Asian Social Science; Vol. 13, No. 2; 2017
ISSN 1911-2017 E-ISSN 1911-2025
Published by Canadian Center of Science and Education

The Application of Discriminant Model in Managing Credit Risk for Consumer Loans in Vietnamese Commercial Bank

Nguyen Thuy Duong[1], Do Thi Thu Ha[1] & Nguyen Bich Ngoc[1]

[1] Faculty of Banking, Banking Academy of Vietnam, Vietnam

Correspondence: Nguyen Thuy Duong, Faculty of Banking, Banking Academy of Vietnam, Vietnam. E-mail: duongnt@hvnh.edu.vn

Received: December 15, 2016 Accepted: December 30, 2016 Online Published: January 19, 2017

doi:10.5539/ass.v13n2p176 URL: http://dx.doi.org/10.5539/ass.v13n2p176

Abstract

The study focus on analysing financial status of consumer credit customers of Vietnamese commercial banks through two group discriminant function. Five independent variables were used in which some variables are related with the loan and the others are related with the demographic and socio-economic conditions of the borrower. Particularly, the variables related with the demographic and socio-economic conditions of the borrower are age; number of dependents; years at present job and salary while the independent variable related with the loan is loan amount. The result indicates that the estimated function is significant at 1 per cent level of significance and could forecast financial health with average 72.3 per cent accuracy. Therefore, in this study, the demographic, socio-economic and loan related variables can be used to determine the expected group membership of the borrowers in Vietnam.

Keywords: consumer credit, demographic and socio-economic characteristics, financial distress, prediction, two-group discriminant analysis

1. Introduction

The idea of consumer credit is extensive. In general, consumer credit is the term stands for the express loan facilities to the common people that have to repay with interest by equal monthly installment and the credit is not used for any commercial purpose. The need of consumer credit today is at it's highest, but at the same time the default rates have risen and from the banks' perspective the riskiness of these loans is usually higher than granted loans they analyzed defaulted. For the lending institution such a default rate affects to its financial performance significantly. So, it is substantially better to use discriminant analysis to determine the expected position or a score for the borrower to make the credit grant decision. In other words, a quantitative effort is made to forecast the expected position of the consumer credit applicant via the discriminant analysis. In the current paper, we use the discriminant analysis to develop predictive models allowing to distinguish between "good" and "bad" borrowers. The data have been collected from commercial Vietnamese banks over a 3-year period, from 2014 to 2016.

The discriminant analysis and the regression analysis have a similar feature about the number of dependent variables (one for both), the number of independent variables (multiple for both) and the nature of independent variables (metric for both). However, both of the analysises are different in terms of the nature of dependent variables. In the regression analysis, the dependent variable is a metric variable whereas in the discriminant analysis, the dependent variable is a categorical variable. In addition, the nature of the dependent variable in the binary logit model and the two-group discriminant analysis is the same. The linear discriminant analysis model involves linear combinations of the equation 1 form:

$$Z = \alpha + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots + \beta kXk$$

In the model, Z = discriminant score, α = constant, β's = discriminant coefficient or weight, X's = predictor or independent variable. The model creates different scores to separate the subjects of analysis into two group. This happens when the ratio- between-group sum of squares to within-group sum of squares is at maximum point. For any other combination, the ratio will be smaller. The figure 1 present by picture of the data collected on the two variables: X1 and X2 for two different groups of G1 and G2. The X1 axis represents X1 variable and the X2 axis represents X2 variable. By the discriminant analysis, a line is drawn to separate the two groups as the below

picture. If the model use more than two variables, drawing a scatter diagram as under is impossible so we have fixed two axes in a graph. In case of using a great number of variables, the discriminant analysis can separate Z scores into positive and negative groups. The lower part of figure 1 represents the group membership by using the estimated discriminant scores (Z) of the groups cases. The shaded section represents the misclassification of the group membership. The smaller the shaded section, the bigger the estimation accuracy is assumed (Uddin, 2013).

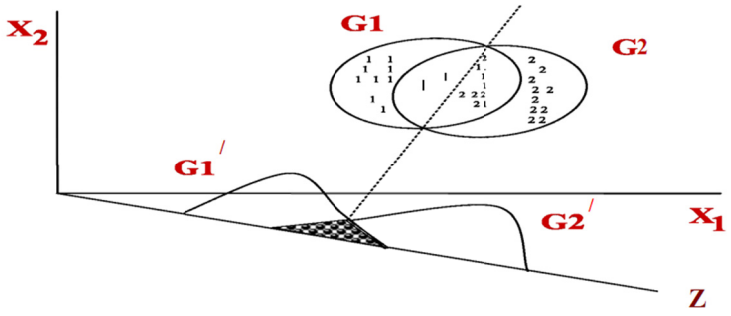

Figure 1. Discriminant Analysis

The broad objective of the study is to determine the consumer credit customers' insolvency by using demographic & socio-economic characteristics and two-group discriminant analysis. In order to reach the broad objective, the specific objectives are as follows: (i) To develop discriminant function or linear combinations of the predictor, or independent variables, which will best discrimate between the categories of the criterion or dependent variable. (ii) Using the values of the predictor variables to classify consumer loan into two different groups. (iii) To evaluate the accuracy of the classification. The first part of this research report is about introduction to the study which comprises prologue, objectives and methodology of the study. Literature review and the variables selection for the study are contained in the second part. Empractical study in Vietnam's commercial banks, findings and analysis are in the third part of the research.

2. Literature Review

2.1 Statistical Methods for Credit Risk Prediction

In the past, many researchers have developed a variety of traditional statistical methods for corporate credit risk prediction, with utilization of Linear discriminant analysis (LDA) and Logistic regression analysis (LRA) being the two most commonly used statistical methods in building corporate credit risk prediction models. Possibly the earliest use of applying LDA to corporate credit risk prediction is the work by Durand (1941). However, Karels and Prakash (1987) pointed that the application of LDA has often been challenged owing to its assumption of the categorical nature of the corporate credit data and the fact that the covariance matrices of the credit risk and non-risk classes are unlikely to be equal. In addition to the LDA approach, LRA is another commonly used alternative to conduct corporate credit risk prediction tasks. Thomas (2000) and West (2000) indicated that both LDA and LRA are intended for the case when the underlying relationship between variables are linear and hence are reported to be lacking in sufficient prediction accuracy. Besides above two statistical methods, Friedman (1991) reported that Multivariate adaptive regression splines (MARS) is another commonly corporate credit risk prediction method. However, the problem with applying these statistical methods to corporate credit risk prediction is that some assumptions, such the multivariate normality assumptions for independent variables, are frequently violated in reality, which makes these methods theoretically invalid for finite samples.

Although these methods are relatively simple and explainable, the ability to discriminate credit non-risk customers from credit risk ones is still an argumentative problem. In recent years, many studies have demonstrated that Artificial intelligence (AI) methods, such as Artificial neural network (ANN) (Herbert, 1992), Decision tree (DT) (Jiang, 2009), case based reasoning (CBR) (Shin & Han, 2001). and Support vector machine (SVM) (Bellotti and Crook, 2007) can be used as alternative methods for corporate credit risk predic-tion. In contrast with statistical methods, AI methods do not assume certain data distributions. These methods automatically extract knowledge from training samples. According to previous studies, AI methods are superior to statistical methods in dealing with corporate credit risk prediction problems, especially for non-linear pattern classification.

2.2 Discriminant Analysis for Consumer Credit

Wiginton (1980) conducted a discriminant analysis by using demographic and economic variables to caculate the consumer credit behavior. There are some demographic variables used such as number of dependents, living status, moved during last year, business use of vehicle and pleasure use of vehicle. The economic variables include employment class, occupation class and years in present employment. It is concluded that years in present employment, living status and occupation type have a significant impact in the credit risk caculating. Grablowsky (1975) carried on a two-group separately discriminant analysis in order to model risk in the consumer loan by using behavioral, financial, and demographic variables. For the behavioral data, two hundred borrowers was asked to answer a questionnaire of credit ratings scale. The loan application form of the two hundred borrowers were used to collect data for the financial and demographic variables. The researcher has started with thirty six variables and after a comprehensive sensitivity analysis, they only used thirteen valuable variables to conduct the analysis. Although the both set of data- analysis sample and holdout sample violated the equal variance-covariance assumptions, the estimated model classified the validation sample 94 per cent correctly. Awh & Waters (1974) conducted a study to determine the bank's active and inactive credit card users by using two group of variables-quantitative (economic and demographic) and attitudinal. The quantitative variables used are: (a) income, (b) age, (c) education, and (d) socio-economic standing. The socio-economic index is based on the respondents' particular position suggested by Reiss (1961). The data of the quantitative variables were collected through the application forms of borrowers, whereas the attitudinal data was collected through a questionnaire answered by the same borrowers. Attitudinal variables are: (a) use or non-use of other credit cards, (b) attitude toward credit, and (c) attitude toward bank charge-cards. The discriminant model can predict the group membership with 78 percent accuracy and significant at 0.01 level. Hand & Henley (1997) reviewed available credit scoring techniques in their article. In addition to the judgmental method, the available quantitative methods are logistic regression, mathematical programming, discriminant analysis, regression, recursive partitioning, expert systems, neural networks, smoothing nonparametric methods, and time varying models. They have concluded that every method have their limitation and the efficience of a good method depends on the structure and characteristics of the data.

Besides, Davis, Edelman & Gammerman (1992) carried on a study to compare various methods and concluded that there are few differences between the accuracy level of the methods, but the neural network algorithms take much longer time to train. According to Hand & Henley (1997), characteristics typical to differentiate the problematic and regular customer are: time at present address, home status, post code, telephone, applicant's annual income, credit card, types of bank account, age, country code judgment, types of occupation, purpose of loan, marital status, time with bank and time with employers, etc. Dinh & Kleimeier (2007) carried on a study for the Vietnam's retail banking market by using logistic regression analysis method. The variables used are age, education, occupation, total time in employment, time in current job, residential status, number of dependents, applicants annual income, total outstanding loan amount, other services used, cash in hand and at bank, etc. They have concluded that the quantitative credit scoring help to minimize the default rate from 3.3 per cent to 2.0 per cent. They also argued that it is useful to set up risk-based pricing in Vietnamese commercial banks. The found some factors such as time with bank, followed by gender, number of loans, and loan duration are most valuable. Based on the above literature review, experience of the researchers and availability of the data, five demographic and socio-economic variables are selected for this study. The data is collected on the variables from the answers of the credit officers by filling up the pre-designed questionnaire.

3. Research Methodology

To be considered as one of the most broadly techniques used to discriminate between two groups (Abdou & Pointon, 2011), discriminant analysis has long been used by researchers and bank's managers for building credit scoring models to distinguish between customers as good credit and bad credit (Caouette et al, 1998; Hand & Henley, 1997 and Desai et al, 1996). Therefore, in this article, discriminant model will also be used to distinguish between two loan borrower classification groups: repayment and non-repayment, in which good borrower is coded as 1 and bad borrower is coded as 0. This use of two groups of customers which are either good or bad ones is also considered as one approach for classification purposes in credit scoring models by many researchers such as Kim & Sohn (2004); Lee et al (2002) and Banasik et al (2001). These two possible states are defined by a number of factors which simultaneously influence on borrower's ability to pay and willingness to pay. In case of this study, information related to age, salary, years at present career, loan amount and number of independents will be used to calculate discriminant score Z for a given customer as follows:

$$Zi = \beta 0 + \beta 1*X1 + \beta 2*X2 + \beta 3*X3 + \beta 4*X4 + \beta 5*X5 + \varepsilon.$$

Where:

Z is the discriminant score that maximizes the distinction between the two groups:

$\beta_{0:}$ constant.

$\beta_{1\text{-}5}$: slopes of independent variables.

X1: Age

X2: Dependents

X3: YAPJ

X4: Salary

X5: Loan amount

ε: random error.

As can be seen from the model, there are two types of variables used in this model, which are dependent and independent variables. The only dependent variable is status of borrower that is a categorical variable. If a borrower's position is default then he is denoted by 0 and if the borrower's position is regular, then he is denoted by 1. By contrast, there are two types of the predictor variables are used in this study. Particularly, some variables are related with the loan and the others are related with the demographic and socio-economic conditions of the borrower. The variables related with the demographic and socio-economic conditions of the borrower are as follows. Age: How old borrower is; Dependents: the number of persons who are dependent on the borrower; YAPJ stands for years at present job; Salary: money earned by the borrower per month. The independent variable related with the loan is loan amount which indicates how much money borrowed by the borrower.

Data used in this study was collected from credit officers in commercial banks in Vietnam. Particularly, a list of information which authors need to know about borrowers was designed and after that was send to directors of Vietnamese commercial banks to ask for help. With the great support of credit officers, authors got data of over 550 consumer credit customers. However, after looking at database we found that for some customers vital information was missed; therefore, only information of 500 borrowers was used. It is clear that by using this source of data which already been available in commercial banks we can save time and money (Ghauri & Grɳhaug, 2006). Moreover, Stewart and Kamins (1993) indicate when comparing between secondary data and own collected data, the quality of former is higher than latter. Finally, secondary data has also been used in many researches on credit scoring conducted by researchers not only in Viet Nam (Duong, Tran & Ho, 2015) but also in other countries like Wiginton (1980); Bartolozzi, Garcıa-Erguın, Deocon, Vasquez & Plaza (2008) and Hörkkö (2010). As a result of that, secondary data collected from commercial banks in Vietnam was used.

Besides, related to sample size, it is said that the larger the sample size, the better the scoring model's accuracy. However, it is also worth noting that "a sample size of at least twenty observations in the smallest group is usually adequate to ensure robustness of any inferential tests that may be made" (Hintze, 1998). Therefore, in case of this model in which the number of independent variables is five, there should be at least 100 cases in smallest group to produce right discriminant function.

According to the World Bank, the proportion of non-performing loans to total gross loans in Vietnam is about 2.94% or in other words the number of non-default borrowers is relatively higher than their counterparts, leading to the number of good and bad borrowers taken from banks in this study is not the same. Therefore, like the way other researchers such as Lee et al (2002) and Desai et al (1996) did, this study also choose the proportion of good borrowers to bad ones used was seven to three. Particularly, in case data of 500 customers will be used in this study, the number of good borrowers will be 350 while their counterpart ones was 150. Moreover, information on 500 customers then will randomly be divided into two different groups named analysis sample and hold out sample. The former including 400 customers will be used to estimate discriminant function while the later including 100 customers will be used to check the validity of the model.

As data used in this study is numerical data, of which value can be measured numerically (Saunders et al., 2007), quantitative approach was applied. Particularly, quantitative approach was used to measure differences in means of independent variables between two groups. Moreover, quantitative analysis was also used to look for connections and spot relationships between independent variables.

Before running discriminant analysis, it is important to describe characteristics of all variables used in this study and check assumptions to make sure that study's findings are accurate. In this study, data was processed by SPSS

21.

Firstly, as data in this study is continuous variables, descriptive was used to explore basic statistics such as mean, maximum, minimum, standard deviation of predictors in each group. Besides, independent sample T test SPSS was also used in this study to compare mean score on predictors between non defaulted and already defaulted group (Pallant, 2013).

Secondly, it is required that data used in discriminant analysis must be independent and normally distributed (Khemakhem and Boujelbene, 2015); therefore, like other researches this study also accesses normality of data's distribution by the Kolomogorov- Smirnov test on SPSS.

Thirdly, not only normal distribution, but outliers and multicollinearity were also tested to make sure results of further tests are accurate (Field, 2009; Pallant, 2013). It is clear that the presence of an outlier, which is defined as cases of which values are quite higher or lower than majority of other cases' ones (Pallant, 2013), might make researchers miss important information and receive confusing results; therefore, it is essential to recognise outlier (Dielman, 2001). Tails of distribution presented in graph named histogram was used to find out there is potential outliers in this study or not. There are some observations are out at the outlier labelling rule, which after that will be eliminated. Besides, the existence of multicollinearity or explanatory variables are correlated might lead to estimates of parameter values are not reliable, and it is difficult for researchers to access the contributions of each independent variable to overall R^2 (Gujarati, 1999). Therefore, this study used results obtained from correlation matrix, which presents not only correlation between dependent variable and predictors, but also between independent variables to test for multicollinearity. Particularly, Pearson producted moment correlation coefficient will be used. The highest absolute value of correlation coefficient between each of independent variable should be less than 0.7 to ensure that multicollinearity does not happen in this study.

After checking and correcting problems related to data, the next step is to apply discriminant analysis to the analysis sample. However, it is worth noting that there are two common methods for discriminant analyses, which are direct method and stepwise discriminant analysis. In this study, which is based on the previous research and theoretical model, the direct method will be used.

4. Results

Table 1. Group statistic

	Ability to pay loan	N	Mean	Std. Deviation	Std. Error Mean
Age	Not good	120	30.719	3.9987	.3161
	Good	280	36.772	5.5364	.2922
Salary	Not good	120	13.0419	4.19951	.33200
	Good	280	14.0351	4.92672	.26410
YAPJ	Not good	120	5.38	2.454	.194
	Good	280	10.26	3.679	.194
Dependents	Not good	120	2.03	.812	.064
	Good	280	1.53	.854	.045
Loan amount	Not good	120	398677156.250	165445876.1431	13079644.9524
	Good	280	469608695.652	261697678.4845	14089329.3907

As can be seen from the table named group statistics, group means and standard deviations are calculated for each variable of the default and the non-default groups, which after that contributes to see whether the variables can differentiate between default customers and regular customers. It is true that, except for salary clear differences are witnessed in group means for the groups for the variables age, years at present job, number of dependents and loan amount. Particularly, average age for creditworthy borrowers, which is about 36 years old, is relatively higher than average age for the bad ones which is only a little above 30 years old. This result supports for conclusion of Peter and Peter (2006) who said that the probability of default is higher with a younger borrower. The same pattern is also witnessed in term of number of dependents. This might be explained by the fact that the more people borrowers have to support financially, the less money they have to pay loan or borrowers are likely not to pay loan in time. Moreover, there is big difference in years at present job between borrowers who are considered as credit worthy and not. Table 1 shows that average value of years at present job of no defaulted borrowers is nearly twice already defaulted borrowers' ones. By contrast, the dissimilarity in

monthly salary between good and bad borrowers is slight, which income among the defaulters is only one million VND lesser than the non-defaulters. More importantly, this difference might contributes to explain why loan amount of non-defaulters is relatively higher than defaulters.

Table 2. Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	Df	Sig.
Age	.095	400	.000	.968	396	.000
Salary	.097	400	.000	.948	390	.000
YAPJ	.079	400	.000	.962	392	.000
Dependents	.317	400	.000	.833	397	.000
Loan amount	.088	400	.000	.910	385	.000

As mentioned above, data used in discriminant analysis should be normally distributed (Khemakhem & Boujelbene, 2015); therefore, K-S test was used to find out whether distribution of data used in study is normal or not.

The test statistic for the K-S test is presented in table 2 showing that the percentage of age D(396) = 0.095, p= .000, which was smaller than 0.05; therefore, the distribution is not normal (Pallant, 2013). The same pattern also was witnessed in salary, years at present job, number of dependents and loan amount. To correct this problem, according to Field (2009), transforming data is one of popular options. Therefore, in this study, all variables were transformed into log transformation, which is as the same as method used by Hörkkö (2010). More importantly, Uddin (2013) proved that discriminant analysis still get good result in case data used is not normally distributed. As a result of that, this problem in this study is not serious.

Besides, by looking at the tails of distribution presented in graph named histogram, this study found that there are potential outliers because there are some observations are out at the outlier labelling rule. However, when considering information in descriptive table, the difference between 5% trimmed mean (4.719) and mean (4.7161) values is extremely small; therefore, outlier problem in this study is not serious and might be solved by eliminating outliers.

According to Pallant (2013), multicollinearity happens when absolute value of correlation coefficient between each of independent variables is 0.7 or more. The correlations between variables used in this study (table 3) showed the first largest bivariate correlation was listed for relationship between age and years at present job. Unfortunately, this pair-wise correlation was only 0.770, which was clearly higher than 0.7; therefore, multicollinearity does happen and age will be omitted from regression.

Table 3. Correlations

	Ability to pay loan	Age	Salary	YAPJ	Dependents	Loan amount
Ability to pay loan	Pearson Correlation	1	.480**	.098*	.560**	-.263**	.139**
Age	Pearson Correlation	.480**	1	.106*	.770**	-.033	.063
Salary	Pearson Correlation	.098*	.106*	1	.131**	.258**	.611**
YAPJ	Pearson Correlation	.560**	.770**	.131**	1	-.193**	.016
Dependents	Pearson Correlation	-.263**	-.033	.258**	-.193**	1	.223**
Loan amount	Pearson Correlation	.139**	.063	.611**	.016	.223**	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Moreover, as the sig. (2-tailed) value for predictors are below the required cut-off of 0.05 (Table 4); there is statistically significant difference in salary, YAPJ, number of dependents and loan amount between the defaulters and non defaulters.

Wilks' lambdas and the F rations are estimated to test the equality of the group means. The value of the Wilks' lambda (λ) varies between 0 and 1. While the large value of λ indicates that group means are not different, small value of λ indicates that the group means are different or in other words the smaller the Wilks's lambda, the more important the independent variable to the discriminant function. Wilks's lambda is significant by the F test for all

independent variables. The lower significant ratio for the corresponding F ratio means - the variable is very significant in the case of determining group membership. Therefore, based on results presented in table 5, it is obvious that dependents and years at present job may best discriminate between the two groups of borrowers.

Table 4. Independent Samples test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	T	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Salary	Equal variances assumed	1.058	.304	-2.207	506	.028	-.993	.449
	Equal variances not assumed			-2.341	358.176	.020	-.993	.424
YAPJ	Equal variances assumed	13.007	.000	-15.342	516	.000	-4.888	.319
	Equal variances not assumed			-17.793	440.828	.000	-4.888	.275
Dependents	Equal variances assumed	7.649	.006	6.220	521	.000	.497	.080
	Equal variances not assumed			6.344	318.735	.000	.497	.078
Loan amount	Equal variances assumed	16.188	.000	-3.148	503	.002	-70931539.402	22531078.827
	Equal variances not assumed			-3.690	457.412	.000	-70931539.402	19224627.818

Table 5. Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Logloanamount	.995	2.543	1	385	.111
Logdependents	.884	64.010	1	385	.000
Logsalary	.997	1.612	1	385	.205
LogYAPJ	.595	331.959	1	385	.000

Table 6. Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.980[a]	100.0	100.0	.704

First 1 canonical discriminant functions were used in the analysis

There are two groups of borrowers which are classified as creditworthy or not; therefore, number of function is 1. The eigenvalue is 0.98 and the canonical correlation, which is 0.695, is the measure of association between the discriminant function and the dependent variable. Besides, the square of canonical correlation coefficient (0.704^2 = 0.4956 = 49.56%) indicates the percentage of variance in the dependent variable is explained by the estimated discriminant function.

Table 7. Wilk's Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1	.505	284.835	4	.000

The Wilks' Lambda = 0.505 (which is equivalent to Chi-square = 284.835 with 4 d.f.) is significant at the 0.000 level. This means that the discriminant function computed in this procedure is statistically significant at the 0.000 level. Only then, we can proceed to interpret the results.

Table 8. Structure matrix

	Function
	1
LogYAPJ	.856
Logdependents	.376
Logloanamount	.075

Pooled within-groups correlations between discriminating variables and standardized canonicalll discriminant functions. Variables ordered by absolute size of correlation within function. Thee structure matrix table shows the correlations of each variable with each discriminant function More importantly, in table 8, predictors are ordered from highest to lowest by the absolute size of the loading. Therefore, years at present job in this study is considered as most important variable in determining the group membership followed by number of dependents and loan amount. By contrast, monthly salary or how much money borrower earns per month is the least important variable.

Table 9. Canonical discriminant function coefficients

	Function
	1
Logloanamount	1.106
Logdependents	3.028
Logsalary	-2.091
Log YAPJ	5.392
(Constant)	-12.999

Unstandardized Coefficients

This table contains the unstandardized discriminant function coefficients. These would be used like unstandardized b (regression) coefficients in multiple regression that is, they are used to construct the actual prediction equation which can be used to classify new cases. Therefore, the model should be like this:

$$Zi = -12.999 + 1.106*loanamount + 3.028*dependents – 2.091*salary + 5.392*YAPJ$$

Table 10. Functions at Group Centroids

Ability to pay loan	Function
	1
Not good	-1.380
Good	.671

Unstandardized canonical discriminant functions evaluated at group means

The group centroids are the averages of the Z values calculated by the estimated model, which can use to evaluate the expected position of the consumer credit customers (Uddin, 2013). As can be seen in table 10, the centroid of not good borrower is -1.380 and the centroid of the regular group is 0.671. Therefore, if the estimated Z value of a customer is negative, then the expected status of this customer is default because the centroid value is negative for default group and if the estimated value of a case is positive then the expected position of the case is good borrower as the centroid value is positive for the regular group.

The classification matrix of the original sample (Table 11) shows that 81.5 percent of the case are predicted by the model correctly. Since at the time of estimating classification matrix of the original cases, the sample for which the prediction is made included in the sample, the classification matrix may be biased. So, cross-validated classification matrix is made based on the activity that the case for which the prediction is being made will be kept out of the analysis and the model is estimated.

The holdout sample is also used to check the validity of the model. After putting the values of the holdout sample on the estimated discriminant function, the Z values are computed for the cases. By using the Z values and centroids, group membership is predicted. The table 12 shows that 72.3 percent of cases are correctly classified.

Table 11: Classification Results[a,c]

Ability to pay			Predicted Group Membership		Total
			0	1	
Original	Count	Not good	92	21	113
		Good	51	225	276
		Ungrouped cases	0	11	11
	%	Not good	81.3	18.8	100.0
		Good	18.5	81.5	100.0
		Ungrouped cases	.0	100.0	100.0
Cross-validated[b]	Count	Not good	92	21	113
		Good	51	225	276
	%	Not good	81.3	18.8	100.0
		Good	18.5	81.5	100.0

a. 81.5% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 81.5% of cross-validated grouped cases correctly classified.

Table 12. Classification Results[a,]

Ability to pay loan			Predicted Group Membership		Total
			0	1	
Original	Count	Not good	21	7	28
		Good	20	47	67
		Ungrouped cases	0	5	5
	%	Not good	76.7	23.3	100.0
		Good	29.6	70.4	100.0
		Ungrouped cases	.0	100.0	100.0

a 72.3% of unselected original grouped cases correctly classified.

5. Conclusion

This study estimates a two-group discriminant analysis in order to determine the expected status of the consumer credit customers of a bank in Vietnam. The estimated function is significant at 1 per cent level of significance and could forecast financial health with average 72.3 per cent accuracy. Thus, the study proposed that the demographic, socio-economic and loan related variables can be used to determine the expected group membership of the borrowers in Vietnam. Discriminant function estimated for an institution or bank cannot be used for other bank or institution because the discriminant function coefficients will vary based on a bank/institution's data set. Hence banks/institutions should use own data base to estimate its own discriminant function to use. By using the estimated function, the consumer credit disbursement decision can be faster, more accurate and cost saving. Moreover, risk based pricing can be adapted in the credit management.

References

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2-3), 59-88. http://dx.doi.org/10.1002/isaf.325

Awh, R. Y., & Waters, D. (1974). A discriminant analysis of economic, demographic, and attitudinal characteristics of bank charge – card holders: A case study. *The Journal of Finance*, *29*(3), 973-980. http://dx.doi.org/10.1111/j.1540-6261.1974.tb01495.x

Banasik, J., Crook, J., Thomas, L. (2001). Scoring by Usage. *Journal of the Operational Research Society, 52*(9), 997-1006. DOI: 10.1057/palgrave.jors.2601178

Bartolozzi, E., Garcıa-Erguın, L., Deocon, C., Vasquez, O., & Plaza, F. (2008). Credit Scoring Modelling for Retail Banking Sector. *Oscar Ivan Vasquez & Fransico Javier Plaza, II Modeling Week, Universidad Complutense de Madrid, 16th–24th June*.

Bellotti, T., & Crook, J. N. (2007). Support vector machines for credit scoring and discovery of significant features. *Journal of the Operational Research Society 56:1082-1088.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.5526&rep=rep1&type=pdf

Caouette, J. B., Altman, E. I., & Narayanan, P. (1998). *Managing Credit Risk: The Next Great Financial Challenge.* New York: John Wiley & Sons Inc

Davis, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, *4*(1), 43-51. http://dx.doi.org/10.1093/imaman/4.1.43

Desai, V. S., Crook, J. N., Overstreet, G. A. (1996). A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment. *European Journal of Operational Research, 95*(1), 24-37. http://dx.doi.org/10.1016/0377-2217(95)00246-4

Dielman, T. E. (2001). *Applied regression analysis for business and economics* (3rd ed.). Pacific Grove, CA: Duxbury/Thomson Learning.

Dinh, T. H. T., & Kleimeier, S. (2007). A credit scoring model for Vietnam's retail banking market. *International Review of Financial Analysis*, *16*(5), 471-495. http://dx.doi.org/10.1016/j.irfa.2007.06.001

Duong, T., Tran, V., & Ho, Q. (2015, January). A Proposed Credit Scoring Model for Loan Default Arobability: a Vietnamese bank case. In *International Conference on Qualitative and Quantitative Economics Research (QQE). Proceedings* (p. 52). Global Science and Technology Forum.

Durand, D. (1941). Risk elements in consumer instalment financing. *National Bureau of Economic Research Books*.

Field, A. P. (2009). Discovering statistics using SPSS: And sex and drugs and rock 'n' roll (3rd ed.). Los Angeles, [Calif.]; London: SAGE.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, *19*(1), 1-141. http://dx.doi.org/10.1214/aos/1176347963

Ghauri, P. N., & Grŋhaug, K. (2006). *Research methods in business studies: A practical guide* (3rd ed.). Harlow: Financial Times Prentice Hall

Grablowsky, B. J. (1975). A Behavioral Model of Risk in Consumer Credit. *The Journal of Finance*, *30*(3), 915-916. http://dx.doi.org/10.1111/j.1540-6261.1975.tb01871.x

Gujarati, D. (1999). *Essentials of econometrics* (2nd ed.). London: McGraw-Hill

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523-541. http://dx.doi.org/10.1111/j.1467-985X.1997.00078.x

Herbert, L. J. (1992) "Using Neural Networks for Credit Scoring", *Managerial Finance, 18*(6),15 – 26. http://dx.doi.org/10.1108/eb013696

Hintze, J. (1998). NCSS statistical software. *NCSS, Kaysville, UT*.

Hörkkö, M. (2010). *The determinants of default in consumer credit market*. (Master's thesis, Aalto University School of Economics). Retrived from http://epub.lib.aalto.fi/en/ethesis/pdf/12299/hse_ethesis_12299.pdf

Jiang, Y. (2009, March). Credit scoring model based on the decision tree and the simulated annealing algorithm. In *Computer Science and Information Engineering, 2009 WRI World Congress on* (Vol. 4, pp. 18-22). IEEE. DOI: 10.1109/CSIE.2009.481

Karels, G. V., & Prakash, A. J. (1987). Multivariate normality and forecasting of business bankruptcy. *Journal of Business Finance & Accounting*, *14*(4), 573-593. http://dx.doi.org/10.1111/j.1468-5957.1987.tb00113.x

Khemakhem, S., & Boujelbene, Y. (2015). Credit risk prediction: A comparative study between discriminant analysis and the neural network approach.*Accounting and Management Information Systems*, *14*(1), 60-78. Retrieved from ftp://ftp.repec.org/opt/ReDIF/RePEc/ami/articles/14_1_3.pdf

Kim, Y. S., & Sohn, S. Y. (2004). Managing Loan Customers Using Misclassification Patterns of Credit Scoring Model. *Expert Systems with Applications 26* (4): 567-573. http://dx.doi.org/10.1016/j.eswa.2003.10.013

Lee, T., Chiu, C. Lu, C., & Chen, I. (2002). Credit Scoring Using the Hybrid Neural Discriminant Technique. *Expert Systems with Applications, 23*(3), 245-254. http://dx.doi.org/10.1016/S0957-4174(02)00044-1

Pallant, J. (2013). SPSS survival manual: A step by step guide to data analysis using IBM SPSS (5th ed.). Maidenhead: McGraw-Hill.

Peter, V., & Peter, R. (2006). Risk management model: an empirical assessment of the risk of default. *International Research Journal of Finance and Economics*, *1*, 42-56. Retrieved from https://www.researchgate.net/profile/Vasanthi_Peter/publication/268345909_Risk_Management_Model_an_Empirical_Assessment_of_the_Risk_of_Default/links/555a9b1208ae980ca6118d86.pdf

Reiss, A. J. Jr. & Albert Lewis Rhodes (1961). The Distribution of Juvenile Delinquency in the Social Class Structure. *American Sociological Review, 26*(5)*,* 720-732. http://www.jstor.org/stable/2090201

Saunders, M., Lewis, P., & Thornhill, A. (2007). *Research methods for business students* (4th ed.). Harlow: Financial Times Prentice Hall.

Shin, K. S., & Han, I. (2001). A case-based approach using inductive indexing for corporate bond rating. *Decision Support Systems*, *32*(1), 41-52. http://dx.doi.org/10.1016/S0167-9236(01)00099-9

Stewart, D. W., & Kamins, M. A. (1993). *Secondary research: Information sources and methods* (2nd ed.). London: Sage.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, *16*(2), 149-172. http://dx.doi.org/10.1016/S0169-2070(00)00034-0

Uddin, N. (2013). Consumer Credit Customers' Financial Distress Prediction by Using Two-Group Discriminant Analysis: A Case Study. *International Journal of Economics and Finance*, *5*(6), 55. http://dx.doi.org/10.5539/ijef.v5n6p55

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11–12), 1131–1152. http://dx.doi.org/10.1016/S0305-0548(99)00149-5