Analyzing Stock Price Prediction Models: A Comparative Study of Linear Regression and Decision Trees During the COVID-19 Pandemic

Caiyuan Yin*

International Economics and Trade, Beijing International Studies University, Beijing, China

* Corresponding Author Email: cy123@berkeley.edu

Abstract. Stock market forecasting plays a vital role in financial decision making. In a volatile and uncertain situation like the COVID-19 pandemic, it is important for us to compare the accuracy of different models for forecasting. This study compares the performance of the Linear Regression (LR) algorithm and the Decision Tree (DT) algorithm in predicting the stock price of Pfizer Inc. before and during a pandemic. The study evaluates the accuracy and stability of the predictions using datasets before and during the pandemic. The results show that the prediction accuracy of LR is better than that of DT in both cases, although both accuracies decrease during the pandemic. The results of the study emphasize the need to consider external factors in the selection of forecasting models and suggest some lessons for future research, leading to the selection of more stable and accurate models under turbulent market conditions. It is worth noting that overfitting occurs when a model learns to capture noise in the data rather than the underlying patterns, leading to reduced generalization performance on unseen data. Future research should address this by exploring techniques such as regularization or cross-validation to mitigate overfitting and enhance the robustness of predictive models in dynamic market environments.

Keywords: Linear Regression, Decision Tree, COVID-19 pandemic.

1. Introduction

Stock markets, being an important part of the global economy, play a vital role in the financial, economic, and decision-making fields. The study of stocks can provide information on market demand and industry trends, helping decision-makers formulate correct corporate strategies and improve the market competitiveness and development potency of enterprises. With the continuous evolution of financial markets and the increasingly strong connection of the global economy, the need to accurately predict stock prices has become very urgent. With the help of new technologies such as data mining and machine learning, it is feasible to learn and analyze large amounts of complex data, and patterns and laws can be automatically learned from the data. This enables automated decision-making and task execution, which greatly improves efficiency without any human error. Therefore, it has become important to utilize advanced technologies such as machine learning algorithms to improve prediction accuracy.

Machine learning models have been widely utilized since their superior feature extraction and prediction ability in many tasks [1-4]. For instance, Liu et al. advanced machine learning techniques for accurate distance measurements with ultra-wideband sensors on engineering structures [1]. Qiu et al. designed a novel machine learning technique called image expression-driven modelling strategy in coke quality prediction [2]. Luo et al. propose a novel utterance-based deep neural network model termed AFF-ACRNN in audio sentiment analysis [3]. All these works mentioned demonstrated the effectiveness of machine learning models and their potential in the finance domain. In addition, the application of machine learning models in financial domain also receives much attention. Qiu and Wang employed sophisticated clustering algorithms for a methodical analysis of credit card user behavior. Their study provides crucial understandings related to economic resilience and underscores the importance of machine learning in financial analysis and risk control [5].

In terms of stock analysis, the unpredictability and complexity of stock market dynamics pose a significant challenge to the analysis. Traditional financial models often fail to fully account for the

complex interactions among the many factors that affect stock prices, which range from economic developments and political policies to human factors. Therefore, many scholars have studied machine learning algorithms for stock prediction, and they have compared the prediction accuracy of different algorithms through a large number of studies [6-11]. Many previous studies on stock price prediction have focused on the application of various machine learning algorithms such as Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and Gate Recurrent Unit (GRU) neural network models, as well as comparative analyses of different algorithms such as Support Vector Machine (SVM) and Random Forest [6-11]. These studies have made significant contributions to the field by demonstrating the potency of machine learning in improving prediction accuracy. However, there are some gaps in the comparative analysis between linear regression (LR) and decision tree (DT) algorithms in stock price prediction.

During the COVID-19 pandemic, some studies also built more sophisticated models to improve the accuracy of stock price prediction in this particular case. For instance, Ardakani et al. utilized big data analytics and LSTM modeling to identify stock price patterns [11]. Sharaf et al. proposed a hybrid prediction system that combines sentiment analysis of COVID-19 news with historical data to predict stock movements [12]. Ronaghi et al. proposed a deep information fusion framework that combines social media trends with market data to mitigate the impact of a pandemic [13]. Faiz et al.investigated machine learning-based price prediction considering changes in consumer behavior during a pandemic, emphasizing the importance of adjusting models according to the importance of adapting models to changing market dynamics [14]. These studies demonstrate various ways to enhance stock price forecasting under unprecedented market conditions. However, there is little research analyzing whether the comparative results of prediction accuracy between Logistic Regression (LR) and Decision Trees (DT) remain consistent across different time periods.

This paper would like to use the stock price data of Pfizer and divide it into two groups, pre-epidemic and epidemic, and model the two groups of data with linear regression and decision tree algorithms for stock price prediction, respectively. Ultimately, the accuracy of these two methods were compared in the two cases to see if the results are consistent and investigate the difference between them.

2. Method

2.1. Data Preparation

2.1.1. Data Source

The datasets used in this study are categorized into pre-epidemic and epidemic: pre-epidemic data from and epidemic data from [15, 16]. Each of the two datasets contains historical stock price data for Pfizer (PFE) and is publicly available for research purposes.

2.1.2. Data Prepossessing

In this study, missing values were removed from both sets of raw data and company-specific (PFE) data were selected. The dates were then converted to datetime format and set as the index of the data. The data were sorted in ascending order by date and the data within the range of specified dates were selected from the data. The two datasets contain 756 rows of data each. The time of the data is from February 8, 2015 to February 7, 2018 and from July 12, 2019 to July 12, 2022 respectively.

Four features 'open', 'high', 'low' and 'volume' from the dataset are selected as input features for the model. By shifting the stock closing price ('close') up num rows, the result is stored in a new column named 'label'. Next, the newly created 'label' column is removed from the processed dataset to generate a new dataset Data. Then, feature columns 'open', 'high', 'low' and 'volume' are selected from the dataset for feature extraction. The study first normalizes the data. This process scales the feature data to ensure that they have similar scales. In this paper, the values of each feature are converted to a distribution with its mean being 0 and standard deviation being 1. The dataset is divided into training

and test sets. 80% of the data is assigned to the training set while the remaining 20% is used as the test set.

2.2. Machine Learning Models

2.2.1. Linear Regression

Linear regression, as a classical machine learning algorithm, has a wide range of applications in solving regression problems. Its main function consists of predicting continuous output variables, which means that it is able to estimate or predict the value of a continuous target variable from the characteristics of the inputs. By fitting the data, linear regression is able to learn from it the relationship between the data and enable prediction of this relationship by building a mathematical model. In addition, linear regression has good interpretability, which means that the relationship between data features and outputs can be understood through the parameters of the model, leading to a better understanding of the meaning and structure of the data. The idea of linear regression lies in finding a best-fit straight line that minimizes the sum of the distances between that line and the data points to achieve an optimal model.

First, a linear regression model object linear_model was created using the Linear Regression function. Then, the training method of the model object was called to pass the training set and the corresponding labels to the model in order for the model to learn the relationship between the features and the labels in the training dataset and fit them.

Next, predictions are made using the model object. First, the test dataset is passed to the predicting method of the model to obtain the linear regression model's prediction of the test set linear_forecast.

Then, the scoring method of the model object is called to evaluate the model's accuracy on the test dataset. The method calculates the model's prediction accuracy on the test data and returns a score, which is between 0 and 1, indicating the model's performance, with scores closer to 1 indicating better predictive power of the model.

Finally, the predictive performance of the model was evaluated using the mean square error (MSE), which is the mean of the squares of the differences between the predicted and true values. The mean square error linear_mse of the linear regression model on the test dataset was calculated through the mean squared error function.

2.2.2. Decision Tree

Decision Tree is based on the known probability of occurrence of various situations, through the composition of the decision tree to find the net present value of the expected value of the probability of greater than or equal to zero, the evaluation of the project risk, to determine the feasibility of the decision-making analysis method is the intuitive use of probabilistic analysis of a graphical method. Because this decision branch is drawn as a graphic very much like the branches of a tree, it is called a decision tree. In machine learning, the decision tree is a predictive model that represents a mapping relationship between object attributes and object values. A decision tree is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category.

First, a decision tree regression model object tree_model was created using the Decision Tree function and the random seed was set to 0 to ensure reproducible results. Then, the fitting method of the model object was called to pass the training set and the corresponding labels to the model in order for the model to learn the relationship between the features and the labels in the training dataset and fit them.

Next, predictions are made using the model object. First, the test dataset is passed to the predicting method of the model to obtain the decision tree model's prediction of the test set tree_forecast.Finally, the predictions are evaluated in the same way.

3. Results and Discussion

The two figures below provide a visualization of the comparison of the difference between the predicted values of the linear regression model and the decision tree model and real prices. Fig. 1 shows the comparison between the predicted and actual stock prices of the two models in the pre-pandemic period. It demonstrates that before the epidemic, the linear regression model predicted a closer stock price, while the decision tree model predicted a result that differed much from the true value and showed a large change in the prediction. Fig. 2 shows the comparison between the predicted and actual stock prices of the two models during the epidemic period. At the time of the epidemic, the linear regression model predicted a much closer stock price, while the decision tree model predicted results that differed significantly from the true value, and the predicted value even occurred at a constant level.

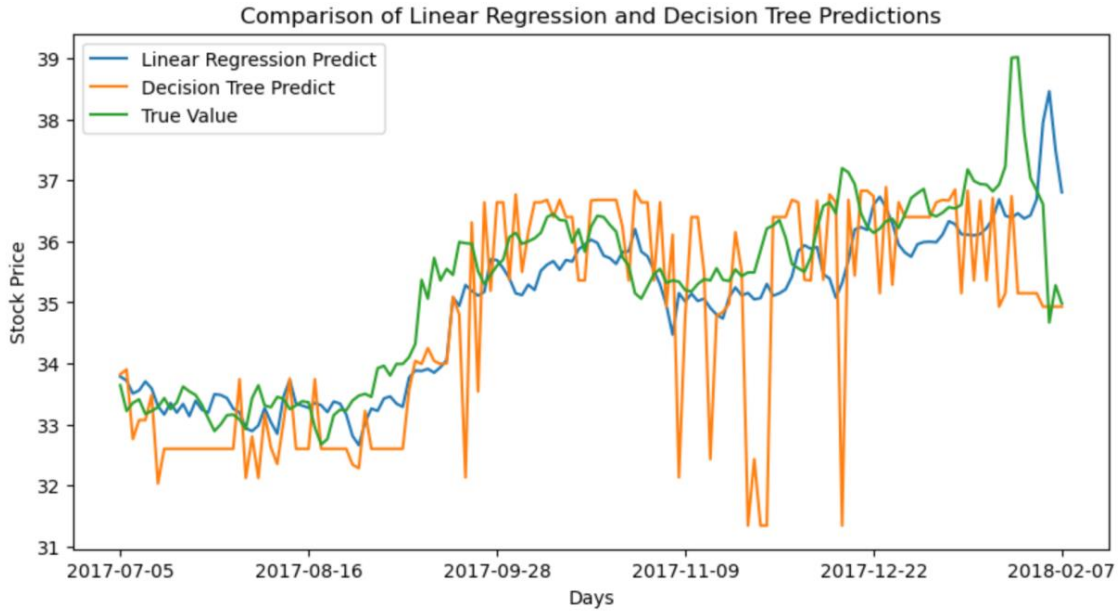

Fig. 1 Comparison of Linear Regression and Decision Tree Prediction (pre-epidemic) (Picture credit : Original).

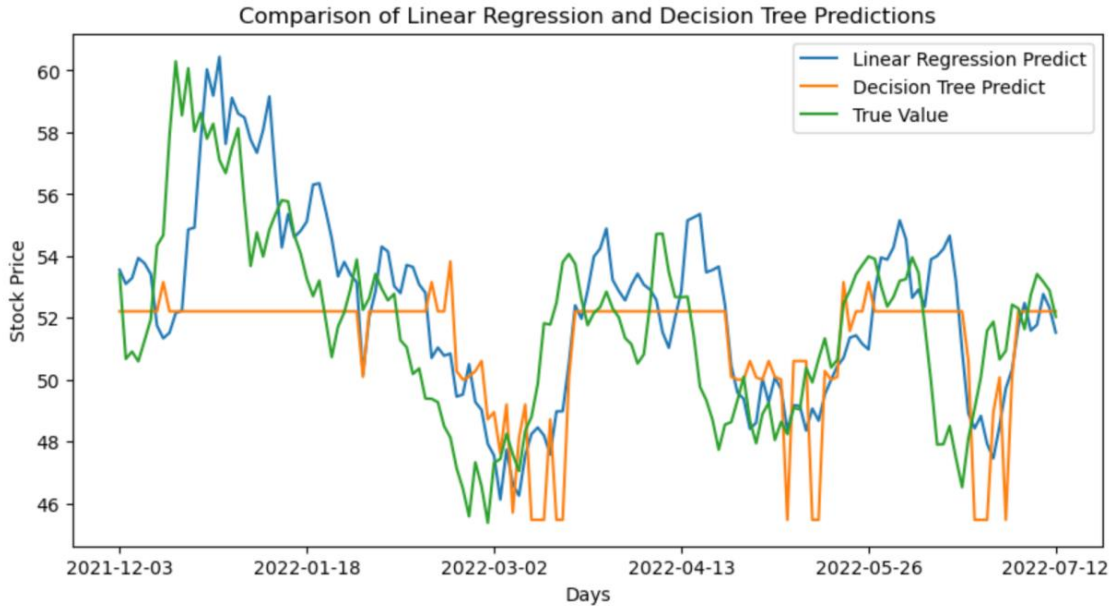

Fig. 2 Comparison of Linear Regression and Decision Tree Prediction (during-epidemic) (Picture credit : Original).

Table 1. Model Evaluation Metrics Data

	LR model score	DT model score	LR model MSE	DT model MSE
pre-epidemic	0.675	0.114	0.653	1.782
during-epidemic	0.199	-0.006	7.198	9.039

The linear regression model had a higher score of about 0.675 shown in Table 1 during the pre-epidemic period, which suggests that during the pre-epidemic period, the linear regression model was more effective in fitting the data compared to the decision tree model. However, during the epidemic period, the scores of both models decreased significantly. The score of the linear regression model dropped to about 0.199, while the score of the decision tree model even became negative. This indicates that the predictive ability of both models was affected to some extent during the epidemic period, especially the decision tree model showed a significant decrease in performance.

Before the epidemic, the mean squared error of the linear regression model was about 0.653 lower than that of the decision tree model, which was about 1.782, which showed that the linear regression model had a smaller prediction error before the epidemic. However, during the epidemic period, the mean square error of the linear regression model increased to 7.198 and the mean square error of the decision tree model increased to 9.039. The prediction errors of both models increased significantly during the epidemic period, and the prediction error of the decision tree model was greater.

Previous studies have extensively explored various machine learning algorithms for stock price prediction, including linear regression, decision trees, and more advanced techniques such as LSTM and GRU neural networks. But few have compared whether the accuracy of these model predictions changes under different time periods.

This study evaluates the performance of linear regression and decision tree models in predicting stock prices over two different time periods. The results show that the predictive accuracy of both models declines during the epidemic period compared to the pre-prevalence period. However, the linear regression model better captures the underlying linear relationships in the data in both cases. The results that linear regression models work better than decision number models generally don't fit with the cognition very well. It is widely recognized that the more complex the model the better the results. This suggests the possibility of overfitting, especially in the case of the decision tree model [17], which performed relatively poorly on the test set. It implies that the model may be too complex to be generalized well to new datasets.

4. Conclusion

The results of this study emphasize the importance of considering external factors (e.g., events such as the COVID-19 pandemic) when evaluating stock price prediction models. Compared to decision trees, linear regression models demonstrated higher stability and predictive accuracy both before and during the pandemic. However, the unprecedented market environment brought about by the pandemic still affected their performance. Future research could explore the integration of other data sources, sentiment analysis, and more sophisticated modeling techniques to enhance the stability of predictive models in volatile market environments. In addition, aggregate approaches or hybrid models that combine the strengths of different algorithms may further improve forecasting accuracy and resilience to market shocks.

References

[1] Liu Y, Bao Y. Real-time remote measurement of distance using ultra-wideband (UWB) sensors. Automation in Construction. 2023 Jun 1;150:104849.

[2] Qiu Y, Hui Y, Zhao P, Cai CH, Dai B, Dou J, Bhattacharya S, Yu J. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy. 2024 Mar 7:130866.

[3] Luo Z, Xu H, Chen F. Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network. InAffCon@ AAAI 2019 Jan 27 (pp. 80-87).

[4] Wang H, Zhou Y, Perez E, Roemer F. Jointly Learning Selection Matrices For Transmitters, Receivers And Fourier Coefficients In Multichannel Imaging. arXiv preprint arXiv:2402.19023. 2024 Feb 29.

[5] Qiu Y, Wang J. A Machine Learning Approach to Credit Card Customer Segmentation for Economic Stability. InProceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA 2023, October 27–29, 2023, Tianjin, China 2024 Jan 19.

[6] Jahan I, Sajal S. Stock price prediction using Recurrent Neural Network (RNN) algorithm on time-series data. 2018 Midwest instruction and computing symposium. Duluth, Minnesota, USA: MSRP, 2018.

[7] Hong S. A study on stock price prediction system based on text mining method using LSTM and stock market news. Journal of Digital Convergence, 2020, 18(7).

[8] Mndawe S T, Paul B S, Doorsamy W. Development of a stock price prediction framework for intelligent media and technical analysis. Applied Sciences, 2022, 12(2): 719.

[9] Rajkar A, Kumaria A, Raut A, et al. Stock Market Price Prediction and Analysis. International Journal of Engineering Research & Technology (IJERT) Volume, 2021, 10.

[10] Shankarlingam G, Reddy K T. Predicting a Small Cap Company Stock Price using Python with Best Accuracy Rate: How the Data Science Working for Predictions and Accuracy Rate. Indian Journal of Science and Technology, 2023, 16(48): 4620-4623.

[11] Pourroostaei Ardakani S, Cheshmehzangi A. Stock Market Prediction During COVID-19 Pandemic: A Time-Series Big Data Analysis Method. Big Data Analytics for Smart Urban Systems. Singapore: Springer Nature Singapore, 2023: 23-39.

[12] Sharaf M, Hemdan E E D, El-Sayed A, et al. An efficient hybrid stock trend prediction system during COVID-19 pandemic based on stacked-LSTM and news sentiment analysis. Multimedia Tools and Applications, 2023, 82(16): 23945-23977.

[13] Ronaghi F, Salimibeni M, Naderkhani F, et al. COVID19-HPSMP: COVID-19 adopted Hybrid and Parallel deep information fusion framework for stock price movement prediction. Expert Systems with Applications, 2022, 187: 115879.

[14] Faiz T, Aldmour R, Ahmed G, et al. Machine Learning Price Prediction During and Before COVID-19 and Consumer Buying Behavior. The Effect of Information Technology on Business and Marketing Intelligence Systems. Cham: Springer International Publishing, 2023: 1845-1867.

[15] Kaggle, S&P 500 stock data, https://www.kaggle.com/datasets/camnugent/sandp500, 2018.

[16] Kaggle, S&P 500 Stocks (Web Scraping), https://www.kaggle.com/code/rprkh15/s-p500-stocks-scraper, 2022.

[17] Liu Y, Liu L, Yang L, Hao L, Bao Y. Measuring distance using ultra-wideband radio technology enhanced by extreme gradient boosting decision tree (XGBoost). Automation in Construction. 2021 Jun 1;126:103678.