Far East Journal of Experimental and Theoretical Artificial Intelligence
Volume 5, Numbers 1-2, 2010, Pages 57-67
Published Online: July 22, 2010
This paper is available online at http://pphmj.com/journals/fejetai.htm

LOGISTIC REGRESSION MODEL IN KNOWLEDGE DISCOVERY

YEN-PING HUANG

Chin Min Institute of Technology
110, Hsueh-Fu Road
Tou-Fen, Miao-li 350, Taiwan, R. O. C.
e-mail: sunny@ms.chinmin.edu.tw

Abstract

Today, many countries in the world are engaged in a massive effort in influencing practice patterns of escaping economic contraction and assigning accountability by constructing systematic mechanisms. As knowledge discovery providers play a critical role in industries and investors, many strategies have been adopted to control the factors that may influence the return of financial services offered by investors. In view of the above, this research employs data-mining technology to predict the probability of patterns for financial database. By using the significant explanatory variables obtained from logistic regression, the relationship between output and input variables can be explained. Finally, the most important factors that affect the probability of returns are the attributes of price, trading volume, turnover rate, size in circulation, and book-to-market equity.

1. Introduction

Nowadays, knowledge discovery in database is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data (Fayyad et al. [10, 11]). Knowledge discovery is a field within the area of

Keywords and phrases: data mining, logistic regression, pattern discovery, time-series analysis.

Communicated by Kyong Joo Oh

Received April 5, 2010; Revised April 30, 2010

data mining. Data mining is a step of the knowledge discovery process, which involves particular algorithms for producing patterns (Chen et al. [5]; Imielinski and Mannila [16]). Data mining, which is also referred to as knowledge discovery in databases, has been recognized as the process of extracting non-trivial, implicit, previously unknown, and potentially useful information from data in databases (Agrawal et al. [1]; Han and Kamber [15]). It is the process of searching automatically large volumes of data and it is also a recent and contemporary topic in computing.

In general, large volumes of time-series data are contained in the financial database and these data have some useful but not easily found patterns in it. Many financial studies on time-series data analysis use linear regression model to estimate the variations and trend of the data. The Logistic Regression (LR) model is one of the most important statistical methods developed for data mining. It is also very popular and widely applied in the fields of health and medicine. The LR model can achieve high accuracy when the data show linear phenomena.

Therefore, the purpose of this study is to develop a LR model for knowledge discovery. The LR model uses a systematic approach in the financial database, and tries to find the patterns for estimating the trend. In addition, it calculates the returns to check the profitable or informative pattern.

The rest of the study is organized as follows: Section 2 introduces the related literature. The LR model is developed in Section 3. Section 4 presents the performance study. Section 5 discusses the results and points out future research directions.

2. Literature Review

Time-series analysis is an important topic in financial management. To achieve idealization and simplification, most traditional methods of time-series analysis have used special types of models or linear models to describe the analyzed data. Regarding time-domain models, Engle described the Autoregressive Integrated Moving Average Process (ARIMA) models (Fama and French [8]), and Bollerslev described the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model (Bollerslev [3]) for the extrapolation of past values into the immediate future. Past values were extrapolated according to the correlations among lagged observations and error terms.

The LR model is one of the most well-known and widely used models of classification. Bell et al. [2] have compared the classifying power of different statistical tools and the LR model. The LR model is appropriate for application when the dependent variables can be grouped into discrete states (Cramer [6]). The LR model identifies Boolean combinations of a given set of predictors that are associated with an outcome. It uses a sigmoid function that provides an output between 0 and 1, which can be easily interpreted as a probability of belonging to one group and is very appropriate for studies on bankruptcy. It is sufficient to assign 0 to those entities that are bankrupt and 1 to those that are solvents. More important is another difference, namely, how this model obtains the coefficients. The LR model uses the cumulative logistic probability as follows:

$$P = \frac{\exp^{(\beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k)}}{1 + \exp^{(\beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k)}}, \tag{1}$$

where β is the set of parameters to be estimated. In a linear regression model, the coefficient β measures the effect on the average value and for a unit change in the value of the repressor. However, the LR model deals with the probability of the event occurring. The parameters of the model are estimated using the maximum-likelihood method where the coefficients that make the observed results most ‘likely’ are selected.

The LR model has been another multivariate statistical model widely used in empirical research (Sinky [24]; Martin [20]). The study conducted by Martin [20] was the first to apply logistic regression to an early warning model for banks. The data sample consisted of US commercial banks that were members of the Federal Reserve System in 1970 and 1974. The variables used in his study can be classified into four categories: asset risk, liquidity, capital adequacy, and earnings. For the 1970 data, the liquidity variable was found to be the most significant. The common variable that contributed significantly to the model for both years was gross charge-off/net operating income. The aim was to find the characteristics of financial ratios of banks using various logistic regression models (Ohlson [21], Haggstrom [13], Lo [19]). To compare the effectiveness of logistic regression with multivariate discriminate analysis (MDA) in predicting financial distress of banks and/or financial institutions, Kwok et al. [18] conducted a study on US commercial banks in the late 1970s and early 1980s. Logistic regression technique was utilized to accomplish with ratios serving as financial distress indicators. Bank data from

Indonesia, South Korea and Thailand were used as separate case studies with time periods ranging from 1995 to 1997.

In the linear regression setting, the score function is the residual sum of squares; while in the classification setting, the scoring function is the misclassification rate. Here, this research uses the LR model in the context of binding site identification. Assume that there are a few interacting transcription factors for experiment of interest and these require binding to different sites on the transcription control regions.

3. Methodology

The LR model has been proposed and studied extensively by Ruczinski et al. (Sharma [23]; Ruczinski et al. [22]). It is one of the frequently used models in pattern recognition, especially in binary classification tasks. The performance of the method was tested in a financial modeling application. In the LR model, if $y = 1$ was set to represent one pattern type versus $y = 0$ for the other pattern type, then the logistic regression model is stipulated as equation (2):

$$P = \frac{\exp^{(\beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k)}}{1 + \exp^{(\beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k)}}, \tag{2}$$

where $\beta_0, \beta_1, ..., \beta_k$ are unknown constants analogous to the multiple linear regression model. The independent variables for the model would be *price*, *trading volume*, *turnover rate*, *size in circulation* and *book-to-market equity*. The dependent variable is rising pattern (1) and falling pattern (0). If $\beta_i = 0,$ then the presence of the corresponding factor has no effect. If $\beta_i < 0,$ then the presence of the factor reduces the odds probability of adoption, whereas if $\beta_i > 0,$ then the presence of the factor increases the odds probability of adoption. The computations for obtaining these maximum likelihood estimates require iterations performed by a computer program.

In the time-domain models, Engle [7] described ARIMA models (Autoregressive Integrated Moving Average Process) (Fama and French [8]), and Bollerslev [3] described GARCH model (Generalized Autoregressive Conditional Heteroscedasticity) (Bollerslev [3]) for the extrapolation of past values into the immediate future. Regarding selection of attributes, this study chooses the attributes

including *return*, *price*, *trading volume*, *turnover rate*, *size in circulation* and *book-to-market equity*. As for the attributes of book-to-market, Fama and French described the portfolio strategies and found out that size in circulation and book-to-market equity could affect stock returns. The book-to-market ratio effect is consistently strong in the bull market and in the bear market (Fama and French [8, 9]). Jegadeesh and Titman reported that *momentum* strategies could also affect the stock returns (Jegadeesh and Titman [17]). Brennan et al. described the *trading volume* factor and it is the cross-section of the expected stock returns (Brennan et al. [4]). Therefore, this research uses these core stock indices, including *price*, *trading volume*, *turnover rate*, *size in circulation* and *book-to-market equity* as dataset attributes.

4. Experiments and Discussion

The architecture is developed using Borland C++ Builder 6, Access database and SPSS 15. A series of experiments have been performed in order to verify the architecture. A mixed time-series dataset have also been designed to show the capability of LR model. In the experiments present the results of the LR model in the financial time-series database.

The database is segmented into the empirical stock indices, which is the Taiwan Stock Exchange Corporation (TSEC). These original datasets cover the daily closing prices from 1/1/2000 to 12/31/2003. The training datasets are from 1/1/2000 to 12/31/2001, while the test datasets are from 1/1/2002 to 12/31/2003.

Among various types of iPOD, PSP and GPS stocks are the most popular. For stock indices, each index can be limited to the types of iPOD, PSP and GPS stocks. There are iPOD, PSP and GPS companies in Taiwan. The GPS companies include Atech (亞元), Eten (倚天), MiTAC (神達), and Leadtek (麗臺); the PSP companies include MiTAC (神達), ASUS (華碩), I-SHENG (鎰勝), AVC (奇鋐), YUFO (育富), and Cyber TAN (建漢); and the iPOD companies include PowerTech (力成), JI-HAW (今皓), Abo (友旺), Foxlink (正崴), Mustang (同協電子), AVID (合邦), Porolific (旺玖), ENIght (英誌), TRIPOD (健鼎), ACON (連展), Transcend (創見), GENESYS (創惟), ASUS (華碩), Etron (鈺創), Milestones (銘異), Foxconn (鴻海) and APCB (競國). Some companies are included in two class levels, MiTAC (神達) is a multi-class company included in the classes of GPS and PSP. ASUS (華碩) is also a multi-class company included in the classes of PSP and iPOD.

The trading signals by the sliding-window calculate the return for days more than the given threshold. The threshold can be defined by the user or the investor. The investor can define which performance pattern liked by the user such that $f(x_i) = \sum_{i=1}^{n}(x_i)$, i is time interval and j is the dataset in each interval data. If $f(x_i) >$ threshold, then it uses these datasets to predict based on pattern discovery analysis. The logistic regression model is proposed and studied extensively by Ruczinski et al. (Sharma [23]; Ruczinski et al. [22]). It is one of the frequently used models in pattern recognition, especially in binary classification tasks. The LR model can achieve high accuracy when the linear variation of data is small. Using a traditional method, this study discusses the use of LR in a financial database.

In time series data, a particular problem has been plagued by multicollinearity. There are several other remedies that solve multicollinearity problems like Factor Analysis (FA) or Principal Component Analysis (PCA) and ridge regression. In this study, the principal component analysis has been employed to solve the multicollinearity (Sharma [23]; Hair et al. [14]; Gujarati [12]). The principal component analysis model is a popular tool for exploratory data analysis or, more precisely, for assessing the dimensionality of sets of items.

Principal component analysis model was used in each case to identify the underlying factors, which were then rotated to obtain the final solution. An oblique rotation was used because the underlying factors were expected to be correlated. The rotated matrix is principal component analysis. Correlation matrices for *price*, *trading volume*, *turnover rate*, *size in circulation* and *book-to-market equity* are presented in Table 1, and their loadings are listed in Table 2. A common rule of thumb for assessing construct validity is that individual items should have a highest loading. For attribute independence, the factor models are Rising_Prin 1, Rising_Prin 2, Falling_Prin 1 and Falling_Prin 2. Rising_Prin 1 and Rising_Prin 2 are the models of sliding window within 5 days and the accumulated returns more than 10%. Falling_Prin 1 and Falling_Prin 2 are the models of sliding window within 5 days and the accumulated returns more than –10%.

Rising_Prin 1 = 0.352 * price + 0.102 * trading volume + –0.110 * turnover rate + 0.343 * size in circulation + 0.376 * book-to-market equity

Rising_Prin 2 = –0.070 * price + 0.578 * trading volume + 0.577 * turnover rate + 0.116 * size in circulation + –0.079 * book-to-market equity

Falling_Prin 1 = 0.305 * price + 0.155 * trading volume + –0.093 * turnover rate + 0.335 * size in circulation + 0.357 * book-to-market equity

Falling_Prin 2 = –0.009 * price + 0.508 * trading volume + 0.678 * turnover rate + 0.001 * size in circulation + –0.080 * book-to-market equity.

Table 1. Correlation matrixes for *price*, *trading volume*, *turnover rate*, *size in circulation* and *book-to-market equity*

	Price	*TV*	*TR*	*SIC*	*BME*
Price	1				
TV	0.203(**)	1			
TR	–0.035(**)	0.421(**)	1		
SIC	0.601(**)	0.491(**)	–0.171	1	
BME	0.797(**)	0.298(**)	–0.183(**)	0.853(**)	1

Table 2. Loadings of *price*, *trading volume*, *turnover rate*, *size in circulation and book-to-market equity* items

	Loadings	
Item	Rising_Prin 1	Rising_Prin 2
Price	0.819	–0.167
Trading volume	0.422	0.802
Turnover rate	–0.127	0.846
Size in circulation	0.917	0.094
Book-to-market equity	0.949	–0.190

Note: Sliding window within 5 days and the accumulated returns more than 10%

	Loadings	
Item	Falling_Prin 1	Falling_Prin 2
Price	0.834	–0.052
Trading volume	0.453	0.711
Turnover rate	–0.125	0.931
Size in circulation	0.918	0.021
Book-to-market equity	0.975	–0.089

Note: Sliding window within 5 days and the accumulated returns more than –10%

In the LR model, if $y = 1$ was set to represent the rising pattern type versus $y = 0$ for the falling pattern type, then the LR model is stipulated as equation (3):

$$P = \frac{\exp^{(\beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k)}}{1 + \exp^{(\beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k)}}, \tag{3}$$

where $\beta_0, \beta_1, ..., \beta_k$ are unknown constants analogous to the multiple linear regression model. The independent variables for the model would be *price, trading volume, turnover rate, size in circulation* and *book-to-market equity*. The dependent variables are rising pattern (1) and falling pattern (0). From the estimated values of the coefficients, we see that the estimated probability of adoption for a *pattern type* with values of *price, trading volume, turnover rate, size in circulation* and *book-to-market equity* for the independent variables. The estimated probability can be calculated as follows. Table 3 shows the estimated number of pattern type for the independent variables.

$$P = \frac{\exp^{(-0.345 + -0.004 * Price + 0.0001 * TV + 0.268 * TR + -0.001 * SIC + 0.557 * BME)}}{1 + \exp^{(-0.345 + -0.004 * Price + 0.001 * TV + 0.268 * TR + -0.001 * SIC + 0.557 * BME)}}.$$

If

$$f(x) = -0.345 + -0.004 * Price + 0.0001 * TV + 0.268 * TR + -0.001 * SIC + 0.557 * BME,$$

then $P = \frac{\exp^{f(x)}}{1 + \exp^{f(x)}}$.

Table 3. Logistic regression analysis with *price, trading volume, turnover rate, size in circulation* and *book-to-market equity* items

Price	*TV*	*TR*	*SIC*	*BME*	$f(x)$	$\exp^{f(x)}$	*P*
105.55	10933	7.62	144	.19	1.34612	3.842488	0.79
111.56	8696	6.06	144	.20	0.8872	2.428321	0.71
108.18	9344	6.51	144	.20	1.0278	2.79491	0.74
108.18	11603	8.09	144	.21	1.4794	4.390311	0.81
115.69	12292	8.57	144	.22	1.59046	4.906005	0.83

To measure the accuracy of the LR model in the test dataset, the accuracy rate and error rate as the accuracy with pattern reappearing are used. In the rising pattern, the accuracy rate is 25% and the error rate is 75%. In the falling pattern, the accuracy rate is 85% and the error rate is 15%. The experimental results on accuracy rate and error rate are shown in Table 4.

Table 4. LR model results calculated with accuracy rate and error rate

		Predicted		
		Rising	Falling	Total
Observed	Rising	1989 (25%)	5931 (75%)	7920
	Falling	1106 (15%)	6117 (85%)	7223
Total		3095	12048	

5. Conclusion

This paper demonstrates that the technology for building classification algorithm from examples is fairly robust, and proposes an efficient LR model for classification. Meanwhile, it achieves high accuracy in the classification. There are evidences supporting the relevance of the LR model to informative pattern discovery. It also shows better support to what-if analysis for the investor.

In future research, there are several directions in which the present architecture can be utilized. The LR model can be applied to an efficient clustering algorithm in the financial database for finding the frequent item sets.

References

[1] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD Conference on Management of Data, 1993, pp. 207-216.

[2] T. B. Bell, G. S. Ribar and J. R. Verchio, Neural nets versus logistic regression: a comparison of each model's ability to predict commercial bank failures, Proceedings of the 1990 Deloitte and Touché University of Kansas Symposium on Auditing Problems, 1990, pp. 29-53.

[3] T. Bollerslev, Generalized autoregressive conditional heteroscedasticity, J. Econometr. 31 (1986), 307-327.

[4] M. J. Brennan, T. Chordia and A. Subrahmanyam, Alternative factor specifications, security characteristics, and the cross-section of expected stock returns, J. Finan. Econ. 49 (1998), 345-373.

[5] M. S. Chen, J. Han and P. S. Yu, Data mining: An overview from a database perspective, IEEE Trans. Knowledge and Data Engineering 8 (1996), 866-883.

[6] C. S. Cramer, Hybridization between diploid and tetraploid Pelargonium hortorum Bailey, Undergraduate Honors Thesis, Pa. State Univ., University Park, 1991.

[7] R. F. Engle, Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation, Econometric 50 (1982), 987-1008.

[8] E. F. Fama and K. French, The cross-section of expected stock returns, J. Finance 47 (1992), 427-465.

[9] E. F. Fama and K. French, Size and book-to-market factors in earning and return, J. Finance 50 (1995), 131-155.

[10] U. M. Fayyad, G. Piatetsky-Shapiro and C. J. Matheus, Knowledge Discovery in Databases: An Overview, AAAI/MIT Press, 1991.

[11] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, Menlo Park, CA, 1996, pp. 37-57.

[12] D. Gujarati, Essentials of Econometrics, 2nd ed., McGraw-Hill, 1999.

[13] G. W. Haggstrom, Logistic regression and discriminate analysis by ordinary least squares, J. Bus. Econom. Statist. 1(22) (1983), 9-238.

[14] J. F. Hair, R. E. Anderson, R. L. Tatham and W. C. Black, Multivariate Data Analysis, 5th ed., Prentice-Hall International, Inc., New Jersey, 1998.

[15] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[16] T. Imielinski and H. Mannila, A database perspective on knowledge discovery, Communications of ACM 39 (1996), 58-64.

[17] N. Jegadeesh and S. Titman, Return to buying winners and selling losers, J. Finance 48 (1993), 65-91.

[18] C. H. Kwok, A. B. Lim and Y. Wang, The Application of Discriminate Analysis (MDA) and Logistic Regression (Logic Analysis) in Predictions of Financial Distress of Banks/Financial Institutions – Issues and Comparisons, Centre for Research in Financial Services, Nanyang Technological University, Singapore, 1992.

[19] A. W. Lo, Logit versus discriminant analysis - A specification test and application to corporate bankruptcies, J. Econometr. 31 (1986), 151-178.

[20] D. Martin, Early warning of bank failure – a logic regression approach, J. Banking Finance 1 (1977), 249-276.

[21] J. M. Ohlson, Financial ratios and the probabilistic prediction of bankruptcy, J. Accounting Research 18(1) (1980), 109-131.

[22] I. Ruczinski, C. Kooperberg and M. L. LeBlanc, Logic regression, J. Comput. Graph. Statist. 12(3) (2003), 475-511.

[23] S. Sharma, Applied Multivariate Techniques, John Wiley, New York, 1996.

[24] J. J. Sinky, A multivariate statistical analysis of the characteristics of problem bank, J. Finance 23 (1975), 21-36.