International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Stock Market Forecasting Using Machine Learning Algorithms

Mr. Harsh Goyal[1]

[1]Btech Scholar in information technology, Maharaja Agrasen Institute of Technology

---

Abstract - Prediction of stock market is a long-time attractive topic to researchers from different fields. In particular, numerous studies have been conducted to predict the movement of stock market using machine learning algorithms such as support vector machine (SVM) and reinforcement learning. In this project, we propose a new prediction algorithm that exploits the temporal correlation among global stock markets and various financial products to predict the next-day stock trend with the aid of SVM. Numerical results indicate a prediction accuracy of 74.4% in NASDAQ, 76% in S&P500 and 77.6% in DJIA. The same algorithm is also applied with different regression algorithms to trace the actual increment in the markets. Finally, a simple trading model is established to study the performance of the proposed prediction algorithm against other benchmarks.

Key Words: Stock price, NSE, SVM

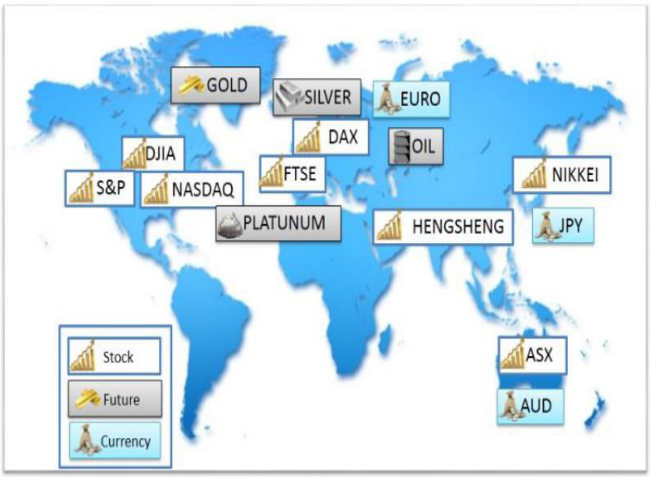
1.INTRODUCTION

Prediction of stock trend has long been an intriguing topic and is extensively studied by researchers from different fields. Machine learning, a well-established algorithm in a wide range of applications, has been extensively studied for its potentials in prediction of financial markets. Popular algorithms, including support vector machine (SVM) and reinforcement learning, have been reported to be quite effective in tracing the stock market and help maximizing the profit of stock option purchase while keep the risk low [1-2]. However, in many of these literatures, the features selected for the inputs to the machine learning algorithms are mostly derived from the data within the same market under concern. Such isolation leaves out important information carried by other entities and make the prediction result more vulnerable to local perturbations. Efforts have been done to break the boundaries by incorporating external information through fresh financial news or personal internet posts such as Twitter. These approaches, known as sentiment analysis, replies on the attitudes of several key figures or successful analysts in the markets to interpolate the minds of general investors. Despite its success in some occasions, sentiment analysis may fail when some of the people are biased, or positive opinions follow past good performance instead of suggesting promising future markets. In this project, we propose the use of global stock data in associate with data of other financial products as the input features to machine learning algorithms such as SVM. In particular, we are interested in the correlation between the closing prices of the markets that stop trading right before or at the beginning of US markets. As the connections between worldwide economies are tightened by globalization, external perturbations to the financial markets are no longer domestic. It is to our belief that data of oversea stock and other financial markets, especially those having strong temporal correlation with the upcoming US trading day, should be useful to machine learning based predictor, and our speculation is verified by numerical results. The rest of the report is organized as following. Section II presents our algorithm in details, including the fundamental principle of our algorithm, data collection and feature selection. Numerical results are shown in Section III followed by analysis and discussions. In Section IV, we established a simple trading model to demonstrate the capability of the proposed algorithm in increasing profit in NASDAQ. Section V summarizes the whole report..

2.ALGORITHMS

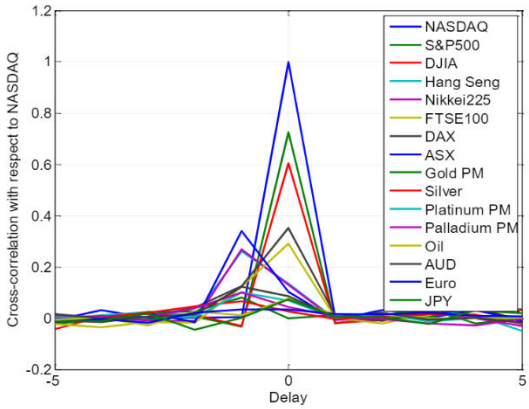
A. Basic Principles Globalization deepens the interaction between the financial markets around the world. Shock wave of US financial crisis hit the economy of almost every country and debt crisis originated in Greece brought down all major stock indices. Nowadays, no financial market is isolated. Economic data, political perturbation and any other oversea affairs could cause dramatic fluctuation in domestic markets. Therefore, in this project, we propose to use world major stock indices as input features for our machine learning based predictor. In particular, the oversea markets that closes right before or at the beginning of the US market trading should provide valuable information on the trend of coming US trading day, as their movements already account for possible market sentiment on latest economic news or response to progress in major world affairs.

B. Data collection The data set used in this project is collected from [3]. It contains 16 sources as listed in Table I and covers daily price from 04-Jan-2000 to 25- Oct -2012: Since the markets are closed on holidays which vary from country to country, we use NASDAQ as a basis for data alignment and missing data in other data sources is replaced by linear interpolation.



C. Feature selection In this project, we focus on the prediction of the trend of stock market (either increase or decrease). Therefore, the change of a feature over time is more important than the absolute value of each feature. We define xi(t), where i ∈ {1,2,..., 16}, to be feature i at time t. The feature matrix is given by F = (X1, X2, ..., Xn) T (1) where Xt = (x1(t),x2(t),...,x16(t)) (2) The new feature which is the difference between two daily prices can be calculated by ∇δxi(t) = xi(t) − xi(t −δ) (3) ∇δ X (t) = X (t) − X (t − δ) = (∇δ x1(t), ∇δ x2(t), · · · ∇δ x16(t))T (4) ∇δ F = (∇δX(δ + 1), ∇δX(δ + 2), ..., ∇δX(n)) (5) Due to the difference in market value and basis of each market, the differential values calculated above can vary in a wide range.

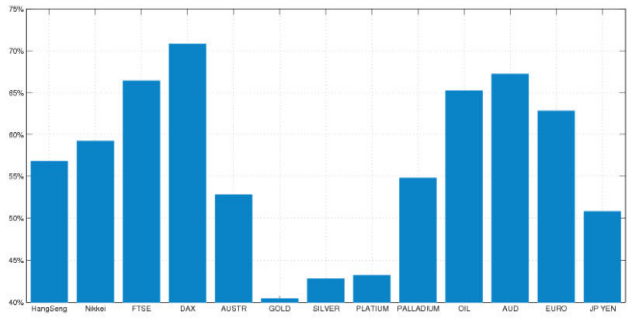
As discussed above, the performance of a stock market predictor heavily depends on theory relation between the data used for training and the current input for prediction. Intuitively, if the trend of stock price is always an extension to yesterday, the accuracy of prediction should be fairly high. To select input features with high temporal correlation, we calculated the autocorrelation and cross-correlation of different market trends (increase or decrease). The results shown in Figure 2 use NASDAQ as the base market.



3.EXPERIMENTAL RESULTS AND DISCUSSION

A. Trend Prediction

1) Single Feature Prediction In section 2, we used cross-correlation to estimate the importance of each feature. To verify the information given by correlation analysis, we use individual feature to predict daily NASDAQ index trend. The prediction accuracy by each single feature



From the figure, we can see that DAX yields the best results, 70.8% accuracy of prediction. Prediction accuracies of Australian dollar, FTSE and oil price are also relatively high, reaching 67.2%, 66.4% and 65.2% respectively.The result of this experiment supports the analysis of crosscorrelation. Hence, we are convinced that index value of other stock markets and commodity prices can provide useful information in the prediction process.

B. Regression

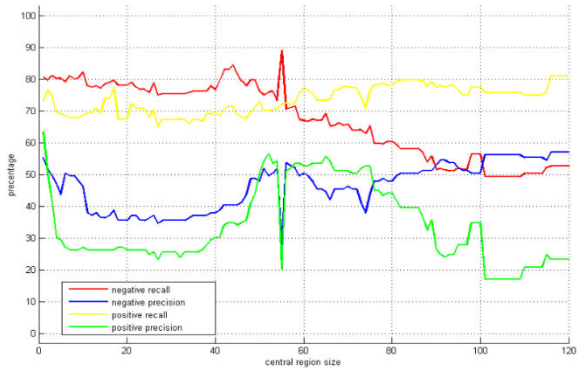
Compared to stock trend, the exact increment in stock index may provide more information for investment strategy. This means the classification problem now evolves to a regression problem. To judge the performance of our model, we use square root of mean square error (RMSE) as criteria, which is defined as $\sum$ (6) We use linear regression, generalized linear model (GLM) and SVM algorithm to predict exact value of daily NASDAQ movement. The RMSE values for different algorithms are reported The baseline predictor in the table is formed by a zeroorder hold filter. From the table, we can see that SVM gives the most accurate prediction. The RMSE given by SVM is 21.6, only half of the average fluctuation, 47.66.

C. Multiclass Classification

In previous part, we explored various methods to improve prediction accuracy and minimize square root mean square error. These efforts can be directly used to maximize the trading profit. However, besides maximizing profit, another aspect of our task is to minimize trading risk. In this part, we will use the SVM regression model and start from the basic intuition in SVM algorithm. In SVM, the further distance between the point and hyperplane is, more confident we are for the prediction we made, whereas, our prediction cannot be

very accurate when the point is close to hyper-plane. To minimize the trading risk, we can pick out these risky points and ignore their prediction labels. Thus, we need to classify the original data into at least three classes, negative, neutral and positive. This is the intuition that leads to the prototype of our multiclass classification model.

To build up the multiclass classifier, we firstly need to define the width of the central region. To evaluate how well our classifier is, we introduced the concepts of precision and recall, which are defined as (7) (8) In the equations above, tp, fp and fn stand for true positive, false positive and false negative respectively. The precision and recall values against different central region widths are plotted in Fig. 6. In the figure, Recall for positive class reflects the proportion of predicted rising days among all rising days while precision indicates the hit rate among rising predictions. Thus, recall directly impacts the frequency of trading and precision impacts the profit / loss at each time. Taking the product of the two, we calculate the F1 score.



4.TRADING MODEL

In this section, we will build up a trading model based on the predictor we find in section 2 and section 3. We compare the simulation results of our model against two selected benchmarks.

A. Basic Settings

We randomly select 5 different time slots for simulation, 50 days inside each time slots. The initial capital at the beginning of the each time slot is 10,000 dollars. At the end of the trading period, all possessed stocks are forced to be cashed. Furthermore, for simplicity, we suppose that there is no Stamp Duty or any kind of tax or fees during the process and short sale is not allowed in our simulation.

B. Model Specification

In our simulation, we use two benchmark models and one model using our predictor. Here, we will describe the three models in detail.

1) Benchmark model 1 In this simple model, we suppose we use all the money to buy stocks on the first day and sell the stocks at the end. Thus, the profit is depend on the trend during this testing period, that is (10)

2) Benchmark model 2 In this model, we assume that closing stock index of tomorrow is higher than today if today's index is higher than that of yesterday. Whenever the prediction is rising, we buy at most shares of stocks. Otherwise, we sell all stocks we have. This model performs well when stock markets go smoothly. But it losses a lot when the markets fluctuate or shocks frequently.

3) Proposed Trading Model We use the prediction results from our SVM learner. The trading principal is the same as Benchmark model 2. That is, we buy stock when prediction is positive and cash all stocks we have when prediction is negative. C. Simulation Results The profit of the three models during the five period is shown in Table 5 below. From the table, we can see that during most of periods, our proposed model wins the most profits. On average, our model gains 814.6 dollars as profits for every 50 days. That is 8% return rate in 50 days. Therefore, we can reach annual interests at about 30%. Besides high profit, our model also has the advantage of low risk.

4) Our model seldom loses in trading period while benchmark model 1 and benchmark model 2 loses in period 3 and 5. Actually, in most cases, our model can get at least 5% profits in the 50 day long trading period. Although we can reach high profit and low risk in our simulation, we shall still remember that we did not take tax .

5. CONCLUSIONS

In the project, we proposed the use of data collected from different global financial markets with machine learning algorithms to predict the stock index movements. Our findings can be summarized into three aspects: 1. Correlation analysis indicates strong interconnection between the US stock index and global markets that close right before or at the very beginning of US trading time. 2. Various machine learning based models are proposed for predicting daily trend of US stocks. Numerical results suggests high accuracy 3. A practical trading model is built upon our well trained predictor. The model generates higher profit compared to selected benchmarks. There are a number of further directions can be investigated starting from this project. The first one is to explore other creative and effective methods that might yield even better performance on stock market forecasting. Second, models can be modified to take care of the tax and fees in the trading process. Finally, we can investigate the

mechanism of short sale and maximize our profit even when the market is bullish.

REFERENCES

[1] W. Huang et al., "Forecasting stock market movement direction with support vector machine," Computers & Operations Research, 32, pp. 2513–2522005, 2005

[2] J. Moody, et al., "Learning to trade via direct reinforcement," IEEE Transactions on Neural Networks, vol. 12, no. 4, Jul. 2001.

[3] www.wikiposit.org

[4] S. Zemke, "On developing a financial prediction system: Pitfall and possibilities," Proceedings of DMLL-2002 Workshop, ICML, Sydney, Australia, 2002. [5] Vatsal H. Shah, "Machine learning techniques for stock prediction," www.vatsals.com.