IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11632

Stock Trend Prediction Based on Machine Learning Approaches

Harinath Paidakula[1], Dhanush B[2], Satish salaman[3], Farhana Kausar[4]

[1-4]Atria Institute of Technology, Visvesvaraya Technological University, India,

Abstract: Research of quantitate investment on stock price prediction is effective to help investors increase profits. Recently, technologies of machine learning have been well applied to explore the issue of stock trading. In this paper, Logistic Regression and Support Vector Machines (SVM) were adopted to solve the problem of predicting the trend of stock movements. The experiment showed that these two models could be effectively used in the stock market of China. Returns based on strategies we constructed were significantly better than the HS300 index. We investigated the relationship between stock returns and various models using various models. It found that the SVM model results are optimal. The annual return of the strategy based on SVM reached 17.13% and the maximum Drawdown was 0.32. In the future, we will not only focus on the stock market, but also plan to apply this strategy to other investment fields, such as trading of digital currency. We will also use other algorithms for research and comparison

1 INTRODUCTION

With the rapid development of China's economy, the stock market one of the hottest trading markets, is constantly enhancing and developing. Companies choose to list financing to look for better development opportunities. On the other hand, investment agencies and investors get the benefit by buying and selling stocks in the stock markets. But 2 investing in stocks is rather risky. If investors are not scientific to blindly invest, it is likely to cause huge losses, and even lead to bankruptcy. Recently, research on stock price prediction based on machine learning algorithm has been used more and more widely. Many scholars can
help investors increase their profits effectively by various algorithms, such as logistic regression, support vector machine, etc.
In the past, many scholars from different areas had explored the issue of stock trading. The technology of

stock prediction using machine learning continued to advance. Choudhry et al. proposed to forecast stock trend by a hybrid machine learning system, which is based on Genetic Algorithm (GA) and Support Vector Machine (SVM) used two stage fusion model comprising three machine learning techniques to predict the values of stock market index. Two indices from the Indian stock market, CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex, were selected for experimental evaluation [2]. In order to deal with the problem of blind investment in stocks, a new method was created by a stock price forecasting model, which is combining Artificial Neural Networks and Decision Trees. DT model can generate some rules to describe the prediction decision while ANN cannot explain clearly [3]. Many scholars introduced the design and architecture of trading platform that employs Extreme Learning Machine (ELM) to predict stock price, and compared it to the algorithms of other models [4]. Abraham used neural network to forecast stocks and used neuro-fuzzy system to analyze the trend of stock values. It proved that it is possible to forecast trend results of stocks by using their proposed hybrid system also put forward some novel ideas, using Twitter users mental state data to improve the accuracy of stock market index prediction [6]. Jan had also investigated this problem.

2. DATA RESEARCH

China stock market is issued by a registered company in China and listed in China. It is denominated in Renminbi and can be subscribed and traded in Renminbi by domestic institutions, organizations or individuals. China stock market are booked electronically and are subject to the 'T + 1' delivery system. They have a 10% rise or fall limit. The monthly stock data of opening price and closing price from January, 2008 to January, 2017

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11632

	000001	000002	000003	000004	000005	000006	000007	000008
count	108	108	108	108	108	108	108	108
mean	6.79	8.31	12.36	3.87	4.50	7.94	2.16	7.22
std	1.80	1.92	5.54	1.82	2.05	3.39	1.97	2.47
min	4.42	5.72	6.77	2.19	2.60	3.27	0.79	4.08
25%	5.57	6.97	9.08	2.74	3.55	5.25	1.21	5.57
50%	6.22	7.82	11.35	3.86	3.93	7.67	1.42	6.26
75%	7.47	8.96	13.08	4.10	4.50	9.75	2.04	9.07
max	13.71	13.56	38.50	15.82	14.92	19.44	8.95	16.47

Table 1 Statistical result of stock closing price (horizontal direction : stock's code

2.1 Logistic Regression

Logistic Regression is a generalized linear regression analysis model, which is often used in data mining, economic prediction and other fields. It has many similarities with multiple linear regression analysis. The implementation of logistic regression is simple and highly efficient (small amount of calculation and low storage consumption), and can be used in big data scenarios. LR belongs to discriminant models, with many regularization methods (L0, L1, L2, etc.). The relevance of different features is always considered in the model. The above Logistic Regression is a linear classification model. In order to compress the output of linear regression from a large range of numbers, output value can be expressed as possibility. There is a good advantage of compressing large values into this range. It can eliminate the influence of particularly sharp variables. To achieve this great function, we need to add a logistic function to the output. In addition, for binary classification, we can simply think that if the probability of sample x belongs to a positive class is greater than 0.5. It is determined to be a positive class, otherwise it is a negative class.

2.2 Regulation

LR is sensitive to multi collinearity of the independent variables in the model. Several ways could be carried out to reduce the correlation between different variables. One important way is selecting representative independent variables by factor analysis. However, the conversion process from logarithm to probability is non-linear. The results show that the effect of variable changes in multiple intervals on the target probability is not obvious, and the threshold cannot be determined. So L1 regularization and L2 regularization are other ways to reduce over fitting phenomenon.
Lasso Regression is regression with L1 regularization. Adding L1 regularization will produce sparse θ parameters, some of which are 0 because of L1 constraint [12]. The equation of cost function is as follows:

Adding L2 regularization can also prevent over fitting, and the regression with L2 regularization is called ridge regression [13]. The expression of the
cost function is

2.3 Support Vector Machine

Support Vector Machine is a classification algorithm. The main idea is to maximize the interval. Many facts have proved that the structural risk minimization (SRM), as one of the most basic ideas of SVM, is superior to the traditional empirical risk minimization (ERM). In the derivation process, the interval maximization is transformed into a convex optimization problem with constraints. The Lagrange multiplier method and dual learning method are introduced to simplify the optimization problem. Finally, the optimization problem with constraints is transformed into a Lagrange multiplier optimization problem. In the whole derivation process, due to the introduction of dual learning, kernel method is naturally introduced. Kernel method can be used to map to high dimensional space and solve nonlinear classification. The final optimization problem in this paper is solved by SMO algorithm [14].
SVM theory provides a way to avoid high dimensional space complexity. It directly uses the inner product function (which is kernel function) of the space. Kernel functions are linear kernel, Gauss kernel, polynomial kernel and so on. Then the solution method under separable conditions is applied to directly solve the decision problem of the corresponding high-dimensional space. When the kernel function is known, it can simplify the solution of high-dimensional space problems. At the same time, SVM is based on small sample statistical theory. It also has better generalization capabilities than neural networks.
Linear separable support vector machine and hard interval maximization:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11632

$$\min \frac{1}{2} \|w\|^2$$
$$s.t.\ y_i(w \cdot x + b) - 1 \geq 0,\ i = 1,2,\ldots,N$$

The optimal w* and b* which satisfy the condition constitute the separation hyperplane: the input feature space to another feature space through a non-linear transformation. We must use the hypersurface separated dataset in the original feature space. The corresponding hyperplane can be perfectly separated after the transformation. The kernel function represents this kind of nonlinear transformation function, which makes our nonlinear separable data set become linear separable through transformation, so as to simplify model learning [15].

2.4 Evaluation

Accuracy, precision and F1 score are used to evaluation different machine learning models in quantitative investment. The equation of accuracy is

$$Accuracy = (TP + TN)/(TP + FP + TN + FN),$$

where TP is the Ture Positive, TN is the True Negative, FP is the False Positive. The higher the accuracy, the better the classifier. Precision is a measure of accuracy, representing the proportion of positive examples in the examples divided into positive examples. The equation is

$$Precision = TP/(FP + TP)$$

And F1 score is a kind of index used to measure the accuracy of two classification model in statistics. It takes into account both the accuracy and recall of the classification model. F1 score can be regarded as a
kind of harmonic average of model accuracy rate and recall rate. Its maximum value is 1 and its minimum value is 0.

3 RESULTS AND DISCUSSION

As stated in Section 3, Lasso regression, Ridge regression, SVM with linear kernel were used to predict the stock movements in China stock market. The statistical results of different models are shown in Table 2. The results showed that Ridge regression and SVM with linear kernel are better than Lasso regression in regard to accuracy, precision and F1 score. So strategies based on Ridge regression and SVM with linear kernel were constructed for China stock market.

Model	Accuracy	Precision	F1 score
Lasso regression	0.71	0.73	0.72
Ridge regression	0.74	0.78	0.76
SVM	0.77	0.79	0.78

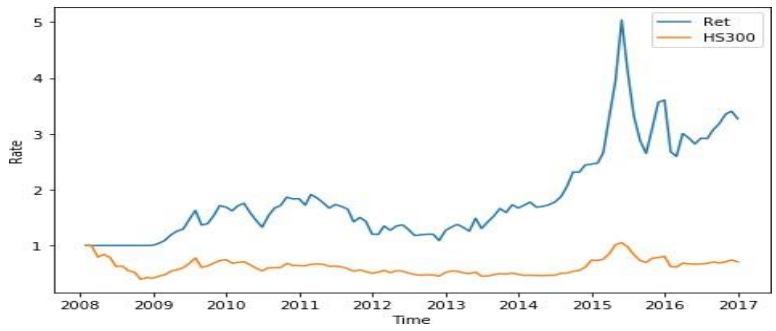

Figure 1 Return we constructed by Ridge Regression and HS300 index

The volatility of the China stock market is very large [16]. If there is no reasonable stock picking, it is likely to lead to losses. Stock selection is a more challenging problem in time-series data prediction. In this paper, we used Ridge Regression for experiments. The total yield of the portfolio we get and the yield of the broader market (HS300 Index) are

DOI: 10.17148/IJARCCE.2022.11632

shown in Figure 1. Blue line is the return we constructed by Ridge Regression; Orange line is the return of HS300 index. From Figure 1 we could see that the return of the portfolio we constructed is significantly better than the market index. In most cases, the return on the portfolio was more than half higher than the yield of the broader market. After the second half of 2015, the return on the portfolio reached 500%. This showed that our investment strategy is basically successful.

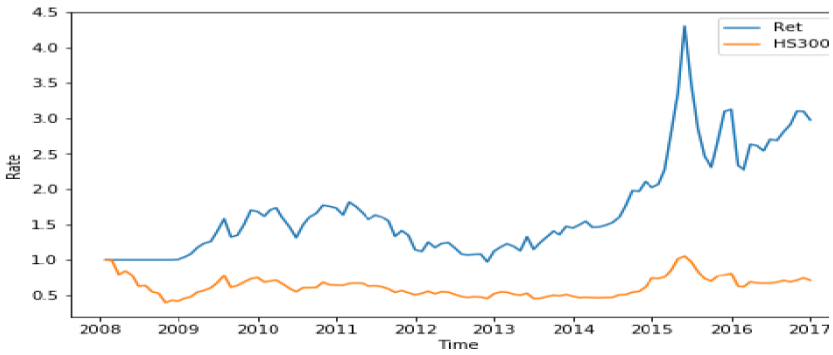

Figure 1 Return we constructed by SVM and the HS300 index

In addition to logistic regression, we also use SVM with linear kernel to predict stock movements. Figure 2 showed the return we constructed by SVM and the market index. The annual return of the strategy based on SVM reached 17.13% and the maximum Drawdown was 0.32. It presented that strategy based on SVM with linear kernel does well in predicting stock movements in China market. There is a certain difference in performance between machine learning models of different structures, which makes certain models have certain gaps in the prediction of the same stock. Therefore, it is important to choose a suitable model for stock selection.

4 CONCLUSION

This paper introduced a stock selection strategy based on Logistic Regression model and SVM model in machine learning. Both models are been widely used in various fields. In the past, many scholars have used them to predict the stock market and achieved good results. After our data mining and forecasting, we had the following conclusions. The Logistic Regression model and the SVM model can effectively predict China stock market. Both of them could be used to select a sufficiently good investment portfolio to obtain an objective rate of return. In the model we built, the return and maximum Drawdown of the SVM model were better than those of the Logistic Regression model. In addition, the investment strategies using Ridge Regression model and SVM model had higher excess return rate. And it was also better than the stock index performance at any time.

In the future, we will use other indicators to construct feature factors, such as market value factors, momentum reversal factors, to obtain higher yields. Besides, other machine learning models will be adopted to build investment strategies, such as random forests, XGBoost. Strategies based on these models may yield higher return effectively. Further, we can also consider how to better apply quantitative investment in bitcoin and other currency circles using the strategy we constructed in this paper.

REFERENCES

[1] Naadun Sirimevan; I.G. U. H. Mamalgaha; Chandira Jayasekara; Y. S. Mayuran; Chandimal Jayawardena (2020). Machine Learning for Stock Market Prediction Learning Techniques in Sri Lanka's Faculty of Computing Malabe, Sri Lanka: Lanka Institute of Information Technology Lanka.10.1109/ICAC49085.2019.9103381

[2] Sukhman Singh; Tarun Kumar Madan; Jitendra Kumar; Ashutosh Kumar Singh (2020) Stock Market Forecasting using Machine Learning: Today and Tomorrow in National Institute of Technology, Kurukshetra, Haryana, India. DOI:10.1109/ICICICT46008.2019.8993160

[3] Ferdiansyah Ferdiansyah; Siti Hajar Othman;Raja Zahilah Raja Md Radzi; Deris Stiawan; Yoppy Sazaki; Usman Ependi(2020).A Case Study of an LSTM-Method for Bitcoin Price Prediction Yahoo Finance Stock Market at Universiti Teknologi Malaysia's School of Computing in Johor Bahru, Johor, Malaysia.10.1109/ICECOS47637.2019.8984

[4] Meghna Misra; Ajay Prakash Yadav; Harkiran Kaur (2020). Department of Computer Science and Engineering, Stock Market Prediction Using Machine Learning Algorithms: A Classification Study (Deemed to be University, Patiala). DOI:10.1016/j. icrieece44171.2018.9009178

DOI: 10.17148/IJARCCE.2022.11632

[5] Mojtaba Nabipour , Pooyan Nayyeri , Hamed Jabani , Shahab S., (Senior Member, Ieee), andAmir Mosavi (2020). A Comparison of Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms on Continuous and Binary Data.10.1109/ACCESS.2020.3015966DOI

[6] B Jeevan, E Naresh, B P Vijaya kumar, Prashanth Kambli (2019). Share Price Prediction Using Machine Learning Technique at RIT, Bangalore 560054. DOI: 10.1109/CIMCA.2018.8739647

7] Sumeet Sarode; Harsha G. Tolani; Prateek Kak; C S Lifna (2019). Stock Price Prediction Using Machine Learning Techniques in Vivekanand Education Society's Institute of Technology, Mumbai, India. DOI: 10.1109/ISS1.2019.8907958

[8] Shao En Gao; Bo Sheng Lin; Chuin-Mu Wang (2019). Prediction of Share Price Trend Department of Computer Science and Information Engineering, National Chin-Yi University of Technology, Taichung, Taiwan, using CRNN with LSTM Structure. DOI:10.1109/IS3C.2018.00012

[9] Ishita Parmar; Navanshu Agarwal; Sheirsh; Ridam Arora; Shikhin Gupta; HimanshuDhiman; Lokesh Chouhan (2019). Stock Market Prediction Using Machine Learning in Department of Computer Science and Engineering, NationalInstitute of Technology, Hamirpur, INDIA. DOI: 10.1109/ICSCCC.2018.8703332

[10] A.J.P. Samarawickrama; T.G.I. Fernando (2018). In the Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura, Nugegoda, Sri Lanka, a recurrent neural network approach to predicting daily stock prices with an application to the Sri Lankan stock market was developed. 10.1109/ICIINFS.2017.8300345

[11] Chawalit Jeenanunta; Rujira Chaysiri; Laksmey Thong (2018). Stock Price Prediction Using a Long Short-Term Memory Recurrent Neural Network at Thammasat University's School of Management Technology in Thailand. 10.1109/ICESITICICTES.2018.8442069DOI

[12] Mehak Usmani; Syed Hasan Adil; Kamran Raza; Syed Saad Azhar Ali (2016). Stock market forecasting using machine learning techniques at Iqra University in Karachi, Pakistan.10.1109/ICCOINS.2016.7783235

[13] Lee, Jae Won (2002). Stock price prediction using reinforcement learning at Sungshin Women's University's School of Computer Science and Engineering.

