2012 International Conference on Network and Computational Intelligence (ICNCI 2012)
IPCSIT vol. 46 (2012) © (2012) IACSIT Press, Singapore

An Introduction to an Approach of Combining and Selecting Independent Variables for Constructing a Predictive Model of Probability of Default

Te-Hsin Liang[1], Jian-Bang Lin [2 +], Yong-Goo Lee [3] and Chih-Hsiung Su [4]

[1] Department of Statistics and Information Sciences, Fu Jen Catholic University, Taiwan
[2] Ph.D student, Graduate Institute of Business Administration, Fu Jen Catholic University, Taiwan
[3] Department of Applied Statistics, Chung Ang University, Korea
[4] Department of Accounting Information, Chihlee Institute of technology, Taiwan

Abstract. In this decade, with advancements in information and data mining technology, many new approaches in extracting knowledge brought forth revolutionary developments in business world as well as other industries. According to New Basel Capital Accord, accurate Probability of Default (PD) prediction is becoming obviously a necessity. In the past, few researchers have considered the inter-correlation between two independent variables when constructing a predictive model of the PD. So, this research proposes an innovative approach of integrating two independent variables, combined-variables, which are inter-correlative and significant to a dependent variable. In order to compare the effectiveness of this prediction, this research also proposes an effectiveness index which is one kind of average Odds Ratio calculated by all values of AR (TNR, RR or PR) from the 9 cut-points and 20 samples. The result shows that all of the average Odds Ratio values are greater than 1 (1.72 to 2.94) in the testing data set. It shows the model used combined-variables can improve the predictive effectiveness of the model for different sampling structures of data.

Keywords: Combined-variables, Inter-correlation, effectiveness index, Probability of default, Basel II

1. Introduction

The Basel Committee proposed the Internal Rating-Based Approach (IRB) in Basel II in 2004. In order to pass muster with the important change of financial environment, the banks in American, Europe, Japan, Singapore, Hong Kong, Taiwan have already adopted IRB to estimate credit risk in 2009 [1]. As this research mentioned above, banks have to use valid models to estimate the average probability of default (PD) of accommodators to implement the IRB rules [2], [3], [4]. So, constructing an effective model to predict the PD has become an extremely important issue for banks.

Researchers have proposed many methodologies to improve the predictive models from the early 1950s. Logistic Regression (LR) model was concluded it is the most powerful [5], [6], [7], [8]. However, in the past, researchers rarely considered inter-correlations between two independent variables (IV), which might affect or counteract the effect of the prediction, when constructing the predictive models of PD. On the other hand, banks' risk management becomes more difficult because the database is gradually extensive.

The aim of this research is that we propose both an innovative approach of creating new significant combined-variables (CV) and an effectiveness index in order to construct an effective predictive model of PD in using data mining (DM). The IT applications in the banks harness the results of DM for becoming more intelligent than ever.

[+] Corresponding author. Tel.: + 886-933-424288; fax: +886-2-2905-2191
E-mail address: jianbang6428@gmail.com

The structure of this paper is as follows. Section II describes the approaches of this research. Section III introduces the results of the empirical data. Section IV concludes this paper.

2. Methodology

2.1. Logistic Regression Model

LR is similar to general linear regression, but its dependent variable (DV, the DV is a 'default' in this research) is binary or polytomous. In addition, the LR model will not only predict classification but also probability [9]. It is represented by the equation (1).

$$\text{logit}\left[\pi(\text{x})\right] = \log\left(\frac{\pi(\text{x})}{1-\pi(\text{x})}\right) = \beta_0 + \beta_1 \text{X}_1 + \ldots + \beta_\text{k} \text{X}_\text{k} \quad (1)$$

Here, $\pi(\text{x})$ is the probability of event, $0 \leqq \pi(\text{x}) \leqq 1$

2.2. Statistical Approach(SA)

For SA, IVs are selected if they are significant (p-value < .05) to the DV by either the Chi-square test or the t-test and no correlation each other. In the past, most researchers have used all significant IVs (by SA) individually when constructing predictive models.

2.3. Combining Variable Approach(CVA)

This research proposes an approach which considers the inter-correlation between two IVs when selecting variables. This approach is termed a combined variable approach (CVA), which uses significant CVs and IVs to construct the LR predictive model. The three major steps of CVA are described below.

- Step 1: Finding and choosing useful CVs.

  This step finds useful CVs by pairing significant IVs selected from SA and testing the interaction of each significant IV for the DV. Then, we choose the CV by using Rule 1.

  Rule 1: If $P_{ik} = \underset{for\ all\ j}{min}(P_{ij})$ and $P_{ik} < .05$, then CV_{ik} will be chosen.

  where P_{ij} is the p-value of the ANCOVA of the ith IV (IV_i) and the jth IV (IV_j), $i < j; j = 2,3,...13$; CV_{ik} is the combined-variable of IV_i and IV_k .
- Step 2: Classifying and testing the chosen CV

  If the original two IVs have a and b classifications, the CV will have an $a \times b$ classification. Then, we test whether or not each classification of CV has a different PD value by Homogeneity test.
- Step 3: Comparing the chosen CV and its IVs

  This step (Rule 2) keeps the input (CV or original two IVs) which is more significant to the DV.

  Rule 2: If $P'_{ik} < min(P'_i, P'_k)$, then IV_i and IV_k will be replaced by CV_{ik} from the LR predictive model. Otherwise, IV_i and IV_k will be kept in the LR predictive model.

  Here, P'_{ik} is the p-value of the Chi-square of the chosen CV_{ik} for DV. P'_i and P'_k are the p-values of the Chi-square of the IV_i and IV_k for DV, respectively.

2.4. Oversampling and Data Partition

The PD in this research is only 6.13%, and default is termed a rare event. Berry and Linoff [11] suggested that applying oversampling to raise the ratio of a rare event makes it easier to construct a better predictive model. This research adopts 3 oversampling proportions (1:1, 1:2 & 1:3) for default and non-default events. There are 60 samples composed from 20 samples from each oversampling proportion. On the other hand, in order to confirm the validity of the predictive model, each of the 60 samples is partitioned into two parts, training data set (80%) and testing data set (20%).

2.5. Confusion Matrix and effectiveness index

This research applies a confusion matrix to evaluate the predictive effectiveness of models (see Table 1.). The evaluating indexes are as following: Accuracy Rate (AR), True Negative Rate (TNR), Recall Rate (RR) and Precision Rate (PR), and the higher the indexes are, the better the model is (see equation (2)-(5)).

Table. 1 Confusion matrix

Prediction \ True	Default	Non-default
Default	A	B
Non-default	C	D

$$AR = \frac{(A+D)}{(A+B+C+D)} \times 100\% \quad (2)$$

$$TNR = \frac{D}{(B+D)} \times 100\% \quad (3)$$

$$RR = \frac{A}{(A+C)} \times 100\% \quad (4)$$

$$PR = \frac{A}{(A+B)} \times 100\% \quad (5)$$

We further proposed an effectiveness index $\overline{O}$ to compare the models' performance using CVA with SA, which is one kind of average Odds Ratio calculated by all values of AR (TNR, RR or PR) from the 9 cut-points (2, 4, 6, 8, 10, 20, 30, 40 and 50) and 20 samples. The $\overline{O}$ must be greater than or equal to zero. An $\overline{O}$ value greater than 1 indicates that the predictive effectiveness of CVA is better than that of SA, The index $\overline{O}$ is as equation (6)

$$\overline{O} = \frac{\sum_{i=1}^{m} O_i}{m} = \sum_{i=1}^{m} \frac{q_i}{1-q_i} \Big/ m \quad (6)$$

Where m; n_s: the number of cut-points and re-sampling samples, in this research m=9; n_s=20.

$O_i = q_i / 1 - q_i$: the Odds that CVA is better than SA for Index AR (TNR, RR or PR) at the i[th] cut-point.

$q_i = \sum_{j=1}^{n_s} y_{ij} \Big/ n_s$: the relative frequency that the value of the CVA is larger than the value of the SA.

$$y_{ij} = \begin{cases} 1 & if \quad I_{CVA}(i,j) > I_{SA}(i,j); i = 1,2,...m, j = 1,2,...n_s \\ 0 & otherwise \end{cases}.$$

$I_{CVA}(i,j)$: the value of the Index AR (TNR, RR or PR) for CVA at the i[th] cut-point and the j[th] sample.

$I_{SA}(i,j)$: the value of the Index AR (TNR, RR or PR) for SA at the i[th] cut-point and the j[th] sample.

3. Results

3.1. Data Structure

The data used in this research comes from one of Taiwan's local banks. Among the 10,997 cases, 674 are default cases and 10,323 are non-default cases. The PD is 6.13%. There are 17 IVs were used in this research.

3.2. Variable selection

Before constructing the LR predictive model, this research applied SA and CVA to select significant IVs. The results of SA and CVA are as follows.

- SA: Using the Chi-square test or t-test, 13 IVs are determined to be significant (having p-values less than .05) to the DV. These 13 variables are called the SA set.
- CVA: Following the step 1~3 introduced as Section 2.3, the 8 IVs found by SA were replaced by their CVs. Table 2. shows the SA set and the CVA set, which includes 5 single IVs and 4 CVs

Table 2 The SA set v.s. CVA set

Type		Variable			N
SA set	IV	Area (.000) Gender (.000) Education (.000) Age_Level (.000) Cross_Selling(.000)	Occupation (.000) Installment_Amt(.000) Auto_TFR_Payment (.000) Account_Open_Quarter(.000) Income (.001)	Early_Payment(.004) Marital_Status (.006) Guarantor(.009)	13
CVA set	IV	Installment_Amt (.000) Auto_TFR_Payment (.000)	Income (.001) Early_Payment (.004)	Guarantor (.009)	9
	CV	Education*Gender (.000) Area*Cross_Selling (.000)	Age_Level*Marital_Status (.000) Occupation*Account_OpenQuarter (.000)		

The value in ( ) is p-value of Chi-square test for DV.

3.3. Modelling

Using the variables in the SA and CVA sets, this research constructed 20 LR models for each of 3 oversampling proportions.

Table 3 shows the values of the average Odds Ratio (i.e. $\overline{O}$). In the training data sets, the accuracy rate for the CVA models (except for the 1:1 oversampling proportion model) is greater than 1, with values ranging from 1.62 to 1.68. The average Odds Ratio values for TNR, RR and PR for the three oversampling proportion CVA models are all greater than 1, with values ranging from 1.29 to 3.02. In the test data set, all of the average Odds Ratio values in the 4 indexes for the 3 oversampling proportions CVA models are greater than 1, with values ranging from 1.72 to 2.94. From this we see that the predictive effectiveness of the CVA model is better than that of the SA model. To sum up, when constructing an LR model, using CVA to select and combine IVs does indeed raise the effectiveness of the prediction.

Table 3 The average Odds Ratios ($\overline{O}$), graded by the 4 indexes, for the models constructed by SA and CVA among 3 oversampling proportions

Index	P-th	Training Data Set (80%)						Testing Data Set (20%)					
		1:1		*1:2*		*1:3*		*1:1*		*1:2*		*1:3*	
		q_i	O_i	q_i	O_i	q_i	O_i	q_i	O_i	q_i	O_i	q_i	O_i
AR	**2**	0.20	0.25	0.60	1.50	0.70	2.33	0.75	3.00	0.50	1.00	0.55	1.22
	4	0.30	0.43	0.75	3.00	0.75	3.00	0.75	3.00	0.65	1.86	0.55	1.22
	6	0.25	0.33	0.65	1.86	0.60	1.50	0.80	4.00	0.60	1.50	0.65	1.86
	8	0.45	0.82	0.65	1.86	0.50	1.00	0.75	3.00	0.60	1.50	0.70	2.33
	10	0.30	0.43	0.50	1.00	0.35	0.54	0.75	3.00	0.65	1.86	0.60	1.50
	20	0.35	0.54	0.65	1.86	0.75	3.00	0.70	2.33	0.80	4.00	0.70	2.33
	30	0.50	1.00	0.15	0.18	0.40	0.67	0.65	1.86	0.75	3.00	0.60	1.50
	40	0.45	0.82	0.60	1.50	0.55	1.22	0.60	1.50	0.65	1.86	0.65	1.86
	50	0.60	1.50	0.65	1.86	0.65	1.86	0.65	1.86	0.70	2.33	0.75	3.00
	$\overline{O}$		0.68		***1.62***		***1.68***		***2.62***		***2.10***		***1.87***
TNR	**2**	0.55	1.22	0.75	3.00	0.70	2.33	0.75	3.00	0.70	2.33	0.70	2.33
	4	0.60	1.50	0.75	3.00	0.85	5.67	0.90	9.00	0.75	3.00	0.75	3.00
	6	0.75	3.00	0.85	5.67	0.70	2.33	0.70	2.33	0.60	1.50	0.70	2.33
	8	0.70	2.33	0.70	2.33	0.50	1.00	0.70	2.33	0.65	1.86	0.70	2.33
	10	0.60	1.50	0.55	1.22	0.30	0.43	0.55	1.22	0.60	1.50	0.65	1.86
	20	0.55	1.22	0.55	1.22	0.65	1.86	0.55	1.22	0.60	1.50	0.65	1.86
	30	0.60	1.50	0.35	0.54	0.45	0.82	0.40	0.67	0.60	1.50	0.65	1.86
	40	0.80	4.00	0.65	1.86	0.50	1.00	0.35	0.54	0.55	1.22	0.60	1.50
	50	0.70	2.33	0.70	2.33	0.60	1.50	0.55	1.22	0.60	1.50	0.70	2.33
	$\overline{O}$		***2.07***		***2.35***		***1.88***		***2.39***		***1.77***		***2.16***
RR	**2**	0.50	1.00	0.80	4.00	0.75	3.00	0.70	2.33	0.70	2.33	0.70	2.33
	4	0.60	1.50	0.80	4.00	0.90	9.00	0.75	3.00	0.75	3.00	0.85	5.67
	6	0.65	1.86	0.85	5.67	0.60	1.50	0.85	5.67	0.70	2.33	0.70	2.33
	8	0.70	2.33	0.60	1.50	0.55	1.22	0.85	5.67	0.75	3.00	0.65	1.86
	10	0.45	0.82	0.60	1.50	0.30	0.43	0.65	1.86	0.60	1.50	0.65	1.86
	20	0.45	0.82	0.70	2.33	0.70	2.33	0.65	1.86	0.80	4.00	0.75	3.00
	30	0.50	1.00	0.25	0.33	0.50	1.00	0.65	1.86	0.70	2.33	0.50	1.00
	40	0.45	0.82	0.60	1.50	0.75	3.00	0.70	2.33	0.60	1.50	0.75	3.00
	50	0.60	1.50	0.65	1.86	0.85	5.67	0.65	1.86	0.60	1.50	0.75	3.00
	$\overline{O}$		***1.29***		***2.52***		***3.02***		***2.94***		***2.39***		***2.67***
PR	**2**	0.65	1.86	0.8	4.00	0.7	2.33	0.8	4.00	0.75	3.00	0.7	2.33
	4	0.55	1.22	0.75	3.00	0.85	5.67	0.8	4.00	0.75	3.00	0.75	3.00
	6	0.75	3.00	0.8	4.00	0.7	2.33	0.7	2.33	0.65	1.86	0.7	2.33
	8	0.7	2.33	0.7	2.33	0.55	1.22	0.65	1.86	0.65	1.86	0.75	3.00
	10	0.65	1.86	0.6	1.50	0.25	0.33	0.55	1.22	0.55	1.22	0.65	1.86
	20	0.5	1.00	0.65	1.86	0.6	1.50	0.45	0.82	0.6	1.50	0.65	1.86
	30	0.75	3.00	0.4	0.67	0.4	0.67	0.45	0.82	0.55	1.22	0.6	1.50
	40	0.8	4.00	0.6	1.50	0.4	0.67	0.35	0.54	0.5	1.00	0.55	1.22
	50	0.75	3.00	0.7	2.33	0.5	1.00	0.4	0.67	0.45	0.82	0.65	1.86
	$\overline{O}$		***2.36***		***2.35***		***1.75***		***1.81***		***1.72***		***2.11***

4. Conclusion

This research proposes a new procedure to compute the probability of default by selecting and combining IVs. Our results prove the effectiveness of this approach. When constructing an LR model, most previous research has relied on SA or stepwise LR models to select IVs. The inter-correlation between 2 IVs also can be considered in a stepwise LR model. Certain IVs which do not show statistical correlation to the PD will not be chosen for use in the LR model. Even IVs that are statistically correlated to the PD may not be chosen for use in a stepwise LR model. When IVs are chosen for the stepwise procedure, the inter-correlation between them would be considered. However, even when IVs are not chosen by a stepwise LR model, all the possible inter-correlation between IVs can be considered using CVA. This provides a wide selection of IVs to construct an LR model. CVA can also be applied to find a significant CV by combining two non-significant IVs or one significant IV and one non-significant IV.

The confusion matrix and ROC are useful tools when comparing the predictive effectiveness of two models. A receiver operating characteristic (ROC) is a graphical plot of the sensitivity vs. (1 - specificity) for a binary classifier system as its discrimination threshold varies. The ROC can be equivalently represented by plotting the fraction of true positives (i.e. RR) vs. the fraction of false positives (i.e. 1-TNR) [12]. When constructing a PD predictive model, the technique of re-sampling is usually used to ensure the stability of the model. There are several ROC curves for each model when re-sampling is used. When comparing two models, those ROC curves might be very similar or overlapping, making them hard to distinguish. In this research, we proposed an index of average Odds Ratios which clearly compares the predictive effectiveness of the two models. This index is a useful tool for comparing predictive effectiveness for re-sampling or for comparing several models.

5. References

[1] FECY (Financial Supervisory Commission, Executive Yuan). http://www.fscey.gov.tw/public/data/612291653771.doc, Accessed on 4 August 2005.

[2] BCBS. *The New Basel Capital Accord*. BIS Press, 2001.

[3] TABF. *Basel II-Chinese version.* Banking, FSC, Executive Yuan, Taiwan, 2004.

[4] R. R. Xue and H. C. Chen. The Research of Quantification of Probability of Default in Basel II. *Currency Observation and Credit Rating*. 2004, Jan.: 74-82.

[5] J. Begley, J. Ming, and S. Watts. Bankruptcy Classification Errors in the 1980s: An Empirical Analysis of Altman's and Ohlson's Models. *Review of Accounting Studies.* 1996, 1 (4): 267-284.

[6] Espahibodi. Identification of Problem Bank and Binary Choice Models. *Journal of Banking and Finance*. 1991, 15 (1): 53-71.

[7] F. E. Harrel and K. L. Lee. A Comparison of the Discrimination of Discriminant Analysis and Logistic Regression under Multivariate Normality. *Statistics in Biomedical, Public Health and Environmental Sciences*. Sen, P.K. ed., Amsterdam: Elsevier, 1985.

[8] A. W. Lo. Logit Versus Discriminant Analysis-A Specification Test and Application to Corporate Bankruptcies. *Journal of Econometrics*. 1986, 31 (2): 151-178.

[9] D. W. Hosmer and S. Lemeshow. *Applied logistic regression*. New York, John Wiley and Sons, 2000.

[10] E. P. Smith, I. Lipkovich, and K. Ye. Weight of Evidence (WOE): Quantitative Estimation of Probability of Impact. *Human and Ecological Risk Assessment.* 2002, 8 (7): 1585-1596.

[11] M. J. A. Berry and G. Linoff. M*astering data mining: the art and science of customer relationship management.* New York, Chichester: Wiley Computer Publishing, 2000.

[12] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado. The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*. 2005, 38 (5): 404-415.