Logistic Regression for Investigating Credit Card Default

Yang, Jeongwon[a] and Sung Ho Ha[b] and Min, JiHong[c]

[a] *Graduate Student, Dept. of Business Admin., Kyungpook National University*
1370, Sankyuk-dong, Book-gu, Daegu-si, Korea
Tel: +82- 53-950-5877, Fax: +82- 53-950-6247, E-mail:jeongwonyang@knu.ac.kr

[b] *Professor, Dept. of Business Admin., Kyungpook National University*
1370, Sankyuk-dong, Book-gu, Daegu-si, Korea
Tel: +82- 53-950-5877, Fax: +82- 53-950-6247, E-mail:hsh@knu.ac.kr

[c] *Srategic Planning Team, LG Electronics*
E-mail:jihmin@lge.com

Abstract

The increasing late-payment rate of credit card customers caused by a recent economic downturn are incurring not only reduced profit of department stores but also significant loss. Under this pressure, the objective of credit forecasting is extended from presumption of good or bad customers to contribution to revenue growth.

As a method of managing defaults of department store credit card, this study classifies credit delinquents into some clusters, analyzes repaying patterns of customers in each cluster, and develops credit forecasting system to manage delinquents of department store credit card using data of Korean D department store's delinquents. The model presented by this study uses Kohonen network, a kind of artificial neural network of data mining techniques to cluster credit delinquents into groups. Logistic regression model is also used to predict repayment rate of customers of each cluster per period. The accuracy of presented system for the whole clusters is 92.3%.

Keywords:

Kohonen Ntwork, Logistic Regression, Credit Forecasting System

Introduction

The Korean department stores, which had been in the center of distribution industry since 1960s, are struggling to escape from the unexpected market diminishment due to the growth of outlets, home shopping, and online shopping channels. What is worse, the successive decreased consumption sentiment due to economic depression since 2000s makes the possibility of insolvency of customers of department store's credit card much higher these days.

Related to this problem, we need to focus on researches about credit forecasting systems. Most previous researches divided customer's credits into two or three groups such as good/bad customers or good/bad/latent bad customers based on bad loan occurrences. And it is difficult to find out researches which considers customer groups who can recover from delinquency to normal credit status [2][6] [8][17][18].

Therefore, this study focuses on recoverable customer groups from credit delinquent state and proposes the credit forecasting system for delinquents of department store credit cards by classifying delinquents into clusters and analyzing credit recovery rate using Kohonen network, a kind of artificial neural network of data mining techniques for clustering, and logistic regression model to predict the rate of credit recovery.

The organization of this study is as follows; <Part 1> addresses the research backgrounds, objectives, methods, and structures. <Part 2> addresses the necessity and types of credit forecasting systems as well as previous relevant researches. <Part 3> develops the framework for development of delinquents' credit forecasting system. In <Part 4>, the developed framework is applied to real data to validate the proposed system. <Part 5> is for conclusion.

Theoretical Studies

Necessities of Credit Forecasting System

Financial institutions have tried to increase service quality such as providing loan services or developing new service products by providing correct credit forecasting information. Specially, credit forecasting system can be used for management of customer's credit level to reduce the possibility of bad loans in advance, provide the customized financial services, and prepare for the financial crisis beforehand. This credit forecasting system will provide the following positive influences on financial institutions. First, it provides the objective information to make the optimized loan decision. Second, it reduces the rate of late-payment and bad loan by managing customers according to their credit levels. Third, it makes possible to

distribute financial assets more efficiently by controlling credit limitations based on customer's current credit status. Finally, it reduces costs of managing credit loans, because it can reduce costs of investigating customer's credit successively in spite of initial high development costs.

Classification of Credit Forecasting System

Credit forecasting systems in consumer finances are classified into 'Credit scoring system', 'Behavior scoring system', 'Recovery scoring system' and 'Survival analysis'. Credit scoring system measures the customer's credits by scores and decides credit level based on scores automatically [15].
Behavior scoring system is introduced to make up the deficits of credit scoring system, because credit scoring system is for the new credit loan customers, so it can not track the changes of customer's credit status. Behavior scoring system overcomes this problem by analysing customer's transaction data after credit loans [2][19][12].
Recovery scoring system is a kind of behavior scoring system, which measures the collection rate of insolvent loans by referring to calculated scores. Survival analysis, which comes from logistic regression of credit scoring system, measures the profits at specified time by reflecting the risks of bad loans at that time.

Methods of Credit Forecasting System

Statistical Methods

The initial credit risk management systems used the statistical methods. James [15] proposed credit scoring system using multi-variate regression analysis method. Logistic regression analysis evolved into special types of regression analysis and made it possible to predict results or infer meanings statistically [4]. Recently, credit forecasting systems using survival analysis methods were proposed [11].

Data Mining Methods

Data mining methods are also used for credit forecasting system; decision tree, neural network, and genetic algorithm. Decision tree method is based on classification technique[3]. Decision tree is created according to analyzed data patterns. The reason of its broad usage in credit forecasting systems is that it makes easier to classify customers into good/bad customer according to their credit status [5][14].
Artificial neural network also shows outstanding results in credit forecasting process because it is appropriate for explaining non-linear models, for all that it does not need a statistical hypothesis [1][7][9][10].
Even through there are lots of credit forecasting systems using data mining techniques, it is not easy to tell which one is a winner [18]. Because data mining techniques cannot guarantee the global optimized solutions because of overfitting or inappropriate learning process. According to [13], other single methods have the same limitations. Therefore, it is thought to be more efficient to make the mixed credit forecasting models by acquiring the merits from other single methods to provide the most optimized solutions [16].

Mixed Methods

Recent studies nearly do not depend on a single method, but they provide credit forecasting systems of higher performance by comparing data mining to other statistical methods or presenting mixed models of previous two methods. Chen and Hung make up for the weak point of neural network, which has analysis problems of results, by using genetic algorithm [6].
The credit forecasting system presented in this study is a kind of recovery scoring system. It combines 'Kohonen network' and 'Logistic regression' to improve forecasting accuracy.

Framework of Development of System

The objective of this study is to develop the credit forecasting system using delinquents' credit information of D department store. As mentioned before, previous researches did not focus on customers who could recover from bad credit to normal credit state. Accordingly, this study proposes the mixed model that combines 'Kohonen network' and 'Logistic regression' to focus on the customer groups that are recoverable from credit delinquency state. 'Kohonen network' is for classifying the credit delinquents into groups with similar characteristics. The resulted clusters of credit delinquents from 'Kohonen network' will be inputs for 'Logistic regression model', which is responsible for analyzing the rates of credit recovery from delinquency.

Classification of Credit Forecasting System

Delinquents' data of this study come from D department store in 2003. The number of data is as follows; Delinquent customer information (50,496), Delinquency transaction information (1,367,506), and Purchasing transaction information (13,561,909). Considering the objective of this study, customers who have bad (unredeemed) debts are not considered from the time of data classification. The proportion of this kind of data is just 0.4% of total delinquents' data.
Collected data goes to next step for pre-processing. 41,831 delinquents' records out of 50,496 are thought to be valid for customer classification. Late-payment transaction data come from delinquency transaction table and 41,831 records are selected from that table, of which customer information is related to customers in collected delinquents. To pre-process the purchasing transaction data of year 2003, purchasing transaction records of arbitrarily selected 160,371 customers are extracted from

purchasing transaction table. Out of them, 21,464 customers have the late-payment records, therefore, 21,464 purchasing transaction records of delinquent customers and delinquent customer records will be used to analyze credit recovery types.

Classifying using Kohonen Network

As mentioned before, Kohonen network is used to classify the delinquent customers into clusters with similar characteristics. Kohonen network is a kind of neural network, which creates a map by extracting characteristics from input data through competitive learning processes. These processes of Kohonen network help to discover the unknown data patterns [Kohonen, 1990].
After selecting and pre-processing data, classification analysis using previously selected input variables will be executed. To improve the forecasting performance, it is important to classify the delinquent customers into several clusters and then, compare the repayment patterns within each clusters. After making clusters to some degree, input vectors are supposed to have more similarities, as nodes in competing layers get closer. This study makes clusters of delinquent customers using these characteristics of Kohonen network. SOM/Kohonen network of SAS Enterprise Miner 6.0 will be used as an analysis tool in this study.

Analyzing Credit Recovery Types using Logistic Regression

This study uses logistic regression to predict the rate of credit recovery and analyze the influencing factors on it. Logistic regression is a kind of statistical method widely used for credit forecasting. Logistic regression used to forecast the probability of event occurrences, in other words, bad loan occurrences, and influencing variables on them, but in this case, we'll create the credit forecasting system and then, forecast the rate of credit recovery using logistic regression.

Implementing Credit Forecasting System

In this step, credit forecasting system is created using clusters and logistic regression model generated in previous three steps. Through Kohonen network and logistic regression, models for each customer cluster of similar delinquency types are created, and then, these models are integrated to make our targeting delinquents' credit forecasting system.

Development of Credit Forecasting System

Making Clusters

New table is created to pre-process the delinquents information, of which invalid data were already excluded, and the number of valid delinquents for processing is 41,831. New candidate variables are created by operation. The resulting variables are presented in <Table 1> and variable names of *italic font* mean that those variables will be used in clustering. The variable of member ID will not be used in clustering. If the variances of variable values are too big, normalized values are inserted in Kohonen network.

Table 1 – Input Variables for Clustering

Column	Description	Column	Description
MEM_NO	Member ID	*AVG_DEL_AMOUNT*	Average amount of delinquency
DEL_TIME	Frequency of Delinquency	*AVG_BET_PERIOD*	Average interval of delinquency
DEL_PERIOD	Period of delinquency	*AVG_DEL_PERIOD*	Average period of delinquency
RETURN_TIME	Frequency of repayment	*AVG_RETURN_TIME*	Average Frequency of repayment
DEL_AMOUNT	Amount of delinquency		

Resulting output is presented in <Table 2> and 9 clusters are classified. Among these clusters, cluster 2,3,6 and cluster 4,5,9 look similar, because the values of repayment frequency, delinquency amount, and delinquency frequency show little differences, even though values of average delinquency period are little different. As a result, total 5 clusters are used for analysis, because cluster 2,3,6 and cluster 4,5,9 can be integrated into one cluster.

Table 2 – Result of SOM/Kohoen 3*3

Cluster	Delinquency Frequency	Avg. Delinquency Period	Avg. Delinquency Amount	Avg. Repayment Frequency	New Cluster
1	7.60	1.19	171,310.36	1.04	A
2	4.08	1.39	151,848.84	1.15	B
3	2.09	1.44	132,145.78	1.08	B
4	1.36	6.77	192,248.32	3.91	C
5	1.54	2.90	146,435.90	2.18	C
6	2.53	1.38	132,811.57	1.13	B
7	1.09	1.08	114,572.35	1.00	D
8	2.21	1.81	2,807,052.06	1.48	E
9	1.00	10.89	236,853.37	7.88	C

According to <Table 3>, which is the result of descriptive analysis, Cluster A seems to become a delinquent group because of the differences between the monthly closing date and repayment date. In the case of Cluster B, it does not have big amount of delinquency, but lots of frequencies of delinquency, therefore, customers in this cluster are thought to be habitual delinquents. Cluster C does not seem to have enough money because customers in this cluster have a long delinquent period, but relatively

frequent repayments with small repayment amount. Customers in Cluster D seem to become delinquents by accident because they repay their delinquent amount at once. Cluster E has relatively big delinquent amount and repays it by average 1.5 installments for two months. In spite of small numbers of records in this cluster, the average amount of delinquency is bigger than other groups by almost 20 times, therefore, it needs a special managerial attention. This study addresses the analysis result of Cluster B, because there is high likelihood that customers in this group become habitual delinquents.

Table 3 – Descriptive Statistics of Each Cluster

	Stats	Del. Frequency	Avg. Del. Period	Avg. Del. Amount	Avg. Del. Interval	Avg. Repay Frequency
A	Mean	7.60	1.19	171,310.36	1.50	1.04
	St.Dev.	1.67	0.23	134,142.36	0.34	0.20
	Max	12.00	2.00	1,311,837.70	2.20	1.80
	Min	6.00	1.00	8,680.00	1.00	0.40
B	Mean	3.06	1.39	139,763.09	3.45	1.13
	St.Dev.	1.05	0.56	138,516.95	2.02	0.45
	Max	6.00	6.00	1,329,524.00	11.00	4.00
	Min	2.00	1.00	2.00	1.00	0.30
C	Mean	1.43	4.77	167,947.96	0.48	3.25
	St.Dev.	0.68	3.12	167,045.29	0.82	2.04
	Max	4.00	12.00	2,146,900.00	6.00	11.00
	Min	1.00	1.50	2.00	0.00	0.50
D	Mean	1.09	1.08	114,464.04	0.09	1.00
	St.Dev.	0.29	0.26	144,009.25	0.29	0.09
	Max	2.00	3.00	1,295,695.00	1.00	1.50
	Min	1.00	1.00	1.00	0.00	0.50
E	Mean	2.24	1.83	2,753,827.22	1.23	1.48
	St.Dev.	1.68	1.87	2,990,921.16	1.83	1.25
	Max	8.00	11.00	20,284,134.50	9.00	10.00
	Min	1.00	1.00	1,300,156.00	0.00	0.50

Logistic Regression to Analyze Credit Recovery Types

Pre-Processing Data

Logistic regression for financial analysis focused on forecasting the rate of accident occurrences such as bad loans, but this study uses logistic regression model for forecasting the rates of credit recovery to analyze the rate of credit recovery.

First of all, variable sets are created using previously created variable sets of each cluster and then, testing variable sets for each cluster are duplicated according to the credit delinquency periods except first month and last month. This is because all customers are delinquents in first month and all customers are out of delinquency in last month. As a result, total 27 variable sets are created for each cluster. And then, 'STATUS' variable is created for each variable set, which represents whether recovering from delinquency or not. Credit forecasting function for each cluster is extracted from these variable sets for each period, therefore, 27 models are created.

Logistic regression of SPSS 12(Korean) will be used as a tool. STATUS variable is set to independent and other variables except 'delinquency period' to categorical variable with options of 'forward' variable input and significance level 0.005.

Creating Logistic Regression Model

As mentioned before, logistic regression model can predict the rates of accident occurrences. This study defines 'accident' as 'credit recovery'. The rate of accident occurrence(E(y)) is presented by equation of $(E(y)=e^z/(1+e^z)=1/(1+e^{-z}))$. Logistic combination, z in this equation ($z=\beta_0+\beta_1\chi_1+\ldots+\beta_k\chi_k$) is formulated by regression coefficient(β_i) and independent variable(x_i) and is used for forecasting the rate of event occurrences.

Logistic regression of SPSS is used to get regression coefficient(β) for cluster B. <Table 4> shows the intercept and variables of logistic regression for cluster B.

Table 4 –Variables of Logistic Regression for B

Selected Variables	(β)			
	2Mon	3Mon	4Mon	5Mon
PUR_YM_CN	0.257	0.366	0.881	2.288
PUR_YMD_CN	-0.009	.	.	.
DEL_TIME	.	0.497	13.964	.
AVG_BET_PERIOD	-0.236	-0.151	.	.
AVG_RETURN_TIME	-5.292	-3.625	-2.294	-1.380
AGE	-0.019	-0.030	.	
Selected Variables	**(β)**			
	2Mon	**3Mon**	**4Mon**	**5Mon**
GENDER(1)	-0.312	-0.766	.	-2.025
MEM_REGI_PERIOD	0.029	.	.	.
CARD_ISSUE_PERIOD	-0.025	.	.	.
JOB(1)	.	.	0.551	.
JOB(2)	.	.	-2.424	.
Intercept	9.080	8.882	-22.222	5.388

<Table 5> presents functions to forecast the rate of combination and event occurrence for 2 month using values of (β) and intercept calculated by logistic regression.

Table 5 –Function of Logistic Regression for B

	z(logistic combination)	Prob(event)
2Mon	9.08+0.257*PUR_YM_CN-0.009* PUR_YMD_CN-0.236*AVG_BET_PERIOD-5.292* AVG_RETURN_TIME-0.019*AGE-0.312* GENDER(1)+0.029*MEM_REGI_PERIOD-0.025 *CARD_ISSUE_PERIOD	$1/(1+\exp(-z_{2Mon}))$

Validating Logistic Regression Model

<Table 6> presents the comparing results of test and validation data set for cluster B. This table shows the very similar value change trends between test and validation data set. <Table 7> is the validation result per customer of logistic regression analysis. The average predicted rate of

credit recovery for cluster B is 93.4% and it means that about 94% of credit delinquent customers can be escaped from default status in 2 months. Since the real rate of credit recovery of this data is 100%, the predicted rate of logistic regression is not bad.

Table 6 – Comparison Between Test and Validation Sets

Data Sets	# of Customers	Delinquency Period	2 Mon	3 Mon	4 Mon	5 Mon
Test	6,803	1.3	0.959	0.999	1.000	1.000
Validation	2,288	1.4	0.961	0.999	1.000	1.000

Table 7 – Validation Result per Customers in B

Member ID	Del. Period	Month of Credit Recovery	Predicted Rate of Credit Recovery	Predicted Rate of Delinquency Upkeep			
				2Mon	3Mon	4Mon	5Mon
A00000039	1	2	99.7%	0.997	1.000	1.000	1.000
A00000211	1	2	99.0%	0.990	1.000	1.000	1.000
A00000586	1.7	2	45.2%	0.452	0.979	1.000	1.000
...	...	...	...	...	...	...	...
Average	1.4	2.2	93.4%	0.855	0.977	0.995	0.999

<Table 8> shows the validation results for each cluster. The average predicted rate of credit recovery is 92.3% for all clusters. The accuracy of this logistic regression is about 92%, which is not a bad score. But the predictive rate of cluster C is relatively lower than other clusters by 71%.

Table 8 – Validation Result for All Clusters

Cluster	Delinquency Period	Month of Credit Recovery	Predicted Rate of Credit Recovery
A	1.2	2.0	99.4%
B	1.4	2.2	93.4%
C	3.5	4.4	71.3%
D	1.1	2.0	94.8%
E	1.9	2.8	90.7%
Total	1.4	2.3	92.3%

Conclusion

Research Summary and Implications

This study presents the credit forecasting system to solve the common problems of credit delinquents in department stores. The credit forecasting system in this study is created by making clusters of credit card delinquents in D department store and analyzing the types of credit recovery for each cluster. The proposed model uses Kohonen network and logistic regression and selects the influencing variables on credit recovery to apply them into our credit forecasting system.

Credit delinquents are classified into 5 groups in this study according to delinquency types such as delinquency frequency, delinquency period, delinquency amount, delinquency interval, and repayment frequency. Using these clusters, the types of credit recovery are analyzed to find out the rate of credit recovery and the influencing factors on that rate. In next step, logistic regression model predicts the rate of credit recovery per period by calculating the values of intercept and coefficients. This credit forecasting system shows 92.3% accuracy.

The implications of this study are as follows; First, this study uses survival analysis to select the influencing variables on the rate of credit recovery and predict the period of credit recovery. Second, this study makes it possible to analyze the recoverable credit delinquent customers, who can recover from the credit delinquency state to normal credit state. On the other hand, previous researches divide customers into just two groups of good/bad credit customers. Even though this credit forecasting system cannot be a perfect solution to credit problems, it is expected to provide the strategic information that will help the management decision-making processes.

Limitation and Future Research

This study has some limitations to be improved in future researches as follows; First, we only use the data of credit delinquents of D department store for 12 months in 2003 due to the difficulty of collecting data. But this limited data area makes it difficult to generalize the research result, therefore, various types of data for different periods need to be collected and analyzed for generalization of this research. Second, variables used in this study are selected because they are thought to be helpful for management of credit delinquency by the help of managerial department in department store. But other new derived variables, which are not handled in this study, can be created, so it is better for future studies to be prepared for these new candidate variables.

Third, this study uses Kohonen network for clustering and logistic regression model for predicting the rate of credit recovery. Over 90% of forecasting rate shows that proposed model in this study has a good performance. But ceaseless development of new credit forecasting system will be also needed. Fourth, this study focuses on the clusters that are recoverable from credit delinquency to be normal credit state, but future study needs to include all customer from good to bad credit customers and enlarge the credit forecasting system to cover all kinds of customers.

References

[1] Altman, E. I., Marco, G. and Varetto. F. (1994), "Corporate Distress Diagnosis : Comparisons Using Linear Discriminant Anlysis and Neural Networks (the Italian Experience)," *Journal of Banking and Finance*, Vol.18, pp. 505-520.

[2] Baesens, B., Egmont-Petersen, M., Castelo, R. and Vanthienen, J. (2002), "Learning Bayesian Network Classifiers for Credit Scoring Using Markov Chain Monte Carlo Search," *Proceedings of the 16th*

International Conference on Pattern Recognition, pp. 49-52.

[3] Berry, M. J. A. and Linoff, G. (2004), *Data Mining Techniques: For Marketing, Sales, and Customer Support*, Wiley & Sons

[4] Bradley, P. S., Fayyad, U. M. and Mangasarian, O. L. (1998), "Data Mining: Overview and Optimization Opportunities," *INFORMS*, Special issue on Data Mining, pp. 17-22.

[5] Carter, C. and Catlett, J. (1987), "Assessing Credit Card Applications Using Machine Learning," *IEEE Expert*, Vol. 2, pp. 71-79.

[6] Chen, M. C. and Huang, S. H. (2003), "Credit Scoring and Rejected Instances Reassigning through Evolutionary Computation Techniques," *Expert System with Applications*, Vol. 24, pp. 433-441.

[7] Cheng, B. and Titterington, D. M. (1994), "Neural Networks: A Review from a Statistical Perspective, *Statistical Science*, Vol. 9, pp. 2-30.

[8] David, W. (2000), "Neural Network Credit Scoring Models," *Computers & Operations Reserch*, Vol. 27, pp. 1131-1152.

[9] Desai, C. S., Convay, D. G., Crook, J. N. and Overstreet, G. A. (1997), "Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms," *IMA Journal of Mathematics Applied in Business and Industry*, Vol.8, pp. 323-346.

[10] Desai, C. S., Crook, J. N. and Overstreet, G. A. (1996), "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment," *European Journal of Operational Research*, Vol.95, pp. 24-37.

[11] Han, J. and Kamber, M. (2004), *Data Mining: Concepts and Techniques*, 2nd Edition, Morgan Kaufmann Publishers, CA.

[12] Hand, D. J. and Henley, W. E. (1997), "Statistical Classification Methods in Consumer Credit Scoring: A Review," *Journal of the Royal Statistical Society*, Vol. 162, pp. 523-541.

[13] Hansen, J. V.(1999), "Combining Predictors: Comparison of Five Meta Machine Learning Methods," *Information Science*, Vol. 119, pp. 91-105.

[14] Imielinski, T. and Mannila, H.(1996), "A Database Perspective on Knowledge Discovery," *Communications of the ACM*, Vol. 40, pp. 214-225.

[15] James, H. M. and Edward, W. F.(1963), "The Development of Numerical Credit Evaluation Systems," *Journal of the American Statistical Association*, Vol. 58, pp. 799-806.

[16] Kim, E., Kim, W. and Lee, Y.(2000), "Purchase Propensity Prediction of EC Customer by Combining Multiple Classifiers Base on GA," *Proceedings of International Conference on Electronic Commerce*, pp. 274-280.

[17] Lee, T. S., Chiu, C. C., Lu, C. J. and Chen, I. F.(2002), "Credit Scoring Using the Hybrid Neural Discriminant Technique," *Expert Systems with Applications*, Vol.23, pp. 245-254.

[18] Thomas, L. C.(2000), "A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers," *International Journal of Forecasting*, Vol. 16, pp. 149-172.

[19] Thomas, L. C., Ho, J. and Soberer, W. T.(2001), "Time Will Tell: Behavioral Scoring and the Dynamics of Consumer Credit Assessment," *IMA Journal of Management Mathematics*, Vol.12, pp. 89-103.