Predictive Analytics for Strategic Investment

Trung Nguyen, Lillian Swan (Lily), Zehao Wang (Steven)

Abstract

The stock market is an important indicator for both proficient technology investors and financial analysts in a time when technology and the financial markets are evolving quickly. Even if this market's natural volatility offers opportunities for profit, it also presents several difficulties, especially for individual investors trying to make their way through these uncertain times. This program uses advanced machine learning techniques, such as K-Nearest Neighbors (KNN) classification and linear regression, to analyze large datasets from Yahoo Finance. The goal of this project is to improve individual investors' ability to make decisions while also making high-quality analytical tools more accessible to a larger population. Our effort establishes the foundation for well-informed decision-making, risk reduction, and the effective distribution of investment resources in the ever-changing stock market by providing insights into market trends, stock performance, and investment opportunities.

Introduction

1.1 Problem Statement

The stock market, an example focused on development investors and high-tech companies, represents the opportunities and risks inherent in the current financial environment. Due to market volatility and the lack of easily understood analytical tools, individual investors who are attracted to the possibility of significant profits frequently find themselves in a tricky situation. It is difficult to accurately predict the growth pattern of growing, industry-correlated stocks and compare the performance of comparable businesses in these emerging industries.

The lack of tools in the typical investor's analytical tool shows the urgent need for an approach change. The planned revolution involves developing easily navigable

and approachable platforms that utilize machine learning to reduce large datasets into insights that can be used. These kinds of tools would not only show investors the way to the future but also provide them with the tools they need to navigate the treacherous waters of the Nasdaq with skill and confidence.

1.2 Background

There is not any doubt about the attraction of the stock market, with its dense of high-tech companies. The stock market is the arena for the majors of technology and innovation, from new startups to massive businesses. Because of this concentration of tech-related companies, investors seeking to capitalize on the next wave of technical innovations and the potential financial rewards they herald are drawn to stock markets like Nasdaq and S&P 500. But the very concentration gives the market a volatility that can be both profitable and restrictive.

Introduction to the Data

2.1 Data summary

Our study uses historical stock price data sourced directly from Yahoo Finance. Our program offers users a user-friendly interface to input their preferred company and starting date for predictive analysis. By entering the stock ticker symbol of the company, our program retrieves the historical daily stock data from Yahoo Finance. The features of the dataset include Date, Open, High, Low, Close, and Volume. These features form the basis of our analysis and provide valuable information about daily market movements and trends.

2.2 Data source and collection

The data was obtained from Yahoo Finance through the “yfinance” library. Yahoo Finance, a reputable financial media website within the Yahoo network, serves as a comprehensive platform offering financial news and stock data, including indices like Nasdaq and S&P 500. From the functionalities of “yfinance” library, our program retrieves the historical stock daily data directly from Yahoo Finance.

Given that the source of the data is public and its origin from an official financial platform, we are confident in its dependability for analysis. Furthermore, the “yfinance” library is distributed under the Apache Software License. The library is an open-source tool that uses Yahoo's publicly available APIs and is intended for research and educational purposes.

2.3 Potential Bias

While our quantitative methodology seeks to reduce subjective biases, we recognize that stock market data has essential biases. Historical bias, for example, assumes that past trends will continue, which can be problematic given the volatile nature of the stock market. Selection bias may also result from our focus on leading technology businesses, thus restricting the generalizability of our results.

When selecting statistical models, we sought to reduce algorithmic biases by using thorough model selection and validation approaches, assuring the accuracy of our findings.

Data Science Approaches

3.1 Importing Libraries

The code begins with the importation of four essential libraries: “yfinance,” “matplotlib,” “numpy,” and “scikit-learn.” Within the scope of our task, "scikit-learn" plays a vital role in data science algorithms, offering key functionalities essential for predictive modeling. “MinMaxScaler” is utilized for data normalization that scales each feature within a specified range, typically between 0 and 1. This normalization is important for many machine learning algorithms, which often operate under the assumption that all features exhibit a similar variance and are centered around zero. “KNeighborsRegressor” is an implementation of the K-Nearest Neighbors (KNN) algorithm for regression tasks. It predicts the value for a new data point by averaging the values of the K-nearest neighbors to this new point. “mean_squared_error” serves as a risk metric, measuring the difference between the values predicted by a model and the values observed from the environment being modeled. The metric is calculated as the expected value of the squared error or loss.

3.2 Data Science Algorithms

In our program, we used the K-Nearest Neighbors (KNN) algorithm to predict stock closing price. KNN algorithm is a popular machine learning technique used for classifications and regression tasks. The idea of the KNN algorithm is that similar data points tend to have similar labels or values. During the training, the KNN algorithm stores the training dataset as a reference. In our program, the code starts by building a data frame with features of date and closing price by fetching the data with “yfinance” library. To prepare the data for modeling, the program normalizes the data to a common scale for machine learning. The program uses 80% of the stock price data as training

data and fits the data to a KNN regressor model. By identifying the 5 closest neighbors to each test data point in the training data and averaging their target values, the program memorizes the patterns between input features and target values of closing prices. As a result, the program predicts future closing prices for the remaining 20% of the data based on the model.

To evaluate the performance and accuracy of the model, the program uses the Root Mean Squared Error (RMSE). The metric calculates the average difference between values predicted by a model and values present in the actual data. Lower RMSE values signify a closer alignment between predicted and actual data, indicating the model's prediction is accurate.

Results and Conclusions

4.1 Results

In our stock prediction analysis, we conducted the analysis using the stock tickers, APPL (Apple Inc.) and NKE (Nike), as examples. Both companies were analyzed using the same start date, January 1st, 2023. The implementation of the K-Nearest Neighbors (KNN) algorithm for stock closing price prediction yielded promising results across both companies.

When analyzing the stock ticker "AAPL" (Apple Inc.) with a start date of January 1st, 2023, the predicted trend closely aligned with the actual stock closing prices. As Figure 1 shows, the prediction of Apple's closing price aligns to the actual stock closing price closely during March to April in 2024. The model's performance is notable, as indicated by a Root Mean Squared Error (RMSE) score of approximately 8.2904. The

RMSE score falls within a reasonable range, considering the randomness of the stock market.

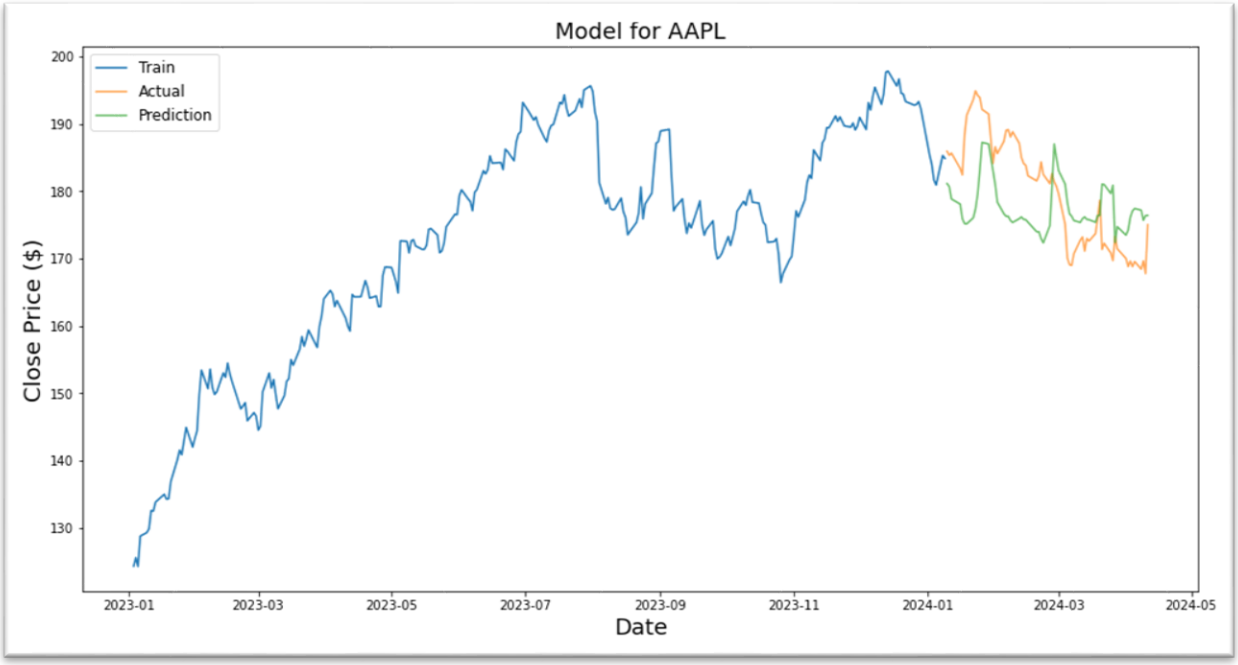

Figure 1: Prediction of AAPL (Apple Inc.) starting 2023-01-01

In another test scenario, we used the stock ticker “NKE” (Nike) with a starting date of January 1st, 2023. As Figure 2 shows, the predictive trend closely aligned with Nike's actual stock prices. Compared to the performance observed with Apple Inc., the model demonstrated an even stronger fitting to Nike's stock price dynamics. The RMSE score for Nike's stock prediction model was approximately 4.3393, indicating a prominent level of accuracy in the model's predictions.

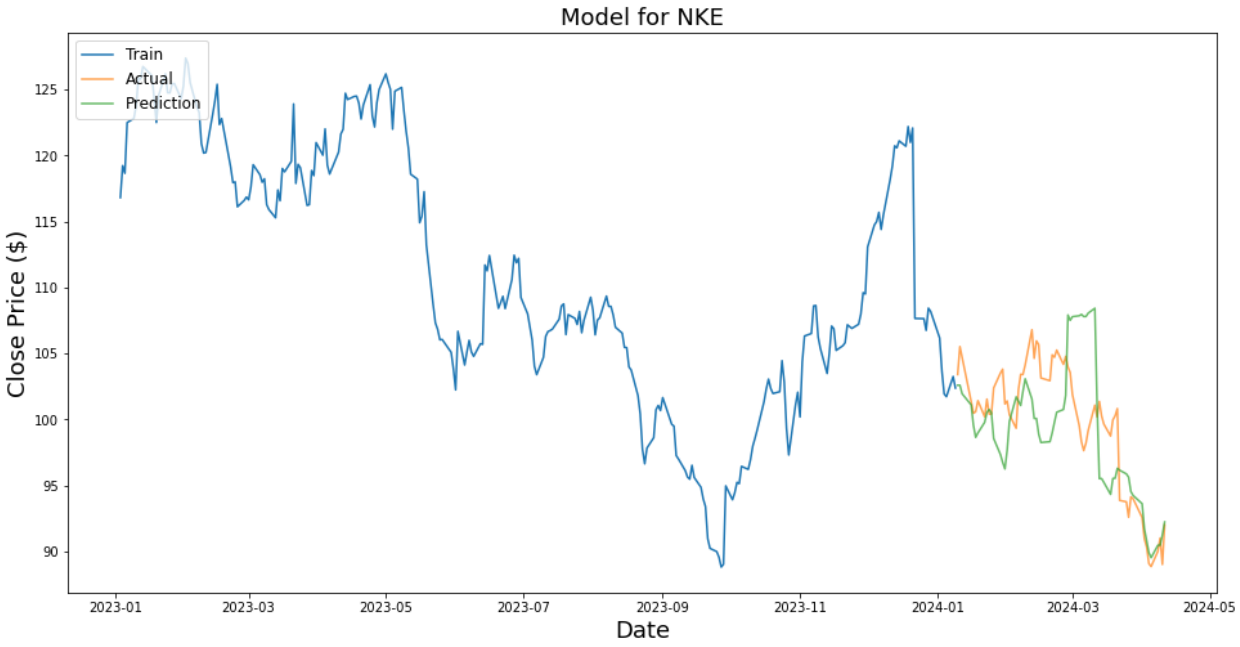

Figure 2: Prediction of NKE (Nike) starting 2023-01-01

4.2 Conclusions

Our experimentation across different stocks has shed light on the behavior of the K-Nearest Neighbors (KNN) algorithm in predicting stock prices. We found that the KNN algorithm performs exceptionally well when trained on historical data showing significant fluctuations in closing prices over time.

In our analysis of stocks such as Apple and Nike, particularly Nike, we noted a significant variability in the closing prices within the training dataset. This variability provided the KNN algorithm with rich and diverse information about the evolving trends in stock prices, enabling it to effectively capture and predict intermediate fluctuations in stock prices.

Conversely, when applying the KNN algorithm to stocks characterized by stable closing prices over time within the training data, the model's performance tends to be

inaccurate. In such cases, the algorithm struggles to discern and predict intermediate significant changes in stock prices accurately. As a result, the prediction often yields higher Root Mean Squared Error (RMSE) scores.

Future Work and Improvements

To advance the robustness and predictive accuracy of our model, several enhancements are proposed for future development. First, incorporating lifetime data can provide a comprehensive view, enabling the prediction of trends over extended periods as well as seeking out patterns that may repeat within those periods. Efforts to decrease the Root Mean Square Error (RMSE) will also be crucial, as a lower RMSE signifies a model with higher precision. Furthermore, integrating advanced techniques such as sentiment analysis and neural networks could offer deeper insights into the data's underlying patterns. Lastly, enriching the model by considering external factors such as inflation, political events, and traders' expectations based on the other competitor companies' performance could yield a more holistic and responsive tool, better equipped to adapt to dynamic market conditions.