Stock Price Prediction Based on Machine Learning and Deep Learning Methods

Hanlin Wang(✉)

South China University of Technology, Xingye Avenue, Panyu District, 511442 Guangzhou, China
202030010152@mail.scut.edu.cn

Abstract. Stock price prediction is one of the most challenging tasks in time series forecasting. Many methods are put forward to explore the nature of the stock market. However, most of them just focus on one kind of model. This paper mainly contributes in two experts: The first is that the author innovatively found that predicting stock price of different companies need to be applied by different methods after data analyses. The second is that the author applies many popular artificial intelligence methods to predict the stock price and makes a summary of their performances. In this paper, the author firstly attempts to apply plenty of methods like linear regression, SVR, Random Forest, KNN, Decision tree, Bagging, AdaBoost, XgBoost, MLP, RNN, LSTM, GRU to predict the stock price of Intel company, Coca-Cola company and Exxon Mobil Corporation. And the results would be evaluated by the metrics of R^2 and accuracy. After conducting out the experiment, it is found that Bagging method is the best model for the Intel company and Exxon Mobil Company and RNN is considered as the best method to predict the stock price of Coca-Cola Company. Due to the fact that these three companies are good representatives for technology, food and drink and energy fields, the approaches corresponding to these three companies can also be transferred to the same kind of other company. And the results prove that the selected methods are effective.

Keywords: Stock market · artificial intelligence · machine learning · deep learning

1 Introduction

Time series prediction is a very popular topic in many fields, such as EEG signal processing, weather forecasting, petroleum production prediction, heating load forecasting and so on. And one of the most fascinating applications of time series problem is in finance market. One of the most challenging issues in the world of time series is stock price prediction. The first reason for it is because the stock price change rapidly and can be affected by many different factors. The second reason is that the model not only focus on the stock price itself, but also the association between daily prices.

In real world, people always tend to predict stock prices of different companies using corresponding strategy intuitively because factors affecting different stock prices

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 1087–1098, 2023.
https://doi.org/10.2991/978-94-6463-198-2_112

are different. In the past, many papers pay more attention to only one method for all the stocks. The author assumes it is unrealistic and unreasonable. Therefore, the author would apply different methods to different stocks. Moreover, many artificial intelligence methods are fancy and complicated in research work. A good summary of different methods for predicting stock prices would also be made.

The current stock prediction techniques may be divided into three categories: fundamental analysis, technical analysis, and time series forecasting [1]. A type of investment study known as "fundamental analysis" looks at a variety of corporate measures, including earnings, sales, financial report, shareholder composition, and other economic variables. As for the basic technical analysis, some simple calculation based on history prices are used to predict the future price. Another category is time series prediction methods. These kinds of method can be summarized as linear based model and non-linear model. Linear based models include Autoregressive-Moving Average Mixed Model (ARMA), Integrating Moving Average Autoregressive Models (ARIMA) and their variations [2]. These kinds of models use some mathematical and pre-defined equations to fit the data but it is only suitable on the stable and linear data. Non-linear models include algorithms like ARCH, GARCH, machine learning algorithm and deep learning algorithm [3]. These methods are variance models describing changes or volatility over time but they may appear oscillation phenomenon due to the any increase in the hidden lag term. Nowadays, more and more methods are based on artificial intelligence method like machine learning and deep learning instead of traditional statistical modelling. The principals of different methods are quite different and they are preferred to different data scenario. Linear regression model is preferred for linear data because of its high confidence value and its model with less parameters can be trained directly and rapidly. SVR is one of the popular data mining techniques used in the machine learning industry and it is suitable on low-resource scenario [4]. Also, SVR can be a good choice to deal with non-linear data. However, it is hard to implement in rich data scenario because calculating a larger order of matrix would consume machine memory and time. Random forest method has a strong ability to resist the overfitting of the training data while it can't output a continuous number in the test phase [5]. KNN is not sensitive to outliers while its complexity of time and calculation is very large [6]. Decision tree is a good method to explain the training process and results. Bagging method resamples the dataset and can effectively decrease the variance and Boosting method minimize the loss function sequentially which can decrease the bias to raise the accuracy. Then as for the deep learning method, the biggest characteristic of these models is that model has plenty of neurons and need a great amount of parameters tuned. Deep learning algorithms have the ability to identify hidden stock price changing regularity and dynamics in the stock price data by training. However, it is hard to explain the principal of its work. Multiple layers perception (MLP) [7] is biomimetic neural network and work well in many assignments including time series forecasting. And the most classic models for solving time series prediction are Recurrent neural network (RNN), long short-term memory (LSTM) and Gated Recurrent Unit (GRU). RNN is built for sequential input. LSTM, an upgraded version of RNN, addresses the issue of gradient vanishing and the loss of long-term information that RNN includes [1]. GRU is a kind of variant of LSTM and achieves the similar accuracy in many tasks compared with LSTM but it contains less parameters [8].

In this paper, this author aims to explore the best method for predicting stock price of different companies. And artificial intelligence methods including linear regression, SVR, Random Forest, KNN, Decision tree, Bagging, AdaBoost, XgBoost, MLP, RNN, LSTM, GRU are applied on the data. Intel company, Coca-Cola company and Exxon Mobil Corporation are targeted objects because technology, food and drink and energy field are always a hot investment topic in the public. And these three influential companies are considered as the representative of the food, energy and technology fields. Then the mentioned methods would be applied for these three companies and the author compares their performance with lots of metrics to find the best model for each of these three companies. Moreover, the author assumes that stock price of companies in the same field change in a similar pattern. Therefore, another three companies in the corresponding field would also be evaluated to assess the efficacy of the raised methods. All the mentioned work can provide positive and effective suggestions to all investors.

2 Method

2.1 Data Sources

S&P500 which is from Kaggle (https://www.kaggle.com/camnugent/sandp500) is a well-structured dataset on a wide range of companies. It provides historical price from 2013 to 2018 for all companies found on the S&P500 index in 2018. The dataset has various characteristics of stocks, including the open price, high price, low price, and volume.

In the data selection part, Intel Company (INTC), Coca-Cola Company (KO) and Exxon Mobil Corporation (XOM) are selected as the representatives of corresponding fields. The models are trained and tested using the data of these three companies. In the testing phase, another three companies (Apple company, Conagra Brands Corporation and Chevron corporation) which are randomly chosen are used to verify the guess.

In the data preprocessing part, this paper chooses close price as the standpost of everyday price. The close prices of observed days are partly shown in Table 1 and the change of prices of these three companies can be observed from Fig. 1. The work is based on a sliding window method of forecasting the stock price of near future. The window size is fixed to be 10 and prediction is made to the next trading day [1].

Table 1. Close Price of observed days

	Intel Company	Coca-Cola Company	Exxon Mobil Corporation
2013–02-08	21.00	38.77	88.61
2013–02-11	21.03	38.61	88.28
…	…	…	…
2018–02-06	44.91	44.67	78.35
2018–02-07	45.20	44.56	76.94

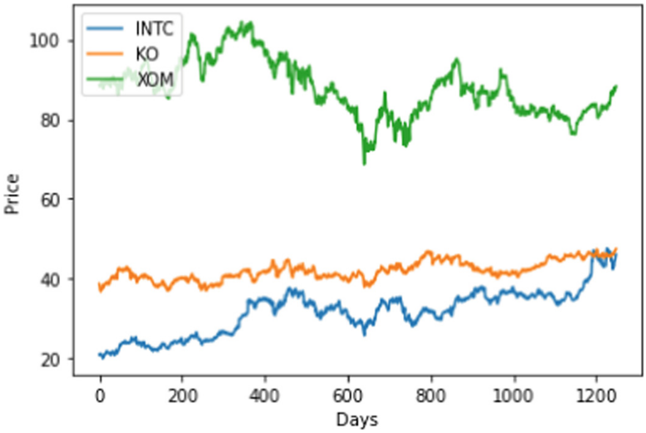

Fig. 1. Stock price of these three companies

2.2 Variable Illustration

Before diving into the models, statistics analysis on the datasets should be carried out first. In the boxplot of Fig. 2, it is observed that these three datasets have different data characteristics. And all data is relatively stable because there isn't any outliers in the boxplot. It indicates that it is meaningful to model and predict.

From Fig. 3, it is observed that the price didn't perfectly follow normal distribution. Because stock prices are impacted by changes in people's psychological states, the evolution of businesses, the economy, and policy trends, the distribution of data can be understood.

From the Fig. 4, it can be observed that autocorrelation of all these companies decreases slowly, which means that they are not stable time series. Therefore, it is necessary to transfer the unstable time series into stable time series before applying the linear regression method.

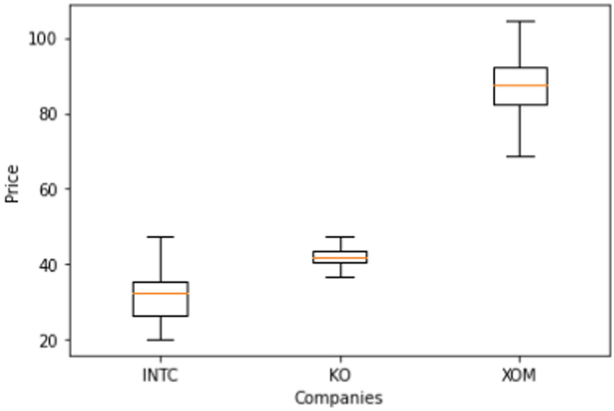

Fig. 2. Box plot of these three companies

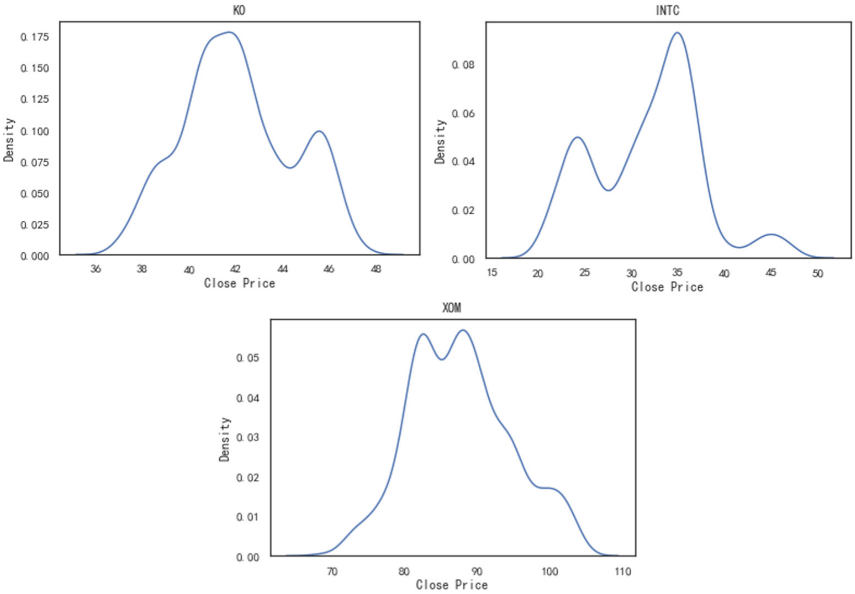

Fig. 3. Probability density distribution plot

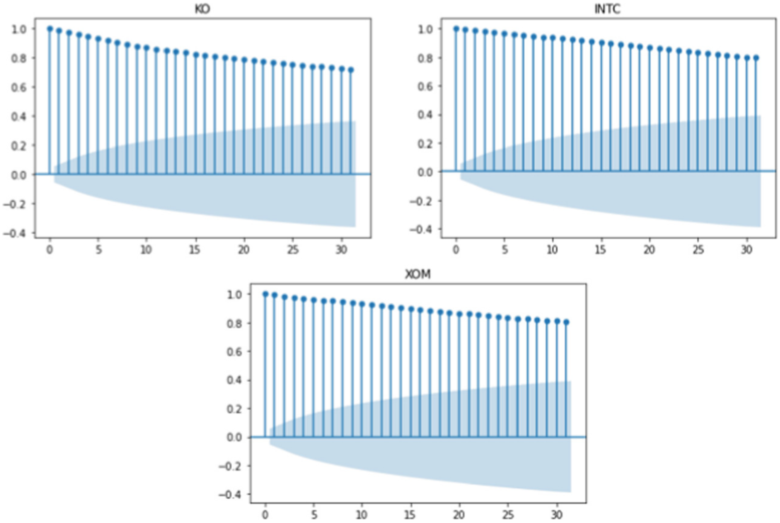

Fig. 4. Autocorrelation of Coca-Cola Company (KO), Intel Company (INTC) and Exxon Mobil Corporation (XOM)

This paper use difference calculation to make time series stationary. Then, ADF (Augmented Dickey–Fuller method) is used to test whether the series is stationary [9]. The outcomes are displayed in Table 2. P-value indicates that this series does not have a time-dependent structure and they are stationary.

Table 2. Results of ADF test

	Intel Company	Coca-Cola Company	Exxon Mobil Corporation
The ADF Statistic	-8.393	-26.658	-19.524
P-value	0.000	0.000	0.000

2.3 Model

The author expects to find the best model for the dataset, so candidate methods are a great number. It includes linear regression, SVR, Random Forest, KNN, Decision tree, Bagging, AdaBoost, MLP, RNN, LSTM, GRU. Details of some important models would be shown in this part.

Linear Regression. Regression is used for predicting an outcome based on a given input. Considering multi-day historical prices, there are more than one descriptive variable and the multiple linear regression is used. The multiple linear regression is used to predict the future price of variable ($\tilde{Y}$) with respect to other variables (X_i) using Eq. 1 [10].

$$\tilde{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1}$$

where $\beta_0, \beta_1, \beta_2 \ldots \beta_n$ are co-efficient that can be calculated using Eq. 2.

$$\beta = \left(X^T X\right)^{-1} X^T Y \tag{2}$$

SVR. SVR creates an interval band with a distance of ϵ (tolerance deviation) on either side of the linear function. The sample that falls into the interval does not calculate the loss, that is, its function model will only be impacted by the support vector. Finally, the model would be optimized by reducing the overall loss and increasing the interval by the method of Lagrange function [4].

KNN. KNN is a form of lazy learning that produces the k records of the training data set that are most similar to the test data set rather than creating a model or function. The class label is then chosen and assigned as a prediction price of the query record using a majority vote among the chosen k nearest neighbors [6].

Decision Tree & Random Forest. Decision tree would build the root, node and leaf according to the training data. And the output of regression tree is a continuous value or real number value. As for Random Forest, there are several decision trees that have been trained on various feature space subspaces. The training data would be recursively split into partitions and none of trees see the entire training data. At a particular node, the split is accomplished by asking a query about an attribute and the splitting criterion is based on some impurity measures such as Shannon Entropy or Gini impurity [5].

Ensemble Method. As for bagging method, the dataset would firstly divided into k equal parts and fed to the k models separately. Then the prediction value would be the average of results of k models. As for the Boosting model, base models are combined linearly through additive models. The weight and probability distribution of the training data for each model would be tuned according to the error rate in each training round.

Multiple layer perception (MLP). In MLP structure, the neurons are called layers. The first layer is input and the last layer is the output. The remaining layers are called hidden layers. In a neural network, each neuron processes inputs from other neurons and then its outputs are passed to the neuron in next layer. Here, the stock price in previous days are passed to the input layer and the output of the neuron in the output layer is assumed as the prediction value [7].

Recurrent neural network (RNN). In RNN architecture, it is a deep learning model for Modeling Sequence Data. It introduces hidden state (h) which can extract the features of sequence data. By some calculations, it can be transferred to the final output y. Hidden state h can be renewed by Eq. 3 in each time step where f is a function, θ is parameters and x is input data [1].

$$h^t = f\left(h^{t-1}, x^t; \theta\right) \tag{3}$$

Long Short-term Memory (LSTM). With LSTM architecture, LSTM cells are used in place of the customary hidden layers. LSTM cells consists of input gate, cell state, forget gate, and output gate. The input gate is used to regulate the amount of data sent to the model. Cell gate has the capacity to add or remove information while operating across the entire network. Forget gate is to decide the fraction of the information should be lost in a long term. Output gate is to decide the output generated by LSTM. The outputs from other gates are used to update each cell gate. Mathematically, it can be described by the following equations.

$$f_t = \sigma\left(W_f.\left[h_{t-1}, x_t\right] + b_f\right) \tag{4}$$

$$i_t = \sigma\left(W_i.\left[h_{t-1}, x_t\right] + b_i\right) \tag{5}$$

$$c_t = \tanh\left(W_c.\left[h_{t-1}, x_t\right] + b_c\right) \tag{6}$$

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{7}$$

$$h_t = o_t * \tanh(c_t) \tag{8}$$

where x_t is the input vector, h_t is the output vector, c_t is the cell state vector, f_t is the forget gate vector, i_t is the input gate vector, o_t is the output gate vector, σ is activation function and W, b are the parameter matrix and vector [1].

Gated Recurrent Unit (GRU). Compared with LSTM, GRU is simplified and only update gate and reset gate are introduced.

Update gate:

$$z_t = \sigma\left(W_z * \left[h_{t-1}, x_t\right]\right) \tag{9}$$

Reset gate:

$$r_t = \sigma\left(W_r * \left[h_{t-1}, x_t\right]\right) \tag{10}$$

where x_t is the input vector, h_t is the output vector, z_t is the update state vector, r_t is the reset gate vector, σ is activation function and W is the parameter matrix.

After resetting the gate and updating the gate, the candidate status value of GRU unit is $\overset{\sim}{h_t}$ and the final output status value is h_t:

$$\overset{\sim}{h_t}= tanh\left(W_{\overset{\sim}{h_t}} * \left[r_t * h_{t-1}, x_t\right]\right) \tag{11}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_{\overset{\sim}{h_t}} \tag{12}$$

3 Results and Discussion

3.1 Self-evaluation

The above model would be firstly evaluated in the original test datasets. It would predict the following one day's price and its trend according to the previous days by sliding window method. And there are many metrics for regression like MSE, MAE, RMSE and for classification like f1 score, recall and precision. This paper chooses R^2 and accuracy as regression and trend classification metrics. The closer R^2 is to 1, the better the model fits. For trend classification, there are two labels which indicate price increasing and decreasing. The larger accuracy is, the better model works. After testing on the original datasets, Table 3 can be obtained.

According to the above analysis, it is known that R^2 ranges from -1 to 1. When R^2 is negative, it means that the trend of predicted value is opposite to the real value. If it is nearly to 0, it indicates that there is no relationship between predicted value and real value. When R^2 is nearly to 1, it indicates that model performs very well. Another metric is accuracy. Accuracy ranges from 0 to 1. If accuracy is more nearly to 1, the model is better. For regression metric, it can be observed that most of models perform well in three companies except for linear regression. Most of them are about 0.9 while linear regression seems to perform bad because R^2 are always negative in three companies. Also, there is a notable phenomenon that deep learning method including MLP, LSTM, RNN and GRU all achieve a high value in these companies compared to machine learning methods. In other words, deep learning method have a better fitting ability. For trend classification metric, most of them are around 50%. The smallest is just over 0.43 while the largest is over 0.7.

In our experiment setting, there are two labels for next day's trend, which is either up or down. There is 50% possibility to get right answer for random guess. However, it could be seen that some of models get accuracy lower than 0.5, which means that it performs bad in this setting and it is not suitable for predicting this company even the regression loss is low. The reason for it is that the above predicting methods are targeted for trading strategy. It means that telling people whether they should buy in or not is the most important. In other words, the trend classification makes a deep influence on people's decisions. And the regression metrics only judge the total situation of the stock and its total loss. It only decides how much money should invest into the stock. For example,

Table 3. Metrics of different models for different companies

Method	Intel		Coca-Cola		Exxon Mobil Corporation	
	R2	Accuracy	R2	Accuracy	R2	Accuracy
Linear Regression	-0.007	0.745	-0.025	0.762	-0.030	0.752
SVR	0.944	0.506	0.977	0.455	0.944	0.506
Random Forest	0.928	0.506	0.950	0.432	0.927	0.513
KNN	0.886	0.526	0.940	0.519	0.886	0.525
Decision Tree	0.911	0.538	0.929	0.506	0.911	0.538
Bagging	0.937	0.542	0.973	0.448	0.937	0.542
AdaBoost	0.943	0.509	0.976	0.458	0.943	0.509
XgBoost	0.908	0.522	0.903	0.441	0.908	0.522
MLP	0.944	0.466	0.977	0.477	0.944	0.466
LSTM	0.945	0.484	0.975	0.539	0.941	0.461
RNN	0.943	0.466	0.976	0.545	0.944	0.474
GRU	0.939	0.477	0.975	0.535	0.939	0.435

if the predicted price or return is very high in a few days later, it certainly deserves a great investment into it. However, classification metrics decides whether people can earn money in this stock. Therefore, Checking the classification accuracy should be the first step to judge the models' performances followed by regression metrics. If a model has a bad performance in trend classification even if its regression loss is very little, it wouldn't be selected to be the best model for this company, for this kind of field.

According to Table 2, most of deep learning method sometimes achieve a classification accuracy lower than 0.5 which even performs worse than random guess. So even if its regression metric (R^2) is high, they wouldn't be selected all the time. In conclusion, the criterion for judging the best model for the company in this field should consider both R^2 and accuracy. And in the condition that accuracy must be over 50% and be as large as possible, the regression metrics (R^2) is more nearly to 1 as possible.

For Intel Company, Linear regression method achieves the highest score (0.745) in terms of accuracy while R^2 is negative and nearly to zero. And as for LSTM, its accuracy is lower than 0.5. Therefore, these two methods are cancelled out. As for the bagging method, its accuracy ranked the second in terms of accuracy and its R^2 is also close to the highest value among all methods. Therefore, bagging method is chosen as the best method for Intel company, which represents for the technology companies in science and technology field.

For Coca-Cola Company, Linear regression method performs similarly as its performance in predicting price of Intel company, so it wouldn't be the candidate of best method. SVR and MLP are also canceled out even if they perform the best in fitting assignment, because their accuracy is lower than 50%. RNN method is considered as the best method for Coca-Cola Company which is the representative of food and drink field. Accuracy is around 0.55 and R^2 is over 0.97 which are all high values compared with other methods.

For Exxon Mobil Corporation, it can be seen that deep learning models perform strongly on regression tasks, but they are weakly on trend classification tasks. After comparing all methods, SVR and Bagging method are selected as the candidates. Due to the fact that SVR is just 0.003 higher than bagging method in terms of regression and bagging method outperforms SVR around 0.04 in terms of accuracy, Bagging is assumed as the best method for this company which represents for companies in energy field.

3.2 Other Evaluation

According to 3.1, it can be known that Bagging method could achieve the best performance in technology companies and energy company. At the same time, RNN method is mostly suitable for food& drink company. In order to verify the performance of our assume, these models are applied to companies in the corresponding field.

After searching for similar companies in dataset(S&P500), Apple company, Conagra Brands Corporation, Chevron corporation are selected to evaluate the model. Apple company is a kind of mobile phone manufacturer and they create plenty of high-tech products to the world. According to the analysis, bagging method would be applied to predict the stock price. Then, Conagra Brands Corporation is a large multinational group company. They integrate the research and development of high-quality health care products and nutritional food. Therefore, it would be applied RNN model. And Chevron Corporation is one of the largest energy companies in the world, so Bagging method would be applied on it according to the above analyses.

In Fig. 5, there are two lines in blue and yellow in each picture. Blue line indicates real stock price while yellow line indicates predicted stock price. And the length of forecasting period is about 300 days. It is obvious that these two lines are nearly coincident which verifies our assumption. It means that the selected model or method is quite suitable to transfer to other companies in the same kind of field. It also can provide a good reference for all investors.

4 Conclusion

In conclusion, this paper applies many methods to predict stock price of Intel Company, Coca-Cola Company and Exxon Mobil Corporation. In terms of machine learning methods, it includes linear regression, SVR, KNN, Decision Tree, Random Forest, AdaBoost, XgBoost. As for deep learning methods, it includes MLP, RNN, LSTM and GRU. The author assumes these methods are classic methods in artificial intelligence and they perform very well in many tasks, thus these methods are selected. And this paper applies sliding windows method to forecast a price for the following day according to the previous days. After experiments, it can be observed that different datasets are preferred for different methods which caters to original assumption. According to the results, Bagging method is mostly suitable for Intel company and Exxon Mobil Corporation compared

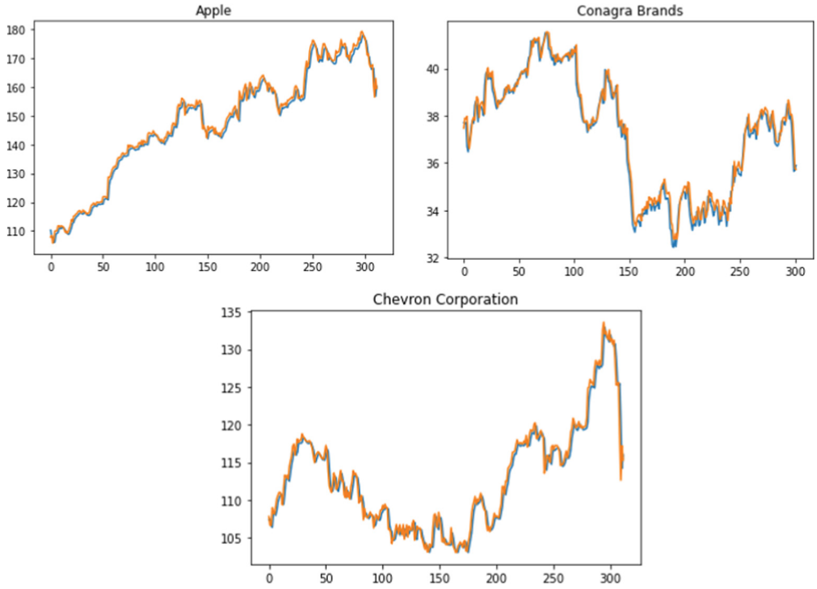

Fig. 5. Prediction on another three companies

with other methods. In other words, when people tend to predict technology and energy companies, bagging method would be a good choice. And RNN would be the best model for Coca-Cola Company which represents for food& drink companies. All models are evaluated both in their own dataset and in other datasets. They all can work well and cater to the guess.

In the future, other factors in the company like profit, shareholder structure would be paid attention to predict the stock price. Also, longer-term stock price would be predicted using more complex models. What's more, models would be applied to more companies and a comprehensive proposal would be proposed.

References

1. S. Selvin, R. Vinayakumar, A. Gopalakrishnan E, et al. Stock price prediction using LSTM, RNN and CNN-sliding window model//2017 international conference on advances in computing, communications and informatics (icacci). IEEE, 2017: 1643–1647.
2. S. M. Huang. Stock Price Analysis and Prediction Based on ARIMA Model -- A Case Study of China Merchants Bank. Small and medium-sized Enterprise Management and Technology, 2022(11):184-187.
3. Y. Wang. Research on Stock Price Prediction Based on ArfMI-Garch-LSTM Hybrid Model. Shanghai normal university, 2022. DOI: 10.27312 /, dc nki. Gshsu. 2022.000752.
4. Y. Zhang. Stock Price Prediction Based on PCA-SVR Model under Random Deletion Mechanism . Yunnan university, 2021. DOI: 10.27456 /, dc nki. Gyndu. 2021.001674.
5. Z. Gao. Share price prediction research based on random forest . China university of political science and law, 2021. The DOI: 10.27656 /, dc nki. Gzgzu. 2021.000061.

6. B. Wang, Y. J. Liu. Research on Stock Price Prediction Based on KPCA-SVM-KNN Algorithm [J]. Computer and Digital Engineering,202,50(04):685–690.
7. R. Achkar, F. Elias-Sleiman, H. Ezzidine, et al. Comparison of BPA-MLP and LSTM-RNN for stocks prediction//2018 6th International Symposium on Computational and Business Intelligence (ISCBI). IEEE, 2018: 48–51.
8. Q. Huang. Research on BiLSTM-GRU Stock Price Prediction Based on Two-stage Feature Extraction. Nanjing information engineering university, 2022. DOI: 10.27248 /, dc nki. GNJQC. 2022.000705.
9. T. T. Fan, Y. T. Kou, C. Liu, H. C. Yan. Time series analysis of the data in the stability judgement study. Journal of modern electronic technology, 2013, 4 (4) : 66–68 + 72. DOI: 10.16652 / j.i SSN. 1004–373 - x. 2013.04.001.
10. T. Yang. Based on the whole function on share prices mixing linear model. Xinjiang university, 2021. The DOI: 10.27429 /, dc nki. Gxjdu. 2021.001330.