PROBABILITY OF DEFAULT MODELLING FOR A RETAIL LOAN PORTFOLIO VIA ANN AND LOGISTIC REGRESSION

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE UNDERGRADUATE DEGREE IN SCIENCE (MATHEMATICS OF FINANCE) OF THE UNIVERSITY OF BOTSWANA

BY
SHOKO TATENDA

Supervised by
Dr. V. GUMBO CBiiPro, CRCMP

May 2022

Contents

Abstract

Arguably, a cornerstone of credit risk modelling is the probability of default modelling. This dissertation aims to model and validate the probability of default (PD) of a personal (retail) loan by examining the relationship between loan characteristics and the likelihood of default. In line with that, several risk retail risk drivers were analyzed. The analysis was conducted via Artificial Neural Networks (ANN) and logistic(logit) regression approaches using SPSS software. The ANN model was applied to improve the accuracy and predictive strength of the logit model. The ANN shows that Loan term is the main driver of risk. The logit model results show that interest rate, loan term and accounts with other banks are the drivers of default, with loan term being the main driver of default. The longer the loan term, the greater the chances of default. The same result was also obtained from the ANN. The model results were tested for accuracy using the ROC Analysis, validity using the Kolmogorov-Smirnov (K-S) test and stability using the Population Stability Index(PSI). These tests confirm that the model is accurate, valid and stable. The findings of this study suggest that there is clear evidence of a relationship between borrower's characteristics and the probability of default. In this study, the higher the interest rate on a loan, the higher the chances of default. Moreover, long-term loans are found to be riskier than short-term ones.

Chapter 1

Introduction

In recent years the financial services industry has experienced a significant development in the understanding of credit risk. Several methodologies were proposed concerning the estimation of key risk parameters like default probabilities. Credit risk affects every financial contract; therefore, its assessment has received much attention from economists, bank supervisors regulators, and financial market practitioners. Profits realized on loans and loan products such as credit cards depend mostly on whether customers pay their debt obligations regularly or miss payments and default. The latter is considered to be a credit risk, which is the major source of risk for lending institutions. The Basel Committee on Banking Supervision (BCBS) states that credit risk default is a failure of a borrower or counterpart to meet its obligations in accordance with agreed terms [11].

The main focus of credit risk is to predict if a borrower will default on loan obligations in the future or to evaluate the probability of defaulting[1]. The PD can be estimated based on the borrower's credit bureau data, such as loan and personal characteristics. A lower predicted probability of default means better creditworthiness. Lending institutions normally set a cut-off threshold and approve credit to those customers that have the predicted probability of default (PD) less than the pre-defined threshold. For credit risk management purposes, the predicted probability will be combined with the other risk factors to determine the amount to be provisioned in case of a default. Provisioning is necessary to account for potential loan defaults and related

expenses and ensure the accurate assessment of a lending institution's financial health. These potential losses are called Expected Credit Losses (ECL) [1]. The PD is important not only for effective risk and capital management but also for the pricing of credit assets, bonds, loans, and more sophisticated instruments such as derivatives.

1.1 Problem Statement

The ability to predict a borrower's default is very important to lenders and investors especially banks and other lending institutions as it provides necessary information about the risk level of an obligor. This therefore means lenders and investors will be able to hedge themselves accordingly against the identified risks chief of which is credit risk.
The International Financial Reporting Standard 9 (IFRS 9) seeks to replace the International Accounting Standard 39 (IAS39) in calculating loan loss provisions or impairments. The IFRS 9 standard is an expected credit loss approach, while the IAS 39 approach is an incurred credit loss approach. It is often argued that IAS 39 promotes a reactionary approach to the management of credit losses that is difficult to budget and plan. Thus, there was a need to introduce a better and more predictive way of managing credit losses to minimize surprises, which banks were often not well prepared for. Under IFRS 9, expected credit losses (ECL) are calculated using the probability of default – PD, the portion of the credit that stands at risk should a default occur (Loss Given Default – LGD), and the expected levels of exposure at default (EAD). The ECL model requires the development of a probability of default (PD) based on observed previous defaults (the minimum period is 5 years).

The global financial crisis of 2008 exposed the weaknesses of the international financial system. It led to the promulgation of Basel III by the Basel Committee on Banking Supervision (BCBS), which includes new and modified regulations that strengthen and improve banking supervision regulations. These revised regulatory frameworks permitted the use of internal modelling approaches for certain risk categories. Basel II / III agreements require banks and lending institutions to report the yearly expected credit loss [10] a fu-

turistic value where PD is one of the major components in its calculation. This research is premised on the realization that there is a need to learn and apply the logistic regression approach to model and validate PD for credit risk quantification.

1.2 Objectives

- To understand the Artificial Neural Networks approach as a supplement to the logistic regression approach in PD modelling.
- To learn and apply the Artificial Neural Networks and logistic regression approach to model and validate PD for a personal loan portfolio.
- To determine the main retail default drivers of default and how they influence PD of retail loans.

1.3 Conceptual Clarifications

- A retail (personal) loan portfolio means a collection of all retail loans on a bank's book. A retail loan is a loan secured by an individual for personal use such as buying property, paying school fees, etc.
- Default means failure to fulfil an obligation, especially to repay a loan.
- Risk means a situation involving exposure to danger.
- Credit risk means the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations.
- A Model is an analytical approximation of reality that is built to simplify complex relationships, usually through an equation or a set of equations.
- Modelling means making a representation of something. Creating a tiny, functioning volcano is an example of modelling.

Chapter 2

Review of literature and concepts

2.1 Definition of default and Probability of Default

Credit risk measurement is based on assessments of the borrower's risk characteristics and the specific transaction type. The probability of default (PD) of a borrower or group of borrowers is the central measurable concept on which the Internal Rating Based (IRB) approach [8] is built. The PD of a borrower does not, however, provide the complete picture of the potential credit loss. Banks also seek to measure how much they will lose should a borrower default on an obligation. This is contingent upon two elements. First, the magnitude of likely loss on the exposure is termed the Loss Given Default (LGD) and is expressed as a percentage of the exposure. Secondly, the loss is contingent upon the amount to which the bank was exposed to the borrower at the time of default, which is Exposure at Default (EAD). These three components (PD, LGD, EAD) are called Basel II risk parameters [8], and they combine to provide a measure of expected intrinsic loss called Expected Credit loss(ECL), which is a product of these three components [11]. According to [9], a default is said to have occurred regarding a particular obligor(borrower) when one or more of the following events has taken place.

- It is determined that the obligor is unlikely to pay its debt obligations (principal, interest, or fees) in full;
- A credit loss event associated with any obligation of the obligor, such as a charge-off, specific provision, or distressed restructuring involving the forgiveness or postponement of principal, interest, or fees;
- The obligor is past due more than 90 days(more than 3 successive payments) on any credit obligation or;
- The obligor has filed for bankruptcy or similar protection from creditors.

The numerical quantification of the propensity to default over a particular time horizon is termed Probability of Default (PD). It estimates the likelihood that a borrower will be unable to meet his/her debt obligations.
PD is used in a variety of credit analyses and risk management frameworks. Under Basel II, it is a key parameter used in calculating economic or regulatory capital for a banking institution.

2.2 Retail Loan PD Approaches

There are several approaches which are used to estimate the PD of a personal loan portfolio. The overview is focused on statistical methods and includes parametric models such as Vintage analysis [4], Survival Analysis [5], Markov chains [16], Discriminant Analysis [3] and Logistic Regression Model [6] among others. The latest developments in Machine learning models, such as Artificial Neural Networks and Decision trees, are also used to estimate PD.

2.2.1 Vintage Data Analysis

Vintage Data Analysis is a technique of evaluating the credit quality of a loan portfolio by analyzing net charge-off in a given loan pool where the loans share the same origination period. In credit risk the term Vintage

refers to the time when account was opened, it could be month or quarter. This method allows to calculate the performance and losses of a portfolio in different periods of time after the loan was granted. Performance can be measured in the form of cumulative charge off-rate, proportion of customers 30/60/90 days past due, utilisation ratio, average balance, etc. The method is widely used in the analysis of retail credit cards and mortgage portfolios. It is also one of the several methodologies financial institutions are using for the current expected credit loss model.
Vintage model has been criticized for its simplicity in that it merely uses estimates of financial ratios to estimate default probabilities; it does not consider the personal borrower's characteristics, which are the major drivers of default for retail borrowers. It is also difficult to validate the vintage data model.

2.2.2 Markov Chains

A Markov chain is a mathematical model that predicts the next state based solely on the previous event state. In credit risk, the model is used to estimate transition rates into default state. The predictions generated by the Markov chain are as good as they would be made by observing the entire history of that scenario. It is mostly used to model the term structure of credit spreads in credit scoring [16].

The model provides transition rates from one state to the other state based on some probability conditions. One characteristic that defines the Markov chain is that no matter how the current state is achieved, the future states are fixed, and then the possible outcome of the next state solely depends on the current state and the time between the states. An absorbing state Markov chain is a Markov chain in which every state can reach an absorbing state. An absorbing state is a state that, once entered, cannot be left, for example, death. The Markovian approach has been criticized in that it predicts the PD of moving into the state (default state), not the exact PD based on what drives PD.

2.2.3 Discriminant Analysis

Discriminant analysis is a default prediction methodology that has been widely used since the work of Beaver who introduced the Univariate Discriminant Model (UDA) and Altman who introduced a Multi-discriminant Model (MDA).
These studies build reduced-form default prediction models using discriminant analysis and provide ordinal rankings of default risk by generating credit scores.
UDA uses a single ratio discriminant analysis in which the default rate is investigated by one ratio in a discrete time. [7] used a dichotomous classification test to identify financial ratios for corporate failure prediction. The model has an easy application and is favourable to studies evaluating the significance of single variables such as cash flow and return on assets [28]; thus, it is mainly used in the corporate probability of default modelling where companies seek to quantify their risk exposure based on financial ratios. UDA was susceptible to problems of endogeneity and omitted variables bias since the correlation between ratios was neglected. The MDA Model pioneer studies are said to investigate a company's credit risk using accounting ratios by [3],[18] and [14]. The original Z-score formula was represented as;

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5 \tag{2.1}$$

where
$X1$=working capital / total assets.
$X2$=retained earnings / total assets.
$X3$=earnings before interest and taxes / total assets.
$X4$=market value of equity/book value of total liabilities.
$X5$=sales / total assets.
However, the Z-score made is criticised for several reasons, being that:

- The assumptions of the MDA accuracy are contested in many studies. The assumption that all variables used to determine failure follow a normal distribution is criticised in practice, as most financial ratios are non-normal [29].

- The MDA is a static model that only predicts corporate default one step ahead and does not incorporate time [20]. This results in failure to capture the risk trend over time.

2.2.4 Survival Analysis

Survival analysis is a statistical area that deals with survival data analysis. The survival data can be collected in medical or reliability studies, for example, when a deteriorating system is monitored and the time until the event of interest is recorded. Credit risk data are very similar to survival data. Survival analysis is a modelling technique for time-to-event or duration data. The model considers the life course of a loan where at each point in time, the loan may enter one of a number of mutually exclusive states, such as performing, default and prepayment. With the passage of time, the loan moves between these states (or it remains static). It is likely that the loan will start in the performing state and later stay in the performing state or move into either default or prepayment. The survival model's point of interest is the length of time the loan spends within the performing state, in other words, how long the loan survives before it defaults or prepays [1]. The model seeks the relationship between loan status and the passage of time, along with other explanatory variables.

[12] proposed the following hazard model as a common formulation for survival analysis data

$$h(t) = h_0(t)e^{(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k)} \tag{2.2}$$

where $h(t)$ is the hazard rate or the conditional probability that a loan survives until time t but fails during the next time interval; time t represents age of loan; $h_0(t)$ is the baseline hazard, which captures the shape of the hazard function and summarizes how the probability of loan termination (either default or prepay) changes over time; $X_1, X_2, \cdots, X_k$ are explanatory variables that also influence risk of loan termination; and $\beta_1, \beta_2, \cdots, \beta_k$ are coefficients that measure the impacts of the explanatory variables on the hazard rate.

Survival analysis is one of the alternative approaches to logistic regression that have not been extensively explored; selected studies include [30], [24].

The survival analysis method is well accepted in studies of default probability because it matches the life course of a loan and its termination process. The estimated output provides forecasts of default probabilities as a function of time (the loan age) and other default determinants. It models both the default probability and time dependence of the probability, which is an advantage over the logit model. In a logistic model, the predicted probability has a fixed time horizon; to have a prediction for a different time horizon, one needs to revise the loan sample and repeat the estimation process. Survival analysis can estimate default risk for any time horizon [5]. However, the survival model has been criticised in that it uses the hazard rate, which is difficult to interpret from a business point of view, and the estimation is carried out using partial maximum likelihood without having to define the base hazard. It is less common among practitioners [1].

2.2.5 Logistic Regression Analysis

Regression is a technique to find the association between explanatory variables and a dependent variable. Logistic regression is a generalised linear model [23], [27] technique that allows one to predict discrete binary outcomes (e.g. death/no death, default/non-default, rain/no rain). The empirical model to be estimated is known as the Logit and can be written as:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mu \tag{2.3}$$

In terms of credit risk variables, Z is given by

$$Z = \beta_0 + \beta_1 * Income + \beta_2 * LoanAmount + \beta_3 * MaritalStatus + \beta_4 * Age + \beta_7 * Gender + \cdots + \mu \tag{2.4}$$

Then the probability of default is given by [22]

$$P(Z = 1|X1, ..., Xk) = \frac{1}{1 + \exp(-Z)}; -\infty < Z < \infty; 0 \leq PD(Z) \leq 1 \tag{2.5}$$

which can also be transformed using the logit distribution to

$$Logit(P(Z = 1|X_1, .., X_k)) = Z; \tag{2.6}$$

Points to note about the logistic regression model

• $X_1, X_2, \cdots, X_k$ are explanatory variables which are retail default drivers in credit risk analysis, and the variable μ is independent of the borrower's characteristics.

• $\beta_1, \beta_2, \cdots, \beta_k$ are coefficients that measure the impacts of the explanatory variables while β_0 is the intercept. Each of the regression coefficients describes the size of the contribution of that risk factor. A positive regression coefficient means that that risk factor increases the probability of the outcome. In contrast, a negative regression coefficient means that the risk factor decreases the probability of that outcome.

• A large regression coefficient means that that risk factor has a strong influence on the probability of that outcome. In contrast, a near-zero regression coefficient means that that risk factor has little influence on the probability of that outcome.

• The variable Z represents the total exposure to the set of risk drivers. In a different way, Z is a linear (non-linear) combination of the risk drivers.

• The graph of $PD(Z = 1)$ is a Sigmoid function, which gives the probability of default.

[26] points out the following advantages of using logistic regression for the construction of PD models:

• The generated model takes into account the correlation between variables, identifying relationships that would not be visible and eliminating redundant variables;

• It considers the variables individually and simultaneously;

• The user may check the sources of error and optimize the model.

In the same text, the author further identifies some disadvantages of this technique:

• The logit model has been argued to be sensitive to the multi-collinearity problem.

• It is a sample-based model; hence, this exposes biasedness in the prediction of a risk sample selection.

• In the case of many variables, the analyst must pre-select the most important ones based on separate analyses.

2.2.6 Artificial Neural Networks (ANN)

ANN are computational techniques that present a mathematical model based upon the neural structure of intelligent organisms who acquire knowledge through experience. The model mimics the human brain.
It was only in the eighties that, because of the greater computational power, neural networks were widely studied and applied. [25] underlines the development of the backpropagation algorithm as the turning point for the popularity of neural networks. An ANN model processes certain characteristics and produces replies like those of the human brain. The neural networks transform information from variables to predict the probability of default. Artificial neural networks are developed using mathematical models.

Previous studies have shown the use of ANN on the Fisher discriminant analysis and probabilistic neural networks to have the best prediction [21]. The method has been used to predict the bankruptcy and default risk of a firm. If applied to PD modelling, ANN approach has one main advantage over other models in that it predicts which variable(s) influences PD the most. However, the ANN model is argued to be a black box on its inputs and processing. This means that the contribution of each of the variables could not be shown directly from neural networks [25]. In other words, it does not provide the coefficient of any risk driver; therefore, it is not easy to calculate default probabilities.

2.2.7 Decision Trees

The Decision Trees predict the default risk using a technique where data is partitioned into sub-classes and then recursively replaces each subset with decision tree nodes until the tree contains unique risk outcomes that bring default or non-default. Even so, decision trees are argued to be a forward selection method and hence may lose some important risk drivers.

Chapter 3

Methodology

3.1 Model Build Process

3.1.1 Data Utilised

Historical data was sourced from Bank X and validated for model development. The retail portfolio that was used in this study is a sample of 925 distinct loan accounts originated in 2012, and the information for each loan is collected monthly over a period from January 2013 to May 2018. Table 3.1 below shows the summary of the data utilised in model development

Table 3.1: Data Summary

Submitted	924
Cleansed Out	70
Percentage Cleansed Out	7.6%
Remaining	854
Non Defaulting	408
Defaulting	446
Default Rate	52.22%
Max Loan Amount	82,591.00
Min Loan Amount	2,500.00

Table 3.2 provides the field names used during the data collection process.

Table 3.2: Retail Credit Risk Drivers considered

Variable	**Units of Measurement**
Loan Amount	US$
Current Borrowings	0 -No, 1 -Yes
Interest Rate	Number
Loan Term	Number in Months
Other Debt	0 -No, 1 -Yes
Age	0 -No, 1 -Yes
Client Type	0 -Staff, 1-Retail
Account with other banks	0 -No, 1 -Yes
Property Ownership	See Table 3.3
Gender	0 - Male, 1 -Female
Number of Dependants	Number
Marital Status	See Table 3.4
Income	US$
Previous Borrowings	0 -No, 1 -Yes
Default_History	0 - Non-Default,1 - Default

Table 3.3: Property Ownership

Status	**Dummy**
Owned	0
Rented	1
Renting	2
Staying with parents	3
Other	4

Table 3.4: Marital Status

Status	**Dummy**
Single	0
Married	1
Widowed	2
Other	3

3.1.2 Observation Period

The observation period for the model build was from 2013 to 2017 providing enough observation period for modeling purposed [8],[11].

3.1.3 Default Definition

The following default definition consistent with the bank reference definition was used:
• accounts written off,
• The obligor is past due more than 90 days(more than 3 successive payments) on any credit obligation or
• borrower bankruptcy.

3.1.4 Data Cleansing

An intense exercise was carried out in Microsoft Excel (horizontal and vertical cleansing) to ensure that the data to be used is arranged in the correct

and desirable format. The data cleaning process was done to ensure that:
• the sample data is free of (obvious) mistakes, and empty rows and columns were removed.
• the data set comprises only homogeneous observations and dummy variables were introduced for qualitative risk drivers such as marital status, residence status, etc.

3.2 Diagnostic tests

3.2.1 Stationarity Tests

Stationarity means that a process's statistical properties that create a time series are constant over time. This statistical consistency makes distributions predictable, enabling forecasting, and is an assumption of many time series forecasting models. To ensure stationarity in the data utilized, numerical variables, which consist of large numerical values (Age, Age-Squared, Loan Amount, Income), were log-normalized.

3.2.2 Multicollinearity tests

After verifying that the underlying assumptions of a logistic regression are valid, the model building process can began. However, although typically a huge number of potential input variables are available when developing a model, from a statistical point of view, it is not advisable to enter all these variables into the logit regression. If highly correlated variables are included in the model, the estimated coefficients will be significantly and systematically biased. Hence, it is preferable to pre-select the most promising explanatory variables by means of the univariate power of and the correlation between the individual input variables. In the Microsoft Excel workbook, the calculations were made, and the appropriate data arrangement was made. This is where most of the data collection and cleansing was done. A test known as the MULTICOLLINEARITY TEST was still carried out in the Microsoft Excel workbook. This test is done to analyze the correla-

tion among the independent variables. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results. The multicollinearity test was capped at a 0.8 correlation cap. Variables with a correlation coefficient higher than 0.8 require further scrutiny.

3.3 Splitting Data

Following data cleansing and diagnostic tests, the available data is divided into one development (Dev Sample) and one validation (Val Sample) by randomly splitting the whole data into two sub-samples. The first one, which typically contains the bulk of all observations, is used to estimate rating models, while the remaining data is left for an out-of-sample evaluation. When splitting the data, it should be ensured that all observations of one client belong exclusively to one of the two sub-samples and that the ratio of defaulting to non-defaulting clients is similar in both data sets. Data cleansing was performed on the original 924 loans to remain with 854 loans fit for modelling purposes. The data was divided into two sets: 70% of 824 = Dev Sample = 598; 30% of 824 = Val Sample = 256.

Table 3.5: Dev and Val Samples

	Dev Sample	**Val Sample**	**Total**
Defaults	312	134	**446**
Non Defaults	286	122	**408**
Total	**598**	**256**	**854**

3.4 Model Selection

3.4.1 Artificial Neural Networks

ANN are computational techniques that present a mathematical model based upon the neural structure of intelligent organisms. ANN predicts which risk

drivers have the most influence in predicting default. For highly correlated variables, ANN depicts which variable drives PD the most but does not give the magnitude of the influence. In this study, ANN was applied as a supplement to the logistic approach to model the default probabilities. The Multilayer Perceptron method in SPSS software was used to determine which variables to include in the logistic regression model building based on their percentage of importance. The Multilayer Perceptron (MLP) procedure produces a predictive model for one or more dependent (default history) variables based on the values of the predictor variables (risk drivers).

3.4.2 Logit Regression

Logit regression attempts to model the relationship between some explanatory variable(s), in this case, retail risk drivers and a response variable (Default status) by fitting a linear equation to observed data. The model was developed in SPSS software using the Backward Wald method. The model can be expressed as:

$$Z = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \mu \tag{3.1}$$

where Z is dependent variable $X_1, X_2, ..., X_k$ are explanatory variables β_0 is the intercept term $\beta_1, ..., \beta_k$ coefficients of explanatory variables μ is the random term. The regression model attempts to find the values of β_0 and β_i that minimise the error in the value of Z. This technique has been used for the following reasons:

- It is a common industry technique known to produce reasonable results;
- It is comparatively easy to calculate the values of the parameter estimates
- The values of the parameter estimates can be easily interpreted in terms of their effect on the value of the dependent variable.

Given the above, the probability of default is then given by

$$PD(Z) = \frac{1}{1 + exp(-z)} \tag{3.2}$$

Chapter 4

Analysis of data and Presentation

4.1 Multicollinearity Results

A correlation matrix of explanatory variables was used to test whether multicollinearity is present or not. Table 4.1 below shows that the risk drivers are free from multicollinearity.

Table 4.1: Correlation Matrix

	CURRENT BORROWINGS	*LN_INTEREST RATE*	*LN_LOAN AMOUNT*	*LN_LOANTERM*	*PREVIOUS BORROWINGS*	*LN_AGE*	*LN_AGE_SQUARED*	*CLIENT TYPE*	*ACCOUNTWITHOTHERBANKS*	*PROPERTY OWNERSHIP*	*GENDER*	*NUMBEROFDEPENDANDS*	*MARITAL STATUS*
CURRENT BORROWINGS	1												
LN_INTEREST RATE	0.036104	1											
LN_LOAN AMOUNT	0.029031	-0.00958	1										
LN_LOANTERM	-0.01918	0.037505	0.084714	1									
PREVIOUS BORROWINGS	0.442259	0.080259	0.030973	-0.02983	1								
LN_AGE	-0.0305	-0.00196	0.060529	-0.10356	-0.02879	1							
LN_AGE_SQUARED	-0.0305	-0.00196	0.060529	-0.10356	-0.02879	1	1						
CLIENT TYPE	-0.50475	-0.00241	-0.03231	0.035482	-0.36025	0.087322	0.087322	1					
ACCOUNTWITHOTHERBANKS	0.03844	-0.01971	-0.30648	0.121275	-0.10323	-0.02506	-0.02506	0.041007	1				
PROPERTY OWNERSHIP	-0.02105	-0.01669	-0.00403	0.328989	0.039026	-0.12697	-0.12697	0.047145	-0.35882	1			
GENDER	0.017227	0.016983	-0.02522	0.095224	-0.017	-0.02166	-0.02166	-0.05057	0.034909	0.062672	1		
NUMBEROFDEPENDANDS	-0.04595	-0.00353	0.034804	0.022125	-0.03453	0.208386	0.208386	0.05909	0.05273	-0.06884	-0.17423	1	
MARITAL STATUS	-0.07196	-0.07241	0.074264	-0.06464	0.01409	0.216382	0.216382	0.104724	-0.06673	-0.01784	-0.40579	0.209082	1

Age squared
The variable age squared is incorporated in building the model because it is generally believed that the behaviour of human beings will not increase to infinite with age but will follow a quadratic function that is concave-shaped. According to theory, the coefficient of this variable is supposed to be positive because the behaviour of old people conforms to that of young ones. This is also complemented by The reason that old people have low incomes is that the majority are pensioners, which increases the probability of defaulting on loans.

4.2 ANN Model Results

The ANN was run in SPSS software based on the combined Dev and Val samples. Table 4.2 and Figure 4.1 shows the results of the importance and normalized importance of risk drivers.

Table 4.2: Independent Variable Importance

	Importance	Normalized Importance
CURRENTBORROWINGS	0.037	13.6%
LN_INTEREST RATE	0.169	61.6%
LN_LOAN AMOUNT	0.045	16.6%
LN_LOANTERM	0.275	100.0%
PREVIOUS BORROWINGS	0.025	8.9%
FACILITYRESTRUCTURINGID	0.014	5.2%
LN_AGE	0.057	20.7%
LN_AGE_SQUARED	0.061	22.4%
CLIENTTYPE	0.107	39.0%
ACCOUNTWITHOTHERBANKS	0.038	13.7%
PROPERTYOWNERSHIP	0.048	17.5%
SECTORPERFORMANCEID	0.021	7.6%
GENDER	0.012	4.5%
NUMBEROFDEPENDANDS	0.048	17.4%
MARITALSTATUS	0.042	15.4%

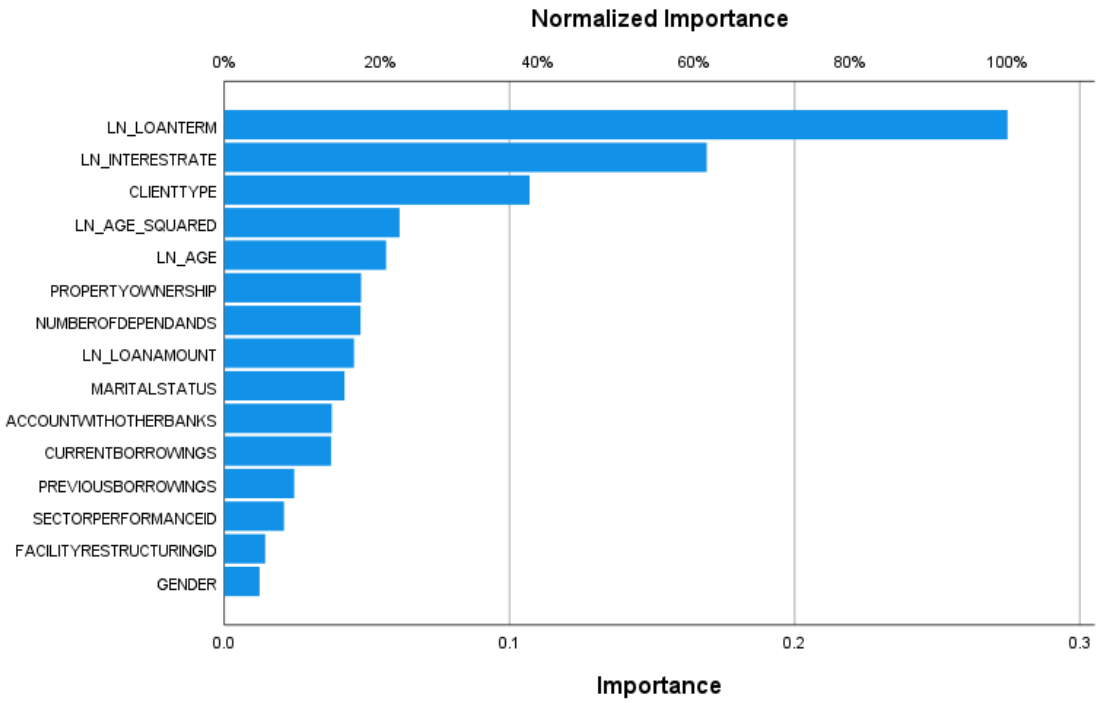

Figure 4.1: Importance and Normalized Performance

Analysis of results

ANN shows which variable influences default the most. ANN was applied to pick variables to include in a model building using the logistics regression. Figure 4.1 shows that loan term has the most influence in predicting default. Interest rate, Age and Age squared also have a high influence on default. Normalized importance was capped at 10%, and therefore, drivers who have normalized importance less than the cap were not included in the model building using the regression analysis since they have negligible influence in default prediction. ANN here was used to improve the accuracy and predictive strength of the logit model by choosing which drivers to include in the model building using the logit regression.

4.2.1 ANN Model Accuracy

The accuracy of the ANN model in selecting which variables have the most influence in predicting default was tested using the Receiver Operating characteristics curve. ROC curve is a two-dimensional graph that visually predicts the performance of the ANN model. The closer the curve to the left-hand border and then the top border of the ROC space, the more accurate the test. Figure 4.2 shows the ROC curves for default and non-default status borrowers. Therefore, it was concluded that the ANN is very accurate in terms of selecting variables that have the most influence in predicting default.

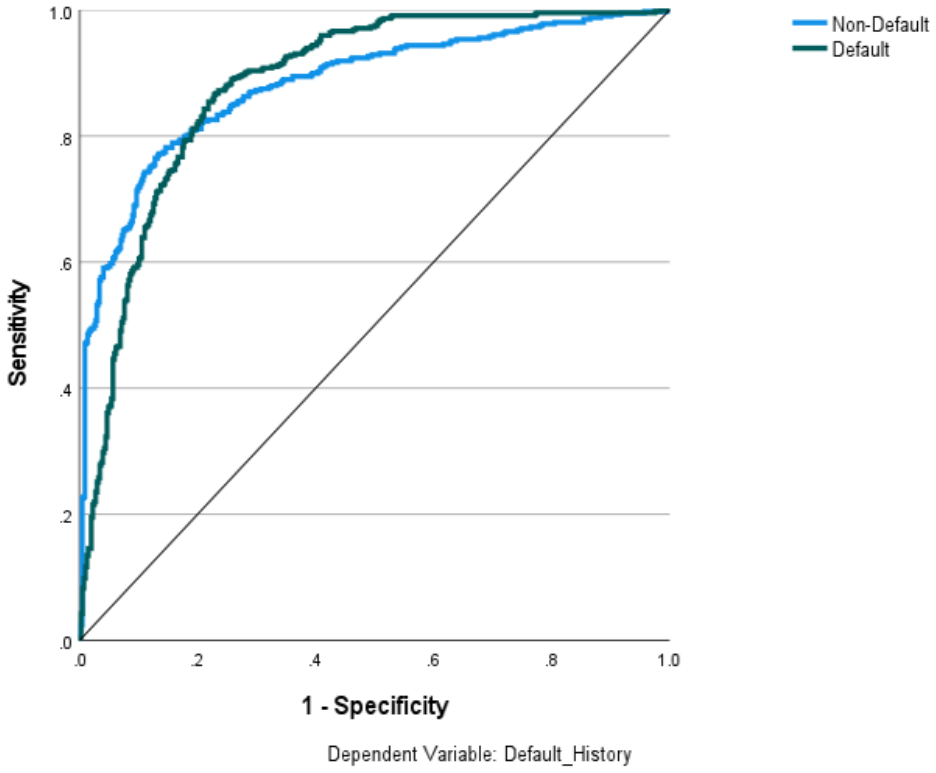

Figure 4.2: Combined ROC Curves

The area under the ROC curve is also a test of accuracy. The closer the area

is to one, the more accurate the model. Table 4.3 shows that the model is 89.6% accurate in terms of predicting default and non-default borrowers.

Table 4.3: Area under ROC curve

		Area
Default_History	Non-Default	0.896
	Default	0.896

4.3 Logit Model Results

4.3.1 The Z Model

Taking the confidence interval of 95% and running the model, the regression results are given in Table 4.5. Any variable whose significance is less than 0.05 is a candidate of the Z model and is therefore critically important [6]. All the variables in Table 4.5 are significant at 5% level.

Table 4.4: Regression Coefficients(the Z Model)

Variables in the Equation						
		B	S.E.	Wald	df	Sig.
Step 10[a]	LN_INTEREST RATE	0.446	0.071	39.627	1	< 001
	LN_LOANTERM	2.289	0.646	12.551	1	< 001
	ACCOUNTWITHOTHERBANKS	0.781	0.311	6.285	1	0.012
	Constant	-12.785	3.133	16.654	1	< 001

From Table 4.5 above, the resultant logistic regression model was deduced to be given by

$$\begin{aligned} Z &= Intercept + \sum(Coeficient \times RiskDriver) \\ &= -12.785 + 0.446 * LN\_INTERESTRATE + 2.289 * LN\_LOANTERM \\ &+ 0.781 * ACCOUNTWITHOTHERBANKS \end{aligned} \tag{4.1}$$

Then the PD model is given by

$$PD(Z) = \frac{1}{1 + exp(-Z)} \tag{4.2}$$

where Z is given by (4.1)

Table 4.5: Regression Coefficients (the Z Model)

Variables in the Equation					
		B	S.E.	Wald	Sig.
Step 10[a]	LN_INTEREST RATE	0.446	0.071	39.627	< 001
	LN_LOANTERM	2.289	0.646	12.551	< 001
	ACCOUNTWITHOTHERBANKS	0.781	0.311	6.285	0.012
	Constant	-12.785	3.133	16.654	< 001

Analysis
Table 4.5 above shows the results of the logistic model to identify personal characteristics that influence the probability of default of an individual borrower. The results showed that all the risk drivers shown in the table are strong, significant default predictors at 95% confidence level and have positive have a positive correlation with PD; hence therefore, they all increase the Probability of default for an individual borrower. Loan Term has the highest coefficient of 2.289. This implies that a monthly increase in loan tenure will lead to a 2.289 increase in the variable Z, hence an increase in the PD of the borrower; it was concluded that in this portfolio, loan term is the major risk driver and longer the term, the bigger the probability of default. The ANN Model also confirms the same result.

4.3.2 Hosmer and Lemeshow goodness of fit test

The Hosmer-Lemeshow test is a statistical test for the goodness of fit for the logit model. The data was divided into 8 groups defined by increasing order of estimated risk as shown in Table 4.6. The observed and expected number of default/non-default cases were calculated, and a Chi-Squared statistic was calculated as follows

$$\chi_{hl} = \sum_{g=1}^{n} \frac{(O_g - E_g)^2}{E_g(1 - \frac{E_g}{n_g})} \tag{4.3}$$

where O_g, E_g and n_g are the observed events, expected events and number of observations for the g-th risk octile group, and n is the number of groups.

Table 4.6: Contigency H-L Test

Contingency Table for Hosmer and Lemeshow Test						
		Default_History = Non Default		Default_History = Default		Total
		Observed	Expected	Observed	Expected	
Step 10	1	93	71.756	2	23.244	95
	2	42	36.749	16	21.251	58
	3	18	20.754	23	20.246	41
	4	22	64.875	112	69.125	134
	5	20	25.747	39	33.253	59
	6	30	29.789	54	54.211	84
	7	32	21.594	33	43.406	65
	8	29	14.737	33	47.263	62

The test statistic follows a Chi-squared distribution with $n - 2$ degrees of freedom. A large value of Chi-squared (with a small p-value < 0.05) means that there is a greater relationship between default and its predictors'. The Chi-squared statistic is 111.319, and the confidence level is < 0.001, which means the model has a good fit.

Table 4.7: Hosmer and Lemeshow Test

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
10	111.319	6	< 0.001

4.3.3 Model Classification

An intuitively appealing way to summarize the results of a fitted logistic regression model is via a classification table [6]. Table 4.8 was the result of cross-classifying the outcome variable Z with a dichotomous variable whose values are derived from the estimated logistic probabilities.

Table 4.8: Classification Table

Observed			Predicted		
			Default_History		Percentage Correct
			Non Default	Default	
Step 10	Default_History	Non Default	153	133	53.5
		Default	41	271	86.9
	Overall Percentage				70.9

From Table 4.8, the overall rate of correct classification was estimated as:

$$Overall\ Percentage = \left(\frac{153 + 271}{598}\right) 100 = 70.9\% \tag{4.4}$$

Thus, the model is 70.9% accurate in distinguishing between default and non-default borrowers, thu which is a fairly good model.

4.4 Model Validation

Model validation is a critical activity to verify that models are working properly as intended and that model usage is in line with business objectives and expectations. A regular model tracking and validation process can ensure that consistent and optimal model-based decisions are being made. It can also serve as an early warning system for identifying when a change may be necessary, whether it be an adjustment to the score cut-off strategy or a full model redevelopment. Under Base III, it is mandatory for banks and lending institutions to regularly validate their models to guard against model risk. Model Validation refers to policies and procedures that must be in place to validate the model appropriately. In this study, outcome 'analysis form' or back-testing was performed, and risk and control assessments were evaluated against actual loss data to determine the validity of the model.

4.4.1 Receiver Operating Characteristics (ROC)

For every possible cut-off point or criterion value was selected to discriminate between defaulter and non-defaulter populations, there will be some defaulters correctly classified as 1 (TP = True Positive fraction), but some will be classified as non-defaulters(FN = False Negative fraction). On the other hand, non-defaulters will be correctly classified as 0 (TN = True Negative fraction), whilst some will be classified as 1 (FP = False Positive fraction). Based on this information, the following statistics were defined:

- Sensitivity: the probability that a test result will be positive when the borrower defaults(true positive rate, expressed as a percentage).
- Specificity: the probability that a test result will be negative when the borrower doesn't default (true negative rate, expressed as a percentage).

The ROC Curve was used to test the accuracy of the model. ROC curve is a two-dimensional graph that visually depicts the performance and performance trade-off of a classification model [25]. ROC curves are industry

standard methods for comparing two or more scoring algorithms [20]. In an ROC curve, the true positive rate (sensitivity) is plotted in function of the false positive rate (1-specificity) for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. It shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. Figure 4.3 shows the ROC of the model, and it was concluded that the model is fairly accurate.

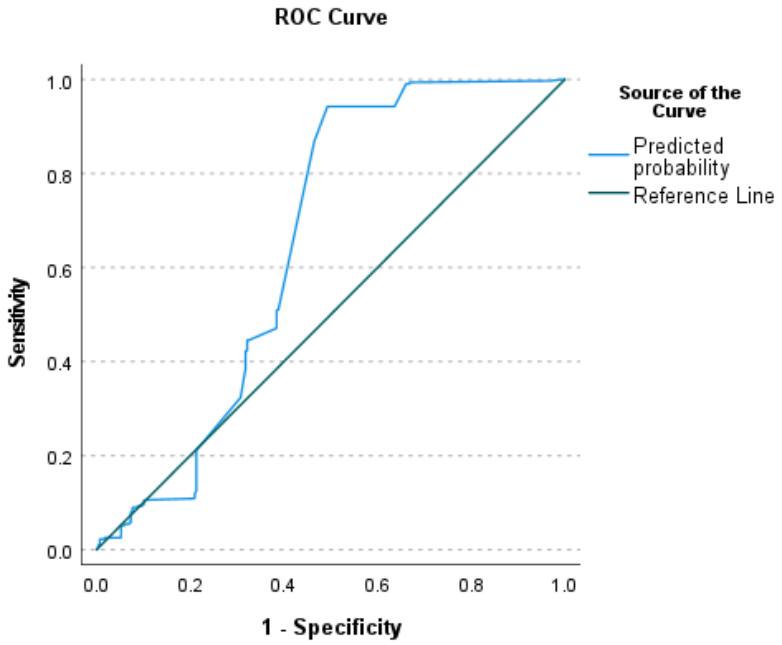

Figure 4.3: ROC Curve

Area Under the ROC Curve (AUROC)
AUROC is a measure of test accuracy. The accuracy of a rating model's performance increases the steeper the ROC curve is at the left end and the closer the ROC curve's position is to the point (0,1). Similarly, the larger the area under the ROC curve (AUROC), the better the model. AUROC which ranges from 0 to 1 provides a measure of the model's ability to discriminate between defaulters and non-defaulters. The accuracy of the model given by the AUROC denoted by A is:

$$A = \int_0^1 HR(FAR)d(FAR) \tag{4.5}$$

where HR is the hit rate, and FAR is the False Alarm Rate. The hit rate is given by:

$$HR = \frac{H}{N_D} \tag{4.6}$$

where
H is the number of defaulters predicted correctly.
N_D is the total number of defaulters in the Development sample.
On the other hand, the false alarm rate FAR is given by:

$$FAR = \frac{F}{N_{ND}} \tag{4.7}$$

where
-F is the number of false alarms, that is, the number of non-defaulters that were classified as defaulters
-N_{ND} is the total number of non-defaulters in the Development sample sample.

A= 68.3% as shown in Table 4.9. The closer the ROC curve is to 1, the better the model. Table 4.9 shows that the model is 68.3% accurate in its ability to discriminate between defaulters and non-defaulters. This is an averagely good model.

Table 4.9: Area under ROC curve

Area Under the ROC CURVE
Test Result Variable; predicted pd
Area
0.683

4.4.2 Kolmogorov-Smirnov (K-S) Test

The aim was to determine if the Validation Sample data would give similar results as the ones found by using the Development(Dev) Sample. The Z found in (4.1) was used to determine the Z model for each borrower in the Validation (Val) Sample and its corresponding PD. If the outcome of two samples is similar, then the assumption is that the samples exhibit the same distribution. The two sample K-S tests were used to test for the validity of the Z model. The two-sample K-S test determines whether the two underlying one-dimensional probability distributions differ. The K-S statistic is defined below:

$$D_{N_1,N_2} = sup|F_{1,N_1}(x) - F_{2,N_2}(x)| \tag{4.8}$$

where $F_{1,N_1}(x)$ and $F_{2,N_2}(x)$ are the empirical distribution functions of the development and validation sample, respectively, and N_1 and N_2 are the sizes of the respective samples (a total number of borrowers in each sample) The Empirical distributions are given in the Table 4.10 and their respective graphs are shown in Figure 4.4.

The null hypothesis is that the two samples belong to the same distribution, and it is rejected at level α if:

$$D_{N_1,N_2} > C(\alpha)\sqrt{\frac{N_1 + N_2}{N_1 * N_2}} \tag{4.9}$$

Table 4.10: Empirical Distribution Function

		K-S Test						
		Development Sample			Validation Sample			
PD	Rating Classification	Freq	RelFreq	CumRelFreqDev	Freq	RelFreq	CumRelFreqVal	K-S
[0.00;0.10)	1	1	0.001672	0.001672	1	0.003906	0.003906	0.0022
[0.10;0.20)	2	11	0.018395	0.020067	9	0.035156	0.039063	0.0190
[0.20;1.30)	3	87	0.145485	0.165552	32	0.125	0.164063	0.0015
[0.30;0.40)	4	36	0.060201	0.225753	23	0.089844	0.253906	0.0282
[0.40;0.50)	5	101	0.168896	0.394649	29	0.113281	0.367188	0.0275
[0.50;0.60)	6	143	0.23913	0.633779	70	0.273438	0.640625	0.0068
[0.60;0.70)	7	141	0.235786	0.869565	66	0.257813	0.898438	0.0289
[0.70;0.80)	8	62	0.103679	0.973244	21	0.082031	0.980469	0.0072
[0.80;0.90)	9	15	0.025084	0.998328	5	0.019531	1	0.0017
[0.90;1.00)	10	1	0	1	0	0	1	0.0000
		598			**256**			**0.0289**

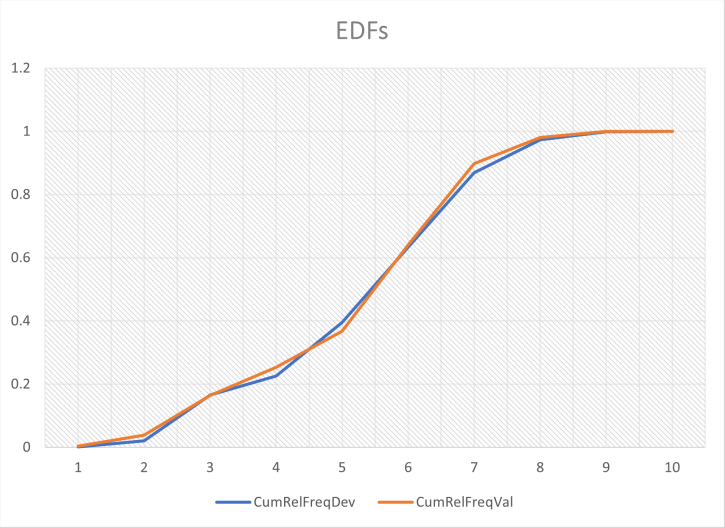

Figure 4.4: Combined ECDF for Development and Validation Samples

Critical Value

The critical value depends on the number of observations ($N_1 = 598$) and ($N_2 = 256$) of the two-sample distributions. The value of $C\alpha$ is given in Table 4.11 for each level of α. The critical value of α is given in Table 4.11

Table 4.11: Critical Value Table

α	0.1	0.05	0.025	0.01	0.005	0.001
$C(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

Table 4.12: Kolmogorov-Smirnov Statistic

D_{N_1,N_2}	**0.028872**
N_1	598
N_2	256
$C(\alpha)$	1.36
$D(\alpha)$	**0.101577**

In this study, the model was developed using 100% of the sample, that is $N_1 = 598$. For validation, a randomly selected of one-third of the total sample was used, that is N_2 =256 as seen in Table 4.12. Therefore at $\alpha = 0.05$ level of significance, $D(\alpha) = 1.36\sqrt{\frac{598+256}{598*256}} = 0.101577$. Since $D_{N_1,N_2} = 0..028872 < 0.101577 = D\alpha$, we fail to reject the null hypothesis. Therefore, this implies that the two samples come from the same distribution. This was also confirmed by the similarity between the combined ECDF for Development and Validation Samples shown in Figure 4.4. It was deduced that the model is valid and fit for purpose.

4.4.3 Population Stability Index

The Population Stability Index (PSI) was used to check for the stability of the model. PSI is a metric to measure how much a variable has shifted in distribution between two samples over time [13]. PSI was developed to monitor the change in the distribution of a score between a validation sample and a development sample. When a model deteriorates in performance, checking distributional changes can help identify possible causes. If at least one of the risk drivers has changed significantly to rebuild the model. A change in a variable distribution can be due to:

• Changes in macroeconomic climate.

• Changes in the source data, e.g., a switch to a different mailing source for marketing campaigns.

• Internal policy changes.

• Issues in data integration.

• Issues in programming, such as model implementation.

PSI is calculated as follows

$$PSI = \sum \left(A - B + LN \left(\frac{A}{B} \right) \right) \tag{4.10}$$

where:

A = Relative frequencies of Development Sample in %

B = Relative Frequencies of Validation Sample in %

Table 4.13: Population Stability Analysis

		Development Sample		Validation Sample				
Score Intervals	**Rating Classification/Decile**	**Freq**	**RelFreq% (A)**	**Freq**	**RelFreq% (B)**	**A-B**	**Ln(A/B)**	**PSI**
	1	1	0.001672241	1	0.003891	-0.00222	-0.84451	0.001874
	2	11	0.018394649	9	0.035019	-0.01662	-0.64384	0.010704
	3	87	0.14548495	32	0.124514	0.020971	0.155658	0.003264
	4	36	0.060200669	23	0.089494	-0.02929	-0.39649	0.011615
	5	101	0.168896321	29	0.11284	0.056056	0.40331	0.022608
	6	143	0.239130435	70	0.272374	-0.03324	-0.13017	0.004327
	7	141	0.235785953	66	0.256809	-0.02102	-0.08541	0.001796
	8	62	0.10367893	21	0.081712	0.021967	0.238097	0.00523
	9	15	0.025083612	5	0.019455	0.005628	0.254098	0.00143
	10	1	0.001672241	1	0.003891	-0.00222	-0.84451	0.001874
		598		**257**			**PSI**	**0.0647**

The PSI rule of thumb is given in Table 4.14. The PSI calculated is 0.067 which is less than 0.01 therefore the model is stable and should continue to be utilised.

Table 4.14: PSI Rule of Thumb

1	PSI < 0.1 - No change. You can continue using existing model.
2	$0.1 \leq$ PSI < 0.2 - Slight change is required.
3	PSI ≥ 0.2 - Significant change is required. Ideally, you should not use this model anymore.

Chapter 5

Conclusion, Recommendations and future steps

5.1 Conclusion

Credit risk modelling and analysis has shown to be an area of low knowledge and practice, especially here in Botswana, following the appropriate regulatory measures of best practice lending (Basel II/III) and best practice reporting(IFRS9) [2]. Thus, there is little to no publication that been made in the area of credit risk modeling for banks and other lending institutions in Botswana.

This study adopted the ANN and logit regression to model and validate the probability of default for a retail portfolio. The ANN model was used as a supplement to the logit model to improve the accuracy of the results and to pick variables with the most influence in terms of predicting default. The model was developed with default history representing the Z-score and retail risk drivers representing the independent variables. From this model, the drivers of default were shown to be Interest rate, Loan term and Account with other banks, all having a positive relationship with the probability of a borrower defaulting loan term was shown to be the main driver of default with a coefficient of 2.289 followed by account with other banks 0.781 and lastly Interest rate with a coefficient of 0.446 . Using the Backward Wald

method of logistic regression in SPSS, it was deduced that the model was 70.9% accurate in terms of distinguishing between defaulters and defaulters; therefore, it's a good model. The accuracy of the model developed was tested using ROC analysis, which showed that the model is 63% accurate, which is, on average, good. The K-S test was performed in Excel to check for the model's validity. It confirms that the Development Sample and Validation Sample give the same results based on the model and, therefore, the two samples belong to the same distribution; hence, the model is valid. PSI was also calculated to check for stability, and the results show that there is no significant change in the model when applying the model results to estimate PDs for the Validation Sample; therefore, the model should continue to be utilized. During the analysis of data, it was recognized that using the logit model inflicted difficulties as it is very prone and sensitive to multicollinearity problems, as stated by [17].

5.2 Recommendations

- ANN does not give the magnitude of the influence of default risk drivers in terms of predicting chances of default, it only shows which variables has most influence. It is also suggested to use ANN and the logit model to improve accuracy when building PD models.

- It was noticed that the logit model output different results when using different software, this is due to the condition of sensitivity to multicollinearity, hence it is suggested that the most optimal and sensitive software to correlation is used for these kinds of logit model.

- It is also suggested to use less number of retail risk drivers to provide more accurate predictions and to find the most significant drivers [6].

- Publications of financials for other companies in Botswana should be made public. This would allow a wide sample of data to be analyzed for the most realistic results. This would also provide those companies with forecasts that could be used to prepare better for the future.

5.3 Future Steps

The goal of this study was to model and validate the PD of a retail portfolio via the logistic regression approach. Future work will consider other methodological approaches, such as survival analysis, in detail and compare their results with the most commonly used logistic regression approach.

Bibliography

[1] Zhang, Qingfen,(2015) *Modeling the Probability of Mortgage Default Via Logistic Regression and Survival Analysis*. Open Access Master's Theses. Paper 541. https://digitalcommons.uri.edu/theses/541.

[2] H. Kably and V. Gumbo(2021) *Bank Distress Prediction Model for Botswana*. Asian Research Journal of Mathematics 17(2):47-59, 2021: Article no.ARJOM.66075.

[3] Altman, E.I. (1968). *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*. The journal of finance, 23(4), 589–609.

[4] Aijun Zhang (2009) *Statistical Methods in Credit Risk Modeling* :https://www.researchgate.net/publication/30864345.

[5] Andrija Durovic(2016) *Estimating Probability of Default on Peer to Peer Market-Survival Analysis Approach.* Journal of Central Banking Theory and Practice, 2017, 2, pp. 149-167 Received: 18 September 2016; accepted: 21 January 2017.

[6] Hosmer, D.W. and Lemeshow S (1989). *Applied Logistic Regression*. New York: John WIley & Sons,Inc.

[7] Beaver, W. H. (1966). "*Financial ratios as predictors of failure*". Journal of Accounting Research, pp. 71–111.

[8] Bernd Engelmann and Robert Rauhmeier *The Basel II Risk Parameters Estimation, Validation, Stress Testing – with Applications to Loan Risk Management.*

[9] Basel Committee on Banking Supervision (BCBS) (2004), Basel II: *International Convergence of Capital Measurement and Capital Standards*: A Revised Framework. http://www.bis.org/publ/bcbs107.htm.

[10] Basel Committee on Banking Supervision (1999), *Credit Risk Modeling: Current Practices and Applications, Bank for International Settlements.*

[11] Basel Committee on Banking Supervision (2001), *The Internal Ratings-Based Approach*, Bank for International Settlements.

[12] Cox, D. R. (1972), "*Regression models and life-tables (with discussion)*" Journal of Royal Statistics Society B, 34, 187-220.

[13] Bilal,(2018) *Statistical Properties of Population Stability Index* . Dissertations. 3208. https://scholarworks.wmich.edu/dissertations/3208.

[14] Blum, M. (1974). *Failing company discriminant analysis.* Journal of Accounting Research, 12(1), 1–25.

[15] Bharath, S.T., Shumway, T. (2008). *Forecasting default with the Merton distance to default model.* Review of Financial Studies,21(3), 1339–1369.

[16] Jarrow, R.A.; D. Lando and S. Turnbull. 1997. *A Markov model for the term structure of credit risk spreads.* Review of Financial Studies, Vol. 10, No. 2, 481-523.

[17] Balcaen, S., & Ooghe, H. (2006). *35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems.* The British Accounting Review, 38(1), 63-93.

[18] Deakin, E.B. (1972). *A discriminant analysis of predictors of business failure.* Journal of Accounting Research,10(2),167–179.

[19] Halperin, M., Blackwelder, W. C., & Verter, J. I. (1971). *Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches.* Journal of chronic diseases, 24(2-3), 125-158.

[20] Thomson, J. B. (1991). *Predicting bank failures in the 1980s. Federal Reserve Bank of Cleveland Economic Review*, 27(1), 9-20.

[21] Ping, C., Yang, L., & Tianshou, M. (1900). *Status and prospect of multi-well pad drilling technology in shale gas.*42(3), 1-7.

[22] Neter, J., Kutner, M. H., Nachtshein, C. J. & Wasserman, W. (1996) *Applied Linear Statistical Models.* Chicago: Irwin.

[23] Dobson, A. (1990) *An Introduction to Generalized Linear Models.* London: Chapman & Hall.

[24] Allen, LN and Rose, LC.(2006) *Financial survival analysis of defaulted debtors*, Journal of the Operational Research Society, , 57, 630-636.

[25] Fausett, L. (1994) *Fundamentals of Neural Networks. Englewood-Cliffs*: Prentice-Hall.

[26] Fensterstock, F. (2005) *Credit Scoring and the Next Step. Business Credit*, 107(3): 46-49. New York: National Association of Credit Management.

[27] Paula, G. A. (2002) *Modelos de Regressão com Apoio Computacional.* Book available in http://www.ime.usp.br/ giapaula/livro.pdf accessed in 12/05/2004.

[28] Shi, C., & Bhargava, B. (1998, September). *A fast MPEG video encryption algorithm. In Proceedings of the sixth ACM international conference on Multimedia* (pp. 81-88).

[29] Halperin, M., Blackwelder, W. C., & Verter, J. I. (1971). *Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches.* Journal of chronic diseases, 24(2-3), 125-158.

[30] Stepanova, Maria and Thomas, Lyn,(2002) *Survival analysis methods for personal loan data*, Operations Research, Vol. 50, No. 2, , pp. 277-289.

[31] Basel Committee on Banking Supervision (2005),*Studies on the Validation of Internal Rating Systems.*