INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact Factor: 6.078

(Volume 7, Issue 3 - V7I3-1911)

Available online at: https://www.ijariit.com

Review on the applications of machine learning models for stock market predictions: A literature survey

Sanjana Hemaraju
sanjanahemaraju99@gmail.com
RV College of Engineering, Bengaluru, Karnataka

Monish Singhal
monishsinghal16@gmail.com
RV College of Engineering, Bengaluru, Karnataka

Dr N. S. Narahari
nsnarahari@gmail.com
RV College of Engineering, Bengaluru, Karnataka

ABSTRACT

Financial stocks values are non-linear, volatile, and chaotic, making them one of the most challenging financial time series to predict. The incentive of financial gain has led many researchers and academia to devise methods to predict the stock market, despite copious uncertainty. Because of their ability to recognize complex patterns in several applications, machine learning models are extensively researched among the most recent methods. In this paper, Support Vector Machine, Artificial Neural Networks and Case-based Reasoning for stock market prediction is surveyed. This paper also reviews sentiment analysis to highlight the behavioral trends of the stock market and its investors with the advent of technology.A generalised modelling methodology for applying machine learning techniques to the stock market is proposed in this paper .

Keywords: *Artificial Neural Network, Case-Based Reasoning, Sentiment Analysis, Stock Market, Support Vector Machine*

1. INTRODUCTION

Significant research is devoted to stock market prediction, one of the most strenuous time series problems owing to the volatile, complex, dynamic and non-linear nature of the market. With the advent of technology and increased availability of information, investing in the stock market has become popular amongst the general public. The potential for financial gain drives the need for a methodical approach to predict the market and understand its trends and movements. Despite the Efficient Market Hypothesis (EMH) [1] and Random Walk Hypothesis [2], numerous techniques have been applied for stock market prediction over many years, stock markets follow chaotic patterns and are unreliable therefore demanding further research on the development of models and algorithms.

Several indices have been formed to assess the relative value of the stocks traded within a market. Fundamental analysis and technical analysis are the two different stock market forecasting approaches widely considered. J. J. Murphy [3] defined technical analysis as "the study of market action, primarily through the use of charts, for the purpose of forecasting future price trend" and believes price, volume and open interest are the three predominant sources of information available. They suggest three ideologies upon which technical analysis is based, namely "market action discounts everything", "price moves in trends", and "history repeats itself" [3]. On the other hand, Fundamental Analysis is a way of determining the value of a stock by considering economic and financial elements including macroeconomic and microeconomic factors like the economics of supply and demand, and the company's performance [4]. Some experts and traders apply both approaches, specifically fundamental analysis to determine what stock to invest in and technical analysis to determine the timing of the investment.

Stock market prediction is a combination of several fields of study including statistics, operations research, computer science, economics, and finance. Traditionally, time series forecasting was used to predict the stock market that included classic regression methods that involved smoothing, moving average and autoregressive techniques [5]. More recently, multiple computational methods such as Bayesian Networks, Logistic Regression, Multiple Linear Regression (MLR), Support Vector Machines (SVM), Artificial Neural Networks (ANN), Genetic Algorithms (GA), and Case-Based Reasoning (CR) have been applied for forecasting the trends and movement of the stock market.

Despite numerous efforts involving data mining and machine learning algorithms, it was observed that the accuracy was not sufficient to rely upon. Investigation into the same has revealed that a huge amount of internet information available, such as Wikipedia usage patterns, news stories from conventional and social media sources, and tweets from Twitter, can have a discernible impact on investor attitudes regarding financial markets. Researchers have therefore initiated the integration of text mining and natural language processing techniques (NLP) to incorporate the effect of societal mood on the stock trends and movement [6]. A combined input of societal inclinations and historical price data has contributed to a significant increase in the accuracy rates of the models used for prediction.

In this paper, the authors aim to survey some of the machine learning techniques, Support Vector Machines (SVM), Artificial Neural Networks (ANNs) and Case-Based Reasoning (CBR) applied and review sentiment analysis and text mining for stock market prediction.

2. MACHINE LEARNING MODELS FOR STOCK MARKET PREDICTIONS:-

2.1 Support Vector Machine (SVM)

Support Vector Machine is a machine learning algorithm used for classification, wherein classification is done through a linear or non-linear separation surface in the input space. For this quadratic optimization problem, the separating function is defined as a linear expression of kernels associated with the support vector [7]. SVMs are not susceptible to overfitting of the model on the training data and are featured by capacity control of the decision function and the use of the kernel functions to ensure the solution's dimensionality. Z. Hu, J. Zhu and K. Tse in their research argue that traditional prediction models are obsolete in the current world situation and suggest a theoretical and empirical methodology to apply Support Vector Machine (SVM) to analyze the interdependence of macro-economic factors (consumer spending, consumer investment, unemployment rate, inflation rate, federal funds rate, Dow Jones industrial average) and company-specific factors (net revenue, net income, price per earnings ratio of a stock, and diluted earnings per share), that affect the stock market movement. The classifier developed is of suitable accuracy of 96.15% and the model works well for input data outside the training set and elucidate improvements for the model by supplementing the algorithm with further company-related variables [8].

SVM has been compared to many algorithms to assess its performance and suitability. Shunrong Shen, et.al., developed a prediction algorithm based on Support Vector Machine (SVM) that uses the temporal link between global stock markets and numerous financial products to anticipate the next-day stock trend. The algorithm indicated a prediction accuracy of 74-78% when applied to different markets and regression models to anticipate the market's underlying increasing or decrease trends. For multiple feature prediction, SVM and Multiple Additive Regression Trees (MART) are applied which suggested accuracy levels as high as 74%. It was noted that SVM is better suited for a larger amount of training data due to the faulty splitting of data by hyper-plane while MART is better suited when less training data is available. The judgement criteria of the model were Root Mean Squared Error (RMSE) and the trading model developed boasts higher profits than existing benchmarks [9]. In another research, authors Wei Huang, et. al., compare and contrast different forecasting models including, Random Walk (RW) model, Linear Discriminant Analysis (LDA), Quadratic Description Analysis (QDA), Eleman Backpropagation Neural Networks (EBNN) to the performance of support vector machines (SVM). In their report, they say they found that SVM has the highest accuracy for forecasting amongst all the forecasting algorithms due to its characteristics of minimisation of structural risk and high generalisation, hence reducing the occurrences of overfitting of the model. It was also observed that incorporating SVM along with other forecasting algorithms improves the performance drastically [10]. By integrating multiple models, the losses and inaccuracies can be balanced by the advantages of another model overall leading to a systematic effect on the prediction accuracy. SVMs also have an advantage over NNs due to their unique and globally optimal solution.

Researchers have attempted to incorporate SVMs with further algorithmic techniques like Genetic Algorithms and Neural Networks to increase the accuracy of the models. In their study, Nti, et. al., present a homogenous ensemble classifier incorporating SVM and Genetic Algorithm (GA) for feature selection and SVM kernel parameter optimization for stock market prediction. The versatile design nature of SVM was narrowed down by optimal selection using GA and applied to 11 years' worth of price dataset from the Ghana Stock Exchange (GSE). The model's performance claims to be a better algorithm than decision tree, random forest and neural network methods in the prediction of 10-day-ahead stock price movement. Since the strenuous parameter optimization process was removed in the practical GA approach, the model offers a prediction accuracy of 93.7% [11]. A neoteric nonlinear combination model was proposed by Fangqiong Luo, et. al., formed on the principles of Support Vector Machine (SVM) regression which was a combination of the conventional linear regression statistical model with neural network (NN) non-linear regression. By integrating both linear and non-linear characteristics of the stock market nature, a forecasting methodology was developed by considering the Shanghai Stock Exchange index. The forecasting model utilised four regression algorithms to identify the linear characteristics and four neural network models are used to identify nonlinear characteristics of the stock market and the final output is presented by the support vector regression model by combining two separate prediction individuals by levying time weights. Judgement of performance is conducted through calculations of RMSE, MAPE, Trend Accuracy and Non-linear regression multiple correlation coefficient [12]. Such experimental results suggest an increase in accuracy and significant potential in the area of time series forecasting.

2.2 Artificial Neural Networks

J. J. Hopfield describes Artificial Neural Networks (ANNs) as computational networks emulating the networks and devices of neurobiology to solve complex problems effortlessly [13]. They are composed of a large number of highly interjoined processing elements connected together through activation functions synonymous with brain synapses [14]. The ability of Neural Networks to identify non-linear tendencies between the independent and dependent variables makes them suitable for modelling dynamic, nonlinear systems such as stock markets. Across literature, ANNs have shown more promise than other modelling techniques for

the prediction of the stock market movement. One of the benefits is the ability to learn associations from the data rather than assuming the relationship's direction.
Given enough data, every relation may be approximated to any level of specificity using NNs, which are known as ubiquitous approximators. It paves way for some tolerance for chaotic and partial data representation.

In the research accurate stock market prediction is the objective brought about by artificial neural networks and feature selection is done through the Boruta method. Regression methodology is used, and the model is evaluated for its performance using the metrics, mean absolute error MAE and Root Mean Square Error (RMSE). Modelled for short-term trading, the paper uses data from the National Stock Exchange (NSE) of India for the algorithm. The ANN prediction model has three layers, with the nodes being technical indicators. Sigmoid functions are used for activation and a threshold value of 0.5 is considered. By considering the stock prices of ICICI bank the model returns an MAE of 15.12 and considering the stock prices of State bank of India, the model returns an MAE of 14.4. The authors define fundamental analysis incorporation, microeconomics, and macroeconomics factors consideration, as the future scope of the project [15].

Although ANNs are suitable for complex and non-linear relationships, such as the workings of the stock market, ANNs do not reveal the degree of influence of the predictor variables on the output variable making it difficult to comprehend the prediction mechanism of the network. ANN also tends to overfit the model to the training dataset making the predictions of price much less accurate. This problem can be overcome by either reducing the complexity of the model with respect to the number of nodes and hidden layers or reducing the number of epochs to a lower value. Training of an NN can be tricky as the network might divulge into the most occurring solution instead of an accurate output. Reducing and transforming irrelevant or duplicate features can speed up the process and produce more generalised results [16].

Often, researchers experiment on hybridisation of modelling techniques in lieu of better performance. Many models have been proposed combining Artificial Neural Networks (ANNs) with fuzzy logic techniques, SVMs, etc. A. Alfa, et, al., in their research explore the relevance of expert systems with Neural Networks in predicting stock price as opposed to a traditional time series method due to non-linear and linear fluctuations of the stock price when multiple features are considered. The authors contrast two foresting systems for stock market prediction: Fuzzy Interference System Expert System (FISES) and Neural Network Expert System (NNES). The FISES is built based on rules and membership functions while NNES is represented by one input layer, one hidden layer and one output layer and trained based on the available dataset. By evaluating Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Average Percentage Error (MAPE) and Relative Absolute Error (RAE) of the two models, it was found that the Neural Network Expert System served the purpose of forecasting with performance twice as better as Fuzzy Interference System Expert System on an average, mainly due to inbuild Rule Base and training and learning [17].

2.3Case-Based Reasoning

Case-based reasoning generates solutions by retrieving and reusing solutions to previously solved situations. Each experience is stored as a case in the case base and is memorised as a pair; problem or circumstance along with the solution or action that goes with it. The experiences does not record how the solution was reached, the retrieval and reuse of retrieved solutions is facilitated by the case base which acts as the memory. CBR derives its functionality from human problem-solving in which new problems are tackled based on the solutions to older and similar problems. Specific experiences are stores and recollected and then re-used when the situation demands, and the methodology can be compared to the rule-based or theory-based problem-solving methods. CBR operates on the fact that similar problems will have similar solutions. CBR is also an example of an instance of lazy learning because there is no trend model to apply to solve new problems. CBR is well suited for circumstances that are not so easily comprehended or where no underlying principles are abundant. A general CBR system consists of four knowledge containers namely vocabulary, similarity knowledge, adaption knowledge, and case base. The system follows a cycle of retrieve, reuse, revise, and retain [18].

Kim, K in their paper proposed a model for stock forecasting using simultaneous optimization method of case-based reasoning (CBR) integrated with Genetic Algorithm (GA). The GA simultaneously selects relevant features sets and optimising feature weights during the CBR system process instead of sequentially, as used in previous works. Their hybrid model showed better performance as compared to other models for stock market prediction. It was found after evaluation that Simultaneous Optimisation with GA CBR exhibited better performance rates than Feature Selection with GA CBR, Feature Weighting using GA CBR or Conventional CBR. The author urges further research on the modelling by deploying GA for relevant instance selection as well [19].

A geometric CBR technique was presented by Chun, Se-Hak& Ko, Young-Woong, wherein a geometric basis was used to choose similar cases that serve as experience for the solution. The shape distance method to extract nearest neighbours is used instead of euclidean distances, where the number of sign changes of the feature for the target case is considered. The CBR models (conventional and geometric) is applied to the prediction of stock market indices (hit rate) which showed promising results with a p-value <0.01. It was observed that geometric CBR, although had a better hit rate, was more suitable for only one nearest neighbour while conventional CBR is better suited when the dataset is large [20].

3. PROPOSED GENARALIZED METHODOLOGY:-

Researchers and modelers often adopt different procedure to apply various machine learning techniques to build a stock market prediction system. In this paper, a generalised methodology for forecasting the stock market is outlined by considering various modelling factors and components as described by Fig 1.

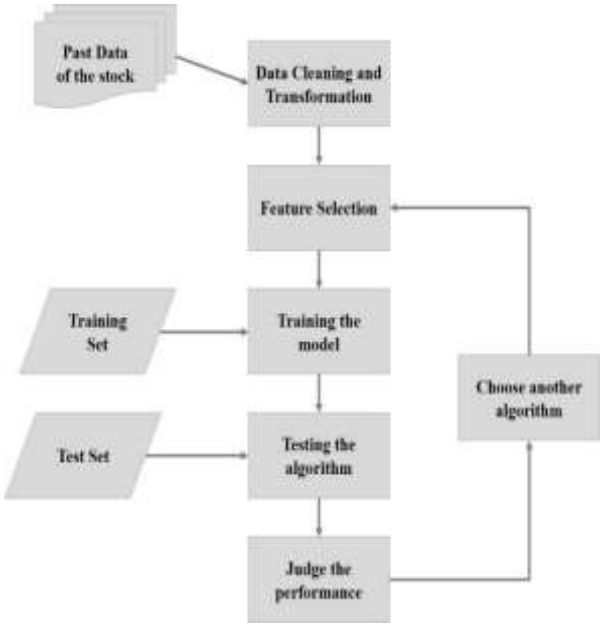

Fig 1. Generalized methodology for using machine learning algorithms, for stock market prediction.

The past information of a stock is loaded onto the chosen analysis environment and is subject to data pre-processing which including cleaning of the data to remove null, unwanted and noisy records. This is followed by transformation of data into a format suitable for analysis, which includes standardizing the currency and calendar dates. Once the data is pre-processed, the relevant features required for the modelling is chosen as predictor and output variables. Once the features are selected, the training data is fed to the chosen machine learning algorithm for training. Next, the learning algorithm is used to predict the output variables by considering the test set as input. The performance of the model is judged through various metrics that highlight the accuracy or error rates of the model. Judgement metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage error (MAPE), p-values, r-squared value, and accuracy. Arriving at the most suitable machine learning algorithm is an iterative process where the above methodology is repeated for various techniques and other different conditions until a satisfactory accuracy or error rate is achieved.

4. SENTIMENT ANALYSIS

While analyzing the stock market, many factors are considered, and they may be dependent on various political and international events reported which may affect the future price of the stock. Studies have shown that the movement of the stock market is dependent on the attitude of the investors, and often, investor behaviour is influenced by news articles and social media. The advent of technology and the internet on a global scale has engendered multiple sources of information. The availability of real-time global and local news and information paved the way for dispensing financial information much quickly as compared to pre-technology times. Articles about the political situation, social conditions, foreign events, government policies, trader psychology, and other topics that we see and comprehend on the Internet are examples of news information [16]. Text mining is required since such information is formulated in the form of texts, also referred to as documents.

Stock traders aim to acquire stocks whose prices are projected to rise and sell stocks whose prices are predicted to fall soonin order to profit from their investment. However, stock markets are unpredictable and hence difficult to anticipate, and external influences such as social media and daily financial news can have a favourable or negative impact on stock values. Financial news articles are seen as a more constant and trustworthy source of data. Many studies have found a strong link between financial news stories and stock market fluctuations; as a result, examining financial news stories can aid in stock market prediction. Recently, tweets are being used for the analysis of the stock market in addition to news articles wherein the 140-character restriction allows to assess the public mood regarding a stock [21]. The attitudes and feelings of investors are automatically extracted from the text through sentiment analysis which uses text mining, natural language processing, and computational techniques [22]. The polarity, i. e., the degree of positive or negative or neutral feeling [21] is recorded either through a lexicon-based approach or machine learning vectorization of the text approach.

Dev Shah and team outline a methodology to predict whether to buy, sell or hold the stock through the means of sentiment analysis that boasted an accuracy of 70.59%. They followed a dictionary-based approach where the raw data from news articles and Twitter was extracted via Application Program Interfaces (APIs) using relevant keywords within the scope of their analysis. The data was then preprocessed and transformed wherein the text was converted into numerical vectors and converted into n-grams and the words were stemmed to their simplest forms. A dictionary was manually created to cater to the specific scope and each entry was tagged with a positive or negative connotation. The n-grams were mapped to the dictionary and the text was given a sentiment score based on which the model predicts a buy, sell, or hold decision for a given stock [23]. M. Qasem and their group developed a Twitter sentiment classification technique for stock markets using logistic regression and neural network classifiers. Bigram term frequency and Unigram term frequency-inverse document frequency was used for the model trained with a dataset of 42000 instances of Twitter, Google, Facebook, and Tesla stocks suing the Twitter search API. Although both unigram and bigram gave an overall frequency of 58%, empirically unigram models outperformed bigram models [24].

5. CONCLUSIONS

Many academics in finance, mathematics, and engineering have been interested in the prediction of financial time series data, despite its nonlinear nature and resistance to following a defined path or trend. Prediction of the stock market has been termed as one of the most difficult analytical techniques, especially due to its highly noisy data. The improved access to real-time information and technology has facilitated the entry of more investors from the general public into the stock market. Given such circumstances and a risen drive to reduce risks in investments, many attempts at predicting the stock market through machine learning techniques have been made with promising results. In this paper, we reviewed Support Vector Machines, Artificial Neural Networks and Case-based reasoning and discussed the advantages and disadvantages of the techniques. Research on Case-based reasoning for stock-market prediction is lesser compared to the other methods but shows great potential for further research. The hybridisation of models is encouraged to negate the disadvantages of the underlying techniques and absorb the beneficial features. This paper also reviews recent developments in stock market prediction to include qualitative data. Text mining and sentiment analysis are popular methods of classification of stocks into a positive, negative, or neutral spectrum. Finally, a generalised modelling methodology for applying machine learning techniques to the stock market is proposed.It can be concluded that there is no single optimal technique for stock market prediction, however, a myriad of techniques is available and the most suitable one can be chosen depending on the scope of analysis.

6. REFERENCES

[1] Malkiel, B. G. "*Efficient Market Hypothesis*", Finance, pp.127-134, 1989.

[2] Malkiel, Burton, G. "*The Efficient Market Hypothesis and Its Critics*." Journal of Economic Perspectives, 17 (1): 59-82. 2003

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

[4] Investopedia,https://www.investopedia.comas accessed on 04/06/2021.

[5] Selvamuthu, D., Kumar, V. & Mishra, A. "*Indian stock market prediction using artificial neural networks ontick data"*FinancInnov 5, 16, 2019.

[6] D. Shah, H. Isah and F. Zulkernine, "*Predicting the Effects of News Sentiments on the Stock Market*" 2018 IEEE International Conference on Big Data (Big Data), pp. 4705-4708, 2018.

[7] S. V. M. Vishwanathan and M. Narasimha Murty, "*SSVM: a simple SVM algorithm*" Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), 3, pp. 2393-2398, 2002.

[8] Z. Hu, J. Zhu and K. Tse, "*Stocks market prediction using Support Vector Machine,"* 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 115-118, 2013.

[9] Shen .S and Tongda Zhang "*Stock Market Forecasting Using Machine Learning Algorithms*" (2012).

[10] Wei Huang, Y. N.-Y. *"Forecasting stock market movement direction with support vector machine"* Computers & Operations Research, 32 2513–2522, *2005.*

[11] Nti, Isaac Kofi, Adekoya, Adebayo Felix and Weyori, Benjamin Asubam. "*Efficient Stock-Market Prediction Using Ensemble Support Vector Machine*" Open Computer Science, vol. 10, no. 1, pp. 153-163, 2020.

[12] Fangqiong Luo, Jiansheng Wu and Kesong Yan, "*A novel nonlinear combination model based on Support Vector Machine for stock market prediction*" 2010 8th World Congress on Intelligent Control and Automation, pp. 5048-5053, 2010

[13] J. J. Hopfield, "*Artificial neural networks*" in IEEE Circuits and Devices Magazine, vol. 4, no. 5, pp. 3-10, Sept. 1988.

[14] SILVA, F. O.*." Soft Computing and Intelligent Systems Design: Theory, Tools and Applications. Pearson Education Limited"* 2004.

[15] N. Naik and B. R. Mohan, "*Optimal Feature Selection of Technical Indicator and Stock Prediction Using Machine Learning Technique*" vol. 985, A. K. Somani, S. Ramakrishna, A. Chaudhary, C. Choudhary, and B. Agarwal, Eds. Singapore: Springer Singapore, 2019, pp. 261–268, 2019.

[16] Paul D. Yoo, M. H. "*Machine Learning Techniques and Use of Event Information for StockMarket Prediction*: *A Survey and Evaluation*" Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference, Intelligent Agents, Web Technologies and Internet Commerce, (CIMCA-IAWTIC'05), IEEE, 2005.

[17] Adewale, O &Misra, Sanjay & Alfa, Abraham. "*Analysis of Fuzzy and Neural Networks Expert Systems in Forecasting Stock Prices*" International Advanced Journal of Natural and Applied Sciences Vol.1. N0. 1. May 2016 (First Edition) Pp 29-39 2016.

[18] Craw S. Case-Based Reasoning. In: Sammut C., Webb G.I. (eds) "*Encyclopedia of Machine Learning*" Springer, Boston, MA, 2011

[19] Kim, K, "*Toward Global Optimization of Case- Based Reasoning Systems for Financial Forecasting*", Applied Intelligence, vol. 21, no. 3, pp. 239-249, 2004.

[20] Chun, Se-Hak& Ko, Young-Woong. "*Geometric Case Based Reasoning for Stock Market Prediction",* Sustainability, 2020.

[21] Faten Subhi Alzazah and XiaochunCheng,."*Recent Advances in Stock Market Prediction Using Text Mining: A Survey"*, E-Business - Higher Education and Intelligence Applications, Robert M.X. Wu and Marinela Mircea, IntechOpen, June 1st 2020.

[22] Agarwal B, Mittal N, Bansal P, Garg S. "*Sentiment analysis using common-sense and context information*" Computational Intelligence and Neuroscience. 2015.

[23] D. Shah, H. Isah and F. Zulkernine, "*Predicting the Effects of News Sentiments on the Stock Market*" 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4705-4708

[24] M. Qasem, R. Thulasiram and P. Thulasiram, "*Twitter sentiment classification using machine learning techniques for stock markets*" 2015 International Conference on Advances in Computing, Communications, and Informatics (ICACCI), pp. 834-840, 2015.

