Stock Price Prediction Based on Machine Learning

Lixing Liu[1, †] Bingxi Peng [1, †] Jieming Yu[2,* †]

[1] *Faculty of Information Technology, Macau University of Science and Technology, Macau, China, 999078*
[2] *Faculty of Information Technology, Beijing University of Technology, Beijing, China, 100124*
[*]*Email: yujieming@emails.bjut.edu.cn*
[†]*These authors contributed equally.*

ABSTRACT

The stock market is riddled with uncertainty and risks, taking one fallacious decision could lead to huge loss. Therefore, stock market prediction is of great interest to many stock investors. The paper adopts four machine learning models including Decision Tree Regression, Linear Regression, Random Forest Regression Support Vector Regression, respectively, to make prediction on the price of Apple Inc. During the experiment, data in the recent three years were used to train the models in order to make prediction. Moreover, by calculating the mean squared error, the comparison between different models were made. The obtained results showed that the Support Vector Linear Regression model shows a better performance than other models, which is instrumental to the related stock investors in financial markets.

Keywords: *Stock Market, Prediction, Machine Learning.*

1. INTRODUCTION

As the rapid development of economic globalization, financial globalization is becoming an inevitable trend, which results an increasing number in investors who diverted their attention to stock markets. Therefore, stock market prediction is of great importance to investment and financial management [1]. However, the stock market is featured a complicated and immense system with enormous factors and fluctuations. The stock price movement is volatile and challenging to extract valid information, which made it difficult to be forecasted. Machine Learning has the potential to revolutionize stock price forecasting [2], and numerous research works about constructing machine learning models have been carried out to do next-day prediction of the close price of stocks so that to make proper investment decisions and bring a satisfactory return. With the aid of applying the machine learning algorithms, machine learning is becoming a class of modern tools that suits features extraction and prediction.

There has been abundant research regarding stock price prediction in capital market with machine learning algorithms. According to Yoo et al. [3], neural network can significantly beat some commonly used methods regarding forecast accuracy. Pahwa et al. analyzed the advantages and disadvantages of several popular machine learning algorithms and tools [4], including Linear Regression, Support Vector Machine, Python and so on. Moreover, Usmani et al. developed a new method in predicting the stock market by several Artificial Neural Network based models and SVM [5]. The results of their study showed that the Multi-Layer Perceptron performed well with the magnitude of 77%. Additionally, introducing the Long short-term memory model, Parmar et al. tested its performance by making comparison between Regression based models [6]. The result revealed that LSTM model outperformed the Regression based model by hitting a higher accuracy. In the study conducted by Kim et al., the results showed when independent variables are continuous, Linear Regression beats Decision Tree and Artificial Neural Networks under all conditions [7]. On top of that, by comparing the accuracy, Ghosh and Maiti revealed that Random Forest is a considerably more complex machine learning method than the traditional Decision Tree [8].

Given the above precedents, however, few studies have been done in predicting the stock price of a high-tech corporation. As a result, in this article, we decided to anticipate the stock price of Apple Inc., a world-renowned corporation. Several machine learning algorithms are adopted in this paper to make prediction, i.e., Linear Regression, Decision Tree Regression, Support Vector Regression and Random Forest Regression, respectively. Then, the mean squared errors of each model was calculated in order to assess their performance. At last, we came to a conclusion that linear

Y. Jiang et al. (Eds.): ICEDBC 2022, AEBMR 225, pp. 1277–1282, 2022.
https://doi.org/10.2991/978-94-6463-036-7_189

models had the minimal error, indicating that Support Vector Linear Regression models had the best performance.

The organization of this paper is as follows: section 2 exhibits the data and techniques chosen by this study, section 3 gives the results, and section 4 is the conclusion.

2. Data and Methods

2.1. Data Analysis

We choose to use the stock data of Apple Inc. for our study because it is the one of the world-famous technology companies with the largest market value of 2.9 trillion dollars. We download the stock data from Yahoo Finance (https://hk.finance.yahoo.com/), between March 4th, 2019, and March 2nd, 2022. The original dataset contains 7 columns. There are 756 entries for each column with no missing data. Additionally, as the data type for "Date" is "object" instead of a number, we will change it to "datetime64" on a copy of the original dataset for further analysis. After that we applied data pre-processing to our dataset, which includes checking missing data, and filling the voids with median or arithmetic mean value of their column. The table below shows the details of the dataset.

Table 1. Dataset Information

	Open	High	Low	Close
Mean	102.941	104.154	101.786	103.030
Maximum	182.630	182.940	179.120	182.010
Minimum	42.580	43.268	42.375	43.125
Standard Deviation	40.623	41.055	40.133	40.602

As shown in Table 1, it may be conducted that during our sample period, the price may change severely as the mean value of the closing price is more than 100, while the minimal and maximum values are 43.125 and 182.010, respectively. Due to the fact that we aim to predict the close price, we dropped all the other columns except 'Close'. Then, the date is to be analysed, thereby making predictions.

2.2. Methods

In our study, four models have been implemented to predict the stock price, namely Linear Regression, Decision Tree Regression, Support Vector Regression and Random Forest Regression, respectively. The follows sub-sections show some details of the methods mentioned above.

2.2.1. Linear Regression

According to Linear Regression, it is assumed that the variables X and Y have a roughly linear relationship [9]. This connection may be written mathematically by two unknown parameters, i.e., β_0 and β_1, that represent the intercept and slope in the linear model. Basic linear model is shown below:

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

As for the stock data, we use "date" as label to predict the feature "close". After adopting training data to fit the model above to get the estimation of β_0 and β_1, we can forecast future close prices on the basis of historical dates.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

2.2.2. Decision Tree

In this study, we apply Decision Tree Regressor to predict stock prices because of the following advantages [10]. First of all, it is useful for exploring data. Decision Tree is a comparatively time-less way to certify the most relevant factors. Besides, the method can easily determine the potential inside variables. Furthermore, it can capture non-linear relationships, which is suitable for predicting stock prices.

Decision tree has become one of the most widely used machine learning algorithms in both contests such as Kaggle and the commercial world. A decision tree estimates a value by seeking answers for several questions till a specific prediction is found. The model determines the order of the question as well as their contents. Besides, the questions are all True/False in nature.

Decision Trees can be applied to either classification or regression problem. The result differentiates between regression trees and categorization trees, although the classification trees are similar to regression trees constructed from Root, Node, and Leaf. A regression tree's output is a continuous or real value rather than a class [11].

A Decision Tree is built based on information gain. For a data set D, |D| is its sample size, in other words the quantity of samples. For instance, there are K different categories C_k, (k=1, 2, …, K). $|C_k|$ is the sample size that belongs to C_k, where $\sum_{k=1}^{K}|C_k| = |D|$. Assume that feature A has n distinct values: {a1, a2, …, an}, according to which divides D into n subsets: D_1, D_2, …, D_n. $|D_i|$ is the sample size of D_i, where $\sum_{i=1}^{n}|D_i| = |D|$. The set containing samples in set Di that belongs to category C_k is D_{ik}, which means $D_{ik} = D_i \cap C_k$, $|D_{ik}|$ is the sample size of D_{ik}. The algorithm for information gain is given below:

Calculate the empirical entropy H(D) of data set D:

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \quad (3)$$

Calculate the entropy of D conditioned on A:

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \quad (4)$$

Calculate information gain:

$$g(D, A) = H(D) - H(D|A) \quad (5)$$

2.2.3. Random Forest

Random Forest Regression integrates several regression trees, which refers to criteria or constraints grouped hierarchically and applied sequentially. The algorithm starts from randomly selected sample and makes replacement inside the training set. Each step is fitted by one certain regression tree. Node in each tree check the input variables, which are selected randomly in the dataset [12]. Empirically, random forests have a better forecast accuracy than decision tree. However, the two models seem worse that gradient enhances trees. Similar with other prediction models, the sampled data shows significant impact on forecast performance [13].

For tree trainees, Random Forest employs bootstrap aggregating or bagging. Considering a training set of $X = x_1, \ldots, x_n$ and the responses represented by $Y = y_1, \ldots, y_n$, bagging chooses a sample randomly. For $b = 1, 2, \ldots, B$:

1. Sample n training cases with substitution within *X*, *Y*, and label the sample X_b, Y_b.
2. Train a classification or regression tree f_b which is based on X_b, Y_b.

After training, projections for undisclosed samples could be made via adding the predictions from all of the different regression trees on *x'*:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \quad (6)$$

Furthermore, the standard deviation of the projections through all of the separate regression trees on x' could be calculated to gauge the prediction's uncertainty:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} (f_b(x') - \hat{f})^2}{B-1}} \quad (7)$$

The quantity of samples or trees is specified by B, which is a free parameter. Hundreds to thousands of trees are commonly utilized, based on the scale and kind of the training set. The appropriate number of trees may be determined using cross-validation error, the median prediction error upon every training sample x_i. The training and test error tend to remain stable once a given quantity of trees have already been made to fit.

2.2.4. Support Vector

The SVR is a general-purpose learning system that can solve function estimate issues. Inside the algorithm, Structural Risk is selected as an indicator for minimization [14]. For thetraining examples of $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$, $\mathbf{x}_i \in R^d$ is regarded as training vector for input, while $y_i \in R^1$ is a target output, detailed formations are shown below [15]:

$$\min_{\mathbf{w},b,\xi,\xi^*} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \xi_i + C\sum_{i=1}^{l} \xi_i^* \quad (8)$$

Subject to:

$$y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \le \varepsilon + \xi_i \quad (9)$$

$$\&\mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \le \varepsilon + \xi_i^* \quad (10)$$

$$\xi_i, \xi_i^* \ge 0, i = 1, \ldots, l \quad (11)$$

The dual is:

$$\min_{\alpha,\alpha^*} \quad \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T\mathbf{Q}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \varepsilon\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \quad (12)$$

Subject to:

$$\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0,\ 0 \le \alpha_i, \alpha_i^* \le C,\ i = 1, \ldots, l, \quad (13)$$

where $Q_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j)$ and *K* is the kernel. In the above equations (8)-(13), vectors $\mathbf{x}_i$ are transferred from lower dimension to a relatively higher (possibly boundless) dimensional feature space. The *C* parameter governs the balance between the smoothness of *f* and the tolerance for level's variations bigger than ε. ε is defined in equation (14) and represents the parameter for Vapnik's ε-insensitive loss function:

$$|\mathrm{x}|\varepsilon = \max(0, |\mathrm{x}| - \varepsilon) \quad (14)$$

The function of SVR regression is as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{l} (-\alpha_i + \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b \quad (15)$$

There are a large number of kernel functions (a method for accepting data as input and converting it to the format required for execution) that can be used, which can be very flexible to solve various regression problems.

3. RESULTS

After training and tuning the models, we feed them with test data to get test results. The predictions are visualized with line graphs and the test error for each

model is calculated with MSE. Then, we compare the errors to determine which model has the best performance among the four of them.

3.1. Predictions

The following plots indicate the close prices that are predicted by the four models and the original close price, providing us with a visual perception of the accuracy of each model by observing these plots. In each plot, the red line represents the initial (training) data, the blue one shows the validation (testing) data, and the purple line is the prediction. As we can see, the trend given by Linear Regression and SVR (Linear Kernel), smaller in values, are quite similar to the actual situation while the sudden variations cannot be predicted. Additionally, SVR (Polynomial Kernel), doing a poor job, gives the opposite trend. On top of that, the results given by the Decision Tree Regressor and the Random Forest Regressor are also quite dufferent from the validation set.

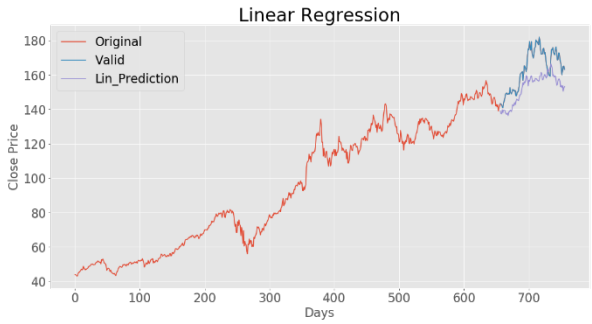

Figure 1 Linear Regression

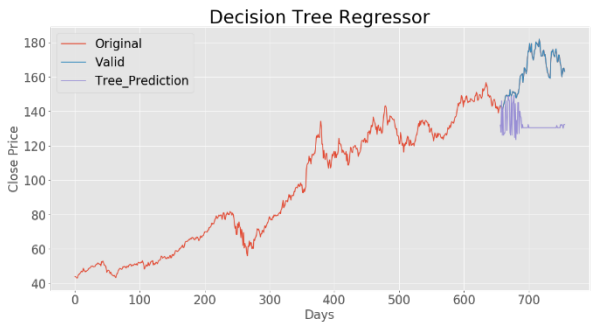

Figure 2 Decision Tree Regressor

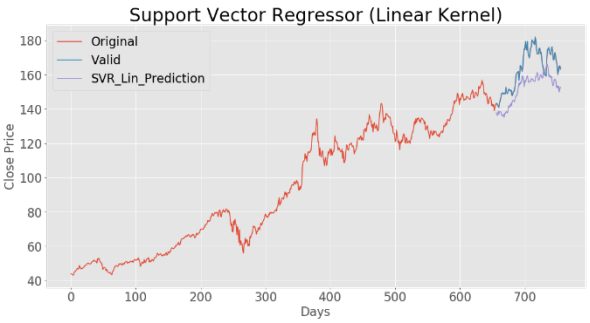

Figure 3 Support Vector Regressor (Linear Kernel)

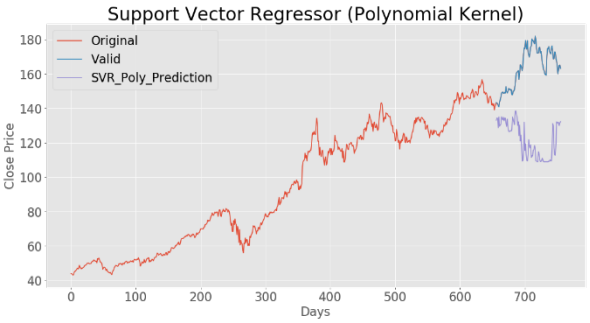

Figure 4 Support Vector Regressor (Polynomial Kernel)

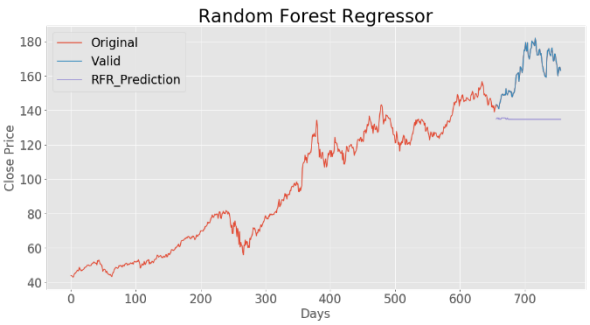

Figure 5 Random Forest Regressor

3.2. Test Error

The above figures give a brief description on the predicted values. However, the comparison seems objective. It is necessary to quantitively get one statistic indicator to compare the above results. In this paper, MSE indicator is adopted. And the results are given in the table below. As we can see, compared with others, the MSE for Support Vector Regressor with linear kernel is the smallest.

Table 2. Test Error

	Linear Regression	Decision Tree Regressor	SVR (Linear)	SVR (Polynomial)	Random Forest Regressor
MSE	126.907	241.585	114.442	722.522	128.008

4. CONCLUSION

Several Machine learning algorithms are adopted in this paper to make prediction, i.e., Decision Tree Regression, Linear Regression, Support Vector Regression and Random Forest Regression, respectively. After calculating the mean squared error of each model, the results revealed that both Linear Regression and SVR with linear kernel had superior performance to the other models, they made rather accurate prediction on Apple price regarding the price and its overall trend. However, the other two models did a less satisfying job, which mis-predicted the stock price. The results remind the related investors to focus more on linear models when predicting stock prices.

In sum, the experiment conducted in this paper was conducive to the related investors in financial markets, whereas there also exist lapses in it. Although the SVR model with linear kernel had the lowest error, it is insufficient to assert that such model best fits the prediction scenario, as the models of Random Forest Regression and SVR with polynomial kernel lack optimization. Despite the listed weakness, the paper is relatively supportive to the stock investors, not to mention those tech lovers. Furthermore, we can utilize more models so as to make comparison between the linear one in the near future.

ACKNOWLEDGEMENT

These authors contributed equally, the authors would like to thank for the reviewers and editors for their helpful suggestions.

REFERENCES

[1] Biswas, M., Nova, A. J., Mahbub, M. K., Chaki, S., Ahmed, S., & Islam, M. A. (2021, August). Stock Market Prediction: A Survey and Evaluation. In *2021 International Conference on Science & Contemporary Technologies (ICSCT)* (pp. 1-6). IEEE.

[2] Sharma, A., Bhuriya, D., & Singh, U. (2017, April). Survey of stock market prediction using machine learning approach. In *2017 International conference of electronics, communication and aerospace technology (ICECA)* (Vol. 2, pp. 506-509). IEEE.

[3] Yoo, P. D., Kim, M. H., & Jan, T. (2005, November). Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)* (Vol. 2, pp. 835-841). IEEE.

[4] Pahwa, N., Khalfay, N., Soni, V., & Vora, D. (2017). Stock prediction using machine learning a review paper. *International Journal of Computer Applications*, *163*(5), (pp. 36-43).

[5] Usmani, M., Adil, S. H., Raza, K., & Ali, S. S. A. (2016, August). Stock market prediction using machine learning techniques. In *2016 3rd international conference on computer and information sciences (ICCOINS)* (pp. 322-327). IEEE.

[6] Parmar, I., Agarwal, N., Saxena, S., Arora, R., Gupta, S., Dhiman, H., & Chouhan, L. (2018, December). Stock market prediction using machine learning. In *2018 first international conference on secure cyber computing and communication (ICSCCC)* (pp. 574-576). IEEE.

[7] Kim, Y. S. (2008). Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Expert Systems with Applications*, *34*(2), (pp. 1227-1234).

[8] Ghosh, A., & Maiti, R. (2021). Soil erosion susceptibility assessment using logistic regression, decision tree and random forest: study on the Mayurakshi river basin of Eastern India. *Environmental Earth Sciences*, *80*(8), (pp. 1-16).

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, (pp. 61). New York: springer.

[10] Drakos, Georgios. (2019, May 23). Decision Tree Regressor Explained in Depth. Retrieved from https://gdcoder.com/decision-tree-regressor-explained-in-depth.

[11] Hindrayani, K. M., Fahrudin, T. M., Aji, R. P., & Safitri, E. M. (2020, December). Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression. In *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 344-347). IEEE.

[12] Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal*, *4*(3), (pp. 212-219).

[13] Piryonesi, S. M., & El-Diraby, T. E. (2020). Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. *Journal of Transportation Engineering, Part B: Pavements*, *146*(2), (pp. 04020022).

[14] Rabe, A., van der Linden, S., & Hostert, P. (2009, August). Simplifying support vector machines for regression analysis of hyperspectral imagery. In *2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing* (pp. 1-4). IEEE.

[15] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), (pp. 1-27).