Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 2, 248 – 253

A Survey on different datasets employed during stock market prediction

Shobhita Singh[1], Dr. DivyaKhanna[2]

[1]*Research Scholar, Department of School of Computing, RIMT University, Mandi Gobindgarh, Punjab, India*
[2]*Assistant Professor, Department of School of Computing, RIMT University, Mandi Gobindgarh, Punjab, India*

Abstract

Stock market prediction is a complex yet an essential task for the investors to earn profits. It is complicated to determine the relation between the input and the output because of its randomness and volatility. Applying the available past and present data and information, one can forecast the stock market cost. Therefore, the historical market datasets play a vital role in the prediction by notifying about technical indicators like daily prices, volume traded information surrounding various stocks, closing price, opening price, highest and lowest intraday price, various fundamental ratios (like PE ratio, liquidity ratio, solvency ratio, activity ratio etc.), last traded price etc. There are several datasets that are been employed in various research papers used for the prediction of numerous stocks price. The most often used datasets for machine learning are the NSE, Chinese stock exchange, S&P 500, NYSE, NASDAQ, Istanbul stock exchange national 100 index, standard & poor's 500 return index, and the stock market return index of Germany. The suggested study's goal is to conduct a review of different datasets used in recent research publications over the last five years in order to help forthcoming researchers keep up with the current developments in the area of datasets for stock market prediction using any approach (machine learning or deep learning).

Keywords: stock market, NSE, S&P 500, NYSE, NASDAQ, PE ratio, solvency ratio, activity ratio.

1. Introduction

Investors have been working perpetually to track down approaches to precisely foresee the securities exchange. As digital currencies keep on blowing everybody away by representing the moment of truth unpredictability, the financial area is scrambling to track down ways of foreseeing and expect these market financialuncertainties.The major uses of financial markets include stock rate and trend prediction, index prediction, risk analysis and return forecasting, as well as portfolio management. ANN, CNN, ARIMA, Decision trees, SVM, LSTM, and other linear, machine learning, and deep learning approaches have been used to estimate the future cost of stocks, including ANN, CNN, ARIMA, Decision trees, SVM, LSTM, and others. By obtaining a link between multiple technical indicators and share price movement, stock market trends and prices may be forecasted.The relationship is determined through the historical data collected viva various dataset sources as in NSE, NYSE, S&P 500, Japan stock exchange, yfinanceetc. therefore data collection additionally supplements to improve the dataset by including more information that are outer. Raw data is then checked for providing coherent configuration to improve the dataset through data mining.Information mining processes have the ability to predict the future direction of stock/list prices clearly or exactly, allowing investors/traders to increase their profits with minimal risk. Information mining algorithms have been shown to be effective in predicting stock value development with high accuracy.

By the amount of contracts exchanged, the National Stock Exchange (NSE) has emerged as the leading exchange in equities derivatives and currency derivatives; the evaluation is based on monetary subject matter specialists. Furthermore, it includes a diverse pool of around 20 million financial subject matter specialists. The securities exchange is the primary location where publicly traded company equities are traded between buyers and sellers, as well as the

source of exchange data. The stock estimation progressions are based on the stock's demand and supply. Because of its chaotic, complicated, and consistent variations in nature (Volatile), which makes it tough to predict steadfast quality, the financial exchange estimate has been a favorite region for investors as well as academic researchers.

Table *1*. analyzing several research papers using different dataset types (stock exchange and indices), Section3collaborates literature survey and Section4 deliberates the conclusion of the study.

2. Assessment table based on different dataset types employed

Table *1*.conveys the assessment table describing profuse research papers of last five years using

The paper is organized in the following manner: Section1 contains introduction to the financial market and its dataset requirement, Section2 includes

A number of studies have been conducted in order to forecast the price and average movement of the stock market. A variety of tactics have been used in the fields of computer science engineering and economics to pick up a piece of this erratic data and profit handsomely from stock market investments.

varying datasets and techniques for stock market prediction.

Table 1. Various models utilizing different datasets

Ref.	Year of publication	Techniques Used	Dataset Types
[1]	2018	ARIMA with Exponential smoothing	NSE (Nifty)
[2]	2018	MLP, RNN, LSTM, CNN	NSE and NYSE
[3]	2019	CNN, feature extraction	S&P 500, NASDAQ, NYSE, DJI
[4]	2019	AR, MA, ARMA, ARIMA	Nifty and Sensex
[5]	2020	Decision tree, Random Forest, adaptive boosting, gradient boosting, ANN, RNN and LSTM	Tehran stock exchange
[6]	2017	Long Short-Term Memory Neural Network	1 Bovespa index from the BM&F Bovespa stock exchange
[7]	2018	CNN, LSTM	S&P's 500, DJIA (NYSE)
[8]	2019	First order fuzzy using CPDA	BSE, NYSE, TAIEX
[9]	2019	ANN, SVM	Germany stock exchange

[10]	2018	Logistic regression, SVM (5 & 10-fold cross validation)	Chinese stock exchange (SSEC index)
[11]	2017	Stock quantitative feature matrix, stock correlation matrix	Chinese A-share stock data, HK stock data
[12]	2018	SMC algorithm, LSTM	78 A-share stocks in CSI 100 and 13 popular HK stocks
[13]	2017	Decision tress, NN, SVM	AAPL (Apple NASDAQ) stock
[14]	2020	CNN with attention based bidirectional LSTM	Chinese stock exchange
[15]	2018	Decision support for financial market	DAX (German prime index), CDAX, STOXX Europe 600

3. Literature Survey

It has been observed that majority of survey papers have not precisely concentrated on dataset survey in the sphere of the stock exchangeprediction applications. This section focuses on the literature survey on related work analyzing several different dataset types utilized during price and trends prediction of stocks.Mentioned below are fifteen research papers and articles portraying diverse techniques as well as datasets.

In 2018, Uma Gurav and Nandini Sidna[1]endeavor to carry out research into many aspects of dynamic stock market forecasting, in view of the fact that minimizing risk in stock market investing is inextricably linked to minimizing prediction mistakes. They discussed several machine learning approaches for stock market prediction, as well as their benefits and drawbacks, and found that EML (Ensemble machine learning) algorithms can be more useful based on specific technical indicators and performance evaluation.Hiransha M, Et al in 2018 [2], on two separate exchanges, NSE (National Stock Exchange) and NYSE (New York Stock Exchange) day-wise closing prices, proposed a model employing four deep learning approaches, namely MLP, RNN, LSTM, and CNN, was presented. The neural network outperformed the linear model ARIMA, according to the results.Ehsan Hoseinzadea, Saman Haratizadeha in 2019[3], presented a CNN-based architecture for extracting characteristics from collected data from diverse sources in order to forecast the market's future. They used the framework to anticipate next day price trends and movement on indexes such as the NASDAQ, S&P 500, NYSE, DJI, and RUSSELL using various sets of starting data. When compared to state-of-the-art baseline methods, the evaluations reveal a considerable improvement in prediction performance.SHEIKH MOHAMMAD IDREES, Et al in 2019[4], proposed a statistical model to analyze the time series of Indian stock market indicesi.e., Nifty and Sensex. They applied AR, MA, ARMA and ARIMA techniques and found ARIMA outperformedamongst all. M. Nabipour and Et al in 2020[5], on 10 years of historical data from the Tehran stock exchange, researchers used decision trees, adaptive boosting (Adaboost), bagging, random forest, gradient boosting, and eXtreme gradient boosting (XGBoost), as well as artificial neural networks (ANN), recurrent neural networks (RNN), and long short-term memory (LSTM). Each prediction model presented used 10 technical indicators as inputs. Based on assessment criteria such as MAPE, MAE, RRMSE, and MSE, the results demonstrated that LSTM is more accurate.David M. Q. Nelson, Et al in 2017[6], proposed a model using LSTM to forecast the

stock market future price and trends through technical indicators based on historical price data.They showed an average of 55.9% of accuracy when applied on 1bovespa index from BM&F bovespa stock exchange.PisutOncharoen and PeeraponVateekul in 2018 [7], Numerical information, such as historical price data and technical indicators, and textual information, such as headlines and news content, were employed as inputs. On the S&P 500 index, the DJIA index, and the NYSE, the suggested model employed CNN and LSTM with event embedding vectors taken from the inputs. They came to the conclusion that using both text information and technological indicators improves prediction accuracy.Shanoli Samui Pal, Samarjit Kar in 2019[8], By applying data discretization based on fuzzistics [1; 2], the cumulative probability distribution technique (CPDA) was employed to generate the intervals for the linguistic values to forecast time series of stock price. The model was tested on the closing prices of three stock indexes from three different time periods: TAIEX, BSE, and NYSE.Dharmaraja Selvamuthu, Vineet Kumar and Abhishek Mishra in 2019[9], For financial market prediction, we used variants of the ANN approach based on three learning algorithms: Levenberg-Marquardt, Scaled Conjugate Gradient, and Bayesian Regularization, and we used tick data as well as 15-min data from an Indian firm. Using tick data, all three algorithms achieved a maximum accuracy of 99.9%.Huiwen Wang, Shan Lu, Jichang Zhao in 2018[10], suggested a system for forecasting the stock market that aggregates three forms of data: scalar variable, compositional variable, and functional variable. The Shanghai Stock Exchange Composite (SSEC) index of the Chinese stock market was used in this study.Xi Zhang, El al in 2017[11], extracted the actions fromnews and sentiments of users from social media, and combined them to see the impact on stock market price movement by applying coupled matrix and tensor factorization framework on Chinese A-share stock data and on HK stock data.Jieyun Huang, Et al in 2018[12], provided a model that employed tensor to combine multisource data, such as financial Web news, quantitative stock data, and investor sentiments gathered from social media. In addition, an improved sub-mode coordinate algorithm (SMC) and LSTM were developed. In the years 2015 and 2016, these strategies were applied to 78 A-share stocks in the CSI 100 and thirteen prominent HK equities.Bin Weng, Et al in 2017[13], on AAPL (Apple NASDAQ), constructed an "inference engine" employing three machine learning models: decision trees, neural networks, and support vector machines. The prediction of the next day's AAPL stock movement was found to be 85 percent accurate.Jiawei Long, Et al in 2020[14], using knowledge graph and graph embeddings approaches, CNN was combined with attention-based bidirectional LSTM to choose the optimal target stocks for synthesizing market and trading information.Stefan Feuerriegela, Julius Gordon in 2018[15], which has implications for commercial applications of decision-support in financial markets, especially given the increasing frequency of index ETFs (exchange traded funds). The test was conducted on the DAX (German main index), CDAX, and STOXX Europe 600 indexes.

Conclusion

The objective of the survey paper was to analyze different indices of various stock exchange used in research papers (from last five years) as dataset types in stock market prediction. This is very important to determine on what dataset type the experiments (techniques) are applied and accordingly outcomes also depend on that.As per the survey, it can be concluded thatNSE, NYSE and Chinese stock exchange indices are frequently operated dataset types and other indices: S&P 500, NASDAQ, DJI, Tehran stock exchange, TAIEX and Germany stock exchange has also been into consideration. **Figure 1** depicts the percentage range of several dataset types used in various research papers in last five years. This figure can help the upcoming researchers to know onwhich stock exchange the experiments are already performed, and which are unexplored areas in indices of stock exchange for future prediction of price and trends movements.

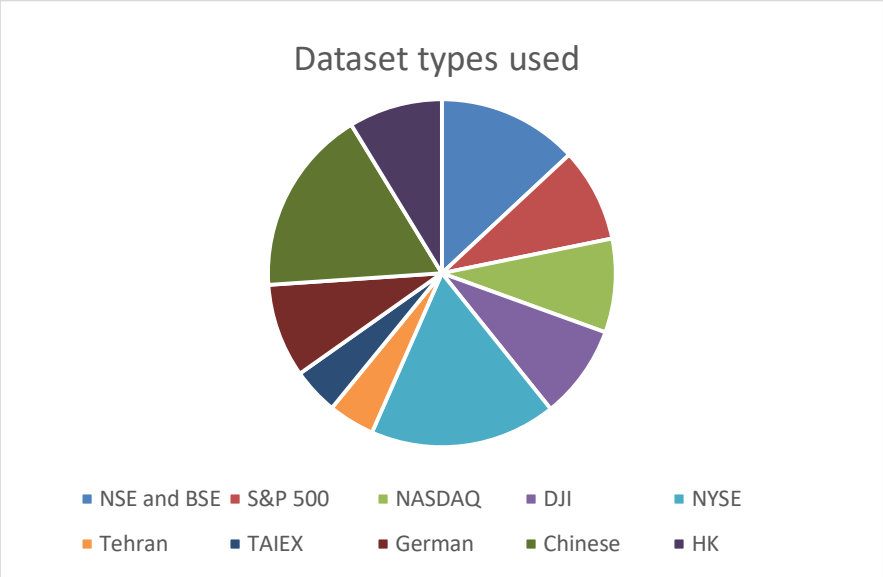

Figure 1: Depicting the percentage of a variety of Dataset types used

References

[1] U. Gurav and N. Sidnal, "Predict Stock Market Behavior: Role of Machine Learning Algorithms," in *Intelligent Computing and Information and Communication*, Springer, 2018, p. 383–394.

[2] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman, "NSE stock market prediction using deep-learning models," *Procedia computer science*, vol. 132, p. 1351–1362, 2018.

[3] E. Hoseinzade and S. Haratizadeh, "CNNpred: CNN-based stock market prediction using a diverse set of variables," *Expert Systems with Applications*, vol. 129, p. 273–285, 2019.

[4] S. M. Idrees, M. A. Alam and P. Agarwal, "A prediction approach for stock market volatility based on time series data," *IEEE Access*, vol. 7, p. 17287–17298, 2019.

[5] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana and others, "Deep learning for stock market prediction," *Entropy*, vol. 22, p. 840, 2020.

[6] D. M. Q. Nelson, A. C. M. Pereira and R. A. de Oliveira, "Stock market's price movement prediction with LSTM neural networks," in *2017 International joint conference on neural networks (IJCNN)*, 2017.

[7] P. Oncharoen and P. Vateekul, "Deep learning for stock market prediction using event embedding and technical indicators," in *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, 2018.

[8] S. S. Pal and S. Kar, "Time series forecasting for stock market prediction through data discretization by fuzzistics and rule generation by rough set theory," *Mathematics and Computers in Simulation*, vol. 162, p. 18–30, 2019.

[9] D. Selvamuthu, V. Kumar and A. Mishra, "Indian stock market prediction using artificial neural networks on tick data," *Financial Innovation*, vol. 5, p. 1–12, 2019.

[10] H. Wang, S. Lu and J. Zhao, "Aggregating multiple types of complex data in stock market prediction: A model-independent framework," *Knowledge-Based Systems*, vol. 164, p. 193–204, 2019.

[11] X. Zhang, Y. Zhang, S. Wang, Y. Yao, B. Fang and S. Y. Philip, "Improving stock market prediction via heterogeneous information fusion," *Knowledge-Based Systems*, vol. 143, p. 236–247, 2018.

[12] J. Huang, Y. Zhang, J. Zhang and X. Zhang, "A tensor-based sub-mode coordinate algorithm for stock prediction," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, 2018.

[13] B. Weng, M. A. Ahmed and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Systems with Applications*, vol. 79, p. 153–163, 2017.

[14] J. Long, Z. Chen, W. He, T. Wu and J. Ren, "An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market," *Applied Soft Computing*, vol. 91, p. 106205, 2020.

[15] S. Feuerriegel and J. Gordon, "Long-term stock index forecasting based on text mining of regulatory disclosures," *Decision Support Systems*, vol. 112, p. 88–97, 2018.