Machine Learning for Stock Prediction by Different Models

Liurui Shi[1,*]

[1] *Department of Mathematics, University College London, London, WC1E 6BT, UK*
[*] *Corresponding author. Email: liurui.shi.20@ucl.ac.uk*

ABSTRACT

Machine learning is a big and popular topic in recent years and is applied wildly in the field of finance to assist researchers in analyzing the tendency of financial assets in the global market as well as the local market. However, predicting stocks or a portfolio is a challenging task due to the uncertainties and randomness of the financial market. Different models have different structures and therefore they have different performances in reducing the uncertainties in the financial field. This paper investigates the impact of Covid-19 on the accuracy of different machine learning techniques and analyzes the effect of walk-forward validation on the stock prediction. The experimental result indicates that the ARIMA model with the use of walk-forward validation has the performance for forecasting the stock price and walk-forward validation improves the accuracy of forecasting and reduces the errors of the models compared to simple time series splitting. So the technique of walk-forward validation is useful to be implemented in the stock price prediction to maximize the capital gain and minimize the analytical error due to uncertainties.

Keywords: *Covid-19; Forecasting; ARIMA; Accuracy; Walk-forward validation*

1. INTRODUCTION

1.1 Background

In the previous years, scientists propose a series of models based on traditional approaches such as the linear regression model and a couple of machine learning techniques in the field of the quantitative financial market such as the random forest model to make an accurate data representation and find precise dependencies between data.

However, the stock price prediction is a challenging job due to the impact of a variety of factors. When the scale of data becomes larger with more dimensions, larger complexity, and randomness, some models do not perform as well as the others which might result in huge loss in the financial market because of an inaccurate stock prediction. For instance, as the breakout of global health issues such as Covid-19, the human society is affected by Covid-19 seriously and many companies' sales are also influenced because of Covid-19 and all these huge changes will reflect on the financial market such as stock price. So this paper discusses several machine learning techniques such as linear regression (LR), random forest (RF), support vector regression (SVR), and ARIMA models in terms of short-term stock price prediction with and without Covid-19 in order to find the extent of influence of Covid-19 on the predictive power of different machine learning techniques.

1.2 Related research

Rundo et al. described different models and techniques and discussed the performance of auto-regressive models developed for financial market applications such as the ARIMA model, KNN model, and Support Vector Machine (SVM). They also discussed some hybrid approaches and illustrated the comparisons between traditional and ML-based approaches, and demonstrated the advantages of Deep Learning models. This research concludes that ML-based algorithms have an overall better performance than the traditional algorithms in terms of accuracy [1].

Bhuriya et al. used the linear regression (LR) model to forecast the stock market price of the TCS data set and compared the Linear Regression (LR) model with the Polynomial model and radial basis function (RBF) model. The experiment chose the open price, high price, low price, and Number of trends of stocks as input-independent variables and chose stock close price as the target-dependent variable to make stock price prediction,

© The Author(s) 2022
Y. Jiang et al. (Eds.): ICEDBC 2022, AEBMR 225, pp. 318–323, 2022.
https://doi.org/10.2991/978-94-6463-036-7_48

then they concluded that linear regression had the best result compared to the other models in terms of confidence value [2]. Siew and Nordin tested regression techniques for the prediction of stock price trends with a transformed data set and compared the result with regression techniques with pre-transformed data set. They concluded that the performance of various regression approaches such as Linear Regression and Sequential minimal optimization (SMO) can be improved by transforming the input data [3].

Ayala et al. investigated an algorithm to generate trading signals with the use of technical indicators and machine learning techniques, and their research constructed a trading decision-making workflow to make practical buy and sell signs, and then the result found that the Linear model and ANN model had the best predictive power compared to Random Forests and Support Vector Regression models [4].

Bini and Mathew proposed an analysis system that used data mining techniques such as clustering and regression. A validation index was used to analyze the performance of various clustering approaches. They concluded that the K-means algorithm in partitioning-based technique and EM algorithm in model-based technique have better performance than other clustering algorithms [5].

Huang et al. investigated three machine learning algorithms for stock prediction based on fundamental analysis, and they applied feature selection and bootstrap aggregation to boost the performance of models and prediction stability. The experiment concluded that the Random Forest (RF) model behaved the best, and the prediction result can be improved by feature selection [6]. Kumar et al. evaluated reviewed a couple of supervised machine learning models and illustrated that the Random Forest (RF) model had the highest predictive accuracy for big datasets and Naïve Bayesian Classifier made the best predictive result in the small datasets. They also observed that the quantitative variation of technical indicators affects the accuracy of each machine learning algorithm [7].

Patel et al. worked on the issue of identifying the moving direction of the stock and its index in the Indian stock markets. The study compared various machine learning techniques for stock prediction and demonstrated that the random forest model obtained the best result in stock prediction with ten technical parameters. They also pointed out that the accuracy of all machine learning models was improved by the use of trend deterministic data as input data [8].

Khaidem et al. determined to regard the prediction problem as a classification problem to reduce the prediction error. They predicted the stock price with the use of ensemble learning and some technical indicators, and the predictive model constructed by a random forest classifier had a very good result in stock prediction [9].

Kumar and Thenmozhi investigated the predictability of the direction/sign of stock index movement and compared various classification techniques related to time series forecasting. The experiment predicted the daily movement of the direction of the S&P CNX NIFTY Index and concluded that the SVM model had the best effect in identifying the tendency of movement of the stock market because of the implementation of the structural risk minimization principle of the SVM model [10].

Xia et al. designed a framework to predict the stock market with the application of the Support Vector Regression (SVR) technique and concluded that SVR had a strong predictive power in recognizing the moving direction of stocks [11]. Ye used wavelet analysis to deal with data and made prediction results for stock prices with the use of the ARIMA model and the SVR model and the experiment showed that the mixed-use of the ARIMA model and SVR model could effectively improve the accuracy of prediction [12].

1.3 Objective

Therefore, this paper will focus on investigating how the performance of different models or techniques differs for multiple stocks before and after the breakout of Covid-19 with the use of the variable-controlling method and error statistics, and find the model with the best performance in predicting the short-term stock price. The study investigates the prediction for AAPL Inc. (AAPL), Pfizer Inc. (PFE), Tesla, Inc. (TSLA) stocks by various models before and during the Covid-19 and does the performance evaluation and statistical metrics such as mean absolute error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) for different models.

2. DATA AND METHODS

2.1 Model

2.1.1 Linear Regression (LR) Model

Linear Regression Model is a statistical approach for modelling the relationship between a dependent variable y and one or more independent variables x. For two variables x and y, we can use the linear equation $y = kx + b$ where k and b are constant to find the relationship between the two variables x and y.

2.1.2 Random Forest (RF) Model

Random Forest is an ensemble learning method for classification and regression that predicts the stock price by constructing multiple decision trees in the training period. An ensemble method means that the random tree

is made up of multiple small trees and each tree is called an estimator and will make its prediction based on its positions. Then random forest makes a more accurate prediction by combining the mean value of all the predictions from small trees. And many hyperparameters can be used for performance optimization such as the number of estimators/trees, the maximum depth, minimum sample split, etc and they can deal with the issue of overfitting in the process of building the model based on training data.

2.1.3 Support Vector Regression (SVR) Model

Support Vector Machine (SVM) is a supervised learning method for data classification and regression and can be used to solve non-linear regression issues. The goal of SVM is to separate data correctly with the maximum margin. Support Vector Regression (SVR) is a variant of Support Vector Machine (SVM) and is a supervised machine learning algorithm that is used for regression tasks for stock prediction. SVR is characterized by kernel function, sparsity of the solution, and the capacity manipulation of decision function and it is equipped with different parameters such as kernel function, degree, gamma, and regularization parameter (C).

2.1.4 Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA model is one of the time series forecasting statistical models that uses time-series data to make predictions about the future stock price. ARIMA model is generated by combining the advantages and solving the disadvantages of previous models such as the moving average model (MA), autoregressive model (AR), and autoregressive moving average model (ARMA). ARIMA model has three parameters, and they are p, d, and q where p is the number of lag observations, d is the number of differences required to make the time series stationary, and q is the order of the moving average term. ARIMA model requires stationarity so if time series is not stationary then it is required to be differentiated until it becomes a stationary time series.

2.2 Data preparation

The sample dataset is the AAPL, PFE, and TSLA stock between 01/2010 and 11/2017 from Kaggle and between 12/2019 and 03/2022 from Yahoo Finance. Historical financial data before Covid-19 for AAPL, PFE, and TSLA stock was retrieved from our downloaded local file in csv format and historical financial data after Covid-19 for the same stocks were obtained by importing data online directly. The data analysis library pandas are used to convert data in csv file to pandas DataFrames which was indexed by dates and the Python NumPy library was used to deal with our data in the pandas DataFrames. The date is converted to daytime and follows the time sequence.

After carefully checking, the original data does not have any missing values so there is no need to fit any missing value by performing data imputation.

2.3 Model construction

For the linear regression model, we convert date to daytime in order to construct our linear model by time series and set X = daytime and y = Close price and split the dataset into X_train, X_test, y_train, y_test where the size of X_test and y_test is 2. Then we use the fit method to build the linear model between X and y with the use of X_train and y_train and then forecast stock prices for the next two days and do a performance evaluation.

In the Random Forest Model, this paper will choose the number of trees to be 500 and the maximum depth in each decision tree to be 3 in order to avoid overfitting, and also choose min_samples_split to be two and min_samples_leaf to be one, and bootstrap is True which means that sampling the data with replacement. This paper uses the same dataset in the same period as the linear regression so that it is possible to compare the performance evaluation of different models. In this model, we set X = the Date, Open, High, Low columns, and y = the Close column. Then the experiment splits the sample data into X_train, X_test, y_train, and y_test where the size of X_test and y_test is 2, and then fits the training data (X_train) and transforms the testing data (X_test) by the function of standard scalar for the sake of avoiding data leakage. The Standardization boosts the stability of the model and increases the training speed. Then using the RandomForestRegressor model to predict the stock price in the next two days and do the performance evaluation.

For the SVR model, this paper considers SVR models with three different kernel functions: polynomial with degree 1, polynomial with degree 2, and radial basis function (RBF) model. In each SVR model, the regularization parameter C = 1000. In the SVR model with the RBF kernel function, gamma = 0.15. Then repeating the same application in the linear regression model to split the dataset into X_train, X_test, y_train, y_test. Then construct the SVR model with X_train and y_train and make predictions for the stock price in the next two days with the use of X_test, and then evaluate the results.

Walk-Forward Validation is where a model is updated at each time step when new data is received.

For ARIMA model, this paper sets the parameters p = 9, d = 1, q = 3 and implements walk-forward validation for ARIMA model. The experiment splits the dataset into the historical close price and actual close price. Fitting ARIMA model with the historical close price of stocks

and then forecasts the stock price for the next trading day and then add the actual stock close price of the next trading day into the historical close price, and train ARIMA model again and then make the stock prediction for the second trading day. Predicted stock close prices for the two trading days in the future are collected to do a performance evaluation.

3. RESULTS

This paper analyzes the two-day prediction for the above companies before and after Covid-19 and makes the below table to show the performance evaluation of different models for different stocks between these two periods (before and after Covid-19).

Table 1. The evaluation for two-days prediction of AAPL stock by various models by error statistics

Stock₩Metrics	Model	MAE	MSE	RMSE	Accuracy (%)
AAPL(Before Covid-19)	Linear model	36.531	1335	36.5324	79.12
	Random forest	18.3219	335.7778	18.3242	89.53
	SVM (with polynomial model of degree 1)	37.4441	1402	37.4454	78.6
	SVM (with polynomial model of degree 2)	17.6833	313	17.6863	89.89
	SVM (with RBF model)	21.082	588	24.255	87.94
	ARIMA model	0.7466	1	0.7467	99.57
AAPL(After Covid-19)	Linear model	1.6159	5	2.2793	99.1
	Random forest	7.6815	61.8276	7.8631	95.68
	SVM (with polynomial model of degree 1)	2.2067	7	2.7357	98.76
	SVM (with polynomial model of degree 2)	11.8527	143	11.9551	93.31
	SVM (with RBF model)	50.3514	2560	50.601	71.62
	ARIMA model	1.9194	4	2.084	98.92

Table 2. The evaluation for two-days prediction of PFE stock by various models by error statistics

Stock₩Metrics	Model	MAE	MSE	RMSE	Accuracy (%)
PFE(Before Covid-19)	Linear model	0.9161	1	0.9163	97.4
	Random forest	1.7553	3.0811	1.7553	95.01
	SVM (with polynomial model of degree 1)	0.9718	1	0.9719	97.24
	SVM (with polynomial model of degree 2)	4.1941	18	4.1941	88.08
	SVM (with RBF model)	3.4908	14	3.8068	90.08
	ARIMA model	0.1253	0	0.1507	99.64
PFE(After Covid-19)	Linear model	3.3164	11	3.3283	93.75
	Random forest	0.27	0.0752	0.2741	99.49
	SVM (with polynomial model of degree 1)	5.3818	29	5.3891	89.85
	SVM (with polynomial model of degree 2)	1.2442	2	1.2781	97.66
	SVM (with RBF model)	10.956	121	10.9823	79.32
	ARIMA model	1.0189	1	1.0229	98.08

Table 3. The evaluation for two-days prediction of TSLA stock by various models by error statistics

Stock₩Metrics	Model	MAE	MSE	RMSE	Accuracy (%)
TSLA(Before Covid-19)	Linear model	16.6976	279	16.6977	94.49
	Random forest	1.6376	2.6819	1.6376	99.46
	SVM (with polynomial model of degree 1)	19.4284	377	19.4285	93.59
	SVM (with polynomial model of degree 2)	59.1979	3504	59.198	80.46
	SVM (with RBF model)	52.1263	3292	57.379	82.8

	ARIMA model	0.7718	1	0.7942	99.75
TSLA(After Covid-19)	Linear model	30.8022	959	30.9749	97.19
	Random forest	15.6886	261.0712	16.1577	98.57
	SVM (with polynomial model of degree 1)	94.7751	8993	94.8327	91.35
	SVM (with polynomial model of degree 2)	81.5956	6665	81.6409	92.55
	SVM (with RBF model)	454.5705	208556	456.6788	58.53
	ARIMA model	43.3306	3402	58.3271	96.03

The result shows that the ARIMA model has the best and most stable performance in predicting the stock price among all machine learning algorithms because the ARIMA model has the lowest MAE, MSE, and RMSE compared to other models and the accuracy of the ARIMA model is very high and stable for different stocks during various periods.

Random forest and linear regression models also have good predictive power for stock price prediction. During the Covid-19, the accuracy of the SVM model with kernel function RBF and ARIMA model is decreased and the accuracy of the other models in this paper is not changed obviously or improved compared to their accuracy before Covid-19.

4. DISCUSSION

ARIMA model has the best performance because ARIMA model is a time series model so it is suitable to deal with time-series data like the stock price and forecast the stock price in the future and the parameters p, d, q in ARIMA(p, d, q) model can be adjusted so that time series is differentiated d times and is shifted from non-stationary time series into stationary time series. And then ARIMA model can make predictions based on the stationary time series. ARIMA model uses walk-forward validation when every time a new data is produced, it is added to the train set, and then the ARIMA model will be trained and fitted again and then make new predictions because the accuracy of the prediction of stocks will be less and less over time for time-series modelling. Therefore, the predictive result of the ARIMA model can be adjusted and become closer to the current tendency of stock due to newly added data. So when predicting time series data such as stock price, walk-forward validation is a very powerful strategy that can reduce the errors and enhance the accuracy of models by continuously training the models with latest data so that the models always make predictions based on current information.

5. CONCLUSION

The stock trend is difficult to be recognized because of the randomness and instability of the global market. This paper investigates the influence of Covid-19 on the errors and accuracy of different models and concludes that although the accuracy of the ARIMA model is reduced by one bit during the Covid-19 period, as a time series model, the ARIMA model still has the best performance in the stock price prediction among all machine learning algorithms in this paper. Compared to a simple time-series split, walk-forward validation greatly improves the accuracy of models. This paper only considers short-term predictions and does not use technical indicator so more experiments can be implemented to investigate the long-term prediction of different models such as the quarterly performance and profit return of stocks and the effect of adding technical indicators in the data of future.

REFERENCES

[1] F., Rundo, F. Trenta, A.L. Stallo, & S. Battiato, Machine Learning for Quantitative Finance Applications: A Survey, Applied Sciences, vol. 9, no. 24, 2019, pp. 5574, DOI: https://doi.org/10.3390/app9245574

[2] D. Bhuriya, G. Kaushal, A. Sharma, & U. Singh, Stock market predication using a linear regression, In 2017 international conference of electronics, communication and aerospace technology (ICECA), IEEE, vol. 2, 2017, pp. 510-513, DOI: 10.1109/ICECA.2017.8212716

[3] H. L. Siew and M. J. Nordin, Regression techniques for the prediction of stock price trend, 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), 2012, pp. 1-5, DOI: 10.1109/ICSSBE.2012.6396535

[4] J. Ayala, M. García-Torres, J. L. V. Noguera, F. Gómez-Vela, & F. Divina, Technical analysis strategy optimization using a machine learning approach in stock market indices, Knowledge-Based Systems, vol. 225, 2021, pp. 107119, DOI: https://doi.org/10.1016/j.knosys.2021.107119

[5] B. S. Bini, & T. Mathew, Clustering and regression techniques for stock prediction, Procedia Technology, vol. 24, 2016, pp.1248-1255, DOI: https://doi.org/10.1016/j.protcy.2016.05.104

[6] Y. Huang, L. F. Capretz & D. Ho, Machine learning for stock prediction based on fundamental analysis, IEEE Symposium Series on Computational

Intelligence (SSCI), IEEE, 2021, pp. 01-10, DOI: 10.1109/SSCI50451.2021.9660134

[7] I. Kumar, K. Dogra, C. Utreja, & P. Yadav, A comparative study of supervised machine learning algorithms for stock market trend prediction, In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), IEEE, 2018, pp. 1003-1007, DOI: 10.1109/ICICCT.2018.8473214

[8] J. Patel, S. Shah, P. Thakkar, & K. Kotecha, Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques, Expert systems with applications, vol. 42, no. 1, 2015, pp. 259-268, DOI: https://doi.org/10.1016/j.eswa.2014.07.040

[9] L. Khaidem, S. Saha, & S. R. Dey, Predicting the direction of stock market prices using random forest, 2016, DOI: https://doi.org/10.48550/arXiv.1605.00003

[10] M. Kumar, & M. Thenmozhi, Forecasting stock index movement: A comparison of support vector machines and random forest, In Indian institute of capital markets 9th capital markets conference paper, 2006, DOI: http://dx.doi.org/10.2139/ssrn.876544

[11] Y. Xia, Y. Liu, & Z. Chen, Support Vector Regression for prediction of stock trend, In 2013 6th international conference on information management, innovation management and industrial engineering, IEEE, vol. 2, 2013, pp. 123-126, DOI: 10.1109/ICIII.2013.6703098

[12] T. Ye, Stock forecasting method based on wavelet analysis and ARIMA-SVR model, In 2017 3rd International Conference on Information Management (ICIM), IEEE, 2017, pp. 102-106, DOI: 10.1109/INFOMAN.2017.7950355

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.