Available online at www.koforiduapoly .edu.gh
International Journal of Technology and Management Research 3 (2014) 8-14

International Journal of Technology and Management Research

Using Multiple Discriminant Analysis to Construct a Statistical Model for Predicting Bank Loan Repay and Default Customers in the Eastern Region of Ghana

Godfred Kwame Abledu

Department of Applied Mathematics, Koforidua Polytechnic, P.O.Box KF 981, Koforidua, Ghana
Email: godfredabledu@gmail.com

Abstract

Banks that lend to small businesses and individuals need to quickly assess the creditworthiness of prospective borrowers so as to reduce the probability of issuing bad loans while attempting to maintain their own profitability. It was for these reasons that credit institutions have made several attempts at modeling and reliably forecasting credit default using numerous statistical approaches. The objective of the study was to develop a model which could be used to identify likely future defaulters. The population for the study was all financial institutions were in the Eastern Region of Ghana. A number of banks that could give the needed data for the study was purposefully chosen and a random sample of 150 customers were randomly selected to provide data on the study variables which include customers' financial standing, reason to loan, employment and demographic information. The statistical model obtained indicated four important influences - total asset, total income, family size and number of years with current employer as the most discriminating variables between the repay and default group. The validity of the model was confirmed using several diagnostic analytical procedures. The importance of examining a model's sensitivity and specificity in the context of one's specific, real-world objectives was also discussed.

Keywords: Loan repay and default; Creditworthiness; Prospective borrowers; Statistical models in banking; Multiple discriminant analysis.

1. Introduction

The process of administering loan varies from one financial institution to the other. However the loan repayment process appears to be the same that is the borrower must repay the lender. The banks monitor the likelihood of corporate default because of its impact on lenders and possible devasting effects on systemic stability of the bank. Financial institutions expect loan losses and therefore include the risk in loan pricing. Unexpected default erode capital to a potentially dangerous degree. Recent research at the bank has been aimed at quantifying the risk of default by individual companies. Some of these studies have used univariate and Multivariate statistical techniques to show how effective financial ratio sets can be when constructing company default prediction models (Altman, 1968, 1993; Beaver et.al, 1989).

The use of statistical methods for credit scoring and prediction of default on credit card accounts is now well-known. In particular, logistic regression has become a standard method for this task (Thomas et.al, 2002)). Recently there has been an interest in using survival analysis for credit scoring. This allows lenders to model not just if a borrower will default, but when the borrower will repay the loan. Survival analysis has been applied in many financial contexts including explaining financial product purchases (Tang et.al, 2007), behavioural scoring on credit customers (Stepanova and Thomas, 2001), predicting default on personal loans (Stepanova and Thomas, 2002) and the development of generic score cards for retail cards (Andreeva, 2006).

Multiple Discriminant Analysis is a much valued tool for market segmentation. Over the years, the estimation of the linear discriminant function has received much theoretical attention (Lopez, 2001; Malhotra and Malhotra, 2002; Crask and Perreault 1977; Morrison 1969).

The lender having fulfilled his part of the contract expects the borrower to fulfill his obligation without any delay. It must be mentioned that in granting loan facilities to customers, the express assumption has been and will continue to be that all parties will fulfill their obligations. However, numerous legal suits are reported in the media of borrowers who have defaulted the repayment of their loan. It is true that there are some categories of customers who deliberately default in loan repayment (Babajide, 2011; BoG, 2011; Suleiman, 2011*)*. It was for these reasons that credit institutions have made several attempts at modeling and reliably forecasting credit default using numerous approaches and methods. The problem of the study was to find a model for monitoring the loan repayment and predicting potential defaulters.

1.1 Objective of the study

The objective of the study was to develop a model which could be used to identify likely future defaulters. The model would develop statistical estimates based upon a cohort of borrowers. Using bank's past data files, a model can be developed around the historical relationships between borrower characteristics and the incidence of default. The resulting model can then be applied to borrowers in order to predict likely defaulters who should be the target of preemptive default prevention efforts.

1.2 The multiple discriminant analysis - conceptual and mathematical model

Multiple Discriminant Analysis is a technique for classifying a set of observations into predefined classes. It refers to all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation (Hair et.al, 2006). Multiple Discriminant Analysis is a multivariate technique which uses several variables simultaneously to classify an observation into priori groups, in this case, repay and default groups of customers. A linear combination of the variables used is formed into an equation, called the discriminant function.

$$D_i = a + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p \qquad (1)$$

The first term ***a***, represents the constant within the equation. The ***b***'s are discriminant coefficients or discriminant weights, and the ***x***'s are the input variables or predictors. The weights and the cutoff score are estimated in such a way to minimize the number of classification errors. The Maximum Likelihood estimator is broadly used in parameter estimation. For the sake of simplicity, it is presented here assuming that the vector of data collected at time t_i is modelled as:

$$y(t_i) = y_m(t_i, \theta^*) + \varepsilon_{i,} \qquad (2)$$

n_t the number of observation times, $y_m(t_i, \theta^*)$ the output of a deterministic model and θ* the true value of the parameter vector. It is assumed that the measurement errors $\varepsilon_{i,}\, i$ (i=1, ...,n_t) are independent, homoscedastic, zero mean and Gaussian, so $\varepsilon_{i,} \sim N(0, \sum\ )$. The likelihood of y is then defined as the probability density $\pi_y(y \mid \theta, \sum)$ of the data y being generated by y_t. The Maximum Likelihood estimator maximizes $\pi_y(y \mid \theta, \sum)$, or equivalently its logarithm. Considering the hypothesis of this research, the Maximum Likelihood estimator is:

$$\text{where, } L(\theta, \sum\ ) = -In\pi_y(y \mid \theta, \sum\ ) \qquad (3)$$

In the framework of Maximum Likelihood estimation, θ is considered as unknown but with a single actual value. Bayesian approaches consider a distribution of possible values for, θ. Hence, θ is assumed to have a known prior probability density $\pi_p(\theta)$. The joint probability density of y and θ satisfies the relation:

$$\pi(y, \theta) = \pi_y(y \mid \theta)\pi_p(\theta) = \pi_p(\theta \mid y)\pi_y(y) \qquad (4)$$

where, $\pi_y(y)$ is the marginal distribution of the observed data, defined by the relation:

$$\pi_y(y) = \int_{O} \pi_y(y \mid \theta)\pi_p(\theta)d\theta \qquad (5)$$

The posterior probability density for $\pi_p(\theta \mid y)$ is given by the Bayes rule

$$\pi_p(\theta \mid y) = \frac{\pi_y(y \mid \theta)\pi_p(\theta)}{\pi_y(y)} \qquad (6)$$

The maximum of a posteriori (MAP) estimator maximizes, $\pi_p(\theta \mid y)$.

2. Methodology

2.1 Population and sample

The population for the study was financial institutions in the Eastern Region of Ghana. A bank that could give the needed data for the study was purposefully chosen.

2.2 Procedure and variables selection

Content analysis as a research method (Elo and Kyngas, 2008; Lauri and Kyngas, 2005) was used to collect and analyse data from the bank's data set. The data set used in this case contains 150 cases and 6 variables (or predictors) with information pertaining to past and current customers who borrowed from a Ghanaian bank for various reasons. The data set contains information related to the customers' financial standing, reasons for obtaining the loan, employment, demographic information, among others. For each customer, the binary outcome "creditability" was also available. This variable contained information about whether each customer's credit was deemed "Good" or "Bad". The data set had a distribution of 89% credit worthy (good) customers and 11% not credit worthy (bad) customers. Customers who had missed 90 days of payment were thought of as bad risks, and customers who had missed no payment were thought of as good risks. Other typical measures for determining good and bad customers were the amount obtained over the overdraft limit, current account turnover, number of months of missed payments, or a function of these and other variables. The following variables were also measured:

(i) Basic personal information (Age, Sex)

(ii) Family information (marital status, number of dependents)

(iii) Employment status (years in current occupation)

(iv) Financial status (Most valuable available assets, number of year with current bank)

(v) Others: (purpose of credit, amount of loan).

The variables listed above were used to develop a model to discriminate between repay and default groups of customers. The assumption was that if the model could discriminate between these two groups, the predictive model can be used to classify or predict new cases where the above mentioned information are provided but credit standing of the borrower is unknown. This would be useful, for example, in deciding whether or not a person qualifies for a loan.

2.3 Data analysis

The data on a sample of 150 customers were analysed using discriminant analysis in the SPSS version 17 programme. The stepwise procedure was used. With this programme, the computer at each stage chose a variable to enter the discriminant function. The Wilks lambda criterion was used for entering the variables in the equation. The variable entered fitted the entry requirements in terms of the associated Wilks lambda value.

Approximately 70% of the customers who were previously given loans were used to create the model. The remaining customers who were previously given loans were used to validate the model results. The classification function was used to assign cases to groups. The discriminant model assigned the case to the group whose classification function obtained the highest score. Using the discriminant analysis function, loan default was predicted for individual loans in the portfolio and the prediction accuracies in terms of the sensitivity (proportion of default cases correctly identified to total number of default in the sample) and the specificity (proportion of non-default cases correctly predicted to total number of non-default cases in the sample).

The data was further analysed using the enter method to determine the best combination of variables that could give the highest prediction accuracy rates taking into consideration that the model and the function constructed and accepted were strong as indicated by the size of the eigen value. The larger the eigen value, the better the discriminating power of the function. Also, the Chi-Square and the Wilk's Lambda values were also assessed to determine discriminating power. SPSS was used to generate χ^2 approximation to in order obtain a significance level. The Wilk's Lambda was used to measure the differences between groups and the homogeneity within groups and to test the null hypothesis that the populations have identical means on *D*. A low Wilk's Lambda and a large Chi-Square with a significant p-value indicated good discriminating power of the discriminant function. Each subject's discriminant score was used to determine the posterior probabilities of being in each of the two groups. The subject was then classified (predicted) to be in the group with the higher posterior probability.

3. Results of the study

About 73(48.7%) of the respondents were males and 77 (51.3%) were females. Out of this number, 50 (33.3%) were teachers, 47 (33.3%) were civil servants, and about 53 (35.4%) were self employed (Table 4.1).

3.1 Background information about the respondents

Over 53 (65%) of the respondents have been with the bank for more than 3 years and 78% had been working for over five years.

Table 1
Sex and occupation of respondents

	Occupation			
sex	Teaches	Civil Servant	Self Employed	Total
Male	27	22	24	73(48.7%)
Female	23	25	29	77(51.3%)
Total	50(33.3%)	47(31.3%	53(35.4%)	150

Table 2 shows the descriptive statistics (i.e. means and standard deviations) of the variables used in the analysis. The mean values for asset, debt, family size and number of years with current employer are higher for the default group than the repay group. On the other hand, the repay group has higher means than the default group for income and number of years with current bank.

Table 2
Mean and standard deviation of the repay/default group

Group	Variables	Mean	Standard Deviation
Default	Asset	45.47	10.411
	Income	19.65	9.899
	Debt	16.35	13.162
	Family size	3.76	1.147
	Number of years with current employer	5.00	2.424
	Number of years with current bank	5.35	3.983
Repay	Asset	32.52	24.988
	Income	29.30	12.553
	Debt	18.50	16.245
	Family size	3.06	0.934
	Number of years with current employer	3.82	2.115
	Number of years with current bank	6.93	4.17

3.2 *Contribution of each variable to the model*

There are several tables that assess the contribution of each variable to the model. In this case study, the tests of equality of group mean and the discriminant function coefficients were used to assess the contributions of the independent variables to the dependent variable. The strength of the functions and discriminating abilities were all assessed by checking the eigenvalues, Wilk's Lambda and Chi-Square and its significance level.

3.3 Tests of Equality of Group Means

The tests of equality of group means measure each independent variable's potential before a model is created. Each test displays the results of a one-way ANOVA for the independent variable using the grouping variable as the factor. Table 3 shows that total asset, total income, family size and number of years with current employer are the most discriminating variables between the repay and default groups. All these four variables are significant at 0.05 level of significance (p=0.05).

Wilks' lambda is another measure of a variable's potential. Smaller values indicate the variable is better at discriminating between groups. Table 3 suggests that income is best in discriminating between groups, followed by years with current employer and asset. The associated chi-square statistic tests the hypothesis that the means of the functions listed are equal across groups. The significance of the chi-value (χ^2 = 47.557, p=0.000) indicates that the discriminant function does better than chance at separating the groups.

Table 3
Comparison of equality of group means

Variables	Wilks' Lambda	F	df1	df2	p- value
Asset	.970	4.445	1	146	.037
Income	.940	9.278	1	146	.003
Debt	.998	.274	1	146	.601
Family size	.948	8.083	1	146	.005
Number of years with current employer	.970	4.552	1	146	.035
Number of years with current bank	.985	2.177	1	146	.142

χ^2=47.557, p=0.000

3.4 Standardized discriminant function coefficients

The standardized coefficients allowed the researcher to compare the variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability. A low standardized coefficient might mean that the groups do not differ much on that variate or it might just mean that the variate's correlation with the grouping variable is redundant with that of another variate in the model.Table 4 shows the estimated standardized discriminant function coefficients.

The standardized coefficients allow for comparison of variables measured on different scales. Parameter values show that a percentage increase in asset, family size and number of years with current employer, ceteris paribus, will decrease the odds of probability of default by almost 4%, 38.3% and 13.4% respectively. On the other hand, a percentage increase in income, debt and number of years with current bank will increase the odds of probability of default by almost 9%, 1.7% and 7.1%respectively.

Table 4
Standardized discriminant function coefficients

Variables	Discriminant Coefficient
Asset	-0.04
Income	0.09
Debt	0.017
Family size	-0.383
Number of years with current employer	-0.134
Number of years with current bank	0.071
(Constant)	-0.221

The results in Table 4.4 give the following estimated discriminant function:

$$D = -1.221 - 0.04A + 0.09I + 0.017D_t - 0.383F - 0.134E + 0.071B \tag{7}$$

where, *A* is Assets; *I* is Income, *D* is Debt, *F* is Family size, E is Number of years with current employer, and B is Number of years with current bank.

The cutting score is zero. Discriminant scores greater than zero (positive scores) indicate a predicted membership in the default group, while negative scores imply predicted membership in the repay group. Correlations between variates and *D* are available in the loading or structure matrix. Generally, any variate with a loading of 0.30 or more is considered to be important in defining the discriminant dimension (Abdi and Williams, 2010).

These correlations may help us understand the discriminant function we have created.

The structure matrix shows the correlation of each predictor variable with the discriminant function. The ordering in the structure matrix is the same as that suggested by the tests of equality of group means and is different from that in the standardized coefficients table.

3.5 Prior and a priori probabilities for membership in groups

Table 5 displays the prior probabilities for membership in groups. A prior probability is an estimate of the likelihood that a case belongs to a particular group when no other information about it is available. The prior probabilities were based on the sizes of the groups. A priori, 88.5% of the cases were non-defaulters, so the classification function was weighted more heavily in favor of classifying cases as non-defaulters.

Table 5
Prior probabilities for groups

		Cases used in Analysis	
Pay/default group	Prior	Number	Percent
default	0.50	17	11.5
repay	0.50	131	88.5
Total	1.0	148	100

There were 17 cases belonging to the default group and 131 cases belonging to the repay group. Table 6 also shows that, 14 (82.4%) cases in the default group and 116 (88.5%) cases in the repay group were correctly classified by the discriminant function. These figures give the prediction accuracy of the discriminant function. The overall success rate or hit ratio of the discriminant function was 82. 5%. The classification results table shows that we correctly classified 88% of the subjects. To evaluate how good this is we should compare 88% with what would be expected by chance.

Table 6
Cross validation of predicted group membership

			Predicted Group Membership		
	Repay/Default Group		default	Repay	Total
Cases Selected	Original	default	14(82.4%)	3(17.6%)	17
		repay	15(11.5%)	116(88.5%)	131
	Cross-validated	default	13(76.5%)	4(23.5%)	17
		repay	15(11.5%)	116(88.5%)	131
Cases Not Selected	Original	default	0(0%)	0(0%)	0
		repay	1(50%)	1(50%)	2

4. Conclusion

The study demonstrated the use of discriminant analysis to identify demographic and behavioral characteristics associated with likelihood to default on a bank loan. The study identified four important influences - total asset, total income, family size and number of years with current employer as the most discriminating variables between the repay and default group. The validity of the model was confirmed using several diagnostic analytic procedures. The overall success rate or hit ratio of the discriminant function was 82. 5%.

The findings showed that using six variables and multiple discriminant analysis, a strong statistical model could be constructed that would be able to predict repay and default customer with very high correct classifications.

References

Abdi, H. and Williams, L. J., (2010). Principal Component Analysis, *Wiley Interdisciplinary Reviews: Computational Statistics*, Issue2, pp. 433–459.

Andreeva, G., (2006). European Generic Scoring Models Using Survival Analysis, *Journal of Operational Research in Social Science,* Issue 57, No.10, pp.1180-1187.

Altman, E., (1993). A Further Empirical Investigation of the Bankruptcy Cost Question, *The Journal of Finance September,* pp.1067-1089.

Altman, E., (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *Journal of Finance,* Vol. 23, No. 4, pp. 589–609.

Babajide, O., (2011). Loan Default Worry Microfinance Banks, *News Agency of Nigeria [online]:*

http://www.ghanamma.com/news/2011/05/21/loan-default-worry-microfinance-banks (Assessed on 13th June, 2011).

Bank of Ghana (BoG), (2011). *Loan Default Rate Surges* [online]:http://www.modernghana.com/news/317528/1/loan-default-rate-surges.html. (Assessed on 13th June, 2011).

Beaver, W. H., Eger, C., Ryan, S. and Wolfson, M., (1989). Financial Reporting, Supplemental Disclosures and Bank Share Prices, *Journal of Accounting Research*, Issue 27, No. 2, pp. 151-178.

Beaver, W. H., Engel, E. E., (1996). Discretionary Behavior with Respect to Allowances for Loan Losses and the Behaviour of Security Prices, *Journal of Accounting and Economics*, Vol.2, No. 3, pp.177- 206.

Cantor, R., (2001). Moody's Investors Service Response to the Consultative Paper Issued by the Basel Committee on Bank Supervision "A New Capital Adequacy Framework", *Journal of Banking and Finance*, Issue 25, pp.171-185.

Crask, M.R. and Perreault, W.D.Jr., (1977). Validation of Discriminant Analysis in Marketing Research, *Journal of Market Research*,Issue 14 (February), pp.60-68.

Elo, S. and Kyngas, H., (2008). The Qualitative Content Analysis Process, *Journal of Advanced Nursing* , Issue 62, No. 1, pp.107–115.

Estralla, E., Sankyun, P.and Peristiani, S., (2000). Capital Ratios as Predictors of Bank Failures, *FRBNY Economic Policy Review* (July) pp. 33-52.

Jacobson, T. and Roszbach K., (2003). Bank Lending Policy, Credit Scoring and Value-at-Risk, *Journal of Banking and Finance*, Issue 27, pp. 615-633.

Lauri S. and Kynga, S. H., (2005). *Developing Nursing Theories (Finnish: Hoitotieteen Teorian Kehitta¨minen).* Werner So¨ derstro¨m, Dark Oy, Vantaa.

Lopez, J.A., (2001). Evaluating the Predictive Accuracy of Volatility Models, *Journal of Forecasting*, Vol.20, Issue 2, pp.87-109.

Malhotra, R. and Malhotra D. K., (2002). Differentiating between Good Credits and Bad Credits Using Neuro-Fuzzy Systems,*European Journal of Operational Research*, Vol. 13, No.6, pp. 190-211.

Morrison, D.G., (1969). On Interpretation in Discriminant Analysis, *Journal of Marketing Research*, Vol.6 (May), pp. 156-163.

Stepanova, M. and Thomas, L.C., (2001). PHAB Scores: Proportional Hazards Analysis of

Behavioural Scores, *Journal of Operational Research in Social Science*, Issue 52, pp. 1007-1016.

Stepanova, M. and Thomas, L.C., (2002). Survival Analysis for Personal Loan Data, *Operational Research*, Issue 50, pp. 277-289.

Suleiman, M., (2011). *Bad Loans Choke Banks* at:http://www.graphic.com.gh/dailygraphic/page.php?news, (Accessed on 13th June, 2011).

Tang, L., Thomas, L.C., Thomas, S. and Bozzetto, J.F., (2007). It's the Economy Stupid:

Modeling Financial Product Purchases, *International Journal of Bank Marketing*, Vol.25, Issue 1, pp.22-38.

Thomas, L.C, Edelman D.B. and Crook, J.N., (2002). *Credit Scoring and Its Applications*. SIAM Monographs on Mathematical Modeling and Computation. Philadelphia, USA: SIAM.

Witten, I.H. and Frank, E., (2005). *Data Mining* (2nd Edn). Elsevier.