Applying Discriminant Model to Manage Credit Risk for Consumer Loans in Vietnamese Commercial Bank

Nguyen Thuy Duong,
PhD, Banking Faculty, Banking Academy of Vietnam
ngocnb@hvnh.edu.vn

Do Thi Thu Ha,
MA, Banking Faculty, Banking Academy of Vietnam
ngocnb@hvnh.edu.vn

Nguyen Bich Ngoc,
MA, Banking Faculty, Banking Academy of Vietnam
ngocnb@hvnh.edu.vn

Abstract. This study estimates a two-group discriminant function to determine the expected financial health of the consumer credit customers' of a bank of Vietnam by using five demographic, socio-economic, and loan characteristics of the sample borrowers. The estimated function is significant at one per cent level of significance and the model estimates financial health/group membership with average seventy-three per cent accuracy. Like developed countries, it is expected that use of the estimated discriminant function in the consumer credit decision making will decrease bad debts, will help to set risk based credit pricing for the clients and will make the credit granting faster and more accurate.

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

Применение дискриминационной модели в управлении риском потребительских кредитов в коммерческом банке Вьетнама

Нгуен Тху Дуонг,
д-р экон. наук, Банковский факультет, Банковская академия Вьетнама, Ханой, Вьетнам
ngocnb@hvnh.edu.vn

До Тхи Тху Ха,
магистр, Банковский факультет, Банковская академия Вьетнама, Ханой, Вьетнам
ngocnb@hvnh.edu.vn

Нгуен Бих Нгок,
магистр, Банковский факультет, Банковская академия Вьетнама, Ханой, Вьетнам
ngocnb@hvnh.edu.vn

В данной работе с помощью бинарной дискриминационной функции проведена оценка ожидаемого финансового «здоровья» пользователей потребительских кредитов, предоставляемых банком Вьетнама, используя пять демографических, социально-экономических видов займов характеристик пробы заемщиков. Оцениваемая дискриминационная функция оказалась достоверной при 1%-ном уровне значимости и применении модели оценки финансового «здоровья» потребителей выбранной группы потребителей, что дало результат с 73%-ной достоверностью. В развитых странах предполагается, что применение оценки с помощью дискриминационной функции при принятии решения в области потребительского кредита будет способствовать снижению числа плохих долгов, а также даст возможность устанавливать оценку платежеспособности с учетом риска. Это поможет ускорить оформление кредита и поднять уровень его обеспеченности.

Ключевые слова: потребительский кредит; финансовое неблагополучие; демографические и социально-экономические характеристики; бинарный дискриминационный анализ.

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 its 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 distinguishing 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 is look like the regression analysis in terms of 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). But, the discriminant analysis and the regression analysis 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/binary variable. Besides, 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:

$$Z = \alpha + \beta 1X1 + \beta 2X2 + \beta 3X3 + ... + \beta kXk. \quad (1)$$

In the model, Z = discriminant score, α = constant, β's = discriminant coefficient or weight, X's = predictor or independent variable. The coefficients of the independent variables are estimated such that the scores differ for the two groups substantially. 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** shows the pictorial presentation of the data collected on the two variables: X1 and X2 for the cases of the two-group G1 and G2. The X1 axis represents X1 variable and the X2 axis represents X2 variable. The discriminant analysis tries to separate the two groups by drawing a line as under. If the data is collected on more than two variables, then it is not possible to draw a scatter diagram as under as we have fixed two axes in a graph. But regardless of the number of variables, the discriminant analysis can generate positive and negative Z scores for the cases of the groups and possible to draw a diagram as a lower part of the **Figure 1**. The lower part represents the group membership by

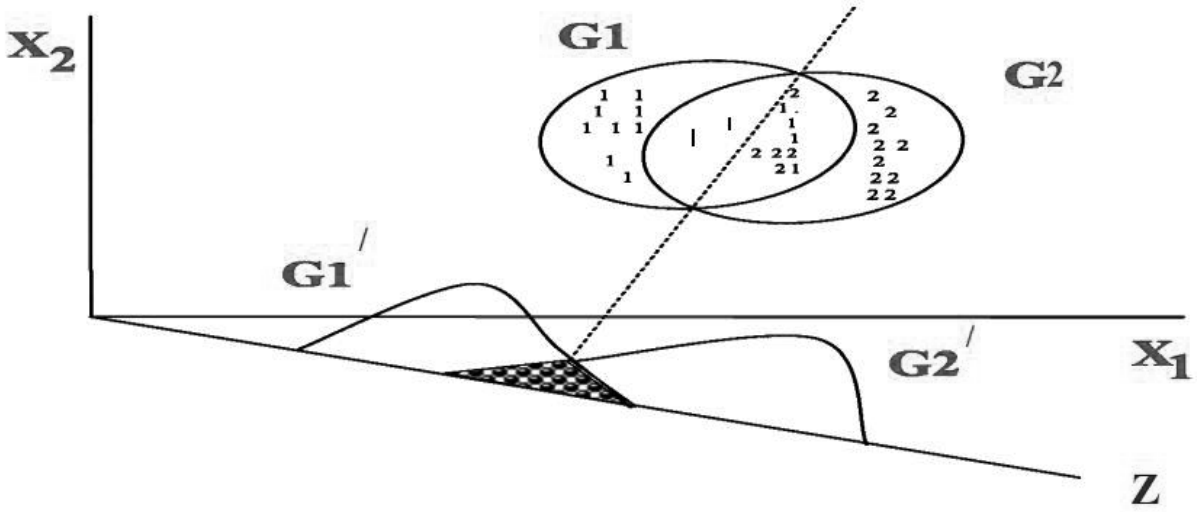

Figure 1. **Discriminant Analysis**

using the estimated discriminant scores (Z) of the groups cases. The shaded proportion represents the misclassification of the group membership. The smaller the shaded proportion, the bigger the estimation accuracy is assumed (Malhotra & Das, 2011; Boyd, Westfall, & Stasch, 2005)

The objectives are divided into two-broad objective and specific objectives. 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 consistent with 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 discriminate between the categories of the criterion or dependent variable. (ii) To examine whether significant differences exist among the groups 'in term of the predictor variables'. (ii) To determine which predictor variables contribute to most of the inter group differences. (iii) To classify cases to one of the groups based on the values of the predictor variables. (iv) To evaluate the accuracy of the classification. The first section of this research report is about introduction to the study which comprises prologue, objectives and methodology of the study. The second section contains literature review and the variables selection for the study. Empirical study in Vietnam's commercial banks, findings and their analysis are in the third section of the report.

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) and Reichert et al. (1983) 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 under-lying 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) (West, 2000), Decision tree (DT) (Jiang, 2009), case based reasoning (CBR) (Shin & Han, 2001) and Support vector machine (SVM) (Schebesch & Stecking, 2005) can be used as alternative methods for corporate credit risk prediction. 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 nonlinear pattern classification (Huang et al., 2004; West, 2000).

2.2. Discriminant Analysis for consumer credit

Wiginton (1980) conducted a discriminant analysis to model the consumer credit behavior by using demographic and economic variables. The demographic variables used are: number of dependents, living status, moved during last year, business use of vehicle and pleasure use of vehicle. The economic variables include industry class of employment, class of occupation and years in present employment. The right prediction power of the model estimated by the researcher is not encouraging and predicting group membership by using logit model provided better forecasting accuracy. It is concluded that years in present employment, living status and occupation type are significantly related to the credit risk rating. Grablowsky (1975) conducted a two-group stepwise discriminant analysis in order to model risk in the consumer credit by using behavioral, financial, and demographic variables. The behavioral data is collected from the two hundred borrowers through a questionnaire of summated ratings scale and the financial and demographic data are collected from the loan application forms of the same two hundred borrowers. The researcher has started analysis with thirty six variables and after a comprehensive sensitivity analysis, found that thirteen variables are enough to model the consumer credit risk. 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 holders by using two types 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 attitudinal variables used are: (a) use or non-use of other credit cards, (b) attitude toward credit, and (c) attitude toward bank charge-cards. The data for the quantitative and attitudinal variables on the same respondent is collected from the loan application forms and by the questionnaires respectively. The discriminant function estimated by them is significant at 0.01 level and forecasted the group membership with 78 per cent accuracy. Hand & Henley (1997) reviewed available credit scoring techniques in their article titled — "Statistical Classification Methods in Consumer Credit Scoring: A Review." 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 there is no best method. What is the best method depends on the structure and characteristics of the data. For a data set, one method may be better than the other method but for another data set, the other method may be better.

In addition, Davis, Edelman & Gammerman (1992) conducted a comparative study of various methods and concluded that all the methods are performed at the same accuracy level 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. The partial list of characteristics those may be useful to determine the group membership given by Capon (1982) includes the variables-telephone at home, own/rent living, age, time at home address, industry in which employed, time with employer, time with previous employer, type of employment, number of dependents, types of credit reference, income, savings and loan references, trade union membership, age difference between man and wife, telephone at work, length of product being purchased, age of automobiles, geographical location, debt to income ratio, monthly installment etc. Dinh & Kleimeier (2007) conducted a study for the Vietnam's retail banking market by using logistic regression analysis method. The variables they have used are age, education, occupation, total time in employment, time in current job, residential status, number of dependents, applicants annual income, family income, short-term performance history with the bank, long-term performance history with the bank, total outstanding loan amount, other services used, cash in hand and at bank, etc. They have argued that by using quantitative credit scoring, the default rate can be minimized from 3.3 per cent to 2.0 per cent. They also argued that by quantifying the credit risk, it is possible to set up risk-based pricing in the retail banking market. Consequently, the bank can become more efficient and competitive in the market. The most important predictors they found are time with bank, followed by gender, number of loans, and loan duration. Based on the above literature review, experience of the researcher and availability of the data, thirteen demographic and socio-economic variables are selected for this study. The variables are the loan amount, number of dependents, years of experiences at present job, salary per month, living status, savings per month, cash in hand and at bank, Net worth, ACT, N-EMI, EMI, interest rate (%), and Guar. The data is collected on the variables from the application forms of the consumer credit customers by filling up the pre-designed questionnaire.

3. RESEARCH METHODOLOGY

3.1. Research design

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 (Abdou & Pointon, 2009; Sarlija et al, 2004; Caouette et al, 1998; Hand 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; Banasik et al, 2001; Boyes et al,1989 and Orgler, 1971. 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:

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

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 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 bor-

rower'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: Dependents mean the number of persons who are dependent on the borrower; YAPJ stands for years at present job; Salary: how much 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.

Secondary data will be used in this study instead of primary data. To explain for this choice, advantages of using secondary data will be analyzed. Firstly, using secondary data, which already been available in commercial banks, might enables me to save time and money (Ghauri & Grṇhaug, 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 Vietnam (Duong, Tran & Ho, 2015) but also in other countries like Wiginton (1980); Elena Bartolozzi, Matthew Cornford, Leticia García-Ergüín, Cristina Pascual Deocón, Oscar Iván Vasquez & Fransico Javier Plaza (2008) and Hörkkö (2010). As a result of that, secondary data collected from commercial banks in Vietnam will be 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); Desai et al (1996); Boritz & Kennedy (1995) and Dutta et al (1994) 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.

3.2. Statistical analysis and checking assumptions

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 are 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 Kolmogorov-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 recognize 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 produced 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

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 Vasanthi and Raja (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 con-

Table 1. **Group Statistics**

	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
Years at present job	Not good	120	5.38	2.454	.194
	Good	280	10.26	3.679	.194
Number of independents	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

Table 2. **Tests of Normality**

	Kolmogorov-Smirnova			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

sidered 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 contribute to explain why loan amount of non-defaulters is relatively higher than defaulters.

As mentioned above, data used in discriminant analysis should be normally distributed (Khemakhem and 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, Reichrt (1983), Hand et al (1996) and 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 (Appendix 6), 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.

As the sig. (2-tailed) value for predictors are below the required cut-off of 0.05; 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 4, it is obvious that dependents and years at present job may best discriminate between the two groups of borrowers.

Table 3. **Correlations**

		Ability to pay loan	Age	Salary	Years at present job	Number of dependents	Loan amount
Ability to pay loan	Pearson Correlation	1	.480**	.098*	.560**	–.263**	.139**
	Sig. (2-tailed)		.000	.028	.000	.000	.002
	N	400	396	390	392	397	385
Age	Pearson Correlation	.480**	1	.106*	.770**	-.033	.063
	Sig. (2-tailed)	.000		.018	.000	.454	.161
	N	396	396	390	392	396	385
Salary	Pearson Correlation	.098*	.106*	1	.131**	.258**	.611**
	Sig. (2-tailed)	.028	.018		.003	.000	.000
	N	390	390	390	390	390	385
Years at present job	Pearson Correlation	.560**	.770**	.131**	1	-.193**	.016
	Sig. (2-tailed)	.000	.000	.003		.000	.715
	N	392	392	390	392	392	385
Number of dependents	Pearson Correlation	–.263**	–.033	.258**	–.193**	1	.223**
	Sig. (2-tailed)	.000	.454	.000	.000		.000
	N	397	396	390	392	397	385
Loan amount	Pearson Correlation	.139**	.063	.611**	.016	.223**	1
	Sig. (2-tailed)	.002	.161	.000	.715	.000	
	N	385	385	385	385	385	385

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

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

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 7) 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. Result presented in Table 7 shows that 81.5% of the cross validated grouped cases are classified correctly. The holdout sample is also used to check

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	95% Confidence Interval of the Difference	
									Lower	Upper
Salary	Equal variances assumed	1.058	.304	−2.207	506	.028	-.99318	.44992	−1.87712	-.10925
	Equal variances not assumed			−2.341	358.176	.020	-.99318	.42423	−1.82748	-.15888
YAPJ	Equal variances assumed	13.007	.000	−15.342	516	.000	−4.888	.319	−5.513	−4.262
	Equal variances not assumed			−17.793	440.828	.000	−4.888	.275	−5.427	−4.348
Number of dependents	Equal variances assumed	7.649	.006	6.220	521	.000	.497	.080	.340	.654
	Equal variances not assumed			6.344	318.735	.000	.497	.078	.343	.651
Loan amount	Equal variances assumed	16.188	.000	−3.148	503	.002	−70931539.4022	22531078.8278	−115198156.3286	−26664922.4757
	Equal variances not assumed			−3.690	457.412	.000	−70931539.4022	19224627.8185	−108711081.8770	−33151996.9274

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	.980a	100.0	100.0	.704

a. First 1 canonical discriminant functions were used in the analysis

Table 7. **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.

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 8 shows that 72.3 percent of cases are 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

1. 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. Available at: http://dx.doi.org/10.2307/2978604.
2. Boyd H.W. Jr., Westfall R., & Stasch S.F. (2005). Marketing Research: Text and Cases (7th ed., pp. 598–603). Richard D. Irwin, Inc., Homewood, Illinois-60430.
3. Capon N. (1982). Credit Scoring Systems: A Critical Analysis. *Journal of Marketing*, 46 (Spring), pp. 82–91. Available at: http://dx.doi.org/10.2307/3203343.
4. Credit Card Redlining. (1979). Hearings Before the Subcommittee on Consumer Affairs of the Committee on Banking, Housing and Urban Affairs, United States Senates, 96th Congress, Fist Session, on 15, June 4 & 5, 1979, Washington DC, U.S. Government Printing Office, pp. 183–184.
5. Davis R.H., Edelman D.B., & Gammerman A.J. (1992). Machine-Learning Algorithms for Credit Applications. *IMA J. Math. Appl.Bus. Industry*, 4, 43–51. Available at: http://dx.doi.org/10.1093/imaman/4.1.43.
6. Dinh T.H. T., & Kleimeier S. (2007). A Credit Scoring Model for Vietnam's Retail Banking Market. *International Review of Financial Analysis*, 16 (5), pp. 571–495. Available at: http://dx.doi.org/10.1016/j.irfa.2007.06.001.
7. George D., & Mallery P. (2006). SPSS for Windows Step by Step: A Simple Guide and Reference, 13.0 Update (6th ed., pp. 278–292), Pearson Education.
8. Glen J.J. (2001). Classification Accuracy in Discriminant Analysis: A Mixed Integer Programming Approach. *The Journal of Operational Research Society*, 52 (3), 328. Available at: http://dx.doi.org/10.1057/palgrave.jors.2601085.
9. 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.
10. Abdou H. & Pointon J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature, *Intelligent Systems in Accounting, Finance & Management,* 18 (2–3), pp. 59–88.
11. Mircea G., Pirtea M., Neamtu M., & Bazavan S. (2011). Discriminant analysis in a credit scoring model. *Paper of Faculty of Economics and Business Administration West University of Timisoara, Romania*.
12. Bank nonperforming loans to total gross loans. Available at: http://data.worldbank.org/indicator/FB.AST.NPER.ZS?locations=VN.
13. Elena Bartolozzi, Matthew Cornford, Leticia García-Ergüín, Cristina Pascual Deocón, Oscar Iván Vasquez & Fransico Javier Plaza. (2008). Credit Scoring Modelling for Retail Banking Sector. II Modelling Week, Universidad Complutense de Madrid, 16th — 24th June 2008. Available at: http://www.mat.ucm.es/momat/2008mw/creditscoring.pdf.
14. Thanh Thi Huyen Dinh, Stefanie Kleimeier, Stefan Straetmans. Bank Lending Strategy, Credit Scoring and Financial Crises. School of Business and Economics, Maastricht University, Maastricht, The Netherlands. Available at: http://stefanstraetmans.com/attachments/File/KD_SK_SS_CreditScoring_final.pdf.
15. Hörkkö M. (2010). The determinants of default in consumer credit market. Available at: http://epub.lib.aalto.fi/en/ethesis/pdf/12299/hse_ethesis_12299.pdf.
16. Duong T., Tran V., & Ho Q. (2015, January). A Proposed Credit Scoring Model for Loan Default Probability: a Vietnamese bank case. In *International Conference on Qualitative and Quantitative Economics Research (QQE). Proceedings* (p. 52). Global Science and Technology Forum.
17. Hintze J. (1998). NCSS statistical software. *NCSS,* Kaysville, UT.