Stock Price Prediction Based on Machine Learning Approaches: A Review

Zeel Janani
Information and
Communication Technology
Marwadi University
Rajkot, India
zeel.janani107893@marwadiuniversity.ac.in

ABSTRACT

Stock price prediction is among the most researched areas, getting interested from both academics and industry. The use of statistics and machine learning algorithms has been used to predict the opening price of the stock the next day or to understand the long-term market in the future. This paper takes a look at a variety of techniques for predicting stock values, ranging from classical machine learning and deep learning to neural networks and graph-based approaches. It offers a comprehensive examination of the methods used to forecast stock values and even the difficulties that come with them. To tackle the difficulty of forecasting stock market trends The experiment demonstrated that these two models may be applied efficiently in China's stock market. The returns from the strategies we built outperformed the HS300 index by a significant margin. We looked at the association between stock returns and various models using various models. It was discovered that the SVM model produced the best results. The annual return of the SVM-based strategy was 17.13 percent, with a maximum drawdown of 0.32 percent. Other algorithms, such as random forests and XGBoost, are being used for research and comparison.

KEYWORDS

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

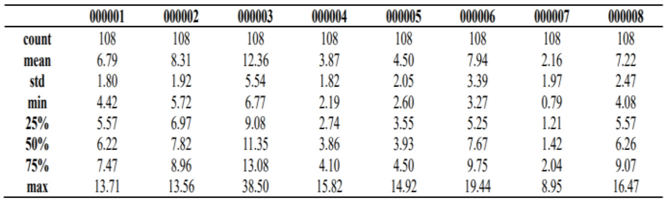
I. INTRODUCTION

Going public and issuing stocks that are then sold in secondary markets, generally known as stock exchanges, is indeed a way for businesses to raise capital for expansion or debt repayment. Instead of borrowing money in the form of cash, the company offers shares, which allows it to avoid losses, obligations, and interest payments. Second, to make money and earn profits for the stockholders. Stock price prediction using a machine learning algorithm has been increasingly popular in recent years. Various techniques, such as logistic regression, support vector machine, and many others, can help investors boost their earnings effectively. Many academics from various fields have looked into the matter of stock trading in the past. Machine learning-based stock prediction technology has advanced further. Choudhry et al. suggested a hybrid machine learning method based on a Genetic Algorithm (GA) and Support Vector Machine (SVM) to estimate the stock trend. In this research, we develop a stock prediction approach based on the Logistic Regression (LR) and SVM models. From January 2018 to January 2019, Section 2 presents statistics data on various China stocks. Section 3 delves into the details of LR and SVM. Equations also display accuracy, precision, and F1 score. The findings of these two stock prediction models are shown in Section 4

II. Data Research

The China stock market is issued and published in China by a Chinese company. It is priced in Renminbi, and domestic institutions, organizations, and people can subscribe to and trade all of this in Renminbi. The 'T + 1' distribution network governs the Chinese stock market, which is booked digitally. They are limited to a 10% rise or fall. To forecast stock fluctuations, monthly stock data of opening and closing prices from January 2008 to January 2017 is used here. The statistics results are given in the table below. There seem to be 3734 stocks available. Figure 1 contains eight instances, 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

[1]

3 Models

3.1 Logistic Regression

Logistic regression is applied when the dependent variable is binary in nature, that is, it can only have two values. Logistic regression is a simple and efficient algorithm that can be utilized in big data scenarios. It is a type of discriminant model that has a variety of regularisation options (L0, L1, L2, etc.). The model considers the significance of many features at all times. It comes with a reasonable probability explanation. Using the gradient descent technique, the model may simply be updated with fresh data. It's widely employed to solve classification problems because it just doesn't require a linear relationship between the dependent and independent variables. Because it uses a non-linear log transformation to predict the odds ratio, it can handle a wide range of relationships. The LR method is a more widely used model in industrial applications. It is primarily used to predict the likelihood and is simple to apply. It's being used to figure out how much of a possibility there is in circumstances where an event is a success and another event is a failure. When the dependent variable is binary (one of two values), that is, it can only have two values, logistic regression can be used.

Sets x, y represents sample characteristics. x is a sample vector with n-dimensional features in this dataset. And y denotes a positive or negative class, denoted by the numbers 0 or 1. If y = 0, sample x belongs to the negative class. Logical functions are shown by the following:

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

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

$$\log it(x)=\ln\left(\frac{P(y=1|x)}{P(y=0|x)}\right)=\ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right)$$
$$=\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m$$ [1]

The Logistic Regression model shown above is a linear classification model. Output value can be represented as a possibility to condense the output of linear regression from a large range of numbers. There is indeed a significant benefit to compressing huge values into this range. It can reduce the impact of extremely sharp inputs/variables.

3.2 Regulation

Overfitting can be avoided by using regularisation. It can be used to train models that generalize better on unseen data, by preventing the algorithm from overfitting the training dataset. The process of converting a logarithm to a probability is nonlinear. The analysis revealed that the impact of variable changes in the many intervals on the target probability is uncertain, and thus the threshold cannot be computed.

L1 regularisation is commonly known as Lasso Regression. ɵ. The addition of L1 regularisation results in sparse parameters, some of which are due to the L1 restriction. The cost function equation is as follows:

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

Overfitting can be prevented by using L2 regularisation, and this type of regression is known as ridge regression [1]. The cost function is expressed as

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

3.3 Support Vector Machine

In SVM, the greater the distance between the point and the hyperplane, the more confident we are in our forecast; nevertheless, when the point is close to the hyperplane, our prediction cannot be very accurate. We can pick out these dangerous places and ignore their forecast labels to reduce trade risk. As a result, the original data must be divided into at least three categories: negative, neutral, and positive. A classification algorithm is the Support Vector Machine. The main goal is to increase the interval as much as possible. Many data have demonstrated that structural risk minimization (SRM), as one of the most fundamental notions of SVM, outperforms standard empirical risk minimization (ERM). Finally, the constraint optimization issue is turned into a Lagrange multiplier optimization problem [5]. The kernel approach is organically included throughout the derivation process as a result of the inclusion of dual learning. The kernel approach can be used to map to a high-dimensional space and tackle nonlinear classification problems. The SMO technique is used to solve the final optimization problem in this research. Hard interval maximization and linear separable support vector machine:

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

Following the transformation, the corresponding hyperplane can be perfectly separated. The kernel function depicts this type of nonlinear transformation function, which transforms our nonlinearly separable data set into a linearly separable data set to ease model learning. The kernel function depicts this type of nonlinear transformation function, which transforms our nonlinear separable set of data into a linear separable data set to ease model learning.

3.4 Evaluation

To assess the performance of our classifier, we introduced the concepts of precision and recall, which are described as

Precision = Tp/(Tp+ Fp) [2]

Recall = Tp / (Tp+ Fn) [2]

From the above calculations, tp, fp, and fn indicate true positive, false positive, and false negative, respectively. The better the classifier, the higher the accuracy. Precision is a measure of accuracy that represents the percentage of positive instances in a set of positive instances. Within statistics, the F1 score is a type of index used to assess the accuracy of two classification models. It considers the classification model's accuracy as well as recall. The F1 score can be thought of as a form of the harmonic average of the model accuracy and recall rates. It has a maximum value of 1 and a minimum value of 0.

4 Results and Discussion

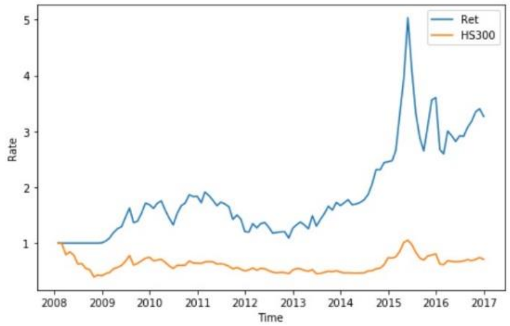
The statistical results of many models are shown in the table below. In terms of accuracy, precision, and F1 score, ridge regression and SVM with linear kernel surpassed Lasso regression. As a result, approaches for the Chinese stock market based on Ridge regression and SVM with linear kernel were created.

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

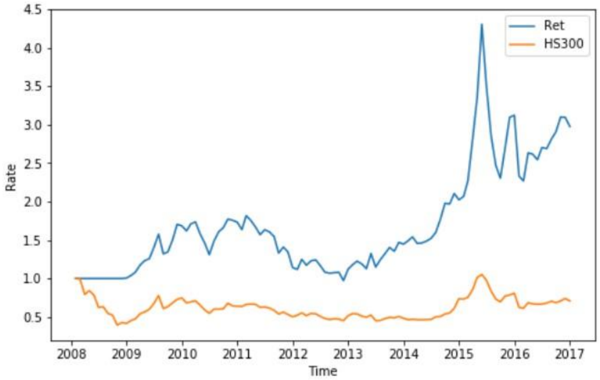
[1]

The Chinese stock market is quite volatile. Ridge Regression was employed for studies in this publication. Figure 3 depicts the total yield of the portfolio as well as the yield of the broader market (HS300 Index). The blue line represents the return generated using Ridge Regression, whereas the orange line represents the return of the HS300 index, below. The figure shows that the return on the portfolio we built outperforms the market index by a wide margin.



[1] Figure 3 Return we constructed by Ridge Regression and HS300 index

The graph below depicts the return we calculated using SVM and the market index. The annual return of the SVM-based strategy was 17.13 percent, with a maximum drawdown of 0.32 percent. It demonstrated that a technique based on SVM with a linear kernel predicts stock movements in the Chinese market well. There is a discrepancy in performance amongst machine learning models of different architectures, which causes certain models to predict the same stock with gaps. As a result, it is critical to select a good model for stock selection.



[1] Figure 4 Return we constructed by SVM and the HS300 index

5 Conclusion

This research presented a stock selection technique in machine learning based on the Logistic Regression model and the SVM model. Both models are widely used in a variety of fields. We arrived at the following conclusions. The Logistic Regression model and the SVM model are both capable of accurately forecasting the Chinese stock market. Both could be used to select a sufficiently good investment portfolio to achieve an objective rate of return. The SVM model outperformed the Logistic Regression model in terms of return and maximum drawdown in the model we created. In addition, various machine learning models, such as random forests and XGBoost [4], will be used to develop investment strategies. Strategies based on these models may produce a higher return.

REFERENCES

[1] H. Wang, "Stock Price Prediction Based on Machine Learning Approaches," 2020.

[2] E. Rosenzweig, "Successful user experience: Strategies and roadmaps," Success. User Exp. Strategy. Roadmaps, pp. 1–344, 2015, DOI: 10.1016/c2013-0-19353-1.

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

[4] 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.

[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