International Journal of
Applied Data Science in Engineering and Health
https://ijadseh.com

Stock Price Forecasting for Netflix with Machine Learning Models

Budak Azangel

[1] *Department of Mathematics, Middle East Technical University, Türkiye (Budak.Azangel@gmail.com)*

Received date: September 8, 2024; Accepted date: October 30, 2024

Abstract

Stock price prediction is a challenging yet valuable task in financial forecasting. This paper explores the application of machine learning for predicting Netflix's stock price, comparing Linear Regression and Random Forest models. Using a five-year dataset, we evaluated model performance based on R^2 Score and Mean Absolute Error (MAE). The Random Forest model outperformed Linear Regression, achieving a higher R^2 Score (0.7202) and lower MAE (0.4092), indicating better predictive accuracy. These results suggest that ensemble methods like Random Forest offer more reliable stock price forecasts, particularly in complex, volatile financial datasets.

Keywords*:* Stock Prediction; Machine Learning; Linear Regression; Random Forest.

Introduction

The application of machine learning techniques to financial forecasting has gained significant traction in recent years, promising advancements in the accuracy and reliability of stock price predictions. Stock markets are complex and often characterized by nonlinear dependencies influenced by factors such as economic indicators, investor sentiment, historical price movements, and external events like geopolitical developments [1-4]. While traditional time-series models like autoregressive techniques have been useful, they frequently struggle to capture the nuances of modern financial data. Machine learning, on the other hand, offers an adaptive approach that can model these complexities with greater flexibility [5-9]. In this study, we focus on predicting the stock price of Netflix Inc., a major player in the entertainment streaming and media production industry. Given the volatility of Netflix's stock, influenced by its business dynamics and broader market forces, accurate predictive models are essential for investors and stakeholders seeking data-driven insights.

Netflix, as a growth-oriented company, faces rapid stock price fluctuations driven by internal factors like content launches and subscriber growth, as well as external pressures from competition, regulatory changes, and economic trends. The company's stock price behavior presents a valuable yet challenging case for prediction, as it reflects both the evolving nature of consumer behavior and the competitive streaming industry. Predicting Netflix's stock price accurately not only aids in investment decisions but also enhances our understanding of how industry-specific and market-wide factors interplay to drive stock movements.

This study uses two machine learning models to predict Netflix's stock price: Linear Regression and Random Forest. Linear Regression, a fundamental approach in statistical modeling, provides a straightforward method for assessing the relationship between historical stock prices and selected predictor variables. By assigning weights to each feature, Linear Regression allows for a clear interpretation of how each variable contributes to stock price predictions. However, due to its linear nature, this model may have limitations in capturing the complex, non-linear relationships that often characterize stock market data. Random Forest, an ensemble learning method, overcomes this limitation by creating multiple decision trees and averaging their predictions. By capturing nonlinear patterns and handling a variety of input data effectively, Random Forest offers a robust approach to financial forecasting. The comparative analysis of these two models highlights the trade-offs between simplicity and predictive power in stock market prediction tasks.

The objective of this study is to evaluate the effectiveness of Linear Regression and Random Forest models in predicting Netflix's stock price. Through a detailed comparison, we aim to identify which model offers superior predictive accuracy and explore the extent to which model complexity and feature selection impact prediction outcomes. By contributing to the broader field of AI-driven financial analytics, this research offers insights into the strengths and limitations of machine learning models in stock price prediction, particularly within the high-growth digital media industry. This paper ultimately aims to demonstrate how machine learning models can inform investment strategies and decision-making in fast-evolving sectors like streaming entertainment.

Methods and Material

Dataset

The dataset used in this study, titled "Netflix Stock Price Prediction," provides comprehensive daily stock data for Netflix Inc. (NFLX) over five years, spanning from February 5, 2018, to February 4, 2022. This period captures a dynamic phase in Netflix's growth trajectory, encompassing market events and company-specific developments that have influenced its stock behavior. The dataset's period allows for a robust analysis of both short-term fluctuations and long-term trends, offering a solid foundation for training predictive models to forecast future price movements [10]. The dataset comprises seven key columns, each serving a unique purpose in capturing various aspects of Netflix's stock trading. The Date column marks each entry in the time series, specifying the exact trading date.

The Open column records the initial trading price of Netflix's stock for a given day, reflecting the initial sentiment of the market. The High and Low columns provide the highest and lowest prices that Netflix's stock reached during the trading day, representing the day's peak and minimum values. The Close column denotes the stock's closing price, adjusted for any stock splits, and is commonly analyzed as it reflects the market sentiment at the end of trading. In addition to the closing price, the Adj Close column shows the adjusted closing price, accounting for stock splits, dividends, and capital gains to offer a normalized historical comparison. Finally, the Volume column tracks the number of Netflix shares traded on a particular day, providing insight into trading activity and liquidity.

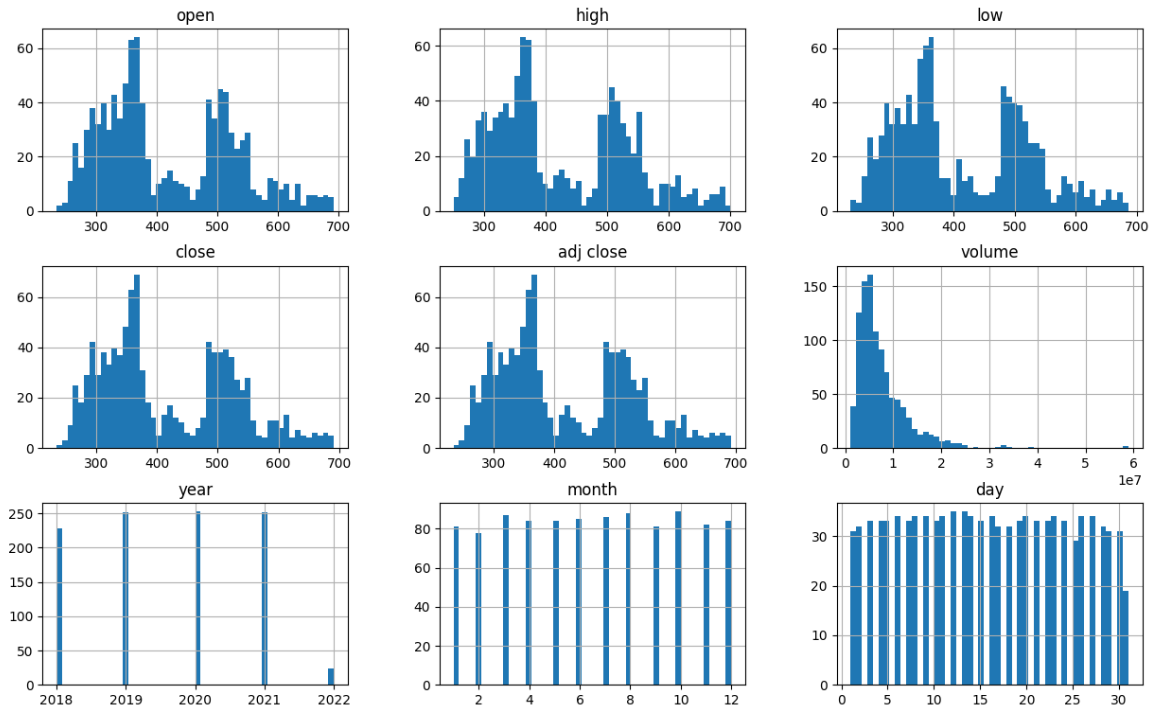

Figure 1 This image presents histograms of key features in the Netflix stock price dataset over a five-year period, from February 2018 to February 2022. The histograms for Open, High, Low, Close, and Adj Close prices reveal a distribution with multiple peaks, indicating periods of price fluctuation and volatility across the observed years. These peaks suggest distinct phases in Netflix's market performance, potentially reflecting the impact of major market events, company-specific developments, or changes in investor sentiment. The Volume histogram, in contrast, shows a right-skewed distribution, with the majority of trading volume concentrated at lower values, while a few days witnessed exceptionally high trading volumes. This spike in trading volume on certain days may correspond to significant events, such as earnings reports or market-wide movements. Additionally, the Year, Month, and Day histograms confirm a uniform distribution of data points across time, ensuring consistent daily, monthly, and yearly coverage throughout the dataset, which is essential for reliable time-series analysis. These visualizations provide a foundational understanding of the stock's historical price behavior and trading activity, offering valuable insights for subsequent predictive modeling.

Linear Regression

Linear Regression is a foundational statistical method commonly used for predicting continuous outcomes based on one or more predictor variables. This study serves as a straightforward approach for forecasting Netflix's stock price by modeling the linear relationship between stock prices and features such as historical prices, volume, and time-related variables. Linear Regression works by fitting a line through the data points in a way that minimizes the sum of squared residuals, effectively capturing the average trend within the data.

This model takes in predictor variables, such as the opening price, trading volume, and previous closing prices, and aims to establish a linear equation that best fits the observed data. In practice, the dataset is split into training and testing sets, with the model trained on the historical data to learn the relationship between these variables and the stock's closing price. Once trained, the Linear Regression model can generate predictions by applying the learned weights to the features in the test data. However, while Linear Regression provides a useful baseline and an interpretable model, it has limitations in capturing the potentially complex, nonlinear patterns often present in stock market data. This limitation makes it less effective in predicting sudden shifts or irregular price movements that are not well-represented by a simple linear relationship.

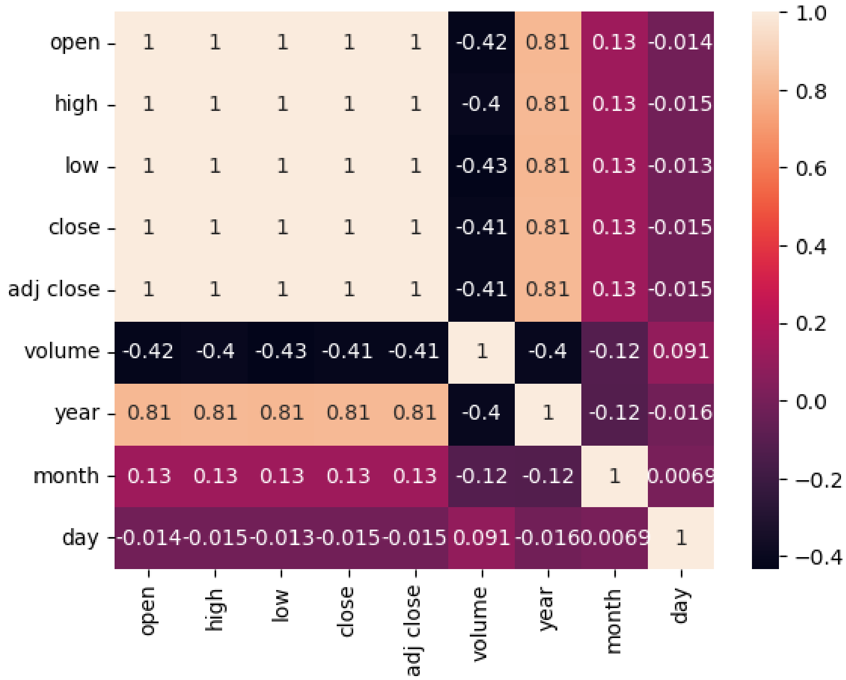

Figure 2 This heatmap shows the correlation matrix for the Netflix stock price dataset, highlighting the relationships between key features. The Open, High, Low, Close, and Adj Close prices exhibit a perfect positive correlation (1.0) with each other, indicating that these price values move in tandem. This strong interdependence is typical in stock data, where opening, high, low, and closing prices are closely linked within a trading day. The Volume feature shows a moderate negative correlation with these prices, suggesting that higher trading volumes are slightly associated with lower price levels. Additionally, the Year feature has a moderately positive correlation (0.81) with price-related variables, which could indicate an upward trend in Netflix's stock price over the observed years. Other temporal variables like Month and Day show low or near-zero correlations with stock prices and volume, suggesting that within this dataset, seasonal or daily patterns have minimal impact on Netflix's stock performance. This correlation analysis provides insights into feature interdependencies, which can guide feature selection and preprocessing steps for predictive modeling.

Random Forest

Random Forest, in contrast, is an ensemble machine learning method that constructs a collection of decision trees to make robust predictions, offering a more sophisticated approach for stock price prediction in this dataset. Unlike Linear Regression, which assumes a linear relationship, Random Forest can model intricate, non-linear interactions between variables, making it well-suited for financial data, which often exhibits irregular patterns. To use Random Forest in this task, the dataset is also split into training and testing sets. During training, the Random Forest model builds multiple decision trees, each trained on random subsets of the data and features.

Each tree independently learns different patterns within the data, and the model's final prediction is obtained by averaging the predictions across all trees, effectively capturing complex dependencies without overfitting to noise in any single decision tree. This ensemble approach enables Random Forest to capture a more nuanced view of the factors driving stock price movements, offering an advantage over simpler linear models.

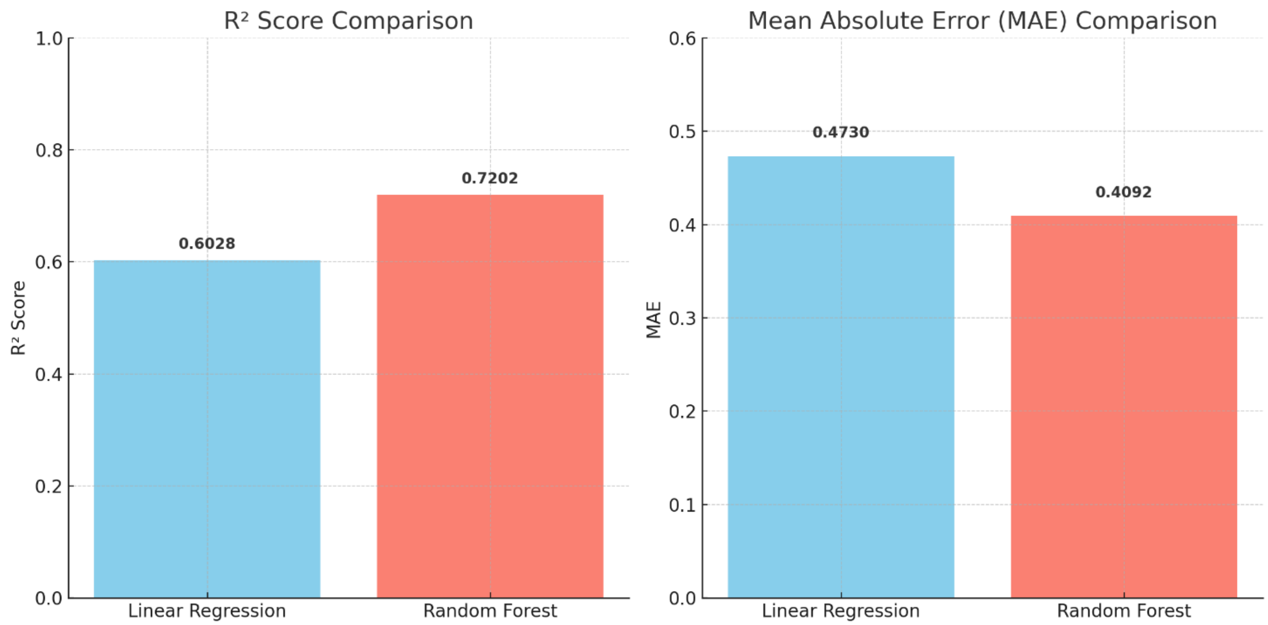

Figure 3 This visualization compares the performance of Linear Regression and Random Forest models based on two key metrics: R^2 Score and Mean Absolute Error (MAE). The chart on the left shows the R^2 Score, where Random Forest achieves a higher value (0.7202) compared to Linear Regression (0.6028), indicating that Random Forest explains a greater portion of the variance in the stock prices. The chart on the right displays the Mean Absolute Error, with Random Forest again performing better, showing a lower MAE (0.4092) than Linear Regression (0.4730). This lower MAE suggests that Random Forest predictions are, on average, closer to the actual stock prices, demonstrating its superior accuracy for this prediction task.

Results

The performance of the Linear Regression and Random Forest models was evaluated based on their R^2 Score and Mean Absolute Error (MAE), two metrics commonly used to assess the fit and accuracy of predictive models. The Linear Regression model achieved an R^2 score of 0.6028, indicating that it explains approximately 60.28% of the variance in Netflix's stock price. This suggests that while the model captures some of the relationship between the features and the stock price, it falls short of fully accounting for the complexity inherent in stock price movements. Additionally, the Mean Absolute Error (MAE) for Linear Regression was 0.4730, showing that on average, the model's predictions are relatively close to the actual stock prices but lack the precision seen in the Random Forest model. The Random Forest model outperformed Linear Regression in both metrics. With an R^2 score of 0.7202, Random Forest explains around 72.02% of the variance in Netflix's stock price, demonstrating a stronger fit to the data. This higher R^2 score reflects the model's capacity to capture more intricate patterns and nonlinear relationships within the dataset. Furthermore, the Random Forest model's MAE was 0.4092, lower than that of Linear Regression, indicating that its predictions are, on average, closer to the actual stock prices. This superior performance highlights Random Forest's ability to produce more accurate predictions, making it a more suitable choice for this stock price prediction task.

In conclusion, the Random Forest model provided better fit and more accurate predictions than the Linear Regression model. Its higher R^2 score and lower MAE suggest that Random Forest is more effective at capturing the complexity of Netflix's stock price data, making it a valuable tool for financial forecasting in this context.

Conclusion

In this paper, we explored the effectiveness of two machine learning models, Linear Regression and Random Forest, for predicting Netflix's stock price over a five-year period. By evaluating these models using key performance metrics—R^2 Score and Mean Absolute Error (MAE)—we assessed their ability to capture the complexity and variability of stock price movements. The results demonstrate that the Random Forest model outperforms Linear Regression, as evidenced by its higher R^2 score of 0.7202 and lower MAE of 0.4092. This indicates that the Random Forest model explains a greater proportion of the variance in Netflix's stock price and provides predictions that are closer to the actual values. Given the superior performance of the Random Forest model, it is the preferred model for this prediction task. Its ability to handle nonlinear relationships and capture intricate patterns within the data makes it better suited for the volatile nature of stock prices. By contrast, the Linear Regression model, while useful as a baseline, lacks the flexibility required to accurately model these complex dependencies.

Conflict of interest

The authors declared no conflict of interest.

References

[1] Shah D, Isah H, Zulkernine F. Stock market analysis: A review and taxonomy of prediction techniques. International Journal of Financial Studies. 2019 May 27;7(2):26.

[2] Kumar D, Sarangi PK, Verma R. A systematic review of stock market prediction using machine learning and statistical techniques. Materials Today: Proceedings. 2022 Jan 1;49:3187-91.

[3] Strader TJ, Rozycki JJ, Root TH, Huang YH. Machine learning stock market prediction studies: review and research directions. Journal of International Technology and Information Management. 2020;28(4):63-83.

[4] Kumar G, Jain S, Singh UP. Stock market forecasting using computational intelligence: A survey. Archives of computational methods in engineering. 2021 May;28(3):1069-101.

[5] Tashakkori A, Talebzadeh M, Salboukh F, Deshmukh L. Forecasting Gold Prices with MLP Neural Networks: A Machine Learning Approach. International Journal of Science and Engineering Applications (IJSEA). 2024;13:13-20.

[6] Rath S, Das NR, Pattanayak BK. An Analytic Review on Stock Market Price Prediction using Machine Learning and Deep Learning Techniques. Recent Patents on Engineering. 2024 Feb 1;18(2):88-104.

[7] Ayyildiz N, Iskenderoglu O. How effective is machine learning in stock market predictions?. Heliyon. 2024 Jan 30;10(2).

[8] Sui M, Zhang C, Zhou L, Liao S, Wei C. An ensemble approach to stock price prediction using deep learning and time series models.

[9] Tashakkori A, Erfanibehrouz N, Mirshekari S, Sodagartojgi A, Gupta V. Enhancing stock market prediction accuracy with recurrent deep learning models: A case study on the CAC40 index. World Journal of Advanced Research and Reviews. 2024;23(1):2309-1.

[10] https://www.kaggle.com/datasets/jainilcoder/netflix-stock-price-prediction