IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 11 (November. 2013), ||V1|| PP 42-48

Stock Portfolio Selection using Data Mining Approach

Carol Anne Hargreaves, Prateek Dixit, and Ankit Solanki
Institute of Systems Sciences, National University of Singapore
25 Heng Mui Keng Terrace, Singapore 119615

Abstract: **-** Once it is decided that investment is to be made in the stock, the obvious question which arises is: which all stocks should be purchased? Past performance will not guarantee the future, but it is still worthwhile to evaluate the investments based on their ability to deliver consistent returns with minimal risk. Therefore, the ability to generate most profitable return from short term stock trading is a crucial factor for traders and investors. The paper focuses on a comparative study between two data mining techniques, Logistic Regression and Neural Network for stock portfolio selection using a set of fundamental and technical parameters.

Keywords: - *Data Mining, Logistic Regression, Neural Network, Stock Portfolio Selection, Stock Trading*

I. INTRODUCTION

Stocks have always been a hot topic of discussion in the financial realm, but the difficulty in understanding the real meaning of stocks still persists. Whenever an investor thinks about investing in stocks he comes across a large pool of stocks, and an important task of his investment decision process is the selection of stocks. It is important for traders and potential investors to have relevant financial information, which will enable them to make a good investment decision in the stock market. For example, there are more than 2000 stocks listed on the Australian Stock Exchange. Selecting a few stocks from all the listed ones requires dealing with innumerable data records, and this work cannot be done by a human brain in a short span of time. Therefore, sophisticated techniques are required to analyze the data and combine the distinct information, which in turn, can potentially affect the returns generated by the stocks.

There isn't any one definitive method for assigning value to a company's stock. The study of advancement in artificial intelligence and its associated technologies to reduce the human efforts have gained an unparalleled importance in research. Applications of machine learning and statistical techniques to automate and improve processes in finance domain have started more than a decade ago.

Not all stocks in a particular market or sector yield a profit, hence, the selection of an optimum set is essential to get the best market results. To solve the problem, nearly all professionals subscribe to one of the two generally accepted approach of stocks analysis: fundamental analysis and technical analysis. So which method of analysis is superior? We believe the answer lies in the effective blending of both the approaches.

Fundamental analysts examine the facts affecting the company's value and growth by following a top-down approach that starts with an economic forecast, group selection, narrowing within the group and company analysis. Whereas technical analysts look at the past for the answers to the future as it helps investors to anticipate what is likely to happen to prices over time by following the trends in the market, often represented in charts that shows price over time like relative strength index, momentum or stochastic oscillator.

The Australian market is chosen as the economic domain to conduct the experiments as Australian market is well placed on global economy and has been ranked 8th by MSCI (Morgan Stanley Corporation International) Global Index Rating. Encouraged by its healthy political and economic position, with a relatively high growth, low inflation economy, and largest trading partners, Australia has emerged as one of the strongest performer among the developed countries during this period.

The main inspiration for the present study is a research paper [1], based on which further modifications have been done along with implementation of new strategies, ideas and concepts. In the present work different models are used for portfolio selection and a completely different trading strategy is employed. The models are run over two sectors (healthcare and financial] of the Australian stock market in different time periods over the year 2013 to test/validate the repeatability of models.

The main purpose of this paper is to develop methods which will help investors to find the strongest of the strong stocks and get maximum returns in a short period of time. Data mining techniques for optimization on portfolio selection are studied in the present work to meet the challenge of good investment decision. An experimental approach has been used by conducting four experiments using the following hypothesis:

- Healthcare sector stock portfolio selected using neural network will outperform the All- Ordinaries Index and Healthcare Sector Index (XHJ) over the twenty days trading period.
- Healthcare sector stock portfolio selected using logistic regression will outperform the All-Ordinaries Index and Healthcare Sector Index (XHJ) over the twenty days trading period.

- Financial sector stock portfolio selected using neural network will outperform the All-Ordinaries Index and Financial Sector Index (XFJ) over the twenty days trading period.
- Financial sector stock portfolio selected using logistic regression will outperform the All-Ordinaries Index and Financial Sector Index (XFJ) over the twenty days trading period.

In the remainder of the paper, the methodology for the present study is explained by discussion of stock selection system, followed by the trading strategy employed. The paper is concluded with the outcomes and the value of this study with possible further research ideas.

II. LITERATURE REVIEW

Data mining is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.

In previous studies, some heuristics based methods were developed to solve the portfolio selection problem. Some of these methods use Genetic Algorithm [2], Swarm Intelligence [3], Local Search [4], Tabu Search [4] and Simulated Annealing [4] to solve the portfolio selection problem. Here we present a comparative study between two different heuristic methods Neural Network and Logistic Regression to solve the problem of portfolio selection focused in the domain of stock market.

When there are no prior expectations about what the data will show, a neural network model may be used to explore possible correlations. Neural Network models are often used for exploration because they can analyze complex relationships between many inputs and outputs. Linear Regression for data with continuous outcomes in (0, 1) or binary outcomes in {0, 1} may not be appropriate. Therefore, Logistic Regression (LR) is an alternative regression technique naturally suited to such data [5].

2.1 Neural Network

Neural Network (NN) is one of the most attractive techniques which is used by finance and investment industries to deal with large amount of data [6]. The human brain can solve complex problems because it can learn and generalize complex problems. Neural Network tries to mimic this nature of the human brain. The main characteristic of neural networks is their ability to learn. The learning process is achieved by adjusting the weights of the interconnections according to some applied learning algorithms. The weights in a Neural Network are the most important factor in determining its output. While learning the data set provided, Neural Network tries to find the best fit for the data set where the error will be minimized.

The Neural Network model learns through the training data with the help of supervised and unsupervised learning, two most commonly implemented learning strategies [7]. Neural Network is especially suited for simulating intelligence in pattern recognition, association and classification activities. These problems frequently arise in areas such as credit assessment, security investment and financial forecasting [8]. For the current research the supervised learning approach is used to build a model that learns from the dataset and selects the portfolio of stocks that will yield maximum profit.

Multilayer Perceptron is used in the current study using back propagation learning algorithm which is also being applied in many other fields. It consists of a set of inputs where the data is fed to the Neural Network and this layer is called an input layer. The data from the input layer is fed to one or more hidden layers and the hidden layer acts as a source to the output computational node [6, 9].

2.2 Logistic Regression

A wide variety of classification algorithms exist in the literature such as Support Vector Machines, K-Nearest Neighbour, Decision Trees, Bayes Classifier [10] which are well understood and widely applied, but the present work mainly focuses on a logistic regression model for data mining and high dimensional data classification. Regression analysis begins with a historical data set in which the target values are known to generate the valuation function. After the valuation functions are determined, they can then be applied to other datasets as a part of the prediction.

The relationship between target and input variables is not always a straight line, so a non-linear or logistic regression model is used. Logistic Regression is a technique for making predictions when the dependent variable is dichotomy, and the independent variables are continuous and/or discrete. The advantage of Logistic Regression is that, through the addition of an appropriate link function to the usual linear regression model, the variables may either be continuous or discrete or any combination of both types and they do not necessarily have normal distributions [11].

Logistic Regression, which is perfect for situations where the aim is to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables, is a multivariate analysis model [12]. It is particularly useful where the dataset is very large, and the predictor variables do not behave in orderly ways, or obey the assumptions required of discriminant analysis [13]. Logistic Regression analysis does

not require the restrictive assumptions regarding normality distribution of independent variables or equal dispersion matrices nor concerning the prior probabilities of failure [14].

III. METHODOLOGY

The method used in this study is more from a top down perspective. Fig. 1 provides an overview of the methodology placed to examine the models over each sector.

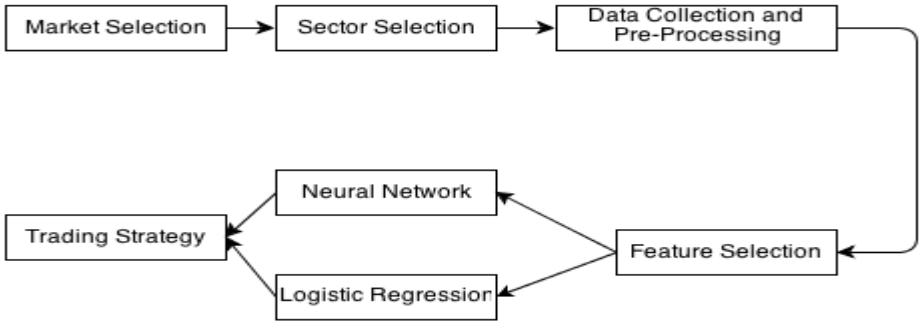

Fig. 1 Framework of stock selection process

3.1 Data Collection & Pre-Processing

To start the stock selection process, a particular sector is to be chosen where the present study can be initiated because the entire Australian market has more than 2000 stocks, and it would have been a difficult task to compare each and every stock from such a huge repository. Therefore, as an iterative process two sectors are chosen, i.e., healthcare and finance, to test the repeatability of the models employed over different time periods.

After a particular sector is chosen, stock count gets reduced from about 2000 to 100-200 which makes it a bit easier to work on the dataset. Historical data and fundamental parameters are collected for respective sector from Yahoo Finance. The collected data is then cleaned and certain pre-processing activities like normalization, handling missing values etc. are performed for all stocks representing that sector to run all the models to select the stock portfolio. Cleaning and other pre-processing activities resulted in reducing some more set of stocks which reduced the daunting task of mining humongous data.

3.2 Feature Selection:

Because there are many extracted variables, and it is not clear which of the extracted variables are relevant for the models employed, it is planned to use some selected variables. Therefore, the Random Forest Importance Algorithm [15] is implemented for feature selection. On implementing the algorithm a set of variables is found in each of the sectors mentioned above with a corresponding score attached to each variable which defines its importance in predicting the output. To verify how many variables should be included as independent variables to obtain the predicted output as close as possible to actual output, the stock portfolio selection models are run on one sector with a set of variables multiple times. In this process we tested several set of combinations of variables in their decreasing level of importance. It is found that on implementing the models with 10 variables into account is giving out the best possible results. One interesting observation is that both the sectors had certain common features which showed high importance for both the sectors. Table 1 shows the list of the independent variables selected for both the healthcare and finance sectors.

Table 1 Selected independent variables of each sector

S.No	Healthcare sector	Financial sector
1	Return on assets	Growth
2	Return on equities	Earnings per share
3	Earnings per share	Book value
4	Revenue per share	Return on equity
5	Quarterly revenue growth	Revenue per share
6	Growth	Analyst opinion
7	Enterprise value per EBITDA	Price per sales
8	Price per book	Enterprise value per revenue
9	Profit margin	Return on assets
10	Price per sales	Current ratio

3.3 Portfolio Selection Models

As discussed above in literature review (2), Neural Network and Logistic Regression models are trained for portfolio selection.

3.3.1 Neural Network

The fact that Neural Networks are inspired by the human brain makes it an efficient methodology for forecasting and decision making. The Multilayer Perceptron Neural Network is trained with the independent variables which were selected by feature selection algorithm.

Conventionally, classification of the training output is done visually by identifying the trend followed by a particular stock in the recent past [1] but in the present study conventional approach coupled with a data-driven approach is followed, in which each of the previous four weeks growth is taken as an important parameter to classify the stock either as 'good' or as 'bad'. Back propagation learning algorithm is employed using two hidden layers to compute the preferred weights for the nodes. Output obtained was not a binary output rather it was a series of continuous values which can be termed as a value corresponding to strength of the stock. Therefore, the top six – preferred stocks for this experiment, are selected based on the strength of the output.

3.3.2 Logistic Regression

The same parameters, as taken for the neural network, are taken for training the Logistic Regression model. The stock strength values are then predicted using the generated model from the training dataset. The output is given as probability score which has range from 0 to 1. If the predicted probability score is more than 0.5 in this study, then those stocks are considered to have a rising trend and make a profit in future. Therefore, as a final result, six stocks with the highest probabilities according to the output of this model are selected again.

3.4 Trading Strategy

Once the stock portfolios are decided, which are expected to give a profitable return as per the two proposed approaches, it becomes a challenge to evaluate a trading strategy (appropriate time to buy and sell) which works in investor's favour. In this study a custom trading strategy is created which will explore various exit methods.

In the trading strategy used in this work, Relative Strength Index (RSI) is selected as a technical indicator to evaluate the favourable buying and selling period, which is usually used for evaluating the stock to be overbought or oversold. In conducting this study 10000 AUD is virtually invested in each of the stocks which sum up to 60000 AUD for each model to compare the result at the end of the trading period on the basis of profitability achieved. Fig. 2 describes the flow diagram for the trading strategy.

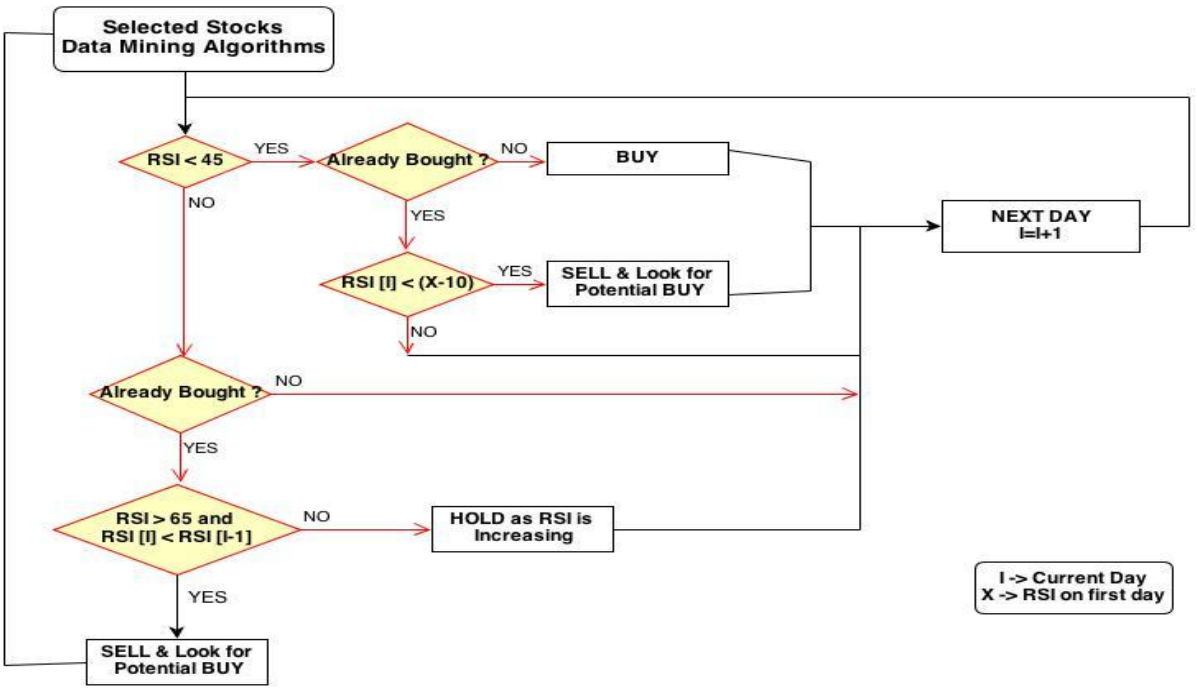

Fig 2 Diagrammatic representation of the flow of our Trading Strategy

As seen in Fig. 2, a stock is bought only when its RSI is below 45, else it is advisable to wait and not buy the stock. When the stock crosses the RSI value of 65 careful monitoring is done since it is then considered as a signal for selling the stock. The stock is kept on hold till the RSI value is gradually increasing, even if it is above 65, but the stock is sold as soon as there is a drop in the RSI value.

At the end of twenty days trading period, the value amount of each stock for each of the models is combined to compare the profitability achieved for each model. This serves as the most important parameter to show results of the present study from the business perspective.

IV. RESULTS & DISCUSSION

The above discussed strategies are implemented and the selected stocks are traded for twenty days trading period. Performance of the model is measured in terms of profit gained while trading with a particular portfolio. The current set of portfolios of each model for each sector is compared with All Ordinaries Index and respective Sector Indexes to analyse the results of each model. Table 2 summarises the percentage change in the invested value for each of the sectors.

Table 2: Percentage Value Change for each Sector

Sector	Neural Network	Logistic Regression	Sector Index	All Ordinaries
Healthcare	18.05%	18.24%	-1.60%	-2.67%
Financials	2.29%	-0.80%	-7.61%	-3.59%

Table 2 describes the percentage change in the amount of invested value for each of the sectors during the twenty day trading period. It can be inferred from Table 2 that the models used in this study clearly outperformed the All Ordinaries and respective Sector Indexes. Given below is a clear description of each of the sectors, i.e., healthcare sector and financial sector.

4.1 Healthcare Sector

At the end of the trading period for healthcare sector, neural network portfolio showed a profit of 18.05% whereas logistic regression showed a profit of 18.24%. Fig. 3 shows the variation of percentage change for both Neural Network and Logistic Regression models, which also outperforms the All Ordinaries Index and Healthcare Index during the course of twenty days trading period.

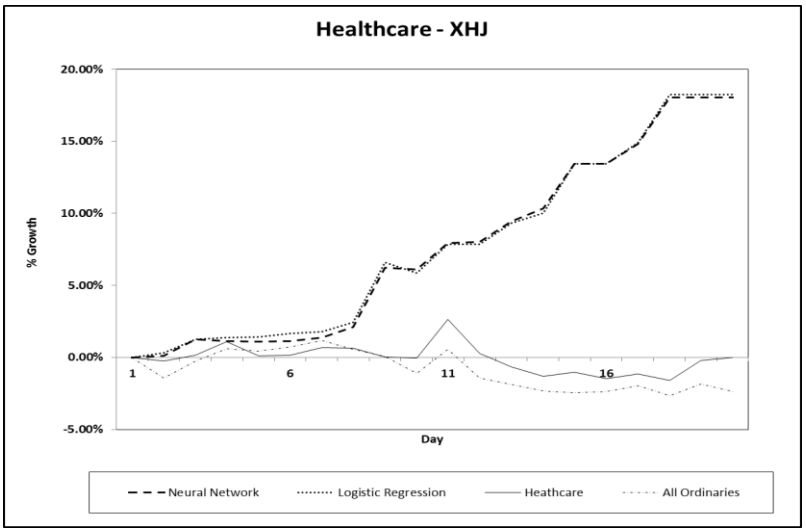

Fig 3 Comparison of data mining models v\s AOI & XHJ for healthcare sector

Fig. 3 shown above compares the performance of models employed over All Ordinaries and Healthcare Index. It can be clearly observed from Fig. 3 that at the end of the trading period in Healthcare Sector, Neural Network has shown a rise of 20.72 and 19.65 percentage points with respect to All Ordinaries index and Healthcare Index respectively. Similarly, Logistic Regression has shown a rise of 20.91 and 19.84 percentage points with respect to All Ordinaries index and Healthcare Index, respectively.

4.2 Financial Sector

During the trading period for financial sector, Australian market was suffering from a loss of 3.76% and Financial Index was suffering from a loss of 6.42%. But during the same period stock portfolio selected by

Neural Network showed a profit of 2.29% but on the contrary Logistic Regression showed a minor loss of 0.80%. Fig. 4 shows the percentage change for both Neural Network and Logistic Regression models and for both the indexes, i.e., All Ordinaries Index and Financial Index, over the course of twenty days trading period.

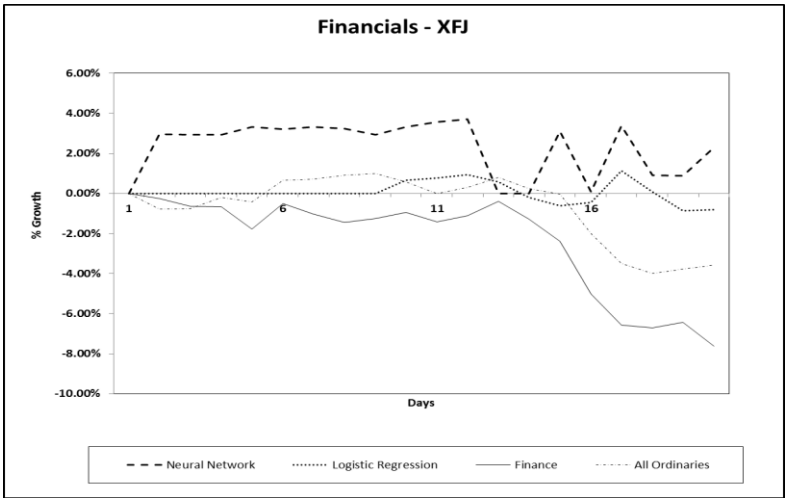

Fig 4 Comparison of data mining models v\s AOI & XFJ for financial sector

Fig. 4 shown above compares the performance of the models employed over All Ordinaries Index and Financial Index. It is clearly observed that the models employed in the present work outperformed both All Ordinaries Index and Financial Index, though one of the models used showed result at a minor loss. In Financial Sector at the end of trading period, Neural Network is showing a rise of 5.88 and 9.90 percentage points with respect to All Ordinaries Index and Financial Index, respectively. Similarly, Logistic Regression is showing a rise of 2.79 and 6.81 percentage points with respect to All Ordinaries Index and Financial Index, respectively.

Thus, Fig. 3 and Fig. 4 validates that the experiments performed justifies the hypothesis presented in this study, i.e., data mining models employed do outperform the market indexes and in turn helps the investors either in gaining fancy monetary returns or in minimizing the loss.

A few constraints were faced while performing the experiment. One was that a lot of missing data are found for many shares which are imputed according to the median values at the end. But due to this some of the prominent shares could not be selected which might have affected the results predicted by the models employed in this work, and in turn, have hindered in bringing much better expected results. The data size was also too small to run the data mining models efficiently. With a larger set of data, predictions done by using data mining models might have been better.

V. CONCLUSION

It is learnt from literature that stock returns may be predicted by various financial and economic variables. This study has employed both the data mining models [Neural Network and Logistic Regression] to determine the preferred stocks for getting most profitable return. Major improvement is observed in prediction of stock returns as both fundamental and technical parameters are used for analysis in the models.

The outcomes in this study justify that each portfolio of stocks predicted from different models employed performed better than Australian All Ordinaries Index and respective Sector Index by maximizing the profits and minimizing the losses. It is also found that in healthcare sector both the models gave an average profit return of more than 18%. In financial sector also, when the market was undergoing huge dip, the models and strategies used in this study gave out an average profit return of almost 1%. The study shows that Neural Network is more consistent in predicting the complex patterns of stock market as compared to Logistic Regression, because Neural Network was performing better than Logistic Regression and giving out profitable returns even in hard situations of the market. The reliability of the success of the models employed and trading strategy used in this study is ensured by testing the hypothesis in different sectors and for different time period.

Many more hybrid models such as ANFIS (Adaptive Neuro-Fuzzy Inference System), Genetic Algorithms coupled with Neural Network and Support Vector Machine (SVM) may also be employed to perform future research for finding the best model to predict stock prices. Integration of all the various components of our framework may be performed to make a suitable product which can be used by investment firms or traders on a commercial scale.

REFERENCES

[1] Carol H and Yi H, Does the use of Technical & Fundamental Analysis improve Stock Choice?, *Int. Journal Statistics in Science, Business, and Engineering, Langkawi, 2012.*

[2] Konstantinos A and Georgios M Multiobjective evolutionary algorithms for complex portfolio optimization problems, *Computational Management Science, Vol. 8 (3),* 2009, 259-279.

[3] Deng G and Tsong W, Swarm Intelligence for Cardinality-Constrained Portfolio Problems, *Int. Proceedings of ICCCI, Kaohsiung, Taiwan, Vol. 3,* 2010, 406-415.

[4] Kellerer H and Maringer D, Optimization of cardinality constrained portfolios with an hybrid local search algorithm, *Journal of Heuristics,* 2001, 481-495.

[5] Paul K, *Logistic Regression for Data Mining and High-Dimensional Classification*, Research Showcase, Carnegie Mellon University, 2004.

[6] Yimin Y, *A comparison of Neural Networks for Stock Selection*, Master Thesis, Oklahoma State University, 1999.

[7] Suraphan T and David E, The adaptive selection of financial and economic variables for use with artificial neural networks, *Neurocomputing, Vol. 56,* 2004, 205-232.

[8] Robert T and Jae L, *Artificial Intelligence in Finance & Investment* (Irwin Professional Publishing, 1996).

[9] Karl N, *Stock Prediction – A Neural Network Approach*, Master Thesis, Royal Institute of Technology, Stockholm, Sweden, 2004.

[10] Hengshan W, and Phichhang O, Prediction of Stock Market Index Movement by Ten Data Mining Techniques, *Modern Applied Science, Vol. 3(12),* 2009.

[11] Shaoyan Z, Christos T, Xiaojun Z, Hong Q, Iain B and John K, Comparing Data Mining Methods with Logistic Regression in Childhood Obesity Prediction, *Information Systems Frontiers, Vol. 11(4),* 2009, 449-460.

[12] Lee S, Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using GIS, *Environmental Management, Vol. 34(2),* 2004, 223-232.

[13] Arun U, Gautam B and Avijan D, Prediction of Stock Performance in the Indian Stock Market using Logistic Regression", *International Journal of Business and Information, Vol. 7,* 2012, 105-136.

[14] Zavgren, C, Assessing the Vulnerability to Failure of American Industrial Firms: A Logistic Analysis, *Journal of Business Finance and Accounting, Vol. 12 (1),* 1985, 19–45.

[15] Breiman, L, Random Forest, *Machine Learning, Vol. 45 (1),* 2001, 5-32.