A LINEAR REGRESSION APPROACH TO PREDICTION OF STOCK MARKET TRADING VOLUME: A CASE STUDY

Farhad Soleimanian Gharehchopogh[1], Tahmineh Haddadi Bonab[2] and Seyyed Reza Khaze[3]

[1]Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran

[2]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, West Azerbaijan, Iran

[3] Department of Computer Engineering, Dehdasht Branch, Islamic Azad University, Iran,

ABSTRACT

Predicting daily behavior of stock market is a serious challenge for investors and corporate stockholders and it can help them to invest with more confident by taking risks and fluctuations into consideration. In this paper, by applying linear regression for predicting behavior of S&P 500 index, we prove that our proposed method has a similar and good performance in comparison to real volumes and the stockholders can invest confidentially based on that.

KEYWORDS

Stock Market, Stock index, S&P 500, Data Mining, Regression, Dataset

1. INTRODUCTION

Predicting the stock market due to its importance and popularity among the masses and also small and large companies due to financial benefits and its low risk is a growing topic in research [1]. Despite the risk of falling too much value per share due to market fluctuations rarely happens, but again, the risk is there. These fluctuations which effect on stock price and trading volume have some difficulties in predicting. The fluctuations effect on the behavior of people in terms of capital savings or investment, the stock price and the increase or decrease of risk for investors. Therefore in general, predicting the stock market behavior through techniques and various methods is a useful tool to assist investors to act with greater certainty and taking the risks and volatility of an investment into consideration and know when to buy the cheapest price and when to sell to highest price [2].

Data mining techniques have a more successful performance in predicting various fields such as economy, policy and engineering compared to traditional statistical methods by discovering hidden knowledge of data [3, 4, and 5]. Experience has shown that machine learning techniques could be successful in predicting daily stock price and its trading volume [6]. In this paper, at first review the prior researches done for predicting stock market and then we describe the importance of trading volume. By using linear regression we predict S&P 500 index [7] behavior and at the end we compared and evaluated the result of our proposed method with other approaches.

2. PREVIOUS WORK

Nowadays the stock market has been called for research in many fields due to its effects on financial challenging and capacity of predicting its various aspects through different scientific methods such as genetic algorithm, Artificial Neural Network (ANN) and other Meta heuristic algorithms. Many institution and academic researchers are trying to propose a method for predicting next day behaviors of stock indexes in order to be better than the other methods, like a research that Majhi and other friends [8] did via applying bacterial foraging optimization technique for predicting stock market and S&P500 indexes in short and long terms, and they made a linear combiner model which its weights updated by BFO and comparing it with Multi-Layer Perceptron (MLP) based method showed that Majhi and other friend's method has less calculative complexity and more precision to MLP method.

Another predicting system [9] in which counting of complex keyword topples and its transformation to predict stock market behavior periodically and doing real-time forecasting on web has been done. Some researchers used text mining approach [10], their findings investigates effects of financial news in predicting stock market. Increasing social networks and their popularity among people have been led into new ideas of investigating of effect of the popularity and application of these social networks that can have on stock market behavior.

Like a work about effect of emotions like hope, fear and worry have on increasing or decreasing amount of Dow Jones on the next day [11] or investigating effects of Facebook [12] on stock market. The relation between the tendencies of investors and activities of stock market found by using a new time scale which operates on updated mood of about 100 million American Facebook users between the periods of 10/09/2007 to 10/09/2010. In this paper, predicting of trading volume is considered which is similar to the introduced methods.

3. IMPORTANCE OF TRADING VOLUME

Stock market is one of the first options of attracting investments and main financial indexes of country [13]. One of the most important parameters affecting the dynamics of the stock market is its trading volume. Stock trading volume includes the number of lots bought and sold which is expressing in daily basis [14]. The more trading volume of a stock is higher, the more the stock is active. Trading volume is an approving to price patterns in technical analysis and it's more important than stock price. If we could predict the moving direction of trading volume of a stock in the future, we can also obtain the prices changing, continuing or finishing of its trend, with more confidence [15].

3.1. LINEAR REGRESSION

Regression predicts a numerical value [16]. Regression performs operations on a dataset where the target values have been defined already. And the result can be extended by adding new information [17]. The relations which regression establishes between predictor and target values can make a pattern. This pattern can be used on other datasets which their target values are not known. Therefore the data needed for regression are 2 part, first section for defining model and the other for testing model. In this section we choose linear regression for our analysis. First, we divide the data into two parts of training and testing. Then we use the training section for starting analysis and defining the model. Scatter plot of 80% out of data has been shown in (figure 1) with taking this into consideration that the (Average) parameter is the mean of the prices of Open, Low, High and close. Scatter plot has been shown with just the Average parameter in order to be simpler.

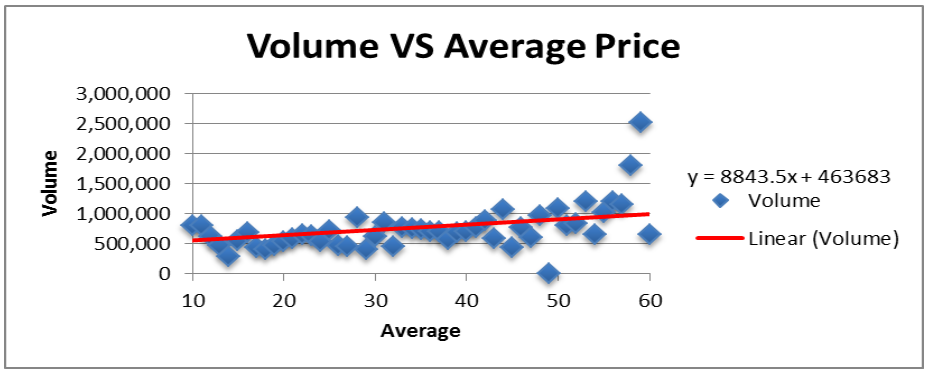

Figure1. Scatter plot of trading volume

Relationship between trading volume (Volume) as the dependent variable and the average price per share (Average) as the independent variable of the regression equation $(Y = 8843x + 46368)$ including the R-squared ($R^2 = 0.358$) that has been calculated in (table 3) and shown with a red trend line in (figure 1). R-squared R^2 shows that the two variables were used for determining the orientation of trend line is 35.8% related. This value is used for analysis based on scatter plot of (figure 1). For the first step correlations or relationship between desired independent parameters for specific the relation between stock prices according to that it is in Open, Low, High, Close and Volume status is obtained, as shown in (table 2). Coefficients have been calculated to 3 decimal (the relations are between 80% out of the whole data).

Table2. The relationship between independent parameters

	Open	***Close***	***Low***	***High***
Lose	0.959	-	-	-
Low	0.975	0.989	-	-
High	0.989	0.976	0.985	-
Volume	-0.383	-0.429	-0.425	-0.391

As it is obvious from the relationship chart of table 2, the relationship of 4 values of S&P 500 indexes are close to 1, and also the independence value of the dependent variable of Volume with other 4 prices is close to 0. By using data analysis that is one of the facilities of Excel which is used for financial analysis and defining predicting patterns, linear regression applied to the data. Summary output of applying the regression analysis has shown in (table 3). (Coefficients are considered simple).

Table3. Achieved regression values by applying regression analysis

Regression Statistics	
Multiple R	0.599
R Square	0.358
Adjusted R Square	0.347
Standard Error	285577
Observations	59

The value of multiple R is 0.599 or 0.6. This value is close to 1 which means that the regression line along with least square value is appropriate and well-adjusted to data. $R\ square = 0.357$

and $Adjusred\ R\ square = 0.347$, this values are close to 0 which means that the average and volume points are close to the trend line shown in figure 1. Since we use linear regression and take independent parameter of average into consideration. Therefore R square value is the value of R^2 . The standard error is equal to 285577 which is the error between real values and estimated value of volume has been calculated from summation of all residual values along with degree of freedom, sum and mean of squares shown in table 4.

Table4. Analysis of variation of linear regression

ANOVA	*df*	*SS*	*MS*
Regression	1	3E+12	3E+12
Residual	57	5E+12	8E+10
Total	58	7E+12	

Table5. Coefficients obtained from applying linear regression

	Coefficients	*Standard Error*
Intercept	4675513	697440
Average	-106938	18953

Using coefficients obtained from the figure 5, correlation of linear regression obtained as:

The linear regression of trading volume = 4675513-106938 * Average

After obtaining coefficients, slope, error and intercept and applying linear regression on sample data, for testing that how much close the formula can predict the trading volume (which is our unknown parameter) to real volume, we applied this formula on the rest 20% of data. The results shown in table 6 obtained (some samples of results have been presented).

Table6. Results of the applying regression formula

Date	*Average Price*	*Predicted Volume*	*Volume*
28/06/2103	33.13	$965675	$1,081,200
27/06/2013	32.96	$963498	$801,800
26/06/2013	32.49	$1019301	$835,100
25/06/2013	32.34	$1042679	$1,196,700
24/06/2013	32.87	$985747	$656,100
21/06/2013	32.94	$964849	$1,017,800
20/06/2013	33.11	$955342	$1,196,100
18/06/2013	33.53	$921192	$1,156,600
17/06/2013	33.52	$919738	$1,794,100
14/06/2013	33.52	$918908	$2,512,100
07/06/2013	35.88	$758848	$645,000

As it's obvious from table 6, the predicted trading volume is very similar to real values. By computing the difference between real and predicted values of proposed approach shown in figure 6, similarities of 61.35% observed

5. DISCUSSION

Financial markets such as stock market are generating constantly great volume of information needed to analysis and to produce any predicting pattern in any time. Therefore they are interesting case of using different scientific methods to development and improvement in generating techniques. Each of the used techniques for predicting financial matters has some benefits and limitations of its own which causes to some weakened or strengthened status. With taking this matter into consideration that our study is a case study on S&P 500 index to compare with other techniques, we checked out 8 most important features for predicting methods. The first feature is ease of encoding, which our method is equal to rule induction and has high degree in it and it is better than ANNs and genetic algorithm.

Second feature is accessibility or availability of off-the-shelf software that for this feature it is equal to rule induction, statistical inference and ANN and has high degree in it. Third feature is flexibility or ability of covering different types of large scales is equal to statistical inference and genetic algorithm and has medium degree in it. Fourth feature is autonomy or independence of prior assumptions from relations between variables and domain theories is equal to rule induction and statistical inference and is weak in this feature.

Fifth feature is optimization capability which tries to generate optimized results, in this feature is equal to genetic algorithms and has medium efficiency in it. Sixth and seventh features are operative complexity and cost of calculation in generating the results which has medium degree in them. In the case of eighth feature which is interpretability or ability of explaining results has high degree like rule induction and data visualization [18].

7. CONCLUSIONS

Each clustering algorithms are solely capable of focusing on particular parts of customers' data in electronic shops. This focus brings better and more detailed results to the same parts. Meanwhile, in analysing other parts, due to the lack of clustering analyses, it brings challenges to them.

So, each algorithm is capable of doing detailed analyses of some parts of customers' data. To provide comprehensive results and clustering analyses, it must be used several integrated and clustering algorithms. We, in this paper, investigate different types of methods and clustering algorithms. Finally, by using K-means, farthest first, EM samples of customers of an E-commerce websites, we made clustering vie Weka software.

We indicated that each algorithm covers the clustering analyses weaknesses of other algorithms for some customers. The integrated data of all algorithms analyses brings detailed results from customers' behavioural method and its relation with shopping basket as well. So, by using integrated collective data, it can be determined marketing policies and customer satisfaction appropriate to all customers' clustering and their orientation which finally lead to increased productivity and incomes.

7. REFERENCES

[1] Enke, D., & Thawornwong, S. (2005), The use of data mining and neural networks for forecasting stock market returns, *Expert Systems with Applications*, *29*(4), 927-940.

[2] oone, L., Giorno, C., & Richardson, P. (1998), *Stock market fluctuations and consumption behaviour: some recent evidence* (No. 208). OECD Publishing.

[3] Gharehchopogh, F.S., Mohammadi, P., & Hakimi, P. (2012). Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study. International Journal of Computer Applications, 52(6), 21-26.

[4] Gharehchopogh, F. S. (2011). Approach and Developing Data Mining Method for Spatial Applications. In *Proceedings of International Conference on Intelligent Systems & Data Processing (ICISD), India* (pp. 342-345).

[5] Gharehchopogh, F.S., & Khaze, S.R. (2012), Data Mining Application for Cyber Space Users Tendency in Blog Writing: A Case Study. *International Journal of Computer Applications*, *47*(18), 40-46.

[6] Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. Wiley. com.

[7] http://au.finance.yahoo.com/q/hp?s=ASX.AX, Last available: 28/06/2013

[8] Majhi, R., Panda, G., Sahoo, G., Dash, P. K., & Das, D. P. (2007, September). Stock market prediction of S&P 500 and DJIA using bacterial foraging optimization technique. In *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on* (pp. 2569-2575). IEEE.

[9] Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998, October). Daily stock market forecast from textual web data. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on* (Vol. 3, pp. 2720-2725). IEEE.

[10] Nikfarjam, A., Emadzadeh, E., & Muthaiyah, S. (2010). Text mining approaches for stock market prediction. In *Computer and Automation Engineering (ICCAE), 2010 the 2nd International Conference on* (Vol. 4, pp. 256-260). IEEE.

[11] Zhang, X., Fuehres, H., & Gloor, P. A. (2011). *Predicting stock market indicators through twitter "I hope it is not as bad as I fear".* Procedia-Social and Behavioral Sciences, 26, 55-62.

[12] Karabulut, Y. (2011). Can Facebook predict stock market activity? *Available at http://bus.miami.edu/umbfc/_common/files/papers/Karabulut.pdf* [last Available 02.07.2013].

[13] Demirguc-Kunt, A., & Levine, R. (Eds.). (2004). *financial structure and economic growth: A cross-country comparison of banks, markets, and development*. MIT press.

[14] Rouwenhorst, K. G. (1999). Local return factors and turnover in emerging stock markets. *The Journal of Finance*, *54*(4), 1439-1464.

[15] Pesaran, M. H., & Timmermann, A. (1994). Forecasting stock returns an examination of stock market trading in the presence of transaction costs. *Journal of Forecasting*, *13*(4), 335-367.

[16] Gharehchopogh, F. S., & Khalifehlou, Z. A. (2012). A New Approach in Software Cost Estimation Using Regression Based Classifier. AWERProcedia Information Technology and Computer Science, Vol: 2, pp. 252-256.

[17] Draper, N. R., Smith, H., & Pownell, E. (1966). *Applied regression analysis* (Vol. 3). New York: Wiley.

[18] Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: data mining in financial application. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *34*(4), 513-522.

Authors

Farhad Soleimanian Gharehchopogh is currently Ph.D. candidate in department of computer engineering at Hacettepe University, Ankara, Turkey. And he works an honour lecture in computer engineering department, science and research and Urmia branches, Islamic Azad University, West Azerbaijan, Iran. He is a member of editorial board and review board in many international journals and international Conferences. His interested research areas are in the Operating Systems, Software Cost Estimation, Data Mining and Machine Learning techniques and Natural Language Processing. For more information please visit www.soleimanian.net

Tahmineh Haddadi Bonab is a M.Sc. student in Computer Engineering Department, Science and Research Branch, Islamic Azad University, West Azerbaijan, Iran. Her interested research areas are Meta Heuristic Algorithms, Data Mining and Machine learning Techniques.

Seyyed Reza Khaze is a Lecturer and Member of the Research Committee of the Department of Computer Engineering, Dehdasht Branch, Islamic Azad University, Iran. He is a Member of Editorial Board and Review Board in Several International Journals and National Conferences. His interested research areas are in the Software Cost Estimation, Machine learning, Data Mining, Optimization and Artificial Intelligence.