Macroeconomic Factors Modeling Optimization in Stock Prediction Using Machine Learning

Zhengao Chen

International Business School Suzhou
Xi'an Jiaotong-Liverpool University
Suzhou, China
Zhengao.Chen18@student.xjtlu.edu.cn

Abstract

Stock prediction has been a focus of research in recent years. Traders are seeking to acquire an effective model to predict the stock prices in the future to make investments and earn arbitrages. Methods in machine learning and deep learning have been broadly used in economic model buildings. However, important factors like macroeconomic environments and government regulations were not considered effective in most cases. With different events happening in various situations, the influences can be extremely different. In this essay, we will use machine learning methods to analyze the impacts of various conditions and how this will optimize prediction accuracy. In the meantime, it will offer a new perspective of view to conduct technical and sentimental analysis based on the premise of fundamental analysis.

Keywords: *artificial intelligence; stock market; price prediction; neural networks; machine learning*

1. Introduction

With the unprecedented developing speed of artificial intelligence and its growing applications, investors and researchers began conducting practices of methods of artificial intelligence and machine learning in the stock market to predict stocks to earn revenue. In these circumstances, the accuracy rate of the prediction system became the decisive element to judge its effectiveness of a prediction system. It has been concluded in the research in this area that stock prediction based on machine learning methods was categorized into three types of research: fundamental analysis, technical analysis, and sentimental analysis [1]. Huang et al. [1] pointed out fundamental analysis was based on published statements and reports from the companies. In their research, they conducted research on the performances of three machine learning methods in fundamental analysis, of which the RF (Random Forest) reaches the most accuracy to recognize the winners. Fundamental analysis was attached less emphasis to in research since FA (fundamental analysis) was not built for short-term investment [2]. Yun et al.'s research [4], pointed out the assertion that more factors should be selected and trained in the machine learning models. They suggested feature selection should be applied in stock prediction in machine learning to improve the anticipation results.

While these researches focus on increasing accuracy in fundamental analysis, they failed to consider simulating the macroeconomic environments in fundamental analysis and how sudden events can influence stock prices in technical and sentimental analysis. This essay will analyze how this will optimize the system accuracy in certain circumstances such as in special markets under regulations like China, In this research, under the background of the Chinese stock market in the past 20 years, we implement the influences of regulations and calculated these effects as a parameter in a monthly period using RF algorithms. Furthermore, monthly parameters will be transformed and applied to the LSTM algorithm for technical analysis and sentimental analysis. it is estimated with this supplement for the former algorithms, prediction algorithms will increase by approximately 6% in TA (technical analysis) and SA (Sentimental analysis).

Since the financial crisis in 2008, investors are raising increasingly aware of the financial trend and recovery methods from potentially harmful events. Covid-19 can be a valid example of this: since 2012, world financial markets like the futures and options and world distribution systems have remained the trend of a slight increase in 7 years, while Covid-19 changed this situation and investors still are not fully aware and prepared for the upcoming event. Although our research cases mainly focused on certain economic environments, these algorithms can be

Y. Jiang et al. (Eds.): ICEDBC 2022, AEBMR 225, pp. 1869–1875, 2022.
https://doi.org/10.2991/978-94-6463-036-7_279

adjusted and suited in separate situations and used by various parameters in different countries respectively. Former stock prediction models mainly emphasize one single perspective or combination of two analyses such as FA&TA or TA&SA. This essay provides a new perspective on using FA as general background and how it will optimize TA and SA. This method will also significantly increase stock prediction models in various limitations and the reliability in these circumstances. Additionally, this method is considered a supplement to former algorithms to improve prediction accuracy.

This essay will be organized as follows: Section II presents previous works in the stock market and machine learning. Section III provides methods and developments, followed by section IV explaining the methods and developments. Discussion and empirical results are evaluated in section V. In section VI, a summary and further suggestions are given.

2.Related Works

In recent years, the application of machine learning has been a focus of study. In most research, the prediction via machine learning is based on three categories: fundamental analysis, technical analysis, and sentimental analysis. In this section, former works and practices will be explained.

In terms of fundamental analysis, machine learning and neural networks form models based on the data sets on the financial statements posted by the underlying companies. Huang et al. [1] conducted a fundamental analysis based on three algorithms. Feature selection and bootstrap aggregation are used to optimize the prediction accuracy result. With the data sets of 22 years of financial data, RF was proved the most accurate machine learning algorithm after the experiment. LSTM algorithm is also brought out by Landis et al. [4] as a valid model to train big data sources and increase the prediction accuracy. Using the financial data in the statement as a basis, overarching ensemble prediction can be conducted.

Technical analysis is the most understandable analysis in stock prediction. Machine learning is used to analyze and study historical data and its parameters to predict its future performance. Porshnev et al. [5] used the data sets with the parameters of historical data in the last 50 years and concluded that the application of neural networks can be of great assistance to prediction, raising the accuracy to 99% for a single stock. Long Short Term Memory (LSTM) can also be used for the technical analysis of the algorithms. With 50 batch-size data for the data set (50 trading period), the percentage error can be minimized [6][7]. In this parameter-based analysis, the choices of parameters can be significant to the learning process. Yun [3] conducted research on factor collecting and pre-processing of the required data. Another classification problem can be used in this consideration and in turn, another machine learning algorithm should be developed for factor searching. Random Forest (RF) can be used for this classification.

Sentimental analysis is attached greater importance to stock prediction algorithms. News and hearsays play a vital role in influencing peoples' investment. Liu et al. [6] researched the relationships between the daily stock average and news headlines. In this research, events like political strategy, trade war, unemployment, and especially the effects of Covid-19 have been taken into consideration for the algorithm. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) were seen as effective for this simulation. Relevant research had also been measured in the aspects of news polarity and top news from Twitter and it is pointed out that deep learning models have been proved more effective than machine learning methods [9] [10][11].

3.Methodology

3.1.Data Set Preparation

The data we use in the prediction models are the 21 representative stocks belonging to various industries' nationalized enterprises with the most market share and 21 personalized companies that firstly issued stock in the market. The reason for choosing these stocks is illustrative: they are the most representative companies in the Chinese stock market and they are the most influenced companies by the regulations in the stock market. In this consideration, it can be an effective data set for analyzing the influences of special events under certain circumstances. In terms of the regulations, we import the publicity notice China Securities Regulatory Commission and use the RF algorithm to categorize each effect and conduct feature selections for these regulations. F23

For each training process, we use the profile of randomly choosing 5 of these 40 stocks, optimization was applied and memorized for not repeating profiles. Financial data was sorted and imported into the CSV. file and the cross-fold verification was applied after the training process. In the data set building period, we found that a great amount of data was missing and unrecoverable. To handle this problem, feature deletion was used to delete the stock data with 20% of its part missing and RF models automate the replacement process with the counterparts in similar roles in other industries. For TA and SA processes, data set selection standards are different: for TA, completeness needs to be reflected in the randomness of selection; for SA, inner information and news are considered more influential for nationalized enterprises. Therefore, we choose a different distribution formula for the portfolios in SA. Detailed standards are stated in table 1 below.

TABLE I. DATA SELECTED IN THE STUDY

Data Set	Method for choosing the data set
Training and Verifying in FA	Randomly select 5 of the 40 stocks, avoiding repetition, with the total number of 8 portfolios
Training and Verifying in TA	Randomly select 4 of the 20stocks with different belongings, with a total number of 10 portfolios.
Training and Verifying in SA	Random select 5 stocks with the proportions of 80% nationalized and 20% personalized with a total number of 5 portfolios

3.2.Local Training and learning

In fundamental analysis emphasizing macroeconomic environments, we employ three methods for different stock portfolios: firstly build an RF model for all the stocks, secondly randomly select 5 stocks of the two types in the RF model, and finally build every single model for the portfolios that are selected in the methods above. In this study framework, we could obtain the overall result and some specific data from selected industries. And in the meantime, machine learning in the selected portfolios would have better performance.

3.3.Model Explanation

We applied the RF (Random Forest) algorithm in the fundamental analysis in consideration of the satisfying performance of its ensembled classifiers. Random Forest is composed of multiple decision trees and each logic classifier can be seen as a supplement to the others. Moreover, the construction of a random forest is conducted in two techniques: random data selection and random characteristic selection. These advantages can be significantly beneficial to the classification of the events and regulations. Those selections during different years in the study process because the effects of the same regulations in different ages. The market situation can be dramatically different and in most cases, a single classifier is not able to tell these differences effectively. Specifically, the number of decision trees and the number of features randomly selected by each node of the decision tree are ought to be set distinctly for different research problems. Based on Qi et al.'s study, [12] we set the number of decision trees to be 500, and the number of features selected randomly by each node of the decision tree is log2 d≈7. LSTM neural network was employed in technical analysis and sentimental analysis in this essay. RNN (recurrent neural network) is effective in simulating and predicting data-moving trends. LSTM as a major method in RNN can considerably process the sequence information tasks. Forget gate, input gate, output gate, and cell state consist of the LSTM model which uses the three doors to control input and forgotten information. Recurrent measurement makes advantages of both the sharpness of short-term memory and accuracy of long-term memory possible. Stock data as constant data, short-term memory, and long-term memories are both needed for analysis in this essay. As mentioned above, most prediction systems only focus on short-term stock interest. Other than the monthly output parameter of the influences which acquires the short-term analysis in CNN (concurrent neural network), the system can be optimized to predict the stock price in a long period such as 5 years with the RNN algorithm. In figure 1 below, the research process of the whole essay will be given in the form of a flow chart for reference.

Figure 1. Flow chart of the study method in this essay

3.4.Evaluation Metrics

The expectation for the study is to develop a tool for selecting profitable stock portfolios for both short-term and long-term investments. Since the prices of selected stocks have significant time series characters, it is significant to choose the data vectors, data, and evaluation time period for the prediction accuracy access.

For data vectors selection, opening price, closing price, the change number, and turnover rate are taken into account for the accuracy access. Since the essay is aimed to discuss how the premise of FA will affect and enhance the prediction model. The confusion matrix is applied to access the quality of binary classifications as a basic assessment method in machine learning which measures the data by MCC (Matthews Correlation Coefficient). Change number and turnover rate are imported and defined into the LSTM model for reference and assistance in price prediction. In terms of the time period in assessment, in consideration of the correlation data span in stock markets, the time step should not be set too large and here we choose to compute the accurate rate daily and compute the average rate once a week. After each week, the temporary analysis basis of change amount and turnover rate is updated with new predictions for these two data vectors. The former basis for these vectors is taken into historical data for analysis.

In the MCC matrix, numbers of true positives, true negatives, false positives, and false negatives are calculated as follows:

$$MCC = \frac{Tp \cdot Tn - Fp \cdot Fn}{\sqrt{(Tp+Fp)(Tp+Fn)(Tn+Tp)(Fn+Fn)}} \quad (1)$$

The value calculated by MCC ranges from -1 to 1, where 1 denotes a 100% accuracy rate; 0 means a rough prediction in the period with a 50% accuracy rate; -1 means a total disagreement between prediction and observation representing a 0% accuracy rate.

For data measurement in the discussion section, the accuracy rate is firstly calculated to its average weekly and then calculated year-based to reduce statistical errors. To avoid data replication for the monthly parameters of macroeconomics, we avoid using monthly parameters.

4. DISCUSSION

This essay researches optimizations for stock prediction accuracy in certain circumstances. Formerly published research on stock and related financial products mainly focus on the sole analysis of one or two perspectives respectively. The relationships between research in three aspects are not discovered enough and attached importance. Predictions nowadays can not effectively predict stock prices in various circumstances such as in markets under regulations analyzed in this essay. This vacancy is growing its urgency when market conditions nowadays changed rapidly under the pandemic environments. This new perspective in the essay can effectively improve the accuracy of stock prediction under the parameterization of the general financial market environment and the LSTM model is effective for supervised learning with this premise.

The research chooses the broad dataset of selected representative stock data in Chinese markets and invested optimization of stock prediction methods from the perspective of macroeconomic elements. Fundamental analysis is measured as a general premise. Further technical analysis and sentimental analysis are conducted on the premise. The selected dataset is trained in the Random Forest model and a monthly macroeconomic parameter was output to the LSTM model for sentimental and technical analysis using Python. The detailed schedule for the research is graphed in figure 1 in the format of a flow chart. Detailed accuracy calculation is represented in (1) for further analysis.

In this research, it is found that the accuracy rate is 6% higher on average than solely conducting TA and SA with the same dataset. In the circumstances like financial crisis between 2007 to 2019 and Covid-19 between 2019-2020, the figure is almost 17% higher in TA and 22% higher in SA. This research is considered an optimization and supplement for Liu et al.'s research on sentimental analysis on news influences in stock prediction [8]. For technical research, it enhances the LSTM-based model using historical data and also conducted the research method in a more specified market circumstance and improves its evaluation metric in a more reasonable way [13].

To have a clearer sight of the influence of the optimization methods, based on the methods stated above, the results of three randomly selected portfolios in the technical analysis have been selected as drawn in the comparison chart below. The chart shows the changes in the accuracy rate trained with or without optimization premises. The solid line represents the results after optimization and the dotted line denotes the counterpart. Round measurement was used in figure 2.

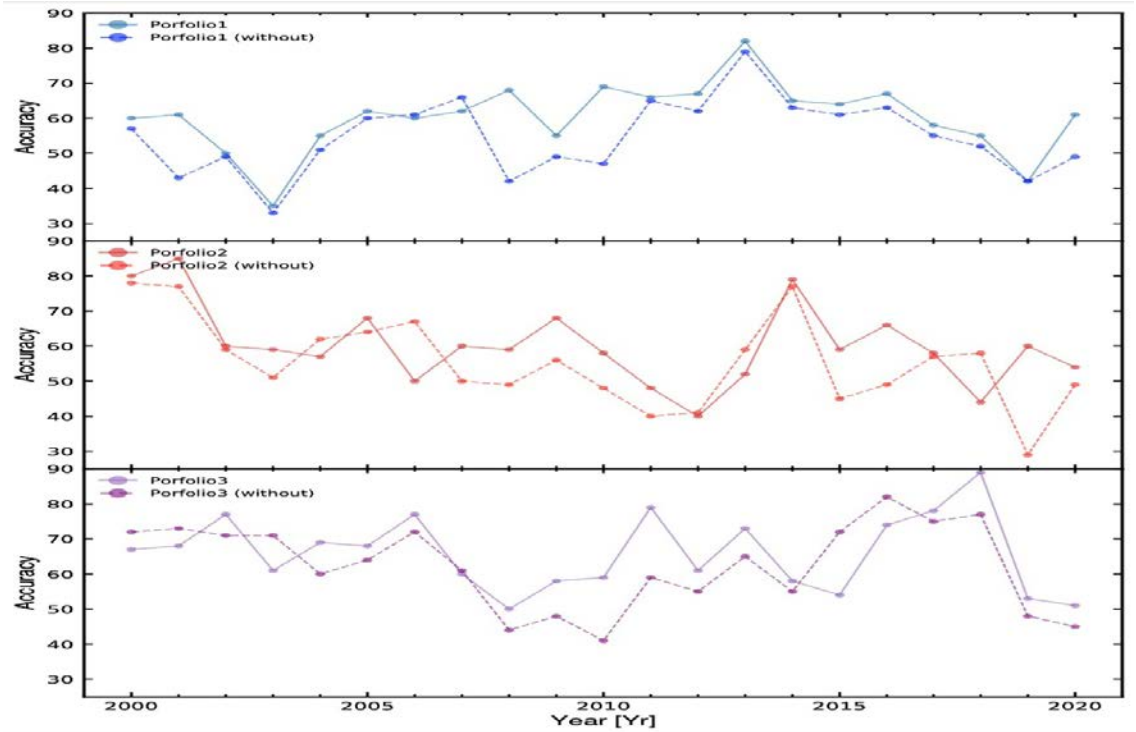

Figure 2. Sample comparison result for technical analysis

In the overall view, it can be clearly concluded that with the optimization premise, the accuracy rate of SA and TA increases more than that without the premise by 7 percent and 6.5 percent respectively. The results for the overall performance comparison are drafted as below in figure 3.

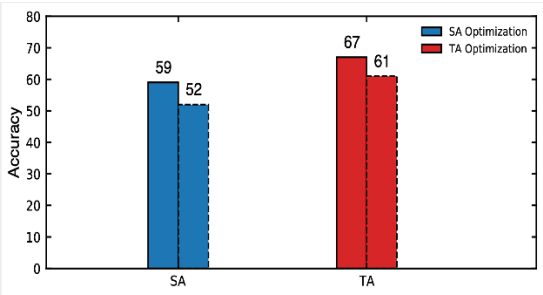

Figure 3. Overall comparison of accuracy performance for SA and TA

According to He et al.'s research [14], larger inflation is expected in sentimental research. In unstable economic inflations, the influences of news become amplified, and investors get more distracted by the news. This results in unprecise predictions based on historical data in the meantime. Qiao [13] also stated this problem. In his research, LSTM is effective in TA in the Covid-19 period and he left questions for the improvements for more precise prediction in special period. Therefore, this essay takes the period of Covid-19(2019-2020) as an example to access how the optimization measurement can improve the prediction of predictions without the premise.

As a result, from 2019 to 2020, the stock prediction has 13% higher prediction accuracy with the optimization measurement for SA and 17% higher for TA. The method proved its capability for analyzing in these circumstances. Data are drafted in figure 4 below to have a better view of the enhancements.

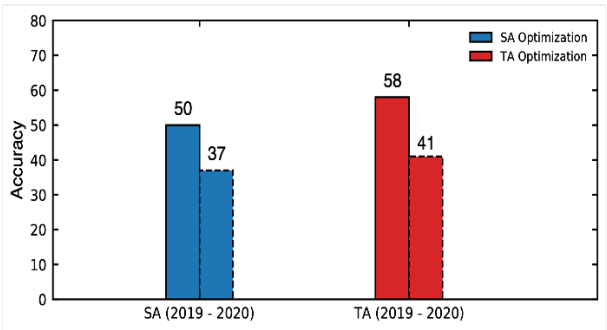

Figure 4. Periodic comparison of accuracy performance for SA and TA during Covid-19

As a more precise and specified model of stock measure, this optimization still needs more improvements to itself for more precise prediction and can be an effective tool to earn revenues. Since it is only a rough qualitative experiment, effective enhancement can be conducted based on the optimization measurement. Firstly, the more precise dataset can be selected for the RF training. In this essay, the analysis was based on all the regulations of the china securities regulatory commission, and for more specified research, detailed classifications can be used for training. Additionally, stock data can be used from different countries and different sections to test its usability in different and make improvements. Secondly, in different steps of this research, the choices of machine learning algorithms can be adjusted and the evaluation matrix can be improved to

better represent prediction accuracy. For example, Lin et al.'s research [15] also pointed out the validity of employing the CNN-based LSTM model which is another opposite perspective compared to this study. Thirdly, more economic factors can be imported into the models training for more parameters to guide the analysis. For instance, the prices of international financial derivatives can be used as references to guide the study. There are still various areas such as employment and testing platforms, data measurement methods, and more detailed deployment in sentimental analysis. More improvements in these areas are expected to be developed in the foreseeable future for this optimization method.

5.Conclusion

Overall, this paper proposes an optimization measurement on stock prediction with the modeling of macroeconomic factors as the basic premise before analyzing. This essay investigates the Chinese security market as an example and uses regulations as the macroeconomic factor. As a result, after the test, it is found that macroeconomic factors optimization can increase the price prediction accuracy by an average number of 6.5% and this optimization measurement is significantly more effective for stock prediction in unstable worldwide circumstances. More developments in algorithms and data measurement are expected to be conducted and more tests in various data are expected to refine the optimized model.

REFERENCES

[1] Y. Huang, L. F. Capretz, D. Ho, "Machine learning for stock prediction based on fundamental analysis," in IEEE Symposium Series on Computational intelligence (SSCI). IEEE, pp. 01-10, December 2021.

[2] A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis," in 2013 IEEE 13th international Conference on Data Mining Workshops. IEEE, pp. 440-444, December 2013.

[3] S. Goswami and S. Goswami, "Stock Market Prediction Using Deep Learning LSTM Model," in 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). IEEE, pp. 01-05, October 2021.

[4] Z. Yun, P. Liu, and B. Wu, "Stock Prediction via Machine Learning and Factor Analysis," in 2020 IEEE MIT Undergraduate Technology Conference (URTC), IEEE, pp. 01-04, October 2020.

[5] M. Misra, A. P. Yadav, and H. Kaur, "Stock market prediction using machine learning algorithms: a classification study," in 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEEECE), IEEE, pp. 2475-2478, July 2018.

[6] Y. Liu, J. Trajkovic, H. G. H. Yeh, W. Zhang, "Machine Learning for predicting stock market movement using news headline," in 2020 IEEE Green Energy and Smart Systems (IGESSC), IEEE, pp. 01-06, November 2020.

[7] G. Ranibaren, M. S. Molin, S.H.Alizadeh, A. Koochari," Analyzing effect of news polarity on stock market prediction: a machine learning approach," in 2021 12th International Conference on Information and Knowledge Technology (IKT). IEEE, pp. 102-106, 2021.

[8] C. Dadiyala and A. Ambhaikar, "Technical Analysis of Pattern Based Stock Prediction Model Using Machine Learning," in 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), IEEE, pp. 01-09, September 2021.

[9] Y. Xu and V. Keselj, "Stock prediction using deep learning and sentiment analysis," in 2019 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5573-5580, December 2019.

[10] W. Landis and S. Cha, "Towards High Performance Stock Market Prediction Methods," in 2020 IEEE Cloud Summit. IEEE, pp. 156-160, October 2020.

[11] K. Puneeth, R. Sagar, M. Namratha, P. Ranispoorti, and W. Rohini, " Comparative Study: Stock Prediction Using Fundamental and Technical Analysis," in 2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC) Mobile Networks and Wireless Communications (ICMNWC), IEEE, pp.01-04, December 2021.

[12] Y. Qi, W. Yu, and Y. Deng, "Stock prediction under COVID-19 based on LSTM," in IEEE Asia-Pacific Conference on Image Processing," Electronics and Computers (IPEC) Image Processing, Electronics and Computers (IPEC), 2021 IEEE Asia-Pacific Conference, IEEE, pp. 93-98, April 2021.

[13] G. Chang, J. Qiao, Y. Liao, Y. Wang and Z. Zhang, "A Perspective of LSTM Based Stock Prediction," in 2021 International Wireless Communications and Mobile Computing (IWCMC) Wireless Communications and Mobile Computing (IWCMC), 2021 International, IEEE, pp. 576-580, June 2021.

[14] P. He, Y. Sun, Y. Zhang, and T.Li, "COVID–19's Impact on Stock Prices Across Different Sectors–An Event Study Based on the Chinese Stock Market," in Emerging Markets Finance and Trade, pp. 2198-2212, July 2020.

[15] S. Lin, W. Xu, and J. Liu, "Two-channel Attention Mechanism Fusion Model of Stock Price Prediction Based on CNN-LSTM," in Transactions on Asian and Low-Resource Language Information Processing, vol 20, pp. 01-12, September 2021.