# Netflix Stock Price Prediction: A Comparison of Linear Regression, k-Nearest Neighbor, and Decision Tree Methods

Jiawei Pang

College of Letters & Science, University of Wisconsin-Madison, Madison, WI, the U.S.

jpang28@wisc.edu

**Abstract.** The implementation of stock prediction has become a major popular topic. However, the effectiveness of different machine learning methods for stock prediction tends to vary. Linear regression, k-Nearest neighbors, and decision trees are three basic machine learning algorithms and are frequently used in the practice of stock prediction. This paper proposes these three methods and compares the effectiveness of applying them to predict the stock price of Netflix by evaluating the mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R-squared). The results of these three methods all display promising predictions, but the linear regression model tends to outperform in this Netflix prediction attempt. This prediction practice finds patterns and trends in past data so as to help investors make more informed decisions in the volatile financial market.

**Keywords:** Machine Learning, Linear Regression, k-Nearest Neighbor, Decision Tree, Netflix.

## 1. Introduction

Reed Hastings founded Netflix, a subscription-based internet movie platform, in 1997. The company started with a DVD rental business using a pay-per-rent model and a monthly membership model, and people could rent a DVD movie and have it delivered directly to their homes [1].

Netflix stock has gone through ups and downs and has experienced multiple stock price spikes, and in 2020 with the occurrence of COVID-19, Netflix's stock price is once again exploding. While other media companies concern about lost revenue from advertising and shutting cinemas, Netflix's reliance on streaming memberships for income has allowed it to record a two-year peak in the industry on Tuesday as the COVID-19 locked people inside [2]. Netflix experienced considerable growth in 2020 and 2021 due to different limitations such as quarantine, which presents a major opportunity for Netflix to grow and generate exciting revenues. However, as the pandemic mandates faded, consumers are moving toward out-of-home activities such as cinemas, restaurants, and amusement centers, and unsurprisingly, Netflix revealed disappointing subscriber numbers and lost the majority of its gain from COVID-19 [3].

Netflix's stock price has arguably seen its share of ups and downs and has been greatly influenced by the environment and society. As one of the world's most famous leaders in video streaming with a large number of subscribers, Netflix's development can be noticed first-hand by the audience. In addition, Netflix brought a lot of levity to the masses during the outbreak, allowing people to find solace in the stressful realities of the world.

The problem with stocks is that they are highly uncertain, and so is Netflix's stock. Financial markets are naturally unstable, and the price of stocks can fluctuate dramatically as a result of a variety of economic, political, and market-specific reasons, as well as news, investor sentiment, interest rates, and a myriad of other factors, as in the case of Netflix, where a sudden quarantine caused by an outbreak of a disease caused a surge in the stock price. Considering these points, creating a powerful predictive model of Netflix's stock would give investors more information, and provide insights into potential future shifts in Netflix's stock, helping investors decide when to buy, sell, or hold their positions to manage risks. Stock prediction models can also benefit Netflix itself, such as helping to determine the best course of action for a company's strategy or policy based on stock predictions. Therefore, providing an effective predictive model helps Netflix to grow and perfect in order to provide better services to the general public.

Machine learning approaches, which incorporate artificial intelligence structures, attempt to extract patterns learned from previous data - a process referred to as training or learning - in order to later generate predictions about newly collected information [4]. Corporations may integrate machine learning technologies to assist customers in making decisions in a range of practical applications. Financial industry has extensively employed this to supply a fresh approach that enables clients to make wiser investment and management choices in order to in order to get greater returns from their securities operations [5]. Generally speaking, it could be challenging for investors to always be informed considering the property of market. Based on this, Obthong [5] further explained that anticipating the movement of the stock market price requires a big amount of information, which is critical for investors as tumultuous stock market can result in a significant reduction of money. As a result, the study of this vast amount of data and financial model is extremely beneficial to investors.

Stock forecast is a widely studied areas in machine learning, and with the development of artificial intelligence and techno; ogy, a variety of algorithms have been now used to forecast changes in the stock market. Hindrayani et al. predicted several stocks in Indonesia and found that decision trees outperformed other methods in the models they used: SVM, multiple linear regression, and KNN, and concluded that the decision tree was a competitive solution for his practice [6]. Polamuri et al. obtain a similar result that decision tree outperformed linear regression [7]. However, Kumar and Kumar suggested that linear regression is more accurate than DT in predicting stock price [8]. Another group of experts, Lawal et al. have practiced a wider variety of supervised machine learning methods and found that ANN, Linear Regression, Support Vector Regression, kNN, and decision tree (DT) all demonstrated predictive results that were worth the surprise [9].

Considering the multiple success stories of using machine learning methods to predict stock prices and the varying effectiveness of different methods, this paper selects three simpler machine learning methods - Linear Regression, k-Nearest Neighbor, and Decision Tree - to perform the prediction and evaluate and compares the accuracy of the three methods to select the one that is the most effective in this Netflix stock price prediction exercise and uses this algorithm to make an adjusted closing price prediction for a period of seven days outside of the dataset.

## 2. Data and Methodology

### 2.1. Data

The dataset is from Kaggle and includes the stock information from February 5, 2018, to February 5, 2022. The adjusted closing price is the target of prediction.

The dataset is first preprocessed. The data information is shown and there is no missing value. Figure 1 - the adjusted closing price from Feb 5th, 2018 to Feb 8th, 2022 - is plotted and it can be seen from the graph that it experienced ups and downs but increased generally, especially after 2020. An outlier check is operated and the result indicates that there are no outliers in the features regarding price. The training set and testing set are separated from the dataset and data standardization is applied to avoid potential error. Cross validation is also implemented to reduce issues such as overfitting. In this case, "Open", "High", "Low" and "Volume" are set to be the independent variables, and the adjusted closing price is the dependent variable. During the training process, three methods - linear regression, k-nearest neighbor, and decision tree - are applied. The outcomes are displayed in graphs and the MAE, MSE, and R squared are calculated for comparison.

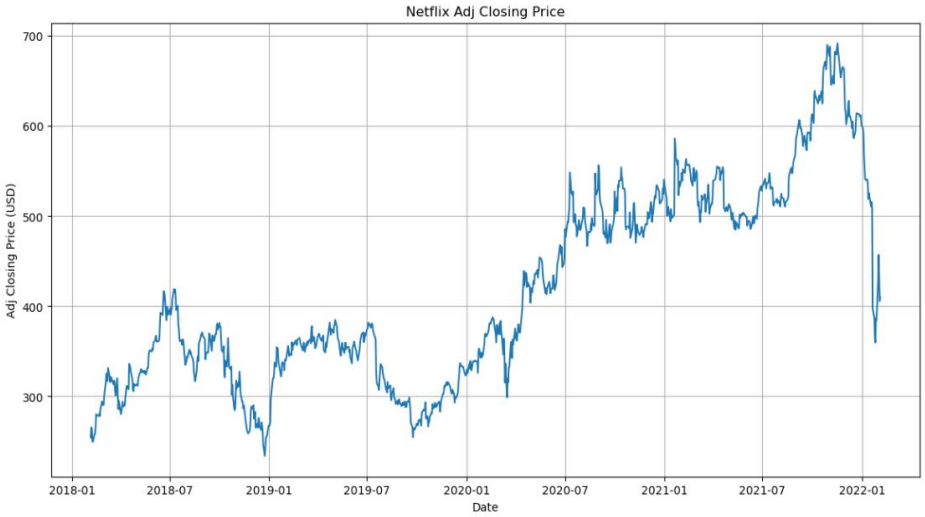



**Figure 1.** Netflix Adjusted Closing Price

## 2.2. Methodology

### 2.2.1 Linear Regression

With the assumption that there is a linear relationship between the input characteristics and the outcome variable, linear regression anticipates an outcome variable from a number of input characteristics [8]. Examples of linear regression models include simple and multiple linear regression, and non-linear or polynomial regression; they are essentially based on the variables that operate as result predictors and the precision of those predictions [10]. In this case, multiple linear regression is applied. The formula (1) indicates how we perform multiple linear regression in the model. Each independent variable corresponds to "$x_i$". All of the remaining features that are taken into account would be considered as $x_i$. The dependent variable, in this case, adjusted price, would be $Y_i$.

$$Y_i = x_i, 1\beta_1 + x_i, 2\beta_2 + \cdots + x_i, n\beta_n \quad (1)$$

### 2.2.2 k-Nearest Neighbor

K-nearest-neighbor (kNN) is one of the most important and efficient algorithms, and it has been widely utilized to predict stock prices. As Singh suggested [11], the fundamental idea underlying kNN is that in a training dataset, kNN may locate an array of k samples that are more comparable to unknown samples. Finding the average of the response variables enables researchers to determine the labels of unknown samples within these k samples. In this practice, k is set to 5, which means that the algorithm considers the five nearest neighbors (data points) to a given data point when making predictions. The five days in the training set that are most comparable to the current day in terms of those features would be evaluated, and then the algorithm would take the adjusted closing stock prices of those five days and average them to make the prediction for the current day.

### 2.2.3 Decision Tree

The decision tree, which functions like a "smart flowchart" based on how a previous series of questions was answered, may evaluate stock market data and forecast whether a stock would gain or fall in value. It is composed of leaf nodes, which stand in for individual outcomes, and inner nodes, which reflect the branch structures that demonstrate the algorithm's judgment. [12]. This process is made up of the node that makes the decision, which has many subdivisions, and the node that represents the leaf, which can be the outcome of the decision nodes but has no extra subdivisions [12]. The prediction process starts and follows the decision path based on the values of features. As it moves down the tree, it will traverse through decision nodes according to the values of the features until it reaches a leaf node, where the prediction is then taken as the predicted adjusted closing price.

# 3. Results and Discussion

## 3.1. Prediction Results of Linear Regression

The line graph reflects a strong outcome. Figure 2 depicts the actual adjusted closing price as the blue line, the train's predicted value as the orange line, and the test's predicted value as the green line. It is clear to see the anticipated price is close to the real value, and the anticipated and real values largely overlap for the most part on the plot.




**Figure 2.** Linear Regression Prediction

## 3.2. Prediction Results of k-Nearest Neighbor

The prediction is also shown on a line graph, which shows a relatively reasonable outcome. From Figure 3, it can be concluded that there is a relatively large gap between many of the predicted results and the actual results, which is even more obvious when compared to the linear regression. However, the accuracy is still guaranteed according to figure 3.




**Figure 3.** k-Nearest Neighbor Prediction

## 3.3. Prediction Results of Decision Tree

Figure 4 depicts the predictive results for the decision tree. It shows expected accuracy, even more accurate than the kNN's prediction. However, the linear regression graph is closer to the actual values than the decision tree ones, especially when the adjusted closing price experienced some obvious fluctuations.




**Figure 4.** Decision Tree Prediction

### 3.4. Performance Comparison

The results of the three methods are all visualized as the graphs as above. According to Figures 2, 3, and 4, we can tentatively see that linear regression has better predictive accuracy than the other two methods. In order to better analyze these three methods, MAE, MSE, and R squared values were calculated for comparison. These three are typical metrics used to assess predictive model performance, notably in the field of regression analysis. They could examine how well a model's predictions resemble the actual values. The lower the MAE and MSE, the closer the prediction is to the actual value. The closer the R squared value is close to 1, the larger the proportion of variability the prediction model can explain. Table 1 contains the calculated results for these three methods. As suggested, kNN has the highest MAE, MSE. Its coefficient of determination is close to one, but it's not as impressive as linear regression and decision tree. When comparing LR and the decision tree, LR has a lower MAE and MSE and has generally a better coefficient of determination. Overall, in this practice about Netflix stock price prediction practice, linear regression performs the best; the decision tree is the second; kNN is the last. The results of this study are more in line with Lawal et al. [9], i.e., for a basic stock prediction model, kNN, LR, and DT are reasonably accurate and also reflect Polamuri et al. [7] point that linear regression is slightly more accurate than the other two.

**Table 1.** Performance Comparison

| | MAE | MSE | R^2 |
|---|---|---|---|
| LR testing | 2.9013 | 16.9936 | 0.9985 |
| kNN testing | 5.1300 | 65.9688 | 0.9941 |
| DT testing | 4.3357 | 40.3518 | 0.9964 |

### 3.5. Prediction Results outside the dataset

Given the LR model's strong reliability, it is used to forecast the adjusted closing price for the following seven days outside of the dataset. The results are displayed in Figure 5. The adjusted closing price for the next seven days floated slightly but generally moved around the $400 to $410 price range.

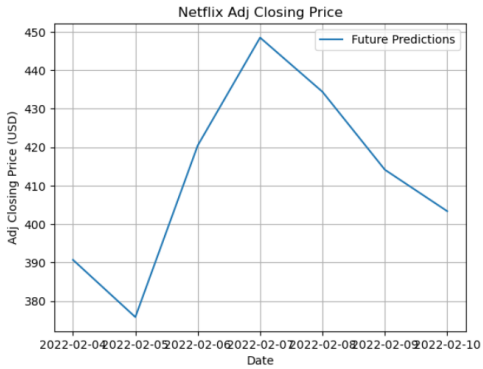



**Figure 5.** Adjusted Closing Price for the Next Seven Days Outside the Dataset

## 4. Conclusion

This paper compares the accuracy of three algorithms LR, kNN and DT for predicting Netflix's stock price, and concludes that all three methods can predict accurately, but linear regression performs slightly better than kNN and decision tree. This conclusion allows investors to refer to the advantages and disadvantages of these three methods, and since linear regression is more accurate in the study, more emphasis can be placed on linear regression in subsequent predictions to help investors make more reliable forecasts. Apart from the guidance of the selection of algorithms, this paper also conducts an out-of-sample prediction to observe the trend of the adjusted closing price in the following seven days and concludes that the price will fluctuate slightly, but the overall change will not be particularly large. This out-of-sample prediction can also serve as a reference for investors to consider whether to buy or sell stocks in the following seven days to assist them in formulating their investment strategies.

However, this is just the result of a simple predictive model and there are far more various factors including data complexity and relationship affecting the effectiveness of methodologies. Additionally, Nabipour's et al. have applied eleven methods and the evaluations show that RNN and LSTM surpass other forecasting techniques including kNN, LR and Decision Tree significantly for continuous data [13]. As a result, for future research, more complex methods such as LSTM, RNN and detailed training processes should be operated to obtain a more solid conclusion.

## References

[1] Wang Yuehan. Netflix: How to Keep a Continued Success. In 2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022). Atlantis Press, 2022, 1215-1219.

[2] Shapiro A. Netflix's stock is now higher than before coronavirus hit U.S.. Forbes, 2020.

[3] Whitten S. Netflix stock has now lost all its gains from the pandemic. CNBC, 2022.

[4] Xiao Yi, Xiao Jin, Lu Fengbin, Wang Shouyang. Ensemble ANNs-PSO-GA approach for day-ahead stock e-exchange prices forecasting. International Journal of Computational Intelligence Systems, 2014, 7 (2), 272-290.

[5] Obthong M., Tantisantiwong N., Jeamwatthanachai W., Wills G. A survey on machine learning for stock price prediction: Algorithms and techniques. 2020.

[6] Hindrayani K. M., Fahrudin T. M., Aji R. P., Safitri E. M. 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), IEEE, 2020, 344-347.

[7] Polamuri S. R., Srinivas K., Mohan A. K. Stock market prices prediction using random forest and extra tree regression. Int. J. Recent Technol, 2019, Eng, 8 (1), 1224-1228

[8] Kumar D. B., & Kumar S. M. Improved accuracy in prediction of stock equity analysis and classification using linear regression and compared with k nearest neighbor. Eur. Chem. Bull. 2023, 12 (S1), 4686-4691.

[9] Lawal Z. K., Yassin H., Zakari R. Y. Stock market prediction using supervised machine learning techniques: An overview. In 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), IEEE, 2020, 1-6.

[10] Priya, K. S. Linear regression algorithm in machine learning through MATLAB. Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET), 2021, 9, 12.

[11] Singh G. Machine learning models in stock market prediction. arXiv preprint arXiv: 2202.09359. 2022.

[12] Bansal M., Goyal A., Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning. Decision Analytics Journal, 2022, 3, 100071.

[13] Nabipour M., Nayyeri P., Jabani H., Shahab S., Mosavi A. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access, 2020, 8, 150199-150212.