# DISCRIMINANT ANALYSIS IN GRANTING LOANS: A CASE STUDY OF KCB-KISUMU BRANCH

by

**BAKKER DANIEL KICHE**

A Project Submitted in Partial fulfillment

of the requirements for the degree of

Master of Science in Applied Statistics

**School of Mathematics, Statistics and Actuarial Science**

MASENO UNIVERSITY

## Abstract

This research study summarizes the loan evaluation method known as credit scoring using Discriminant Analysis. Credit scoring is a technique that helps banks decide whether to grant credit to applicants who apply to them or not. Credit department is faced with higher risk in making decisions. There is no general study that had been conducted in leading them toward making correct decisions. The main objective of this project was to determine the factors that can affect the bank's decision to grant a loan. The study estimated a discriminant function to determine the expected financial health of the consumer credit of customers of KCB Kisumu Branch by using eight demographic, socio-economic, and loan characteristics of the sampled borrowers. The data was analyzed using SPSS. The estimated function was found to be significant at one per cent level of significance and the model estimated a group membership with more than seventy-five per cent accuracy. This may decrease bad debts, and help to set risk based credit pricing for the clients and may also make the credit granting faster and more accurate.

# Chapter 1

# INTRODUCTION

This Project describes an approach to the decision to grant loan in a bank. The model presented was developed with a view toward incorporating the factors most relevant to the decision. Loan outcomes are first considered to be either default or non-default and are later expanded to include a detailed description of delinquent behavior. Loan characteristics (period, amount, interest rate, etc.) are used to determine rewards for correct classification. The resulting loan granting decision rules have as their objective the maximization of expected net present value.

This project is outlined in five parts as follows: The first chapter of this research report is about introduction to the study which comprises background information, objectives, methodology and the variables selected

for the study and significance of the study. The second chapter contains published literature in the area that is relevant to the loan granting decision. Basic concepts are in the third chapter of the report that gives the insight into the area of loan processing. Fourth chapter consists of findings and their analyzes of the study. Lastly we have the conclusions and recommendations for the policy makers.

## 1.1 Background Information

Loan granting, or consumer credit evaluation as it is sometimes popularly called, has been a topic of interest to the statistics community for many years. As a result of the increasing availability to the consumer of a number of forms of credit and the concurrent widening appeal of management science technique, credit managers have begun to view quantitative credit evaluation tools as both helpful and necessary.

The customer submits a completed loan application form which contains all information relevant to the loan granting decision. This information usually includes information such as age, amount, occupation, years at present job, car ownership among others. In addition, past and current credit history is included, as well as a mention of other accounts (sav-

ings, net worth, e.t.c.) that s/he may have with the institution. Given this initial information on the application form, a lending officer will conduct whatever credit investigation seems appropriate. He then makes a decision to accept or reject the application based on the application and credit investigation information and any additional information he may have. Each credit granting decision will affect the customer's use of related services of the institution. For example, if the customer's application is accepted, he will be more likely to open a new savings account at a bank than if his application is rejected. This cross selling effect plays a significant role in the credit granting decision.

Lending decision of a bank is very important because it determines the future profitability and performance of the bank. Recently banks are becoming more and more conscious in customer (borrower) selection to avoid the negative impact of bad loans or non-performing loans. The issue of non-performing loans (NPLs) has gained increasing attentions in the last few decades. Amounts of bad loans are alarmingly increasing in not only the developing and under-developed countries but also in developed countries. The immediate consequence of large amount of NPLs in the banking system is bank failure as well as economic slowdown [7].

Technically higher probabilities of indebtedness imply a higher level of risk for a bank. So the selection of borrowers must depend on credit risk of the borrower. Additional to the mandatory information included in standard CRB (Credit Risk Bureau) all commercial banks in Kenya developed their own loan application form asking for some extra information (personal and financial) about the borrower to increase likelihood of an appropriate loan decision and to strengthen their risk management systems there off.

At the moment there is no universal scoring model that could be used by all the financial institutions, due to the fact that each institution preserves its strategy in dealing with its customers. The scoring model in this project is based on discriminant analysis and it is pointed to the usage by the bank, by creating a tool that corresponds to variables analyzed simultaneously.

## 1.2 Statement of the Problem

In banking domain to know what are the best decisions to make is a concern for managers. An active banking area with higher risk is represented by credit department. Here, credit officers analyze the customers' application credit forms and calculate a score. The factors considered can influence more or less the credit scoring model. No study has been conducted in Kenya on those factors with high importance in loan granting.

## 1.3 Objective of the Study

The main objective of this project was to develop a model which best discriminates between the groups of customers. The specific objectives were:

1. To examine whether significant differences exist among the groups of customers in terms of predictor variables.

2. To classify the cases to one of the groups on the values of the predictor variables.

## 1.4 Significance of the Study

The use of the estimated discriminant function in the consumer credit decision making may decrease bad debts and losses by the bank, and may help to set risk based credit pricing for the clients and may make the credit granting faster and more accurate.

## 1.5 Research Methodology

### 1.5.1 Data Collection

For this study, the variable of interest was the information required in loan application form used by KCB Kisumu branch. Though it is very tough to reach to such level of confidential information, however some leverage has created the opportunity to collect a sample of 50 filled loan applications from the branch. Out of 50 filled loan applications, some 20 filled loan applications has been rejected due to lack of enough information and finally a total of 30 loan applications were selected for the research. The primary interest was whether the loan has been approved or not based on the information in the loan application. The loan application contains dependent variable loan decision-status (yes or no) and the explanatory variables.

The data was collected on 15 bad cases and 15 good cases. A set of data was formed called-analysis sample by combining 10 good and 10 bad cases and a set of data is formed called-holdout sample or validation sample by combining the remaining 5 good cases and 5 bad cases. The analysis sample is used to estimate the discriminant function and the holdout sample is used to check the validity of the model.

### 1.5.2 Data Analysis

To analyze the collected data and to answer the research objective, the direct method discriminant analysis was 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. The alternative of this approach is the stepwise discriminant analysis. This is where variables are included in the model according to their discriminating power. This was done by considering demographic and socio-economic characteristics of the applicant. The software used in this study to analyze the data was SPSS.

### 1.5.3 Description of the Variables

The variables used in this study were divided into two types: dependent variable and independent variables. If a borrower's position is bad then s/he was denoted by 1 and if the borrower's position is good, then s/he was denoted by 2. There were two types of the independent/predictor variables used in this study: variables that were related with the loan and the others that were related with the demographic and socio-economic conditions of the borrower.

The independent variables related with the loan were as follows: Amount: The loan variable *Amount* indicates the amount of loan borrowed by the applicant. The duration the loan shall take is the *Period*. The amount of equal monthly installment paid by the borrower per month is the *Repayment*. The *Rate* is the interest rate determined by the branch for the loan.

The variables related with the demographic and socio-economic conditions of the borrower were as follows: Salary: The variable *Salary* denotes the salary drawn by the borrower per month. Savings: The amount of money

saved per month by applicant represented by *Savings*. The amount of money present in hand and at the bank account of the applicant is the *Cash*. The variable *Net* means personal net worth of the borrower. Net worth is calculated by subtracting the total liabilities from the total assets.

Figure 1.1 denotes the variable and their notations as described above.

Table 1.1: Variables and Notations used.

| Variable | Notation |
|---|---|
| Period | $X_1$ |
| Amount | $X_2$ |
| Salary | $X_3$ |
| Cash | $X_4$ |
| Net Worth | $X_5$ |
| Repayment | $X_6$ |
| Rate | $X_7$ |
| Savings | $X_8$ |

where $X_1$, $X_2$ ,..., $X_8 > 0$.

In Table 1.2 we present the first five rows of the data collected and used in this analysis. It contained information about the loan period, amount, salary, cash, net worth, repayment, payment rate (percentage), savings, and credit approval (target variable). The first 8 variables will be further

on denoted with $X_1$ to $X_8$ and the target variable with $V_T$.

Table 1.2: Credit Data

| No. | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $V_T$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | $1.02M$ | $120K$ | $75K$ | $125K$ | $133K$ | 20 | $30K$ | $Y$ |
| 2 | 48 | $1.5M$ | $70K$ | $60K$ | $150K$ | $28K$ | 18 | $15K$ | $Y$ |
| 3 | 12 | $300K$ | $20K$ | $15K$ | $25K$ | $18K$ | 20 | $7K$ | $N$ |
| 4 | 42 | $450K$ | $30K$ | $20K$ | $35K$ | $15K$ | 18 | $10K$ | $Y$ |
| 5 | 24 | $500K$ | $40K$ | $80K$ | $100K$ | $20K$ | 19 | $10K$ | $Y$ |

Where $K$ denotes thousand, $M$ denotes Million, $N$ denotes No and $Y$ denotes Yes.

# Chapter 2

# LITERATURE REVIEW

## 2.1 How Credit Scoring has Developed in Importance

It is believed that credit scoring, can seriously help to answer some key questions. However, [2] has argued that while a lot of credit scoring models have been used in the field, these key questions have not been yet answered conclusively: What is the optimal method to evaluate customers? What variables should a credit analyst include to assess their applications? What kind of information is needed to improve and facilitate the decision-making process? What is the best measure to predict the loan quality (whether a customer will default or not)? To what extent can a customer be classified as good or bad?

One of the key components of risk management is that associated with the personal credit decision. This is one of the most critical banking

decisions, requiring a distinction between customers with good and bad credit. The behavior of former and current customers can provide a useful historical data-set, which can be crucial in predicting new applicant's behavior. With the fast growth of the credit industry all over the world and portfolio management of huge loans, credit scoring is regarded as one of the most important techniques in banks, and has become a very critical tool during recent decades. Credit scoring models are widely used by financial institutions, especially banks, to assign credit to good applicants and to differentiate between good and bad credit. Using credit scoring can reduce the cost of the credit process and the expected risk associated with a bad loan, enhancing the credit decision, and saving time and effort [25]. Decision-making involving accepting or rejecting a client's credit can be supported by judgmental techniques and/or credit scoring models. The judgmental techniques rely on the knowledge and both the past and present experiences of credit analysts whose evaluation of clients includes their ability to repay credit, guarantees and client's character [30]. Due to the rapid increase in fund-size invested through credit granted, and the need for quantifying credit risk, financial institutions including banks have started to apply credit scoring models.

Weingatnar [20] conducted a discriminant analysis to model the consumer credit behavior by using demographic and economic variables. The demographic variables used are: number of dependants, living status, moved during last year, business use of vehicle and pleasure use of vehicle. The economic variables included: industry class of employment, class of occupation and years in present employment.

[19] 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. The financial and demographic data are collected from the loan application forms of the same two hundred borrowers. The researcher started the analysis with thirty six variables and after a comprehensive sensitivity analysis, found that thirteen variables were enough to model the consumer credit risk. Although both sets of data- analysis sample and holdout sample violated the equal variance-covariance assumptions, the estimated model classified the validation sample 94 per cent correctly.

## 2.2 Key Determinants of Credit Scoring

The objective of credit scoring models is to assign loan customers to either good credit or bad credit or predict the bad creditors. Therefore, scoring problems are related to classification analysis [3, 14].

Credit scoring was primarily dedicated to assessing individuals who were granted loans, both existing and new customers. Based on pre-determined scores, credit analysts reviewed customer's credit history and creditworthiness to minimize the probability of delinquency and default [2]. The categorization of good and bad credit is of fundamental importance, and is indeed the objective of a credit scoring model. The need of an appropriate classification technique is thus evident. But what determines the categorization of a new applicant? From the review of literature, characteristics such as gender, age, marital status, dependants, having a telephone, educational level, occupation, time at present address and having a credit card are widely used in building scoring models [10, 18]. Time at present job, loan amount, loan duration, house owner, monthly income, bank accounts, having a car, mortgage, purpose of loan, guarantees and others have been also used in building the scoring models [17].

Classification models for credit scoring are used to categorize new applicants as either accepted or rejected with respect to these characteristics. These need to be contextualized to the particular environment, as new variables are appropriately included, for example, the inclusion of corporate guarantees and loans from other banks within the Egyptian environment in the investigation by [1].

While most of the authors have agreed about the importance of credit scoring methodology and the utmost necessity of developing a system "model" with a strong predictive ability, there has been disagreement about what is the most appropriate cut-off score in evaluating customer credit. The determination of the cut-off point(s) is central to the usefulness and value of credit scoring models.

## 2.3 Discriminant Analysis

This is a simple parametric statistical technique, developed to discriminate between two groups. Many researchers have agreed that the discriminant analysis approach is still one of the most broadly established techniques to classify customers as good credit or bad credit. This technique

has long been applied in the credit scoring applications under different fields. Therefore, credit scoring model based on a discriminant approach is basically used for statistical analysis to classify group's variables into two or more categories. The steps involved in conducting discriminant analysis are: the formulation, estimation, determination of meaning, interpretation, and validation of the results.

### 2.3.1 Assumptions of Discriminant Analysis

The major underlying assumptions of DA are:

1. The observations are a random sample;
2. Each predictor variable is normally distributed;
3. Each of the allocations for the dependent categories in the initial classification are correctly classified;
4. There must be at least two groups or categories, with each case belonging to only one group so that the groups are mutually exclusive and collectively exhaustive (all cases can be placed in a group)[13].

### 2.3.2 Purpose of DA

Discriminant Analysis is used to:

1. Determine the most parsimonious way to distinguish between groups;

2. Maximally separate the groups;

3. Discard variables which are less related to group distinctions [6].

The aim of the statistical analysis in DA is to combine (weight) the variable scores in some way so that a single new composite variable, the discriminant score, is produced. Therefore Discriminant analysis creates an equation which will minimize the possibility of misclassifying cases into their respective groups or categories.

## 2.4 Confusion Matrix (Average Correct Classification Rate Criterion)

It is one of the most widely used criteria in the area of accounting and finance (for credit scoring applications) in particular, and other fields, such as marketing and health in general. The average correct classification rate measures the proportion of the correctly classified cases as good credit and as bad credit in a particular data-set. The average correct classification rate is a significant criterion in evaluating the classification capability of the proposed scoring models.

It is believed that the average correct classification rate is an important criterion to be used, especially for new applications of credit scoring, because it highlights the accuracy of the predictions. Specifically, it ignores different misclassification costs for the actual good predicted as bad and the actual bad predicted as good observations. In the real field it is believed that the cost associated with Type II errors is normally much higher than that associated with Type I errors [5]. The estimated misclassification cost criterion simply measures the relative costs of accepting customer applications for loans that become bad versus rejecting loan applications that would be good. It is based on the confusion matrix; this criterion gives an evaluation of the effectiveness of the scoring models performance, which can cause a serious problem to the banks in the case of the absence of these estimations, especially with the actual bad predicted good observations. The estimated misclassification cost criterion, is a crucial criterion to evaluate the overall credit scoring effectiveness, and to find the minimum expected misclassification cost for the suggested scoring models.

## 2.5 Credit Scoring

Credit evaluation is one of the most crucial processes in bank's credit management decisions. This process includes collecting, analyzing and classifying different credit elements and variables to assess the credit decisions. The quality of bank loans is the key determinant of competition, survival and profitability of the bank. One of the most important kits, to classify a bank's customers, as a part of the credit evaluation process to reduce the current and the expected risk of a customer being bad credit, is credit scoring. [14] stated that "the process (by financial institutions) of modeling creditworthiness is referred to as credit scoring."

Anderson [3], suggested that to define credit scoring, the term should be broken down into two components, credit and scoring. Firstly, simply the word "credit" means "buy now, pay later". It is derived from the Latin word "credo", which means "I believe" or "I trust". Secondly, the word "scoring" refers to "the use of a numerical tool to rank order cases according to some real or perceived quality in order to discriminate between them, and ensure objective and consistent decisions". Therefore, scores might be presented as "numbers" to represent a single quality, or "grades" which may be presented as "letters" or "labels" to represent one

or more qualities.

Consequently, credit scoring can be simply defined as "the use of statistical models to transform relevant data into numerical measures that guide credit decisions. It is the industrialization of trust; a logical future development of the subjective credit ratings first provided by nineteenth century credit bureau, that has been driven by a need for objective, fast and consistent decisions, and made possible by advances in technology" [3]. Furthermore, "Credit scoring is the use of statistical models to determine the likelihood that a prospective borrower will default on a loan. Credit scoring models, are widely used to evaluate business, real estate, and consumer loans" [9]. Also, "Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the lenders" [19].

## 2.6 Judgmental Systems versus Credit Scoring Systems

The overall idea of credit evaluation is to compare the features or the characteristics of a customer with other earlier customers, whose loans they have already paid back. If a customer's characteristics are adequately similar to those, who have been granted credit, and have consequently defaulted, the application will normally be rejected. If the customer's features are satisfactorily like those, who have not defaulted, the application will normally be granted. Generally, two techniques can be used: "Loan officer's subjective assessment and credit scoring" [12].

Bailey [4] argues that in a judgmental technique evaluation, each credit application includes information contained within it, to be evaluated individually by a decision-maker "creditor". The success of a judgmental process depends on the experience and the common sense of the credit analyst. As a result, judgmental techniques are associated with subjectivity, inconsistency and individual preferences motivating decisions; and judgmental methods have some strengths, such as taking account of qualitative characteristics and having a good track record in evaluating past credit by utilizing the wealth of the credit analyst's past experience [2,

9].

## 2.7 Benefits of Credit Scoring

Credit scoring requires less information to make a decision, because credit scoring models have been estimated to include only those variables, which are statistically and/or significantly correlated with repayment performance; whereas judgmental decisions, prima facie, have no statistical significance and thus no variable reduction methods are available. Credit scoring models attempt to correct the bias that would result from considering the repayment histories of only accepted applications and not all applications. They do this by assuming how rejected applications would have performed if they had been accepted. Judgmental methods are usually based on only the characteristics of those who were accepted, and who subsequently defaulted [12].

Credit scoring models consider the characteristics of good as well as bad payers, while, judgmental methods are generally biased towards awareness of bad payers only. Credit scoring models are built on much larger samples than a loan analyst can remember. Credit scoring models can be seen to include explicitly only legally acceptable variables whereas it

is not so easy to ensure that such variables are ignored by a loan analyst. Credit scoring models demonstrate the correlation between the variables included and repayment behavior, whereas this correlation cannot be demonstrated in the case of judgmental methods because many of the characteristics which a loan analyst may use are not impartially measured. A credit scoring model includes a large number of a customer's characteristics simultaneously, including their interactions, while a loan analyst's mind cannot arguably do this, for the task is too challenging and complex. An additional essential benefit of credit scoring is that the same data can be analyzed easily and clearly by different credit analysts or statisticians and give the same weights. This is highly unlikely to be so in the case of judgemental methods [9, 12].

Previous studies in this field have generally been directed at a single important factor in the credit granting decision. The development of credit scoring formulas has received much attention. Its application was limited because of assumptions of linearity, inadequate modeling of the economics of the process, lack of attention to the effects of the delinquency behavior, and failure to consider potential profit from subsequent loans and other bank services.

In summary, it would seem from the literature that there is no optimal credit scoring model procedure, including specific variables or number of variables, particular cut-off point, exact sample size and meticulous validation, which can be applied to different banks in different environments. This was also the conclusion reached by other authors, e.g. [2], who came to a similar conclusion that there is no best scoring model holding explicit variables that can be used in different markets.

# Chapter 3

# BASIC CONCEPTS

## 3.1 Credit Scoring Fundamentals

We assume that a bank has access to information about its customers, regarding both the good payers (reimbursing loan without problems) and the bad payers (those who had problems with repayment over time). This information may relate to age, salary, social status, job stability and other reimbursement problems of individuals and to financial statements of legal persons. When a new customer is applying for a loan, the bank must decide whether to grant him the requested loan or not by applying a discrimination rule. As a result of this process, the applicant will receive a score which classifies the application in one of the existing categories (e.g. bad payers, good payers). The discrimination rule offers support for decision of granting or not granting a loan, by attending at the background of the applicant and providing the required risk assessment.

In this study, an effort was made to model the consumer credit of a bank in Kenya 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 model involves linear combinations of the equation of the form:

$$Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{3.1}$$

where $Z$ is the discriminant score, $\beta$ is the discriminant coefficient or weight for that variable, $X$ is respondent's score for that variable $\alpha$ is a constant and $k$ is the number of predictor variables.

Figure 3.1 below shows the pictorial presentation of the data collected on the two variables: $X$ and $Y$ for the cases of the two-group $G_1$ and $G_2$. The $X$ axis represents $X$ variable and the $Y$ axis represents $Y$ variable. The discriminant analysis tries to separate the two groups by drawing a line as above. If the data is collected on more than two variables, then it is not possible to draw a scatter diagram as above as we have fixed two axes in a graph. But regardless of the number of variables, the discrimi-

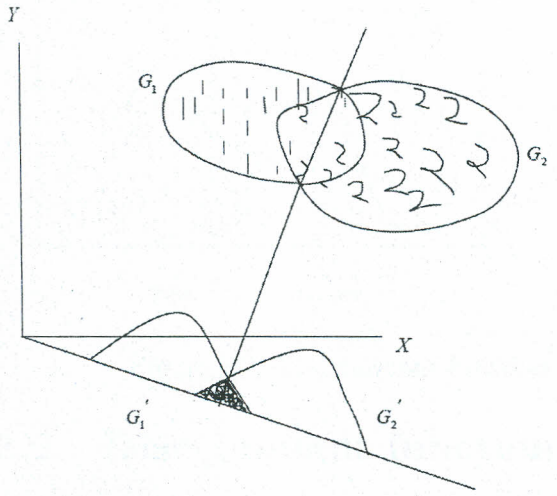


Figure 3.1: Pictorial Presentation of DA.

nant 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. The lower part represents the group membership by using the estimated discriminant scores ($Z$) of the group cases. The shaded proportion represents the misclassification of the group membership. The smaller the shaded proportion, the bigger the estimation accuracy is assumed [8].

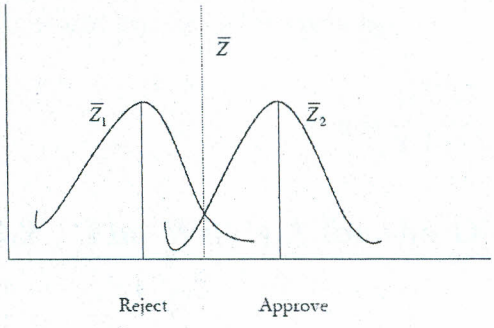


Figure 3.2: Discriminant function with two groups.

## 3.2 Discriminant function with two groups

Figure 3.2 shows how the two group can be partitioned. $\bar{Z}_1$ and $\bar{Z}_2$ are called the centroid. $\bar{Z}_1$ is the mean discriminant score (centroid) for group one, $\bar{Z}_2$ is the mean discriminant score (centroid) for group two and $\bar{Z}$ is the mean discriminant score (centroid) for the two groups.

### 3.2.1 The Canonical Correlation ($\eta$)

The canonical correlation of the predictor(s) with the discriminant scores produced by the model is given by:

$$\eta = \sqrt{\frac{\lambda}{1+\lambda}} \tag{3.2}$$

### 3.2.2 The Wilk's $\Lambda$ for the Discriminant Model

The Wilk's $\Lambda$ for the discriminant scores is given by

$$\Lambda = 1 - \eta^2 \tag{3.3}$$

Or

$$= \frac{1}{1+\lambda} \tag{3.4}$$

by replacing $\eta$ with equation (3.2). Where $\eta^2$ is the coefficient of determination and 1- $\eta^2$ is coefficient of non-determination.

## 3.3 The Hit Ratio of the Model and its Significance

$$\text{Hit ratio} = \frac{\text{Total primary diagonal}}{\text{Total number of cases observed}}$$

Hit ratio is currently the most common metric for measuring the accuracy of classifiers.

## 3.4 Cutting score

It can be used to sort the cases into either group based on their discriminant scores.

When $n_1 = n_2$, then

$$Z_{cutting} = \frac{\bar{Z}_1 + \bar{Z}_2}{2} \tag{3.5}$$

When $n_1 \neq n_2$, then

$$Z_{cutting} = \frac{n_1\bar{Z}_1 + n_2\bar{Z}_2}{2} \tag{3.6}$$

## 3.5 How SPSS Classifies Cases

In SPSS, case classification is accomplished by calculating the probability of a case being in one group or the other (i.e. good or bad), rather than by simply using a cutting score. This is accomplished by using a posterior probability of group membership using the Bayes' Theorem.

$$P(G_i|D) = \frac{P(D|G_i)P(G_i)}{\sum P(D|G_i)P(G_i)}$$

Where $D$ is the Discriminant Score (i.e. $Z$), $P(G_i|D)$ is the Posterior Probability that a case is in group $i$, given that it has a specific discriminant score $D$. $P(D|G_i)$ is the conditional probability that a case has a discriminant score of $D$, given that it is in group $i$. $P(G_i)$ is the Prior Probability that a case is in group $i$.

## 3.6 Estimating Misclassification Rates

To judge the ability of classification procedures to predict group membership, we usually use the probability of misclassification, which is known as the error rate. We could also use its complement, the correct classification rate. A simple estimate of the error rate can be obtained by trying out the classification procedure on the same data set that has been used to compute the classification functions. This method is commonly referred to as resubstitution. Each observation vector is submitted to the classification functions and assigned to a group. We then count the number of correct classifications and the number of misclassifications. The proportion of misclassifications resulting from resubstitution is called the apparent error rate. The results can be conveniently displayed in a classification

table or confusion matrix, such as Table 3.1.

Table 3.1: Classification for two groups.

| | | Predicted Group | |
|---|---|---|---|
| Actual Group | Number of observations | 1 | 2 |
| 1 | $n_1$ | $n_{11}$ | $n_{12}$ |
| 2 | $n_2$ | $n_{21}$ | $n_{22}$ |

Among the $n_1$ observations in $G_1$, $n_{11}$ are correctly classified into $G_1$ and $n_{12}$ are misclassified into $G_2$ , where $n_1 = n_{11} + n_{12}$ . Similarly, of the $n_2$ observations in $G_2$ , $n_{21}$ are misclassified into $G_1$ , and $n_{22}$ are correctly classified into $G_2$ ,where $n_2 = n_{21} + n_{22}$ therefore,

$$\text{Apparent error rate} = \frac{n_{12} + n_{21}}{n_1 + n_2}. \tag{3.7}$$

Similarly, we can define

$$\text{Apparent correct classification rate} = \frac{n_{11} + n_{22}}{n_1 + n_2}. \tag{3.8}$$

Therefore, the relationship between Apparent error rate and Apparent correct classification rate is:

$$\text{Apparent error rate} = 1 - \text{Apparent correct classification rate}. \tag{3.9}$$

## 3.7 Sensitivity and Specificity

Sensitivity relates to the test's ability to identify positive results correctly.

Hence

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$$

.

Specificity relates to the test's ability to identify negative results correctly.

Hence

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{Number of false positives}}$$

. Their relationships can be in the Table below:

Table 3.2: Relationship between Sensitivity and Specificity.

| | Positive | Negative |
|---|---|---|
| Positive | a | b |
| Negative | c | d |

By using the definitions above and Table 3.2,

$$\text{Sensitivity} = \frac{a}{a+c}. \tag{3.10}$$

Similarly,

$$\text{Specificity} = \frac{d}{d+b}. \tag{3.11}$$

In general, Positive means accepted application and negative means rejected application. Type I error is often referred to as a 'false positive', and is the process of incorrectly rejecting the null hypothesis in favor of the alternative. Type II error is the opposite of a Type I error and is the false acceptance of the null hypothesis. In most fields of science, Type II errors are not seen to be as problematic as a Type I error. With the Type II error, a chance to reject the null hypothesis is lost, and no conclusion is inferred from a non-rejected null hypothesis. The Type I error is more serious, because you have wrongly rejected the null hypothesis.

# Chapter 4

# RESULTS AND DISCUSSIONS

## 4.1 Group Means

In discriminant analysis we are trying to predict a group membership. Firstly we examine whether there are any significant differences between groups on each of the independent variables using group means and ANOVA. The Group Statistics and Tests of Equality of Group Means tables provide this information. If there are no significant group differences, it is not worthwhile proceeding any further with the analysis. A rough idea of variables that may be important can be obtained by inspecting the group means.

Group means are calculated for each variable of the bad and the good groups. By examining the difference between the group means, it is pos-

sible to see whether the variables can differentiate between bad customers and good customers.

Table 4.1 shows that group means are different for the groups for the variables- amount, period, monthly salary, savings, cash, net-worth, monthly payment and interest rate. These variables can differentiate the group membership successfully.

Table 4.1: Group Statistics

| Status | Bad | Good | Total |
|---|---|---|---|
| Variables | Mean | Mean | Mean |
| Amount | 601,000 | 1,072,000 | 836,500 |
| Salary | 63,682 | 115,849 | 89,765 |
| Savings | 37,856 | 140,620 | 89,238 |
| Cash | 743,600 | 313,000 | 528,300 |
| Net | 10,227,584 | 4,211,100 | 7,219,342 |
| Period | 54 | 55.2 | 54.6 |
| Repayment | 16,114 | 28,711 | 22,412 |
| Rate | 18.97 | 16.69 | 17.83 |

## 4.2 Tests of Equality of Group Means

In order to test the equality of the group means, the Wilks' $\Lambda$ and the F ratios are estimated and reported as in Table 4.2. The Wilks' $\Lambda$ varies between 0 and 1 inclusive. The large value of Wilk's $\Lambda$ indicates that group means are not different. On the other hand, small value of Wilk's $\Lambda$ indicates that the group means are different. In general, Wilks' $\Lambda$ is acceptable when its value is less or equal to 0.95. So, if we eliminate the variables having Wilks' $\Lambda$ greater or equal 0.95, our result should not change. The test also shows that some predictors like interest rate, savings, monthly repayment, net and loan amount have significant role to distinguish bad and good applicants. The lower significant ratio for the corresponding F ratio means that the variable is very important in the case of determining group membership. Conversely, the very high significant ratio for the corresponding F ratio means that the variable is not important in the case of predicting group membership.

Therefore, the smaller the Wilks' $\Lambda$, the more important the independent variable to the discriminant function and Wilks' $\Lambda$ is significant by the F test for all independent variables (See Table 4.2).

Table 4.2: Test of Equality of Group Means

| Variables | Wilks'Λ | F |
|---|---|---|
| Amount | 0.923 | 1.508 |
| Salary | 0.968 | 0.599 |
| Savings | 0.911 | 1.752 |
| Cash | 0.963 | 0.685 |
| Net | 0.915 | 1.666 |
| Period | 0.994 | 0.101 |
| Repayment | 0.915 | 1.673 |
| Rate | 0.696 | 7.877 |

## 4.3 Determining the Significance of the Discriminant Function

The eigenvalues provide information on each of the discriminant functions (equations) produced. The maximum number of discriminant functions produced is the number of groups minus one. We are only using two groups here, namely "bad" and "good", so only one function is displayed.

In Table 4.3 , one discriminant function is estimated as we have two groups in the dependent variable. The higher the value, the better the estimation of the function and the minimum acceptable eigenvalue is more than one. The eigenvalue of the estimated function is 21.8 that counts

for 100% variance explained. 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% of the variance in the dependent variable is explained by the estimated discriminant function. The Wilks' $\Lambda$ associated with the estimated function is 0.044 which is used to check the significance of the estimated function. Smaller values of Wilks' $\Lambda$ indicate greater discriminatory ability of the function. The chi-square is 35.92 with 8 degrees of freedom. The chi-square statistic tests the hypothesis that the means of the functions listed are equal across groups. The p-value (Significance) associated with chi-square function is 0.00 which means that the null hypothesis is rejected at 1 % level of significance. Therefore, estimating and interpreting the discriminant function are significant.

Table 4.3: Determining the significance of the discriminant function

| $\lambda$ | $\eta$ | $\Lambda$ | $\chi^2$ | $df$ | $Sig.$ |
|---|---|---|---|---|---|
| 21.8 | 0.978 | 0.044 | 35.92 | 8 | 0.00 |

Because $p < 0.01$, we can say that the model is a good fit for the data. This multivariate test is a goodness of fit statistic.

Therefore,

$$\text{canonical correlation}, \eta = \sqrt{\frac{21.8}{(1+21.8)}} = 0.9778$$

by equation (3.2).

Similarly,

$$\text{The Wilk's } \Lambda, = (1 - 0.9778^2) = 0.0439$$

by equation (3.3) or,

$$\text{Wilks' } \Lambda, = \frac{1}{1+21.8} = 0.0439$$

by equation (3.4), which confirms with the results from the table.

## 4.4 Interpreting the Results

### 4.4.1 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 Table 4.5.

Table 4.4: Structure matrix.

| Variables | Function 1 |
|---|---|
| Rate | 0.142 |
| Savings | -0.067 |
| Repayment | -0.065 |
| Net | 0.065 |
| Amount | -0.062 |
| Cash | 0.042 |
| Salary | -0.039 |
| Period | -0.016 |

The Table shows the positions of the variables in determining the group membership according to the most important variable to the least important variable. The most important variables that can determine the group membership are interest rate followed by savings, monthly repayment, net-worth, loan amount, and cash. The least important variables are salary and repayment period. Therefore, coefficients with large absolute values correspond to variables with greater discriminating ability.

Table 4.5: Canonical discriminant function coefficients (unstandardized coefficients).

| Variables | Function 1 |
|---|---|
| Amount | 0.0000106 |
| Salary | 0.0000093 |
| Savings | -0.0000121 |
| Cash | 0.00000156 |
| Net | 0.00000012 |
| Period | -0.1667751 |
| Repayment | -0.000464 |
| Rate | 1.683505 |
| (Constant) | -20.77995 |

## 4.5 The Discriminant Function

Estimating the discriminant function coefficients was 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, using equation (3.1), is as shown below:

$$\begin{aligned} Z = &-20.77995 + 0.0000106 \text{ Amount} + 0.0000093 \text{ Salary} \\ &- 0.0000121 \text{ Savings} + 0.00000156 \text{ Cash} + 0.00000012 \text{ Net} \\ &- 0.166775 \text{ Period} - 0.000464 \text{ Repayment} + 1.68351 \text{ Rate} \end{aligned} \tag{4.1}$$

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

## 4.6 Group Centroids

A further way of interpreting discriminant analysis results is to describe each group in terms of its profile, using the group means of the predictor variables. The group centroids are the averages of the $Z$ values calculated by the estimated model and reported in the last column of Table 4.6 for the bad and good groups. 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. Like in our case, there are two centroids in a two-group discriminant analysis-one for each group.

In this study, the centroid of the bad group is 3.53 and the centroid of the good group is -3.53. The group centroids are used to evaluate the expected position of the consumer credit customers. If a consumer credit customer applies for a loan his raw values for the variables will be substituted in the estimated discriminant function. The function will generate a positive or a negative value. The bigger the value the better the forecasting.

Note that unstandardized canonical discriminant functions are evaluated at group means.

Table 4.6: Functions at Group Centroids.

| Customer Type | Function |
|---|---|
| Bad | 3.53 |
| Good | -3.53 |

The $Z$ scores are computed as follows using the group means of individual variables from Table 4.1 and inserting in equation (4.1). For the bad group:

$$Z = -20.77995 + 0.0000106(601,000) + 0.00000933(63,682)$$
$$- 0.0000121(37,856) + 0.00000156(743,600) + 0.00000012(10,227,584)$$
$$- 0.166775(54) - 0.000464(16,114) + 1.68351(18.97) = 3.5656$$

Similarly, for the good group:

$$Z = -20.77995 + 0.0000106(1,072,000) + 0.00000933(115,849)$$
$$- 0.0000121(140,620) + 0.00000156(313,000) + 0.00000012(4,211,100)$$
$$- 0.166775(55.2) - 0.000464(28,711) + 1.68351(16.69) = -3.4740$$

If applicant's score on the discriminant function is closer to 3.53, then

probably s/he is a bad applicant. If the person's score on the DF is closer to -3.53, then s/he is probably a good applicant. In practical terms, we usually figure out which group a person is in by calculating a cut score halfway between the two centroids (See Figure 4.1). By using equation (3.5):

$$\text{Cut Score} = \frac{3.53 - 3.53}{2} = 0.000$$

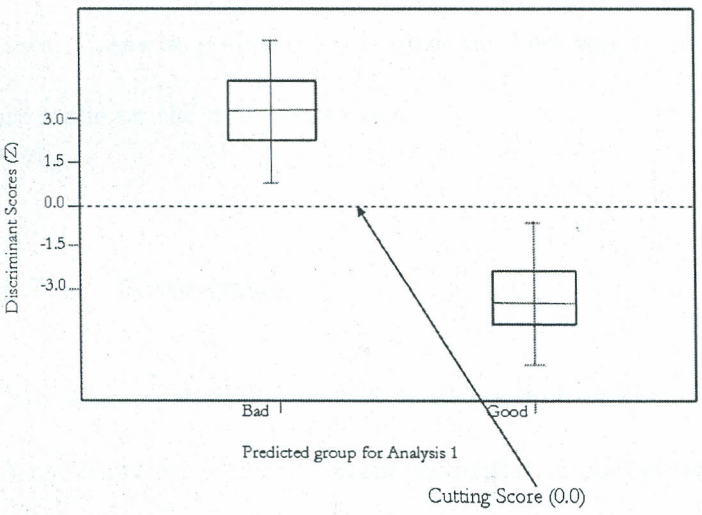


Figure 4.1: Box-Whisker plot illustrating the discriminant scores for the two groups.

## 4.7 Casewise Statistics

Figure 4.2 provides a summary of the analysis of the groups. In the casewise statistics, the actual group means the actual position of the consumer credit customer on which the data is collected and shows where the case actually belongs to. 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. Casewise statistics table gives the best way to see the predictions are made on the individual cases.

### 4.7.1 New cases

Only some of the columns of this table are of interest. First we have the "Case Number" which is just the sequential number of each subject. Then we have "Actual Group" which is the group number to which this subject actually belongs. Next we need to note that there are two - "Highest Group" and "Second Highest Group". Each of these refer to a group,

but which group they refer to is dependent on which group the discriminant analysis model assigned the highest probability to. "Highest Group" could refer to bad or good depending on which group the model assigned the highest probability to for this subject (and thus assigned the subject to) – likewise, of course for the Second Highest Group. In addition, suppose we have more than two groups, SPSS gives us only the Highest and Second Highest.

Case 1 we see from "Actual Group" was in group 1 (which in this case was the bad group). In addition, we see from "Predicted Group" that this subject was actually classified into group 1 (bad), and, obviously, bad was the group that the discriminant analysis model assigned the highest probability to. It then follows that the Second Highest group is 2, or good for this subject. If we look at the $P(G = g|D = d)$ columns in the Highest Group and Second Highest Group column, we see that, for each subject, these are two numbers that add to 1.000. These are the probabilities of group membership assigned by the discriminant analysis model. Thus our model is telling us that the probability of case 1 being a bad payers is 1.000 and the complementary probability of this subject being a good payers is 0.000. This is added information that we can use over-and-above

the model's decision to classify this subject as a bad payer; and we can have pretty good confidence in doing so.

However, an additional piece of information in the $P(D > d|G = g)$ column deserves note. This is often called the "typicality" probability, and represents the probability that this subject is from the assigned group. This is a consideration that this subject may be so extreme as to be an outlier, and thus not a member of any of the candidate groups. In the case of 17 for example, the typicality is only 0.040, thus we might want to consider whether the subject really belongs to the analysis. Considerable justification is needed beyond a small typicality to exclude a subject from the analysis however, (See Figure 4.2).

The $Z$ values estimated for the analysis samples in the last column of Figure 4.2 are presented in the bar diagrams. The first bar diagram is prepared for the bad group.The minimum $Z$ value is 1.851, the maximum $Z$ value is 4.804, and the average value of this group is 3.53. The estimated $Z$ values are substantially higher than 0. This indicates that the model forecasted the group membership of the samples of the bad group in the analysis sample very accurately (See figure 4.3). The second bar

| | | Highest group | | | | | Second highest group | | | Discrim-inant Scores |
|---|---|---|---|---|---|---|---|---|---|---|
| Case number | Actual group | Predicted group | P(D>d \| G=g) p | df | P(G=g \| D>d) | Squared Mahalanobis Distance to Centroid | Group | P(G=g \| D>d) | Squared Mahalanobis Distance to Centroid | Function 1 |
| original 1 | 1 | 1 | 0.857 | 1 | 1.000 | 0.033 | 2 | 0.000 | 81.452 | 3.714 |
| 2 | 1 | 1 | 0.776 | 1 | 1.000 | 0.081 | 2 | 0.000 | 83.345 | 3.818 |
| 3 | 1 | 1 | 0.539 | 1 | 1.000 | 0.377 | 2 | 0.000 | 89.471 | 4.148 |
| 4 | 1 | 1 | 0.803 | 1 | 1.000 | 0.062 | 2 | 0.000 | 82.706 | 3.783 |
| 5 | 1 | 1 | 0.129 | 1 | 1.000 | 2.303 | 2 | 0.000 | 53.684 | 2.016 |
| 6 | 1 | 1 | 0.440 | 1 | 1.000 | 0.597 | 2 | 0.000 | 92.492 | 4.306 |
| 7 | 1 | 1 | 0.659 | 1 | 1.000 | 0.195 | 2 | 0.000 | 70.611 | 3.092 |
| 8 | 1 | 1 | 0.093 | 1 | 1.000 | 2.829 | 2 | 0.000 | 51.301 | 1.851 |
| 9 | 1 | 1 | 0.788 | 1 | 1.000 | 0.072 | 2 | 0.000 | 83.049 | 3.802 |
| 10 | 1 | 1 | 0.204 | 1 | 1.000 | 1.614 | 2 | 0.000 | 102.309 | 4.804 |
| 11 | 2 | 2 | 0.485 | 1 | 1.000 | 0.487 | 1 | 0.000 | 91.052 | -4.23 |
| 12 | 2 | 2 | 0.715 | 1 | 1.000 | 0.133 | 1 | 0.000 | 84.811 | -3.897 |
| 13 | 2 | 2 | 0.846 | 1 | 1.000 | 0.038 | 1 | 0.000 | 81.698 | -3.726 |
| 14 | 2 | 2 | 0.960 | 1 | 1.000 | 0.002 | 1 | 0.000 | 79.105 | -3.582 |
| 15 | 2 | 2 | 0.935 | 1 | 1.000 | 0.007 | 1 | 0.000 | 79.675 | -3.614 |
| 16 | 2 | 2 | 0.232 | 1 | 1.000 | 1.428 | 1 | 0.000 | 100.794 | -4.727 |
| 17 | 2 | 2 | 0.040 | 1 | 1.000 | 4.213 | 1 | 0.000 | 46.132 | -1.48 |
| 18 | 2 | 2 | 0.110 | 1 | 1.000 | 2.555 | 1 | 0.000 | 52.505 | -1.934 |
| 19 | 2 | 2 | 0.326 | 1 | 1.000 | 0.967 | 1 | 0.000 | 96.583 | -4.515 |
| 20 | 2 | 2 | 0.932 | 1 | 1.000 | 0.007 | 1 | 0.000 | 79.737 | -3.617 |

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

Figure 4.2: Casewise Statistics.

diagram shows the $Z$ values of good group. The minimum value is -4.727, the maximum value is -1.480, and the average of this group is -3.53. The $Z$ values are substantially negative which indicate that the accuracy of the model for the good group is very high (See figure 4.4).

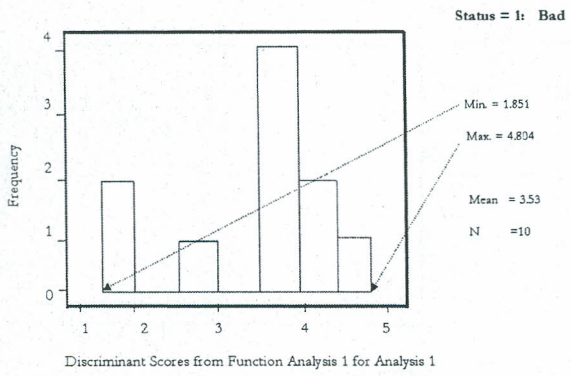


Figure 4.3: Histogram of $Z$ values of status = 1 (bad).

## 4.8 Assessing the Validity of the Model

### 4.8.1 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 predicted rate is also known as hit ratio.

The classification matrix of the original sample (See Table 4.7) shows that 100% of the cases are predicted by the model correctly. The cross validated set of data is a more honest presentation of the power of the

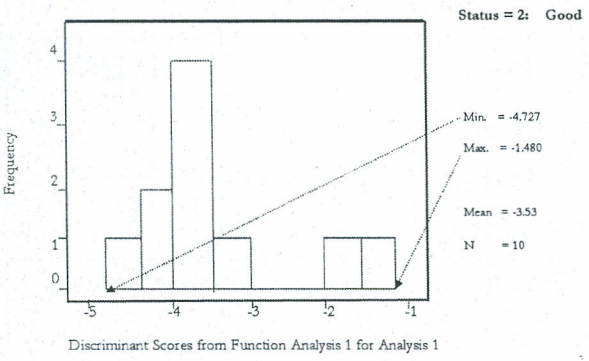


Figure 4.4: Histogram of $Z$ values of status = 2 (good).

discriminant function than that provided by the original classifications which often produces a poorer outcome. 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. Finally, the classification matrix is made. It also shows that 85% of the cross-validated grouped cases are classified correctly. The cross validated hit ratio should be considered first compared to original hit ratio in order to assess the validity of the model. The cross validation is often termed a "jack-knife" classification, in that it successively classifies all cases.

Table 4.7: Classification results

| | | Customer Type | Bad | Good | Total |
|---|---|---|---|---|---|
| Original | Count | Bad | 10 | 0 | 10 |
| | | Good | 0 | 10 | 10 |
| | % | Bad | 100 | 0 | 100 |
| | | Good | 0 | 100 | 100 |
| Cross-Validated | Count | Bad | 9 | 1 | 10 |
| | | Good | 2 | 8 | 10 |
| | % | Bad | 90 | 10 | 100 |
| | | Good | 20 | 80 | 100 |

### 4.8.2 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. Table 4.8 shows that 70 per cent of cases are correctly classified.

Table 4.8: Classification results-holdout sample.

| | | Customer Type | Bad | Good | Total |
|---|---|---|---|---|---|
| Original | Count | Bad | 5 | 0 | 5 |
| | | Good | 3 | 2 | 5 |
| | % | Bad | 100 | 0 | 100 |
| | | Good | 60 | 40 | 100 |

### 4.8.3 Casewise Statistics of the Holdout Sample

By putting the values of the hold out sample in the estimated discriminant function, Table 4.9 of casewise $Z$ values is constructed. Here, we see from Table 4.8 in the holdout category, 5 bad customers out of 5 are classified correctly and 3 good customers out of 5 are incorrectly forecasted. In total, 7 out of 10 are classified correctly and 3 out of 10 are incorrectly classified. To sum up, 70 per cent of the cases are classified correctly.

Table 4.9: Casewise statistics- holdout sample.

| No. | Customer Type | $Z$ Value | Predicted Type |
|---|---|---|---|
| 1 | Bad | 6.419573 | Bad |
| 2 | Bad | 6.613362 | Bad |
| 3 | Bad | 0.886851 | Bad |
| 4 | Bad | 1.963649 | Bad |
| 5 | Bad | 0.355264 | Bad |
| 6 | Good | 0.011423 | Bad ** |
| 7 | Good | -3.65053 | Good |
| 8 | Good | -4.18221 | Good |
| 9 | Good | 5.496882 | Bad ** |
| 10 | Good | 2.637537 | Bad ** |

Table 4.10: Classification results- holdout sample as analysis sample

| | | Customer Type | Bad | Good | Total |
|---|---|---|---|---|---|
| Original | Count | Bad | 5 | 0 | 5 |
| | | Good | 0 | 5 | 5 |
| | % | Bad | 100 | 0 | 100 |
| | | Good | 0 | 100 | 100 |
| Cross-Validated | Count | Bad | 5 | 0 | 5 |
| | | Good | 1 | 4 | 5 |
| | % | Bad | 100 | 0 | 100 |
| | | Good | 20 | 80 | 100 |

### 4.8.4 Classification Matrix Using Holdout Sample as Analysis Sample

When the holdout sample is used as the analysis sample, the prediction matrix is displayed on Table 4.10. The matrix shows that 100% of the original grouped cases and 90% of the cross-validated grouped cases are classified correctly.

### 4.8.5 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 in Table 4.11. It reveals that 86.7 per cent of the original grouped cases and 76.7

per cent of the cross-validated grouped cases are classified correctly.

Table 4.11: Classification results- total sample as analysis sample

| | | Customer Type | Bad | Good | Total |
|---|---|---|---|---|---|
| Original | Count | Bad | 13 | 2 | 15 |
| | | Good | 2 | 13 | 15 |
| | % | Bad | 86.7 | 13.3 | 100 |
| | | Good | 13.3 | 86.7 | 100 |
| Cross-Validated | Count | Bad | 12 | 3 | 15 |
| | | Good | 4 | 11 | 15 |
| | % | Bad | 80 | 20 | 100 |
| | | Good | 26.7 | 73.3 | 100 |

The Hit Ratio of the original grouped cases is given by

$$\text{The hit ratio} = \frac{13+13}{15+15} = 0.867$$

or 86.7 %

Similarly, Hit Ratio for cross validated grouped cases is given by

$$\text{The hit ratio} = \frac{12+11}{15+15} = 0.767$$

or 76.7 %

## 4.9 Testing the Sensitivity and Specificity

To test the sensitivity and specificity of the grouped cases, for the original,

$$\text{Sensitivity} = \frac{13}{13+2} = 0.867$$

by equation (3.10).

$$\text{Specificity} = \frac{13}{13+2} = 0.867$$

by equation (3.11).

Therefore, its overall % correctly classified = 86.7%.

Similarly, for Cross-validated grouped cases,

$$\text{Sensitivity} = \frac{12}{12+4} = 0.75$$

by equation (3.10).

$$\text{Specificity} = \frac{11}{11+3} = 0.786$$

by equation (3.11).

Therefore, its overall % correctly classified = 76.7%.

This shows how accurately the customers were classified into these groups. 75% sensitive test means that there are few false negative results (Type II error) while 78.6% specific test means that there are few false positive results (Type I error).

## 4.10 Misclassification rates

### 4.10.1 Apparent error rate

By equation (3.7),

for Original: $n_1 = 15$, $n_2 = 15$, $n_{12} = 2$ and $n_{21} = 2$. Therefore,

$$\text{Apparent error rate} = \frac{2+2}{15+15} = 0.1333$$

For cross-validated: $n_1 = 15$, $n_2 = 15$, $n_{12} = 3$ and $n_{21} = 4$ Therefore,

$$\text{Apparent error rate} = \frac{3+4}{15+15} = 0.2333$$

### 4.10.2 Apparent correct classification rate

By equation (3.8), For Original: $n_1 = 15$, $n_2 = 15$, $n_{11} = 13$ and $n_{22}= 13$. Therefore,

$$\text{Apparent correct classification rate} = \frac{13+13}{15+15} = 0.8667$$

For cross-validated: $n_1 = 15$, $n_2 = 15$, $n_{11} = 12$ and $n_{22}= 11$. Therefore,

$$\text{Apparent correct classification rate} = \frac{12+11}{15+15} = 0.7667$$

Alternatively, by equation (3.9),

$$\text{Apparent error rate} = 1 - \frac{13+13}{15+15} = 0.13333$$

or 13.3%.

This solution confirms the same as the first solution. Therefore it shall be true for the rest of the other solution considerations.

In conclusion, an error rate of 0.1333 is a less optimistic (more realistic) estimate of what the classification functions can do with future samples.

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. If the groups are equal in size, then the hit ratio is 1/number of groups. In this study, there are two groups, good and bad, so, if the decision is made randomly, the hit ratio is 50 per cent. There is no specific rule/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 cent of the hit ratio that is obtained by chance [13, 15, 19]. In addition, [10] mentioned that more than 70% accuracy is justified to conduct discriminant analysis. For this study, the average hit ratio is more than 75% and hence, the validity is satisfactorily justified.

# Conclusions and Recommendations

This study estimated a discriminant analysis in order to determine the expected status of the consumer credit customers of KCB - KISUMU BRANCH. The estimated function was significant at 1 % level of significance and could forecast financial health with more than 75% 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 KCB- Kisumu Branch.

Future research studies should use the advanced credit scoring techniques like genetic algorithms, fuzzy discriminant analysis and neural networks. The bank should also take evaluation of differential misclassification costs and use ROC information to choose a cutoff point which minimizes the total misclassification costs.

# References


[1] **Abdou, H.** Credit scoring models for Egyptian banks. *The University of Plymouth, UK*, 2009.

[2] **Al Amari, A.** The credit evaluation process and the role of credit scoring: A case study of Qatar. Ph.D. Thesis,. *University College Dublin*, 2002.

[3] **Anderson, R.** The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. *New York: Oxford University Press*, 2007.

[4] **Bailey, M.** Consumer credit quality: underwriting, scoring, fraud prevention and collections.. *Kingswood, Bristol: White Box Publishing*, 2004.

[5] **Banasik J., Crook J.** Reject inference, augmentation, and sample selection. . *European Journal of Operational Research* ,183 (3): 1582−1594, 2009.

[6] **Bierman, Harold, Jr. and Hausman, H.** The Credit Granting Decision. Management Science, Vol. 16, No. 8.

[7] **Bishnu Kumar Adhikary.** Nonperforming Loans in the Banking Sector of Bangladesh. *Realities and Challenges. Bangladesh Institute of Bank Management*, 2006.

[8] **Boyd, H. W. Jr., Westfall, R., Stasch, S. F** Marketing Research: Test and Cases (7th ed., pp. 598-603). *Richard D. Irwin, Inc. Homewood, Illinois-60430*, 2005.

[9] **Chandler, G. G., Coffman, J. Y** A comparative analysis of empirical vs. judgemental credit evaluation.. *The Journal of Retail Banking*, 1 (2): 15−26, 2009.

[10] **Chen, M., Huang, S.** Credit scoring and rejected instances reassigning through evolutionary computation techniques.. *Expert Systems with Applications* 24(4): 433−441, 2009.

[11] **Chuang, C., Lin, R.** a reassigning credit scoring model.. *Expert Systems with Applications* 36 (2/1): 1685−1694, 2009.

[12] **Crook, J.N.** Credit Scoring: An Overview. *Working paper no. 96/13, Department of Business studies, The University of Edinburgh*, 1996.

[13] **Glen, J. J.** Classification Accuracy in Discriminant Analysis: A Mixed Integer Programming Approach.. *The Journal of Operational Research Society* 52(3), 328, 2001.

[14] **Hand, D. J., Jacka, S. D** Statistics in Finance. *Arnold Applications of Statistics: London*, 1898.

[15] **Leonard, K. J.** Information Systems and Benchmarking in the Credit Scoring Industry. *Benchmarking for Quality Management and technology* 3 (1): 38-44, 2009.

[16] **OkoreAja** ). Major Determinants of Loan Repayment in a Developing Economy. *Empirical Evidence from Ondo State, Nigeria, Savings and Development* No. X, 1, 89–98, 1986.

[17] **Orgler, Y. E.** Evaluation of Bank Consumer Loans with Credit Scoring Models. *Journal of Bank Research* 2 (1): 31–37, 1971.

[18] **Sarlija, N., Bensic, M., Zekic-Susac, M** Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications* 36 (5): 8778–8788, 2009.

[19] **Thomas, L. C., Edelman, D. B., Crook, L. N.** Credit Scoring and Its Applications. *Philadelphia: Society for Industrial and Applied Mathematics*, 2002.

[20] **Weingartner, H. Martin** Concepts and Utilization of Credit-Scoring Techniques. *Banking*, 2006, pp. 51–58, 2006.