The Use of Data Mining in Stock Market Predictions

Thanh Nguyen-Duong
Bellevue University
Bellevue, NE 68005, USA
tnguyenduong@my365.bellevue.edu

Courtney Sehn
Bellevue University
Bellevue, NE 68005, USA
csehn@my365.bellevue.edu

Rojan Khatri
Bellevue University
Bellevue, NE 68005, USA
rkhatri@my365.bellevue.edu

Girija Singh
Bellevue University
Bellevue, NE 68005, USA
gsingh@my365.bellevue.edu

Tai Ngo
Bellevue University
Bellevue, NE 68005, USA
tngo@my365.bellevue.edu

Abstract

The stock market refers to the collection of markets and exchanges where there is a total aggregation of buyers and sellers of stocks, in the form of shares, which represent fractions of ownership claims on businesses. Stocks are great investment for both long term and short term. However, due to the volatility in the stock market, many have experienced high losses which have led to tragedies of life in countless people. With the emergence of Machine Learning models and other Data Mining techniques, we can precisely predict stocks course by combining the stock historical data performance, stock news releases, and Machine Learning algorithms. These techniques, if correctly implemented, can aid investors in their investment process and provide many benefits to their long-term growth funds. This paper evaluates the current state of process in how data mining techniques can impact stock market performance prediction and whether these Machine Learning algorithms are performing effectively and efficiently.

Author Keywords

Stock market; market prediction; machine learning; predictive modeling; data science; big data

ACM Classification Keywords
I Computing Methodologies, I.2 ARTIFICIAL INTELLIGENCE, I.2.6 Learning

Introduction

"Buy low, sell high"- a common saying regarding the stock market and how profit is made buying and selling stocks. In reality, it is easier said than done. There are two types of analysis that every investor should weigh in before investing in certain stock. First, fundamental analysis where the basic value of stock is measured in relation to economic and financial factors like performance of the industry, effectiveness of the management, political climate, etc. On the other hand, the technical analysis is forecasting the direction of prices of stock by studying analysis of historical market data such as price and volume. These factors make the stock market highly volatile and difficult to predict.

In recent years, the popularity and advancement in machine learning has expanded its scope to predictive systems in stock market. Traders are evolving from traditional methodology, to a state where sophisticated machine learning techniques are applied to make an investment decision. Machine learning has been proven to ease the entire process by analyzing big portions of data and identifying relevant patterns which help traders make their decisions based on predicted stock prices and market trend.

Does Data Science Influence Stock Market Predictions?

Yes, Data Science absolutely can influence the market performance prediction. There are many factors involved in predicting market – physical factors and physiological, rational and irrational behavior, and global factors but the core of stock value generally is driven by fundamental math of assets minus liabilities equals equity [1]. Therefore, data sets such as macroeconomic, fundamental and price data all can be potentially be utilized in building algorithms to predict the Market Performance as best as possible [5].

This paper will focus on assessing feasibility of using various sources and tools in predicting stock market. It will focus on following five main aspects:

1. Feasibility of using machine learning algorithms in predicting stock market.
2. Feasibility of using data gathered from social media such as Facebook and twitter in predicting stock market.
3. Feasibility of using daily up-to-date financial information and available news in predicting stock market.
4. Feasibility of using historical data in predicting stock market.
5. Feasibility of the current data science technologies in stock market predictions by combining all the indicators gathered from above aspects along with a recommendation for potential investors.

It should be noted that there are most likely many more methodologies currently being researched in the domain of stock market prediction. However, these five concentrations provide a concise overview of the current state or research in the field and assist in determining whether one would choose to make use of existing data science tools in their market investment approach.

Results – What is the Current Emphasis in Research Studies?

The process involves using a few different machine learning models. After testing different models and methods, there is no clear reliable model that can predict the stock price trajectory with an accuracy above 70%. The confidence interval is found to be between 50% - 70%. Below is what happens with each tested model:

1. Stocker Module: This Python Library module tested on two stocks: GOOGL and FB. For both stocks, this module provides worse returns than the buy and hold strategy be [15].
2. Linear Regression: this method is used to predict the trajectory of stock price. It is on track to grasp the general trend of FB stock but fails to capture the moments that investors can make big money during the spike [20].
3. K-Nearest Neighbors (KNN): This model shows an overfitting result, which is the worst result out of three models. Its principle is to calculate the distances from one point to another, but it is unable to figure out where the new point or trend is going to be [15].

Out of three models, the Stocker Module gives the best prediction while the K-Nearest Neighbors provides the worst result. Machine learning models can be developed further to achieve better accuracy [17]. This, however, has not stopped researchers from attempting to accurately predict stock price trajectory.

What Data is Being Used in Price Prediction?

Prior to the development of the internet, data for predicting the stock market was quantitative data [16]. This type of data includes typical time series forecasting analysis. The main objective of time series forecasting is to carefully study past observations to develop an appropriate model. This model is then used to generate future forecasts. Time series forecasting can be described as the act of predicting the future by understanding the past [8]. Specifically, financial time series forecasting uses important daily stock data over an extended period of time including, high price, low price, open price, close price & volume [21]. Volume is defined as the number of shares that are sold, or traded, over a certain period of time. [19].

While it may be easier to gather and analyze quantitative data for forecasting, it may not be reliable enough to predict stock trajectory upon [11]. The activity of a market is conflicting and dependent on multiple factors, making it very dynamic in its nature. Furthermore, most conventional time-series models utilize one variable – the previous day's stock price – when there are many influential factors such as market indexes, technical indicators, economics, politics, investor psychology, and the fundamental financial analysis of companies that can influence forecasting performance [14].

Qualitative data can be used in trend forecasting or predicting the premonition of the immediate future trends of the market based off of real time top news [12]. Top news and social media are primary focuses because these are engaged by the majority of the population and can have the most influence on investors [2][4].The sources for this type of data can vary depending on the research. Examples of qualitative data explored in this research included data from twitter, analysis of google trends, expert investor advice, impacts of news and the time public requires to process news, and risks based off financial reports [7]. Some of the most popular techniques used are natural language processing, sentiment analysis, support vector regression(SVR), and k-nearest neighbors (KNN) classifier algorithm [8].

With the emergence of technological advancements like statistical analysis, machine learning and other analytical data science techniques, forecasting models can be produced that give humans a chance to make more educated trading decisions, giving them a better chance at making a profit verse relying on good luck [10]. Here are some examples:

1. The authors from the College of Engineering in Mumbai India selected 6 stocks under the banking sector for predictions. The analysis then focused on the use of KNN, genetic algorithm, support vector regression SVR and sentiment analysis [6].
2. A research group from the college of Engineering in New Delhi, India used the 25 top articles from the Reddit World News Channel and tried to correlate their impact on the Dow Jones Industrial Average’s (DIJA) close price from June 2008 to July 2016. They used a lexicon-based approach for sentiment analysis & Count Vectorizer language processing [10].
3. Taiwanese researchers used data from both Google trends and Twitter, as well as historical trading details to make predictions on stock closing price. Five stock markets and three companies: Apple Corporation (APPL), Alphabet Corporation (GOOGL), and Microsoft Corporation (MSFT) were forecasted by least squares support vector machines models with the hybrid data collected [15].
4. A researcher from Oracle Data science blog did an exemplary study of performing a time series analysis on a S&P 500 stock index. This example used 21 years (1995-2015) of S&P stock index data at a monthly frequency from Yahoo Finance. The research utilizes the Auto Regressive Integrated Moving Average, or ARIMA model, and performed between an 80-95% confidence interval [3].
5. This study utilizes Taiwan's stock as the data set. This study explores combinations of four different methods including multivariate adaptive regression splines, genetic algorithm, support vector regression & stepwise regression [14].

Overall, the research presented here shows the forecasting of stock markets relied heavily on historical trading data, social media and news trend [13]. It is beneficial to incorporate data that includes historical time series analysis but also consider the probable trends a market can experience utilizing real-time top news from the internet.

Recommendations

There are many reasons the machine learning models did not perform up to our expectations. Here are a few things that we believe can improve the machine learning models in the future:

1. Data: With more financial data, the models should yield lower percent errors, and it may drop to the point that seems acceptable [9].

2. Short-term and long-term investments: Short-term investments will imply a buy or sell signal depending on the news trend and other financial events.
3. Parameters: stock is dictated by many factors ranging from unpredictable factors such as politics and economy to a more predictable factors such as earning report and company management. A good model should account for all known and unknown factors. Otherwise, if one factor is left out, it can easily skew the result of the entire model [18].

Conclusion

In conclusion, all the research work lean towards the importance of this approach in understanding the market feasibility. Various research models discussed here showed that in addition to historical data, many other factors were used in forecasting the stock market. With how closely related between attributes such as political news and stock patterns, outcomes of stocks are getting more accurate with the help of machine learning and data mining. However, investors still need to be in a driving seat and control their own actions regarding their investments. Machine learning is not yet perfect, and it was determined that this method should not be used as the decision making of an investor's investment strategy. For now, it is recommended to use machine learning algorithms and its prediction capability only as a guidance since it can perform lots of tasks that investors normally do daily basis such as pattern tracking.

Acknowledgements

We would like to acknowledge all the stock investors for giving us the inspiration to research on the current topic. Our professor, Shankar Parajulee, for broaden our knowledge in the realm of Data Science. Lastly, our group members for contributing to this project.

References

[1] Adhikari, R. K., & Agrawal, R. K. (2013). Introductory study on time series modeling and forecasting. Germany: LAP Lambert Academic Publishing. doi: 10.13140/2.1.2771.8084

[2] Atkins, A., Niranjan, M., & Gerding, E. (n.d.). Financial news predicts stock market volatility better than close price. The Journal of Finance and Data Science, 4(2), 120–137. doi: 10.1016/j.jfds.2018.02.002

[3] Eulogio, R. (2018, January 30). Retrieved from https://blogs.oracle.com/datascience/performing-a-time-series-analysis-on-the-sandp-500-stock-index

[4] Garcia-Lopez, F., Batyrshin, I., & Gelbukh, A.(2018). Analysis of relationships between tweets and stock market trends. Journal of Intelligent & FuzzySystems, 34, 3337-3347. doi:10.3233/JIFS169515

[5] Ghosh, P. (2019). How Is Machine Learning Used in the Stock Market? PC Quest, 32(2), 28.

[6] Gupta, A., Bhatia, P., Dave, K., & Jain, P. (2019). Stock Market Prediction Using Data Mining Techniques. SSR Electronic Journal. doi: 10.2139/ssrn.3370789

[7] Kavšek, B. (2017). Using Words from Daily News Headlines to Predict the Movement of Stock Market Indices. Managing Global Transitions: International Research Journal, 15(2), 109–121.

[8] Kumar, L., Pandey, A., Srivastava, S., & Darbari, M.(2011). Record Title: A Hybrid Machine Learning System for Stock Market Forecasting. Journal of International Technology & Information Management, 20(1), 39-48.

[9] Lee, T. K., Cho, J. H., Kwon, D. S., & Sohn, S. Y. (2019). Global stock market investment strategies based on financial network indicators using machine learning techniques. Expert Systems with Applications, 117, 228–242.doi: 10.1016/j.eswa.2018.09.005

[10] Leekha, A., Wadhwa, A., Jain, N., & Wadhwa, M.(2018). Understanding the Impact of News on Stock Market Trends Using Natural Language Processing and Machine. International Journal of Knowledge Based Computer Systems, 6(2), 23-30.

[11] Li, X., Xie, H., Wang, R., Cai, Y., Cao, J., Wang, F.,. . . Deng, X. (2016). Empirical Analysis: Stock Market Prediction via Extreme Learning Machine, 27(1), 67-78.

[12] Li, Xiaodong; Xie, Haoran; Wang, Ran; Cai, Yi; Cao, Jingjing; Wang, Feng; Min, Huaqing; Deng, Xiaotie. Neural Computing & Applications. Jan2016, Vol. 27 Issue 1, p67-78. 12p. DOI: 10.1007/s00521-014-1550-z.,Database: Military & Government Collection

[13] Liang, Q., Rong, W., Zhang, J., Liu, J., & Xiong, Z. (2017). Restricted Boltzmann machine based stock market trend prediction. 2017 International Joint Conference on Neural Networks (IJCNN). doi: 10.1109/ijcnn.2017.7966014

[14] Ming-Chi Tsai, C.-H. C. (2018, December 31). Forecasting leading industry stock prices based on a hybrid time-series forecast model. Retrieved October 29, 2019, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209922

[15] Pai, P.-F., Hong, L.-C., & Lin, K.-P. (2018). Using Internet Search Trends and Historical Trading Data for Predicting Stock Markets by the Least Squares Support Vector Regression Model. Computational Intelligence and Neuroscience, 2018, 1–15. doi: 10.1155/2018/6305246

[16] Ratnaparkhi, S. (2019, January 14). Machine Learning -Predict Stock Prices using Regression. Retrieved from https://www.quantinsti.com/blog/machine-learningtrading-predict-stock-prices-regression

[17] Singh, A. (2019, March 11). Predicting the Stock Market Using Machine Learning and Deep Learning. Retrieved from https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learningtechniques-python/

[18] Song, Yuan. (2018). Stock Trend Prediction: Based on Machine Learning Methods. UCLA: Statistics 0891. Retrieved from: http://www.escholarship.org/uc/item/0cp1x8th

[19] Staff, M. F. (2016, June 30). What is Volume in Stock Trading? Retrieved October 29, 2019, from https://www.fool.com/knowledge-center/what-is-volume-in-stock-trading.aspx.

[20] Xu J. (2019). How to use machine learning to possibly become a millionaire: Predicting the Stock Market. Retrieved from https://towardsdatascience.com/how-to-use-machine-learning-to-possibly-become-a-millionaire-predicting-the-stock-market-33861916e9c5

[21] Zhang, K., Zhong, G., Dong, J., Wang, S., & Wang, Y. (2019). Stock Market Prediction Based on Generative Adversarial Network. Procedia Computer Science, 147, 400–406.doi: 10.1016/j.procs.2019.01.256