ISSN: 0976-2876 (Print)
ISSN: 2250-0138 (Online)

MACHINE LEARNING APPLICATION IN THE FINANCIAL MARKETS INDUSTRY

CAROL ANNE HARGREAVES[a1], VALLARU CHANDANA REDDY[b] AND RAGHUPATHY VISHNWARDHAN REDDY[c]

ABSTRACT

Machine Learning is increasingly prevalent in Stock Market trading. The goal of this paper is to investigate whether the machine learning technique is able to retrieve information from past prices and predict price movement and future trends. We explore using trend trading indicators in a machine learning based model. We propose algorithms that combine different technical and fundamental indicators in order to provide accurate positive indicators for stock price movements. In this paper, machine learning techniques such as the Logistic Regression, Decision Tree, Neural Networks and Artificial Intelligence are applied to big data from the stock market that is of high volume, high velocity, high variety and high variability using real time and off-line data of different time granularities. The results of predictive algorithmswereanalysed and the results presented. Experimental results confirmed that the use of machine learning and artificial intelligence methods can help to select top performing stock portfolios that outperform the stock market.

KEYWORDS: Algorithms, Decision Trees, Financial Markets, Logistic Regression, Machine Learning, Neural Networks, Stock Market Trading.

The stock market is one of the most important and attractive markets in the financial industry.Though the stock market fluctuates randomnly from day to day,experienced traders know that much of the stock market fluctuations are not random.Ithas been demonstrated that onecan accurately predict daily stock market activity a lot of the time. Jheng-Long Wu, Liang-Chih Yu and Pei-Chann Chang (2014) showed that predictive automatic trading decisions using comprehensive features are significantly better than traditional methods using using numeric features alone. The primary challenges with stock trading is the identification of profitable stocks and trading the stocks without human error and interferencr of personal sentiments in order to reap better returns. ChandrikaKadirvel Mani and Carol Anne Hargreaves (2016) combined the usage of both fundamental and technical variables for the prediction of profitable stocks using thesupport vectors machine learning algorithm. They systematically identified high returning healthcare stocks, and traded them with the help of an auto trading application without human error and sentiment interference and yielded 16.64% revenue at the end of three months trading. Further, Carol Anne Hargreaves, PrateekDixti and AnkitSolanki investigated whether a healthcare sector stock portfolio selected using the logistic regression will outperform the ASX All-Ordinaries Index and Healthcare Sector Index (XHJ) over the twenty days trading period. The healcthare stock portfolio returned 18.24%.

In this paper, we focus on the Small Cap Stocks that trade on the Australian Stock Market and again combine technical and fundamental data using machine learning applications and a proprietary trading method.

The purpose of this paper is to determine whether a data driven machine learning approach could be used to select good stocks that will outperform the stock market index. The three top training stocks from March 2017 werepapertraded in April 2017. Similarly, thethreetop training stocksfrom April 2017 were paper traded in May 2017 and topthreetrainingstocks from May 2017 were paper traded in June 2017 to validate the selection of good profitable stock portfolios and the consistent and reliable methodology.

In this paper,important algorithms such as the logistic regression,decision tree and neural networks are applied on the stock data to understand the future trends and price movements.

This paper hasfive sections. While Section 1 is the introduction, Section 2, a brief overview of the mdthodology, Section 3, the statistical analysis results, Section 4, the conclusion and Section 5, the referencs.

METHODOLOGY

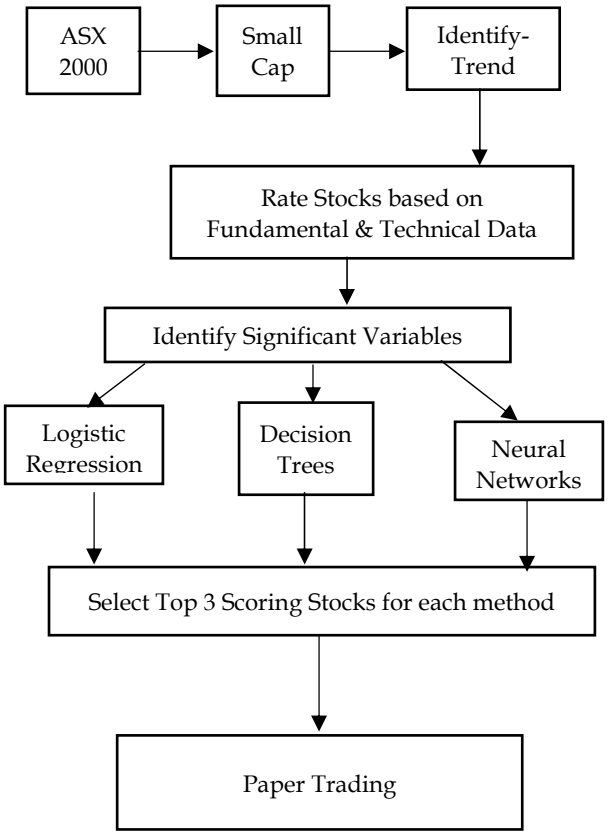
Data Capture

Daily stock data such as open, high, low, close and volume was downloaded from www.eoddata.com for two thousand stocks for the period of 2 January 2016 to 30 June 2017. The Small Cap Indexand All Ordinaries Index daily stock datacamefrom au.finance.yahoo.com. Small Cap stocks are the focus of this paper as we believe that Small Cap Stocks have the most price growth potential and we would like to focus on short term trading of

one month. While Small Cap Stocks are likely to be higherrisk stocks, we manage this risk by using financially health stocks for trading. In the following section, section 2.2.,we present the data analysis procedure.

Data Analysis

The following flow explains how data was captured and steps on further data analysis.



Data Modelling

As the target variable is 'Trend', a categoriacal-variable,three algorithms were used for data modelling.They are Logistic Regression, Decision Trees, Neural Networks. Twenty-four combinations of the five variables, open, high, low, close, and volume were derived and used as inputs to the model. Variables such as the '20 Moving Average', '200 Moving Average', '50 Moving Average', and similar were derived for modelling purposes.

Logistic Regression

TheLogistic Regression is a special type of regression where the binary response variable is related to a set of explanatory predictor variables which can be wither continuous or discrete.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.

In logistic regression,probability of taking that response is modelled based oon the combination of values taken by predictor variables.Logistis Regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of characteristic of interest:

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

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

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values. 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. 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.Arun U, Gautam B and Avijan D (2012) 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.

Decision Trees

Decision trees are an important type of algorithm for predictive modelling. In particular, Tsai, C.F. and Wang, S.P. (2009) have shown that decision trees have excellent ability to describe cause and effect relation of information. Their experimental decision tree and neural network hybrid model had 77% forecasting accuracy in electronic industry and supplied reliable rules to assist investors to determine when to perform their investment. Decision trees are among the most powerful techniques used in machine learning.

The algorithm is recursive in nature as the groups formed cab be sub-divided using the same strate-

gy. Due to this process,this algorithm is also known as greedy algorithm, as there is excessive desire of lowering the cost.

As this paper is focused on binary classification, the cost function used in classification to find most homogeneous branches or branches having groups with similar responsesis theGiniscore.AGini score gives an idea of how good a split is by how mixed the response classes are in the groups created by the split.

G=sum(pk*(1-pk))

Here, pk is the proportion of same class inputs present in a particular group. A prefect class purity occurs when a group contains all inputs from the same class,in which case pk is either 1 or 0 and G=0,where as a node having a 50-50 split of classes in a group has the worst purity,so for a binary classification it have pk=0.5 and G=0.5.

Neural Networks

The application of neural networks in machine learning in stock trading has predominantly gained importance in recent times.The theory of a neural network computation providesinteresting techniques that replicate the human brain andnervous system.

Neural networks are typically organized in layers and layers are made up of a number of interconnected nodes which contain an activation function. Generallyspeaking, a neural network is a set of connected input andoutput units where each connection has a weight associatedwith it. During the learning phase, the network learns byadjusting the weights so as to be able to correctly predict orclassify the output target of a given set of input samples. Given numerous neural network architecture, multi-layer feed-forwardneural networks were implemented in this study to comparetheir predictive ability against the logistic regression, decision trees model and an artificial intelligence system.

Neural networks have beenwidely used for financial forecasting due to its ability tocorrectly classify and predict the dependent variable. Tong-SengQuah (2008) used neural networks for stock selection and showedthat there is positive relationship between predictions of the trained networks with the equities appreciation, which may result in better earnings for investment. During the training phase,the predictor variables are fed making it an input layer.The weighted inputs are fed to the hidden layer and the weighted inputs of the hidden layer are fed into the other hidden layers.This process continues till the last hidden layer in the network.The weights of the last hidden layer are fed to the output layer which gives predictions for a given set of input samples.

Paper Trading

Each stock portfolio had an investment of $3 000, 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.

RESULTS

Modelling and Paper Trading Results

The ‘area under the curve’ and ‘recall’ are key metrics for evaluating machine learning model results. As a rule of thumb ‘recall’ and ‘area under the curve should be at least 0.7.The ‘recall’ and ‘area under the curve’ measures were very good, so paper trading went ahead.

Table 1: Paper Trading Results for April 2017

ALGORITHMS	RECALL	AREA UNDER CURVE	RETURN ON INVESTMENT (%)
LOGISTIC REGRESSION	0.97	0.926	0.39
DECISION TREES	0.85	0.935	0.94
NEURAL NETWORKS	0.93	0.969	0.39

Table 1 above, presents the results for the paper trading of the 3 stock portfolios, Neural Networks, Decision Trees and Logistic Regression. The market return on investment was 0.64%. The decision tree portfolio outperformed the stock market performance. But the Logistic Regression and Neural Network Model Portfolios were less than the market performance but not significantly different from the market performance.

Again, the ‘recall’ and ‘area under the curve’ metrics were generally very good. The market return on investment for May 2017 was -3.56%. All 3 models predicted the same top 3 stocks which performed extremely well.

Table 2: Paper Trading Results for May 2017

ALGORITHMS	RECALL	AREA UNDER THE CURVE	RETURN ON INVESTMENT (%)
LOGISTIC REGRESSION	0.95	0.936	8.13
DECISION TREES	0.71	0.969	8.13
NEURAL NETWORKS	1	0.985	8.13

Table 3: Paper Trading Results for June 2017

ALGORITHMS	RECALL	AREA UNDER THE CURVE	RETURN ON INVESTMENT (%)
LOGISTIC REGRESSION	0.95	0.936	3.56
DECISION TREES	0.71	0.969	2.62
NEURAL NETWORKS	0.92	0.985	5.74

Again, the 'recall' and 'area under the curve' metrics were generally very good. The market return on investment for June 2017 was -0.15%. All 3 stock portfolios outperformed the market by at least 2.5 times.

CONCLUSION

This study investigated whether a data driven approach could be used for selecting good small cap stocks that will outperform the stock market index. Machine Learning models such as the Logistic Regression, Decision Tree and Neural Network was used and the experimental results confirmed that the machine learning models were very good at identifying top performing stock portfolios that outperform the stock market. In all three consecutive time periods, stock portfolio performance was positive.

In the April 2017 period, the performance of the decision tree stock portfolios was better than the stock market but quite similar to the stock market for the logistic regression and neural network stock portfolios. In the May 2017 and June2017 periods the stock portfolios performed at least 2.6 times better than the stock market.

This paper has contributed to the fact that data driven approaches such as machine learning techniques can accurately predict which stocks are likely to go up. It is recommended that many more experiments be performed using different sectors and market cap size stocks.

REFERENCES

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.

Carol Anne Hargreaves, Prateek Dixit, AnkitSolanki. (2013). Stock Portfolio Selection using Data Mining Approach. IOSR Journal of Engineering. Vol 3, Issue 11, 42-48.

ChandrikaKadirvel Mani, Carol Anne Hargreaves (2016). Stock Trading using Analytics. American Journal of Marketing Research, Vol 2, No 2, 27-37.

Jheng-Long Wu, Liang-Chih Yu, Pei Chann Chang (2014). An intelligent stock trading system using comprehensive features. Applied Soft Computing (23), 39-50.

Tong-SengQuah(2008). DJIA stock selection supported by neural network. Expert systems with applications (35), 50-58

Tsai, C.F, Wang, S.P. (2009). Stock Price Forecasting by Hybrid Machine Learning Techniques.Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2009 Vol I, ISBN: 978-988-17012-2-0