Stock Price Prediction Based on Machine Learning Approaches

Hongming Wang
College of Information and Control Engineering
Qingdao University of Technology
Shandong,China
whming1999@163.com

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. In different models, we analyzed the relationship between stock returns and different 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, such as andom forests, XGBoost.

CCS CONCEPTS

• Computing methodologies~Machine learning~Machine learning algorithms • Computing methodologies~Machine learning~Machine learning approaches • Mathematics of computing~Mathematical analysis • Computing methodologies~Modeling and simulation

KEYWORDS

Stock price prediction, Logistic Regression, Support Vector Machine, Annual return, Maximum Drawdown

ACM Reference format:

Hongming Wang. 2020. Stock Price Prediction Based on Machine Learning Approaches. In *Proceedings of 2020 International Conference on Data Science and Information Technology (DSIT' 20). Xiamen, China, 5 pages.* https://doi.org/10.1145/3414274.3414275

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 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) [1]. Patel et al. 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 (ANN) and Decision Trees (DT). 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 et al. 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 [5]. Porshnev et al. also put forward some novel ideas, using Twitter users' mental state data to improve the accuracy of stock market index prediction [6]. Jan et al. had also investigated this problem. They introduced recent developments in stock market forecasting models, and discussed their advantages and disadvantages [7].

DSIT 2020, July 24–26, 2020, Xiamen, China

ACM ISBN 978-1-4503-7604-4/20/07…$15.00
https://doi.org/10.1145/3414274.3414275

In this paper, we construct a strategy based the Logistic Regression (LR) model and SVM model to predict stock movements. Section 2 describes statistical results of different China stocks from January 2018 to January, 2019. Details of LR and SVM are shown in Section 3. Accuracy, precision and F1 score are also presented by equations. Section 4 is the results of these two models for stock prediction.

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 are used to predict stock movements. The statistical results are shown in following table. There are 3734 stocks. Eight examples are shown in Table 1, including count, mean, standard value, quantile and so on.

Table 1 Statistical results of stock closing price (horizontal direction: stock’s codes)

	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

3 Models

3.1 Logisttic 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. It has a good probability explanation. The model can update with new data easily using gradient descent algorithm [8]. LR is a widely used classification algorithm. LR algorithm is a more commonly used model in industrial applications and is mainly used to predict likelihood, which is not hard to implement. In classification, the computational complexity of LR is very small, the speed is very fast, and the storage resources are low. The probability score of samples can be observed conveniently [9-11]. We supposed that features of samples are denoted by sets {x, y}. In this dataset, x is a sample vector with n-dimensional features. And y means positive or negative class indicated by the number 0 or 1. If y =0, the sample x is a negative class. Logical functions are shown by the following:

$$p(y=1|x\colon\theta)=\sigma(\theta^T x)=\frac{1}{1+\exp\left(-\theta^T x\right)}$$

Where θ is the regression coefficient, and σ is the sigmoid function. In fact, this function is obtained by the following logarithmic probability:

$$\log it(x)=\ln\left(\frac{P(y=1|x)}{P(y=0|x)}\right)=\ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right)$$
$$=\theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_mx_m$$

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. The expression of the cost function is:

$$J(\theta)=\frac{1}{2}\sum\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$$

Where is the logistical function.

3.2 Regulation

LR is sensitive to multicollinearity 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:

$$J(\theta)_{L1} = J(\theta) + \frac{1}{C}\sum_{i=1}^{N}|\theta_i|$$

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

$$J(\theta)_{L2} = J(\theta) + \frac{1}{C}\sqrt{\sum_{i=1}^{N}(\theta_i)^2}$$

3.3 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:

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

The optimal w* and b* which satisfy the condition constitute the separation hyperplane:

$$w^* \cdot x + b^* = 0$$

The classification decision function is:

$$f(x) = sign(w^* \cdot x + b^*)$$

In fact, the core technique of this model is the high-dimensional space where the input features of the current dataset are located cannot be separated by the linear hyperplane. Then we transform 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 non-linear separable data set become linear separable through transformation, so as to simplify model learning[15].

3.4 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.

4 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.

Table 2 Evaluation of Different models

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

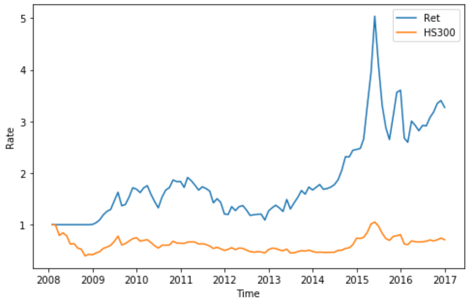

Figure 1 Return we constructed by Ridge Regression and HS300 index

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 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. The annualized rate of return of the portfolio reached 16.38% and the maximum retreat rate was 0.48.

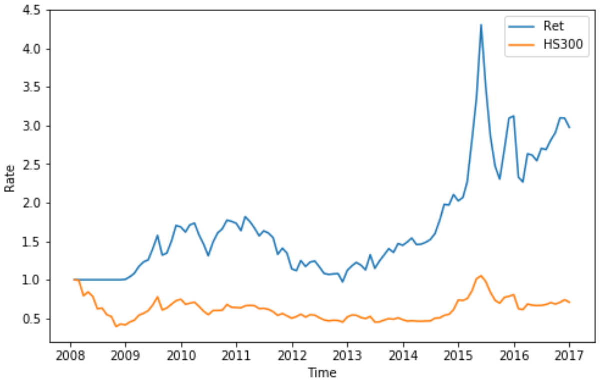

Figure 2 Return we constructed by SVM and the HS300 index

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.

5 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] Choudhry R, Garg K. A hybrid machine learning system for stock market forecasting[J]. World Academy of Science, Engineering and Technology, 2008, 39(3): 315-318.

[2] Patel J, Shah S, Thakkar P, et al. Predicting stock market index using fusion of machine learning techniques[J]. Expert Systems with Applications: An International Journal, 2015, 42(4): 2162-2172.

[3] Tsai C F, Wang S P. Stock price forecasting by hybrid machine learning techniques[C]//Proceedings of the International MultiConference of Engineers and Computer Scientists. 2009, 1(755): 60.

[4] Li X, Xie H, Wang R, et al. Empirical analysis: stock market prediction via extreme learning machine[J]. Neural Computing and Applications, 2016, 27(1): 67-78.

[5] Abraham A, Nath B, Mahanti P K. Hybrid intelligent systems for stock market analysis[C]//International Conference on Computational Science. Springer, Berlin, Heidelberg, 2001: 337-345.

[6] Porshnev A, Redkin I, Shevchenko A. Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis[C]//2013 IEEE 13th International Conference on Data Mining Workshops. IEEE, 2013: 440-444.

[7] Yoo P D, Kim M H, Jan T. Machine learning techniques and use of event information for stock market prediction: A survey and evaluation[C]//International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). IEEE, 2005, 2: 835-841.

[8] Hilbe J M . Logistic Regression Models[M]. CRC Press, 2009.

[9] Shevade S K, Keerthi S S. A simple and efficient algorithm for gene selection using sparse logistic regression[J]. Bioinformatics, 2003, 19(17): 2246-2253.

[10] Komarek P, Moore A W. Making logistic regression a core data mining tool with tr-irls[C]//Fifth IEEE International Conference on Data Mining (ICDM'05). IEEE, 2005: 4 pp.

[11] Mehta C R, Patel N R, Senchaudhuri P. Efficient Monte Carlo methods for conditional logistic regression[J]. Journal of the American Statistical Association, 2000, 95(449): 99-108.

[12] Ravikumar P, Wainwright M J, Lafferty J D. High-dimensional Ising model selection using ℓ1-regularized logistic regression[J]. The Annals of Statistics, 2010, 38(3): 1287-1319.

[13] Ng A Y. Feature selection, L 1 vs. L 2 regularization, and rotational invariance[C]//Proceedings of the twenty-first international conference on Machine learning. ACM, 2004: 78.

[14] Saunders C, Stitson M O, Weston J, et al. Support vector machine[J]. 2002, 1(4):1-28.

[15] Soman K P, Loganathan R, Ajay V. Machine learning with SVM and other kernel methods[M]. PHI Learning Pvt. Ltd., 2009.

[16] Fagan M, Zhao S. SME financing in China: the current situation, problems and possible solutions[J]. International Journal of Entrepreneurship and Small Business, 2009, 8(2):