Research on Predicting Amazon Stock Price Based on Linear Regression and Decision Tree Algorithms

Gengyuan He*

School of Data Science, the Chinese University of Hong Kong, Shenzhen 518172, China

*Corresponding author: 120090886@link.cuhk.edu.cn

Abstract. Stock price prediction is really popular in finance. A good stock price prediction can benefit investors. In this study, the stock price of Amazon was predicted using decision tree and linear regression. Besides this study compared the accuracy and prediction results of two algorithms. The prediction models use the previous stock price and volume as the predictor. This study finds that the Decision Tree algorithm fit a better model than the Linear Regression algorithm. It indicates that the Decision Tree algorithm fits a better model than the Linear Regression algorithm. Moreover, this study forecasts the Amazon's stock price for the next 60 days based on the two models. This research helps investors choose whether and when to buy Amazon stock.

Keywords: Amazon Stock Prediction, Linear Regression, Decision Tree.

1. Introduction

Stock price prediction is really popular in finance. Investors are interest in predicting stock price. Predicting the future development of stocks allows investors to choose the best stocks to invest in at the most favorable time.

Machine learning (ML), as an emerging artificial intelligence (AI) technology, also plays a role in stocks price prediction. The fundamental goal of the prediction is to make investing decisions less uncertain. The forecast of stock prices is extremely challenging, nevertheless, because of the stock market's turbulence and instability. Therefore, a suitable and effective model is helpful of stock price prediction.

This study uses Linear Regression and Decision Tree algorithms in machine learning to fit models to predict Amazon's stock price. This study enriches the theoretical basis for machine learning to predict stock price. The study fills the research gap about using Decision Tree algorithm to predict stock price. In addition, this study also evaluates and forecasts Amazon stock for investors' reference.

For this paper, Section 1 is about the introduction. Section 2 is about the literature review. Section 3 is about the methodology and data. Section 4 is about the results. Section 5 is about the discussion. Section 6 is about the conclusion.

2. Literature References

Patel J. et al. (2015) states that there are two types of stock analysis. The fundamental analysis comes first. When deciding whether or not to invest, investors consider the inherent worth of the stocks, the performance of the market and economy, the political environment, etc. [1]. The appraisal of equities using technical analysis, on the other hand, involves looking at market data such as historical prices and volume. Machine learning is undoubtedly a part of technical analysis.

There are many machine learning techniques have been used to predict stock price. Artificial neural networks were initially used to predict stock price patterns by Zhang G. et al. in 1998 [2]. The support vector machine (SVM), a highly specialized subclass of learning algorithms created by Vapnik (1999), is characterised by the ability to constrain the decision function's size, the use of kernel functions, and the rarity of the solution. SVM was utilized by Kim K. in 2003 to predict stock price. SVM was used in this study to compare case-based reasoning (CBR) and back-propagation neural networks (BPN) [4]. The experimental findings made it clear that SVM outperformed BPN and CBR. Kavitha S. et al. (2016) compared Support Vector Regression with Linear Regression using historical data. As a result, linear regression is best suited by the LeastMedSq function[5].

In order to predict the behavior of the financial markets, Hassan, M. et al. (2007) proposed and put into practice a fusion model that combines the Hidden Markov Model (HMM), Artificial Neural Networks (ANN), and Genetic Algorithms (GA). The daily stock prices were converted using ANN into separate sets of variables that were then supplied to HMM [6]. To forecast the price movement of the Hang Seng index on the Hong Kong stock market, Ou and Wang (2009) used ten famous algorithms. As a result, the SVM and LSSVM were found to be adequate.

Using the classifier ensembles technique, Studies by TSAI, Lin, Yen, and Chen (2011) demonstrate that multiple classifiers outperform single classifiers in terms of forecast accuracy and ROI[8]. The probabilistic neural network was researched by Chen, Leung, and Daouk in 2003[9]. The results demonstrated that PNN method generated higher returns than other types of investment methods. Using an upgraded Legendre neural network and the assumption that investors choose their investment positions by looking at the company's historical data, Liu and Wang (2012) forecasted the price fluctuation. They also included a random time strength function in the forecasting model. The study demonstrates that the suggested random temporal strength function is suitable for boosting prediction accuracy [10].

However, there are little study about Decision Tree for stock price prediction. Thus, this study aims to use Linear Regression and Decision Tree to fit models to predict Amazon stock price and evaluate the models.

3. Methodology and Data

The study aims to predict the Amazon stock price. Linear Regression and Decision Tree are two machine learning methods that are used to predict the Amazon stock price.

3.1 Methodology

3.1.1 Machine Learning and Supervised Learning

Machine learning is a method of training a model from data and then using the model to make predictions. Classified by training methods, machine learning is mainly divided into supervised learning, unsupervised learning and reinforcement learning. Supervised learning is a method of obtaining an optimal model by training sample dataset with known labels. The sample dataset must be split into a training dataset and a test dataset in order to create and evaluate the model. The test dataset is used to evaluate the model once it has been built using the training dataset.

Supervised learning algorithms are mainly divided into two categories: regression and classification. The basic idea of regression is to analyze the interrelationship between the prediction object and related factors, select the appropriate model and then predict according to the model. The basic idea of classification is to group data with known labels into different categories, and then the unlabeled data will be grouped into the corresponding categories. The biggest difference between them is that regression is to deal with continuous data, and classification is to deal with discrete data.

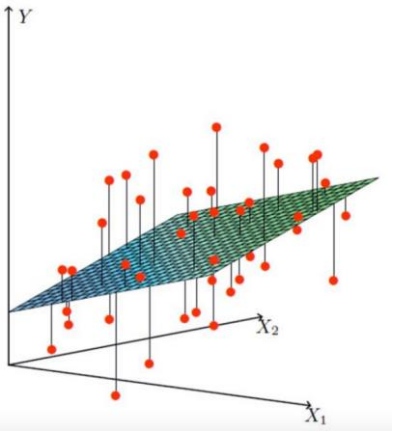
3.1.2 Linear Regression

Linear regression is a traditional and suitable algorithm of stocks price prediction. It belongs to regression algorithms in supervised learning. In mathematical statistics, regression analysis is used to examine the relationship between two or more variables.

Linear regression algorithm can be represented like that. Given n features $X = (X_1, X_2, \dots\dots, X_k)$, where X_i is the value of X on the ith property, the linear model attempts to learn a function that makes predictions from the linear combination of each property, like:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (1)$$

where β_i are coefficients, β_0 is interception, $X_1, X_2, \dots\dots, X_k$ are independent variable, y is dependent variable, $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \dots\dots, \widehat{\beta_k}$ are estimate of the parameter $\beta_0, \beta_1, \beta_2, \dots\dots, \beta_k$, $\hat{y}$ is the estimate of the dependent variable y. Below is the illustration of linear regression.

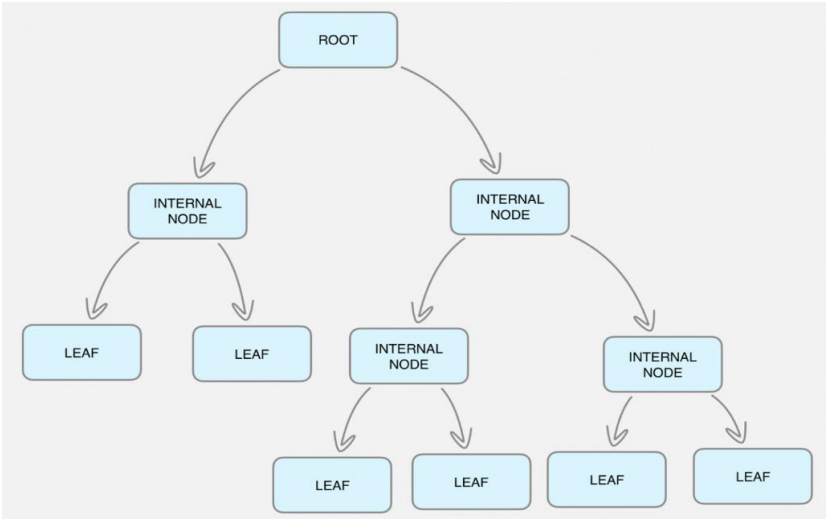


Source: http://www.jsledd.cn/2018/08/05/linearregression/

Figure 1. Linear Regression

3.1.3 Decision Tree

The Decision Tree represents a form a tree. It presents observations of sample data in internal node,obtaining the classification results in leaf node. Below is the illustration of decision tree.



source: https://medium.com/@madanmaram/decision-tree-in-machinelearning-b9c0eed1964b

Figure 2. Decision Tree

The growth of a decision tree is the increase of internal nodes. When a branch does not need to grow, it becomes a leaf. When all the branches are not growing, a complete decision tree is established.

3.2 Data and New Variables

3.2.1 Description of Data

The historical Amazon stock price information is obtained in Kaggle. The sample spans the years 5/15/1997 through 3/24/2022. The sample contains some data on the Amazon stock, including the high, low, open, close, adjacent close prices and volume.

3.2.2 New Variables

For the purpose of predicting stock price, two new variables have been established. These are the new variables:

1. The percentage change in the high and low prices of the stock (HL_PCT). Calculated as: HL_PCT = (High- Low) / Low × 100%
2. The percentage change in the close and open prices of the stock (PCT_change). Calculated as: PCT_change = (Close - Open) /Open × 100%

3.3 Model Evaluation

In this study, coefficient of determination (R^2) is used to evaluate the model.

It is a measure of estimating the goodness of fit of the regression equation. Suppose a dataset includes a total of $y_1, y_2, \ldots\ldots, y_k$observations, and the predicted values given by the fitted model are $\widehat{y_1}, \widehat{y_2}, \ldots\ldots, \widehat{y_k}$. Then the Sum of Squares for total (SST) is:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{2}$$

The Sum of Squares for regression (SSR) is:

$$SS_{reg} = \sum_i (\widehat{y_i} - \bar{y})^2 \tag{3}$$

Where the $\bar{y}$ is the mean of $y_1, \ldots, y_n$, its formular is:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4}$$

Then the R^2 is:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{\sum_i (\widehat{y_i} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \tag{5}$$

For the training dataset, the R^2 range is between [0,1]. The value of the test dataset can also be negative. In general, the higher the R^2, the more the independent variable interprets the dependent variable.

3.4 The Learning Environment

The learning environment of predicting Amazon stock price is shown as Figure 3.

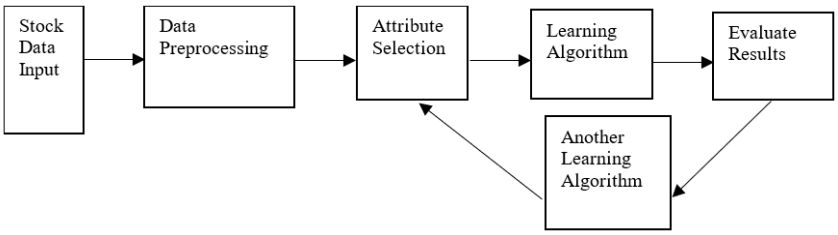


Figure 3. Learning Environment

4. Results

4.1 Steps for Stock Price Prediction

There are 5 steps in the prediction, which is described as follows

Step 1: Download data from the net and import data into the Jupyter Notebook.

Step 2: Data Preprocessing. The general steps are as follow:

(I) Fill the missing data by using the average of value of the previous day and value of the next day.

(II) Change the stock information "Date" type from float to time series data. Set the "Date" as index.

(III) Construct the "Prediction" column. Copy the data in the "Close" column into the "Prediction", and shift up 60 rows.

Step 3: Attribute Selection. Use "HL_PCT", "PCT_change" and "Volume" as independent variables. Use "Prediction" as dependent variables.

Step 4: Learning Algorithm. The general steps are as follow:

(I) Set "HL_PCT", "PCT_change" and "Volume" to X, set "Prediction" to Y. The 60 rows below of X and Y named "X predict" and "Y predict" is used to predict stock price for the next 60 days. The "Y predict" is empty and it is used to receive the prediction for the next 60 days. The rest of X named "X model" is used to train the model. And the rest of Y named "Y model" is used to test the model.

(II) Divide the data of "X model" and "Y model" into train data and test data.

(III) Use the train data and the test data to build and test the model.

Linear Regression algorithm and Decision Tree algorithm will use different methods to fit the models.

(IV) Use the "X predict" and the model to fill the "Y predict" . "Y predict" is the prediction about Amazon stock price of the next 60 days.

Step 5: Evaluate Results. Use R^2 to evaluate and compare different models from Linear Regression and Decision Tree.

4.2 Prediction Results of Linear Regression

Prediction results of Linear Regression is shown as Figure 4 and Figure 5. As shown in the Figure 4, From 1997 to the end of 2020, Amazon's stock price was on an overall upward trend. Starting in 2021, Amazon's stock price will become more volatile. By 2022, Amazon's stock price had fallen, from a high of 3731.40 to a low of 2720.29. The predicted future stock price will continue to fluctuate. As shown in the Figure 5, the range of the prediction of Amazon stock price for the next 60 days is [3482.23, 2854.30]. Amazon stock prices fluctuate greatly. Its amplitude is large.

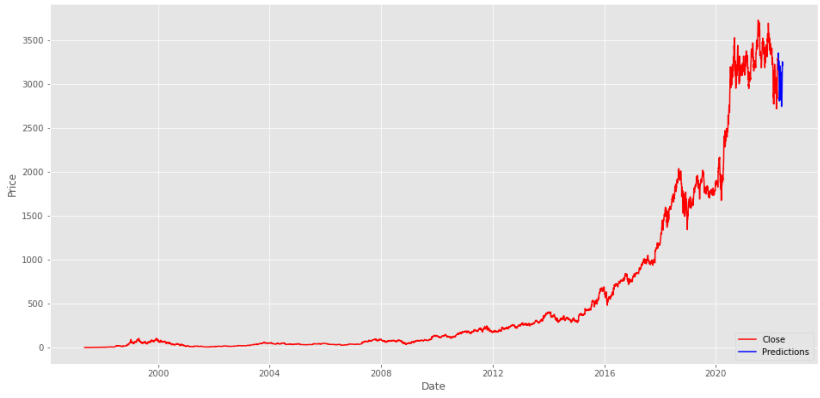


Figure 4. The prediction of Linear Regression

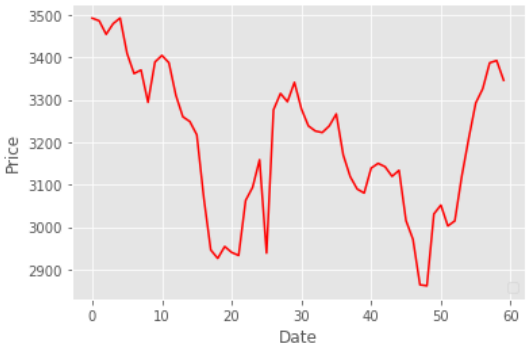


Figure 5. details of the prediction results by Linear Regression

4.3 Prediction Results of Decision Tree

Prediction results of Linear Regression is shown as Figure 6 and Figure 7. As shown in the Figure 6, The predicted future stock price will rebound. As shown in the Figure 7, the range of the prediction of Amazon stock price for the next 60 days is [3596.06, 2954.90].

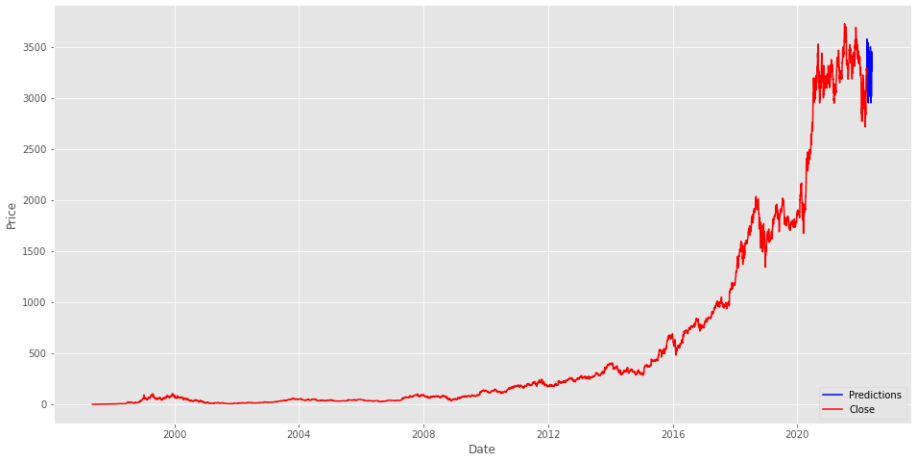


Figure 6. The prediction of Decision Tree

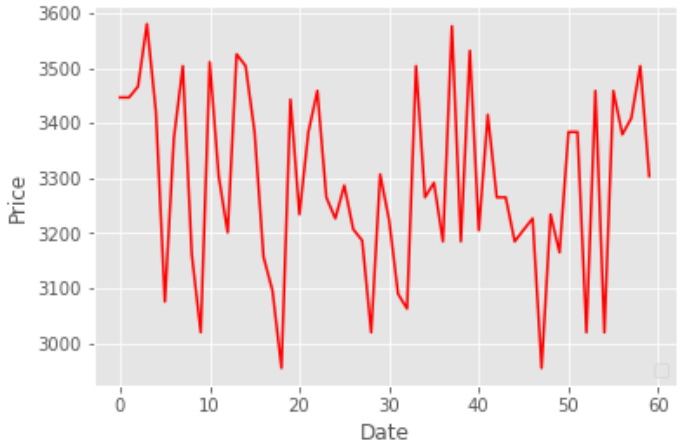


Figure 7. details of the prediction results by decision tree

Both of the prediction of Linear Regression and Decision Tree, Amazon stock prices fluctuate greatly. Its amplitude is large. In contrast, the prediction results of Decision Tree fluctuate more frequently than Linear Regression. The amplitude of the prediction results of the Decision Tree is smaller than Linear Regression.

4.4 Evaluation Results of Two Predictions

For the evaluation about coefficient of determination (R^2). The coefficient of determination of Linear Regression is 0.9807. The coefficient of determination of Decision Tree is 0.9983.

Judging from the evaluation results, for Amazon stocks, the Decision Tree algorithm fits a better model than the Linear Regression algorithm. Therefore, the Decision Tree algorithm gives relatively better predictions than Linear Regression algorithms.

5. Discussion

Linear Regression and Decision Tree, as two well-known machine learning algorithms, have their own advantages and disadvantages. For Linear Regression, it is fast to fit the model. It doesn't need

complicated math. With a lot of data, it can operate quickly. But it can't accurately match nonlinear data.For Decision Tree, it is able to handle irrelevant features. But it is prone to overfitting.

According to the above analysis, the price of Amazon stock fluctuates frequently and greatly. It means that the Amazon stock is an unstable stock. Buying Amazon stock is risky and may also bring benefits. Buyers need to choose carefully whether to buy or not.

6. Conclusion

This study uses Linear Regression and Decision Tree algorithms to predict stock price. There are two conclusions. The first is the Decision Tree algorithm gives relatively better predictions than Linear Regression algorithms. The second is Amazon stock is an unstable stock. It implicates that buying Amazon stock is risky. The limitation of this study is that it only bases on two machine learning algorithms, including linear regression and decision tree. As for the future work, academics can use more machine learning algorithms to predict stock price.

References

[1] Patel, J., Shah, S.R., Thakkar, P., & Kotecha, K. Predicting stock market index using fusion of machine learning techniques. Expert Syst. Appl., 2015, 42, 2162-2172.

[2] Zhang, G.P., Patuwo, B.E., & Hu, M.Y. FORECASTING WITH ARTIFICIAL NEURAL NETWORKS: THE STATE OF THE ART. International Journal of Forecasting, 1998, 14, 35-62.

[3] Vapnik, V. N. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 1999, 10, 988–999.

[4] Kim, K. Financial time series forecasting using support vector machines. Neurocomputing, 2003, 55, 307-319.

[5] Kavitha, S., Varuna, S., & Ramya, R. A comparative analysis on linear regression and support vector regression, 2016.

[6] Hassan, M.R., Nath, B., & Kirley, M. A fusion model of HMM, ANN and GA for stock market forecasting. Expert Syst. Appl., 2007, 33, 171-180.

[7] Ou, P., & Wang, H. Prediction of Stock Market Index Movement by Ten Data Mining Techniques. Mathematical Models and Methods in Applied Sciences, 2009, 3, 28.

[8] Tsai, C., Lin, Y., Yen, D.C., & Chen, Y. Predicting stock returns by classifier ensembles. Appl. Soft Comput, 2011, 11, 2452-2459.

[9] Chen, A., Leung, M.T., & Daouk, H. Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index. S&P Global Market Intelligence Research Paper Series, 2003.

[10] Liu, F., & Wang, J. Fluctuation prediction of stock market index by Legendre neural network with random time strength function. Neurocomputing, 2012, 83, 12-21.