Application of Machine Learning Algorithms in the Stock Market Analysis

Chunjiang Li *

Colgate University, New York, 13346, United States

* Corresponding Author Email: cli@colgate.edu

Abstract. With the development of deep learning and machine learning, more new methods have been produced in the economic and financial fields. When talking about machines, one thing that comes to people's minds is what they can do with machines to solve problems that need machines. The people in the stock market always want to find ways to forecast the stock trend, the pattern of stock, and the stock value. Before the development of machine learning algorithms, stock market predictions could be made in limited ways, and those methods usually did not produce accurate predictions. However, machine learning algorithms changed the phenomenon and offered people novel ways to analyze the stock market. This paper will discuss three research in which authors have implemented machine learning algorithms into stock market analyses. From analyzing the research, this paper tries to investigate the extent of applying machine learning algorithms in the stock market and how the algorithms have helped investors make improvements in stock market analysis.

Keywords: Machine learning; stock market; analysis.

1. Introduction

When machine learning algorithms started to develop, one of the questions that puzzled stock market traders was whether these fancy algorithms could be used for analyzing the stock market trend. However, the nature of the stock market is unpredictable, which makes the stock market analysis difficult for machines as there are so many factors that machines cannot analyze or yield reliable data collections for people to analyze. In other words, the application of machine learning algorithms might need to be improved in the stock market analysis. This paper is focused on analyzing how far people can apply machine learning algorithms in stock market analysis to help them better understand the trend of the stock market and the potential difficulties involved in the process. This paper provides a literature review of how different scholars have been applying machine learning algorithms in the stock market and analyzes the extent to which machine learning algorithms could be applied.

People usually choose nonlinear models when applying machine learning algorithms to stock market analysis. The reason for using nonlinear models is that the complexity and multi-factorial dependencies of stock price predictions make nonlinear models work better than linear models. This paper will focus on analyzing three studies done by different scholars who try to apply machine learning algorithms to stock market analysis and makes an analysis, from their research, on the extent of the application of machine learning algorithms in stock market analysis.

2. Relevant methods and analysis

2.1. Abnormal return in an inefficient stock market

2.1.1 Background

The research investigates whether Fama's hypothesis about efficient market holds; provides a novel trading strategy made with a linear and a nonlinear model. The study uses support vector machine(SVM) and logistic regression to perform stock trading [1]. It is important to clarify the terms the authors have mentioned. The weak efficient market is a type of market mentioned in Fama's Efficient Market Hypothesis. What Fama means by efficiency is that the market can produce rapid absorption of information. Information is news that can affect prices in the stock market, and the news

is unpredictable. In other words, an efficient market means a market that is responsive to the changes in the world that can affect the market prices and absorb information quicker. From that, it is reasonable to deduce that market prices in the efficient market can disclose information that affects the market prices as the efficient market is responsive to the information that affects the prices. The weak efficient market in Fama's EMH is the market in which a security's price reflects historical data about a security's price, including stock price and trading volume [2]. That is to say, unlike the most efficient market, where the prices can disclose any information, past stock market price changes do not affect the current stock market prices in the weak efficient market.

Logistic regression in this study refers to the statistical technique describing the relationship between independent and binary dependent variables. As shown in Table 1. It is noteworthy that logistic regression is a linear model, and the authors in the research combine the logistic regression(linear model) with the SVM(nonlinear model) for analyzing the abnormal return. The Support Vector Machine(SVM) algorithm was proposed by Vapnik and Lerner to solve the classification problem [3]. The rolling window approach refers to analyzing the stability of parameters in the model. The rolling window technique can accurately predict the parameters of a given model developed by machine learning algorithms [4-6].

Table 1. Comparison table

Ticker	Logistic	SVM	Ticker	Logistic	SVM	Ticker	Logistic	SVM
BID	59.49	92.55	POW	56.85	93.36	MWG	53.84	91.51
BVH	58.4	92.91	REE	62.21	92.53	NVL	58.29	93.72
CTG	61.93	93.53	SBT	62.2	93.6	PDR	59.98	92.76
FPT	58.67	90.64	SSI	58.9	91.75	PLX	56.3	91.41
GAS	57.55	92.9	STB	63.32	92.19	PNJ	57.86	91.64
HDB	57.61	93.71	TCB	55.53	95.73	VIC	59.87	91.57
HPG	58.93	91.22	TCH	54.88	91.97	VIC	58.57	86.66
KDH	61.89	93.36	TPB	58.52	96.94	VNM	61.83	91.34
MBB	62.15	93.67	VCB	58.6	93.25	VPB	57.5	88.03
MSN	62.48	93.05	VHM	57.84	94.85	VRE	55.62	91.95
Average accuracy			SVM			Logstic		
			92.48			58.93		

2.1.2 Analysis

The research starts with researching data from 30 companies. Then the study observes the data, each observation including variables listed in the table above. Then the study would test the Fama;s hypothesis that the the prices in the inefficient stock market appear randmomly so the past price is not useful in predicting the future price. The authors can determine if Fama's efficient hypothesis holds using the Wald-Wolfowitz test to test the randomness hypothesis on a two-state data series. After using the run test, the authors use logistic regression for all data in the research period to test the influence of factors affecting price movements. After testing Fama's efficient-market hypothesis, the authors use logistic regression and SVM to forecast price movement direction in the stock market.

The study examines the fixed training data of 365 observations to make forecasts by applying the "rolling window" method, supposing that the data's maximum value is 1 [7]. The study used algorithms to indicate the ideal parameters of the observations and make forecasts of the observations. At the end of the research, the study compares the performance of the logistic regression model and SVM [8]. To summarize the research, the authors use the runs-test to test the hypothesis, a linear and a nonlinear model for forecasting the price and compare the results generated from two different models. With the application of linear and deep learning models in forecasting price movements, the authors proposed a novel securities trading method [9].

2.1.3 Results

After using the linear logistic regression model and the SVM model to analyze the data, the authors refute the weak efficient market hypothesis. Furthermore, the authors found that technical analysis in research can obtain returns that are different from normal returns on an inefficient stock market. In the study, the logistic regression and SVM models predict increases or decreases in stocks. The shares are then traded following the forecasts made by the logistic regression and SVM models. After analyzing all the stock data, the authors concluded that the price action expectation indicator was the most helpful. In addition, the logistic regression model and SVM filter the variables listed in the introduction and help researchers conclude which variables will affect stock trading [10].

After using the linear logistic regression model and the SVM model to analyze the data, the authors refute the weak efficient market hypothesis. Furthermore, the authors found that technical analysis in research can obtain returns that are different from normal returns on an inefficient stock market. In the study, the logistic regression and SVM models predict increases or decreases in stocks. The shares are then traded following the forecasts made by the logistic regression and SVM models. After analyzing all the stock data, the authors concluded that the price action expectation indicator was the most helpful. In addition, the logistic regression model and SVM filter the variables listed in the introduction and help researchers conclude which variables will affect stock trading.

2.2. An intelligent stock trading system

2.2.1 Background

The research aims to propose an intelligent stock trading system, and the authors use data to test the system's performance. The fundamental analysis estimates a company's intrinsic value by analyzing internal and external financial factors(macroeconomic, market conditions, and competitors). Blue-chip stocks are stocks that are the most valuable in the stock market. Blue-chip stocks are the shares of companies that are very large, well-recognized, and low-risk for investors. They pay dividends relatively more frequently to the investors due to their sound financial performance. The proposed system will use the Stacked Long Short-Term Memory Network model to detect the trading signal. This model can process short-term and long-term memories, enabling the model to apply in the data analysis of sequences. After equalling weighing the shares through the proposed system, the system would form a portfolio for an investor and an index to manage the risk called the risk control index. A base index in the risk control index can illustrate the performance of an investment. The risk control index is a method to manage the risk in a portfolio formed after using the proposed stock market trading system to create the portfolio.

2.2.2 Analysis

The authors select blue-chip stocks for the proposed system for fundamental analysis. The authors would use the technical indicators and a deep neural network to predict the trading signals for the selected ten blue-chip stocks. The chart below shows what the proposed system tries to achieve. As the chart shows, the proposed trading system gathers data on the ten best blue chip stocks before doing an analysis. Then, the system would analyze each stock separately until all ten blue chip stocks have been analyzed. In the analysis process, the proposed system calculates the technical indicators, then uses the feature selection to reduce the data dimensions, and uses the stacked LSTM to generate a trade signal. After the proposed system finished the above steps, the portfolios for each ten blue

chip stocks would be constructed. Then proposed a trading system that calculated the risk control index and rebalanced the portfolio to manage the risk involved in each portfolio. At last, the proposed trading system would evaluate the portfolios. Investors can better understand each blue chip stock with the proposed trading system's assistance. As shown in Figure 1.

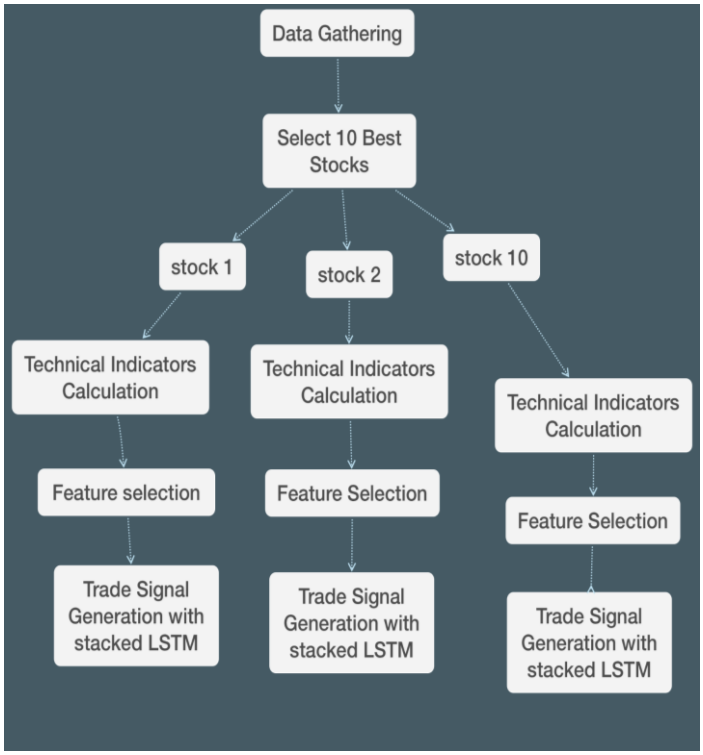

Fig 1. Stock trading system flow chart

2.2.3 Results

The evaluation indicates that the smart stock trading system can create distinguished portfolios for investors with different risk appetites. In developing the system, the authors performed both fundamental and technical analyses [11]. The main limitation of the research is that the proposed system should have dealt with the short sales in Tehran Stock Exchange.Moreover, the research only used one weighting method, and other weighting methods can be implemented in future studies. The study also adds daily reviews to calculate the risk control index, which led to the daily portfolio rebalancing. The movement could increase the transaction costs, which may influence the rebalancing of the portfolio .

2.3. Machine learning system for the stock market

2.3.1 Background

Stock price data has very significant characteristics such as nonlinear, non-stationary, high noise, strong time-varying, etc., which has certain challenges to the prediction of stock prices. Throughout the existing research, early research mainly used technical analysis. Among them, technical analysis is to determine the trend of stock prices by combining common market indicators such as trading volume and transaction price of stocks. Researchers often use time series models to predict stock prices. The commonly used models include the difference integration moving average autoregressive (ARIMA) model, the generalized autoregressive conditional heterosquare (GARCH) model for the volatility aggregation effect of financial data, and the vector autoregressive (VAR) model, a variant of ARIMA model. In addition to the traditional econometric model, grey model, BP neural network model and fuzzy theory also have many applications in stock price prediction. This paper proposes a marking method called N-cycle min-max (NPMM), which can be effectively used in stock market analysis. The NPMM tag model uses XGBoost to develop the transaction system and realize

transaction automation. XGBoost is a nonlinear machine learning model and a library written in C++language to improve the performance of gradient lifting (a regression and classification technology).

2.3.2 Analysis

The above graph shows three phases of the research, building learning data, model training, and simulation. In phase one, the author tries to build learning data for the system. To do that, the authors examine the technical indicators. The technical indicator is a mathematical calculator based on historical price information that aims to predict stock market trends. (Murphy, 1999) The formula for the N-Period Min-Max labeling proposed by the author is: *Lt*+1 = 0, *if Ct == Nmax, Lt*+1=1 *if Ct == Nmin*. The formula shows the feature of the labeling that the labeling only offers labels for the minimum and maximum periods in the window period. The research uses the window-based labeling proposed by Sezer and Ozbayoglu (2018), so when the window period refers to the period assessed by the labeling method. The feature of the NPMM labeling is that it is insensitive to small changes as it only assesses the minimum and maximum window periods. Then the authors move on to the next research phase, model training with XGBoost. As the authors move from phase 1 and phase 2 to phase 3, the authors make label predictions to automate the stock market. In phase 3 of the simulation, the authors apply the learned XGBoost model to generate stock market signals and automate stock market operations. After automating stock market operations, the authors do the performance evaluation. As shown in Figure 2.

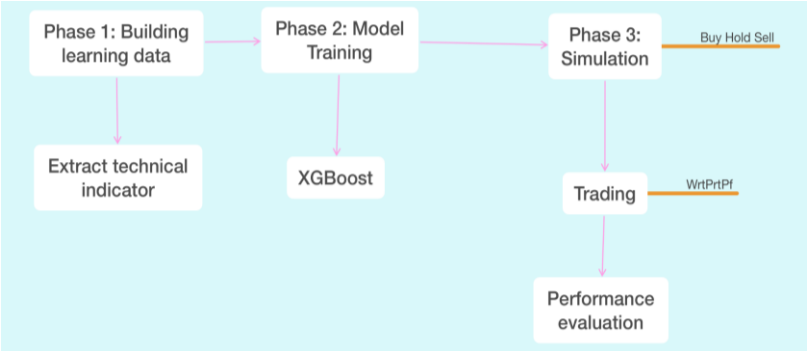

Fig 2. Stock market system flow chart

The performance evaluation aims to evaluate the trading system developed under the NPMM labeling method proposed by the authors. The buying and selling system developed using the NPMM labeling method can work well if the evaluation is good. The author evaluates the trading system using learning data figures, win ratio (Wr), payout ratio (Pr), and profit factor (Pf), as the graphs above show [4-6].

2.3.3 Results

After experimenting with the NPMM labeling method and a trading system was developed based on the method. Empirical analysis of 92 stocks listed on the NSDAQ (to which the proposed system is implemented) shows that the learning data measure decreases as the N value increases in the NPMM labeling, but overall trading performance improves. Overall, the system works fine.

3. Discussions and Suggestions

From analyzing the performance of machine learning algorithms in the above research, machine learning algorithms are helpful in stock market analysis. We also know how to apply them in the stock market analysis from the above examples. However, the limitations of machine learning algorithms are that it is hard to get accurate results when analyzing large amounts of data. The first research using an approach based on machine learning in stock trading to investigate if it is possible to search for abnormal returns in an inefficient stock market shows differences between linear and nonlinear models. The nonlinear models outperform the traditional statistical linear models as the

stock prices are nonlinear and often appear randomly. Therefore, applying linear models in the stock market generates a lower accuracy than applying nonlinear deep learning models. However, nonlinear models are not helpful when analyzing large data sets.

All research listed above applied nonlinear models to analyzing small stock samples to make stock market price predictions or automate the stock market. The reason for this is that as the dimensions of data increase, the data spreads all over places, and it becomes hard to analyze the scattered data. In the second research, the authors use feature selection (a method to reduce the dimensions in unsupervised machine learning) for reducing the dimensions to analyze each blue-chip stock. Even though all research shows the efficiency of stock market analysis using machine learning algorithms, they all have limitations in that they only analyze portions of the data in the stock market. The results produced under these studies should be further analyzed in larger data samples. The first research shows what people can do with the logistic regression linear and SVM models to generate more efficient trading strategies. As shown in Table 2. It also shows the limits of using a linear model in analyzing nonlinear stock markets. However, by combing the logistic regression with SVM, the authors generated excellent imitated trading results and found the factors that affect the stock market price movement. The first research indicates that although linear models might not be suitable for analyzing nonlinear data, people can combine them with nonlinear models in analyzing nonlinear data. This is because linear models are more helpful in determining patterns or relationships between factors and results. The authors used the linear model in the first study to find the factors affecting stock market price movements. The results were reliable and helpful. The table below shows what logistic regression model has contributed to the study.

Table 2. Comparison and analysis table

Deviance residuals				
Min	IQ	Median	3Q	Max
-2.2376	-1.0705	-0.9763	1.2765	1.9416
	Estimate	Std. error	Z value	Pr(>lzl)
(intercept)	-1.54e+--	8.77e-02	-17.589	<2e-16…
close	3.57e-03	4.41e-03	0.81	0.4179
HL	5.53e-02	9.29e-03	5.951	2.67e-09
LO	6.07e-02	1.11e-02	5.47	2.67e-09
variation	1.00e-02	6.00e-03	1.67	0.0939
vnicbb	-2.51e-02	2.18e-03	-11.543	<2e-16***
vnipc	2.00e-01	1.63e-02	12.295	<2e-16***
insec	7.45e-05	5.18e-06	14.389	<2e-16***
ma7	4.88e-03	7.81e-03	0.625	0.5321
ma14	-1.34e-03	9.15e-03	-0.146	0.8838
ma21	-8.35e-03	5.35e-03	-1.562	0.1184
sd7	3.98e-02	5.54e-03	7.184	6.76e-13***

The authors used the logistic regression to investigate how each variable affect the stock market price movement, and the results have been shown in the above table. The second and third research show how people can use machine learning algorithms to build trading systems or the labeling necessary for a trading system. These studies show how effective machine learning algorithms can be in developing trading systems that can help investors to make investment decisions. With the smart trading system proposed by the second research, investors can have different portfolios for the different blue-chip stocks that are worthwhile for investors to invest in. The new labeling method proposed by the third research can increase the accuracy of stock market trends in the stock market analysis. The trading system developed under it can also help investors make better investment decisions.

4. Conclusion

Application of machine learning algorithms to stock market analysis produces many unprecedented ways of making stock market predictions, trading systems, and testing the hypothesis of efficient markets. However, more studies need to focus on improving the system's efficiency built with machine learning algorithms and overcoming the problem of reducing dimensions when applying machine learning algorithms to the stock market analysis. In the first research, the authors test the hypothesis of a weak form efficient market with runs-test and provide trading strategies for short-term investors. The second research proposed a trading stock system built with machine learning algorithms. The system produced different portfolios for investors with different risk appetites. The third research proposed a new labeling method, a system built under that method, which performs well in the stock market analysis of the 92 samples of stock data in the stock market.

References

[1] Eskandari H, Sadegheih A, Zare H K, et al. Developing a smart stock trading system equipped with a novel risk control mechanism for investors with different risk appetites[J]. Expert Systems with Applications, 2022, 210: 118614.

[2] Fama E F. Efficient capital markets: A review of theory and empirical work[J]. The journal of Finance, 1970, 25(2): 383-417.

[3] Han Y, Kim J, Enke D. A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost[J]. Expert Systems with Applications, 2023, 211: 118581.

[4] Harris M. Profitability and Systematic Trading: A Quantitative Approach to Profitability, Risk, and Money Management[M]. John Wiley & Sons, 2008.

[5] Kim Y, Enke D. Developing a rule change trading system for the futures market using rough set analysis[J]. Expert Systems with Applications, 2016, 59: 165-173.

[6] Khoa B T, Huynh T T. Is it possible to earn abnormal return in an inefficient market? An approach based on machine learning in stock trading[J]. Computational Intelligence and Neuroscience, 2021, 2021.

[7] Murphy J J. Technical analysis of the financial markets: A comprehensive guide to trading methods and applications[M]. Penguin, 1999.

[8] Schöneburg E. Stock price prediction using neural networks: A project report[J]. Neurocomputing, 1990, 2(1): 17-27.

[9] Sezer O B, Ozbayoglu A M. Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach[J]. Applied Soft Computing, 2018, 70: 525-538.

[10] Stridsman T. Tradings Systems That Work: Building and Evaluating Effective Trading Systems[M]. McGraw Hill Professional, 2001.

[11] Vapnik V N, Lerner A Y. Recognition of patterns with help of generalized portraits[J]. Avtomat. i Telemekh, 1963, 24(6): 774-780.