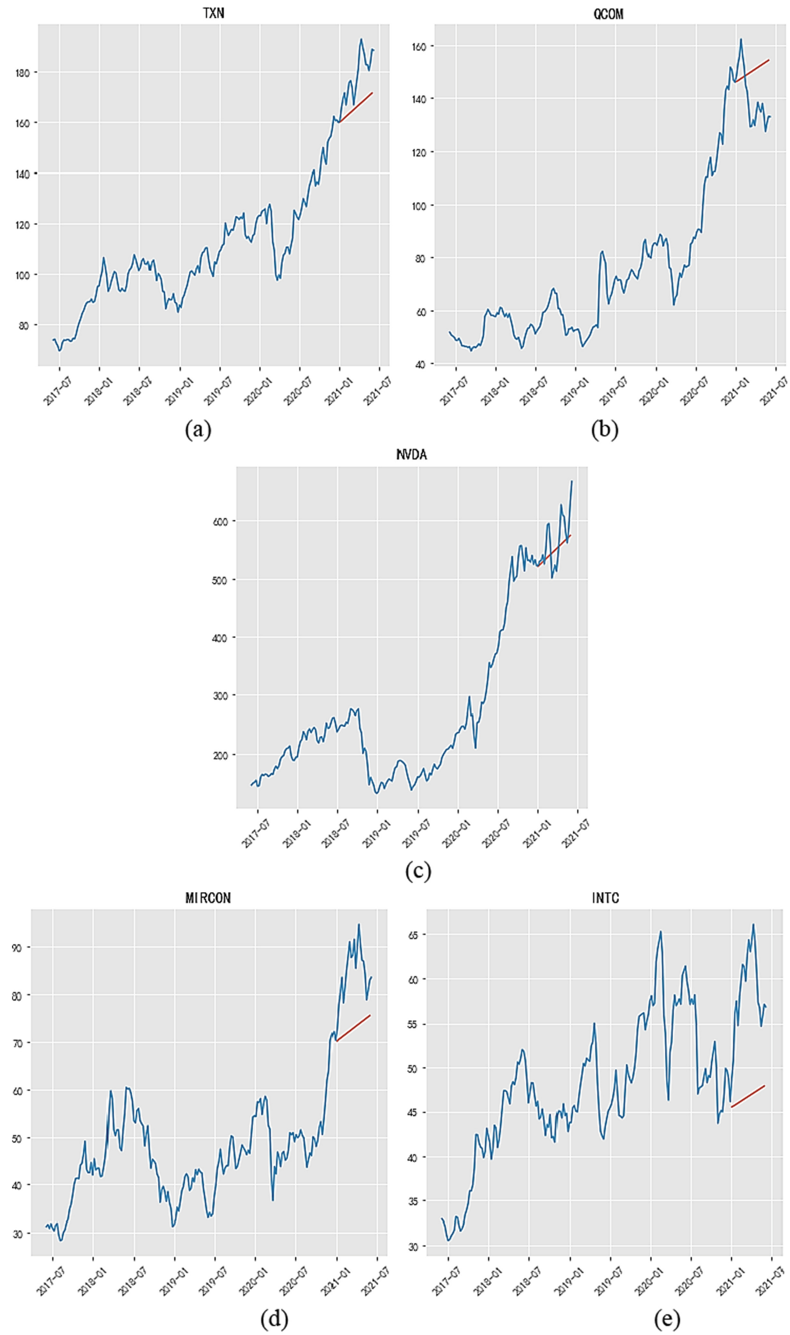

Fig. 1. Trend Prediction of Stocks Price using Arima-based Method. Red line is the prediction of company stock trend. Blue line is the actual trend a) Texas Instrument Incorporated (b) Qualcomm Incorporated (c) Nvidia Corporation (d) Micron Technology (e) Intel Corporation

INTC: Although it has an upward trend, it has quite a difference in the real stock situation between January and June. Therefore It is not very accurate for this company to use arima by now.

MICRON: Although the prediction is largely correct about the general trend, it fails to consider the drastic fluctuations in the six months. Only limited information could be derived from this prediction.

Nvidia: As we can see, this company also has an increasing incline, it corresponds to the data that it provides, so It is somewhat useful for predicting.

Qualcomm: As we can see, this company also has an increasing incline, it corresponds to the data that it provides, so It is somewhat useful for predicting.

TXN prediction: Almost describing the trend between five months, but as we can see, the incline of these models is similar among these five companies. It is also not that accurate.

3.2 Results of Machine Learning-Based Method for Trend Prediction

The results of trend prediction using machine learning-based methods can be found in Fig. 2.

In Fig. 2, subfigure (a) represents the result of INTC; subfigure (b) represents the result of MIRCON; subfigure (c) represents the result of QCOM; subfigure (d) represents the result of TXN; subfigure (e) represents the result of NVDA.

Linearity is the advantage of linear regression which simplifies the estimation procedure and make interpretations of equations and results easy on a modular level.

However, linear regression makes assumptions on the dataset (e.g., data are i.i.d, no significant outliers). If these assumption are violated, results from linear regression will be unreliable (e.g., outliers could distort the prediction). Linear regression only takes into account the relationship between mean of the dependent variable and the independent variables.

Flexibility is the advantage of random forest. Suitable for both classification and regression problems, random forest demonstrates the relative importance it assigns to the input features clearly.

Random forests excel at predicting non-linear relationship. However, it usually suffers from overfitting and its interpretation could be problematic since it's composed of many decision trees. In practice, random forest algorithm is usually computational costly and inefficient to implement.

Comprehensiveness is the major advantage of decision tree because its algorithm searches through all possible outcomes of a decision and traces each path to a conclusion. Then it identifies nodes that need further investigations.

The decision trees also have its disadvantages: Their instability means that a tiny change in the data can cause a huge change in the structure of the optimal decision tree. Usually, they are relatively inaccurate. By contrast, with similar data, many other predictors perform better.

The advantage of the gradient boost is that it has plenty of flexibilities, which can optimize on different loss functions and provide some hyper parameters tuning options that make the function fit extremely flexible.

Fig. 2. Comparison of different machine learning-based Trend Prediction. The yellow, green, blue and red line each represents the results derived from linear regression, decision tree, random forest and gradient boosting methods.

Of the disadvantage, boosting is sensitive to outliers because every classifier is obliged to fix the errors in the predecessors. Therefore, the outliers can influence this method in a great extent.

4 Conclusion

In this paper, the comparison of the Arima-based method and the machine learning method is used to analyze which method is more suitable in applying trend prediction of stocks. After several evaluation tests, we can conclude that the Arima-based method better predicts stock price trend prediction than the machine learning methods. In addition, our research has a reference significance in the field of stock price trend prediction in the US semi-conductor stock market.

References

1. Huang C J, Yang D X, Chuang Y T. Application of wrapper approach and composite classifier to the stock trend prediction. Expert Systems with Applications, 2008, 34(4): 2870-2878.
2. Kara Y, Boyacioglu M A, Baykan Ö K. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert systems with Applications, 2011, 38(5): 5311-5319.
3. Kumar I, Dogra K, Utreja C, et al. A comparative study of supervised machine learning algorithms for stock market trend prediction. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2018: 1003–1007.
4. Kim K, Han I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert systems with Applications, 2000, 19(2): 125-132.
5. Vapnik, V. N. Statistical learning theory. New York: Wiley, 1998
6. Kim K. Financial time series forecasting using support vector machines. Neurocomputing, 2003, 55(1-2): 307-319.
7. Hassan M R, Nath B, Kirley M. A fusion model of HMM, ANN and GA for stock market forecasting. Expert systems with Applications, 2007, 33(1): 171-180.
8. Abraham A, Nath B, Mahanti P K. Hybrid intelligent systems for stock market analysis. International Conference on Computational Science. Springer, Berlin, Heidelberg, 2001: 337–345.
9. Website: https://finance.yahoo.com/
10. Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE, 2014: 106–112.
11. Pai P F, Lin C S. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 2005, 33(6): 497-505.
12. Merh N, Saxena V P, Pardasani K R. A comparison between hybrid approaches of ANN and ARIMA for Indian stock trend forecasting. Business Intelligence Journal, 2010, 3(2): 23-43.
13. Nochai R, Nochai T. ARIMA model for forecasting oil palm price. Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and applications. 2006: 13–15.
14. Meyler A, Kenny G, Quinn T. Forecasting Irish inflation using ARIMA models. 1998.
15. Breiman L. Random forests. Machine learning, 2001, 45(1): 5-32.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.