Optimisation of the Linear Probability Model for Credit Risk Management

[1]Kanelo Tsepiso Nyathi
Capital Bank Corporation
ICT Services Department
Harare, Zimbabwe

[2]Siqabukile Ndlovu [3]Sibonile Moyo [4]Thambo Nyathi
National University of Science and Technology
Computer Science Department
Bulawayo, Zimbabwe
[2]Email: siqabukile.sihwa {at} nust.ac.zw

Abstract— **One of the aims of the banking business is to provide loans to applicants. Credit risk management plays an important role in banks, as loans generally account for half to three-quarters of the total value of bank assets. Credit scoring is a systematic method for evaluating credit risk and assists decision makers determine whether or not to provide loans to applicants. Scoring models are systematic means of evaluating the creditworthiness of a loan applicant. However, existing scoring models cause some loan applications to be rejected unnecessarily as their credit rates are lowered to rejection levels due to lack of information such as previous loan payment data. This might be refusal of good credit, which potentially can cause the loss of future profit margins. This study aims at optimising one such credit scoring model to ensure that it uses only the critical scoring criteria to determine a credit score. The optimised model will not only reduce the proportion of unsafe borrowers, but also identify profitable borrowers.**

Keywords- credit risk; loans; linear probability; scoring model; optimisation

I. Introduction

Risk is the potential for a negative or positive future reality that may or may not occur. Risk is the possibility of adversity or loss [1]. It is defined by two characteristics of a possible negative future event: probability of occurrence (likelihood) and consequences of occurrence (impact). Risk is inherent in any walk of life in general and in financial sectors in particular, therefore risk management is an essential part of any financial institution. Risk management is a process that enables an organisation to manage and keep risk at levels that will protect its solvency and maintain its stability. The survival and success of a financial organisation depends critically on the efficiency of managing risks [2]. More importantly, good risk management is highly relevant in providing better returns to the shareholders [3; 4].

The objective of risk management is not to prevent risk taking activities, but to ensure that the risks are consciously taken with full knowledge, purpose and clear understanding so that they can be measured and mitigated. It also prevents an institution from suffering unacceptable loss causing it to suffer or damage its competitive position. Risk management involves the setting aside of funds, known as capital provisions or reserves, that will enable an institution to recover from the materialization of a risk. One aim of the banking business is providing loans to applicants. During this process, in order to make a decision whether to approve or reject a loan, the bank examines the credit worthiness of an applicant. This decision may be based on information contained in the financial statements of the borrower. Credit scoring is the system helping the decision maker such as the credit manager to determine whether or not to provide a loan to an applicant, basing on a set of predefined criteria.

A Credit Risk

Credit risk is the probability that some bank's assets, especially loans, will decline in value or become worthless [5]. According to [6], credit risk is simply defined as the potential that a bank borrower or counterparty will fail to meet their obligations in accordance with agreed terms. It plays a major role in a bank as loans are by far the largest asset. Loans generally account for half or three-quarters of the total value of all bank assets. The goal of credit risk management is to maintain credit risk exposure within acceptable levels, so that the bank can maximise risk-adjusted returns. According to [7] the effective management of credit risk is a critical component of a comprehensive approach to risk management and essential to the long-term success of any banking organisation.

B Risk Analysis

Risk analysis or assessment is the development of a quantitative or qualitative estimate of risk based on engineering evaluation and mathematical techniques [2, 8]. Credit Risk assessment is very important for any type of financial institution for avoiding huge amount of losses that may be associated with any type of inappropriate credit approval decision [8].

C Credit Scoring

The most important tool for the assessment of credit risk is credit scoring. Credit scoring can be defined as a systematic method for evaluating credit risk that provides a consistent analysis of the factors that have been determined to cause or affect the level of risk [9]. Scoring methods traditionally estimate the creditworthiness of a credit applicant. They predict the probability that an applicant or existing borrower will default or become delinquent [10]. A credit score is thus a numerical value that represents the level of risk of being defaulted. It rates how risky a borrower is; the higher the score, the less risk the person poses to creditors. A credit score (probability of default) is determined by a formula that takes into account many different factors. Credit scoring models compute a borrower's score primarily from information contained in their credit report [11].

Credit loans have a risk of being defaulted. To understand risk levels of creditors, bankers collect information on borrowers. Statistical predictive analytic techniques are used to analyse the information to determine the levels of default risk. Banking institutions require effective credit risk analysis tools to provide decision makers with the best possible information about the possibilities of losses. The credit rating tools should not only use an applicant's data for the rating process, but also make use of historical credit data. Predictive models can be developed from past historical records of loans. The models can learn the patterns of different credit default ratios from past credit information, and can be used to predict default risk levels of future credit loans.

D Linear Probability Models

Credit scoring is a statistical method used to classify applicants for credit into good versus bad risk classes [12]. Several statistical methods are used to develop credit scoring systems, including linear probability models, logit models, probit models, and discriminant analysis models. The linear probability model estimates the probability of default based on the characteristics of the borrower. It assumes that the probability of default varies linearly with these. [13]. Using this model, to determine the applicant's credit score, sixteen (16) criteria are used: product type, company type, credit history, number of years the borrower has dealt with the institution, number of years company has been in business, liquidity, turnover, gearing, debtor/creditor analysis, financial statements, collateral, cash flow, ability to pay, company management, and, the company's net worth. Each of these criteria for an applicant is assessed and is awarded points (weights). These weighted values are aggregated to produce a score. If the score reaches a certain level, the loan application may be approved; otherwise rejected. The following questions on an applicant's credit history are particularly interesting to note:

- Has the applicant been given a loan before?
- How much did they borrow?
- Has the applicant ever delayed payment?

II. PROBLEM STATEMENT

Current models heavily depend on the applicant's past credit history to approve applications. This leads to the rejection of applicants with no credit history. Thus, applicants with no credit history are unfairly penalised. It follows from this that if the rejected applicants are thought of as being a bad risk, then individuals with similar characteristics will potentially be rejected from receiving credit in any scoring system based on this data. An improved model that will attempt to infer the true risk status of the rejected applicants would address this problem. The building of such a model forms the basis of this research.

A Proposed Solution

To ensure fairness in credit scoring, a model is required, that is able to identify a relationship between previous credit scores and a set of linear probability evaluation factors (criteria), and uses this relationship on a new applicant's data to determine the credit score of the new applicant. In this paper we present an optimised linear probability model which provides the characteristics stated above.

III. RELATED MODELS

A Linear Probability Model/Linear Regression Model

Linear regression is the process of establishing a relationship between one dependent variable with one independent variable (simple linear regression) or between multiple independent variables (multiple linear regression). These models are developed by regressing a selection of quantitative and/or qualitative variables describing borrower characteristics against a dependant variable that takes a value of 1 if a loan is in default or 0 otherwise. Making assumptions about linearity and normally distributed target variables the predicted probabilities could lie outside the (0, 1) range [14]. This implies that there is no guarantee that a borrower can be classified as being either good or bad. This could be due to the borrower not being a customer long enough for their risk status to become clear. This may be problematic for application screening and capital adequacy purposes. That having been said, linear probability models remain the most common means of credit scoring/rating. Linear regression model provides a fairly robust estimate of the actual probability, given available information [14,15].

B Logit/Probit Models/ LogisticRegression Models

These utilise more sophisticated regression techniques that constrain estimated default probabilities within a 0-1 range. This is achieved by assuming that default probabilities are distributed in particular ways within that range e.g. probit models assume that default probabilities are normally distributed while logit models assume that a logistic distribution is more appropriate. Application scoring requires characteristics to be categorical, that is, to have two discrete classes of accept and reject applicants [16]. This is not possible with discriminant analysis and linear regression as the normal distribution is violated when categorical characteristics are used. Logistic regression addresses this problem. The objective of the logistic model in credit scoring is to determine the conditional probability of a specific observation belonging to a class given the values of the independent variables of the credit applications [17].

C Discriminant Analysis Models

These models seek to establish a linear classification rule or formula that best distinguishes between particular groups of borrowers, specifically defaulters and non-defaulters. Discriminant analysis models differ from the previous two in that instead of estimating a borrower's probability of default, it divides borrowers into high and low default-risk classes.
Discriminant analysis (DA) is a statistical technique that is used in modelling classification tasks [15] to predict group membership from a given set of predictors [18]. A DA model can be built step by step, where all available variables are reviewed and evaluated at each step to determine which contributes the most to discriminating between groups. It attempts to derive the linear combination of two or more independent variables that will best discriminate between a priori defined groups, for example, good and bad credit risk.

IV METHODOLOGY

A. *Theoretical Framework*

Credit scoring uses a technique that awards scores to borrowers as a means of evaluating the performance of their future loans [19]. It follows the framework presented in Figure 1.

The methodology used in this research is the design and creation methodology. In this methodology the following stages apply: awareness, suggestion, development, evaluation and conclusion. In the awareness stage the problem was identified as indicated in the problem statement. In the development stage, an optimised model was developed as indicated in Section B. In the evaluation stage the model was compared with existing model and it proved to perform better

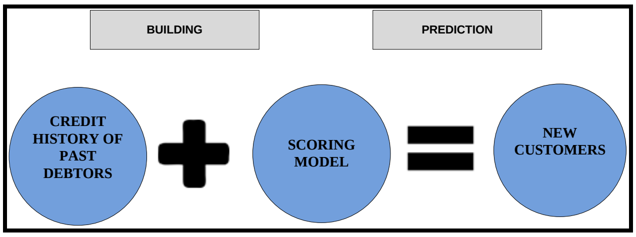

Figure 1. Credit Scoring Framework

The relationship between the dependent and the independent variables in the current model is linear. The model can be defined by Equation 1a.

$$\boldsymbol{S} = \boldsymbol{f_1} + \boldsymbol{f_2} + \boldsymbol{f_3} + \cdots \boldsymbol{f_n} \qquad \textit{Equation 1a}$$

Where $\boldsymbol{S}$ is the credit score and $\boldsymbol{f_i}$ are the independent variables, which are the scoring factors shown in Table 1. This model can also be represented as Equation 1b:

$$\mathbf{S} = \sum_{i=1}^{n} \boldsymbol{f}_{\mathbf{i}} \qquad \textit{Equation 1b}$$

Each of the factors is given equal weighting and the score $\boldsymbol{S}$ should fall on or above a set threshold value for the applicant to be granted a loan. This threshold is defined by the bank's credit policy. Based upon the value of $\boldsymbol{S}$, a credit decision is made, also based on the bank's credit policy.

TABLE I. CREDIT SCORING INDEPENDENT VARIABLES

FINANCIAL	CASH FLOW MANAGEMENT	GENERAL	MANAGEMENT
Turnover	Ability to service facility	Industry and products	Management Structure
Liquidity	Ability to repay facility	Ownership	
Gearing		Credit Check	
Debtors and Creditor Analysis		Years in Business	
Financial Statement		Years with Bank	
Security			
Net worth			
Stock obsolescence			

B. Optimised Model

The optimised model makes use of logistic regression to predict the probability of default of each applicant. Logistic regression is ideal for the scoring model as the model makes use of qualitative independent variables. Logistic regression models the relationship between multiple independent variables and a dependent variable. It provides a coefficient, which measures each independent variable's partial contribution to variations in the dependent variable. The independent variables are all potentially relevant parameters to credit risk [20]. A logistic regression model is represented in Equation 2.

$$\boldsymbol{Y} = \boldsymbol{a} + \sum_{k=1}^{n} \boldsymbol{\beta}_{\mathbf{k}} \boldsymbol{X}_{\mathbf{k}} \qquad \textit{Equation 2}$$

Where Y is the dependent variable being predicted, $\boldsymbol{X}_{\mathbf{k}}$ are the independent variables used in the prediction, β denotes the coefficients associated with $\boldsymbol{X}_{\mathbf{k}}$, or multipliers that describe the size of the effect the independent variables are having on the dependent variable Y, and $\boldsymbol{a}$ is the intercept, or the value

Y is predicted to have when all the independent variables are equal to zero. The relationship between these variables and the dependent variable can be denoted by Equation 3a.

$$\boldsymbol{S = f_1\omega_1 + f_2\omega_2 + f_3\omega_3 + \cdots f_n\omega_n} \quad \textit{Equation 3a}$$

where $\boldsymbol{S}$ is the credit score given the $\boldsymbol{n}$ input variables $\boldsymbol{f_i}$. $\boldsymbol{\omega}$ is the weight of variable $\boldsymbol{f_1}$ and is synonymous to coefficient $\boldsymbol{\beta}$, which is a set of unknown parameters weighting the influence of the independent variables. The input variables are the credit scoring factors presented in Table 1. This model can be represented logistically as Equation 3b:

$$\boldsymbol{S = \sum_{k=1}^{n} f_k\omega_k} \quad \textit{Equation 3b}$$

The credit score can now be determined by the model as presented in Figure 2.

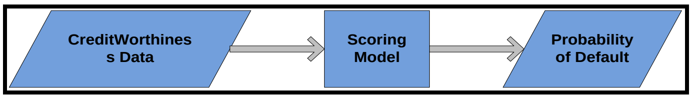

Figure 2 Optimised Credit Scoring

IV. IMPLEMENTATION

```
public void loadModel(){

classFile cfl = new classFile();
String fname = cfl.fileName;//scoring data file

try{
//read scoring data ARFF file
Instances data = null;
BufferedReader breader = new BufferedReader (new FileReader(fname));

data = new Instances (breader);
data.setClassIndex(data.numAttributes()-1);

//loading the model:
Classifier cls = (Classifier) weka.core.SerializationHelper.read("C:/Weka/CreditScoreModel.model");
Evaluation evaluation = new Evaluation(data);
evaluation.evaluateModel(cls, data);
```

Listing 1.Java code to load the model

V. MODEL EVALUATION

When the model is executed it displays the coefficients for each independent variable. Each coefficient represents the contribution its respective variable has on the overall credit score, which we may term as the Probability of Default (PD). The PD is the probability that a borrower will fail to repay their loan within the agreed period. The probability varies between 0 and 1 with 0 representing no credit risk and 1 representing high risk. Considering that grades A, B and C are granted loans, we shall represent these as P, for Pass. Grades D, E and F are denied loans and therefore will be represented as F, for Fail.

Figure 3 is an implementation of the independent variables used in the model together with their coefficients. In Figure 3, we note that some coefficients are positive whilst others are negative. The negative coefficients cause the PD to approach 0, indicating low risk, whilst the positive coefficients cause the PD to approach 1, indicating high risk. Therefore, company ownership, gearing, security, service ability and company net worth are high risk factors. These are the characteristics that, according to the model, lead to loan default. To get a loan approved, a customer must therefore have good company structure, low gearing, have adequate security, be able to service the loan and have a significant net worth value.

VI. CONCLUSION

One of the objectives of the study was to determine borrower characteristics that lead to loan default. With this in mind, we can assert that the optimised model will ensure that deserving loan applicants are granted loans. If the applicant meets the important characteristics determined by the model then they will be granted credit. The new model offers banks an opportunity to adopt a credit risk analysis model that not only focuses on minimising the percentage of customers who default but also identifies customers that are most profitable and those that that would be erroneously classified as high risk due to missing past history.

The model however does not consider credit bureau data. For example, in Zimbabwe the Financial Credit Bureau (FCB) is one such bureau. Access to this data improves the effectiveness of the model as it can then compare characteristics of customers in the credit bureau with those of the applicants.

Areas for future research include extending the research to make use of credit bureau data as well as bank customer data to enhance the performance of the model. The inclusion of macro-economic characteristics in the model would also significantly improve the model and the use of common scoring characteristics in the model would do away with characteristics that may otherwise not contribute significantly to the probability of default.

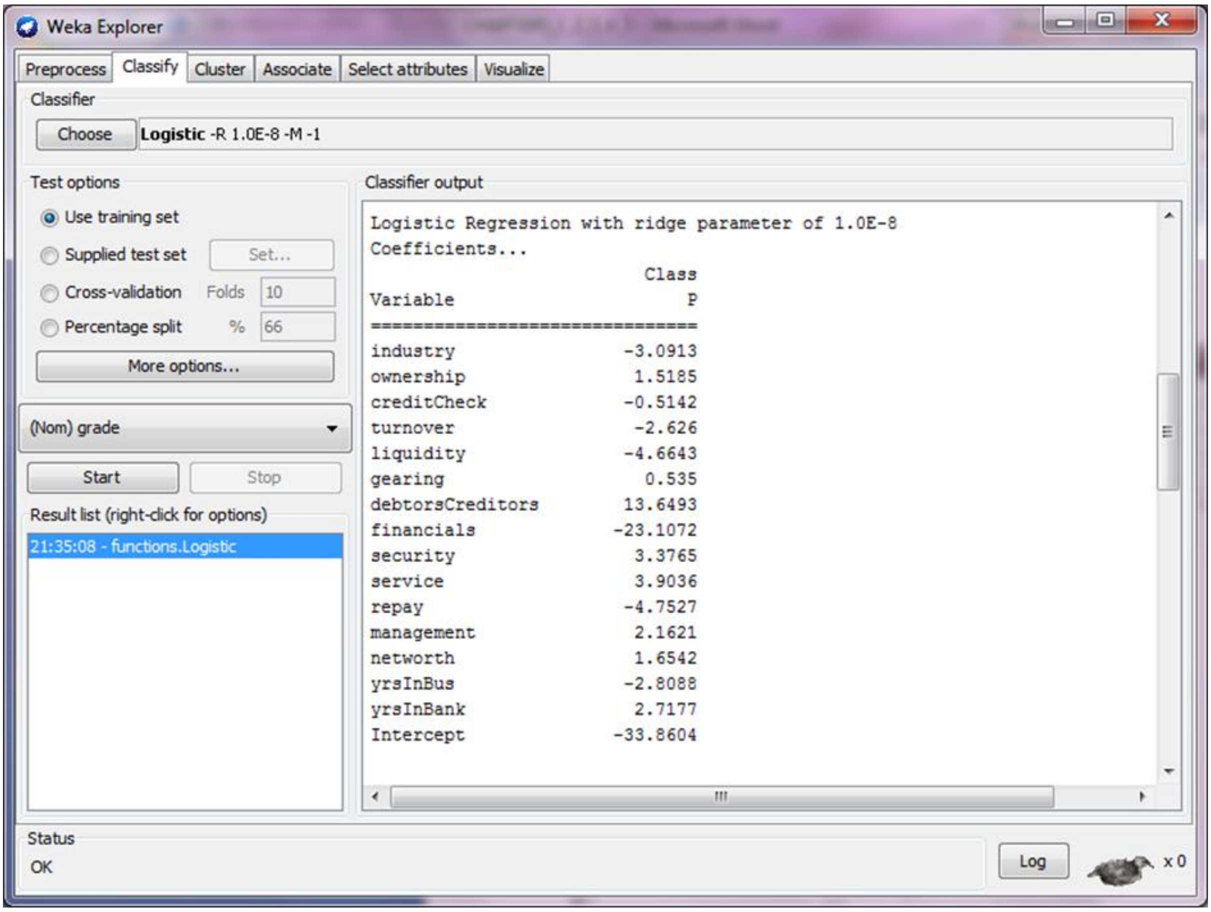

Figure 4 Building the logistic model

REFERENCES

[1] Harwood, J.R. Heifner, K. Coble, T. Perry, and A. Somwaru " Managing Risk in Farming: Concepts, Research and Analysis," Agricultural Economic Report No. 774. Market and Trade Economic Division and Resource Economics Division, Economic Research Service U.S. Department of Agriculture. March, 1999.

[2] Khan, T. and Ahmed, H. 2001, Risk Management: An Analysis of Issues in Islamic Financial Industry IRTI/IDB Occasional Paper, No. 5.

[3] Akkizidis, I. and Khandelwal, S.K. 2008, Financial Risk Management for Islamic Banking and Finance Palgrave Macmillan, First Edition.

[4] Al-Tamimi, H. and Al-Mazrooei M. 2007, "Banks' risk management: a comparison study of UAE national and foreign banks", The Journal of Risk Finance, Vol. 8

[5] Rose, P. (2002) Commercial Bank Management, 5th edition, Mc Graw-Hill/Irwin

[6] BIS, 2006, Basel Committee on Banking Supervision, International Convergence of Capital Measurement and Capital Standards, A Revised Framework. Technical Report, Bank for International Settlements, Basel.

[7] BIS, 2000, Bank for International Settlements,4http://www.bis.org/publ/bcbs75.htm accessed 12 February 2014.

[8] Yu, L., Wang, S., Lai, K. K. & Zhou, L. (2008) Bio-Inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machines, Springer.

[9] L.Yu, SH.Wang, K.K.Lai, (2009) An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring, European Journal of Operational Research

[10] Lieli, R. P., & White, H. (2010). The Construction of Empirical Credit Scoring Models Based on Maximization Principles. Journal of Econometrics.

[11] Ong, C., Huang, J., Tzeng, G. (2005) Building Credit Scoring Models Using Genetic

[12] Programming. Expert Systems with Applications 29 (1)

[13] Verstraeten, G. and Van den Poel, D. (2004). The impact of sample bias on consumer credit scoring performance and profitability. Journal of the Operational Research Society, 56(8):981–992.

[14] Saunders, 2000

[15] Crook, J.N., Edelman, D.B. & Thomas, L.C., 2007, Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447-1465.

[16] Lee, G., Sung, T.K. and Chang, N.(1999), Dynamics of modelling data mining: interpretive approach to bankruptcy prediction, journal of management information systems

[17] Thomas, L. (2000) A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers, International Journal of Forecasting, Vol. 16, USA.

[18] Lee, T., Chiu, C. Lu, C., Chen, I.(2002) Credit Scoring Using the Hybrid Neural Discriminant Technique. Expert Systems with Applications

[19] Tabachnick, B.G. and Fidell, L.S.(2006) Using multivatiate statistics. 5^{th} edn,Boston,Pearson Education

[20] Schreiner, M. Benefits and Pitfalls of Statistical Credit Scoring for Microfinance, Center for Social Development, Washington University in St. Louis, USA, 2004.

[21] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.