The Progress and Challenges of Machine Learning Technology in Stock Analysis

Anlan Wang[a]
College of Cyberspace Security, Nankai University, Tianjin, China

Keywords: Machine Learning, Stock Market Prediction.

Abstract: With the advancement of human society and technology, economic and financial connections have become increasingly intertwined, and technological development has significantly transformed business practices. The analysis of the stock market, widely regarded as an important investment and financial market barometer of the substantial economy, is crucial to financial activities. While traditional methods such as fundamental and technical analysis remain prevalent, machine learning algorithms have gained substantial traction in stock market forecasting over the past decade, enhancing both the accuracy and efficiency of analysis. In order to gain a better understanding of the development and application of machine learning approaches, this paper provides a review of the progress and challenges related to applying machine learning techniques in stock analysis and prediction. Despite the considerable progress and growing adoption of these methods, there is still a long way to go in further advancing their application in stock prediction.

1 INTRODUCTION

With the continuous development and advancement of human society, economic and financial connection have become increasingly close and integrated over the past few decades. This trend is evidenced not only by the growing volume of international trade but also by the increased cooperation between companies, the integration of financial markets, and the widespread application of financial technology. As a result, the interdependence of the economy and finance has become widely recognized.

Advancements in information technology over the last few decades have significantly transformed the business landscape. Among the most impactful innovations are financial markets, which have a profound effect on a nation's economy (Rouf et al., 2021). As one of the most important investments of the substantial economy and the barometer of the financial market (Zhou, 2021), the stock market plays a role of growing significance in social financial, and economic activities.

Stock analysis refers to the analysis which applies methods and techniques to study and predict the price and movement of certain stocks, in order to provide investors with decision-making strategies. Stock analysis is mainly about fundamental analysis and technical analysis. Fundamental Analysis refers to the evaluation of certain stocks and the prediction of their future price according to information involving financial statements, profitability, industry position, and the macroeconomic environment of the company. On the other hand, technical analysis evaluates market trends and predicts future price movement of certain stocks by analyzing data on their historical period price and volume (Rouf et al., 2021).

These analysis methods are mostly used in stock analysis in past decades by traditional methods including fundamental analysis and statistics-based technical analysis. These methods have a high accuracy indeed provided by highly professional specialists but may make mistakes without the provision.

Nowadays, methods based on machine learning are more and more widely applied in stock analysis by stock investors. Thanks to the progress in technologies, participants in the stock market have more access to stock investment analysis. This article is going to mainly discuss the stock analysis based on machine learning technologies in the past decade, introducing the main methods, latest progress, challenges, and future direction of the technologies.

[a] https://orcid.org/0009-0002-8635-0345

Wang, A.
The Progress and Challenges of Machine Learning Technology in Stock Analysis.
DOI: 10.5220/0013212900004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 205-208
ISBN: 978-989-758-726-9

The article will mention the application of technologies including Regression Analysis (Geyer-Klingeberg et al., 2018; Elmahgop & Sayed, 2020; Zhang & Yang, 2023; Ma, 2024), Support Vector Machine (Bhosle & Galande, 2023; Liu & Dong, 2024), Clustering Algorithm (Renugadevi et al., 2016; Sáenz et al., 2023), and Neural Network (Wang, 2024; Huang, 2024) in stock analysis, to illustrate the methodologies and recent developments of the fainter financial technology.

2 APPROACHES

This section provides an introduction to machine learning approaches in stock forecasting and provides readers with insights into recent applications.

2.1 Regression

Regression is a predictive approach that models the relationship between a dependent variable and independent variables. Previous studies have used various regression approaches, including simple linear regression, multiple regression, decision tree regression, logistic regression, support vector regression, and ensemble regression (Rouf et al., 2021). New techniques of regression models have been recently utilized in the research over the past decade. The techniques include linear regression, meta-regression, linear autoregression distributed lag regression, and polynomial regression.

The linear regression algorithm is a frequently used method for solving estimation problems. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. (Zhang & Yang, 2023) In other words, the model used in linear regression is a linear equation in (1):

$$y = wx + b \quad (1)$$

where y refers to the output or label, x refers to inputs, and w are the model parameters. The goal of linear regression is to iteratively adjust the parameters to train the model and fit the dataset accurately. Once the model is trained, it can be used for prediction tasks.

Regression models are used in many researches. In research on Chinese stock prediction under COVID-19 (Zhang & Yang, 2023), attributes Nominal Price, High, Low, Previous Close, and Share Volume in the dataset were used as features of the model. The dataset was split into training and testing sets. The former was used for training the Linear Regression classifier in Python Scikit-Learn and the latter was to test the accuracy of the trained model. The model showed an accuracy of over 85% as a result.

In the research (Elmahgop & Sayed, 2020), the linear model was first defined with the variables of stock return (KSE), inflation rate (INF), exchange rate per USD (EX), nominal money supply growth rate(M2), and profit margin (MPM).

Taking distributed lag into consideration, the linear model can be rewritten as a linear autoregression distributed lag model in accordance with the bounds testing approach (Pesaran et al., 2001). The model was then applied in regression analysis to identify the relationship between stock returns and inflation.

In the research (Ma, 2024), the regression model was also tested for its effectiveness. The linear model was defined with output Close price and attributes including Date, High, Low, Open, and Volume. After the training and testing process, many metrics were used including mean square error and R squared score, in order to measure the model which turned out to be effective.

2.2 Support Vector Machine

Support Vector Machine (SVM) makes an effort to draw a line or hyperplane from a set of training examples to identify the category to which new data examples belong based on the training results. (Bhosle & Galande, 2023) In practice, models may not be able to do classification tasks between categories exactly due to the random noise and outliers. Thus, relaxation variables are applied for a certain degree of classification fault tolerance. (Liu & Dong, 2024) There are two types of Support Vector Machines: linear and nonlinear.

In the research (Bhosle & Galande, 2023), the model was supposed to predict the future stock's closing price based on the current closing price only. Three types of support vector models were used in the test, including linear function, polynomial function, and radial basis function (RBF). The dataset was divided into dependent and independent data sets before it was split into training, and testing sets and then used for the training and validating procedure of machine learning. Finally, the RBF model shows the most robustness over the other two models. Its efficiency and confidence coefficient increase over time.

In the research (Liu & Dong, 2024), it was the aim to verify the support vector model newly constructed. Similarly, the stock's closing price is predicted

according to the closing price of the previous days. In order to enhance the persuasiveness of the research, 2 other models were used as comparisons. Mean square error was tested to show the reliability of the models. As a result, the new model showed less error in comparison to the other models.

2.3 Clustering

Clustering is one of the typical unsupervised machine learning approaches, meaning learning from a dataset without labels. The algorithm executes classification by calculating the similarity of data in the dataset. There are several types of clustering algorithms, including K-means Clustering, Hierarchical Clustering, and Density-Based Spatial Clustering of Applications with Noise.

Hierarchical clustering is a connectivity-based clustering, which results in clusters with high within-cluster similarity and low inter-cluster similarity. (Renugadevi et al., 2016) Initially, each data will be distributed to a cluster. Then the algorithm is going to integrate data clusters with the most similarity by iteration until the number of clusters meets the requirement of classification.

The K-means algorithm aims to distribute data according to the distance between the data and the cluster center. The algorithm will initialize k cluster centers, and each data will be distributed to the cluster whose center is the most nearby. Consequently, the algorithm will iterate to reset the position of each cluster center based on the mean value of the data in the cluster unless the cluster center becomes stable or the iteration time reaches the setting. Generally, the K-means algorithm is more sensitive to outliers. (Renugadevi et al., 2016)

In the hierarchical clustering of the research (Renugadevi et al., 2016), an agglomerative approach is used, which means the approach would gather the clusters bottom up. Specifically speaking, the clustering model takes the percentage of different stock price increases as the independent variable. The model prioritizes gathering the stocks with more similar percentages into the same cluster in the following classification. In the K-means clustering part (Renugadevi et al., 2016), the approach uses the result of previous hierarchical clustering as the initial cluster centers for further classification. As the approaches above were applied to forecast the value of stocks with the samples from the national stock exchange, the conclusion was reached that the models could also be extended to other exchanges.

In the research (Sáenz et al., 2023), the data of 240 stocks was fetched from the Rusell 3000 including the information of prices and reports, which was then smoothed with the average of a sliding window of 5 days. In the following step, the companies were classified by clustering according to features like prices, daily returns, and financial ratios respectively, in order to form the classification allowing companies with the same characteristics in the same cluster. The clustering result was used for reinforcing the training of models like neural network and ARIMA model consequently. In other words, specifically, the data of other companies in the same cluster was added as features, to enhance the stock prediction of the very company.

2.4 Artificial Neural Network

Artificial Neural Network (ANN) is an artificial model made up of connected nodes layer by layer. In other words, a node so-called artificial neuron, is connected with nodes in the former layer and the later layer, in which the first layer that receives input data is called the input layer, the last layer that provides model output is called the output layer and the other layers are hidden layers. Methods like weighted summing, and activation function processing are used, in order to convey and proceed information between connected nodes.

In the research (Wang, 2024), the research selects 14 macroeconomic factors from aspects of the financial market, real economy, and investor sentiment as the independent variables and the other 5 variables to fit the excess stock return. The variables are used to train the artificial neural network with the algorithms of backpropagation and random gradient descent to adjust the inner parameters iteratively. Finally, the R-squared value is used for validating.

In the research (Huang, 2024), researchers use convolution neural networks for stock prediction tasks. The research fetches the data of the highest daily price of stocks within 10 years and evaluates the accuracy of the predicting approach with the R-squared score. It turns out that the model makes a close prediction of the actual results.

3 CHALLENGE

In spite of technical progress, machine learning approaches are exposed to limitations and challenges in aspects of technology and society.

Technically, though the principle of approaches differs from each other, they are encountered with some challenges in common. With the use of iteration, parameters are easy to reach and stop at a

locally optimal solution in approaches like support vector machine, and neural network. Besides, most approaches are limited by the cost of the training process and require a dataset with high quality and large scale.

In particular, approaches are faced with especial challenges. The linear regression cannot perform well in long-term prediction. (Zhang & Yang, 2023) Consequently, it takes much time for the model to train and adapt to the new data. The K-means Clustering is sensitive to the initial cluster centers. (Renugadevi et al., 2016) Different initializations could lead to different results of classification. The Artificial Neural Network can also be greatly impacted by the initialization. Besides, it gets harder to explain how the model works, as the network structure of the model becomes more and more complicated.

Socially, machine learning is still not widely put into use nowadays, despite the boom in machine learning algorithms, and the effectiveness of these techniques in stock forecasting. Corporate and individual investors still rely more on the past experience and subjective judgment of individual analysts in their forecasting and analysis of the stock market. Unlike machine learning, which is an objective analysis based on the characteristics of the data. There is still a long way before machine learning technology is widely accepted and adopted by the market.

4 CONCLUSIONS

Generally speaking, the Stock market is both an important indicator and an influential factor. Having reviewed the machine learning technologies of stock prediction, the approaches developed and put into use are exposed systematically, including regression, support vector machine, clustering, artificial neural network, and naive Bayes. All these approaches are proven to be effective for their analysis based on principles of statistics and exploitation of data features. That is the main reason these methods are gradually becoming more widely used.

However, the machine learning approaches have encountered challenges, despite the great progress researchers have made. On one hand, challenges such as locally optimal solutions, running costs, and the requirement of high-quality datasets prevent algorithms from high precision and efficiency, which limits the model performance. On the other hand, the actual application of stock analysis and prediction is still mostly based on a subjective judgment of individual analysts according to traditional methods like fundamental analysis. That means acceptance of new technologies is still on the way to be improved.

REFERENCES

Bhosle, S., Galande, S., 2023. Development of Stock Market Analysis and Prediction Methods Based on Support Vector Machine Algorithm. Bryan House Publishing, 5(11).

Elmahgop, F. O., Sayed, O.A., 2020. The Effect of Inflation Rates on Stock Market Returns in Sudan: The Linear Autoregressive Distributed Lag Model. Asian Economic and Financial Review, 10(7), 808–815.

Geyer-Klingeberg, J., Hang, M., Walter, M., Rathgeber, A., 2018. Do stock markets react to soccer games? A meta-regression analysis. Economics, 50:19, 2171-2189.

Huang, S., 2024. Big data processing and analysis platform based on deep neural network model. Systems and Soft Computing, 6.

Liu, L., Dong, M., 2024. Research on Stock Trend Prediction Method in Financial Markets Based on Support Vector Machines, Francis Academic Press, 6(6).

Ma, J., 2024. Regression and Classification in Stock Price Trend Forecasting. Computer Knowledge and Technology, 20(12), 12-14+23.

Pesaran, M. H., Shin, Y., Smith, R. J., 2001. Bounds Testing Approaches to the Analysis of Level Relationships.Journal of Applied Econometrics, 16(3), 289-326.

Renugadevi, T., Ezhilarasie, R., Sujatha, M., Umamakeswari, A., 2016. Stock Market Prediction using Hierarchical Agglomerative and K-Means Clustering Algorithm. Indian Journal of Science and Technology, 9(48).

Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., Kim, H.-C., 2021. Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. Electronics, 10, 2717

Sáenz, J. V., Quiroga, F. M., Bariviera, A. F., 2023. Data vs. information: Using clustering techniques to enhance stock returns forecasting. International Review of Financial Analysis, 88.

Wang, C., 2024. Stock return prediction with multiple measures using neural network models. Financ Innov, 10(72).

Zhou, Y., 2021. A Review of Economic Growth and Stock Returns. Economy and Trade Updates, (05), 62-65.

Zhang, Y., Yang, Y., 2023.Chinese stock price prediction under COVID-19 period based on linear Regression Model. (eds.) Proceedings of the 5th International Conference on Computing and Data Science(part2) (pp.285-290). The department of artificial intelligence, Beijing university of chemical technology; The department of computer science, Xiamen university of technology.