Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/57/20230633

The Current Status and Future Prospects of Machine Learning in the Chinese Stock Market

Tingwei Li[1,a,*]

[1]*WeBank Institute of FinTech (SWIFT), Shenzhen University, 3688 Nanhai Avenue, Nanshan District, Shenzhen,518060, China*
a. litingwei2020 @email.szu.edu.cn
**corresponding author*

Abstract: This article explores the current status and prospects of machine learning techniques in the stock market of china. It begins by providing an overview of the relevant concepts of machine learning and analyzing its strengths and weaknesses in the stock market context. The article also introduces the current application of machine learning techniques in the gobal stock market. In the methodology section, several machine learning models that have performed well in the stock market are listed. The article then summarizes and analyzes the specific applications of machine learning and its algorithms in the stock market of China, highlighting the advantages and limitations due to the unique characteristics of the market.Finally, the article concludes by summarizing the aforementioned content and providing an outlook on the future development direction of machine learning technology in the stock market of chinese. While the performance of machine learning technology in the stock market has its pros and cons, it is undeniable that it holds an important position in the future development of the financial market, especially within the wave of innovation and progress in the financial industry.

Keywords: machine learning, China stock markets, financial technology

1. Introduction

Due to the continuous increase in trading volume, the stock market is generating a massive amount of data, including both structured and unstructured data. The scale and complexity of the dataset are continuously rising, while traditional forecasting methods are inadequate in effectively processing these datasets. However, machine learning technology can effectively extract critical information from this vast and complex data.

Through data analysis and induction, machine learning, a subfield of artificial intelligence, enables computer systems to automatically improve their algorithms and handle challenging issues. It uses statistical models, learning algorithms, and data mining techniques to allow computer systems to learn from data without explicit programming and perform specific tasks assigned by humans [1]. It is a method of data analysis that makes use of algorithms to find patterns and rules in the data so that the computer system may learn from the information and enhance its learning algorithms to solve challenging issues.

Current research in machine learning focuses on feature understanding, analysis of relevant data, autonomous selection of suitable learning algorithms for training, and utilization of trained models for data analysis and predictions. Through this process, machine learning gains sharp insights at the data level and makes predictions based on existing data.

DOI: 10.54254/2754-1169/57/20230633

Compared to traditional forecasting methods like linear regression, machine learning is more suitable for making predictions in the modern stock market. It exhibits greater flexibility, allowing simultaneous analysis of multiple feature variables and selecting the most appropriate algorithms among various functions [2]. Additionally, machine learning models can automate data processing, feature extraction, and model training. They can make quick decisions and adjust investment portfolios promptly. Lastly, machine learning can identify potential risk factors through large-scale data analysis, enabling comprehensive and accurate risk assessment and prediction.

However, machine learning also has some limitations in its performance in the stock market. Firstly, compared to other application domains, the stock market data has a relatively low signal-to-noise ratio and machine learning models can easily overfit due to the inclusion of more data and features, leading to model inefficiency [2]. Secondly, insufficient expertise among practitioners can result in unexpected model algorithm performance. Machine learning algorithms are considered black-box models for many professionals and investors, as they might not understand the mathematical principles and operating mechanisms behind them, making it challenging for them to improve model performance by adjusting parameters or changing functions. Lastly, machine learning models have higher requirements for the quantity and quality of historical data, which can be a limiting factor in their performance.

In conclusion, machine learning efficiently analyzes critical information from large and complex datasets and trains algorithms to make predictions. For the stock market of China, its unique characteristics in investor structure, market system, and data volume make traditional forecasting methods face significant challenges. The flexibility and autonomous learning ability of machine learning can better adapt to the characteristics of the stock market of China and accomplish related tasks [3]. This is one of the significant reasons why machine learning has been rapidly developing in the Chinese stock industry, making research on its application in the stock market highly meaningful.

This article mainly provides an overview of past literature, introduces several machine learning algorithm models that have performed well in the stock market, and outlines the specific application methods and current development status of machine learning techniques in the stock market of China. Finally, it analyzes the advantages and limitations and concludes by looking forward to the future development direction of machine learning technology in the stock market of China.

2. Review of Several Existing Machine Learning Models

2.1. Principal Component Regression (PCR)

PCR is a statistical modeling. In PCR, the input data are first downscaled using PCA to transform the high-dimensional independent variables (input features) into a new set of principal components (projection variables) that are linear combinations of the distinctive characteristics. Following that, regression models are constructed using linear regression using these major components as predictor variables. The goal of PCR is to reduce the dimensionality of the data while preserving the key patterns of change in the data. By reducing dimensionality, PCR can handle high-dimensional datasets, reduce multicollinearity, and increase the model's capacity for explanation.

The PCR model consists of two main steps, firstly, the feature variables are combined into a linear combination by the PCA method, in which the covariance structure between the variables will be maintained in a better way [1]. Secondly, in the model, PCR uses several leading components to normalize the prediction problem by using coefficients that zero out the low-variance components [2].

DOI: 10.54254/2754-1169/57/20230633

2.2. Neural Network Model

Neural network is a machine learning model that consists of basic units called neurons (or nodes, units) that are connected and transmit information. These neurons are usually organized into different hierarchical structures, including three different layers.

Neural networks basically work by learning to extract features from input data and using these features to make predictions or classifications. Each neuron takes information from the layer below it, weights and changes it using an activation function, and then sends the modified input to the layer below that. Such transfer and transformation of information is repeated in the network until it reaches the output layer, which produces the final prediction.

The weights and bias parameters in a neural network are learned, and the learning process usually uses backpropagation algorithms (backpropagation) and gradient descent optimization methods. During the training process, the neural network adjusts the weights and biases by comparing them with the labeled data to minimize the disparity between the outcomes as anticipated and the actual numbers. In this way, the network gradually learns patterns and regularities in the data and can make predictions on unseen data.

2.3. Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) is an integrated learning method that iteratively trains a series of decision tree models and progressively improves the performance of the models in a gradient descent manner. The basic principles and steps of Gradient Boosting Decision Tree can be divided into three parts.

The first is initializing the model: a simple model (e.g., a decision tree or a constant) is used as the initial model for training. Next is iterative training: for each iterative step, residuals (the difference between actual and predicted values) are calculated based on the predictions of the current model. Then, using these residuals as targets, a new decision tree model is trained which will try to capture the part of the residuals that the previous model could not explain. The new model is added to the current model and used to update the predictions. Scaling the predictions of each new model using a learning rate parameter allows one to regulate the level of contribution of each new model. The performance of the overall model is gradually improved by iteratively training a series of decision tree models, each building on the residuals of the previous models. Finally for the integrated prediction: the prediction results of all decision tree models are weighted and summed to get the final integrated prediction result.

The key idea of gradient boosting decision trees is to use gradient descent optimization methods to gradually improve the model's fitting ability by iteratively training a series of decision trees. Each new decision tree attempts to capture the residual portion of the data that could not be explained by the previous model, thus improving the overall model performance. By integrating the predictions from multiple decision trees, a more accurate prediction structure can be obtained.

GBDT performs well in many machines learning tasks, especially in regression and classification problems. It is highly flexible and generalizable and is very good at modeling nonlinear relationships.

2.4. Random Forest

Random Forest (Random Forest) is likewise an integrated learning method that combines the concepts of decision trees and randomness. It consists of multiple decision trees, each constructed by random sampling of the dataset and random selection of features.

The implementation of the Random Forest model is divided into three main parts, the first is the random sampling of the feature variables, from the original dataset using a method with put-back sampling (bootstrap), to generate several different training sets, every one of which has the same

DOI: 10.54254/2754-1169/57/20230633

amount of samples as the initial dataset. In this way, each training set is generated by taking samples from the original dataset, but each training set may contain duplicate samples and missing samples. Next is decision tree construction for each training set, an independent decision tree model is constructed by using decision tree algorithms (e.g., ID3, CART). During the construction of the decision tree, each node no longer considers all the features when choosing to divide the features, but randomly selects a portion of the features from all the features to be evaluated. This random selection of features helps to increase the variability between decision trees and improve the diversity of the integrated model. Finally, there is the integration of prediction results. For the classification problem, the random forest combines the classification results of each decision tree into a final prediction result through a voting mechanism (majority voting). That is, the category that receives the most votes are selected as the final prediction result. In order to get the final prediction for regression issues, Random Forest averages the forecasts of each decision tree.

By combining the prediction outcomes of various decision tree models, random forest increases the resilience and generalizability of the model. With the introduction of random sampling and random selection of features, it reduces the risk of variance and overfitting of the model and evaluates the importance of features. Random forest is widely used in machine learning with good performance and ease of use. It is suitable for various tasks including classification, regression and feature selection.

3. Discussion

3.1. Key Application Areas of Machine Learning in the Stock Market of Chinese

Object categorization, value prediction, and other tasks may all be accomplished using a broad variety of machine learning approaches [4]. Emerging applications based on these domains are revolutionizing the traditional financial industry as well as society in a major way [4,5]. A large number of financial institutions or companies are tilting their resources towards the research and application of machine learning techniques, from fund companies, banks, and investment service providers in the traditional financial industry to many Internets financial companies providing fintech services [4, 6]. This trend also appears in stock investment market of China, the use of artificial intelligence in the stock market of China is growing, and many investment institutions take the initiative to introduce machine learning into their investment system, and based on machine learning technology to develop new investment products, machine learning is mainly used in these aspects.

First of all, machine learning is frequently employed to create quantitative trading methods in the stock market of China. Due to the development of technology, new characteristic factors are constantly being discovered, that predict stock prices more accurately [7]. Therefore, many institutions and firms use machine learning techniques to develop and optimize quantitative trading strategies to make trading decisions and execute them in an automated and intelligent way. These tactics, which seek to increase investment returns and lower risk, are based on the study and forecasting of market data using machine learning algorithms.

In order to predict stock price movements and market trends, predictive analytics for the stock market uses machine learning, which is the second application of predictive analytics. By using machine learning algorithms and models, historical data and relevant factors can be utilized to predict future stock price movements. These predictive analytics can provide decision support and market outlook for investors.

In addition, machine learning is also applied to investor sentiment analysis, based on the theory of behavioral finance, investor psychology and behavior will have a certain impact on stock prices [6]. Machine learning can use text mining technology, and correlation analysis methods to comprehensively analyze the social media containing finance-related keywords in the emotional text, it is high possible to increase the precision of stock market trend forecast by investor and market

DOI: 10.54254/2754-1169/57/20230633

sentiment analysis, which helps investment institutions to better understand the market dynamics, emotions and investor behavioral preferences, to make the appropriate investment decisions [6].

Turn to risk management, machine learning is employed in assess the risk of investment portfolios by analyzing market data and related factors, and providing optimization suggestions to achieve the purpose of reducing risk, optimizing asset allocation, and improving investment returns.

Finally, machine learning technology is used to develop intelligent trading investment advisor systems, through the analysis of data, machine learning can realize automatically provide clients with asset portfolios or automated trading decisions and execution based on data.

3.2. The Advantages of Machine Learning in the Stock Market of China

Machine learning is a science that relies heavily on data, and within the Chinese market, frequent trading activities can generate a large amount of trading data and market information. This provides machine learning with rich data samples that can be used for model training and optimization to improve prediction accuracy. Meanwhile, compared to the rest of the global stock markets, the stock market of China has a variety of data sources, including exchange data, company financial reports, news reports, and social media. Such diverse data sources can provide machine learning algorithms with more comprehensive and multi-dimensional information, enhancing the analytical capabilities of the models. The Chinese government attaches great importance to the development of AI and machine learning technologies, promoting relevant policies and funding. This provides a favorable policy environment and resource support for the application of machine learning in China's stock market and promotes the innovation and application of the technology. China has a huge reserve of high-tech talents, including data scientists, machine learning experts, algorithm engineers, etc. These talents have rich experience and technical capabilities in the field of machine learning, providing solid support for machine learning applications. These applications in stock market of China have benefited from the innovative atmosphere and the development of the startup ecosystem. Many startups and tech companies are dedicated to the technology development and use in machine learning, driving the continuous advancement of related technologies and the expansion of application scenarios.

3.3. Disadvantages of Machine Learning in the Stock Market of China

While machine learning techniques perform well in the stock market of China, they also have the same shortcomings. Compared to the rest of the stock market, there are more individual investors in the Chinese market, and compared to institutional investors, the trading logic and strategy of these investors are quite different, and their behavior will cause greater volatility in the stock market of China [7]. And compared to the rest of the stock market where investors mainly rely on dividend profits, Chinese investors prefer to speculate on stocks by trading spreads to gain profits, which may lead to a large number of anomalous data, making the data inaccurate and untrue, and further leading to model failure [3]. From the perspective of the trading system, the unique up-and-down stop-plate system of the stock market of China sets upper and lower limits on intraday fluctuations, which may seem to stabilize the trading price, but once an extreme market occurs, it will take several days to be released, which tends to widen the fluctuation amplitude, a reason that may also lead to model failure in some cases. An extremely important feature of China's financial system is that state-owned capital occupies a significant position in the stock market, and due to their unique position and importance, state-owned enterprises are often criticized for their lack of information transparency, and because state-owned enterprises may deviate from the corporate norm of profit maximization due to relevant political factors, thus impairing the formula performance and generating anomalous data [3]. This situation is a very difficult challenge for machine learning. The large number of illiquid share capitals in the stock market of China may make it difficult for some important data features to reflect the real

DOI: 10.54254/2754-1169/57/20230633

corporate situation, thus making machine learning models perform poorly. Finally, the trading changes brought about by machine learning and other artificial intelligence technologies have increased financial efficiency while also giving rise to many new types of financial risks. Under China's traditional securities regulatory system, there are still deficiencies in the regulation of programmed trading brought about by such technologies [8-11], and the "black-box" problem brought about by the opacity of machine learning code has also increased the risk of regulatory failure. The "black box" problem caused by the opacity of machine learning code also increases the difficulty of regulation, which may also lead to the risk of unfairness and discrimination caused by the application of machine learning in the stock market [10].

3.4. More Widely Used Machine Learning Algorithms in Market of China

In the methodology section of the previous paper, several machine learning algorithms that perform well in the international stock market were introduced, but when applied to the stock market of China, due to the unique advantages and disadvantages of the Chinese market mentioned in the previous paper, two of the algorithms, the neural network model and the random forest algorithms, have a better performance due to their greater flexibility and accuracy and therefore have been widely researched and applied in various fields within the Chinese market. The neural network model and the random forest algorithm have been widely used in various fields in the Chinese market.

Many studies have shown that neural network models have better analyzing and predicting abilities for the nonlinear part of the financial market [12]. In addition, neural network models also have strong generalization ability [13]. Compared to other markets, China's stock market is more volatile and is more likely to produce aberrant data., which makes the neural network model more outstanding than other models. This is shown in particular by the fact that trading strategies driven by neural network models typically have larger trading profits [15] and that they perform better at predicting returns than standard trading strategies [14, 15]. Similarly, in terms of stock indices, according to relevant studies, neural network models have better performance in tracking the simulation of the SZSE 100 index compared to previously used methods [16].

Random forest algorithm, on the other hand, has better performance in stock portfolio price prediction as well as stock portfolio selection. First of all, stock price prediction, due to the chaotic nature and high volatility of stock prices, it is more appropriate to treat stock price prediction as a classification problem than a regression problem, therefore, in the Chinese market, where the degree of chaos and volatility are both high, Random Forest is popular due to its advantages in classification problems compared to other regression-based machine learning algorithms. Random Forests are popular for their advantages in classification problems. Secondly, in terms of stock portfolio selection, the Random Forest model has better fault tolerance and robustness due to the randomized training set and randomized eigenvectors, which also makes it perform better in quantitative stock selection [17].

3.5. Prospects for Machine Learning in China's Stock Market

To summarize, machine learning has been sought after by a large number of domestic financial institutions in China's stock market because of its unique advantages and many benefits, and it also has an excellent development status in China's stock market, which is reflected in a wide range of application fields, more technological achievements, abundant available data, and strong support from the government, etc. However, because of the negative situation of inaccurate data and insufficient supervision, machine learning also has poor performance. But at the same time, because of the Chinese stock market's inaccurate data, volatility, insufficient regulation and other negative situations, machine learning also has a poor performance, and this is an important problem we need to solve and improve in the future.

DOI: 10.54254/2754-1169/57/20230633

Because of the current development of machine learning in the stock market of China, I believe that in its future development and application, we should focus on the following aspects.

Firstly, improve the quality and diversity of data and features: As previously said, machine learning is a science that greatly depends on data. The upper bound of machine learning is determined by quality of data and features, while proper algorithms can assist us in exploring the potential of the data and assisting in our continual approach to the upper limit. The quality and diversity of data and features in the stock market of China still need to be improved, and the accuracy of prediction can be effectively improved by developing and analyzing some indicators in the stock market that have not been paid attention to in the past [9].

Secondly, application of Deep Learning and Reinforcement Learning: As an important part of machine learning, Deep Learning and Reinforcement Learning have their unique advantages in data analysis and prediction, Deep Learning can help identify more complex patterns and correlations, and Reinforcement Learning can be used to optimize trading decisions and asset allocation, and its high-precision prediction can bring more help to investors or overdraft institutions in the future.

Thirdly, further application of natural language processing: Natural language processing technology can help machine learning systems understand and analyze large amounts of text data, such as financial news, analyst reports and social sentiment media comments. By combining natural language processing and machine learning, the impact of market sentiment and opinion trends on stock prices can be better captured.

Fourth, improvements in Model Interpretation and Interpretability: There has been a big problem with how interpretable machine learning models are. In order for investors and regulators to better understand the decision-making process and risk factors of the models, reducing risks and further improving financial efficiency, it is essential to keep enhancing the comprehensibility of machine learning models and to try to open the “black box” in machine learning.

Fifth, improvement and perfection of legal regulation and supervision: As mentioned above, loopholes in regulation and supervision may lead to a greater possibility of unfair and discriminatory risks, therefore, in the future, we can improve the existing legal regulation and supervision system by strengthening data supervision, establishing algorithmic evaluation system, and subdividing the field of application, so as to ensure the safety and stability of machine learning algorithmic systems in the complex and changeable risk environments. Safety and Stability

Finally, cross-disciplinary cooperation and innovation: The combination of knowledge and abilities from many domains is necessary. Promoting cooperation and innovation in the fields of finance, AI, and data science to advance the application and development of machine learning in China’s stock market is an integral part of our future development and application of machine learning.

4. Conclusion

In summary, machine learning technology in the stock market of China has a more mature application, but also because of its excellent performance for many financial institutions and Internet companies favoured, but due to the objective reasons of the Chinese market, machine learning technology in the application of the level of shortcomings, need to be improved, but there is no doubt that in the future wave of financial market development, machine learning technology has an important position.

References

[1] *Mahesh, B., (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR), 9(1), 381-386.*

[2] *Gu, S., Kelly, B. and Xiu, D., (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273.*

DOI: 10.54254/2754-1169/57/20230633

[3] *Leippold, M., Wang, Q. and Zhou, W., (2022). Machine learning in the Chinese stock market. Journal of Financial Economics, 145(2), 64-82.*
[4] *Wall, L. D., (2018). Some financial regulatory implications of artificial intelligence. Journal of Economics and Business, 100, 55-63.*
[5] *Li, X., Tang, P., (2020). Stock index prediction based on wavelet transform and FCD-MLGRU. Journal of Forecasting, 39(8), 1229-1237.*
[6] *Dai, D., et al., (2019). Research on stock index prediction and decision-making based on Text mining and machine learning. China Soft Science, 4, 166-175.*
[7] *Fang, Y., Chen, Y. Z. and Wei, J., (2022). Artificial Intelligence and the Chinese Stock Market: A Quantitative Study of Investment Portfolio Based on Machine Learning Prediction. Industrial Technology and Economy.*
[8] *Lv, T. T., (2023). Improving the System of Programmed Transaction Supervision in the Age of Artificial Intelligence. Modern Economic Exploration.*
[9] *Xu, H. R., et al., (2020). A Review of the Application of Machine Learning in Stock Forecasting. Journal of Computer Engineering & Applications, 56 (12).*
[10] *Cui, C. C. and Xu, Z. X., (2020). Legal Regulation of Machine Learning Algorithms. Journal of Shanghai Jiao Tong University: Philosophy and Social Sciences Edition, 28 (2), 35-47.*
[11] *Ye, W., (2014). Risk Control Strategies for Programmed Transactions in China's Capital Market. Securities Market Guide, 8, 46-52.*
[12] *Qiu, M., Song, Y. and Akagi, F., (2016). Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. Chaos, Solitons & Fractals, 85, 1-7.*
[13] *Sun, B., Li, T. K. and Wang, B. L., (2011), A Neural Network Prediction Model Based on Sensitivity Analysis of the Stock Market. Computer engineering and Application, 47(1), 26-31*
[14] *Motiwalla, L. and Wahab, M., (2000). Predictable variation and profitable trading of US equities: a trading simulation using neural networks. Computers & Operations Research, 27(11), 1111-1129.*
[15] *Enke, D. and Thawornwong, S., (2005). The use of data mining and neural networks for forecasting stock market returns. Expert Systems with applications, 29(4), 927-940.*
[16] *Liu, L., (2010). Index tracking optimization method based on genetic neural networks. Systems Engineering Theory and Practice, (1), 22-29.*
[17] *Wang, S. Y., et al., (2016). The application of Random forest in quantitative stock selection. Operations Research and Management, 25(3), 163-168.*