

Consumer Credit Customers' Financial Distress Prediction by Using Two-Group Discriminant Analysis: A Case Study

Nasir Uddin¹

¹ Department of Business Administration, Yokohama National University, Japan

Correspondence: Nasir Uddin, Department of Business Administration, Yokohama National University, Tokiwadai 79-4, Hodogaya-Ku, Yokohama 240-8501, Japan. Tel: 81-45-339-3659. E-mail: uddin-nasir-xc@ynu.ac.jp

Received: October 30, 2012

Accepted: April 26, 2013

Online Published: May 21, 2013

doi:10.5539/ijef.v5n6p55

URL: <http://dx.doi.org/10.5539/ijef.v5n6p55>

Abstract

This study estimates a two-group discriminant function to determine the expected financial health of the consumer credit customers' of a bank of Bangladesh by using thirteen 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-five 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

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. In the US, in the year 1979, 20-30 per cent of all consumer credit decisions are made based on the discriminant analysis and most of the large institutions in the sectors: banks, finance companies, oil companies, retail merchants, and travel and entertainment cards used the discriminant analysis for their credit granting decision making (Credit Card Redlining, 1979). Unlike the US, in Bangladesh, banks and other financial institutions sanction loan to their client with the help of traditional credit approval method- based on the human assessment and the experience of the previous decisions. In this way, the various aspects of the consumer credit application are manually evaluated and based on that the decision is made about whether to grant credit. It is not possible to generate a concrete score about the new applicants whether to grant or not grant credit by using this conventional process. 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 this study, an effort is made to model the consumer credit of a bank of Bangladesh by using socio-economic, demographic, loan characteristics and discriminant analysis for reliable and efficient loan operations and to minimize the consumer credit risk. In other words, a quantitative effort is made to forecast the expected position of the consumer credit applicant via the discriminant analysis. 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:

$$Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k \quad (1)$$

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: X_1 and X_2 for the cases of the two-group G_1 and G_2 . The X_1 axis represents X_1 variable and the X_2 axis represents X_2 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, than 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 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).

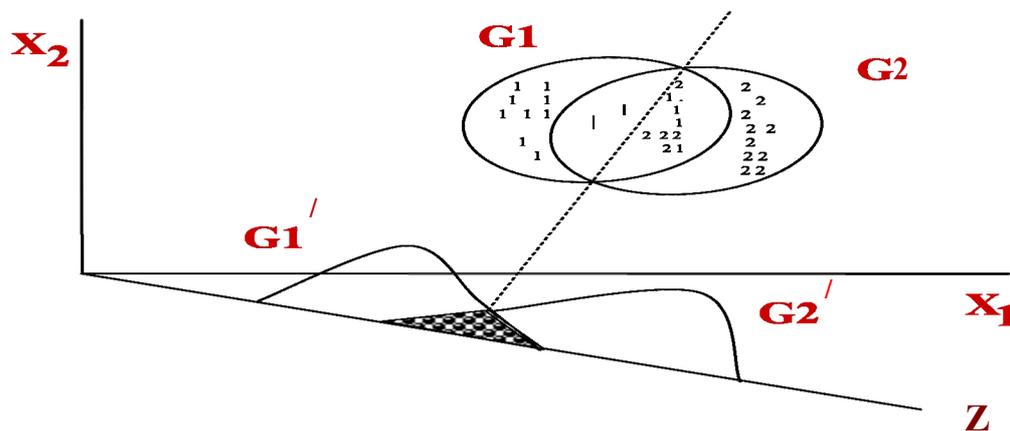


Figure 1. Discriminant analysis

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'. (iii) To determine which predictor variables contribute to most of the inter group differences. (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. Findings and their analysis are in the third section of the report. Fourth section consists of recommendations for the policy makers and conclusion of the study.

2. Literature Review

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 the

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 & Gammernan (1992) conducted a comparative study of various methods and concluded that all of 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. Methodology

3.1 Data

Both primary and secondary data are used in this study. The primary data is collected by a pre-determined questionnaire from the loan application forms of a private bank of Bangladesh and the secondary data is collected from the published journal articles, books, www, and SPSS manual. The primary data is collected on 15 default cases and 15 regular cases. A set of data is formed called-*analysis sample* by combining 10 regular and 10 default cases and a set of data is formed called-*holdout sample or validation sample* by combining the remaining 5 regular cases and 5 default cases. The analysis sample is used to estimate the discriminant function and the holdout sample is used to check the validity of the model. If possible, it is wise to collect the data for a

large sample size and to split the sample into two parts-analysis sample and holdout sample and to use the analysis sample to estimate the function and to use the holdout sample to check the validity of the model. After that, reverse the role of the data sets, to estimate the function by using the holdout sample and to use the analysis sample to check the validity of the model. This process is known as *double cross-validation*.

3.2 Data Analysis Technique, Software Used and Cautions

To analyze the collected data and to answer the research questions, the *direct method* discriminant analysis is used as data analysis technique for this study. According to the direct method of discriminant analysis, all of the variables are included in the study simultaneously without considering the discriminant power of the variables. This method is used when based on the previous research or a theoretical model, researcher wants that discrimination should be based on the all variables. The alternative of this approach is *stepwise discriminant analysis*. According to this approach, variables are included in the model according to their discriminating power. The softwares used in this study to analyze the data are SPSS, and MS-Excel. Like regression analysis, the *sample size* should be large enough to estimate a discriminant function. Inadequate sample size may produce wrong discriminant function. A substantially larger sample size than that is used in this study is expected for a true discriminant function to use in real life decision making. The *multicollinearity* problem is handled professionally. The author was very careful about selecting independent variable and ensured that any *unnecessary independent variable* is not included in the study. The *quality of the dependent variable* is ensured in this study. Sometimes, the quality of dependent variable may be poor. For instance, if the dependent variable is successful and unsuccessful salesman and the target to be successful salesman was set unrealistically high, the quality would be poor.

3.3 Description of the Variables

The variables used in this study are divided into two types: dependent variable and independent variables. The only dependent variable is *status* of the borrower that is a categorical variable. Based on the historical data, if a borrower's position is default then s/he is denoted by 1 and if the borrower's position is regular then s/he is denoted by 2. There are two types of the independent/predictor variables used in this study. Some variables are related with the *loan* and the others are related with the *demographic and socio-economic* conditions of the borrower. The *independent variables* related with the *loan* are as follows. *Loan*: The loan variable indicates the amount of loan borrowed by the borrower. *N-EMI*: The number of equal monthly installment. *EMI*: The amount of equal monthly installment paid by the borrower per month. *Interest*: The interest rate determined by the bank for the loan. and *Gua.*: The Gua. represents personal guarantor of the borrower. If the borrower provided personal guarantor then it is denoted by 1; otherwise denoted by 0.

The variables related with the *demographic and socio-economic* conditions of the borrower are as follows. *Dependents*: Dependents mean the number of persons who are dependent on the borrower. *Y-P-J*: Y-P-J stands for years of experience in the present job. *Salary*: Salary variable denotes the salary drawn by the borrower per month. *Living*: Living means status of living where the borrower resides. It may be rental or own. If own then it is denoted by 1 and if rental it is denoted by 0. *Savings*: Savings represent amount of money saved per month. *Cash*: Cash denotes amount of money present in hand & at bank of the borrower. *Net worth*: Net worth means personal net worth of the borrower. Net worth is calculated by subtracting the total liabilities from the total assets. *ACT*: Total number of bank accounts belonging to the borrower in other banks. and *Designation*: Designation of the present job of the borrower. Although data is collected on the designation, the variable is not included in the study because of extreme diversity in the designation.

4. Conducting the Discriminant Analysis

4.1 Group Means

Group means and standard deviations are calculated for each variable of the default and the regular groups. By examining the difference between the group means and the standard deviations, it is possible to see whether the variables can differentiate between default customers and regular customers. The groups statistics of the two-group can be used as *characteristics profile* of the two-group. The table 1 shows that group means are different for the groups for the variables- loan amount, dependents, monthly salary, savings, cash, net-worth, EMI and interest rate. So, these variables can differentiate the group membership successfully. Other variables: Y-P-J, living, ACT, N-EMI, and Guar. look similar in terms of magnitude- means that those variables do not play significant role in the case of determining group membership. The pooled within group correlations matrix is not reported here because of space problem shows very low correlations between variables-which indicates that there is no multicollinearity problems in the data. In the table-1 and paragraph-4.2, we have statistically tested whether the group means are different or same.

Table 1. Group statistics

Status	Default		Regular		Total	
	Mean	Deviation	Mean	Deviation	Mean	Deviation
Loan	6,01,000	4,24,013	10,72,000	11,36,403	8,36,500	8,69,059
Dependents	1.1	1.45	1.7	1.57	1.4	1.5
Y-P-J	7.3	4.64	7.4	4.62	7.35	4.51
Salary	63,682	57,750	1,15,849	2,05,246	89,765	1,49,165
Living	0.3	0.48	0.3	0.48	0.3	0.47
Savings	37,856	42,410	1,40,620	2,41,852	89,238	1,77,025
Cash	7,43,600	15,63,365	3,13,000	5,12,750	5,28,300	11,53,719
Net worth	1,02,27,584	1,42,89,700	42,11,100	36,23,888	72,19,342	1,06,05,220
ACT	1.2	0.42	1.2	0.63	1.2	0.52
N-EMI	54	8.49	55.2	8.39	54.6	8.24
EMI	16,114	10,516	28,711	28,950	22,412	22,162
Interest (%)	18.97	0.03	16.69	2.57	17.83	2.12
Guar.	0.6	0.52	0.5	0.53	0.55	0.51

4.2 Tests of Equality of Group Means

In order to test the equality of the group means, the Wilks' lambdas and the F ratios are estimated and reported as under. The Wilks' lambda (λ) for each predictor is the ratio of the within-group sum of squares to the total sum of squares. Its value varies between 0 and 1. The large value of λ indicates that group means are not different. On the other hand, small value of λ indicates that the group means are different. Sometimes, Wilks' λ is known as *U statistics*. The table 2 shows that the values of Wilks' λ are equal to 1 for the variables: Y-P-J, living and ACT. Consequently, these variables are insignificant in the case of determining group membership. In general, Wilks' λ is acceptable when its value is less or equal to 0.95. So, if we eliminate the variables having Wilks' λ greater or equal 0.95, our result of analysis should not be changed. The tests also shows that some predictors-interest rate, savings, EMT, net-worth, loan and dependents have significant role to distinguish default and regular borrowers. F values are calculated from a one-way ANOVA where the group variable serve as the categorical independent variable and each predictor variable serve as the metric dependent variable. The lower significant ratio for the corresponding F ratio means- the variable is very significant in the case of determining group membership. Conversely, the very high significant ratio for the corresponding F ratio means- the variable is very insignificant in the case of predicting group membership.

Table 2. Tests of equality of group means

Variables	Wilks' λ	F	Sig.
Loan	.923	1.508	.235
Dependents	.958	.790	.386
Y-P-J	1.000	.002	.962
Salary	.968	.599	.449
Living	1.000	.000	1.000
Savings	.911	1.752	.202
Cash	.963	.685	.419
Net worth	.915	1.666	.213
ACT	1.000	.000	1.000
N-EMI	.994	.101	.754
EMI	.915	1.673	.212
Interest (%)	.696	7.877	.012
Guar.	.990	.184	.673

4.3 Estimate the Discriminant Function Coefficients

Test of Equality of Covariance Matrices by Using Box's M: To estimate a valid discriminant function, an important assumption is that each of the groups is a sample from a multivariate normal population and the two groups have equal co-variance matrices although the two groups have different mean values. The Rank, in the table 3, means the size of the covariance matrices. The 13 means that this is a 13X13 matrix, the number of variables in the Discriminant function. The log determinants mean the natural log of the determinant of the

covariance matrices. In addition, the pooled within-groups is a matrix composed of by taking the average of each corresponding value within the two 13X13 covariance matrices of the two levels of the groups. The Box's M is a measure of the multivariate normality of the data which is based on the similarity of the log determinant of the two groups' covariance matrices. A transformed value of the Box's M is F ratio which tests the equality of the log determinants of the two covariance matrices. The F is conceptually equal to the F ratio in ANOVA which is the ratio of between group variability to within group variability. A significance value of .000 indicates that the data differ significantly from multivariate normal. However, a value less than 0.05 do not automatically disqualify the estimation of the discriminant analysis. Although the assumption is violated, the estimation is worthwhile which is validated in assessing the validation of the model section. This is surprising true for many cases. However, since the significance ratio is very low, it is justifiable to check the uni-variate normality of the variables.

Table 3. Test of equality of covariance matrices by using box's M

Status	Rank	Log Determinant	Box's M	Approx. F	df1	df2	Sig.
Default	13	125.757	423.287	2.248	91	2457.1	.000
Regular	13	142.772					
Pooled within-groups	13	149.382					

Note: Tests null hypothesis of equal population covariance matrices.

4.4 Determine the Significance of the Discriminant Function

Function 1, in the table 4, means that one discriminant function is estimated as we have two groups in the dependent variable. The eigen value means a ratio of between group sum of squares to within group sum of squares. The higher the value, the better estimation of the function and the minimum acceptable eigen value is more than one. The eigen value of the estimated function is 21.8 that counts for 100 per cent variance explained. The cumulative percent is also the same-100 per cent. The canonical correlation measures the association between the discriminant scores and the groups. The canonical correlation associated with the estimated function is 0.978. The coefficient of determination is equal to the square of the correlation coefficient that is $(0.978)^2 = 0.9565$ which means that 95.65 per cent of the variance in the dependent variable is explained by the estimated discriminant function. The Wilks' λ associated with the estimated function is 0.044 which is used to check the significance of the estimated function. The transformed χ is 35.92 with 13 degrees of freedom. The p-value (Sig.) associated with chi-square function is 0.00 which means that the null hypothesis is rejected at 1 per cent level of significance. So, estimating and interpreting the discriminant function are significant.

Table 4. Determine the significance of the discriminant function

Function	Eigen value	% of Variance	Cumulative %	Canonical r	Test of Function(s)	Wilks' λ	χ	df	Sig.
1	21.8(a)	100.0	100.0	.978	1	.044	36	13	.00

Note: a First 1 canonical discriminant function is used in the analysis.

4.5 Interpreting the Results

4.5.1 Structure Matrix

The structure correlations are also referred as *discriminant loadings*. The structure correlations represent the simple correlations between the predictors and the discriminant function. These correlations are used to determine the relative importance of the variables in predicting the group membership. The variables are ordered by absolute size of the correlations between the discriminating variables and the un-standardized canonical discriminant function in the table 5. The table 5 shows, the positions of the variables in determining the group membership according to the most important variable to the least important variable. According to the table, the most important variables those can determine the group membership are interest rate followed by savings, EMI, net-worth, loan, dependents, and cash. The least important variables are ACT followed by living, Y-P-J, N-EMI, and Gur.

Table 5. Structure matrix

Variables	Function
Interest (%)	.142
Savings	-.067
EMI	-.065
Net-worth	.065
Loan	-.062
Dependents	-.045
Cash	.042
Salary	-.039
Guar.	.022
N-EMI	-.016
Y-P-J	-.002
Living	.000
ACT	.000

4.5.2 The Function

Estimating the discriminant function coefficients is our main concern of this study. The discriminant function coefficients (unstandardized) are the multipliers of the variables, when the variables are in the original units of measurement. By using the estimated discriminant function coefficients, the required discriminant function, often called- “the discriminator” is as equation 2:

$$Z = -23.19092749 + .00001055Loan - 2.71769730Depen + .15850856Y-P-J + .00000933Salary - 1.34342089Living - .00001210Savings + .00000156Cash + .00000012Net\ worth + 3.92075936ACT - .16677507N-EMI - .00046422EMI + 1.68350521Interest + 1.29099513Guar \quad (2)$$

Table 6. Canonical discriminant function coefficients (unstandardized coefficients)

Variables	Function1
Loan	.00001055
Dependents	-2.7176973
Y-P-J	.15850856
Salary	.00000933
Living	-1.3434209
Savings	-.00001210
Cash	.00000156
Net worth	.00000012
ACT	3.9207594
N-EMI	-.16677507
EMI	-.00046422
Interest (%)	1.68350521
Guar.	1.29099513
(Constant)	-23.190928

The variable values of a new loan applicant will have to be substituted in the above equation 2 from the loan application form. If the estimated Z score of a loan applicant is positive, then the expected position of the applicant is default as the centroid is positive for the default group and the application should be rejected. The larger the distance between positive Z and 0, the default risk of the borrower is higher. Consequently, the management should look for higher risk premium. And if the estimated Z score of the credit applicant is negative, then the expected position is regular as the centroid is negative for the regular group and hence the loan should be allowed to the borrower. The larger the distance between negative Z and 0, the default risk of the borrower is lower. Consequently, the management should look for lower risk premium. Thus, management can use Z scores to set risk-based interest rate.

4.5.3 Group Centroids

The group centroids are the averages of the Z values calculated by the estimated model and reported in the last column of the table 8 for the default and regular groups. In other form, if the average values of the variables are substituted in the estimated discriminant function, the function generates the centroids. There are as many centroids as there are groups. There are two centroids in a two-group discriminant analysis-one for each group.

In this study, the centroid of the default group is 4.422 and the centroid of the regular group is -4.422. The group centroids are used to evaluate the expected position of the consumer credit customers. Now, if a consumer credit customer applies for a loan his raw/un-standard values for the variables will be substituted in the estimate discriminant function, the function will generate a positive or a negative value. The bigger the value the better forecasting is made. If the estimated Z value of a case is positive then the expected status of the case is default because the centroid value is positive for the default group and if the estimated value of a case is negative then the expected position of the case is regular as the centroid value is negative for the regular group case. The centroids are reported in the table 7:

Table 7. Functions at Group Centroids

Customer Type	Function1
Default	4.422
Regular	-4.422

Note: Unstandardized canonical discriminant functions evaluated at group means.

4.5.4 Casewise Statistics

The table 8 provides an excellent summary of the analysis. In the casewise statistics, the actual group means the actual position of the consumer credit customer on which the data is collected and the predicted group means the predicted position of the actual group member by the estimated discriminant model. The highest group means the highest possibility of being in a group according to the estimated discriminant model. The second highest is the alternative of the highest group as our analysis is the two group discriminant analysis. The last column is the estimated z values of the analysis sample cases.

Table 8. Casewise statistic

Case Number	Actual Group	Highest Group						Second Highest Group			Discriminant Scores
		Predicted Group	P(D>d G=g)		P(G=g D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g D=d)	Squared Mahalanobis Distance to Centroid		
			p	df						Function 1	
Original	1	1	.857	1	1.000	.033	2	.000	81.452	4.603	
2	1	1	.776	1	1.000	.081	2	.000	83.345	4.707	
3	1	1	.539	1	1.000	.377	2	.000	89.471	5.037	
4	1	1	.803	1	1.000	.062	2	.000	82.706	4.672	
5	1	1	.129	1	1.000	2.303	2	.000	53.684	2.905	
6	1	1	.440	1	1.000	.597	2	.000	92.492	5.195	
7	1	1	.659	1	1.000	.195	2	.000	70.611	3.981	
8	1	1	.093	1	1.000	2.829	2	.000	51.301	2.740	
9	1	1	.788	1	1.000	.072	2	.000	83.049	4.691	
10	1	1	.204	1	1.000	1.614	2	.000	102.309	5.693	
11	2	2	.485	1	1.000	.487	1	.000	91.052	-5.120	
12	2	2	.715	1	1.000	.133	1	.000	84.811	-4.787	
13	2	2	.846	1	1.000	.038	1	.000	81.698	-4.616	
14	2	2	.960	1	1.000	.002	1	.000	79.105	-4.472	
15	2	2	.935	1	1.000	.007	1	.000	79.675	-4.504	
16	2	2	.232	1	1.000	1.428	1	.000	100.794	-5.617	
17	2	2	.040	1	1.000	4.213	1	.000	46.132	-2.370	
18	2	2	.110	1	1.000	2.555	1	.000	52.505	-2.824	
19	2	2	.326	1	1.000	.967	1	.000	96.583	-5.405	
20	2	2	.932	1	1.000	.007	1	.000	79.737	-4.507	
Cross-validated ^a	1	1	.001	13	1.000	34.018	2	.000	110.299		
2	1	1	.002	13	1.000	32.524	2	.000	114.184		
3	1	1	.838	13	1.000	8.088	2	.000	97.549		
4	1	1	.000	13	1.000	60.140	2	.000	141.239		
5	1	1	.003	13	1.000	31.835	2	.000	59.788		
6	1	1	.000	13	1.000	67.261	2	.000	199.661		
7	1	1	.378	13	1.000	13.944	2	.000	73.899		
8	1	2**	.000	13	1.000	59.040	1	.027	66.244		
9	1	1	.000	13	1.000	226.335	2	.000	319.740		
10	1	1	.001	13	1.000	33.729	2	.000	180.813		
11	2	2	.212	13	1.000	16.733	1	.000	113.597		
12	2	2	.000	13	1.000	1296.554	1	.000	1590.733		
13	2	2	.000	13	1.000	1712.859	1	.000	1774.486		
14	2	1**	.000	13	1.000	848.669	2	.000	905.802		
15	2	1**	.000	13	1.000	3185.500	2	.000	3519.422		
16	2	2	.000	13	1.000	74.397	1	.000	264.910		
17	2	2	.005	13	1.000	30.019	1	.000	59.155		
18	2	2	.000	13	1.000	50.832	1	.013	59.495		
19	2	2	.000	13	1.000	291.115	1	.000	651.588		
20	2	2	.338	13	1.000	14.527	1	.000	88.218		

For the original data, squared Mahalanobis distance is based on canonical functions.
 For the cross-validated data, squared Mahalanobis distance is based on observations.

** Misclassified case

a. 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.

4.5.5 Histogram of Z Values of Status-1 (Default) & Status-2 (Regular)

The Z values estimated for the analysis samples in the last column of the above table 8 are presented in the bar diagrams-figure 2. The first bar diagram is prepared for the default group. The bar diagram and the above table 8 show that the minimum Z value is 2.74, the maximum Z value is 5.69, the average value is 4.42 and the standard deviation is 0.952. The estimated Z values are substantially higher than 0, indicates that the model forecasted the group membership of the samples of the default group in the analysis sample very accurately. The bar diagram in the right hand side shows the Z values of regular group. The bar diagram and the above table 8 show that minimum value is -5.62, the maximum value is -2.37, the average is -4.42 and the standard deviation is 1.05. The Z values are substantially negative which indicate that the accuracy of the model for the regular group is very high.

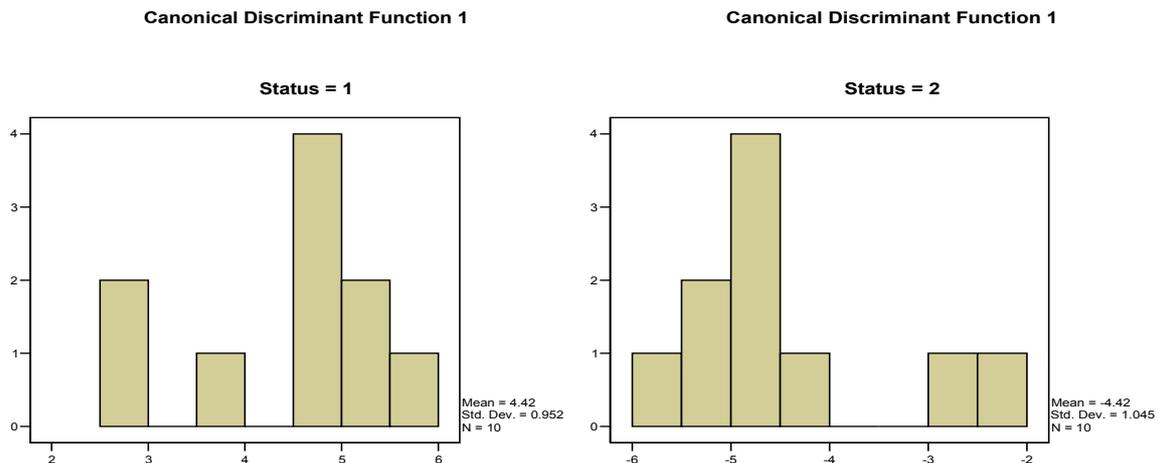


Figure 2. Histogram of Z values of status-1(default) & status-2 (regular)

4.6 Assessing the Validity of the Model

4.6.1 Classification Matrix of the Analysis Sample

The classification matrix is also known as *confusion or prediction* matrix and the matrix is used to check the validity of the model. The primal diagonal shows the correctly predicted cases and the off-diagonal shows the wrongly predicted group membership. The total of the primal diagonal element divided by the total number of cases used in the study is the correctly predicting rate-which is also known as *hit ratio*.

The classification matrix of the original sample (table 9) shows that 100 per cent of the cases 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 sample and the model is estimated. After that, the model is used to predict the membership of the case which was out of the sample at the time of the estimation of the function. The process is continued as many times as many cases in the analysis sample. Finally, the classification matrix is made. The lower part of the table 9 shows that 85 per cent of the cross-validated grouped cases are classified correctly. The cross validated hit ratio should be considered first compare to original hit ratio in order to assess the validity of the model.

Table 9. Classification results (b,c)

		Customer Type	Predicted Group Membership		Total
			Default	Regular	
Original	Count	Default	10	0	10
		Regular	0	10	10
	%	Default	100.0	.0	100.0
		Regular	.0	100.0	100.0
Cross-validated(a)	Count	Default	9	1	10
		Regular	2	8	10
	%	Default	90.0	10.0	100.0
		Regular	20.0	80.0	100.0

Notes: a 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. b 100.0% of original grouped cases correctly classified. c 85.0% of cross-validated grouped cases correctly classified.

4.6.2 Classification Matrix of the Holdout Sample

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 10 shows that 70 percent of cases are correctly classified.

Table 10. Classification results-holdout sample

		Customer Type	Predicted Group Membership		Total
			Default	Regular	
Original	Count	Default	5	0	5
		Regular	3	2	5
	%	Default	100	0	100
		Regular	60	40	100

Note: a. 70.0% of cases correctly classified.

4.6.3 Casewise Statistics of the Holdout Sample

By putting the values of the hold out sample in the estimated discriminant function, the table 11 of casewise Z values is constructed. Here, we see, in the holdout category, 5 default customers out of 5 are classified correctly and 3 regular customers out of 5 are incorrectly forecasted. In total, 7 out of 10 are classified correctly and 3 out of 10 are incorrectly predicted. To sum up, 70 per cent of the cases are classified correctly.

Table 11. Casewise statistics- holdout sample

SL No.	Status	Z Value	Predicted Status
1	Default	6.419573	Default
2	Default	6.613362	Default
3	Default	0.886851	Default
4	Default	1.963649	Default
5	Default	0.355264	Default
6	Regular	0.011423	Default**
7	Regular	-3.65053	Regular
8	Regular	-4.18221	Regular
9	Regular	5.496882	Default**
10	Regular	2.637537	Default**

Note: ** Misclassified Case.

4.6.4 Classification Matrix Using Holdout Sample as Analysis Sample

When the holdout sample is used as the analysis sample, the prediction matrix, table 12, is found. The matrix shows that 100 per cent of the original grouped cases and 90 per cent of the cross-validated grouped cases are

classified correctly.

Table 12. Classification results (b,c)- holdout sample as analysis sample

		Status	Predicted Group Membership		Total
			Default	Regular	
Original	Count	Default	5	0	5
		Regular	0	5	5
	%	Default	100.0	.0	100.0
		Regular	.0	100.0	100.0
Cross-validated(a)	Count	Default	5	0	5
		Regular	1	4	5
	%	Default	100.0	.0	100.0
		Regular	20.0	80.0	100.0

Notes: a 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. b 100.0% of original grouped cases correctly classified. c 90.0% of cross-validated grouped cases correctly classified.

4.6.5 Classification Matrix Using Total Sample as Analysis Sample

In this section, the analysis sample and the holdout sample is used as analysis sample again and the confusion matrix is constructed as under (table 13). It reveals that around 87 per cent of the original grouped cases and around 77 per cent of the cross-validated grouped cases are classified correctly.

Table 13. Classification results (b,c) - total sample as analysis sample

		Status	Predicted Group Membership		Total
			Default	Regular	
Original	Count	Default	13	2	15
		Regular	2	13	15
	%	Default	86.7	13.3	100.0
		Regular	13.3	86.7	100.0
Cross-validated(a)	Count	Default	12	3	15
		Regular	4	11	15
	%	Default	80.0	20.0	100.0
		Regular	26.7	73.3	100.0

Notes: a 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. b 86.7% of original grouped cases correctly classified. c 76.7% of cross-validated grouped cases correctly classified.

It is also wise to compare the hit ratio estimated based on the discriminant analysis and the hit ratio if the decision would be made by chance-randomly. If the groups are equal in size, then the hit ratio is 1/number of groups. In this study, there are two groups, so, if the decision is made randomly, the hit ratio is 50 per cent. There is no specific rules/guide line when the discriminant analysis should be conducted. However, some researchers argued that the hit ratio of the discriminant analysis should be higher at least by 25 per of the hit ratio that obtained by chance (Joseph, William, Barry, & Ralph, 2010; Glen, 2001). In addition, Boyd et al. (2005) mentioned that more than 70 percent accuracy is justified to conduct discriminant analysis. For this study, the average hit ratio is more than 75 per cent and hence, the validity is satisfactorily justified.

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 Bangladesh. The estimated function is significant at 1 per cent level of significance and could forecast financial health with average 75 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 Bangladesh. 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 it's 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

- 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/10.2307/2978604>
- Boyd, H. W. Jr., Westfall, R., & Stasch, S. F. (2005). *Marketing Research: Test and Cases* (7th ed., pp. 598-603). Richard D. Irwin, Inc. Homewood, Illinois-60430.
- Capon, N. (1982). Credit Scoring Systems: A Critical Analysis. *Journal of Marketing*, 46(Spring), 82-91. <http://dx.doi.org/10.2307/3203343>
- 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, First Session, on S15, June 4 & 5, 1979, Washington DC, U.S. Government Printing Office. pp. 183-184.
- 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. <http://dx.doi.org/10.1093/imaman/4.1.43>
- Dinh, T. H. T., & Kleimeier, S. (2007). A Credit Scoring Model for Vietnam's Retail Banking Market. *International Review of Financial Analysis*, 16(5), 571-495. <http://dx.doi.org/10.1016/j.irfa.2007.06.001>
- 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.
- Glen, J. J. (2001). Classification Accuracy in Discriminant Analysis: A Mixed Integer Programming Approach. *The Journal of Operational Research Society*, 52(3), 328. <http://dx.doi.org/10.1057/palgrave.jors.2601085>
- Grablowsky, J. B. (1975). A Behavioral Risk in Consumer Credit. *The Journal of Finance*, 30(3), 915-916. <http://dx.doi.org/10.2307/2326880>
- Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. The Open University, MK 7 6AA, UK. <http://dx.doi.org/10.1111/j.1467-985X.1997.00078.x>
- Joseph, F., Hair, Jr., William, C. B., Barry, J. B., & Ralph, E. A. (2010). *Multivariate Data Analysis with Readings* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Malhotra, N. K., & Dash, S. (2011). *Marketing Research: An Applied Orientation* (6th ed., pp. 552-582). Prentice Hall.
- Reiss, A. J. Jr. (1961). Socio-economic Index for Occupations in the Detailed Classification. *Occupations and Social Status* (pp. 114-138). New York: The Free Press.
- Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*, 15(3), 757-770. <http://dx.doi.org/10.2307/2330408>