Stock Price Analysis and Prediction Method Based on Machine Learning: Taking Apple Inc as an Example

Yixuan Jin *

Shandong Zibo Experimental High School, Shandong, China

* Corresponding Author Email: jinyixuan2005@gmail.com

Abstract. Stock forecasts are analyses of Apple's future performance based on financial data, market dynamics and macroeconomic factors. However, there are conflicting arguments that the wider the time horizon of the data, the more accurate the forecast. These forecasts are crucial for investment decisions, risk management and corporate governance. Therefore, in this paper, we will use vector autoregressive modelling to compare nine training sets with different time horizons and evaluate these nine sets of predictions by calculating the weights of the corresponding variables in the predictions. Knowledge of machine learning and graphical visualization is used to evaluate the share of five factors affecting stock prices as well as the training time horizon. This paper demonstrates that in the field of stock prediction the closer the time horizon is to the prediction the closer it is to the actual value. At the same time investors should consider multiple factors to diversify the risk.

Keywords: Stock Price Analysis, Vector Autoregression model, Time Series Prediction.

1. Introduction

People forecast the stock market in order to try to understand and interpret the future movements of the stock market in order to make informed investment decisions. This involves studying and analyzing a wide range of factors such as macroeconomics, company fundamentals, technical indicators and market psychology. Despite the uncertainty surrounding the future development of the stock market, by observing and analyzing market dynamics, one hopes to gain a better understanding of the risks and opportunities in order to develop a more effective investment strategy. Meanwhile stock prices are influenced by a variety of factors whose importance changes over time and market conditions. The first and foremost factor is the company's fundamentals, including profitability, financial condition and market position, which have a direct impact on the share price. Next in line are macroeconomic factors such as GDP, inflation and interest rate policies, which have a significant impact on overall market sentiment and company earnings. Factors such as market psychology and sentiment, company news and events, geopolitical risks, industry competition, technological factors and monetary policies also shape share prices to varying degrees. [1] Stock market analyses and stock forecasts are critical for investors and market participants as they provide insight and decision support regarding the future movement of the stock market. By analyzing in-depth factors such as company fundamentals, macroeconomic conditions, market psychology and technical indicators, investors are able to better understand the market, identify investment opportunities, mitigate risks, as well as achieve financial goals. In addition, stock forecasting has important applications for parties such as financial institutions, government policy makers and corporate managers, helping to make decisions on capital allocation, risk management and strategic decisions. However, it is important to bear in mind that stock market forecasting is subject to uncertainty and that successful investment still requires prudent and diversified strategies.

In the field of finance, there exists a wide variety of stock price forecasting models and stock market analysis models that aim to predict future stock price movements based on historical data and various factors. These models can be classified into different types such as fundamental based analysis, technical analysis and quantitative analysis. Below are some common stock price prediction models and some of their results and conclusions:

Fundamental based analysis models. DCF Model (Discounted Cash Flow Model): this model estimates the fair value of a stock by estimating the future cash flows of the company and discounting them to present value. The result depends on the accurate forecast of future cash flows and the discount rate chosen; PE Comparative Analysis: This method compares a company's P/E ratio with its industry or market average to determine the stock's valuation level.

Technical Analysis Models. Moving Averages (Moving Averages): Technical analysts use moving averages of different durations to identify trends and levels of support and resistance in the market; Relative Strength Indicator (RSI): RSI measures the relative strength of a stock price and is used to determine if it is an overbought or oversold signal [2]

Quantitative analysis models. Machine Learning Algorithms: These include Decision Trees, Random Forests, Neural Networks, etc. These algorithms can process large amounts of data, identify patterns and generate predictions; Statistical Arbitrage Strategies: use statistical methods to find price differences or correlations to construct arbitrage trading strategies [3]

The results and conclusions of the models vary depending on the model type, input data and parameter settings. It should be emphasized that stock market forecasting is a very complex and uncertain task, and no model can predict stock prices with 100% accuracy. In addition, markets change rapidly, and new information and events may change stock price movements at any time. Studies have found that investment strategies based on long-term fundamentals usually perform well over the long term, whereas technical analyses and short-term quantitative strategies may be more suitable for short-term trading. At the same time, some investors and fund managers also adopt hybrid strategies, which use a combination of models and methods to reduce risk. [4]

In this paper, it focusses on applying knowledge points related to machine learning to fit more realistic results using different time ranges brought into a vector-variable autoregressive model [5] in a python environment. Research has shown that the results predicted using small recent time ranges will be more in line with the real data.

2. Metrology

2.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) [6] serves to provide an in-depth understanding and interpretation of data sets, revealing patterns and trends inherent in the data through visualisation and statistical analysis methods. It helps data scientists identify outliers, deal with missing data, explore relationships between features, and provide a basis for problem definition and hypothesis formation. This process provides key insights and direction for the subsequent stages of data analysis and modelling. Through the initial exploration of the data, EDA helps to identify features and distributions in the dataset, discover potential associations and patterns, and prepare the data for further analysis. It also helps in cleaning and preparing the data to improve the quality of the data. Mainly through the two steps of the heat map of correlation between features and feature pairplot provided the direction for my model prediction later on, allowing me to drop the variable of volume.

In data analysis and exploratory data analysis (EDA), I employ two common visualization tools involved in feature correlation analysis, feature correlation heat map [7] and pairplot of features [8], which have different principles and roles." Pairplot of features" is a common chart type in the field of data visualisation and is a feature in tools such as the seaborn library that is used to create a matrix containing multiple scatter plots and histograms to represent the relationships between the six variables of my original dataset. Specifically, the pairplot combines multiple numerical features from my stock dataset two by two, and then plots the relationships between those features on a scatterplot. Histograms or kernel density estimates are typically plotted on the diagonal and are used to show the distribution of each feature. Plots on non-diagonal lines show scatter plots between different features for observing relationships between them, such as linear correlation, clustering structure, etc. This visualization method helps to quickly identify patterns, trends and correlations between features. A feature correlation heat map is also used to show the correlation between different features (variables)

in a dataset. Typically, it is a rectangular matrix where the color or color scale of each cell indicates the strength of the correlation between the corresponding features. Combining these two plots with the correlation results from EDA allows variables with little correlation to be eliminated, thus providing useful insights for further data analysis and fitting functions using vector autoregressive models.

2.2. Vector Autoregression model

Through the vector autoregressive model, which aims to analyze and model the interrelationships between multiple time series variables, the following objectives are achieved: firstly, it allows us to simultaneously consider the dynamic interactions between multiple relevant variables affecting the price of a stock, which is essential for revealing the correlations and trends within the system. Secondly, the VAR model can be used to predict the values of each variable at future points in time, which helps in decision making and planning. In addition, it can be used for shock analysis to simulate the impact of external and internal shocks on the system, which helps in risk management and policy formulation. By training the dataset for different time horizons, the magnitude of the impact of each variable on the opening price of the stock at the next moment is explored and the weights of each variable are derived separately.

The vector autoregressive (VAR) model is a statistical model used to model the dynamic relationships between multiple time series variables. The core idea of the model is to describe the linkages between variables by linearly combining the past values of each variable at the current moment and introducing an error term. The key steps of the VAR model include determining the appropriate lag order, estimating the model parameters, carrying out the model diagnosis and selection, and finally for predicting and explaining the interactions of multiple time series variables. The VAR model has flexibility and a wide range of areas of application and is particularly suitable for macroeconomics, financial market analysis and other time series data analyses that need to capture complex dynamic relationships among variables.

3. Experiment

3.1. Data

Apple inc. [9] is a globally recognized technology giant headquartered in Cupertino, California, USA. Since its inception, Apple Inc. has been enjoying great success worldwide with its innovative products and services. Its products include iPhone, iPad, Mac, Apple Watch, and a variety of software and services, such as the iOS operating system, App Store and iCloud [10]. Apple's stock is the center of attention for investors around the world, and its market capitalization and share price volatility have a significant impact on the stock market. My dataset was found from Kaggle containing a dataset of Apple stock prices from 1980 to 2021.It was collected from Yahoo Finance. This dataset includes the opening price, closing price, high price, low price, adjust close, and volume of Apple stock for every day for over 40 years. It is worth noting that there are no missing values or outliers in this dataset. After getting a basic understanding of the data, I dropped the "Volume" column, leaving me with 5 columns: Open, High, Low, Close, and Adj Close. I divided the data into 9 training and test sets with different time ranges. The test set is "2022.5.17-2022.6.17" for one month, and the training set is one year, 5 years, 10 years, 15 years, 20 years, 25 years, 30 years, 35 years and 40 years from "2022.5.17", 35 years and 40 years. This gives a better idea of the difference between the predicted and actual values for different time ranges.

3.2. EDA(Exploratory Data Analysis)

I loaded the collected dataset into a series of libraries such as Python's Pandas library. I also performed a series of data overviews (looking at the first few rows of the data, column names, data types, etc.), data cleaning (dealing with missing values, outliers, and duplicates), data visualisation (plotting histograms, scatter plots), and statistical summarisation (calculating statistical metrics such

as the mean, median, and standard deviation of the data) in order to facilitate an understanding of the basic structure of the data and to ensure the quality and accuracy of the data. Especially through the heat map of correlation between features (Fig.1) and pairplot of feature (Fig.2), I graphically represent the relationship between each two variables, which can be clearly seen " Volume" is not very relevant to the other five variables, so I dropped this variable.

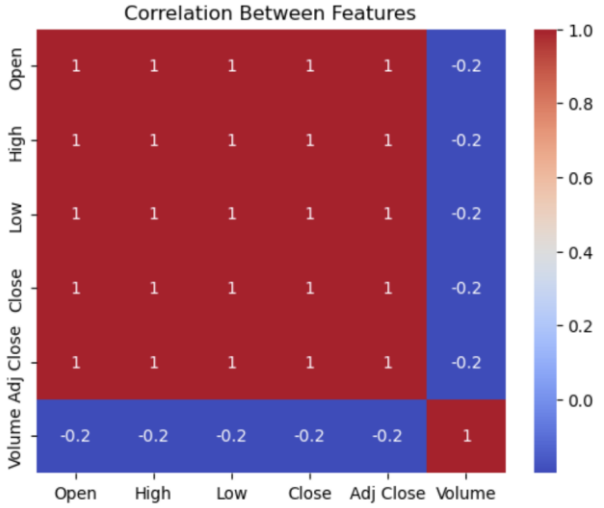

Fig. 1 Correlation between features

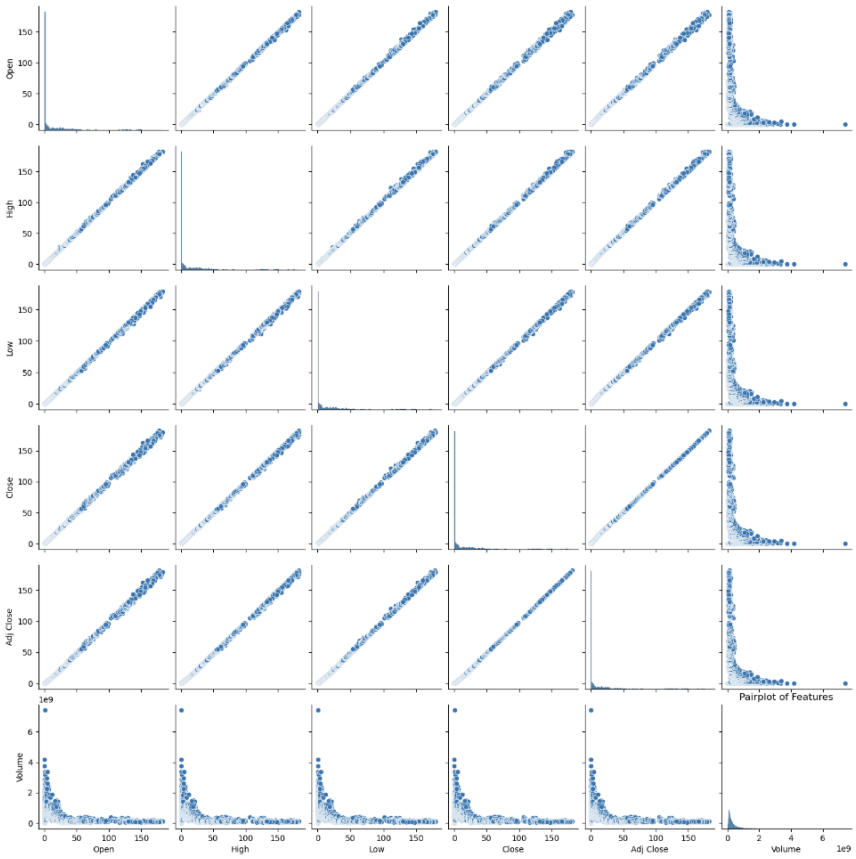

Fig. 2 Pairplot of features

3.3. Vector Autoregression model

I renamed the new dataset after dropping the "Volume" column. After I have labelled the 9 datasets according to the different time ranges, I bring each dataset into the VAR model and label them clearly according to the corresponding number. The forecast function in python was used to forecast each of them and output the 9 corresponding results.

4. Result

After I obtained the predictions corresponding to each of the 9 data sets, I derived the Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R2), (Table 1) for analyses and comparisons. By visualising the comparison of the changes in each test set and predicted values over a one-month period, I used nine discount plots. Then I used box plots to represent the predicted and actual values (Fig. 3). After these steps, I found that the first training set using a one-year time horizon predicted the closest results to the actual data. Of course, I used the opening prices for these comparisons. I then derived and visualised the weights of each of the 5 variables for each set of data when asking for the opening price, high price, low price, closing price and Adj close price respectively. I have shown the weights of the first set in the form of a bar chart and also this is the set of data where the predicted values are closest to the actual values (Fig. 4). This helped me to understand the intrinsic relationship of these five variables.

Table 1. The predictions to each of the 9 data sets

Group Number	MSE	RMSE	MAE	R^2	MAPE
1	77.48498	8.80256	7.85099	-0.86858	5.55938
2	94.42959	9.71749	8.71846	-1.27720	6.05333
3	106.96970	10.34262	9.07017	-1.57961	6.26432
4	116.61423	10.79881	9.36417	-1.81219	6.45163
5	118.76833	10.89809	9.42862	-1.86414	6.49320
6	119.24255	10.91982	9.44251	-1.87558	6.50215
7	119.44057	10.92889	9.44829	-1.88035	6.50587
8	119.53138	10.93304	9.45092	-1.88254	6.50757
9	119.59480	10.93594	9.45278	-1.88407	6.50877

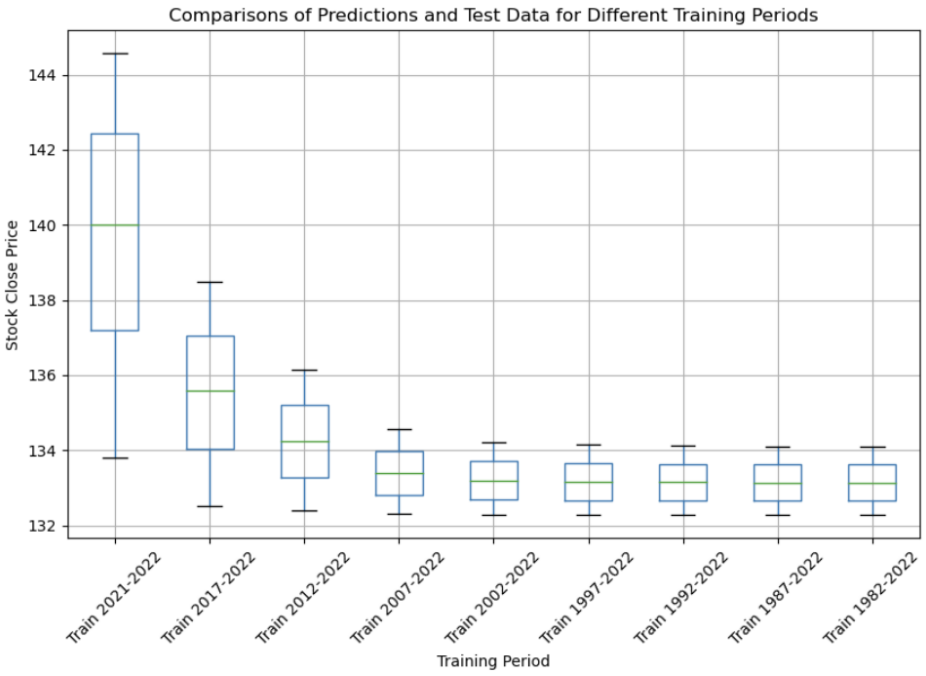

Fig. 3 Box plots of comparisons

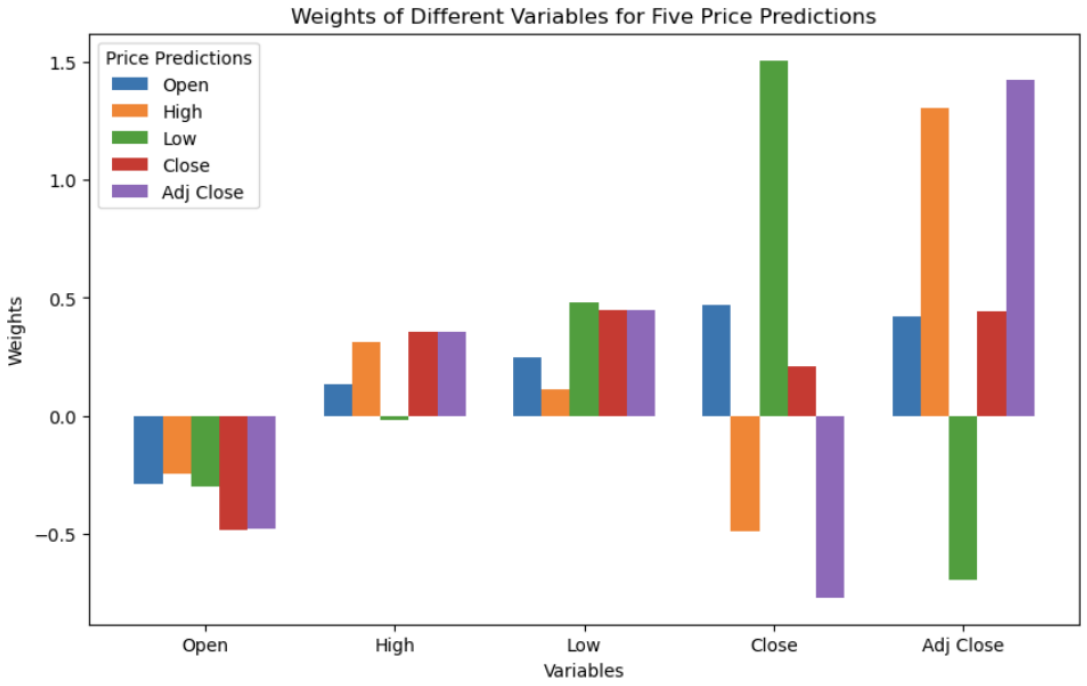

Fig. 4 Weighting of the first (best) set of data.

5. Conclusion

In this paper, knowledge about machine learning is utilised to predict the Apple stock price. Vector autoregressive modelling is mainly employed to explore the relationship between the accuracy of the prediction and the time span of the training set using datasets of different time horizons. It can be seen that as the time span increases, the uncertainty of the uncertainties and external influences increases, so the predictions become more and more inaccurate, but in the longer time range, the difference in the predicted values is not very large. Therefore, in the process of stock prediction, we can try to take recent data to train and predict, rather than the more data the more accurate the prediction. On the basis of this article, one can also add sentiment models to consider more impressions of multiple internal and external factors to predict more accurate values. At the same time stock investing requires careful and informed decision making. Investors construct diversified portfolios to diversify risk after defining their investment objectives, risk tolerance and timeframe, conduct adequate research and analysis to understand the company's fundamentals and market sentiment, etc. Equity investment involves risks and there is no guarantee of success, so careful decision-making is required.

References

[1] JI Hongyun, SUN Yaxuan. What factors affect stock prices:a literature review[J]. Productivity Research,2013(10):193-196.

[2] Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. Artificial Intelligence Review, 53(4), 3007-3057.

[3] Pavithya, M. B. D., Perera, G. S. D., Munasinghe, S. L., & Karunarathna, S. N. (2021, August). Quantitative analysis and sentiment analysis for stock price forecast: the case of Colombo Stock Exchange. In 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS) (pp. 512-517). IEEE.

[4] Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. International Journal of Crowd Science, 5(1), 55-72.

[5] Lütkepohl, H. (2013). Vector autoregressive models. Handbook of research methods and applications in empirical macroeconomics, 30. https://www.geeksforgeeks.org/data-visualization-with-pairplot-seaborn-and-pandas/

[6] Chatfield, C. (1986). Exploratory data analysis. European journal of operational research, 23(1), 5-13.

[7] Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847-2849.

[8] Bisong, E., & Bisong, E. (2019). Matplotlib and seaborn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, 151-165.

[9] Yoffie, D. B., & Rossano, P. (2012). Apple Inc. in 2012. Harvard Business School.

[10] https://www.apple.com/