APPLYING MACHINE LEARNING TECHNIQUES TO IDENTIFY TOP STOCKS

CAROL ANNE HARGREAVES

Department of Statistics and Applied Probability
National University of Singapore Singapore,
E-mail: stacah@nus.edu.sg

Abstract- In this paper, we focus on the application of machine learning algorithms such as the hierarchical clustering technique, the principle component analysis technique and the logistic regression technique. The main objective is to firstly, identify the most important technical factors that may be associated with "good" stocks and also to identify a cluster of stocks that may be classified as "good" stocks. Using only technical data, two principle components were identified, "Price High" and "Price Growth", contributing 46% and 29% to explaining the variability in the stock data. Hierarchical Clustering was used to understand which stocks were similar. Notably, stocks that scored high on the factor analysis were also grouped together by the hierarchical clustering method. Further, a machine learning model, the logistic regression model predicted the top five stock to go up in price, these same five stocks were identified by the hierarchical clustering technique. The logistic regression model had a sensitivity of 80%.

Keywords- Machine learning model; logistic regression; principle component analysis; hierarchical clustering; stocks; stock trading

I. INTRODUCTION

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 identify the good stocks.

Not all stocks in the stock market or sector will yield a profit, hence, the investor needs to analyse the stock data, so that an informed decision on which stocks to buy is made. To solve the problem, nearly all professionals subscribe to one of the two generally accepted approaches for stock analysis: fundamental analysis and technical analysis. Fundamental analysis includes the industry conditions, financial conditions and management of the organizations whereas in technical analysis instead of measuring the intrinsic value it mainly studies the stock charts to identify patterns and the trend of the stock market. It is always a debate which methodology is best [1].

Technical analysis is the foundation for automated trading. Currently there are decision making algorithms such as Pairs Trading, Automated Gaming and Execution Algorithms such as Iceberg, Volume Weighted Average Price been implemented for algorithmic trading [2]. A wide variety of classification algorithms exist in the literature such as Support Vector Machines, K-Nearest Neighbour, Decision Trees, Bayes Classifier [3] which are well understood and widely applied, but the present work mainly focuses on the logistic regression model, principle component analysis and hierarchical clustering.

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. A stock portfolio using the data mining approach was performed using the Australian Stock Market [4], where results demonstrated successfully, that data mining techniques can model the trend of stock prices which are nonlinear.

In this study, a short-term stock trading growth strategy of one month has been adopted where innovative performance indicators were developed. The performance of the top stocks identified were assessed using a paper trading strategy. In [5], principal component analysis was demonstrated to identify the key factors that identify good stocks. The purpose of this paper is to determine whether machine learning techniques, both unsupervised (Hierarchical Clustering and Principle Component Analysis) and supervised (Logistic Regression) can identify good stocks for trading by using trading rules in a semi-automated approach in the Australian Stock Market. This paper is structured into 5 sections. While Section 1 is the introduction, Section 2 gives a brief literature review, Section 3, a brief overview of the methods used, Section 4 the analysis results, after which Section 5 presents the conclusion.

II. LITERATURE REVIEW

Many research papers have predicted the pricing of the stock index as well as stock performance across many European markets (e.g., UK, France, and Germany), [6] observed that stock returns are predictable. Both fundamental and technical variables

were combined for the prediction of profitable stocks using the support vectors machine learning algorithm [7]. [7] systematically identified high returning healthcare stocks and traded them with the help of an automated trading application without human error and sentiment interference and yielded 16.64% revenue at the end of three months trading.

The logistic regression was used by [8] as a comparative method in order to build a better model for predicting stock returns effectively and efficiently. The logistic regression technique yields coefficients for each independent variable based on a sample of data [9]. The logistic regression also
has the capacity to analyze a mix of all types of predictors
[10]. 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. Further, [4] investigated whether a healthcare sector stock portfolio will outperform the ASX All-Ordinaries Index and Healthcare Sector Index (XHJ) over the twenty days trading period using the logistic regression model. The healthcare stock portfolio returned 18.24%. [11], 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.

Principal Component Analysis has been demonstrated [12] to identify the key factors that identify winning stocks. [13], explored the application of principle components of Shanghai stock exchange 50 index by means of functional principal component analysis (FPCA) and reduced the dimension to a finite level using FPCA and extracted the most significant components of the data and some relevant statistical features of related datasets. FPCA has proved to be successful at extracting the main variance factors. Further, [14], also used the principal component analysis technique to reduce 19 stock data input variables to 9 stock data input variables for a stock prediction system for the Nigerian stock exchange.

In addition, [15], applied the principle component analysis on daily frequency observations on spot exchange rates, stock market indexes, and long term and short term interest rates for nine countries. [15] showed that the principal components analysis may be used to reduce the effective dimensionality of the scenario specification problem in several cases. [16] applied the principle component analysis to the Korean composite stock price index (KOSPI) and the Hangseng Index (HIS) to reduce the data points into two components.
[17] used a hierarchical single-link agglomerative clustering method to convert feature vectors to form clusters based on maximum similarity. The hierarchical cluster analysis results were visualized using a dendrogram.
All the authors above have successfully demonstrated that machine learning techniques can accurately identify good stocks for trading. They have all used only the technical indicators based on past price and volume to predict the stock performance. It is a fact that good technical indicators on their own, do not always lead to better returns. Same holds for fundamental stock indicators. A stock may have an upward trend in its stock movement at the time of analysis but if it may not have good fundamentals such as an effective management, and therefore, the upward trend may not be a
long standing one.

In our research, we have used the principle component analysis to understand the most important factors associated with an upward trend, we also apply the hierarchical cluster analysis to determine which stocks are similar and then apply the machine learning algorithm, the logistic regression model to predict the trend of the stocks using the technical stock data. Finally, we hypothetically trade the top 3 stocks suggested by the principle component analysis, hierarchical clustering and logistic regression technique, and, compare the returns with the stock market returns.

III. METHODS USED

1.1 Principal Component Analysis (PCA)

Principal component analysis is important for identifying the relevant features and used to filter redundant features out. By filtering out redundant features, we run a simpler model that is easier to interpret. With 24 technical input variables, there were many pairwise correlations between variables. To interpret the data in a more meaningful way, we used the dimension reduction technique, the Principal Component Analysis, in order to reduce the 24 input variables to a fewer number of factors. The resulting factors are an interpretable linear combination of the input variables.

The initial eigenvalues and the scree plot were used to understand the approximate number of components (factors) that could be extracted. The sample size should be adequate for a PCA analysis to run successfully. The Kaiser-Meyer-Olkin (KMO) is a measure of sampling adequacy and verifies whether the original variables can be factorised efficiently by comparing the correlation values between variables and those of the partial correlation. If the KMO measure is above 0.5, we may conclude that the sample size is adequate.

Bartlett's test also tests the variables relationship strength. It

does this by testing the null hypothesis that the correlation
matrix is an identity matrix. An identity matrix is a matrix in
which all the diagonal elements are 1 and all off diagonal
elements are 0. The expected result is to be significant so that
we can reject the null hypothesis [18].

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.

1.2 The Machine Learning Technique: Logistic Regression

The logistic regression model was selected for use in this study because it is the most basic and robust classification algorithm. The logistic regression is a special type of regression where the binary response variable is related to a set of explanatory predictor variables which can be continuous or discrete. The logistic regression model is perfect for situations where the aim is to predict the presence or absence of a characteristic or outcome based on the values of a set of predictor variables. The relationship between the target and input variables is not always a straight line, and so a non-linear or logistic regression model is used. Further, the Logistic Regression model was chosen as it required little running time compared to other complicated machine learning algorithms and its output is also easy to interpret. As the dependent variable of the Logistic Regression Model is binary, the dependent variable selected was “Upward Trend” and was used to classify the stocks.

The main assumption for building a logistic regression model is that the independent variables do not have significant multi-collinearity [19]. Initially, there were twenty-four input variables. After multi-collinearity was handled, we were left with 14 input variables for our Logistic Regression Model. Our Logistic Regression model was carried out using the backward stepwise regression method. Variables with a p-value greater than 0.05 was deemed as insignificant to “Upward Trend” and were removed from the model. The process stopped when the model was left with only variables with p-values less than 0.05.

1.3 Hierarchical Cluster Analysis

Clustering is one of the most useful tasks for discovering groups and identifying interesting distributions and patterns in the underlying data. The clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. Cluster Analysis is a method of unsupervised learning to partition a data set into a set of clusters.

In this study, single-link agglomerative clustering was used. The clusters were formed based on maximum similarity and the clusters were visualized using a dendrogram.

IV. RESULTS AND FINDINGS

1.4 Principle Component Analysis Results

The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy test was first performed. The KMO measure was 0.785, as it was above 0.5, we concluded that the sample size was adequate. Next, Bartlett's test was performed. A p-value of 0.000, which is less than 0.05, suggesting that a principle component analysis can be performed on this stock dataset.

There were two factors that results from the twenty-four input variables with eigenvalues greater than one. These two factors explained 81.2% of the total variance. See table 1 below.

TABLE 1: Principle Component Analysis Total Variance Explained

Total Variance Explained

	Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %
1	4.873	48.725	48.725
2	3.250	32.498	81.223

Extraction Method: Principal Component Analysis.

Factor 1 is the most important factor, explaining 48.7% of the total variance, while factor 2, explained 32.5%. Factor 1 had six of the twenty-four input variables loading on it, while factor 2 had four variables loading on it. We termed factor 1, ‘price high’ and factor 2 we termed, ‘price growth’. Principal Component Analysis helped us to identify from the twenty - four input variables, the ten most important variables for identifying which stocks have an “Upward Trend” and are likely to go up.

Using Cronbach Alpha, factor 1 has a 0.947 reliability coefficient, factor 2, a reliability coefficient of 0.854 and the reliability and consistency for this study overall was 0.888. All three reliability values were very high, that is, close to 1, demonstrating consistency and confidence in the principle component model results. Factor 1 scores range between -2.11 and 1.73. Since factor 1 is known to be the most important factor, we assume that stocks that have a factor 1 score greater than 1.4 (9%) are good

stocks to buy. Stocks, SPO, EBO and CNU were selected for our paper trading exercise.

1.5 The Hierarchical Cluster Analysis Results

We input 3 variables that loaded onto factor 1 of the principle component analysis, into our hierarchical cluster model to define our clusters. Figure 1 below, shows the positioning and similarity of our 3 stocks selected using the principle component analysis, CNU, EBO and SPO.

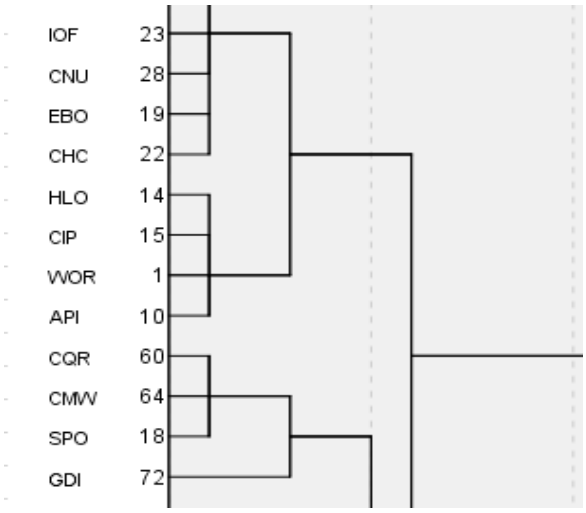


Figure 1: Part of the Dendrogram for the 155 stocks.

The Hierarchical clustering analysis suggests that CNU, EBO and SPO may be good stocks to buy.

1.6 The Logistic Regression Model Results

Insights about the input variables and their effect on Trend was gained by understanding the signs and size of the coefficients of the Logistic Regression model results and the resulting odds ratios. The size of the coefficient indicates the main drivers of "Upward Trend". The larger the coefficient, the higher the significance of the variable to "Upward Trend". A positive coefficient indicated that the predictor variable had a positive relationship with "Upward Trend".

"IndGrowthMonth" (2.2), "FourWeekHigh" (1.31) and "52WeekHigh" (0.628) have the highest positive relationship with "Upward Trend" and were the only significant variables in the logistic regression model. See figure 2 below.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	IndGrowthMonth	2.200	.778	7.989	1	.005	9.024
	Fifty2WeekHigh	.628	.269	5.443	1	.020	1.873
	FourWeekHigh	1.310	.518	6.409	1	.011	3.708
	Constant	-28.373	8.796	10.404	1	.001	.000

a. Variable(s) entered on step 1: IndGrowthMonth, Fifty2WeekHigh, FourWeekHigh.

Figure 2: Logistic Regression Coefficient Summary

An added advantage of using the Logistic Regression is that it calculates the logarithmic odds of a stock going up. The odds ratio is calculated to study the effect of each variable in affecting the odds of the "Upward Trend" [20]. The odds ratio for "IndGrowthMonth", "FourWeekHigh" and "52WeekHigh" was 9.024, 3.708 and 1.873 respectively.

This means that holding all other variables constant, stocks with "Growth in price in the Month" are 9 times more likely to have an "Upward Trend" than stocks without "Growth in price in the Month". Similarly, for stocks which have a "Four Week High" are 3.7 times more likely to have an "Upward Trend" than stocks without a "Four Week High". Lastly, stocks with a "52WeekHigh" are 1.9 times more likely to have an "Upward Trend" than stocks without a "52 Week High".

Other than understanding the association of different variables to an "Upward Trend", there is a need to evaluate the training model to find out its predictive power. This is done by measuring the accuracy of the model. Accuracy measures such as the confusion matrix, recall, overall accuracy, and Area under the curve were calculated and are shown below.

The confusion matrix measured the ability of the model to classify the stocks into their correct class. From the confusion matrix, 144 plus 8 stocks were classified correctly as downward trend and upward trend stocks respectively. However, one downward trend stock was classified as an upward trend stock and two upward trend stocks were classified as downward trend stocks. TABLE 2 below shows the "Recall" and "Overall" Accuracy.

TABLE 2: Recall & Overall Accuracy

Recall Accuracy	80%
Overall Accuracy	98%

The logistic regression model results were considered good.

CONCLUSION & DISCUSSION

Three machine learning techniques (Principle Component Analysis, Hierarchical Cluster Analysis, 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 are likely to go up.

Thirty thousand dollars was invested in the top three stocks, SPO, EBO, and CNU and at the end of the two weeks the portfolio was worth $31 705.81, which is a return of 6% versus the All Ordinaries stock

market index return of 1%. The stock portfolio performed six times better than the stock market.

In a nutshell, machine learning applications are valuable for investors as the analytical results can accurately identify which stocks are good and likely to go up. True profitability was realized for our stock portfolio.

Most importantly, this study demonstrated that the 24 variables that were 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 1% versus the stock portfolio returns of 6%.

REFERENCES

[1] Fundamental and Technical Analysis www.investopedia.com/ask/answers/131.asp

[2] https://www.sungard.com/~/media/fs/capitalmarkets/resources/white-papers/Valdi-Algo-Trading-Complex-Map.ashx?sfdcCampaignId=70150000000Yyzo.

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

[4] Hargreaves, C.A; Dixit, P; Solanki, A (2013). Stock Portfolio Selection using Data Mining Approach. IOSR Journal of Engineering. Vol 3, Issue 11, 42-48.

[5] Hargreaves, C.A; Hao,Yi (2012)." Does the use of Technical & Fundamental Analysis improve Stock Choice? A Data Mining Approach applied to the Australian Stock Market". IEEE Explore

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

[7] Kadirvel Mani,C, Hargreaves, C.A (2016). Stock Trading using Analytics. American Journal of Marketing Research, Vol 2, No 2, 27-37.

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

[9] Huang; Q. Cai, Y. Peng, J. "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.

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

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

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

[13] 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

[14] Mbeledogu, N.N., Odoh, M., Umeh, M.N. (2012). "Stock feature extraction using Principle Component Analysis". International Conference on Computer Technology and Science. IACSIT Press, Singapore DOI: 10.7763/IPCSIT.2012.V47.44

[15] Loretan, M. (1997). "Generating market risk scenarios using principle component analysis: methodological and practical considerations". Federal Reserve Board. https://www.bis.org/publ/ecsc07c.pdf

[16] 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

[17] 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.

[18] KMO & Bar tlett's sphericity test. http://eric.univlyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_KMO_Bartlett.pdf, http://staff.neu.edu.tr/~ngunsel/files/Lecture%2011.pdf

[19] Menard, S.W., NetLibrary,I. "Applied logistic regression analysis". Thousand Oaks, Calif: Sage Publications.

[20] Hosmer, D.W., Lemeshow, S. & Sturdivant, R.X. "Applied Logistic Regression (third ed.) Hoboken, N.J: Wiley.

★★★