Regression Techniques for the Prediction of Stock Price Trend

Han Lock Siew
Malaysian Institute of Information Technology
Universiti Kuala Lumpur
Kuala Lumpur, Malaysia
lshan@miit.unikl.edu.my

Md Jan Nordin
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia
jan@ftsm.ukm.my

***Abstract* — This paper examines the theory and practice of regression techniques for prediction of stock price trend by using a transformed data set in ordinal data format. The original pre-transformed data source contains data of heterogeneous data types used for handling of currency values and financial ratios. The data formats in currency values and financial ratios provide a process for computation of stock prices. The transformed data set contains only a standardized ordinal data type which provides a process to measure rankings of stock price trends. The outcomes of both processes are examined and appraised. The primary design is based on regression analysis from WEKA machine learning software. The stock price movement in Bursa Malaysia is used as our research setting. The data sources are corporate annual reports which included balance sheet, income statement and cash flow statement. The variables included in the data set were formed based on stock market trading fundamental analysis approach. Classifiers in WEKA were used as algorithms to produce the outcomes. This study showed that the outcomes of regression techniques can be improved for the prediction of stock price trend by using a dataset in standardized ordinal data format.**

Keywords- regression techniques; ordinal data type; machine learnin; fundamental analysis; classifiers; linear regression

I. INTRODUCTION

Stock markets provide opportunities along with associated risks for investors to make profits. Scholars and professional investors have done many research and developed theories and hypothesis on stock market timing techniques. This paper research the usage of regression techniques as predictive analytic on stock price trends. The formulation of dataset is based on fundamental analysis approach which is a dominant school of thought in investing. The features used in the dataset are variables made up of statistical ratios. In this study, all data in numerical values are transformed into ordinal or enumerated values to form the dataset. Regression based classifiers from WEKA are then used as predictive analytics to test the ordinal data. The outcomes were compared and evaluated.

II. LITERATURE REVIEW

Since the existence of stock markets, a lot of research had been done in developing models to make predictions on stock price movements. Two models are well known and they are the efficient market hypothesis and the random walk theory.

In the nineteen sixties, Eugene Fama wrote a PhD dissertation on the efficient market hypothesis. Fama's argument indicated that stock price will be priced appropriately and reflect all available information in an active market. Since market is efficient, there is no quality analysis can be carried out to result in superior performance of an appropriate benchmark [1].

Louis Bachelier's PhD dissertation titled "The Theory of Speculation" was a research on the random walk hypothesis's concept. The hypothesis argues that stock market price evolves according to random walk and that the market stock price cannot be predicted. The hypothesis is in line with the efficient-market hypothesis [2].

Professional investors favor two dominant schools of thought on investing which are fundamental analysis and technical analysis.

Fundamental analysis approach identifies prospective stocks by analyzing their fundamental attributes. Statistics from financial reports such as balance sheet, cash book and profit and loss statement are used to study the intrinsic values of companies [3]. Financial ratio statistics that include operating performance, corporate valuation, growth equilibrium, financial leverage and corporate liquidity form the basis of fundamental attributes [4].

A related strategy of fundamental analysis is called the contrarian strategy. This strategy combines factor of human emotional biases with fundamental analysis [5]. A contrarian believes that over reaction of greed and fear in crowd behavior leads to exploitable mispricing in stock prices. A contrarian investor takes a contrary position in buying shares of stocks that are performing poorly and then selling them when they perform well [6].

Technical analysis approach identifies chart patterns based on a company's historical share price. This approach does not gain insight into the business side of a company; it assumes the available public information does not offer a competitive trading advantage. This technique predicts trends in advance through chart patterns [7].

The research on stock market prediction techniques has eventually moved into the technological realm. Machine learning approach is one of the common techniques. The

approach of machine learning is by examining a potentially linear or non-linear relationship exists with the availability of enough indicators [8]. Machine learning is a branch of artificial intelligence. This approach find patterns in training datasets and form their own rules which are then used for making forecasts in testing datasets [9].

Regression techniques are part of the machine learning approach. In 1805, Legendre published the method of least squares, which was the earliest form of regression. In 1821, Gauss published a further development of the theory of least squares which include the Gauss-Markov theorem. In the nineteenth century, Francis Galton used the term "regression" to describe a biological phenomenon. Galton's work was later developed into the statistical context by Udny Yule and Karl Pearson [10].

Common regression analysis involves inputs of numerical data which may consist of infinite or a wide range of values. In this research, we start by gathering numerical data in real-valued format using the fundamental analysis approach. After that we apply a new transformation process to convert the numerical values into ordinal values. The ordinal values contain only a range of categorical enumerated values. The relationships between the dependent and the independent ordinal variables are correlated based on the enumerated values.

III. Data and Methodology

A. Data

Data was collected from companies in Bursa Malaysia.

TABLE 1 – Data Collections

No.	Stock Code	Duration (year)
1	3689	2003-2010
2	3255	2004-2011
3	3921	2003-2010
4	4707	2003-2010
5	4065	2003-2010
6	7084	2004-2011
7	4588	2003-2010
8	5584	2003-2010
9	3107	2004-2011
10	9466	2005-2010
11	6033	2004-2011
12	6599	2004-2010
13	5032	2003-2010

The datasets are structured by the following features and data types.

TABLE 2 – Dataset 1

Feature	Data Type
NTA	Real-Valued
LA	Real-Valued
DE	Real-Valued
ZS	Real-Valued
AT	Real-Valued
Price	Real-Valued

TABLE 3 – Dataset 2

Feature	Data Type
NTA	Ordinal
LA	Ordinal
DE	Ordinal
ZS	Ordinal
AT	Ordinal
PriceRank	Ordinal

The types of variables under investigations were identified based on fundamental analysis approach. The selected independent variables are Net Tangible Asset (NTA), Liquid Asset (LA), Debt to Equity (DE), Altman Z-Score (ZS) and Asset Turnover (AT). Net Tangible Asset is a measure of the tangible worth of a company, minus any intangible assets. This is one possible measure of a company's share worth. Liquid Asset is an asset that can be quickly converted into cash. It is a good indicator on the financial strength of a company. Debt to Equity ratio measures the level of debt relative to the equity of a company. This measurement indicates whether a company has little or over exposure to debt. Altman Z-Score combines five financial ratios to determine the probability of bankruptcy for a company. Asset Turnover measures the ability of a company in turning out sales based on its available asset. A company which is able to generate sales efficiently indicates higher exposure to profitability [11].

The dependent variable is named Price for Dataset 1. The Price variable contains the real-valued data of the predicted price. The dependent variable is named PriceRank for Dataset 2. The PriceRank variable contains the ranking of the predicted price trend in categorical ordinal value. All the variables contain either a positive or a negative value for Dataset 2. A positive value implies positive correlation on price trend (price ranking) and vice versa for a negative value. The outcomes of both Price and PriceRank variables are dependent on their relationships with the other variables.

B. Limitations of the Data

The data for all the variables in Dataset 2 consist of only the values in the range of {-2, -1, 1, 2}. The range of values for research outcomes could increase if the data range consists of a Zero value. The same is true if more values are included in the data range.

The beginning of the financial reporting period differs among the companies. This is common since the accounting practice allows companies to choose their own accounting periods as long as each of the accounting periods consists of twelve months.

Most collected data were dated from year 2003 to year 2011. Some data were dated from 2004 or 2005 because the data were not available prior to those periods.

C. Methodology

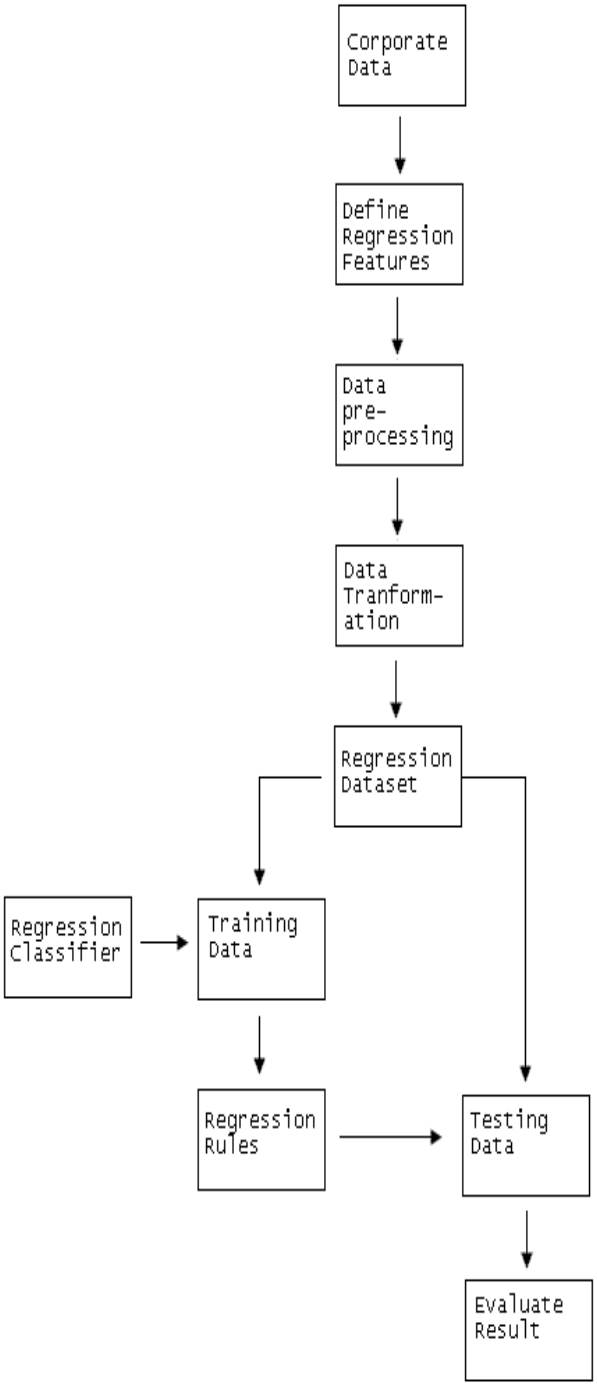

Figure 1 – Methodology

Statistics on corporations and dataset features are generated through fundamental analysis. Data was screened and pre-processed to remove out-of-bound values. This process can prevent problems of producing misleading results [12].

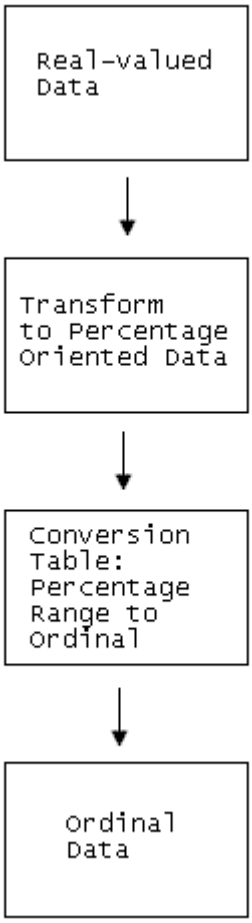

Figure 2 – Data Transformation Method

The objective of the transformation process is to make the data more structured. Pre-processed data contains general real-valued data which include the currency amount and percentage formats. In the data transformation process, the pre-processed data is standardized into percentage oriented data. A percentage to ordinal conversion table contains the ranges of percentage values associated with their ordinal enumerated values. Each enumerated value is assigned to the dataset based on the conversion table. This approach provides identical categories of enumerated values for different variables even though the range of values for one variable differs from another variable. This approach also clarifies the categories within a variable where its numerical values swing widely from one range to another range.

During the data training stage, one after another each regression classifier was used as predictive analytic on the dataset. A percentage split specifies a regression classifier to split the dataset into training data and testing data proportionally. Training data provides learning process for each classifier to formulate its own regression rules. The regression rule was used on the testing data for predictions of future stock price trends. The test result was then evaluated [13].

In a regression model, a predicted value Y is related to a function of x and b.

$$Y = f(x, b) \quad (1)$$

Y is the dependent variable, x is the independent variable and b is the unknown parameter. A linear regression model takes the form

$$Y = b_0+b_1 x_1+ ... +b_nx_n+ e \quad (2)$$

where x_1 to x_n are independent variables, and e is called the error term. A linear regression equation written in vector form [10] is

$$Y= a+bx+e \quad (3)$$

IV. FINDINGS AND DISCUSSION

A. Findings

The following shows the test results and evaluation measures for regression classifier performances based on two datasets with different data types:

TABLE 3 – Dataset 1 with Real-valued Data Type

Regression Algorithm	Correlation Coefficient	Mean Absolute Error	Root Mean Square Error
Additive Regression	0.3747	15.2275	19.0615
Linear Regression	0.3755	14.4419	18.5975
Regression by Discretization	0.0397	18.8868	24.8674
Simple Linear Regression	0.3049	15.3287	19.3164
SMO Regression	0.4173	14.2697	19.2356

TABLE 4 – Dataset 2 with Ordinal Data Type

Regression Technique	Correlation Coefficient	Mean Absolute Error	Root Mean Square Error
Additive Regression	0.5336	0.7317	0.8641
Linear Regression	0.5742	0.7141	0.8360
Regression by Discretization	-0.0378	0.9491	1.3084
Simple Linear Regression	0.5742	0.7141	0.8360
SMO Regression	0.6079	0.5462	0.8164

Dataset 1 contains the original source of data in real numbers. Dataset 2 contains the transformed values in ordinal form from dataset 1. Test on Dataset 2 demonstrates that result has improved when transformed ordinal data is used. Compare to Dataset 1, all but one regression techniques used on Dataset 2 show higher correlation coefficient rates and lower error rates. Among the regression techniques used on Dataset 2, the SMO Regression technique demonstrated a reasonable result. The technique obtained a correlation coefficient of 0.6079. Also, the technique has the lowest error rates among the regression techniques with a mean absolute error of 0.5462 and a root mean square error of 0.8164. The extra ordinary incident of the 2008 global financial crisis had contributed to the wild swing of stock prices though out the world during the period. As a result, the correlations among stock prices had been deeply affected and distorted.

B. Discussion

Rule was formed when each regression classifier learned from the training data. The classifier's rule is a regression function of the independent variables which produces the estimation target. The following is the equation of the SMO Regression model.

Prediction =

- 0.042 * (normalized) NTA
- 0.0184 * (normalized) LA
+ 0.0208 * (normalized) DE
+ 0.2803 * (normalized) ZS
- 0.0409 * (normalized) AT
+ 0.7783 (4)

The function of "(4)" is a probability distribution that describes the probability of random variables taking certain values. In this SMO Regression model, initially the PriceTrend takes the linear regression form from "(2)" as shown below:

$$Y = b_0+b_1 x_1+ ... +b_nx_n+ e$$

The value of b_0 is zero and the value of e is 0.7783. On top of the linear regression equation, the SMO Regression normalized the variables where the outputs are based on the standardized data [14]. As a result, the function of "(4)" is formed. The "+" signs denote positive correlations for the parameters, whereas "-" signs denote negative correlations for the parameters in "(4)". The interpretations of the correlations vary from one investing approach to another. Contrarian approach has tendency to go directly against the research findings by the conventional stock analysts who broadcast their views [6].

Regression techniques are useful tools for predicting the value of a dependent variable based on the values of other independent variables. The process involves a linear transformation of the selected parameters into the predicted variable. The selection of parameters is based on the least squares criterion to achieve an optimal decision rule [15].

V. CONCLUSION AND IMPLICATIONS

This research shows that the outcomes of regression techniques can be improved when the input data was standardized into a common data type through a customized transformation process. The use of an ordinal data type for prediction based on ranking system provides a different dimension for predicting outcomes.

. In this research which utilized the WEKA regression techniques, SMO Regression technique has outperformed the other regression techniques in the experiment. Data transformation process opens up another opportunity to be discovered by the targeted algorithms. The use of a different data type by transforming real numbers into categorical ordinal data can improve the outcomes of the regression techniques. The outcomes are favorable when less structured data are transformed into more structured data in ordinal form. Since there are many other data types, further research can be conducted to compare the effects of transforming various forms of data types in regression techniques used for prediction of stock price trend.

REFERENCES

[1] Beechey M, Gruen D, Vickrey J. (2000). The Efficient Markets Hypothesis: A Survey. Reserve Bank of Australia

[2] Lo, A.W. and Mackinlay, A.C. A Non-Random Walk Down Wall Street 5th Ed. Princeton University Press, 2002.

[3] Graham, Benjamin; Dodd, David (December 10, 2004). Security Analysis. McGraw-Hill. ISBN 978-0071448208.

[4] Walsh, Ciaran (2003) Key Management Ratios, Third Edition, Prentice Hall.

[5] Shefrin, Hersh (2002) Beyond Greed and Fear: Understanding behavioral finance and the psychology of investing. Oxford University Press.

[6] O'Shaughnessy, James (2009). Predicting the Markets of Tomorrow: A Contrarian Investment Strategy for the Next Twenty Years, Penguin Group. ISBN 1591841089.

[7] Kirkpatrick and Dahlquist. Technical Analysis: The Complete Resource for Financial Market Technicians. Financial Times Press, 2006, page 3. ISBN 0-13-153113-1.

[8] MacKay, D.J.C. (2003). Information Theory, Inference, and Learning Algorithms, Cambridge University Press. ISBN 0-521-64298-1.

[9] Alpaydın, Ethem (2004) Introduction to Machine Learning (Adaptive Computation and Machine Learning), MIT Press, ISBN 0-262-01211-1.

[10] Freedman, David (2005) Statistical Models: Theory and Practice, Cambridge University Press.

[11] Bodie, Zane; Alex Kane and Alan J. Marcus (2004). Essentials of Investments, 5th ed. McGraw-Hill. ISBN 0-07-251077-3.

[12] Kotsiantis, S.; Kanellopoulos, D. ; Pintelas, P. (2006) "Data Preprocessing for Supervised Leaning", International Journal of Computer Science.

[13] Theodoridis, Sergios; Koutroumbas, Konstantinos (2009) "Pattern Recognition", 4th Edition, Academic Press, ISBN 978-1-59749-272-0.

[14] Smola, Alex; Scholkopf, Bernhard (1998). A Tutorial on Support Vector Regression. NeuroCOLT2 Technical Report Series - NC2-TR-1998-030.

[15] Tofallis, C (2009). "Least Squares Percentage Regression". Journal of Modern Applied Statistical Methods 7: 526–534.