Transactions on Computational and Scientific Methods | Vo. 4, No. 11, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Leveraging Machine Learning for Stock Market Prediction: A Comprehensive Analysis of Market Trends and Practical Applications

HongJie Qui
School of Data Science and Analytics, Florida International University, USA
quillj47@fiu.edu

Abstract: The rapid advancements in big data and artificial intelligence have significantly reshaped financial markets, particularly in stock price forecasting. This study delves into the increasing influence of big data on the Chinese stock market, emphasizing the role of machine learning algorithms in enhancing prediction accuracy and decision-making efficiency. Beginning with an exploration of China's stock market evolution and the strategic importance of big data, the paper highlights the challenges traditional financial models face in navigating the nonlinear complexities of stock price trends. By analyzing existing algorithms, such as artificial neural networks and support vector machines (SVM), the research underscores their advantages and limitations, particularly in addressing noisy and multidimensional data. The study further proposes a novel voting method that combines feature selection with classifier optimization to enhance prediction performance. While machine learning shows promise in improving stock price trend predictions and revenue certainty, the paper also identifies limitations stemming from data dependency and the diverse economic factors influencing market behaviors. Recommendations for future research include expanding data sets, refining industry-specific analyses, and adapting machine learning models to dynamic trading scenarios. This work aims to bridge theoretical insights and practical applications, contributing to the broader discourse on financial technology integration in stock market forecasting.

Keywords: Stock price forecasts, machine learning, prediction model.

1. Introduction

Nowadays, with the profound changes in the international and domestic situation, the position of the stock market in China has been significantly elevated to a new strategic level compared to the real estate market. Since the outbreak of the COVID-19, the secondary market's dependence on the substantial economy comes to a higher degree, which objectively improves the status of the stock market and raises people's expectations.

With the rapid development of the Internet technology, it's easier for people to access data with terminal devices. For professional organizations, traditional PC-side data interaction is the main method. But for the majority of the individual stock investors, mobile communication tools are the main means. At the same time, people are fully aware that big data plays a very important role, as important as Gold Ores. There are so many individual stock investors in the market, which makes big data play a decisive role in the stock market. A considerable number of the individual stock investors select and trade stocks according to their own understanding of big data. As a result, when some professional institutions prefer the phenomenon of psychology professionals when recruiting core decision-making tiers.

A large amount of economic data appears, which is related to international relations, economic policies, market entities, etc. are generated every day. Among them, there will be some direct factors. Buffett has focused on economic news for decades and catches major investment opportunities according to the news. There is also a large amount of financial data every day, which directly or indirectly affects the trend of the secondary market, such as MLF data, FTSE A50, RMB exchange rate, etc. The data generated by the stock market itself, such as stock's up and down, trading volume, changes in northbound funds, and peripheral market trends, as well as opinions published by some public platforms, also affects stock market transactions in the short or long term.

According to the background of the stock market's policy in China, big data has a decisive influence on stock selection and trading. For example, last year's ETC concept stock Jinyi Technology is a typical trend of policy markets. Judging from the background of the stock market sentiment market, big data has caused the consistency of individual stock investors and led to the style switching in stock market. For example, the phenomenon of indiscriminate speculation of low-priced stocks after the recent reform of the ChiNext. Judging from the background that foreign investment in stock market is getting deeper and deeper in China, the trend of northbound funds, known as smart funds, has even caused phased changes in the stock market. Sometimes even a wave of market conditions is jokingly called "foreign bulls".

People are getting easier to access to big data, and people's emotions and behaviors are increasingly affected by big data. For the stock market, studying the impact of big data on behavior can undoubtedly improve decision-making to a large extent and improve the certainty of revenue. At the same time, the rise of artificial intelligence has also created beneficial conditions for us to obtain and process big data. Big data is already a real "golden sea". The establishment of a scientific acquisition and analysis model will definitely increase the certainty of stock market transactions to a greater extent.

In the 1980s, some state-owned enterprises in China began to raise funds in the form of shares and publicly issued them to the society. The first stock of New China, Fei Yue Acoustics, was quietly born in the Chinese mainland. However, compared with the jeans and Fei Xiang, which swept in our country overnight, the beginning of A-shares was facing slump. In December 1986, the State Council formally loosen the trial shareholding system for large and medium-sized enterprises owned by the individual people. On December 19, 1990, the Shanghai Stock Exchange entered China's financial capital market. China's stock market has had experienced substantial fluctuations for a long time. As it is a significant part of the stock market, the stock market plays a pretty important role in the economic development and social stability of a country.

First of all, I have to say that the stock market has played a certain role in promoting the construction and development of the national economy.

The benefits include, firstly, it is beneficial to the financing of domestic capital groups, which is of course beneficial to the expansion of domestic capital and the development of domestic enterprises. Secondly, it is also beneficial to the national finance, because the growth of listed companies will undoubtedly increase the growth of national finance and tax. Securities companies and other related enterprises also need to pay taxes, as well as stamp duty, which are beneficial to the government burden. Thirdly, the addition of a financial derivative instrument not only provides a platform for domestic enterprises to seek growth through capital operation, but there is also a stage for them to seek development and prosperity.

But, I have to say is the stock market is a high-risk and high yield of a place, the investors may utilize all of his contacts or resources to find the currently the most suitable stock operation, such as news in the stock market, securities and the company's website, which are the majority of shareholders are familiar with the way to find information about it.

But, as we have learned python ,such algorithm, which would provide the much convenience for us to engage the analysis and research of the stock market in later, I think it also could be the first-hand information in foreign public, due to the resources are no longer processed and polished by the company.

The studies of stock price trend predicting have never stopped since the birth date of stock market.

Predict stock price trend accurately may bring huge returns for investors under fewer investment risks. The stock market is a complex nonlinear system. Traditional analysis of financial time series has limitation in stock predicting. Machine learning methods have powerful ability for nonlinear problems. In recent years, machine learning methods are widely applied to financial time series predicting. It has become popular in the study of stock price trend predicting. Compared with traditional modeling methods, the methods based on machine learning algorithm have unique advantages in stock trend predicting.

In this study, first, the feasibility and advantages of the machine learning method in the research of stock price trend predicting were discussed and pointed out the advantages and disadvantages of the existing learning algorithms used in stock predicting. The theories of statistical learning and support vector machine were introduced.

In this paper, we will use one of the algorithms to start from the current situation of China's stock market, and take two or three large companies as a reference case for analysis and comparison to get the results.

The research on the stock market prediction has been more popular in recent years. Numerous researchers tried to predict stock prices or indices based on technical indices with various mathematical statistics models and machine learning techniques. Although these researches exhibit satisfactory prediction accuracy, the prediction accuracy of whether stock market goes or down is seldom analyzed.

Stock market forecasting is considered as a challenging task of financial time series forecasting. There is a lot of research in this area using artificial neural networks. Many successful applications show that artificial neural networks are a very useful tool for time series modeling and prediction. Early researchers focused on using artificial neural networks to predict the stock market, and recent research tends to hybridize several artificial intelligence techniques. Later, genetic algorithm was proposed to carry out feature discretization, and artificial neural network connection weight decision was used to predict stock price index. These methods reduced the dimension of feature space, but enhanced the prediction performance. However, some of these studies suggest that artificial neural networks have some shortcomings in learning patterns because of the loud noise and complex dimensions of stock market data. Therefore, artificial neural networks exhibit inconsistent and unpredictable performance over noise data. However, the BP neural network, the most popular neural network model, has difficulty selecting a large number of control parameters, including the size of the relevant input variables, the learning rate, and the momentum constant.

Recently, a new neural network algorithm has been invented. Many traditional SVM neural network models implement the empirical risk minimization principle, while SVM implements the structural risk minimization principle. The former seeks to minimize misclassification or deviation from the correct solution to the training data, while the latter seeks to minimize an upper bound generalization error. In addition, the SVM solution may be globally optimal, while other neural network patterns may tend to fall into the locally optimal solution. Therefore, SVM cannot be fitted. In addition, a new voting method is proposed which combines the different classification algorithms with the feature set elected by the Wrapper method of each classifier. Compare with the difference between common voting methods, which is called stacking. The voting method proposed by the author is that the common stacking scheme only combines several different classifiers to reach a consensus. In this method, the Wrapper feature selection algorithm is further used to find the best feature set for each specified classification adopted in the voting method.

2. Research Methodolgoy and Analysis

2.1 Intro

We obtained historical stock data of many domestic and foreign companies through the yfinance module, and in the process of applying SVR model to make prediction of the future stock prices, we found out that for some specific stocks such as Amazon (stock code: AMZN) and Tian Chang Group Holdings Ltd (code: 2182.HK), the results SVM model simulated have large bias. The bias is reflected on the RMSE and the decisive factor R2. We let the true value been y_i , and expected value been y_i with a total of m data in the test set. Then the formula of RMSE will be the following:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{y_i} - \widehat{\mathrm{y_i}})^2}$$

From a machine learning perspective, a lower RMSE will result in a less degree of dispersion for the bias, and thus a better prediction. The coefficient of determination reflects the proportion of the change in the dependent variable that can be explained by the independent variable through the regression relationship. Using the mean as the error benchmark, we will see if the predicted error is higher or lower than the mean error, and the formula is the following:

$$\mathrm{R2} = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{y_i} - \widehat{\mathrm{y_i}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{y_i} - \overline{\mathrm{y_i}})^2}$$

We first analyze the reasons for the excessive deviation of the RMSE and the coefficient of determination R2 for SVR model on Amazon and other stocks. By comparing the curve of changes in the stock prices of these stocks with other stocks, we found that the range of changes in the stock prices of these stocks is very large, unlike other stocks that change relatively smoothly. This means that the accuracy of using the SVR model to model stock price forecasts is greatly affected by the distribution characteristics of the data. Using a single model to analyze different forms of data set cannot guarantee a good fitting. Thus, our improved method is to apply ensemble learning to integrate multiple models within an ensembled learner to predict stock prices.

2.2 Ensemble learning

Ensemble learning is to combine multiple individual learners, also called basic learners, with a certain strategy to form a learning committee in order to obtain a better comprehensive strong learner. If all individual learners are of the same kind, for example, decision trees or neural networks, then this kind of ensemble is called Homogeneous; conversely, if there are both decision trees and neural networks, then is called heterogeneous. When choosing individual learners, we generally pay attention to the following two criteria: (1) Accuracy: The selected individual learner must have a certain degree of accuracy. (2) Diversity: There must be certain differences between individual learners.

The idea of ensemble learning is that the errors of one learner can be corrected by another learner. Therefore, when selecting individual learners, we need to choose different types of learners, so that these learners can complement each other, and the final prediction result is better than the single learner result. The methods of learner ensemble can be divided into two categories: (1) Sequence integration, a single learner is trained and generated in sequence (for example, AdaBoost). The principle is to use the dependency relationship between the basic learners to improve the overall prediction results by assigning higher weights to the samples that were incorrectly marked in the previous training. (2) Parallel integration method, a single learner generates in parallel (for example, Random Forest). The principle is to use the independence between basic learners to reduce errors and improve prediction accuracy through averaging.

For the output results of different learners, the fusion methods mainly include the average method, the voting method, and the weighting method.

2.3 Intro to Models

According to the fact that the performance of a single learner cannot be too bias and the differences between the learners are required, we integrate the linear regression model (LR) and the k nearest neighbor model (KNN) on the basis of the support vector regression model (SVR). Next, we will introduce these three machine learning models.

The traditional SVM is a classifier. Its basic idea is to find a partitioning hyperplane in the sample space so that the interval between the positive and negative samples closest to the hyperplane is maximized. SVR believes that the absolute value of the error between the predicted regression model f(x) and the true value y is tolerable within ϵ. That is to say, a 2ϵ-wide interval is constructed with f(x) as the center. If the sample falls within this interval, then the prediction is considered correct. Through the introduction of the first chapter, we can also know that SVR is widely used in stock price forecasting.

Linear regression is to use lines or curves to fit the distribution and trajectory of data points in space, that is, to use linear combinations of features to represent the fitted lines or curves. The commonly used fitting methods for linear regression models include least squares approximation and gradient descent method. In linear regression, the data is modeled using a linear predictive function, and unknown model parameters are also estimated through the data. Linear regression modeling is usually used for data forecasting, time series models and discovering the relationship between variables. It is a simple and effective regression method.

The idea of the K nearest neighbor model is that in a sample data set, there are many samples with known categories and feature attributes, then for a sample to be classified, its category is determined by the category of the K samples that are most similar to it in the feature space. In other words, if most of the K nearest samples belong to a certain category, then this sample also belongs to this category. In the classification decision, this method only determines the category of the sample to be classified based on the categories of the nearest samples. The three basic elements of the K-nearest neighbor model are the selection of k value, distance measurement, and classification decision rules.

When choosing a basic learner, our main basis is the two criteria mentioned earlier, namely accuracy and diversity. In addition to the SVR model, the accuracy of the LR model and the KNN model in different stock price predictions is not low. The linear regression model is a mathematical statistical model. The classification idea of the K-nearest neighbor algorithm is based on the categories of a small number of adjacent samples. These are two completely different models. Therefore, our ensemble model is a heterogeneous ensemble model. At the same time, both the LR model and the KNN model are simple models. We know that the simpler the model, the stronger its generalization ability. This is consistent with our original intention to improve the SVR model's weaker generalization ability in stock price prediction. When integrating learners, this article uses a parallel integration method to train different models separately and combine the final predicted values together to obtain the final predicted value.

3. Discussion

It is necessary to have an accurate understanding of the index and method selection under different forecasting demand scenarios of stock trading, and machine learning algorithm can assist trading strategy through a certain range of index evaluation. By selecting appropriate parameters and evaluation indicators combined with specific trading scenarios, the conversion from price trend prediction to actual revenue promotion can be realized. However, in the actual environment, the accuracy of model prediction is data dependent, so it is impossible to get a model which is suitable for all scenarios. Considering that the stock sample is still not comprehensive enough and there are complex economic factors in the real stock market which cannot be predicted, in order to get a model with generalization ability, it is necessary to further classify and deeply investigate the stock industry and specific trading scenarios, conduct more comprehensive data analysis and model adjustment, and combine the idea of machine learning algorithm and the practical application.

Through the brief introduction of machine learning and its role in stock price predicting, there is no doubt that people can better understand it and use it to create wealth and appreciate the science and technology in the new era. Stock price forecasting is also a research field which needs to be enriched. The content of this paper also needs to be further explored. Machine learning has been developing rapidly and combining with many fields, and it radiates vigorous vitality.

Reference

[1] Ding X, Zhang Y, Liu T, Duan J. Deep learning for event-driven stock prediction. In: Twenty-fourth international joint conference on artificial intelligence; 2015.

[2] Zhang L, Aggarwal C, Qi GJ. Stock price prediction via discovering multi-frequency trading patterns. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining;

2017. p. 2141–2149.

[3] Refenes A, Zapranis A, Francis G. Stock Performance Modeling Using Neural Networks: A Comparative Study With Regression Models. Neural Networks.1994;7:375–388.

[4] Dong W, Zhao C. Stock price forecasting based on Hausdorff fractional grey model with convolution and neural network, 2021 04 15;18.

[5] Moghaddam AH, Moghaddam MH, Esfandyari M. Stock market index prediction using artificial neural network. J Econ Financ Adm Sci. 2016;21(41):89–93.

[6] Wei, Y., Xu, K., Yao, J., Sun, M., & Sun, Y. (2024). Financial Risk Analysis Using Integrated Data and Transformer-Based Deep Learning. Journal of Computer Science and Software Applications, 7(4), 1-8.

[7] Xu, Z., Pan, J., Han, S., Ouyang, H., Chen, Y., & Jiang, M. (2024). Predicting Liquidity Coverage Ratio with Gated Recurrent Units: A Deep Learning Model for Risk Management. arXiv preprint arXiv:2410.19211.

[8] Xu, K., Wu, Y., Xia, H., Sang, N., & Wang, B. (2022). Graph Neural Networks in Financial Markets: Modeling Volatility and Assessing Value-at-Risk. Journal of Computer Technology and Software, 1(2).