TAF Journal of Administrative and Business Studies
2016, 2(3): 107-112 JABS

PRIMARY RESEARCH

The prediction of stock returns with regression approaches and feature extraction

Chrits Liew [1, *], Tsung-Nan Chou [2]

[1, 2] Chaoyang University of Technology, Taiwan

Keywords:
Financial ratios
Feature extraction
Regression approach

Received: 29 March 2016
Accepted: 22 April 2016
Published: 21 June 2016

Abstract. Value investing is one of the most popular investment strategies for investors to search for the undervalued stocks based on their financial reports and balance sheets. However, the numerous metrics derived from the financial statements are not easy for the investor to analyze and determine the financial health of a company. The main purpose of this study is to employ feature extraction to identify a smaller number of financial ratios for the prediction of stock return which reflects the quality of a company. Two regression approaches, including Multilayer Perceptron model and Meta Regression by discretization model, were incorporated with feature extraction to evaluate the forecast performance for two different industries in Taiwan. The results demonstrated that the prediction errors were improved for both models by the feature extraction strategy which reduced the original 16 variables to5 variables. Besides that, both models achieved better prediction result in concrete industry rather than rubber industry. Finally, this paper concluded that the overall performance of the Multilayer Perceptron model is better than the other model.

INTRODUCTION

The reason why fundamental analysis was used in this paper is because it was inspired by Benjamin Graham and Warrant Edward Buffet. They discuss about value investment. Both of them used fundamental analysis to analyze actual value of a company, and gain huge amount of profit. According to the successful study of Graham and Buffet, I believe that study about fundamental analysis to predict stock price is worth to study. Besides that, many articles come out with financial information such as P/E, cash flow rate, asset turnover, ROA, net profit, equity growth rate and et cetera that will affect the future stock price. Our finding is revealed to be significant between fundamental variable and future stock return in Taiwan equity market. The past article shows that Taiwan equity market frequently rose and fell according to the information release, means that Taiwan equity market is predictable. In this paper, we use Multilayer Perceptron model, Meta Regression by discretization model and fundamental variable to build a model to predict future stock price. We added fundamental variables such as asset utilities, profitability, liquidity, growth and valuation to test the correlation between the influences of prediction model. This paper focuses on comparing which model is better in stock return prediction, and employs featur

*Corresponding author: Chrits Liew
E-mail: chrits901016@gmail.com

extraction to identify a smaller number of financial ratios for the prediction of stock return which reflects the quality of a company. The research framework was summarized in Figure 1.

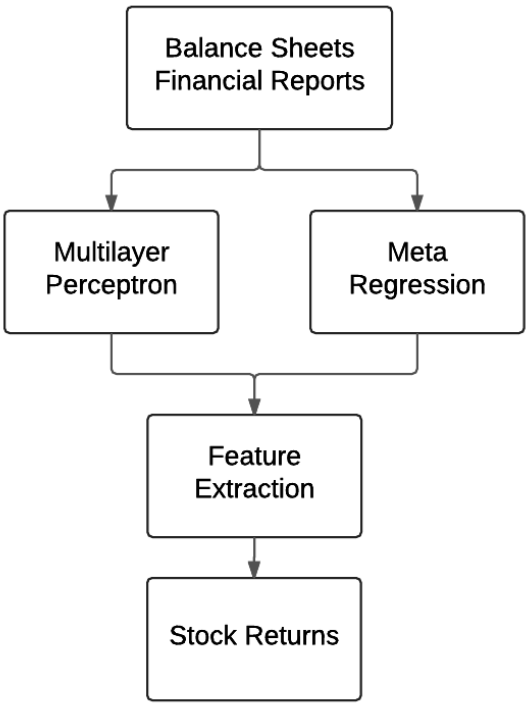

FIGURE 1. Research framework

LITERATURE REVIEW

From a human's point of view, it is very difficult to detect trends and extract patterns from data which are imprecise and complex. However, some research works have reported that neural network helps us to work on it. Hafezi, Shahrabi & Hadavandi (2015) propose a new intelligent model in a multi-agent framework called Bat-Neural Network Multi-Agent System (BNNMAS) to predict stock price. The results show that BNNMAS significantly performs accurately and reliably, so it can be considered as a suitable tool for predicting stock price especially in long term periods. In addition, (Siddiqui & Abdullah, 2015) present that using Artificial Neural Network (ANN) model to predict stock price in India, the preliminary data testing yielded encouraging results for the model. Finally the predicted values of stock have a testing accuracy of more than 85%.Some researches applied hybrid model to increase the prediction accuracy. Salehi, Mousavi Shiri, & Bolandraftar Pasikhani (2016) gave description of prediction of the financial distress of Iranian firms, with four techniques: support vector machines, artificial neural networks (ANN), k-nearest neighbor and naïve. Besides that it is also the first study in Iran which used such methods for analyzing the data. Therefore the results might help in the Iranian condition as well for other developing nations. Barak & Modarres (2015) showed that the proposed hybrid model is a proper tool for effective feature selection and these features are good indicators for the prediction of risk and return. Patel, Shah, Thakkar & Kotecha (2015) presented the comparison model between four prediction models, Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and naive-Bayes with two approaches as the input to these models. Experimental results also reveal that the performance of all the prediction models improves when these technical parameters are represented as trend deterministic data. On the other hand (Ince & Trafalis, 2008) developed two models in order to forecast the short term stock price movements by using technical indicators. Comparison shows that support vector regression (SVR) outperforms the multilayer perceptron (MLP) networks for a short term prediction in terms of the mean square error. However, if the risk premium is used as a comparison criterion, then the support vector regression technique is as good as the multilayer perceptron method.

METHODOLOGY

In order to investigate the performance of a number of learning algorithms for predicting the stock return, two different models and related performance measures are briefly described as follows. Normally, the research works might be unable to derive efficient prediction results through the machine learning models based on high dimensional and voluminous data. The performance of prediction models will be limited if the analysis applies the raw data directly to train, verify and test their models. As a result, to reduce the complexity of computation of data analysis, and increase the model accuracy, the original data need to be transformed from high-dimensional space to a space of fewer dimensions. The Principal Component Analysis (PCA)was applied in this study to reduce the original 16 variables to5 variables during the stage of feature extraction (Witten & Frank, 2005). The one of PCA consistence equations of 16 variables can be portion shown as below:

0.432F+0.425E+0.422D+0.405G+0.313M…

In this equation F= Net profit margin, E= ROA, D= EPS, G= Net rate of return, M= Net income growth rate. The first model evaluated in this study is the Multilayer Perceptron

ISSN: 2414-309X
DOI:10.20474/jabs-2.3.1

(Witten & Frank, 2005) which is a feed forward artificial neural network model based on back propagation learning algorithm. The model maps a set of independent variables onto a set of appropriate outputs which represent the decision variables or dependent variables. The mapping process is fulfilled through a network structure that consists of multiple layers of neural nodes with a nonlinear activation function for each. In contrast to the linear regression model, the output of neural network is a nonlinear function of the independent variables. In addition, the sigmoid function applied as activation function for network is a hyperbolic tangent which ranges from -1 to 1. The hidden layer of the neural network is defined as the average of the number of independent variables and number of dependent variables. For speeding up the convergence of the training process, the learning rate and momentum are set to 0.5 and 0.2 respectively. Then set the Nominal to Binary Filter as TRUE. This could help to improve performance if there are any nominal attributes in the data. Validation Threshold set as 20, Validation Threshold is used to terminate validation testing. The value here determines how many times in a row the validation set error can get worse before training is terminated. Normalize Attributes is to improve the performance of the network, furthermore this will normalize nominal attributes as well, therefore I set the Normalize attribute as TRUE. Besides that, the number of decimal places is signed as 2. It is used for the output of numbers in the model. The setting of batch size indicates the preferred number of instances to process if batch prediction is being performed. More or fewer instances might be provided, but this gives implementations a chance to specify a preferred batch size. Hence I set the Batch size as 100.ValidationSetSize set as 0 means that no validation set will be used and instead the network will train for the specified number of period. After that, the training time is set as 500, and training time means the number of period to train. For the purpose of improving general performance of the network training, the decaying learning rate started from the original value of 0.5 applied to help the network reduce the divergence between desired and actual output. Meanwhile, the input variables are also normalized to improve performance of the network. Another applied methodology is Meta Regression by discretization (Witten & Frank, 2005) which is an extension to subgroup analyses that divides a set of numerical values into a set of intervals for the class attribute. In other words, the continuous attribute will be divided into a number of bins using equal-width discretization, and then specified base regression learners will be employed to predict outcomes. The predicted outcomes based on the synthesis of each discretized interval will allow the effects of multiple factors to be investigated simultaneously. To implement the Meta Regression by discretization, the C4.5 algorithm of decision tree was used as the based learner. The number of discretized interval was set to 10 bins.

```
A <= 0.42
| N <= -28.5
| | I <= 62.89
| | | E <= -1.92: '(9.73018-25.1409]' (3.0/1.0)
| | | E > -1.92
| | | | D <= 0.11
| | | | | O <= 0.39
| | | | | | D <= -0.75: '(-21.09126--5.68054]' (2.0)
| | | | | | D > -0.75: '(40.55162-55.96234]' (2.0)
```

FIGURE 2. Total asset turnover

In figure 2,if total asset turnover is smaller or equal to 0.42, operating margin growth rate smaller or equal to -28.5, quick ratio smaller or equal to 62.89 and ROA smaller or equal to -1.92, imply that the output of Meta regression is between 9.73018 and 25.1409. Finally, the resulting size of pruned decision tree was 197 with 99 leave nodes.

```
>=-16.64
M <= -2.75
| O <= 131.66
| | C <= 8.2
| | | O <= -164.76
| | | | C <= 6.48: '(-40.79422--20.29938]' (3.0/1.0)
| | | | C > 6.48: '(-61.28906--40.79422]' (2.0/1.0)
| | | O > -164.76
| | | | F <= -2.54
| | | | | C <= 6.08: '(-40.79422--20.29938]' (5.0/1.0)
```

FIGURE 3. Total asset turnover

In figure 3, if net profit margin is bigger or equal to -16.64, net income growth rate smaller or equal to -.2.75, net growth rate smaller or equal to 131.66, receivable

ISSN: 2414-309X
DOI:10.20474/jabs-2.3.1

turnover smaller or equal to 8.2, net growth rate smaller or equal to -164.76 and receivable turnover smaller or equal to 6.48, imply that the output of Meta regression is between -40.79422 and 20.29938. Finally, the resulting size of pruned decision tree was 231with 116 leave nodes. After building the Meta-Regression model and Multilayer Perceptron model, several criteria are applied to evaluate and compare the performance of models. Both the mean absolute error (MAE) and root mean squared error (RMSE) can be used to measure the accuracy of forecast result for continuous variables. The MAE is calculated through the absolute values of the differences between actual target and the corresponding predicted target. On the contrary, the differences are each squared and then averaged over the sample prior to the square root calculation taken for the RMSE. The mean absolute error is given by the following equation, where ai is the actual target and pi represents the predicted target.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|p_i - a_i|}{n} \quad (1.3)$$

Normally the RMSE will be larger or equal to the MAE as the prediction errors are calculated by squared rather than absolute calculation. The root mean squared error is very common error metric and is described as following equation:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \quad (1.4)$$

Unlike the MSE and RMSE looking at the average difference between actual target and predicted target, the relative absolute error (RAE) and root relative squared error (RRSE) try to divide the differences by the variation of actual target, and result in a measure scale from 0 to 1 which can be used for the variables with different units. Basically, the RAE and RRSE can be described as the following equations:

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n}|P_{ij} - a_i|}{\sum_{i=1}^{n}|\bar{a}_{ij} - a_i|} \quad (1.5)$$

$$\mathrm{RRSE} = \frac{\sum_{i=1}^{n}(P_{ij} - a_i)^2}{\sum_{i=1}^{n}(\bar{a}_i - a_i)^2} \quad (1.6)$$

EMPIRICAL RESULTS

The dataset used in our research work was gathered from the database of Taiwan Economic Journal (TEJ). The dataset consists of 16 quarterly financial variables form 1995 until 2012 including Total assets turnover, Inventory turnover, receivable turnover, EPS, ROA, Net profit margin, Net rate of return, Current ratio, Quick ratio, Debt ratio, Revenue growth rate, Total growth rate, Net income growth rate, Operating margin growth rate, Net growth rate, P/B and stock return. As the amount of such data is very large and complicated, the decision was taken to randomly choose two different industries, rubber and concrete, to run this experiment. Totally, eight companies were included from which 9,656 data were collected for rubber industries, and another seven companies were included from which 9,656 data were collected for concrete industries respectively. At the beginning, when the data were collected, we removed the abnormal value in the data. After that we separated the data into training data set during the period from 1995 through 2009, and test data set for the period from 2010 to 2012.After the data had been prepared, the next step was to build the Multilayer Perceptron model and Meta Regression model. Both models were selected because their construction does not require any domain knowledge and can handle high dimensional data. Another benefit is that the result inductions of Multilayer Perceptron model and Meta Regression are simple and fast. Moreover, for the purpose of increasing the accuracy of the result, feature extraction is used in this paper to reduce the original 16 variables to5 variables, and prove that whether feature extraction strategy is capable of improving the accuracy of the result. The final step of the experiment is to compare which model is better in the prediction of stock return. The criteria used to evaluate and compare performance in this paper include MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), RAE (Relative Absolute Error) and RRSE (Root Relative Squared Error).The prediction results for the rubber industries were summarized in Table 1. The MAE values of Multilayer Perceptron and Meta Regression are 19.12 and 23.59 respectively. Therefore, the Multilayer Perceptron performs better than Meta Regression by 4.47. As a result, RMSE of Multilayer Perceptron is 24.54 and Meta Regression is 30.14, Multilayer Perceptron is better than Meta Regression by 5.60. RAE of Multilayer perceptron is 168.72% and Meta Regression is 208.16%, Multilayer perceptron is better than Meta Regression by 39.44%. RRSE of Multilayer perceptron and Meta Regression is 169.85% and 208.59% respectively. RRSE of Multilayer perceptron is better than Meta Regression by about 38.74%. Furthermore, table demonstrates the MAE,

ISSN: 2414-309X
DOI:10.20474/jabs-2.3.1

RMSE, RAE and RRSE of Multilayer Perceptron model from 23.39, 106.38% and 108.62% after using the feature extraction strategy. In comparison, the MAE, RMSE, RAE and RRSE of Meta Regression model also from 23.59, 30.14, 208.16% and 208.59% improve to 20.92, 28.56, 143.26% and 132.61% after using the feature extraction strategy. The prediction result for the concrete industries was summarized in Table 2. The MAE values of Multilayer Perceptron and Meta Regression are 11.56 and 16.49 respectively. Therefore, the Multilayer Perceptron performs better than Meta Regression by 4.93. RMSE of Multilayer Perceptron and Meta Regression is 16.59 and 20.41 respectively, Multilayer Perceptron is 3.82 better than Meta Regression. RAE of Multilayer perceptron and Meta Regression is 135.19% and 192.91% respectively, Multilayer perceptron is 57.73% better than Meta Regression. RRSE of Multilayer perceptron and Meta Regression is 150.08% and 184.65% respectively, 19.12, 24.54, 168.71% and 169.85% improve to 15.54, Multilayer perceptron is 34.57% better than Meta Regression. As the result, table 2 shows that Multilayer Perceptron model is still better than Meta Regression model. As illustrated in Table 1 and 2, the Multilayer Perceptron also outperforms the Meta Regression in all performance metric the result from this experiment shows that Multilayer Perceptron model and Meta Regression model are capable to be used in stock return prediction. However the result shows that the MAE, RMSE, RAE and RRSE of Multilayer perceptron model are better than Meta Regression model.This implies that Multilayer Perceptron is better in stock return prediction. Furthermore, Table 1 and Table 2 also show that after using PCAAnalysis results in this paper have improved. This implies that the (PCA) is able to improve the accuracy of the prediction in this paper.

TABLE 1. Prediction results of the rubber industries

	Multilayer Perceptron			**Meta Regression by discretization**		
	Original	**Feature extraction**	**Difference**	**Original**	**Feature extraction**	**Difference**
MAE	19.12	15.54	3.58	23.59	20.92	2.67
RMSE	24.54	23.39	1.15	30.14	28.56	1.58
RAE	168.72%	106.38%	62.33%	208.16%	143.26%	64.90%
RRSE	169.85%	108.62%	61.23%	208.59%	132.61%	75.98%

TABLE 2. Prediction results of the concrete industries

	Multilayer Perceptron			**Meta Regression by discretization**		
	Original	**Feature extraction**	**Difference**	**Original**	**Feature extraction**	**Difference**
MAE	11.56	9.07	2.49	16.49	11.93	4.56
RMSE	16.59	11.23	5.36	20.41	15.92	4.50
RAE	135.19 %	105.83%	29.36%	192.91%	139.26%	53.65%
RRSE	150.08%	101.72%	48.36%	184.65%	144.11%	40.54%

CONCLUSION AND FUTURE WORK

In this research we proposed to find out which model is better in stock return prediction by using data mining technique, Multilayer Perceptron model or Meta Regression model. We used both classifiers on the historical financial data of the industries to compare the result. The prediction results on this paper suggest that Multilayer Perceptron model is better in stock return prediction. However, both models can be useful tools for the investors to make a right decision on their stocks based on the analysis of the historical financial data. By the way, the model used in this paper is still not prefect because many financial factors including, but not limited to, politics issue, economic factor, direction of the institutional investor and investors' expectations influence stock market. Future work in this paper still offers huge space for examination and improving the model by assessingall the industries or any companies in the capital market in Taiwan. In addition, the evaluation of a larger collection of learning techniques such as rough sets,

ISSN: 2414-309X
DOI:10.20474/jabs-2.3.1

decision tree and Bayes net can show a great field for the future research. Furthermore, thinking over the factorsthat will be affecting the behavior of the capital market, such as direction of the institutional investor, trading volume, historical stock price, and economic issue which might influence stock market can be another great field for future studying.

REFERENCES

Barak, S., & Modarres, M. 2015. Developing an approach to evaluate stocks by forecasting effective features with data mining methods. ***Expert Systems with Applications***, 42(3): 1325-1339. **DOI:**10.1016/j.eswa.2014.09.026

Hafezi, R., Shahrabi, J., & Hadavandi, E. 2015. A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price. ***Applied Soft Computing***, 29: 196-210. **DOI:**10.1016/j.asoc.2014.12.028

Ince, H., & Trafalis, T.B. 2008. Short term forecasting with support vector machines and application to stock price prediction. ***International Journal of General Systems***, 37(6): 677-687. **DOI:**10.1080/03081070601068595

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. ***Expert Systems with Applications***, 42(1): 259-268. **DOI:**10.1016/j.eswa.2014.07.040

Salehi, M., Mousavi Shiri, M., & Bolandraftar Pasikhani, M. 2016. Predicting corporate financial distress using data mining techniques: an application in Tehran Stock Exchange. ***International Journal of Law &Management***, 58(2): 216-230. **DOI:**10.1108/ijlma-06-2015-0028

Siddiqui, T.A., & Abdullah, Y. 2015. Developing a nonlinear model to predict stock prices in India: An artificial neural networks approach [dagger]. ***IUP Journal of Applied Finance***, 21(3): 36-49

Witten, I.H., & Frank, E. 2005. ***Data mining practical machine learning tools and techniques***. San Francisco, CA: Morgan Kaufmann.

— This article does not have any appendix. —

ISSN: 2414-309X
DOI:10.20474/jabs-2.3.1