An Automated Stock Investment System Using Machine Learning Techniques: An Application in Australia

Carol Anne Hargreaves

***Abstract*—**A key issue in stock investment is how to select representative features for stock selection. The objective of this paper is to firstly determine whether an automated stock investment system, using machine learning techniques, may be used to identify a portfolio of growth stocks that are highly likely to provide returns better than the stock market index. The second objective is to identify the technical features that best characterize whether a stock's price is likely to go up and to identify the most important factors and their contribution to predicting the likelihood of the stock price going up. Unsupervised machine learning techniques, such as cluster analysis, were applied to the stock data to identify a cluster of stocks that was likely to go up in price – portfolio 1. Next, the principal component analysis technique was used to select stocks that were rated high on component one and component two – portfolio 2. Thirdly, a supervised machine learning technique, the logistic regression method, was used to select stocks with a high probability of their price going up – portfolio 3. The predictive models were validated with metrics such as, sensitivity (recall), specificity and overall accuracy for all models. All accuracy measures were above 70%. All portfolios outperformed the market by more than eight times. The top three stocks were selected for each of the three stock portfolios and traded in the market for one month. After one month the return for each stock portfolio was computed and compared with the stock market index returns. The returns for all three stock portfolios was 23.87% for the principal component analysis stock portfolio, 11.65% for the logistic regression portfolio and 8.88% for the K-means cluster portfolio while the stock market performance was 0.38%. This study confirms that an automated stock investment system using machine learning techniques can identify top performing stock portfolios that outperform the stock market.

***Keywords*—**Machine learning, stock market trading, logistic principal component analysis, automated stock investment system.

I. Introduction

STOCK selection is an important step before portfolio construction [8]. Moreover, one of the key issues and difficulties on stock selection is how to identify the most informative fundamental factors and technical factors that can explain the excess return. A stock portfolio using the data mining approach was performed using the Australian Stock Market [2], where results demonstrated successfully that data mining techniques are able to model the trend of stock prices which are nonlinear. A comprehensive review of stock algorithmic trading methods [6] found that most models perform effectively in downtrend markets and poorly in uptrend markets, and transaction costs are a significant factor.

It has been demonstrated that investors can select winning stocks in a systematic way and make a profit [4]. Most investors only use technical analysis based on price movement to predict the future stock performance. Using only technical analysis will not always lead to selecting winning stocks. The same holds true for the use of fundamental data analysis to choose winning stocks. A stock may have an upper trend in the stock movement currently but if it does not have a good financial performance, such as a high return on equity (ROE) value or a high return on asset (ROA) value, the stock's upward trend may not be a long-standing one. In our research we use both fundamental and technical data analysis in order to reap the benefits of both the methodologies [7].

This research considers the 2 000 stocks listed on the ASX stock market (Australia) to identify the best portfolio that an investor can maintain for risk reduction and high profitability. Machine learning techniques accurately predicted stock performance in Korea [12]. In particular, [13] demonstrated that the logistic regression technique accurately predicted the stock market index.

Further, [5] justified the use of analytics for the classification and prediction of stock prices while using only five input variables, (ROE, ROA, analyst opinion, growth this year, price). They confirmed that the metrics were good predictors for classifying stocks into two groups, namely, stocks that were highly likely to increase in price and stocks they were not likely to increase in price. In addition, [10] demonstrated the effectiveness of principal component analysis using the Shanghai Stock Exchange stock.

In this study, we explore how well principal component analysis contributes to identifying stocks that will perform well, using a short-term stock trading strategy of one month and innovative performance indicators.

The purpose of this paper is to determine whether an automated stock investment system using machine learning techniques can be used to select good stocks that will outperform the stock market index in Australia. To validate whether our automated stock investment system can identify and select good profitable stock portfolios that perform better than the stock market index we experimented using three different stock portfolios.

For each of our three stock portfolios, (i) Principal Component Analysis (PCA) Portfolio, (ii) Logistic Regression Portfolio, (iii) K-Means Cluster Portfolio, the three top training stocks from 15 January 2018 to 15 February 2018 were paper traded from 15 February 2018 – 15 March 2018.

Carol Anne Hargreaves is with the National University of Singapore, Singapore (e-mail: stacah@nus.edu.sg).

The returns for all four stock portfolios were 23.87% for the PCA stock portfolio, 11.65% for the logistic regression portfolio and 8.88% for the K-means cluster portfolio while the stock market performance was 0.38%

This paper has five sections. While Section I is the introduction, Section II is a brief overview of the methodology, Section III, the statistical analysis results, Section IV, the conclusion and Section V, the references.

II. Methodology

A. Data Capture and Transformation

Daily stock data such as open, high, low, close and volume were downloaded from au.finance.yahoo.com for 2000 stocks for the period of 2 January 2017 to 15 February 2018. The All Ordinaries Index daily stock data also came from au.finance.yahoo.com. 24 new variables were derived from the open, high, low, close and volume variables. All stocks were then rated by sector on a scale of 1-10 for each of the 30 variables. 24 fundamental stock financial data variables such as ROA, ROE, etc. were also captured from au.finance.yahoo.com. The data were also rated by sector on a scale of 1-10.

All stocks have some risk of losing money, some stocks more than others. We manage stock risk by using only financially health stocks for trading. In other words, even if the stock has a high likelihood of going up, we will only select them as a good stock for trading if their fundamental stock rating is high, at least 7 on a scale of 1-10. We further manage the risk of our stocks by diversifying our stock portfolio, that is, we buy more than one stock at a time. In particular, 3 stocks were selected for our stock portfolio but at the same time; the three stocks do not all come from the same sector. In Section II *B*, we present the data analysis procedure.

B. Data Analysis Flow Chart

Fig. 1 explains how data were captured and the steps on further data analysis.

C. Data Modelling

As the target variable is ‘Trend’, a categorical variable, four methods were used for data modelling. They are Logistic Regression, K-Means Cluster Analysis, Statistical Hypothesis Testing, and PCA. 24 combinations of the five variables, open, high, low, close, and volume were derived and used as inputs to the model. Variables such as the ’FiftyTwoWeekHigh’, ‘Standard Deviation’, ‘FourWeekChange’, and similar were derived for modelling purposes.

1. Logistic Regression

The Logistic Regression is a supervised machine learning technique used for classification problems with a binary response variable as a function of explanatory variables which can be continuous or discrete.

The Logistic Regression is represented by the following logit function:

$$\text{logit}(p)=b_0 + b_1X_1+b_2X_2+b_3X_3+b_4X_4+\ldots..+b_kX_k$$

$$\text{logit}(p)=\ln(p/1-p)$$

where p is the probability of a success for the response class of interest and the ‘X” variables are the explanatory variables, with b_0, b_1, b_2, etc are the coefficients of the explanatory variables.

Arun et al. [1] predicted stock performance in the Indian Stock Market using Logistic Regression. We would like to determine whether prediction of stocks in the Australian Stock Market using the Logistic Regression will also result in selecting stocks that perform well.

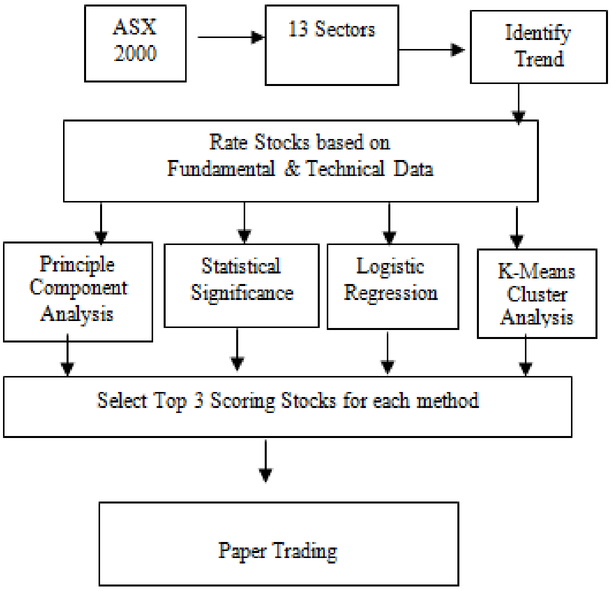
ASX 2000
13 Sectors
Identify Trend
Rate Stocks based on Fundamental & Technical Data
Principle Component Analysis
Statistical Significance
Logistic Regression
K-Means Cluster Analysis
Select Top 3 Scoring Stocks for each method
Paper Trading

Fig. 1 Data Analysis Flow Chart

2. K-Means Cluster Analysis

Clustering is one of the most useful tasks for discovering groups and identifying interesting distributions and patterns in the underlying data. The objective of the clustering technique is to group the data in such a way that similar data points form a cluster and data points in different clusters are not similar.

Cluster Analysis is a method of unsupervised learning to partition a data set into a set of clusters. Partitive algorithms like K-means partition the data set into k-clusters, in which every prototype is clustered based on its closest mean. Results obtained from K-means represent clusters of prototypes, which in turn are a map of the underlying stocks. Reference [9] used K-Means to aggregate similar technical charts in order to construct trading rules. We adopted representatives of the defined clusters as our data of interest. Thus, data compression is achieved.

K-Means cluster analysis involves making several clustering trials with different values for the number of clusters, until there is a cluster having the best underlying stocks. Each cluster is profiled according to the key features that best describe that cluster. The cluster with high ratings on the features that best relate to the likelihood of the stock price going up is then selected for choosing the top three stocks for trading. At the end of the month, the return on investment for each portfolio is compared with the All Ordinaries Index market return.

3. PCA

PCA is important for identifying the relevant features and used to filter redundant features. By filtering out redundant features, we run a simpler model that is easier to interpret. With a large number of variables, there are many pairwise correlations between variables. To interpret the data in a more meaningful way, we use the dimension reduction technique, PCA, in order to reduce a large number of variables to a fewer number of factors. The resulting factors are an interpretable linear combination of the input variables. IBM Statistics 23 was used to perform the PCA.

Initially, all the 24 technical variables were used for the PCA analysis. The initial eigenvalues and scree plot were used to identify the approximate number of components (factors) to be extracted. There is a rule of thumb that the meaningful factors should contribute at least 70% of the total variance. The rotation method helps to make the final factors that are extracted, more interpretable. For this study we used the varimax rotation method. Variables which had low (less than 0.6) or no loadings were removed one by one and the PCA procedure was repeated until we achieved all variables loading onto a factor with an eigenvalue greater than one, their proportion of total variance explained at least greater than 70% and the variables loaded in the rotation matrix is greater than at least 0.7.

Once we are satisfied with the factors we have identified, we need to test the reliability of the factors. The Cronbach Alpha Coefficient helps to measure the internal consistency of the factors. The Cronbach Alpha Coefficient value is expected to be greater than 0.7 in order to confirm that the factor is reliable and consistent.

D. Paper Trading

Each stock portfolio had an investment of $30 000, with each stock having almost $10 000 invested in. At the end of each month the stocks are sold and the return on investment for the portfolio is computed and compared with the stock market performance.

III. Statistics Analysis Results

A. Logistic Regression Analysis Results

A correlation analysis was performed for all pairs of the 24 input variables to identify which input variables were highly correlated with each other. Nine of the variables were highly correlated with each other and removed as input variables for the Logistic Regression model, to overcome the multicollinearity problem.

Three variables were significant for the Logistic Regression Model, ‘Standard Deviation’, ‘Fifty2WeekHigh’, ‘FourWeekChange’. The variable, ‘FourWeekChange’ was five times more important for predicting the likelihood of the stock price going up than the other two variables, ‘Fifty2WeekHigh’ and ‘Standard Deviation’, see Table I.

The logistic regression model was evaluated using the overall accuracy of the model, the sensitivity measure and the specificity measure. The overall accuracy for the Logistic Regression model was 88.3%, sensitivity measure was 94.4% and the specificity measure was 88.2%, see Table II.

TABLE I
Logistic Regression Coefficient Summary Table

Variable	B	S.E.	Wald	df	Sig.
Standard Deviation	0.442	0.117	14.262	1	0.000
Fifty2WeekHigh	0.488	0.178	7.495	1	0.006
FourWeekChange	1.622	0.407	15.912	1	0.000
Constant	23.722	4.712	25.348	1	0.000

TABLE II
The Logistic Regression Confusion Matrix

			Predicted		
	Observed		Trend		Percentage Correct
			0	1	
Step 1	Trend	0	603	81	88.2
		1	1	17	94.4
	Overall Percentage				88.3

All three measures were very high and close to one, so the logistic regression model results were considered to be good.

Thirty thousand dollars was invested in the top three stocks, MIN, TWE, and A2M and at the end of the month the portfolio was worth $33 495.80, which is a return of 11.65% versus the All Ordinaries stock market index return of 0.38%.

B. K-Means Cluster Analysis Results

Seven clusters were identified. Cluster 1 was identified as the cluster with the high ratings on the features associated with the likelihood of the stock price going up. There were nine feature variables that had a rating of ‘9’ compared to other clusters with ratings lower than 9 on the same input variables. The top three stocks selected for this portfolio were MIN, TWE and BAL. Thirty thousand dollars was invested in these three top stocks and at the end of the month the portfolio was worth $32 663, which is a return of 8.88% return versus the All Ordinaries stock market Index of 0.38%.

C. PCA Results

There were two factors that result from the 24 input variables with eigenvalues greater than one. These two factors explained 85.17% of the total variance, see Table III.

Factor 1 is the most important factor, explaining 47.84% of the total variance, while factor 2, explained 37.33%. Factor 1 had nine of the 24 input variables loading on it, while factor 2 had six variables loading on it. We termed factor 1, ‘price gain’ and factor 2 we termed, ‘week high’.

PCA helped us identify from the 24 input variables, the nine most important variables for identifying which stocks are likely to go up.

Using Cronbach’s Alpha, factor 1 has a 0.91 reliability coefficient, factor 2, a reliability coefficient of 0.90 and the reliability and consistency for this study overall was 0.90. All three reliability values are very high, that is, close to 1, demonstrating consistency and confidence in the model outcomes.

The top three stocks for the PCA portfolio were A2M, IMF and BAL. Thirty thousand dollars was invested in this portfolio for 1 month and at the end of the month the portfolio was worth $37160.19, which is a return of 23.87% versus the All Ordinaries stock market Index of 0.38%.

IV. Conclusion

Three machine learning techniques (K-Means Cluster Analysis, PCA, Logistic Regression Analysis) were used to identify which stocks are likely to be the most profitable. The outcome from this study confirms firstly that machine learning techniques can be used to predict which stock is likely to go up. Further, this study successfully demonstrated that using technical and fundamental stock data, in conjunction with machine learning techniques, we were able to trade a portfolio of stocks that consistently (all three portfolios) outperformed the All Ordinaries Index. The Logistic Regression Portfolio return on investment was 11.65%, the K-Means Cluster Portfolio return on investment was 8.88%, and the PCA Portfolio return on investment was 23.87% versus the All Ordinaries stock market index return of 0.38%.

In this study we demonstrated that the approach of rating stocks on a scale of 1-10, based on the likelihood of the stock price going up and also rating the stocks on a scale of 1-10 in terms of their financially health, true profitability was realized for 3 independent machine learning methods and stock portfolios.

Most importantly, this study demonstrated that the 24 variables derived innovatively from the typical stock variables,' High', 'Low', 'Open', 'Close', and 'Volume', proved to be statistically significant predictors that accurately predicted the likelihood of the stock price going up and delivered the returns that any investor would be more than happy with as the stock market index returns was only 0.38% versus the three stock portfolio returns of 23.87%, 11.65% and 8.88% respectively.

References

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

[2] C.A. Hargreaves, P. Dixit, A. Solanki, "Stock Portfolio Selection using Data Mining Approach", *IOSR Journal of Engineering*, Vol 3, Issue 11, 2013, pp. 42-48.

[3] C.A. Hargreaves, C. Kardivel Mani, "The Selection of Winning Stocks using Principal Component Analysis", *American Journal of Marketing Research*, Vol 1, No.3, 2015, pp. 183-188.

[4] C.A. Hargreaves, Y. Hao," Does the use of Technical & Fundamental Analysis improve Stock Choice? A Data Mining Approach applied to the Australian Stock Market", *International Conference on Statistics in Science, Business and Engineering (ICSSBE), IEEE Explore,* 2012, pp. 1-6.

[5] C.A. Hargreaves, Y. Hao, "Prediction of Stock Performance Using Analytical Techniques", *Journal of Emerging Technologies in Web Intelligence,* Vol 5, No. 2, 2013, pp. 136-142.

[6] R. Peachavanish, "Stock selection and Trading Based Cluster Analysis of Trend and Momentum Indicators," *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2016, Vol 1,

[7] C. Kardivel Mani, C. A. Hargreaves, "Stock Trading using Analytics", *American Journal of Marketing Research,* Vol.2, No. 2, 2016, pp. 27-37.

[8] S. Thawornwong, S and D. Enke, "The adaptive selection of financial and economic variables for use with artificial neural networks," *Neurocomputing*, vol. 56, 2004, pp. 205-232.

[9] K. Wu, K; Y. Wu, H. Lee, "Stock trend prediction by using K-means and Apriori algorithm for sequential chart pattern mining", *Journal of Information Science and Engineering*, 30, 2014, pp. 653-667.

[10] Wang, Z., Sun, y., Stockli, P. (2014). "Functional Principal Components Analysis of Shanghai Stock Exchange 50 Index". Discrete Dynamics in Nature and Society Volume 2014 Article ID 365204, 7 pages

[11] Renugadevi, T., Ezhilarasie,R., Sujatha, M and Umamakeswari, A. (2016). Stock Market Prediction using Hierarchical Agglomerative and K-Means Clustering Algorithm. Indian Journal of Science and Technology,Vol.9,(48),DOI:10.17485/ijst/2016/v9i48/108029

[12] Wang, Y., In-Chan Choi. (2013)." Market Index and stock price direction prediction using Machine Learning Techniques: An empirical study on the KOSPI and HSI". Science Direct. Pages 1-13. http://arxiv.org/pdf/1309.7119v1.pdf

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