Prediction of Stock Performance Using Analytical Techniques

Carol Hargreaves
Institute of Systems Science
National University of Singapore, Republic of Singapore
Email: carol.hargreaves@nus.edu.sg

Yi Hao
School of Business Economics
Wilfrid Laurier University, Waterloo, Canada
Email: haox5200@mylaurier.ca

Abstract— **With an easy access to share information and data nowadays, many investors worldwide are interested in predicting stock prices. The prediction of stock prices using data mining techniques applied to technical variables has been widely researched but not much research to date has been done in applying data mining techniques to both technical and fundamental information. This paper is based on a personal approach to stock selection, using both technical and fundamental information. In this paper we construct a framework that enables us to make class predictions about industrial stock performances. In order to have a systemized approach for the selection of stocks and a high likelihood of the performance of the stock price increasing, several analytical techniques are applied. A trading strategy is also designed and the performance of the stocks evaluated. Our two goals are to validate our stock selection methodology and to determine whether our trading strategy allows us to outperform the Australian market. Simulation results show that our selected stock portfolios outperform the Australian All-Ordinaries Index. Our findings justify the use of analytics for classification and prediction purposes. Further, in conclusion, we can safely say that our stock selection and trading strategy outperformed the Australian Ordinary index.**

Index Terms— **Stock price prediction, stock selection, stock market, analytics, decision trees, neural networks, logistic regression, trading strategy.**

I. Introduction

Many financial companies such as stock markets produce large datasets and are looking to find efficient ways to discover useful information about stocks and the market for investment decisions. Further, with the easy access to stock information and data, many private investors worldwide are interested in predicting stock prices and hope to maximize on the opportunities in the market and become rich. The problem is mainly due to the following two reasons: There are so many stocks in the market, and large amounts of stock information easily available on the internet, through newspapers, magazines, radio and television. How can investors go about selecting stocks in a systematic way so that they select winning stocks and make a profit with minimum risk?

A. Aim of the Present Study

The focus of this paper is to investigate whether stocks selected by the application of a personal systematic approach or with several analytical techniques, when monitored and managed using a trading strategy, will outperform the Australian stock market.

More specifically, we would like to explore and extend the study in [11] and see whether the Logistic Regression model is also able to predict stock performance and assist with stock selection as other data mining techniques, such as CHAID, C5.0 and Neural Networks have demonstrated in [11].

B. Hypothesis

The following hypotheses will be tested in this study:

- H1: The stocks selected through a personal systematic approach outperforms the Australian stock market,
- H2: The stocks selected by decision tree methods,
- H3: The stocks selected by neural network methodology,
- H4: The stocks selected by logistic regression models,

H1, H2, H3 and/or H4 will outperform the Australian stock market.

In the remainder of the paper, we explain the methodology for our study, by discussing our stock selection system, followed by our trading strategy. We conclude our paper with the results and the value of our study with possible further research ideas.

II. Literature Review

A large number of research papers predict the pricing of the stock index as well as stock performance. Across a number of European markets (e.g., UK, France, and

doi:10.4304/jetwi.5.2.136-142

Germany), [7] observed that stock returns are predictable. The application of the logistic regression has repeatedly been used in the area of investments, corporate finance and banking. The logistic regression was used by [10] as a comparative method in order to build a better model for predicting stock returns effectively and efficiently. The effectiveness of neural network models, which are known to be dynamic and effective in stock market predictions were evaluated by [8]. The logistic regression technique yields coefficients for each independent variable based on a sample of data [12]. For a detailed technical explanation of the application of two decision trees applied (CHAID and C5.0) in this study, the reader is referred to [14] for C5.0 and [4] for CHAID. Existing literature indicates that very little research has been done on the use of analytical techniques to predict stock performance in the Australian stock market. In this context, the present study will provide valuable information to investors and enable them to make scientific decisions regarding investments and not make decisions on pure gut feel.

III. Research Objective and Methodology

Over the last few decades, increasingly huge amounts of past data have been stored electronically and this volume is expected to grow considerably in the future. The objective of this paper is to apply several analytical methods to stock data in order to classify the stocks into two categories, (1) the stocks that are likely to increase in price (2) the stocks that are likely to decrease in price, over the next 20 trading days. We used decision trees, neural networks and logistic regression to classify our stocks as to whether they will increase or decrease in price. These stock predictions assisted us in our decision of whether to buy a stock or not.

A. Decision Trees

Decision trees are a form of multiple variable analyses and are powerful and popular tools for classification and prediction. The attractiveness of decision trees lies in their ease of interpretation, relative power, robustness with a variety of data and levels of measurement, and ease of use. Decision trees attempt to find a strong relationship between input values and target values in a group of observations that form a data set [10]. In contrast to neural networks, decision trees represent rules and rules can readily be expressed so that humans can understand them or even directly used in database access language like SQL so that records falling into particular category may be retrieved [14].

In our study, the decision trees generated some rules of how the performance of different industrial company stocks were predicted as data became available.

B. Neural Network

Why use Neural Networks for forecasting stock price? There are numerous reasons why Neural Networks offer an advantage in the quest to forecast stock prices. Firstly, there is no need for any assumptions to be made used by Efficient Market Hypothesis (EMH) and no need for Normality assumptions. The EMH [5, 6] assumes that the price of a stock reflects all of the information available and that everyone has same degree of access to information and that whenever a change in financial outlook occurs, the market will instantly adjust the stock price to reflect the new information.

Many studies [1] have established that non-linearity exists in financial data and Neural Networks can be successfully used to model the relationship among this data. So the applications of Neural Networks to financial forecasting have become very popular over the last few years. Whether randomly or fully predictable, a correctly designed Neural Network will theoretically converge to an optimal result. Most other models do not have this luxury!

- The main advantage of Neural Networks is that they can approximate any nonlinear function to a degree of accuracy with a suitable number of hidden units
- Neural Networks can cope with “fuzzy patterns” – patterns that are difficult to reduce into precise rules.
- Neural Networks can be retrained and thus can adapt to changing market behaviour.
- Neural Networks can play a crucial role in deciding which technical variables to follow when analyzing past prices.
- Even when a data set is noisy or has irrelevant inputs, the networks can learn important features of the data.

Based on the above information, we ran a Neural Network with our Australian industrial stock data using a multi-layered perceptron Neural Network model, trained with a back propagation algorithm, using the Hyperbolic Tangent Solution.

C. Logistic Regression

Simple linear or multiple linear regression is applicable when the relationships between variables are assumed to be linear [3]. A number of nonlinear techniques can be used to obtain a more accurate regression if the relationship between variables is not linear in parameters. The logistic regression is preferred when the response variable takes on binary values (yes or no). It also has the advantage of being less affected when the normality of the variable cannot be assumed. Logistic regression also has the capacity to analyze a mix of all types of predictors [9]. For example, a logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these.

III. The Stock Selection Process and The Experiment

Our method is made from a top-down perspective. Figure 1 below, illustrates the path taken to arrive at our 6 stocks for selection consideration. Since there are more than 2000 stocks available in the Australian stock market, our first step is to narrow down the universe of stocks we want to choose from.

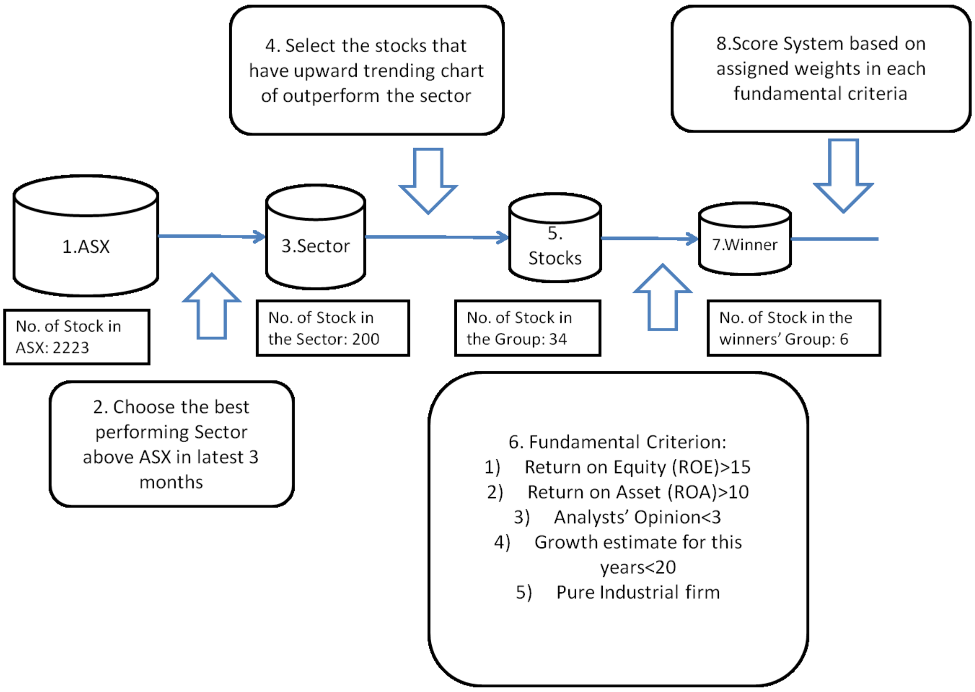
4. Select the stocks that have upward trending chart of outperform the sector
8.Score System based on assigned weights in each fundamental criteria
1.ASX
3.Sector
5. Stocks
7.Winner
No. of Stock in ASX: 2223
No. of Stock in the Sector: 200
No. of Stock in the Group: 34
No. of Stock in the winners' Group: 6
2. Choose the best performing Sector above ASX in latest 3 months
6. Fundamental Criterion:
1) Return on Equity (ROE)>15
2) Return on Asset (ROA)>10
3) Analysts' Opinion<3
4) Growth estimate for this years<20
5) Pure Industrial firm

Figure 1. Stock Selection Process.

To do this, we first identify the best performing sector by comparing all sector price trends against the ASX (Australian Ordinary) index for the latest 3 months using charts on the Australian stock market website [16]. In our experiment, the industrial sector turned out to have the best sector performance when compared to the Australian Ordinary Index. So, we instantly reduced our universe of stocks from 2223 to 200 as there are about 200 stocks in the Industrial sector. Of these 200 industrial stocks, 69 of them had an upward trend. Financial data on the 69 industrial stocks was then collected using Yahoo Finance [15].

The five stock selection strategies were applied in this study. One personal trading strategy, two based on Decision Trees (CHAID & C5.0) and one based on Neural Networks and one based on the Logistic Regression. Each selection strategy selected six stocks and roughly ten thousand dollars was invested in the market for each selected stock. To keep the input variables consistent, five variables were used for all of the strategies:

- Return on Equity
- Return on Assets
- Analyst Opinion
- Growth this year
- Price

In the next few sections we will describe each of the four stock selection processes. For each of the 5 stock selection processes, 6 winning stocks were selected from 69 industrial stocks to form 6 Stock Portfolios.

A. *Personal Trading Strategy (Stock Portfolio 1)*

We then selected all those industrial stocks that had an upward trend in the market and were outperforming the Australian Ordinary Index. There were 34 of these stocks. We then selected the best six stocks based on the following criteria:

- Return on Equity (ROE) >15%,
- Return on Assets (ROA)>10%
- Analysts' opinion < 3
- Growth estimate for the current year>20%
- Pure Industrial Firm

As a result, six stocks were selected for our personal strategy. Our scoring system assigned weights to stocks based on their fundamental ratios, price, analysts' opinion and growth estimates, where the sum of all weights is one. We next ranked our stocks from 1 to 6 with 1 as the highest preferred score to 6 as the lowest preferred score (for example, the higher the ROE the lower the rank). The final score is the sum product of the stock ranking and assigned weights. The assumption made is that the smaller the score, the more likely the stock will make profit.

B. *Price Trading Strategy (Portfolio 2)*

We ran a C5.0 decision tree with the five input metrics on the 55 industrial stocks. The 'price' metric was the most important. More specifically the rule generated was 'price >1.05'. So the six stocks that had the highest C5.0 probability of increasing their price and meeting this criterion the closest, was selected for this portfolio.

C. *Growth Trading Strategy (Portfolio 3)*

We ran a CHAID decision tree with the five input metrics on the 55 industrial stocks. The only metric that was important was 'growth this year', where the rule generated was 'growth this year >13.9%'. So the six stocks with the highest probability of increasing their price using the CHAID probability scores were selected for this portfolio.

D. All Mixed Strategy (Portfolio 4)

We ran a Neural Network with the five input metrics on the 55 industrial stocks. All 5 metrics were important; The most important metric being 'analyst opinion', followed by 'growth this year', followed by 'price' and then equally 'return on equity' and 'return on assets'. So the six stocks with the highest probability of increasing their price using the neural network probability scores, was selected for this portfolio.

E. Growth & Value Trading Strategy (Portfolio 5)

We ran a Logistic Regression with the five input metrics on the 55 industrial stocks. 3 metrics were important. The most important metric being 'return on equity, followed by 'price, followed by 'growth this year'. So the six stocks with the highest logistic regression probability scores of increasing their price, was selected for this portfolio.

IV. Trading Strategy

The goal of our trading strategy is to reduce the universe of stocks to a manageable few. We use technical analysis (charts) and fundamental analysis (financial ratios) to assist us in our selection of stocks with good indicators for growth & value respectively. After the stock selections process in Figure 1, we set the exit strategy as either a stock gain of 10% or more, or a loss of 5% or more. To evaluate whether our trading strategy performs well, we compare it with the Australian Ordinary Index.

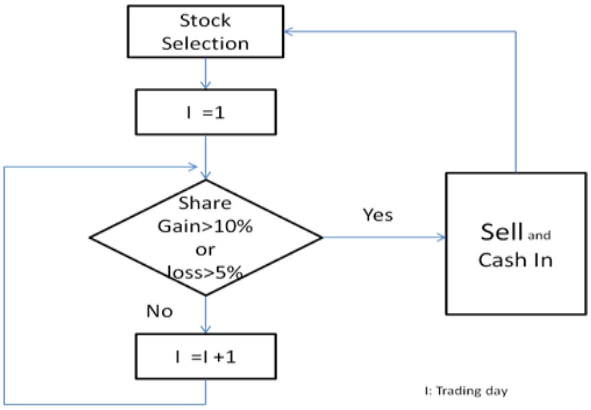

Figure 2. Trading Strategy Flowchart.

V. Results

Accuracy can be compared in many ways. In the simplest way, a decision tree is grown and the predicted classification is tested against the data set used to train the form of the tree. This is called a re-substitution test [4]. The performance of CHAID is determined by a test of significance of 0.05. For our CHAID, C5.0 and Neural network models, we obtained the following results. The results are shown in the Table 1, Table 2, Table 3, Table 4 and Table 5. The results are generated by using different analytical methods like those aforementioned.

A. Model Accuracy Results

TABLE I.
C5.0 Accuracy Table

	Actual		
Predicted	***Increase Stock Price***	***Decrease Stock Price***	***Sum***
Decrease Stock price	1	23	24
Increase Stock Price	42	3	45
Sum	43	26	69

TABLE II.
CHAID Accuracy Table

	Actual		
Predicted	***Increase Stock Price***	***Decrease Stock Price***	***Sum***
Decrease Stock price	6	12	18
Increase Stock Price	37	14	51
Sum	43	26	69

Both Table I and Table II show the C5.0 and CHAID decision tree classification techniques that can achieve high accuracy rates.

For the C5.0 decision tree model, of the 43 stocks that actually increased their price, 42 (97.7%) of these were predicted by the C5.0 model to increase their price. For the CHAID decision tree model, of the 43 stocks that actually increased their price, 37 (86%) of these were predicted by the CHAID model to increase their price.

TABLE III.
Neural Network Accuracy Table

	Actual		
Predicted	**Increase Stock Price**	**Decrease Stock Price**	**Sum**
Decrease Stock price	3	8	11
Increase Stock Price	40	18	58
Sum	43	26	69

From Table III, for the Neural Network model, of the 43 stocks that actually increased their price, 40 (93%) of these were predicted by the Neural Network Model to increase their price.

From Table 4, for the Logistic regression model, of the 43 stocks that actually increased their price only 35 (81%) of these were predicted by the Logistic Regression Model to increase their price.

Typically, two measures, "Sensitivity" and "Specificity" help one to assess the accuracy and performance of a model.

$$Sensitivity = \frac{\#True\ Positives}{\#True\ Positives + \#False\ Negatives}. \quad (1)$$

$$Specificity = \frac{\#True\ Negatives}{\#True\ Negatives + \#False\ Positives}. \quad (2)$$

TABLE IV.
LOGISTIC REGRESSION ACCURACY TABLE

	Actual		
Predicted	**Increase Stock Price**	**Decrease Stock Price**	**Sum**
Decrease Stock price	8	17	25
Increase Stock Price	35	9	44
Sum	43	26	69

The accuracy rates for the analytical models in terms of sensitivity (1) and specificity (2) are presented in Table V below.

TABLE V.
SUMMARY ACCURACY TABLE

Accuracy Measure	C5.0	CHAID	Neural Network	Logistic Regression
Sensitivity	98%	86%	93%	80%
Specificity	88%	46%	31%	65%

Overall the models performed very well, with average sensitivity of 89% and specificity of 58%. The C5.0 decision tree had the best overall performance. Sensitivity is the more important measure for us as we are more interested in correctly classifying stocks with increasing price trends. So, the overall accuracy level (sensitivity of 89%), provides support that our analytical techniques are reliable and have high accuracy.

B. *Portfolio Result*

The personal strategy, portfolio 1 yielded a higher return, than the Australian Ordinary Index (AOI) (7.9% versus 1.7%) during a 20 day trading period. This yield is more than 4.5 times higher than the AOI return.

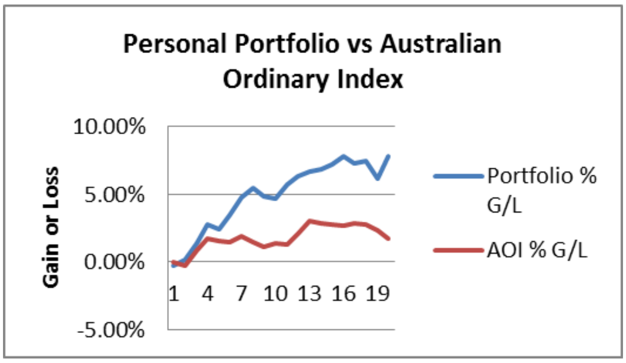

Figure 3. Personal Strategy Portfolio.

The price trading strategy (using a C5.0 decision tree), portfolio 2 also yielded a higher return, than the Australian Ordinary Index (AOI) (19.9% versus 1.7%) during a 20 day trading period. This yield is more than 11.7 times the AOI return. One thing to take note with the low price strategy is the volatility. The volatility is higher than the personal strategy. It confirms the notion that the higher the volatility, the higher the possible returns. Stocks with lower prices are more volatile than others but may produce higher returns.

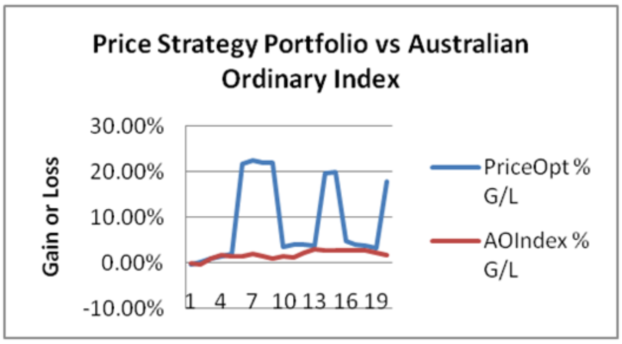

Figure 4. Price Strategy (C5.0) Portfolio.

Similarly, the growth strategy (using a CHAID decision tree), portfolio 3 also yielded a higher return, than the Australian Ordinary Index (AOI) (20.8% versus 1.7%) during a 20 day trading period. This yield is twelve times the AOI return.

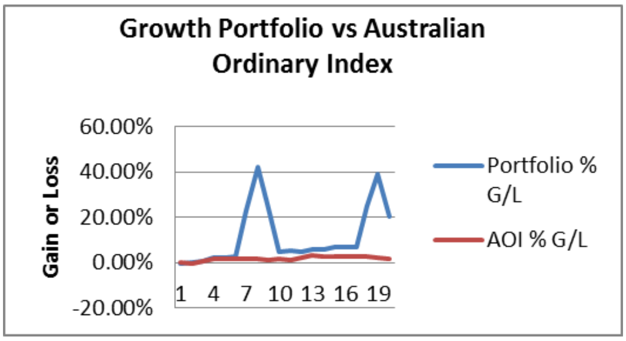

Figure 5. Growth Strategy (CHAID) Portfolio.

The “All mixed strategy”, portfolio 4, also yielded a higher return than the Australian Ordinary Index (AOI) (23.2% versus 1.7%). This yield is almost 14 times more than the AOI return.

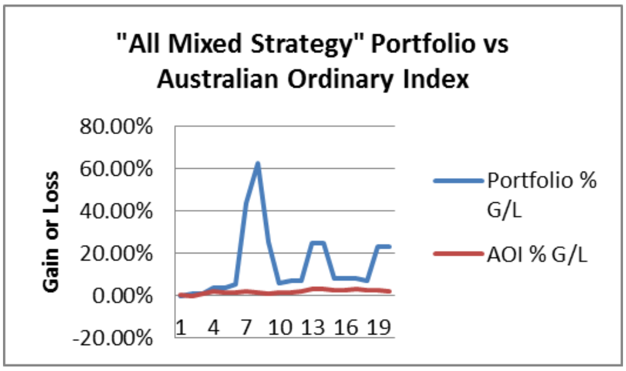

Figure 6. All Mixed Strategy (Neural Network) Portfolio

Lastly, the growth & value strategy, portfolio 5, also yielded a higher return than the Australian Ordinary Index (AOI) (5.4% versus 1.7%). This yield is 3 times more than the AOI return.

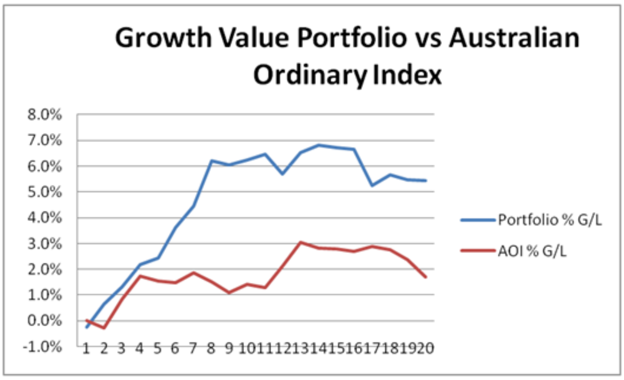

Figure 7. Growth Value Strategy (Linear Regression) Portfolio

It is quite clear from the above results as shown in Figures 3 to 7, that our hypothesis hold true:

H1: The stocks selected through a personal systematic approach outperforms the Australian stock market significantly

H2, H3, H4: The stocks selected through analytical techniques such as C5.0 decision tree, CHAID decision tree, Neural Networks and the Logistic Regression Model outperform the Australian stock market significantly

Furthermore, a statistical test was performed at the 5% level and it confirmed that the returns of all 5 portfolios were significantly different to the returns of the Australian stock market.

C. *Limitations*

The data set size is smaller than usual and missing data was imputed using the mean value. Another limitation is that there are many other learning algorithms in the Neural Network to be explored. Secondly, we focused on the industrial sector so our stock selection criteria and trading strategy is currently limited to this sector.

VI. Conclusion

In this paper we showed how several analytical techniques can be used to build classification models with stock data from the Industrial Sector of the Australian Stock Market. In conclusion, we can safely say that our stock selection and trading strategy outperformed the Australian Ordinary index. If we annualize our portfolios under the assumptions that we could consistently trade in this way each month, then expected returns for each portfolio would be 94% for our personal strategy, 216% for the price strategy, 244% for the growth strategy, 263% for the "all mixed" strategy, and 65% for the growth value strategy.

Further, we have obtained high accuracy rates for all four analytical models under study, the C5.0 decision tree, the CHAID tree, the Neural Network model, and the Logistic Regression Model. The C5.0 decision tree model performed the best in classifying the stocks.

Our results are good with regards to the validation criteria. We have confirmed that the metrics, ROE, ROA, estimated growth this year, price, and analyst opinion, are good predictors for classifying stocks into two groups, namely, stocks that are highly likely to increase in price and stocks they are not likely to increase in price.

Our findings justify the use of analytical techniques for classification and prediction purposes. Lastly, our trading strategy consistently produced results that outperformed the Australian Ordinary Index, whether our personal strategy or the pricing strategy or the growth strategy or the growth value strategy or the all mixed strategy was chosen, because our trading strategy was employed we consistently outperformed the Australian market.

VII. Future Work

Future researchers may include more methods for finding the best model for predicting stock prices. We used stocks from the Industrial Sector however it would be interesting to expand our study to see whether our stock selection and trading strategy will work in other sectors. Another interesting idea we have in mind, is to build an application tool that applies our stock selection and trading strategy. Finally, our trading strategy has a 10% upper limitation on the portfolio and an exit strategy on a loss of 5% or more. Other trading strategies may be investigated.

References

[1] R.Apostolos-Paul, "Neural Networks in the Capital Markets," 1995.

[2] A. Costea, "New Approached and Perspectives in eth use of Quantitative Methods". Journal of Applied Quantitative Methods. Vol.1 No.2 Winter, 2006.

[3] D. Davis. "Business research for Decision–making", 1st ed., Belmont, CA: Thomson Brooks/Cole, 2005

[4] B. de Ville. "Decision trees for Business Intelligence and Data mining: Using SAS Enterprise Miner. SAS Institute Inc., Cary, NC, USA.ISBN-13:978-1-59047-567-6, 2006.

[5] E. Fama. "The Behaviour of Stock Market Prices." Journal of Business 38, pp.34-105, 1965.

[6] E. Fama. "Random Walks in Stock Market Prices". Financial Analysts Journal 21, pp.55-9, 1965.

[7] W.R. Ferson; C.R Harvey. "The Risk and Predictability of International Equity Returns". Review of Financial Studies 6, pp.527-66, 1993.

[8] E Guresen. "Using Artificial Neural Network Models in Stock Market Index Prediction. Expert Systems with Applications", 38 (8), pp.10389-10397.

[9] J.F. Hair. "Multivariate Data Analysis with Readings" 4th ed, Englewood Cliffs, NJ:prentice Hall, 1995.

[10] E. Hajizadeh; H.D. Ardakani; J. Shahrabi. "Application of Data Mining Techniques in Stock Markets: A Survey.", Journal of Economics and International Finance. Vol. 2 (7), pp.109-118, 2010.

[11] C.A. Hargreaves; H Yi; "Does the Use of Technical & Fundamental Analysis Improve Stock Choice? : A Data Mining Approach applied to the Australian Stock Market". Statistics in Science, Business and Engineering (ICSSBE)

International Conference Proceeding. IEEE Explore Digital Library, pp.1-6, 2012

[12] Q. Huang; Y. Cai; J. Peng. “Modeling the Spatial Pattern of Farmland using GIS and Multiple Logistic Regression: A Case study of Maotiao River Basin, Guizhou Province, China. Environmental Modeling and assessment, 12 91), pp.55-61.

[13] H. Li. “Predicting Business Failure Using Classification and Regression Tree: An Empirical Comparison with Popular Classical Statistical Methods and Top Classification Mining Methods”. Expert Systems with Applications, 37 (8), pp.5895-5904, 2010.

[14] J.R. Quinlan. “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers, 1993.

[15] http://sg.finance.yahoo.com/q?s=asx.ax&ql=1

[16] http://hfgapps.hubb.com/asxtools/Charts.aspx